fermentation

Article

A Statistical Workflow to Evaluate the Modulation of Wine Metabolome and Its Contribution to the Sensory Attributes

José Manuel Muñoz-Redondo [1,*], Belén Puertas [2], Gema Pereira-Caro [1], José Luis Ordóñez-Díaz [1], María José Ruiz-Moreno [1], Emma Cantos-Villar [2] and José Manuel Moreno-Rojas [1,*]

[1] Department of food Science and Health, Andalusian Institute of Agricultural and Fisheries Research and Training (IFAPA), Alameda del Obispo Avda, Menéndez Pidal, s/n, 14004 Córdoba, Spain; mariag.pereira@juntadeandalucia.es (G.P.-C.); josel.ordonez@juntadeandalucia.es (J.L.O.-D.); mariaj.ruiz.moreno@juntadeandalucia.es (M.J.R.-M.)

[2] Department of Food Science and Health, Andalusian Institute of Agricultural and Fisheries Research and Training (IFAPA), Cañada de la Loba, 11471 Jerez de la Frontera, Spain; mariab.puertas@juntadeandalucia.es (B.P.); emma.cantos@juntadeandalucia.es (E.C.-V.)

* Correspondence: josem.munoz.redondo@juntadeandalucia.es (J.M.M.-R.); josem.moreno.rojas@juntadeandalucia.es (J.M.M.-R.); Tel.: +34-666-306-651 (J.M.M.-R.); +34-671-532-758 (J.M.M.-R.)

Abstract: A data-processing and statistical analysis workflow was proposed to evaluate the metabolic changes and its contribution to the sensory characteristics of different wines. This workflow was applied to rosé wines from different fermentation strategies. The metabolome was acquired by means of two high-throughput techniques: gas chromatography–mass spectrometry (GC-MS) and liquid chromatography–mass spectrometry (LC-MS) for volatile and non-volatile metabolites, respectively, in an untargeted approach, while the sensory evaluation of the wines was performed by a trained panel. Wine volatile and non-volatile metabolites modulation was independently evaluated by means of partial least squares discriminant analysis (PLS-DA), obtaining potential markers of the fermentation strategies. Then, the complete metabolome was integrated by means of sparse generalised canonical correlation analysis discriminant analysis (sGCC-DA). This integrative approach revealed a high link between the volatile and non-volatile data, and additional potential metabolite markers of the fermentation strategies were found. Subsequently, the evaluation of the contribution of metabolome to the sensory characteristics of wines was carried out. First, the all-relevant metabolites affected by the different fermentation processes were selected using PLS-DA and random forest (RF). Each set of volatile and non-volatile metabolites selected was then related to the sensory attributes of the wines by means of partial least squares regression (PLSR). Finally, the relationships among the three datasets were complementary evaluated using regularised generalised canonical correlation analysis (RGCCA), revealing new correlations among metabolites and sensory data.

Keywords: data-integration; untargeted; multivariate analysis; metabolomics; chemometrics; rosé wines; yeasts; sequential inoculation

check for updates

Citation: Muñoz-Redondo, J.M.; Puertas, B.; Pereira-Caro, G.; Ordóñez-Díaz, J.L.; Ruiz-Moreno, M.J.; Cantos-Villar, E.; Moreno-Rojas, J.M. A Statistical Workflow to Evaluate the Modulation of Wine Metabolome and Its Contribution to the Sensory Attributes. *Fermentation* **2021**, 7, 72. https://doi.org/10.3390/fermentation7020072

Academic Editor: Sergi Maicas

Received: 23 March 2021
Accepted: 30 April 2021
Published: 5 May 2021

1. Introduction

Wine quality perception by consumers is a complex concept involving a sensory experience and consumers' expectation [1]. The sensory experience in wine tasting is related to its organoleptic attributes, such as aroma, taste and mouthfeel, which are further linked to its chemical composition [1,2]. In recent years, the advent of robust and high-throughput analytical techniques such as gas chromatography–mass spectrometry (GC-MS) and liquid chromatography–mass spectrometry (LC-MS) has prompted the study of the metabolic profile of foods and beverages. In this sense, untargeted strategies, that aim to simultaneously measure as many compounds as possible present in samples [3], have been successfully used to untangling the wine metabolome in the last few years [4–9]. However, the data generated by these analytical platforms are complex, with multi-scale (different orders of magnitude) and multivariate (different chemical substances) properties [10], requiring the use of proper statistical methodologies to correctly process and extract the relevant information.

The complete workflow in metabolomics studies to extract the biological information from the raw files generated by the high-throughput analytical techniques generally includes a pre-processing (convert the raw instrumental files into organised and tabulated file formats), pre-treatment (refinement of the pre-processed data for posterior data analysis) and statistical analysis of the data (application of appropriate algorithms to find the relevant biological information) [11].

Pre-processing of GC-MS and LC-MS chromatograms must fulfil the correct identification of the mass spectrum of individual metabolites and the accurate determination of metabolites abundance in each sample, being challenging due to coelution of analytes within single chromatographic peaks and retention time shifts between samples [12]. For this reason, several bioinformatics tools such as PARAFAC2, AMDIS, ChromaTOF, eRah, XCMS, MS-Dial, MZmine 2, MarkerView, or Compound Discoverer among others have been developed to automatically pre-process chromatographic data. For GC-MS data, the so-called PARAllel FACtor analysis2 (PARAFAC2) algorithm [13] demonstrated to successfully handle complex situations with coelutions, low signal-to-noise (S/N) ratio and retention time shifted chromatographic peaks, with the advantage that only the number of components of the model in a selected region of the chromatogram is required to be set, diminishing the risk of modelling problems [12]. Meanwhile, for pre-processing of LC-MS data, the XCMS framework implementing the feature detection algorithm *centWave*, the time alignment algorithm *OBI-warp,* and the grouping of features across samples algorithm *group.density* [14], has yielded better metabolite identification abilities than other widely-used similar methods, while keeping a high quantification performance [3].

During the pre-treatment step, the pre-processed data are filtered, cleaned, drift corrected, missing value imputed, normalized, transformed and/or scaled [15] with the aim to make it adequate for the subsequent statistical analysis.

Finally, statistical analyses are performed on the data to discover the relevant biological information. Different computational and statistical methods have been developed for this task, comprising unsupervised and supervised analysis [11]. In metabolomics studies, unsupervised analyses are fundamentally used to explore and unravel the structure of the data, aiming at detecting sample clusters or systematic trends and errors. One of the most popular and powerful unsupervised statistics used in metabolomics studies is principal component analysis (PCA), based on dimensionality reduction [16]. Meanwhile, supervised learning methods are used for classification, prediction and biomarker discovering, dealing with datasets that have response variables either continuous (regression problems) or discrete (classification problems) [16]. Partial least squares (PLS) and its discriminant extension PLS-DA are one of the most popular methods used for regression and classification problems, respectively, in metabolomics studies due to its high performance and easy interpretation [17]. These models are useful when the data are acquired from the same omics platform (i.e., LC-MS or GC-MS or nuclear magnetic resonance (NMR), etc.), since simple concatenation of different omics data ignores their heterogeneity and a single type of omics is mainly highlighted [18]. However, metabolomics studies usually involve measurements from different omics platforms that may contain latent biological relationships and the independent assessment of each biological domain may not reveal that information. To overcome this, several frameworks implementing multi-omics data integration have emerged [19], including the unsupervised regularised generalised canonical analysis (RGCCA) and its sparse discriminant version sGCC-DA (DIABLO), which are useful multiblock methods to study the relationships between blocks of data measured in different platforms and to identify the main subsets of variables involved in those relationships [18].

The aim of this study was to establish a data processing and statistical analysis pipeline to (1) evaluate the impact of different elaboration processes on the metabolome of wines, acquired by means of two high throughput analytical platforms in an untargeted approach: GC-MS (volatile metabolites) and LC-MS (non-volatile metabolites), and to (2) relate those changes with their sensory descriptors. An optimal pre-treatment and pre-processing of the data generated from both analytical platforms was proposed to ensure its quality. Variable selection procedures were used to reveal the metabolites more impacted by the different wine elaboration processes. Each single omics data set and the sensory data were independently assessed by means of univariate and multivariate statistical analyses to study the impact of the wine elaboration process. The relationships between each single analytical platform and the sensory descriptors were evaluated by means of partial least squares regression (PLSR) models. Finally, the data obtained from the three sources were treated as a whole by means of data integration models to reveal hidden correlations.

2. Materials and Methods

2.1. Wine Samples

The red grape varieties *Vitis vinifera* L. cv. Garnacha and L. cv. Cabernet Sauvignon were used to elaborate rosé wines with three fermentation strategies (classes of the factorial design). Both grapes were picked at optimum ripeness, destemmed, and crushed. The maceration process was performed for 24 h with the addition of 40 mg/L of sulphur dioxide (SO_2) (Enartis, La Rioja, Spain). During the maceration tank filling the enzymes (Enartiszym Arom MP, Enartis, La Rioja, Spain) were added according to the supplier's recommendations (3 g/100 kg). The resulting musts were softly pressed homogenized and dejuiced at 4 °C for 24 h with the addition of 2.5 mL/hL of pectolytic enzymes (EnartisZym Blanco L, Enartis, La Rioja, Spain). At the beginning of the monoculture fermentation the *Saccharomyces cerevisiae* yeasts (Red Fruit, Enartis, La Rioja, Spain) were added at 25 g/hL, while in the sequential inoculations the *non-Sc* yeasts (*Torulaspora delbrueckii* TD291 Biodiva^{TM} and *Metschnikowia pulcherrima* MP346 Flavia$^{®}$) at 25 g/hL were used as starters of the fermentation, and 1 day later *Sc* yeasts were also added at 25 g/hL. The *non-Sc* yeasts were supplied by Lallemand as activated dried yeasts (ADY). The alcoholic fermentation was carried out at 17–18 °C. To control the kinetics of the alcoholic fermentation, density and temperature were daily monitored. The alcoholic fermentation finished when residual sugar concentrations were under 2 g/L. The wines we <re racked 15 days after the final alcoholic fermentation by cold settling using 60 g/hL of bentonite (Enartis, La Rioja, Spain) and 15 mL/hL of liquid gelatin at 30% (Enartis, La Rioja, Spain). After two months of cold stabilization at 0 °C, the wines were filtered, first using CKPV 16 plates (Cordenons, Milan, Italy) and then using a membrane cartridge of 1.0, 2.0 and 0.45 µm (Millipore), and then bottled using agglomerated cork caps (Selecork model, Tapones del Sur, Jerez de la Frontera, Spain).

2.2. Analysis, Data Acquisition and Processing of Headspace Solid-Phase Microextraction Coupled with Gas Chromatography-Mass Spectrometry (HS-SPME-GC-MS) Data

The volatile metabolome of the wines was determined using headspace solid-phase microextraction coupled with gas chromatography-mass spectrometry (HS-SPME-GC-MS). Samples of 15 mL were diluted 1:1 with EDTA solution (200 mM and pH 7.0 adjusted with NaOH 1.0 M) and 10 mL of this dilution were transferred to a 20 mL SPME vial closed with a 18mm magnetic PTFE/Sil headspace cap after the addition of 3.5 mg of NaCl. The samples processed were then homogenized in a vortex shaker for 30 s and placed in the autosampler tray (CTC Analytics, Zwingen, Switzerland). The fibre used for the HS-SPME was DVB/CAR/PDMS (1 cm, StableFlex/SS) supplied by Supelco (Bellefonte, PA, USA). The vials were incubated at 500 r.p.m for 2 min at 40 °C. The volatile fraction was extracted at 40 °C for 38 min followed by the desorption at 250 °C for 15 min into a Trace GC ultragas chromatograph (Thermo Fisher Scientific S.p.A., Rodano, Milan, Italy) coupled to an ISQ Single Quadrupole MS spectrometer (Thermo Fisher Scientific, Austin, TX, USA). Injection was set to splitless mode for 0.75 min. GC separation was performed on a BP21 column (50 m × 0.32 mm × 0.25 µm) (SGE Analytical Science, Milton Keynes, United Kingdom). The carrier gas used was helium at a flow rate of 1.7 mL/min. The oven temperature started at 40 °C for 5 min, then raised to 220 °C at 3 °C/min and maintained for 30 min. The MS transfer line and ion source temperatures were 230 °C and 200 °C, respectively. The mass spectrometer operated in electron ionization mode at 70 eV. Mass spectra were recorded in the 50–400 m/z value range and the scanning frequency was 5 scans/s. The Thermo Xcalibur v. 2.2 software was used to control both the Combipal autosampler and GC-MS.

Raw untargeted GC-MS data from rosé wines were acquired in Xcalibur file format (.raw) and converted to the international ANDI file format (.cdf) by means of the file converter tool implemented in Xcalibur. They were then pre-processed (Figure 1) to convert the chromatograms into extracted data (tables of metabolite chromatographic areas). The raw .cdf files with a three-way array structure (elution time × mass spectra × samples) were deconvoluted by means of the PARAFAC2 algorithm. The chromatograms were first divided into 335 intervals in the elution time dimension in order to reduce the complexity and a single model was built for each interval. For each model, one to eight components were fitted and the selection of the optimal number of components was based on model fit (%), core consistency, distribution of the residuals, comparison of the resolved mass spectral profiles against the raw profiles, and retention times of the weighted elution profiles. The freely-available platform PARADISe v. 3.88 was used to implement this framework [12]. To verify the quality of the data, pooled quality control (QC) samples were repeatedly injected throughout the whole sequence. Different levels of metabolite identification confidence were categorized in accordance with the proposed minimum reporting criteria defined by the metabolomic standard initiative [20]. Two orthogonal properties: linear retention index (LRI) calculated with the equation of Van den Dool and Kratz, and fragmentation mass spectrum, were used for metabolite identification. Three levels of identification were considered. The resolved PARAFAC2 mass spectrums and linear retention index were compared against an authentic chemical standard analysed under the same experimental conditions for definitive identification (level 1). Level 2 was considered for metabolites showing a match factor (MF) > 850 and LRI ± 30 compared to the NIST database. Level 3 was considered for metabolites with 650 $\leq$ MF $\leq$ 85 a and LRI ± 30 compared to the NIST database. Since close values of MF and reverse match factor (RMF) are related to low levels of background, RMF was used to control the quality of each metabolite background subtraction.

2.3. Analysis, Data Acquisition and Processing of Using Ultra-High-Performance Liquid Chromatography High-Resolution Mass Spectrometry (UHPLC-HRMS) Data

The relative non-volatile metabolome of rosé wines was determined using ultra-high-performance liquid chromatography coupled to an Exactive Orbitrap mass spectrometer (UHPLC-HRMS) controlled by the Thermo Xcalibur v. 2.2 software. A total of 1 mL of each rosé wine sample was centrifuged at 15,000 rpm for 10 min at 20 °C (Eppendorf™ 5424 Microcentrifuges, Hamburg, Germany) and 200 µL of the supernatant were transferred into a 1.5 mL amber glass vial, mixed with 380 µL of acetonitrile and placed in the autosampler refrigerated at 10 °C of a Dionex Ultimate 3000 RS UHPLC system coupled to an Exactive Orbitrap mass spectrometer detector (Thermo Fisher Scientific, San Jose, CA, USA). Hydrophilic interaction separation was carried out on a 2.1 mm × 150 mm ACQUITY UPLC 1.7 µm BEH amide column, equipped with an ACQUITY UPLC BEH amide 1.7 µm van-guard pre-column (Waters, Barcelona, Spain). The temperature of the column was maintained at 40 °C and two mobile phases were used: A: acetonitrile and B: acetonitrile-water (2:98) + 1 mM ammonium formate. The injection volume was 5 µL and the separation was obtained at a flow rate of 0.4 mL/min with a 40 min gradient. The gradient started at 5% B for 4 min, raised to 28% B in 25 min, raised to 60% B in 30 min, maintained for 0.5 min before rising to 80% B in 33 min. After that, the column was equilibrated to the previous conditions within 10 min. Full scans were recorded in the *m/z* range from 100 to 1000 with a resolution of 50,000 Hz and with a full AGC target of

100,000 charges, using 2 microscans. Analyses were also based on scans with in-source collision-induced dissociation (CID) at 25.0 eV. MS experiment condition with HESI in positive ionization (separately analysed) mode was: (i) capillary temperature was 325 °C, the heater temperature was 300 °C, the sheath gas was 25 units, the auxiliary gas was 4 units, and the spray voltage was 4.0 kV. While the MS experiment condition with HESI in negative ionization (separately analysed) mode was: (i) capillary temperature was 325 °C, the heater temperature was 300 °C, the sheath gas was 20 units, the auxiliary gas was 2 units, and the spray voltage was 4.0 kV.

Raw untargeted UHPLC-ESI-MS data from rosé wines were acquired in the Xcalibur file format (.raw) and converted to the international ANDI file format (.cdf) by means of the file converter tool implemented in the Xcalibur software. Then, the data were pre-processed (Figure 1) using the R package XCMS v. 3.4.4 [21]. Peak picking, retention time correction and grouping parameters from XCMS were optimised by means of a Box-Behnken design using the R package IPO v. 1.8.1 [22]. The centWave algorithm was selected for peak picking, while the OBI-warp and density methods were used for retention time corrections and grouping, respectively. Then, the features derived from the same metabolite and annotation of the ion species were grouped with the R package CAMERA v. 1.38.1 [23]. Scripts including XCMS optimised parameter settings and CAMERA are given in the Supplementary material R scripts. Metabolite annotation was performed on the basis of the minimum reporting criteria of the four identification levels defined by the metabolomic standard initiative [20] and described in [24]. The mass-based searching tool MetaboQuest was used to find putative metabolite identities from the HMDB and METLIN databases. The results shown in Supplementary Tables S1 and S2 were compared against the feature list from the WinMet database [9] by using an automated R routine (available in the Supplementary material R scripts) that matches neutral masses of WinMet with the putative metabolites obtained in MetaboQuest and selects the candidates present in both lists. Finally, the metabolites were manually inspected and annotated for up to 5 ppm error.

2.4. Sensory Analysis

The sensory analysis of the wines was performed at IFAPA Rancho de la Merced (Jerez) by means of a tasting panel composed of 10 trained panellists between 35 and 55 years. The panellists were trained following the AENOR (UNE-EN ISO 8586:2014 and UNE 87022:1992) standards and the procedure was the same than that described in [25]. The wine description terminology agreed by consensus included 5 attributes scored in an olfactory phase (scent intensity, red fruit, black fruit, citrus fruit and tree fruit) and 6 attributes scored in a taste phase (taste intensity, acidity, alcohol, complexity, balance and persistence).

2.5. Statistical Analysis

The data from the different sources were pre-treated in different ways (Figure 1). The sensorial scores were only autoscaled before the integration approach. Meanwhile, the data from both analytical platforms were first cleaned, replacing features with non-detected single classes but detected in other classes by the minimum peak area reported in the data matrix divided by 2. Next, the data were peak filtered by removing rows (samples) or columns (features) with missing values ≥30% across all classes. To reduce the contribution of the experimental and analytical variability, the data were drift corrected by fitting a smoothed cubic spline onto the QC samples as described by [26]. The R function smooth.spline was used, and the smoothing parameter was set to 1 after its optimization using leave-one-out cross validation at different values of this hyperparameter. The drift correction was performed as followed (Equation (1)):

$$X_{corrected}\ (i) = X_{original}\ (i) + mean(X_{QC}) - f_{drift}(i) \tag{1}$$

where i is the injection order of each sample, $X_{corrected}$ is the corrected value of a feature, $mean(X_{QC})$ is the mean value of all the QC samples used to fit the smoothed cubic function and f_{drift} is the smoothed cubic function. Subsequently, features with an RSD > 20 and RSD > 30 in QC samples for LC-MS and GC-MS data respectively were removed. Random forest imputation was applied to impute missing values using the R package missForest [27]. QC samples were not included in the data to ensure a proper imputation based on patterns of the real wine data. Next, data were probabilistic quotient normalised (PQN) using the median of the QC samples, since it is less sensitive to outliers than the mean [28]. Finally, GC-MS and LC-MS data were autoscaled. Throughout this pre-treatment, principal component analyses (PCA) including QCs were conducted to unravel the structure of the data and to detect systematic trends and errors.

Univariate analyses were performed on the data without autoscaling. Normality and heteroscedasticity were checked with the Shapiro–Wilk's and Levene's tests respectively, and the variables that failed the parametric assumptions were box-cox transformed. Two-way analysis of variance (ANOVA) and Tukey's honestly significant difference (HSD) post-hoc tests were then applied to identify differences in the metabolites and sensory descriptors due to the factors studied.

Partial least squares discriminant analysis (PLS-DA) was used for the selection of the most discriminative features (potential markers) and for classification tasks, which were further used for the biological interpretation. The optimization and validation of the PLS-DA models were based on a double cross-validation described by [29]. Briefly, the cross-validation consisted of two nested loops, referred to as inner and outer loops. In the former, the model was optimised using leave-one-out exhaustive cross-validation. In the latter, a 4-fold cross-validation made it possible to assess the performance of the final discriminant model. The optimization and validation of the models were based on the balanced error rate (BER) measured at the Mahalanobis distance. The significance of the PLS-DA models was obtained through the following equation (Equation (2)):

$$p\text{-}value = 1 + (HP_P \leq HP)/N \tag{2}$$

where HP is a hyperparameter: BER, number of misclassification (NMC) or the area under receiver operating characteristic curve (AUROC), and ($HP_P \leq HP$) is the number of elements in the H_0 distribution which are smaller or equal to the hyperparameter of the original data. H_0 distribution was generated by random permutation of sample labels. N = 1000 permutations were simulated, since this number was large enough to sample the distribution tails and attain a p-value up to 0.001. The BER, NMC and AUROC of the original data were calculated by averaging the values of each hyperparameter obtained from the sub-models fitted during the cross-validation. The selection of the most discriminative metabolites (potential markers) followed an iterative workflow based on variable importance in projection (VIP) described in [30]. Alternatively, the data from the three different sources was integrated using the framework DIABLO, which is the supervised extension of sparse generalized canonical correlation analysis (sGCC-DA) [18].

Partial least squares regression (PLSR) was used to predict the sensory attributes scores from the volatile and the non-volatile profile separately. Afterwards, the simultaneous integration of the three datasets was carried out by means of RGCCA implemented in the R package mixOmics [18]. To this end, the R package MUVR [31] was first used to extract the all-relevant variables from the LC-MS and GC-MS platforms, i.e., all features influenced by the studied factor (in this case the fermentation strategy) once the non-informative features were removed. A PLS-DA and random forest (RF) models were independently fitted in each single dataset using a double cross-validation scheme. The all-relevant variables from both discriminant analyses were merged to perform the PLSR and RGCCA models.

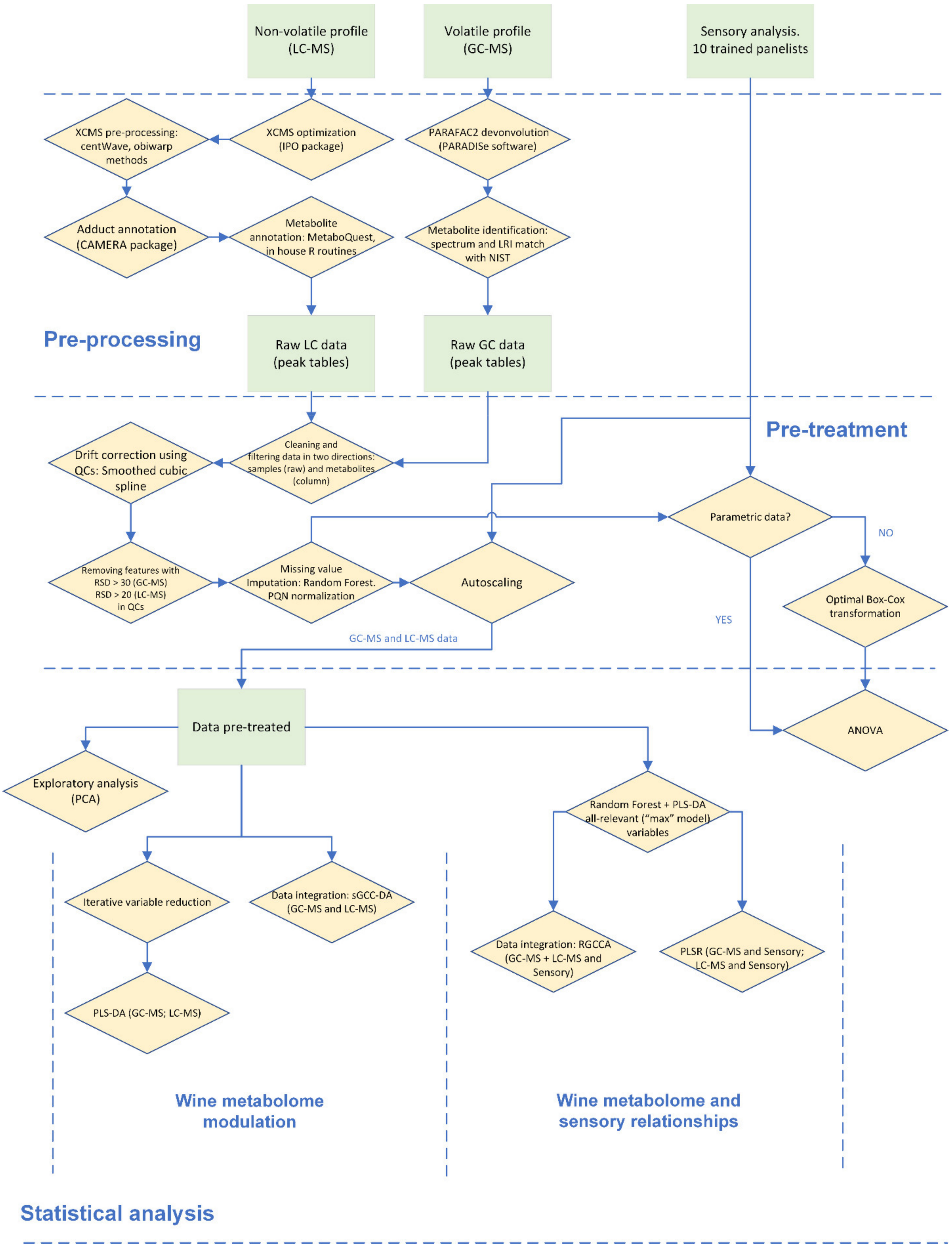
Non-volatile profile (LC-MS)
Volatile profile (GC-MS)
Sensory analysis. 10 trained panelists
XCMS pre-processing: centWave, obiwarp methods
XCMS optimization (IPO package)
PARAFAC2 devonvolution (PARADISe software)
Adduct annotation (CAMERA package)
Metabolite annotation: MetaboQuest, in house R routines
Metabolite identification: spectrum and LRI match with NIST
Pre-processing
Raw LC data (peak tables)
Raw GC data (peak tables)
Pre-treatment
Drift correction using QCs: Smoothed cubic spline
Cleaning and filtering data in two directions: samples (raw) and metabolites (column)
Parametric data?
NO
Removing features with RSD > 30 (GC-MS) RSD > 20 (LC-MS) in QCs
Missing value Imputation: Random Forest. PQN normalization
Autoscaling
Optimal Box-Cox transformation
YES
GC-MS and LC-MS data
Data pre-treated
ANOVA
Exploratory analysis (PCA)
Random Forest + PLS-DA all-relevant ("max" model) variables
Iterative variable reduction
Data integration: sGCC-DA (GC-MS and LC-MS)
Data integration: RGCCA (GC-MS + LC-MS and Sensory)
PLSR (GC-MS and Sensory; LC-MS and Sensory)
PLS-DA (GC-MS; LC-MS)
Wine metabolome modulation
Wine metabolome and sensory relationships
Statistical analysis

Figure 1. Flowchart illustrating the step-by-step processes followed to convert raw data into processed data to extract biological interpretations.

3. Results

3.1. Volatile Profile

Prior to GC-MS injection of the real samples, pooled QCs were used to optimize the dilution factor and the extraction fibre employed. Dilution ratios of 1:1, 1:2, 1:4 and non-dilution were tested in three different fibres, which covered a vast range of polarities, such as PDMS, CAR/PDMS and DVB/CAR/PDMS. The combination of the DVB/CAR/PDMS fibre and non-dilution yielded the best results in terms of total chromatographic peak signal (sensitivity). However, the DVB/CAR/PDMS fibre and 1:1 dilution ratio were selected, since 10% more features could be detected with low deviation coefficients (20% RSD), improving stability. Subsequently, the real rosé wine samples were analysed with the optimised methodology. PARAllel FACtor analysis2 (PARAFAC2) deconvolution algorithm [13] was used to pre-process the converted data, with a total of 505 molecular features assigned to detected peaks. After the pre-treatment step, 429 features remained in the data set, among which 126 were definitive—or putatively identified (Supplementary Table S3). PCAs including QC samples were used during data pre-treatment to check the quality of the data (Figure 2). The quality of the raw data (before pre-treatment) was quite satisfactory, since QCs were allocated in a narrow region and samples were grouped according to the fermentation strategy. However, the pre-treatment workflow proposed in this study led to a slight improvement of the final data, with QCs and samples belonging to the same fermentation strategy being allocated in closer areas.

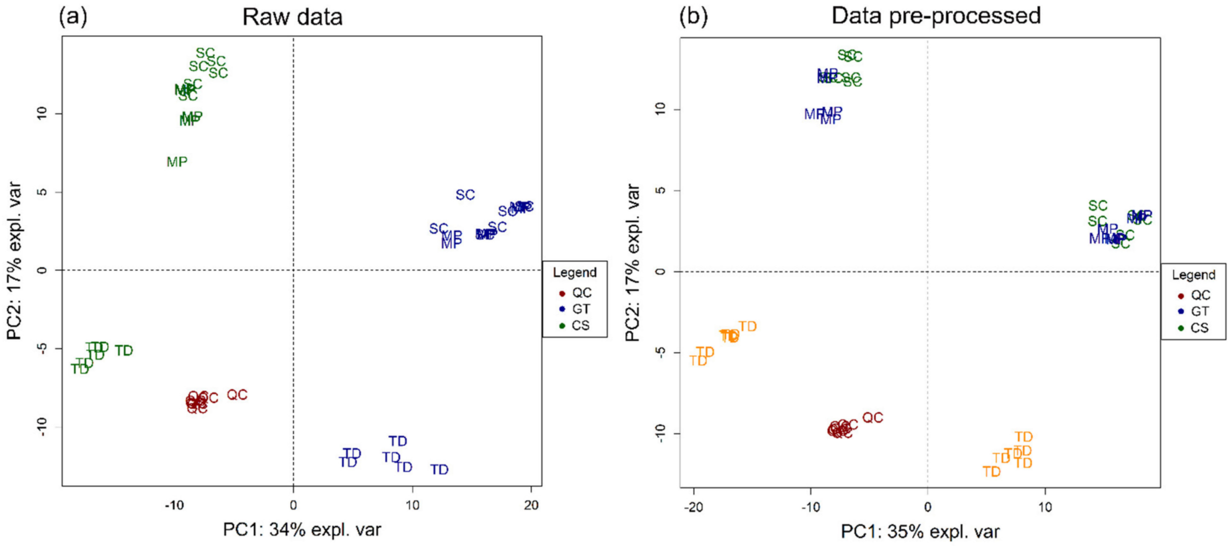

Figure 2. Principal component analysis (PCA) performed for the gas chromatography (GC) data during pre-treatment. (**a**) Raw data and (**b**) pre-treated data. Quality control samples (QCs) are included to verify the quality of the data. *TD*: wines fermented with sequential inoculation of *Torulaspora delbrueckii* and *Saccharomyces cerevisiae*. *MP*: wines fermented with sequential inoculation of *Metchnikovia pulcherrima* and *Saccharomyces cerevisiae*. *SC*: wines fermented with monoculture of *Saccharomyces cerevisiae*.

Afterwards, a two-way ANOVA (Supplementary Table S3) was performed to study the effect of the fermentation strategy, that impacted to different chemical groups, especially esters, alcohols and terpenes, independently of the grape variety of the wine (Supplementary Table S3). Rosé wines that followed a sequential inoculation with the Flavia *Metschnikowia pulcherrima* yeasts (*Sc-Mp*) showed statistically significant differences compared with those fermented with the monoculture Red Fruit *Saccharomyces cerevisiae* yeasts (*Sc*). However, the sequential inoculation with Biodiva *Torulaspora delbrueckii* yeasts (*Sc-Td*) produced rosé wines with the most distinct volatile profile, with overall lower contents of esters and aldehydes, but an increase in several alcohols and terpenes (Supplementary Table S3). Subsequently, the main wine volatile metabolites linked to the fermentation strategy were selected by means of a variable reduction procedure based on variable importance in projection (VIP) from the PLS-DA model. These metabolites were considered as potential markers of the fermentation strategy and a new model was fitted using the reduced data. The performance of the PLS-DA including the potential volatile markers displayed a mean overall BER of 0.08 ± 0.06 and *p*-values from the permutation test below 0.05 in the three diagnostic statistics (Table 1). This resulted in a slight improvement compared to the model fitted with all the variables (BER of 0.11 ± 0.08), and the complexity of the model was reduced from 3 to 2 components. The X-variate 1 explained a high 65% of the total variance found in the rosé wine samples, that separated the *Sc-Td* wines from the rest (Figure 3). This result supported the more distinct volatile profile of *Sc-Td* wines previously observed in the ANOVA. Considering the annotated compounds, 3-ethoxy-1-propanol, ethyl decanoate, 1-dodecanol, methyl 4-decenoate, (E)-nerolidol and ethyl 2-methylpropanoate (ethyl isobutyrate) were selected as the most discriminative metabolites of this fermentation strategy by displaying higher overall content. Meanwhile, *Sc* and *Sc-Mp* wines were separated in X-variate 2, that explained the 20% of the total variance of the samples (Figure 3). The volatile metabolites involved in such differentiation were pentyl-cyclopropane, isopentyl hexanoate and ethyl heptanoate, found in overall higher concentrations in *Sc-Mp* wines, and 3-isopropyl-1-pentanol (3-ethyl-4-methylpentanol), ethyl trans-3-hexenoate, (E,E)-farnesol and 3-furaldehyde, that were found in higher contents in *Sc* wines.

Table 1. Model performance of the partial least squares discriminant analysis (PLS-DA) models built to discriminate yeast fermentation strategy and grape variety used to elaborate the rosé wine samples.

Analytical Platform	Model	Mean Overall BER [(1)]	Ncomp	Class	Mean Class Error [(2)]	*p*-Value [(3)]
GC-MS	Allvariables	0.11 ± 0.08	3	*Sc-Td* *Sc-Mp* *Sc*	0.01 ± 0.03 0.15 ± 0.16 0.18 ± 0.17	BER: <0.001 NMC: 0.024 AUROC: 0.020
GC-MS	Variable reduction	0.08 ± 0.06	2	*Sc-Td* *Sc-Mp* *Sc*	0.00 ± 0.00 0.12 ± 0.11 0.14 ± 0.11	BER: <0.001 NMC: 0.007 AUROC: 0.003
LC-MS	Allvariables	0.17 ± 0.10	4	*Sc-Td* *Sc-Mp* *Sc*	0.14 ± 0.09 0.07 ± 0.13 0.29 ± 0.21	BER: <0.001 NMC: 0.120 AUROC: 0.035
LC-MS	Variable reduction	0.02 ± 0.04	2	*Sc-Td* *Sc-Mp* *Sc*	0.00 ± 0.00 0.05 ± 0.11 0.00 ± 0.02	BER: <0.001 NMC: 0.027 AUROC: 0.002

[1] Mean overall balanced error rate (BER) values with the standard deviation calculated on the basis of 200 PLS-DA sub-models in a double-cross validation scheme. [2] Mean class error values with the standard deviation calculated on the basis of 200 PLS-DA sub-models in a double-cross validation scheme. [3] Model statistical significance was calculated on the basis of a permutation test (N = 1000) using the mean overall BER, number of misclassifications (NMC) and the area under receiver operating characteristic curve (AUROC) as diagnostic statistics. Statistically significant models (*p*-value ≤ 0.05) were highlighted in bold. *Sc-Td*: wines fermented with sequential inoculation of *Torulaspora delbrueckii* and *Saccharomyces cerevisiae*. *Sc-Mp*: wines fermented with sequential inoculation of *Metchnikovia pulcherrima* and *Saccharomyces cerevisiae*. *Sc*: wines fermented with monoculture of *Saccharomyces cerevisiae*.

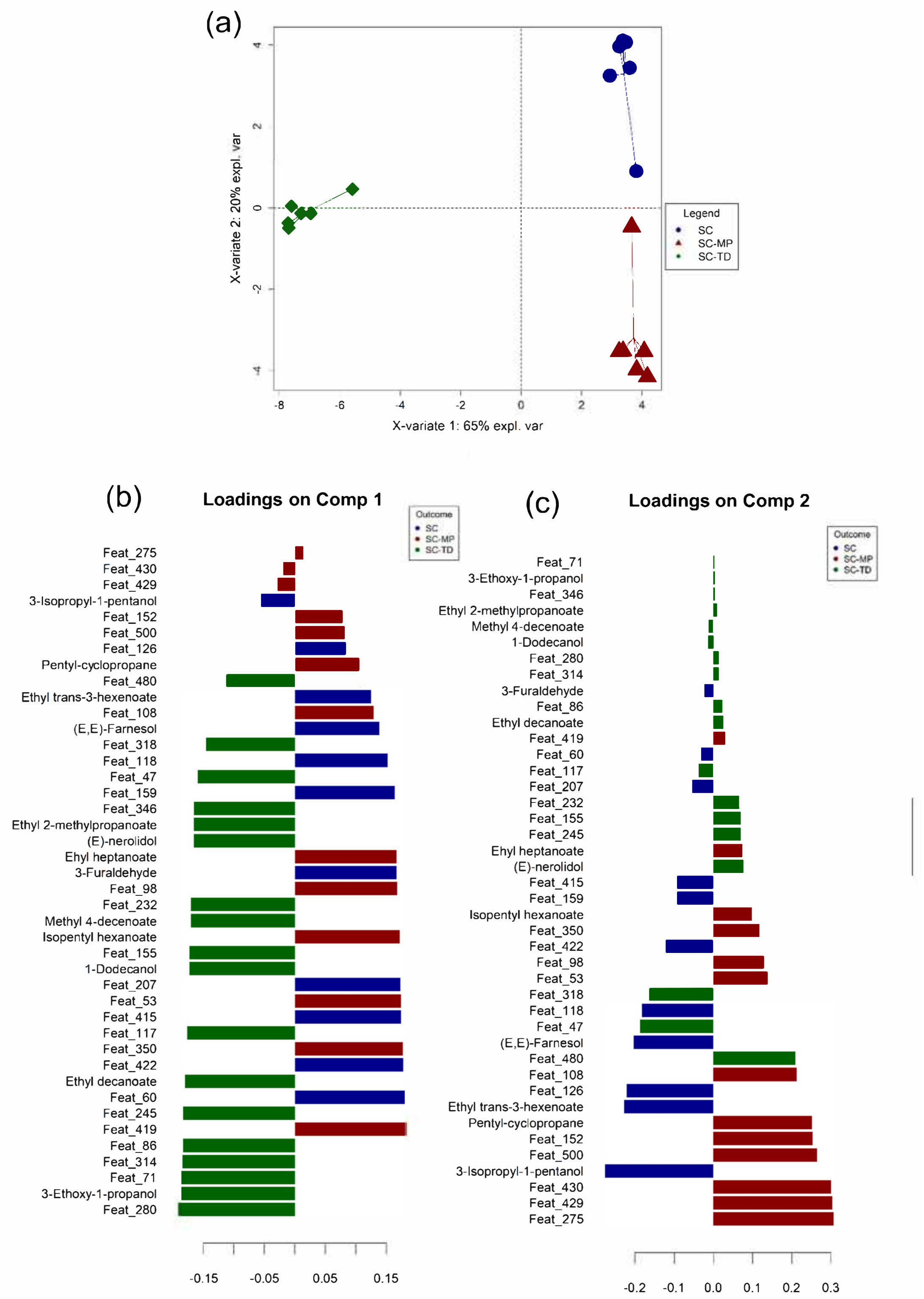

Figure 3. Graphical outputs of the partial least squares discriminant analysis (PLS-DA) performed to classify rosé wines according to the fermentation strategy on the basis of their volatile profile. (**a**) Scores plot for the components 1 and 2; (**b**) and (**c**) loading contribution barplot on components 1 and 2. Colour indicates the class for which the compound has a maximal mean value. Bar length represents the multivariate regression coefficient with either a positive or negative sign for that particular feature of each component, i.e., the importance of each variable in the model. *Sc-Td*: wines fermented with sequential inoculation of *Torulaspora delbrueckii* and *Saccharomyces cerevisiae*. *Sc-Mp*: wines fermented with sequential inoculation of *Metchnikovia pulcherrima* and *Saccharomyces cerevisiae*. *Sc*: wines fermented with monoculture of *Saccharomyces cerevisiae*.

3.2. Non-Volatile Profile

The optimization of the dilution procedure for the wine samples was achieved using pooled QC samples. Direct injection and dilution ratios of 1:1, 1:2 and 1:5 were tested. Samples were analysed and processed to extract the features, and two parameters were considered to select the best dilution conditions: total sum area (response) and number of features with RSD <30 (stability). The dilution of the samples yielded a reduction in the response but increased the stability of the features. In positive ionization mode, the compounds were most stable at a dilution ratio of 1:2, which captured close to 2-fold more compounds with RSD <30 than direct injection. Meanwhile, in negative ionization mode, the dilution ratio of 1:5 displayed slightly better results than 1:2 in terms of response. However, this dilution ratio was discarded due to loss of sensitivity (around 50% compared to 1:2). Therefore, the best consensus in terms of sensitivity and stability was obtained for a dilution ratio of 1:2, which was used for the analysis of the real samples. Respective totals of 1801 and 4275 features for positive and negative ionization modes were obtained after XCMS pre-processing, and 1167 and 1701 features remained respectively after data pre-treatment step. PCAs including QC samples were again used during the pre-treatment process to verify the quality of the data (Figures 4 and 5 for negative and positive ESI modes, respectively). Data acquired in negative ESI mode displayed an important batch effect which was highlighted by the QC samples, and after data pre-treatment this effect was successfully eliminated (Figure 4). Although a batch effect was not observed in positive ionization mode, an improvement in the final quality of the data was also achieved after the pre-treatment process, since QCs and samples were grouped closer and better separated (Figure 5).

As for the volatile data, ANOVA and PLS-DA models were used to disclose the relevant biological information contained in the data. The ANOVA showed clear differences in the non-volatile profile of the rosé wines due to the fermentation strategy regardless of the grape variety, with amino acids and organic acids being the most impacted compound families (Supplementary Table S4). The variable reduction procedure allowed us to obtain a PLS-DA with improved classification performance compared to the model fitted with all the metabolites, displaying a BER of 0.02 ± 0.04 and *p*-values from the permutation test below 0.05 using the three diagnostic statistics (Table 1). A total of 38 metabolites were selected as markers and the model was optimized for two components accounting for 88% of the total variance found in the samples, which were clearly grouped according to the fermentation strategy (Figure 6). As found for the volatile profile, *Sc-Td* wines displayed the most distinct non-volatile profile, since these wines were separated from the rest on component 1 (X-variate 1), that explained the 71% of the total variance of samples. Among the annotated compounds, glutamic acid and D-mannitol/dulcitol were selected as potential markers of the fermentation strategy, displaying higher overall concentrations in *Sc-Td* wines (Figure 6). Meanwhile the variance explained in component 2 made it possible to differentiate *Sc-Mp* from *Sc* wines, being the non-volatile metabolites quinic acid and imidazole acetic acid selected as potential markers of *Sc-Mp* wines (Figure 6).

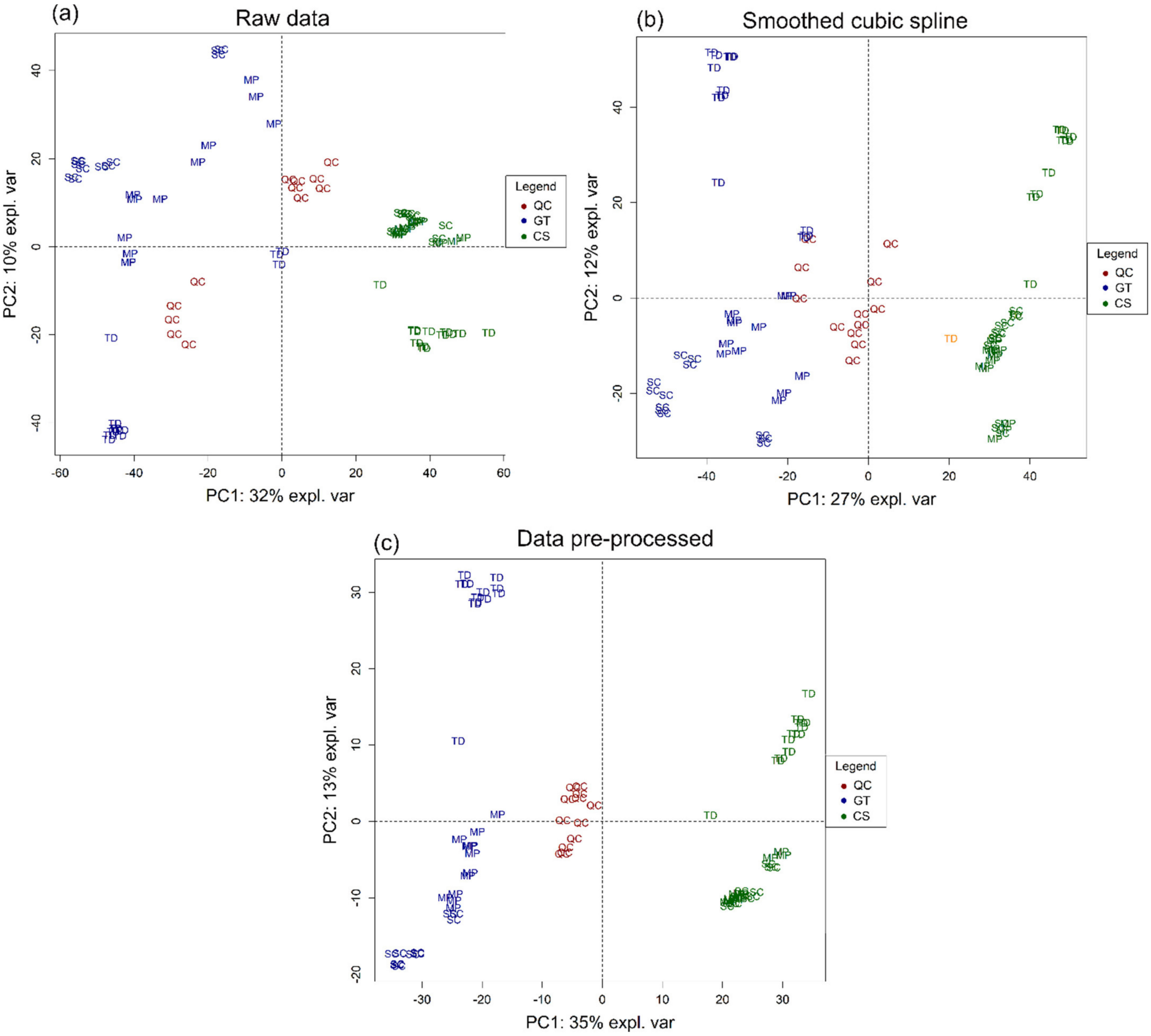

Figure 4. Principal component analysis (PCA) performed for the liquid chromatography (LC) negative ESI mode data during pre-treatment. (**a**) Raw data, (**b**) after smoothed cubic spline and (**c**) pre-treated data. Quality control samples (QCs) are included to verify the quality of the data. *TD*: wines fermented with sequential inoculation of *Torulaspora delbrueckii* and *Saccharomyces cerevisiae*. *MP*: wines fermented with sequential inoculation of *Metchnikovia pulcherrima* and *Saccharomyces cerevisiae*. *SC*: wines fermented with monoculture of *Saccharomyces cerevisiae*.

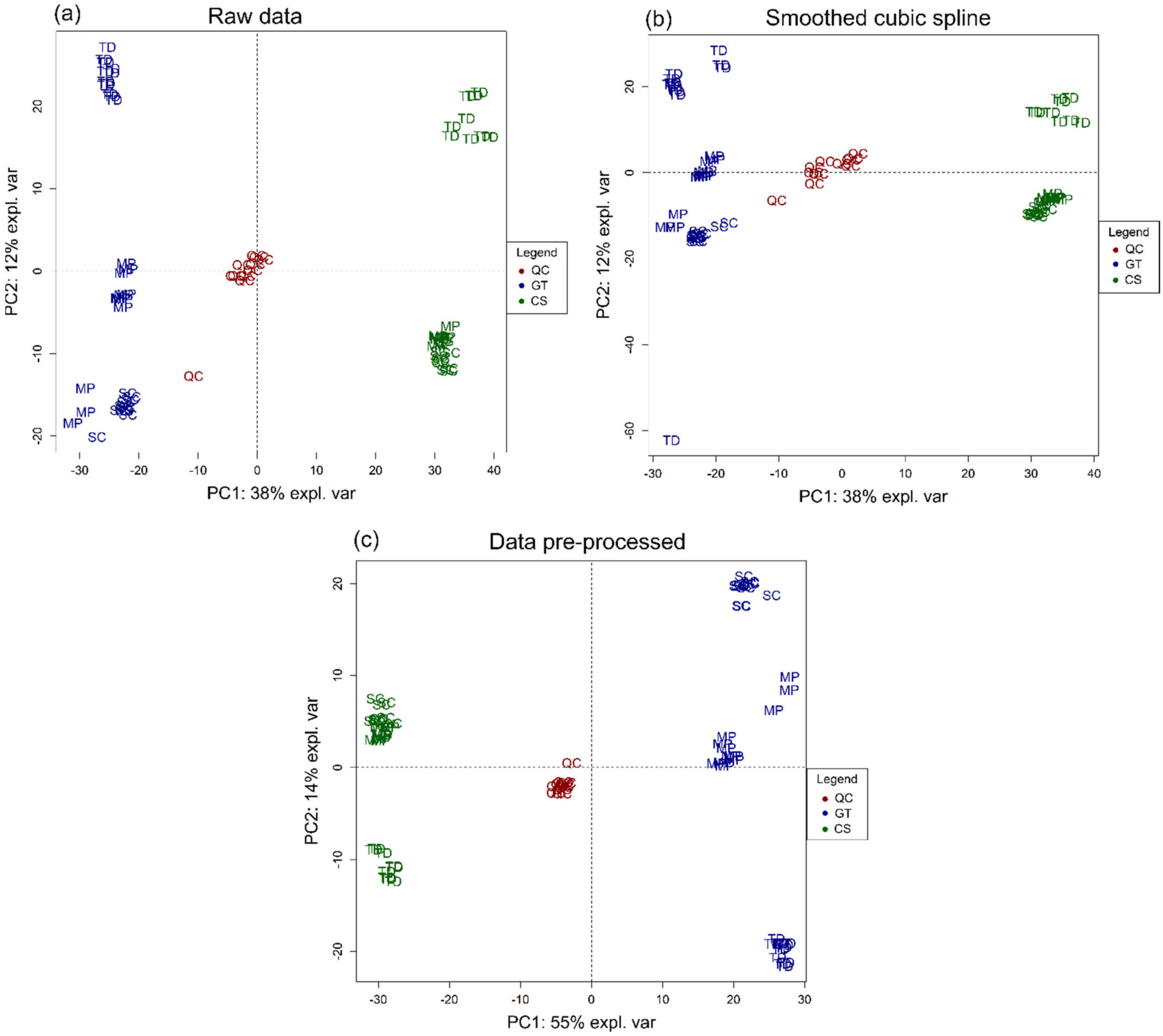

Figure 5. Principal component analysis (PCA) performed for the LC positive ESI mode data during pre-treatment. (**a**) Raw data, (**b**) after smoothed cubic spline and (**c**) pre-treated data. Quality control samples (QCs) are included to verify the quality of the data. *TD*: wines fermented with sequential inoculation of *Torulaspora delbrueckii* and *Saccharomyces cerevisiae*. *MP*: wines fermented with sequential inoculation of *Metchnikovia pulcherrima* and *Saccharomyces cerevisiae*. *SC*: wines fermented with monoculture of *Saccharomyces cerevisiae*.

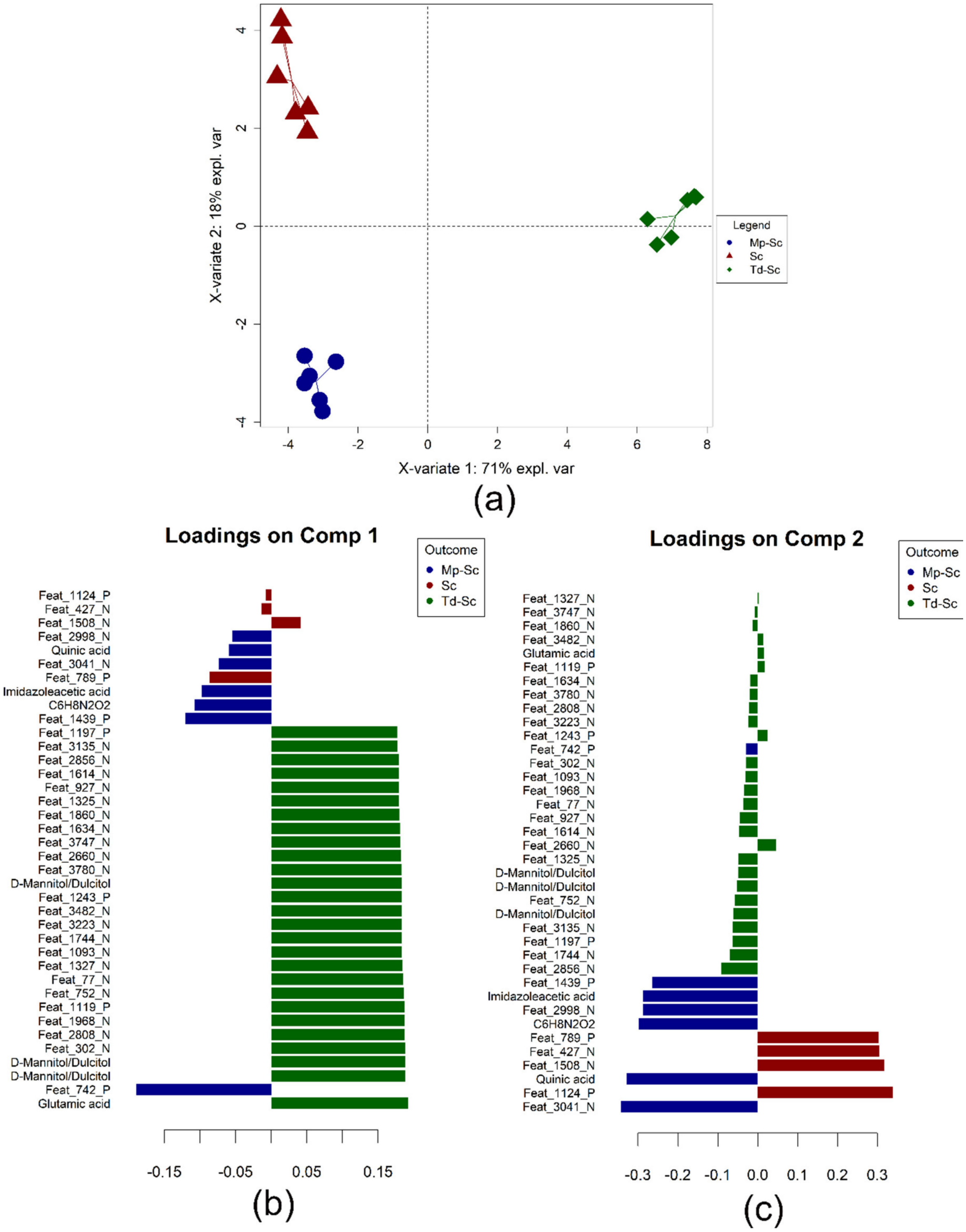

Figure 6. Graphical outputs of the partial least squares discriminant analysis (PLS-DA) performed to classify rosé wines according to the fermentation strategy on the basis of their non-volatile profile. (**a**) Scores plot for the components 1 and 2; (**b**,**c**) loading contribution barplot on components 1 and 2. Colour indicates the class for which the compound has a maximal mean value. Bar length represents the multivariate regression coefficient with either a positive or negative sign for that particular feature of each component, i.e., the importance of each variable in the model. *Sc-Td*: wines fermented with sequential inoculation of *Torulaspora delbrueckii* and *Saccharomyces cerevisiae*. *Sc-Mp*: wines fermented with sequential inoculation of *Metchnikovia pulcherrima* and *Saccharomyces cerevisiae*. *Sc*: wines fermented with monoculture of *Saccharomyces cerevisiae*.

3.3. Sensory Attributes

The impact of the fermentation strategy on the sensory attributes of wines was assessed by means of a two-way ANOVA. All the sensory attributes except taste intensity were affected by the fermentation strategy (Table 2). Overall higher scores were assigned to *Sc* and *Sc-Mp* wines regarding scent intensity, red fruit, black fruit, citrus fruit, tree fruit, acidity, alcohol and complexity. In addition, *Sc* wines displayed the highest scores in balance, followed by *Sc-Mp* wines. Persistence was the only sensory descriptor with the highest score in *Sc-Td* wines (Table 2).

Table 2. Two-way analysis of variance (ANOVA) performed on the sensory scores of the rosé wines from different fermentation strategies and grape varieties.

Sensory Descriptor	*Sc-Td* [1]	*Sc-Mp* [2]	*Sc* [3]	*p*-Value [4]	CS [5]	GT [6]	*p*-Value	Interactions (*p*-Value)
Scent intensity	6.0b	6.2ab	6.6a	*	6.4	6.1	ns	***
Red fruit	1.35b	1.97a	2.01a	***	1.90	1.66	ns	ns
Black fruit	1.47b	1.99a	2.03a	*	2.95a	0.71b	***	***
Citrus fruit	0.86b	1.65a	1.84a	***	1.00b	1.90a	***	***
Tree fruit	1.10b	1.78a	2.12a	***	1.58	1.75	ns	**
Taste intensity	5.8	5.5	5.6	ns	5.9a	5.4b	**	ns
Acidity	4.5b	4.9a	5.0a	**	4.6b	5.0a	*	ns
Alcohol	4.7b	5.0a	4.9a	**	4.8	4.9	ns	ns
Complexity	4.1b	4.6a	4.7a	***	4.6a	4.3b	*	ns
Balance	4.1c	5.0b	5.7a	***	4.5b	5.3a	***	**
Persistence	5.1a	4.2b	4.3b	***	4.7	4.4	ns	*

[1] Sc-Td: wines fermented with sequential inoculation of *Torulaspora delbrueckii* and *Saccharomyces cerevisiae*. [2] Sc-Mp: wines fermented with sequential inoculation of *Metchnikovia pulcherrima* and *Saccharomyces cerevisiae*. [3] Sc: wines fermented with monoculture of *Saccharomyces cerevisiae*. [4] *p*-value from the two-way ANOVA. ns: not significant, * : $0.05 > p\text{-value} > 0.01$, ** : $0.01 > p\text{-value} > 0.001$, and *** : $p\text{-value} < 0.001$. Mean values with different letters differ significantly. [5] CS: Cabernet sauvignon. [6] GT: Garnacha Tinta.

3.4. Data Integration: Sparse Generalised Canonical Correlation Analysis Discriminant Analysis (sGCC-DA) Approach

To look for the most important metabolites and sensory descriptors linked to the wine fermentation strategy, sparse generalized canonical correlation discriminant analysis (DIABLO) was also performed integrating the three datasets. This multi-block data analysis framework maximises the covariance between the data from the different analytical platforms and identifies the multi-omics signature that better discriminates the target outputs [18]. A consensus between maximising the correlation of datasets and the discrimination was adopted by setting the connection of blocks to a link of 0.7 [18]. The sGCC-DA model was optimized for three components, displaying a satisfactory error rate of 0 for both LC and GC analytical platforms and 0.222 for the sensory data. The Pearson's correlation coefficient revealed a good connection between the different blocks, with values above 0.85, being observed the highest correlations for the volatile and non-volatile metabolites blocks (Figure 7).

This approach made it possible to identify additional markers of the fermentation strategy which were not previously selected during the independent assessment of each analytical platform (Figure 8). The new potential markers selected were 1-undecanol and shikimic acid for Sc-Mp wines, which were respectively in around 9-fold and 6-fold higher concentrations compared to the Sc-Td wines, and duplicated in both cases the concentrations of Sc wines (Supplementary Tables S3 and S4). Meanwhile, linalyl anthranilate, ethyl 5-hexenoate, hexyl acetate and proline betaine for were selected as potential markers for Sc wines, found in around 1.5 to 4-fold higher concentrations in comparison to Sc-Td and Sc-Mp wines (Supplementary Tables S3 and S4).

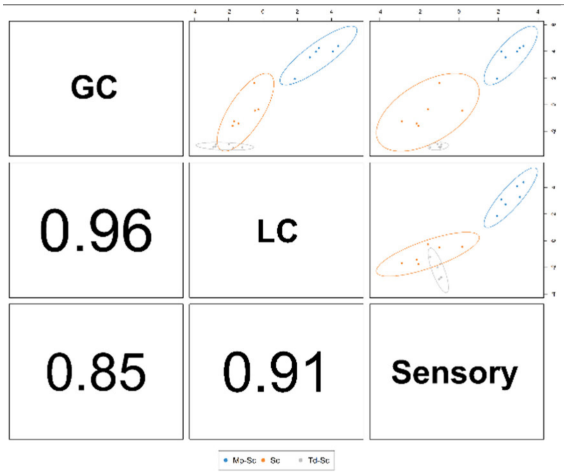

Figure 7. Pearson correlation plot obtained from the sparse generalised canonical correlation analysis discriminant analysis (sGCC-DA) model.

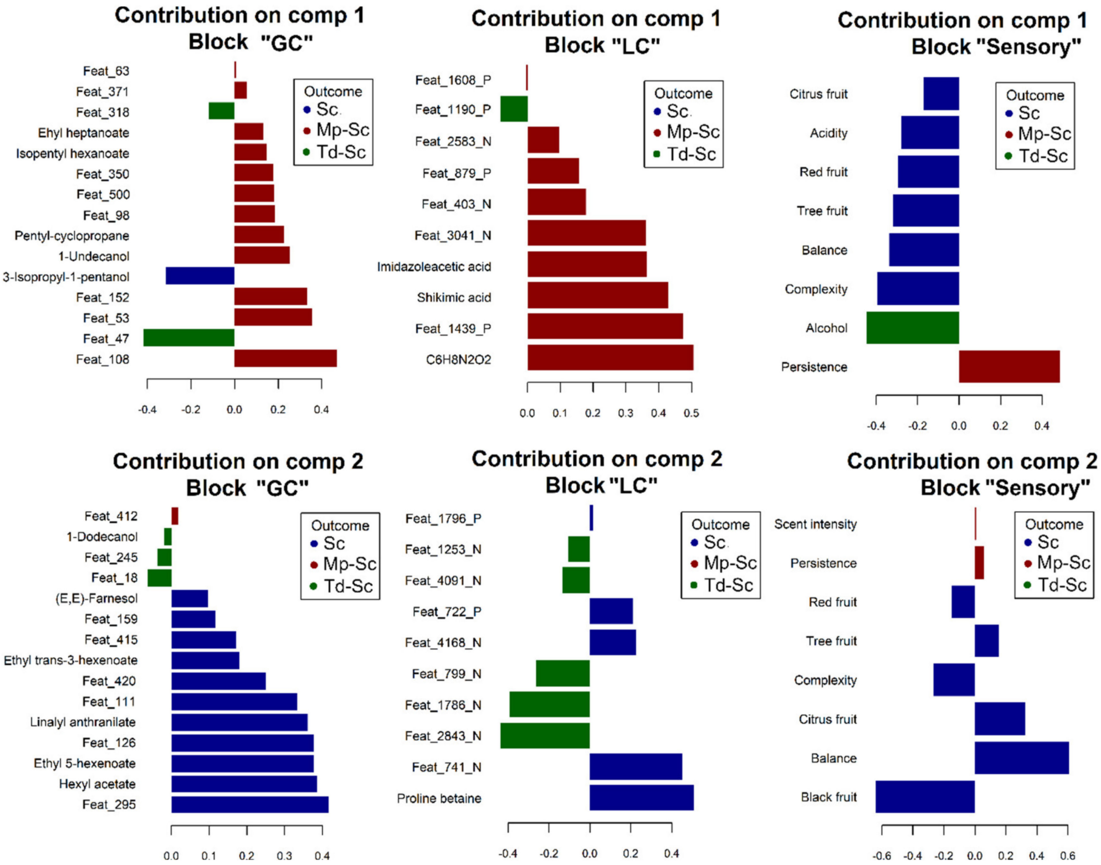

Figure 8. Loading contribution barplot on components 1 and 2 for the different blocks of the sGCC-DA. Colour indicates the class for which the compound has a maximal mean value. Bar length represents the multivariate regression coefficient with either a positive or negative sign for that particular feature of each component, i.e., the importance of each variable in the model.

3.5. Partial Least Squares (PLS) Regression

The relationships between the sensory attributes and the metabolite composition of wines was studied by means of partial least squares (PLS) regression. To that end, PLS1 and PLS2 regression was computed for each analytical platform (LC-MS and GC-MS), which were used as independent variables to predict the response of each sensory attribute (dependent variables). The models were fitted using the all-relevant variables previously selected by means of the MUVR package implemented in R (Supplementary Figure S1). A total of 74 and 247 metabolites for GC and LC platforms were, respectively, selected as relevant, which were obtained by combining the results from a PLS-DA and random forest (RF) models. For both platforms, a better performance was observed for PLS-DA over RF to discriminate and select a major number of relevant compounds.

The root mean squares error (RMSE) and R^2 were used to optimise the number of component of the PLS models. Using PLS1 regression, the sensory descriptors balance, black fruit and citrus fruit were satisfactory predicted by the volatile and non-volatile profiles with R^2 in cross validation higher than 0.72, while good predictions of the descriptors complexity and tree fruit were also obtained using the non-volatile data (Supplementary Figures S2 and S3).

To study the correlation between the sensory data and the volatile profile, PLS2 regression was subsequently performed. The model fitted with the volatile and sensory data was optimised for two components, which explained a 66% of the total variance found in the X matrix. In Supplementary Figure S4, the correlation loadings plot from the PLS2 highlights the contribution of each variable to each component. Greater distances from the origin implies stronger correlations. Thus, variables falling within the inner circumference displayed light associations. In addition, if variables are represented as vectors sourcing from the origin, acute angles between vectors of two variables indicate positive correlations, while obtuse angles indicate negative correlations. The sensory attributes balance, citrus fruit, acidity, tree fruit and alcohol were highly correlated with 3-isopropyl-1-pentanol. Meanwhile, red fruit, complexity and black fruit were positively correlated with 3-methylpentanol, methyl 4-decenoate, 3-ethoxy-1-propanol, (E)-nerolidol, ethyl decanoate, ethyl 2-methylpropanoate, trans-3-hexen-1-ol and 1-dodecanol. Finally, positive correlations were observed between the sensory descriptor persistence and isopentyl hexanoate, ethyl heptanoate and pentyl cyclopropane. Only two sensory descriptors were poorly correlated with the volatile profile data (taste intensity and scent intensity). Only scent intensity fell within the inner circumference, meaning weak correlations of this descriptor with the volatile data. For the LC-MS data, a PLS regression was optimized for 2 components, which explained 80% of the total variance in the X matrix. The sensory attributes balance, citrus fruit, acidity and alcohol were slightly and positively correlated with several amino acids (alanine, glutamic acid, 2-Phenylglycine, L-Alanine, N-propyl-/L-Alloisoleucine/L-Isoleucine/L-norleucine, L-leucyl-L-proline, N-acetyl-L-ornithine, N-acetylproline, phenylalanine, proline betaine and suberylglycine). Meanwhile, persistence, taste intensity and to a lower extent black fruit displayed light positive correlation with cycloleucine, uracil and two organic acids (imidazole acetic acid and shikimic acid). The remaining sensory descriptors were placed within the inner circumference, being again the scent intensity descriptor in this area.

3.6. Data Integration: Regularised Generalised Canonical Correlation Analysis (RGCCA) Approach

Once the relationships between the sensory attributes and the metabolite composition of wines were independently assessed (GC-sensory and LC-sensory), regularized generalized canonical correlation analysis (RGCCA) was performed to integrate the three datasets. The connection of the three blocks was set to a link of 1 to maximise their correlations. As stated in the previous section, a total of 74 and 247 metabolites for GC and LC platforms were, respectively, selected as relevant and were used to built the RGCCA model. Using this approach, all the sensory attributes fell between the two circumferences of the correlation circle plot, showing important correlations with several metabolites (Figure 9). As stated in the PLSR models, citrus fruit, balance, acidity, tree fruit and alcohol descriptors were highly correlated with 3-isopropyl-1-pentanol (A5). In addition, other compounds such as ethyl 5-hexenoate (E3), hexyl acetate (E7), linalyl anthranilate (T3), proline betaine (AA11) and alanine (AA2) were found to be also related with these sensory descriptors. In this approach, scent intensity was allocated between the two circumferences, showing correlations with 2-pyrimidine acetic acid (OA1), 2-phenylglycine (AA4) and trans-3-hexen-1-ol (A7), as well as with complexity, black fruit, red fruit and taste intensity. Meanwhile, the sensory descriptor persistence was negatively correlated with most of the sensory descriptors and positively correlated with imidazole acetic acid (OA4), L-alanine, N-propyl-/L-alloisoleucine/L-isoleucine/L-norleucine (OA6), uracil (M2), 1-undecanol (A3), isopentyl hexanoate (E8), ethyl heptanoate (E1), pentyl-cyclopropane (M1), 1-nonanol (A2) and ethyl hexanoate (E5). The rest of volatile and non-volatile metabolites were allocated within the inner circumference and therefore, no correlations were established.

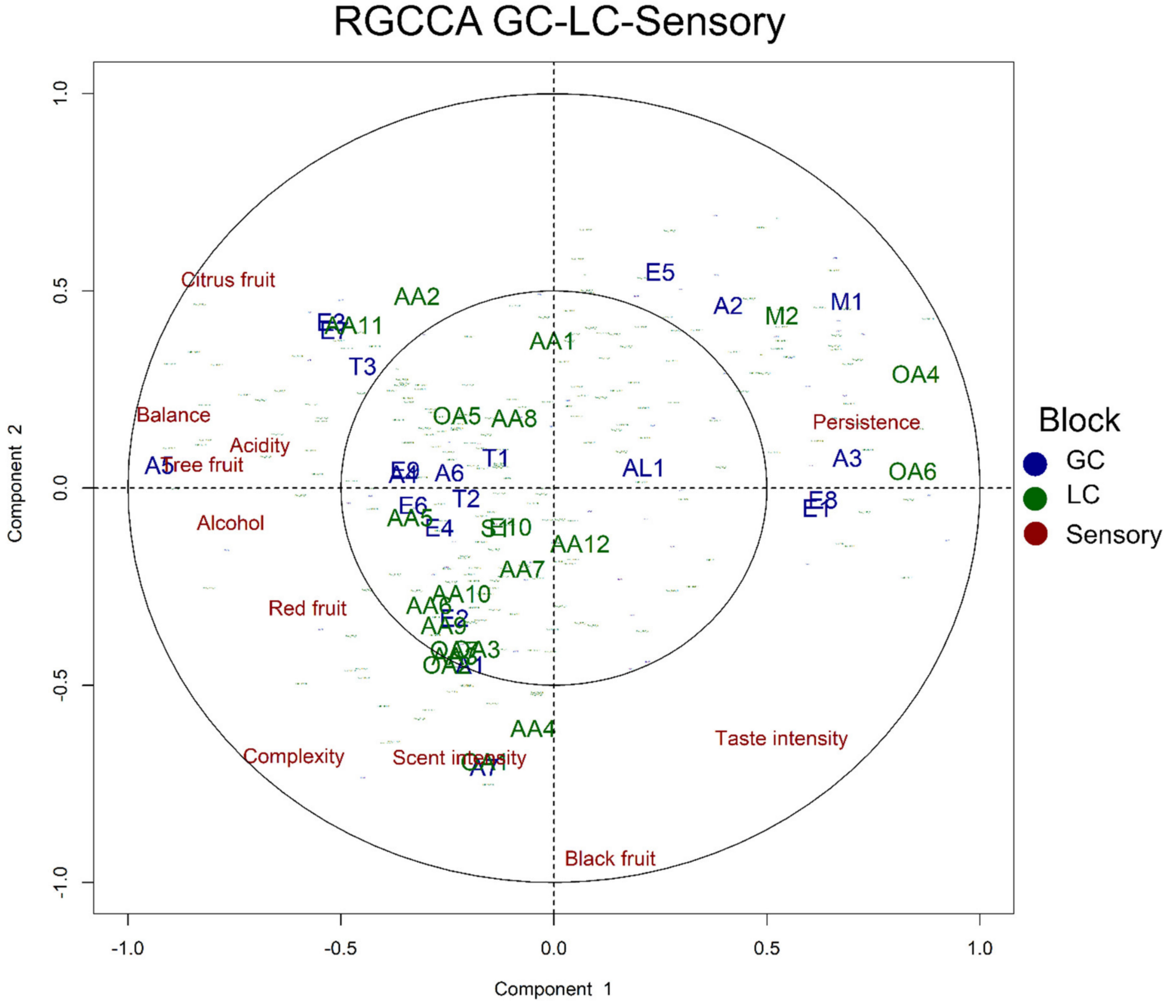

Figure 9. Circle correlation plot from the regularised generalised canonical correlation analysis (RGCCA) performed on the volatile, non-volatile and sensory data. A: alcohol, E: ester, M: miscellaneous, AA: amino acid, OA: organic acid.

4. Discussion

The pre-processing and data pre-treatment workflow proposed in this study for the volatile and non-volatile profile made it possible to extract the signal of a large number of metabolites with low deviation coefficients. The use of automatic pre-processing softwares (IPO-XCMS-CAMERA for LC-MS data and PARAFADISe implementing PARAFAC2 for GC-MS data) with low number of parameters required to be set, diminished the risk of modelling problems. The pre-treatment step included cleaning and filtering of the data (samples and metabolites), drift corrections based on QCs, removing molecular features with high RSD, missing value imputation, PQN normalization and autoscaling, and the quality of the data was checked by means of different PCAs (Figures 2, 4 and 5). The proposed workflow allowed us to adequate the data for the subsequent statistical analyses. The use of quality control samples (QCs) analysed along the analytical sequences, combined with PCA was demonstrated to be a strong tool to verify the quality of the final data. In our case, this methodology was especially useful for the non-volatile data since higher variability in those data was observed (Figures 4 and 5). In particular, LC-MS data acquired in negative ionization mode suffered of an important drift that was revealed by two well separated clusters of QCs in the PCA (Figure 4). On this basis, we recommend the systematic use of QCs for metabolomics studies to both verify the quality of the data and to correct for drifts or batch effects.

The independent (PLS-DA and PLSR) and integrative (sGCCA-DA and RGCCA) statistical analysis led to complementary results of modulation of the wine metabolome and the relationships with their sensory descriptors, obtaining robust potential markers of the fermentation strategy which had important contribution on the sensory characteristics of the wines. To study the link between wine metabolome and sensory descriptors, a previous selection of the all-relevant metabolites involved in the fermentation strategy was performed by means of PLS-DA and RF. This methodology reduced the complexity of the final PLSR and RGCCA models by focusing on the compounds that were modulated by yeasts, making it easier the interpretation.

Interestingly, the markers selected from the sGCCA-DA approach were highly correlated with the sensory descriptors. Several metabolites such as 3-isopropyl-1-pentanol, isopentyl hexanoate, ethyl heptanoate, pentyl-cyclopropane, ethyl trans-3-hexenoate, (E,E)-farnesol, 1-dodecanol and imidazole acetic acid were selected as markers in both the independent and integrative assessments (Figures 3, 6 and 8). However, others such as 1-undecanol (A3) and shikimic acid (OA6), which were very positively correlated with persistence of the wines and to a lower extent with taste intensity, and linalyl anthranilate (T3), ethyl 5-hexenoate (E3), hexyl acetate (E7) and proline betaine (A11), highly correlated with citrus fruit, balance, acidity, tree fruit and alcohol descriptors (Figure 9), were not selected during the independent assessment. Therefore, the integration of both data sets revealed hidden biological information, supporting this approach to complement the classical independent assessment.

Several sensory descriptors displayed high correlation between predicted and observed sensory attribute scores in the PLSR1 assessment (Supplementary Figures S2 and S3), highlighting the quality of the metabolic data to make sensory predictions. The RGCCA also led to additional information hidden during the independent assessments with PLSR. While scent intensity fell within the inner circumference in the PLSR model, intermediate correlations were established with 2-pyrimidine acetic acid (OA1), 2-phenylglycine (AA4) and trans-3-hexen-1-ol (A7), complexity, black fruit, red fruit and taste intensity (Figure 9). These results highlight the potential of using PLSR in combination with RGCCA to extract subtle information that could be ignored in an independent assessment.

The results obtained in this work were in accordance with literature regarding the lower contents of isoamyl acetate, hexyl acetate and medium chain ethyl fatty acids as well as the higher contents of diethyl succinate found in *Sc-Td* wines [32,33]. *Sc-Td* also favoured the production of some terpenes such as linalool and geraniol, in agreement with previous findings [34]. The selected potential marker of *Sc-Td* wines 3-ethoxy-1-propanol, has been described as a compound highly dependent on yeast strain and species, and it was also found in higher concentrations in the wines elaborated with *Td* yeasts [35]. Three of the selected potential markers of *Sc-Td* were esters (ethyl decanoate, methyl 4-decenoate and ethyl 2-methylpropanoate). The formation of esters during fermentation and their concentration relationships with the predominant yeast strains have been widely described in the literature [36]. These compounds are present at low concentrations in wine, often below the aroma threshold concentration, but make a strong contribution to the aroma of wine through synergistic interaction effects. In agreement with our findings, previous studies have reported increases in ethyl decanoate and ethyl 2-methylpropanoate in wines fermented with *Td* yeasts and its contribution was related to the fruity character of wines [37,38]. In addition, we found positive correlations between these esters and the sensory descriptors red fruit, complexity and black fruit in the PLSR model (Supplementary

Figure S4). Several esters not selected as potential markers of the fermentation strategy, such as ethyl 2-methylbutyrate and ethyl phenylacetate, were also found in significantly higher concentrations in rosé wines fermented with *Sc-Td* (Supplementary Table S3). The larger production capacity of *Td* yeasts, a synergistic effect between *Td* and *Sc* [39] or modifications in the nitrogen available for *Sc* species due to *non-Sc* yeasts activity could explain these results. Another volatile compound selected as a potential marker of the fermentation strategy was (E)-nerolidol, a sesquiterpene derived from farnesyl pyrophosphate, which can be produced by *Td* yeasts [40] and was found in significantly higher concentrations in the *Sc-Td* fermented rosé wines (Supplementary Table S3). In addition, several terpenes such as geranyl ethyl ether, (Z)-β-farnesene, nerolidyl acetate, linalool and citronellol were found in higher contents in these wines (Supplementary Table S3). The impact of terpenoids on the grape aroma typicity highlights the importance of the fermentation strategy and the yeasts selected for winemaking to obtain wines with specific sensory characteristics. The selected potential marker 1-dodecanol has pleasant flowery descriptors at low concentrations, but it may give unpleasant aromas to the wine at high concentrations [41]. In our study this compound displayed slightly positive correlations with red fruit, complexity and black fruit. (E,E)-farnesol, which was selected as a potential marker of *Sc* wines is a isoprenoid alcohol with lemon descriptors [42]. This compound was found to be slightly related to the citrus fruit descriptor in the PLSR, although in the integrative approach fell within the inner circumference. The selected potential marker mannitol is a polyol mainly present in high-quality full bodied wines [43], although no important correlations of this metabolite and sensory descriptors were observed. Interactions between grapes and yeasts were found, and in general, the wines fermented with *Sc-Td* had lower fruity descriptors in agreement with literature [26,44]. This reduced fruitiness could be explained by the lack of ageing in the wines analysed, since *Td* tend to enhance fruitiness with time (aged wines) [37].

5. Conclusions

A data processing and statistical analysis pipeline was proposed to study the metabolome modulation of wines with different characteristics and to relate it to their sensory descriptors. This methodology was then applied to study the effect of different fermentation strategies in the elaboration of rosé wines with two grape varieties. A classical independent assessment of the volatile, non-volatile and sensory data made it possible to identify the most relevant metabolites and wine descriptors influenced by the fermentation strategy. Afterwards, this approach was complemented with the statistical integration of the entire metabolome and the sensory descriptors, revealing new potential markers that were highly correlated with the sensory attributes of the wines. A parallel workflow was followed to study the relationships among the metabolome and the sensory attributes of the wines, first by assessing the independent volatile and non-volatile profiles and finally by statistical integration of the three data sets. Strong correlations among sensory descriptors and metabolites were found by means of partial least squares regression and data integration strategies, demonstrating the suitability of combining both methodologies for making robust interpretations of the interrelationships in these complex biological systems.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/fermentation7020072/s1, Figure S1: Balanced error rates (BER) obtained for the partial least squares discriminant analysis (PLS-DA) and random forest (RF) double-cross validated models used to select for the maximum number of relevant variables in GC data (a and b) and LC data (c and d), Figure S2: Partial least squares regression (PLSR) performed to predict the sensory descriptors from the volatile profile, Figure S3: Partial least squares regression (PLSR) performed to predict the sensory descriptors from the non-volatile profile, Figure S4: Correlation loadings plot from the partial least squares regression (PLS2) performed on the GC-sensory and LC-sensory data sets, Table S1: non-volatile metabolites identification results from MetaboQuest using the Metlin database. Positive ionization ESI mode, Table S2: Non-volatile metabolites identification results from MetaboQuest using the Metlin database. Negative ionization ESI mode, Table S3: Two-way analysis of variance (ANOVA) for the inoculation strategy and variety using both the identified and non-identified volatile metabolites, Table S4: Two-way analysis of variance (ANOVA) for the inoculation strategy and variety using both the identified and non-identified non-volatile metabolites.

Author Contributions: Conceptualization, J.M.M.-R. (José Manuel Muñoz-Redondo), B.P., E.C.-V. and J.M.M.-R. (José Manuel Moreno-Rojas); methodology, J.M.M.-R. (José Manuel Muñoz-Redondo), J.L.O.-D., G.P.-C., M.J.R.-M.; software, J.M.M.-R. (José Manuel Muñoz-Redondo); validation, J.M.M.-R. (José Manuel Muñoz-Redondo); formal analysis, J.M.M.-R. (José Manuel Muñoz-Redondo); investigation, B.P., E.C.-V., G.P.-C., J.M.M.-R. (José Manuel Muñoz-Redondo), M.J.R.-M. and J.L.O.-D.; resources, G.P.-C.; data curation, J.M.M.-R. (José Manuel Muñoz-Redondo); writing—original draft preparation, J.M.M.-R. (José Manuel Muñoz-Redondo); writing—review and editing, B.P., E.C.-V., and J.M.M.-R. (José Manuel Moreno-Rojas); visualization, J.M.M.-R. (José Manuel Muñoz-Redondo) and J.M.M.-R. (José Manuel Moreno-Rojas); supervision, J.M.M.-R. (José Manuel Moreno-Rojas); project administration, J.M.M.-R. (José Manuel Moreno-Rojas), G.P.-C., B.P. and E.C.-V.; funding acquisition, B.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Andalusian Institute of Agricultural and Fisheries Research and Training (IFAPA) through the Project "Investigacioón e Innovación Tecnoloógica en Vitivinicultura" (PR.AVA.AVA2019.016) and the European Rural Development Fund (ERDF, EU). J. M. Muñoz-Redondo and J. L. Ordoñez-Diaz were awarded a research contract funded by the Andalusian Institute of Agricultural and Fisheries Research and Training (IFAPA), within the National Youth Guarantee System funded through the European Social Fund (ESF) and the Youth Employment Initiative (YEI). G. Pereira-Caro was supported by a postdoctoral research contract funded by IFAPA and the ESF (03/2014 to 03/2017) and by a postdoctoral "Juan de la Cierva-Incorporación" research contract funded by the Spanish Ministry of Economy, Industry and Competitiveness (FJCI-2015-26433; 04/2017 to 03/2019). M. J. Ruiz-Moreno was supported by a research contract funded by IFAPA and the ESF (09/2012 to 08/2017).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Siegrist, M.; Cousin, M.-E. Expectations Influence Sensory Experience in a Wine Tasting. *Appetite* **2009**, *52*, 762–765. [CrossRef] [PubMed]
2. Sherman, E.; Coe, M.; Grose, C.; Martin, D.; Greenwood, D.R. Metabolomics Approach to Assess the Relative Contributions of the Volatile and Non-Volatile Composition to Expert Quality Ratings of Pinot Noir Wine Quality. *J. Agric. Food Chem.* **2020**, *68*, 13380–13396. [CrossRef] [PubMed]
3. Li, Z.; Lu, Y.; Guo, Y.; Cao, H.; Wang, Q.; Shui, W. Comprehensive Evaluation of Untargeted Metabolomics Data Processing Software in Feature Detection, Quantification and Discriminating Marker Selection. *Anal. Chim. Acta* **2018**, *1029*, 50–57. [CrossRef]
4. Arapitsas, P.; Ugliano, M.; Marangon, M.; Piombino, P.; Rolle, L.; Gerbi, V.; Versari, A.; Mattivi, F. Use of Untargeted Liquid Chromatography–Mass Spectrometry Metabolome to Discriminate Italian Monovarietal Red Wines, Produced in Their Different Terroirs. *J. Agric. Food Chem.* **2020**, *68*, 13353–13366. [CrossRef]
5. Šuklje, K.; Carlin, S.; Stanstrup, J.; Antalick, G.; Blackman, J.W.; Meeks, C.; Deloire, A.; Schmidtke, L.M.; Vrhovsek, U. Unravelling Wine Volatile Evolution during Shiraz Grape Ripening by Untargeted HS-SPME-GC × GC-TOFMS. *Food Chem.* **2019**, *277*, 753–765. [CrossRef] [PubMed]
6. Pinu, F.R. Grape and Wine Metabolomics to Develop New Insights Using Untargeted and Targeted Approaches. *Fermentation* **2018**, *4*, 92. [CrossRef]
7. Rocchetti, G.; Gatti, M.; Bavaresco, L.; Lucini, L. Untargeted Metabolomics to Investigate the Phenolic Composition of Chardonnay Wines from Different Origins. *J. Food Compos. Anal.* **2018**, *71*, 87–93. [CrossRef]
8. Whitener, M.E.B.; Stanstrup, J.; Panzeri, V.; Carlin, S.; Divol, B.; Du Toit, M.; Vrhovsek, U. Untangling the Wine Metabolome by Combining Untargeted SPME–GCxGC-TOF-MS and Sensory Analysis to Profile Sauvignon Blanc Co-Fermented with Seven Different Yeasts. *Metabolomics* **2016**, *12*, 53. [CrossRef]
9. Arbulu, M.; Sampedro, M.C.; Gómez-Caballero, A.; Goicolea, M.A.; Barrio, R.J. Untargeted Metabolomic Analysis Using Liquid Chromatography Quadrupole Time-of-Flight Mass Spectrometry for Non-Volatile Profiling of Wines. *Anal. Chim. Acta* **2015**, *858*, 32–41. [CrossRef]

10. Castro, C.C.; Martins, R.C.; Teixeira, J.A.; Ferreira, A.C.S. Application of a High-Throughput Process Analytical Technology Metabolomics Pipeline to Port Wine Forced Ageing Process. *Food Chem.* **2014**, *143*, 384–391. [CrossRef]
11. Goodacre, R.; Broadhurst, D.; Smilde, A.K.; Kristal, B.S.; Baker, J.D.; Beger, R.; Bessant, C.; Connor, S.; Capuani, G.; Craig, A. Proposed Minimum Reporting Standards for Data Analysis in Metabolomics. *Metabolomics* **2007**, *3*, 231–241. [CrossRef]
12. Johnsen, L.G.; Skou, P.B.; Khakimov, B.; Bro, R. Gas Chromatography–Mass Spectrometry Data Processing Made Easy. *J. Chromatogr. A* **2017**, *1503*, 57–64. [CrossRef]
13. Amigo, J.M.; Skov, T.; Bro, R.; Coello, J.; Maspoch, S. Solving GC-MS Problems with Parafac2. *TrAC Trends Anal. Chem.* **2008**, *27*, 714–725. [CrossRef]
14. Mahieu, N.G.; Genenbacher, J.L.; Patti, G.J. A Roadmap for the XCMS Family of Software Solutions in Metabolomics. *Curr. Opin. Chem. Biol.* **2016**, *30*, 87–93. [CrossRef] [PubMed]
15. Considine, E.C.; Salek, R.M. A Tool to Encourage Minimum Reporting Guideline Uptake for Data Analysis in Metabolomics. *Metabolites* **2019**, *9*, 43. [CrossRef]
16. Ren, S.; Hinzman, A.A.; Kang, E.L.; Szczesniak, R.D.; Lu, L.J. Computational and Statistical Analysis of Metabolomics Data. *Metabolomics* **2015**, *11*, 1492–1513. [CrossRef]
17. Worley, B.; Powers, R. Multivariate Analysis in Metabolomics. *Curr. Metab.* **2013**, *1*, 92–107.
18. Rohart, F.; Gautier, B.; Singh, A.; Le Cao, K.-A. MixOmics: An R Package for ‘omics Feature Selection and Multiple Data Integration. *PLoS Comput. Biol.* **2017**, *13*, e1005752. [CrossRef] [PubMed]
19. Wanichthanarak, K.; Fahrmann, J.F.; Grapov, D. Genomic, Proteomic, and Metabolomic Data Integration Strategies. *Biomark. Insights* **2015**, *10*, BMI–S29511. [CrossRef]
20. Sumner, L.W.; Amberg, A.; Barrett, D.; Beale, M.H.; Beger, R.; Daykin, C.A.; Fan, T.W.-M.; Fiehn, O.; Goodacre, R.; Griffin, J.L.; et al. Proposed Minimum Reporting Standards for Chemical Analysis. *Metabolomics* **2007**, *3*, 211–221. [CrossRef] [PubMed]
21. Smith, C.A.; Want, E.J.; O’Maille, G.; Abagyan, R.; Siuzdak, G. XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification. *Anal. Chem.* **2006**, *78*, 779–787. [CrossRef] [PubMed]
22. Libiseller, G.; Dvorzak, M.; Kleb, U.; Gander, E.; Eisenberg, T.; Madeo, F.; Neumann, S.; Trausinger, G.; Sinner, F.; Pieber, T.; et al. IPO: A Tool for Automated Optimization of XCMS Parameters. *BMC Bioinform.* **2015**, *16*. [CrossRef] [PubMed]
23. Kuhl, C.; Tautenhahn, R.; Böttcher, C.; Larson, T.R.; Neumann, S. CAMERA: An Integrated Strategy for Compound Spectra Extraction and Annotation of LC/MS Data Sets. *Anal. Chem.* **2012**, *84*, 283–289. [CrossRef]
24. Arapitsas, P.; Della Corte, A.; Gika, H.; Narduzzi, L.; Mattivi, F.; Theodoridis, G. Studying the Effect of Storage Conditions on the Metabolite Content of Red Wine Using HILIC LC–MS Based Metabolomics. *Food Chem.* **2016**, *197*, 1331–1340. [CrossRef] [PubMed]
25. Puertas, B.; Jimenez-Hierro, M.J.; Cantos-Villar, E.; Marrufo-Curtido, A.; Carbú, M.; Cuevas, F.J.; Moreno-Rojas, J.M.; González-Rodríguez, V.E.; Cantoral, J.M.; Ruiz-Moreno, M.J. The Influence of Yeast on Chemical Composition and Sensory Properties of Dry White Wines. *Food Chem.* **2018**, *253*, 227–235. [CrossRef]
26. Kokla, M.; Klåvus, A.; Noerman, S.; Koistinen, V.M.; Tuomainen, M.; Zarei, I.; Meuronen, T.; Häkkinen, M.R.; Rummukainen, S.; Babu, A.F. “NoTaMe”: Workflow for Non-Targeted LC-MS Metabolic Profiling. *Metabolites* **2020**, *10*, 135.
27. Kokla, M.; Virtanen, J.; Kolehmainen, M.; Paananen, J.; Hanhineva, K. Random Forest-Based Imputation Outperforms Other Methods for Imputing LC-MS Metabolomics Data: A Comparative Study. *BMC Bioinform.* **2019**, *20*, 492. [CrossRef] [PubMed]
28. Noonan, M.J.; Tinnesand, H.V.; Buesching, C.D. Normalizing Gas-Chromatography–Mass Spectrometry Data: Method Choice Can Alter Biological Inference. *BioEssays* **2018**, *40*, 1700210. [CrossRef]
29. Szymańska, E.; Saccenti, E.; Smilde, A.K.; Westerhuis, J.A. Double-Check: Validation of Diagnostic Statistics for PLS-DA Models in Metabolomics Studies. *Metabolomics* **2012**, *8*, 3–16. [CrossRef]
30. Muñoz-Redondo, J.M.; Ruiz-Moreno, M.J.; Puertas, B.; Cantos-Villar, E.; Moreno-Rojas, J.M. Multivariate Optimization of Headspace Solid-Phase Microextraction Coupled to Gas Chromatography-Mass Spectrometry for the Analysis of Terpenoids in Sparkling Wines. *Talanta* **2020**, *208*, 120483. [CrossRef]
31. Shi, L.; Westerhuis, J.A.; Rosén, J.; Landberg, R.; Brunius, C. Variable Selection and Validation in Multivariate Modelling. *Bioinformatics* **2019**, *35*, 972–980. [CrossRef] [PubMed]
32. Sadoudi, M.; Tourdot-Maréchal, R.; Rousseaux, S.; Steyer, D.; Gallardo-Chacón, J.-J.; Ballester, J.; Vichi, S.; Guérin-Schneider, R.; Caixach, J.; Alexandre, H. Yeast–Yeast Interactions Revealed by Aromatic Profile Analysis of Sauvignon Blanc Wine Fermented by Single or Co-Culture of Non-Saccharomyces and Saccharomyces Yeasts. *Food Microbiol.* **2012**, *32*, 243–253. [CrossRef]
33. Renault, P.; Miot-Sertier, C.; Marullo, P.; Hernández-Orte, P.; Lagarrigue, L.; Lonvaud-Funel, A.; Bely, M. Genetic Characterization and Phenotypic Variability in Torulaspora Delbrueckii Species: Potential Applications in the Wine Industry. *Int. J. Food Microbiol.* **2009**, *134*, 201–210. [CrossRef]
34. Benito, S. The Impact of Torulaspora Delbrueckii Yeast in Winemaking. *Appl. Microbiol. Biotechnol.* **2018**, *102*, 3081–3094. [CrossRef] [PubMed]
35. Velázquez, R.; Zamora, E.; Álvarez, M.L.; Ramírez, M. Using Torulaspora Delbrueckii Killer Yeasts in the Elaboration of Base Wine and Traditional Sparkling Wine. *Int. J. Food Microbiol.* **2019**, *289*, 134–144. [CrossRef] [PubMed]
36. Sumby, K.M.; Grbin, P.R.; Jiranek, V. Microbial Modulation of Aromatic Esters in Wine: Current Knowledge and Future Prospects. *Food Chem.* **2010**, *121*, 1–16. [CrossRef]

37. Oliveira, I.; Ferreira, V. Modulating Fermentative, Varietal and Aging Aromas of Wine Using Non-Saccharomyces Yeasts in a Sequential Inoculation Approach. *Microorganisms* **2019**, *7*, 164. [CrossRef] [PubMed]
38. Renault, P.; Coulon, J.; de Revel, G.; Barbe, J.-C.; Bely, M. Increase of Fruity Aroma during Mixed T. Delbrueckii/S. Cerevisiae Wine Fermentation Is Linked to Specific Esters Enhancement. *Int. J. Food Microbiol.* **2015**, *207*, 40–48. [CrossRef] [PubMed]
39. Gobert, A.; Tourdot-Maréchal, R.; Morge, C.; Sparrow, C.; Liu, Y.; Quintanilla-Casas, B.; Vichi, S.; Alexandre, H. Non-Saccharomyces Yeasts Nitrogen Source Preferences: Impact on Sequential Fermentation and Wine Volatile Compounds Profile. *Front. Microbiol.* **2017**, *8*, 2175. [CrossRef]
40. King, A.; Richard Dickinson, J. Biotransformation of Monoterpene Alcohols by Saccharomyces Cerevisiae, Torulaspora Delbrueckii and Kluyveromyces Lactis. *Yeast* **2000**, *16*, 499–506. [CrossRef]
41. Jiang, B.; Zhang, Z. Volatile Compounds of Young Wines from Cabernet Sauvignon, Cabernet Gernischet and Chardonnay Varieties Grown in the Loess Plateau Region of China. *Molecules* **2010**, *15*, 9184–9196. [CrossRef] [PubMed]
42. Coelho, E.; Coimbra, M.A.; Nogueira, J.M.F.; Rocha, S.M. Quantification Approach for Assessment of Sparkling Wine Volatiles from Different Soils, Ripening Stages, and Varieties by Stir Bar Sorptive Extraction with Liquid Desorption. *Anal. Chim. Acta* **2009**, *635*, 214–221. [CrossRef] [PubMed]
43. Soetaert, W.; Vanhooren, P.T.; Vandamme, E.J. The production of mannitol by fermentation. In *Carbohydrate Biotechnology Protocols*; Springer: Berlin/Heidelberg, Germany, 1999; pp. 261–275.
44. Azzolini, M.; Tosi, E.; Lorenzini, M.; Finato, F.; Zapparoli, G. Contribution to the Aroma of White Wines by Controlled Torulaspora Delbrueckii Cultures in Association with Saccharomyces Cerevisiae. *World J. Microbiol. Biotechnol.* **2015**, *31*, 277–293. [CrossRef] [PubMed]