



HAL
open science

Preliminary authentication of apple juices using untargeted UHPLC-HRMS analysis combined to chemometrics

Katy Dinis, Lucie Tsamba, Freddy Thomas, Eric Jamin, Valérie Camel

► To cite this version:

Katy Dinis, Lucie Tsamba, Freddy Thomas, Eric Jamin, Valérie Camel. Preliminary authentication of apple juices using untargeted UHPLC-HRMS analysis combined to chemometrics. *Food Control*, 2022, 139, pp.109098. 10.1016/j.foodcont.2022.109098 . hal-04096835

HAL Id: hal-04096835

<https://hal.science/hal-04096835>

Submitted on 9 Jul 2023

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.

Journal Pre-proof

Preliminary authentication of apple juices using untargeted UHPLC-HRMS analysis combined to chemometrics

Katy Dinis, Lucie Tsamba, Freddy Thomas, Eric Jamin, Valérie Camel



PII: S0956-7135(22)00291-2

DOI: <https://doi.org/10.1016/j.foodcont.2022.109098>

Reference: JFCO 109098

To appear in: *Food Control*

Received Date: 8 December 2021

Revised Date: 4 April 2022

Accepted Date: 8 May 2022

Please cite this article as: Dinis K., Tsamba L., Thomas F., Jamin E. & Camel Valé., Preliminary authentication of apple juices using untargeted UHPLC-HRMS analysis combined to chemometrics, *Food Control* (2022), doi: <https://doi.org/10.1016/j.foodcont.2022.109098>.

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2022 Published by Elsevier Ltd.

Credit Author Statement:

Katy Dinis: Conceptualization, Investigation, Formal analysis, Writing – Original Draft, **Lucie Tsamba:** Conceptualization, Validation, Writing – Review and editing, Project administration **Freddy Thomas:** Funding acquisition **Eric Jamin:** Writing – Review and editing, Funding acquisition, Project administration **Valérie Camel:** Writing – Review and editing, Supervision

Journal Pre-proof

1 **Preliminary authentication of apple juices using untargeted UHPLC-HRMS analysis**
2 **combined to chemometrics**

3

4 Katy Dinis ^{a,b}, Lucie Tsamba ^{a*}, Freddy Thomas ^a, Eric Jamin ^a, Valérie Camel ^b

5

6 ^a Eurofins Analytics France, 9 rue Pierre Adolphe Bobierre, B.P. 42301, F-44323, Nantes Cedex 3,
7 France

8 ^b UMR SayFood, Université Paris-Saclay, INRAE, AgroParisTech, 91300 Massy, France

9

10 * Corresponding author: Eurofins Analytics France, 9 rue Pierre Adolphe Bobierre, B.P. 42301, F-
11 44323 NANTES Cedex 3, France, tel.: +33 2 51 82 55 39, fax: +33 2 51 83 21 11. E-mail address:
12 LucieTsamba@eurofins.com

13 Abstract

14 In this work, apple juice samples from different farming and production processes (direct and
15 concentrated juices; organic and conventional juices) were analyzed by ultra-high performance liquid
16 chromatography coupled to high resolution mass spectrometry (UHPLC-HRMS). A workflow was
17 developed and implemented for data processing using the Workflow4Metabolomics (W4M)
18 platform. First, features were detected using XCMS, and next data filtration steps were applied
19 leading to the removal of nearly 50% of the detected features. Intra- and inter-batch correction was
20 then performed, followed by chemometric tools (PCA, PLS-DA, OPLS-DA, ANOVA). The developed
21 approach successfully discriminated apple juice samples in two distinct scenarios simultaneously
22 (direct vs. concentrated juices and organic vs. conventional juices). PCA highlighted the
23 reproducibility of the method and confirmed the efficiency of batch corrections. OPLS-DA models
24 showed good quality metrics, particularly after feature selection for organic vs. conventional juices
25 discrimination (almost 80% of predictive ability). Based on ANOVA and OPLS-DA results, 24 features
26 were retained as significantly discriminant. Among them, some compounds were identified as amino-
27 acids and derivatives, using additional MS/MS experiments and online databases. An independent
28 data set was used to evaluate their potential as marker compounds, with promising results obtained.
29 Further investigation is needed to validate such an untargeted method and its routine application to
30 detect apple juice adulteration and confirm its authenticity.

31

32 Keywords

33 Food authenticity, High resolution mass spectrometry, Liquid chromatography, Metabolomics, PCA,
34 PLS-DA, OPLS-DA

35

36 Highlights

- 37 ○ Development of an UHPLC-HRMS metabolomics approach with great potential in juice
- 38 authentication
- 39 ○ Discrimination between sample groups in two distinct authentication applications
- 40 ○ Relevant markers selected by OPLS-DA and ANOVA were tentatively identified
- 41 ○ Main discriminant compounds were identified as amino-acids and derivatives

42

43 1. Introduction

44 Food fraud is a worldwide issue, and recent crises (like the horse meat scandal in 2013) have sparked
45 interest on food authentication among consumers and food industries [Brooks et al., 2017].
46 Concerning food fraud and Economically Motivated Adulteration between 1980 and 2010, fruit juices
47 are one of the top ten products most at risk, particularly apple and orange juices [Moore et al.,
48 2012]. Typical frauds on fruit juices include (1) dilution with water, (2) addition of sugars or organic
49 acids, (3) addition of foreign fruits (mostly cheaper ones) and (4) false labeling of the product
50 (cultivar or geographical origin, as well as production mode such as organic) [Vaclavik et al., 2011].

51 To facilitate the detection of food fraud, the Codex Alimentarius, the Association of the Industry of
52 Juices and Nectars of the European Union (AIJN) and the European Commission have established
53 guidelines and standards to define permitted practices and evaluate the quality and authenticity of
54 juices [Directive 2012/12/EC; CODEX STAN 247-2005; AIJN Code of practice]. However, fruit juices

55 authentication may be challenging due to their complex chemical composition influenced by several
56 factors such as variety, geographical origin, stage of maturity, storage conditions and processing
57 techniques [Jandric et al., 2014; Cubero-Leon et al., 2018; Dasenaki et al., 2019].

58 Juice authentication is routinely performed by conventional analytical methods (called targeted
59 methods) that are usually described and validated by the IFU (International Fruit and Vegetable Juice
60 Association) [IFU website]. For example, sugars, organic acids, minerals, phenolic compounds and
61 several volatile compounds are analyzed to authenticate direct apple juice samples [AIJN Code of
62 practice, Wolter et al., 2008]. These methods are sensitive and usually provide low limits of detection
63 and quantification as they have been developed to detect specific compounds or classes of
64 compounds (e.g., molecular markers of foreign fruits or low fruit content). However, these targeted
65 approaches generally focus on a specific fraud and may fail to reveal more sophisticated frauds such
66 as false organic claims [Knolhoff and Croley, 2016; Dasenaki et al., 2019]. The illegal addition of
67 vegetable water (such as water obtained during the grape juice concentration process) to orange
68 juice concentrate, with the false claim of "orange juice not from concentrate", is another illustrative
69 example since conventional $^{18}\text{O}/^{16}\text{O}$ isotope ratio analysis fails to detect this fraud; in that case, there
70 is an additional health concern related to the presence of allergenic sulphur dioxide [Rinke and
71 Jamin, 2018].

72 Therefore, it is important to move toward untargeted methods to detect adulteration and confirm
73 authenticity [Dasenaki et al., 2019; Rinke, 2016]. Untargeted methods allow to have an overview of
74 the sample, also called a fingerprint [Medina et al., 2019]. Thousands of compounds can be detected,
75 making them more holistic than the conventional methods [Dasenaki et al., 2019]. These untargeted
76 methods have emerged with the improvements of analytical techniques (e.g., the development of
77 high resolution mass spectrometers) and the use of advanced statistical methods. Nuclear magnetic
78 resonance (NMR) and mass spectrometry (MS) are widely used in the assessment of food
79 authentication using untargeted methodology, in particular liquid chromatography coupled to high
80 resolution mass spectrometry (LC-HRMS) [Cubero-Leon et al., 2014; Sobolev et al., 2019; Esteki et al.,
81 2018; Danezis et al., 2016].

82 Metabolomics-based approaches using LC-HRMS have already been used in food safety assessment
83 [Knolhoff et al., 2016; Delaporte et al., 2019] and revealed their potential. In the field of food
84 authentication, untargeted LC-HRMS analysis coupled to chemometrics has been used to attest the
85 geographical origin of saffron (100% of the investigated samples were correctly classified) [Rubert et
86 al., 2016]. Using a similar methodology, Cavanna and co-workers have assessed the authentication of
87 durum wheat based on geographical origin, with approximately 90% of samples correctly classified
88 [Cavanna et al., 2020]. Moreover, metabolomics-based methodology using LC-HRMS has already
89 been successfully implemented for juice authentication regarding geographical origin [Diaz et al.,
90 2014] or for adulteration detection and classification of juices types and varieties [Vaclavik et al.,
91 2012; Jandric et al., 2014; Jandric et al., 2017]. Similarly, Dubin et al. used this methodology to
92 authenticate blackcurrant, specifically to detect adulteration with aronia, with a detection limit of 5%
93 aronia concentrate in blackcurrant concentrate [Dubin et al., 2017]. Furthermore, the untargeted
94 methodology was also used for pomegranate juice authentication allowing detection of 1%
95 adulteration [Dasenaki et al., 2019].

96 Thereby, the metabolomics-based methodology appears to be a method of choice for juice
97 authentication. However, this trend deserves confirmation and further methodological development.
98 In particular, studies with large data sets and/or with models that offer broad applicability are
99 needed to validate the potential of this methodology for food authentication [Cubero-Leon et al.,
100 2018]. Moreover, the authentication of organic food also requires further work due to limited

101 number of studies regarding this topic, especially in the juice sector [Cuevas et al., 2017; Mihailova et
102 al., 2021], and the lack of reliable analytical techniques to confirm the organic production of a sample
103 [Cuevas et al., 2019].

104 In this work, apple juice samples from different farming and production processes (organic and
105 conventional, direct and from concentrate juices) were analyzed using untargeted UHPLC-HRMS
106 analysis. A data processing workflow was developed to select relevant features after peak detection.
107 Based on these features, models were built for sample groups discrimination using chemometric
108 tools in two distinct scenarios (direct juice vs. concentrated juice, and organic juice vs. conventional
109 juice). Chemical markers allowing the discrimination were then tentatively identified, using online
110 and in-house databases as well as UHPLC-HRMS/MS analyses. The discriminant potential of these
111 marker compounds was evaluated using an independent set of samples.

112 2. Materials and methods

113 2.1. Reagents and chemicals

114 Methanol (MeOH), water and formic acid (FA), all LC-MS grade, were purchased from Fisher
115 Scientific.

116 Some compounds known to be present in apple juice, and routinely analyzed by targeted methods,
117 were purchased from Sigma-Aldrich: alpha-terpineol (purity: >99%), hexyl acetate (purity: 99%), ethyl
118 2-methylbutyrate (purity: > 98%), limonene (purity: 97%), phloridzin (purity: > 99%) and 2-
119 methylbutyl acetate (purity: > 99%). Hydroxymethylfurfural (purity: 100%) was purchased from
120 ACROS. These commercial standards were considered to assess the ability of our untargeted method
121 to detect them. In addition, they may be considered as possible candidates for markers responsible
122 of the discrimination between our sample groups. With the aim to develop a real untargeted
123 method, these target compounds were not included in our inclusion list in our MS/MS experiments.
124 Individual standard solutions were prepared in methanol with a concentration of 0.2 mg/L for most
125 compounds, and of 0.5 mg/L for hydroxymethylfurfural and phloridzin. These solutions were
126 analyzed using the UHPLC-HRMS analytical conditions described in section 2.3, in order to determine
127 the m/z and retention time (RT) of the compounds which will be used to highlight their potential
128 presence in the analyzed samples.

129 2.2. Samples description and preparation

130 One hundred and ten apple juice samples from several geographical origins and farming processes
131 were collected (organic and non-organic juices; direct juices, concentrated juices and juices from
132 concentrate). Samples were stored in the freezer until analysis. After thawing, aliquots (5 ml) of
133 samples were centrifuged for 10 min at 4,500 rpm. The supernatant was collected and diluted with
134 water directly into a vial before analysis. Three replicates per sample were prepared. Sample vials
135 were randomized in the analytical sequence. Quality Control (QC) samples (pool of apple juice
136 samples) and diluted QC samples were also prepared and analyzed every 10 injections. The repeated
137 injections of QC samples were used to evaluate analytical performance. Also, analytical blanks were
138 analyzed regularly to check for carry over (every 20 samples). Moreover, these blanks were useful to
139 detect residual peaks corresponding to the mobile phases used.

140 Samples were analyzed in different batches. The first one contained 24 samples (12 organic and 12
141 conventional apple juices). The second batch contained 30 samples (15 direct apple juices and 15
142 concentrated apple juices). The third batch contained 26 samples including organic and conventional
143 juices as well as direct and concentrated juices. Another set of samples coming from a different
144 harvest year was also analyzed in a fourth batch in which MS/MS acquisition was performed; this

145 batch contained 30 samples (10 concentrated juice samples, 10 conventional direct juice samples
146 and 10 organic direct juice samples). A detailed list of the samples is presented in Table A.1
147 (Supplementary material).

148 *2.3. Analytical method*

149 Analyses were performed on a ThermoFisher® Vanquish Flex UHPLC system, composed of a binary
150 pump, refrigerated sampler and column oven, connected to a ThermoFisher® QExactive Plus
151 Orbitrap® high resolution mass spectrometer (version 2.9) with a heated electrospray ion source
152 (HESI). The UHPLC separation was achieved using a C18 Hypersil Gold column (150 x 2.1 mm, 1.9 µm)
153 at a 0.3 mL/min flow-rate. The column temperature was set to 30°C. The mobile phases were water
154 acidified with 0.1% FA (A), and MeOH acidified with 0.1% FA (B), with the following linear gradient
155 elution: 0-2 min, B: 3%; 2-20 min, B: 3-98%; 20-24 min: B: 98%; 24-24.1 min, B: 98-3%; 24.1-32 min,
156 B: 3%. The injection volume was 1 µL.

157 Raw data were acquired using TraceFinder software (version 3.1, ThermoFisher®). MS data were
158 acquired in positive ion mode (ESI+) with a mass range set at m/z 120-1000 in full scan mode and
159 with a resolution of 70,000. The parameters applied on the electrospray ion source are presented in
160 Table A.2 (Supplementary material); MS data was acquired in centroid mode. The MS detector was
161 weekly calibrated using the Pierce™ positive and negative ion calibration solution purchased from
162 Thermo Fisher Scientific.

163 For MS/MS acquisition, full scan data-dependent analyses were carried out using an inclusion list.
164 This inclusion list was established after the data processing of the first three batches where several
165 features were identified as discriminant (24 features for both studies). The resolution was set at
166 17,500. An isolation window of ± 1 uma was used to select the m/z of interest at the expected
167 retention time of the features (± 1 min). Three normalized collision energies were applied (10; 30 and
168 60 eV) for the MS/MS spectrum acquisition.

169 *2.4. Data processing*

170 Raw data files were analyzed using the Workflow4Metabolomics (W4M) platform (version 3.0)
171 [Giacomini et al., 2015] after conversion of the data files to mzXML format using ProteoWizard
172 [Chambers et al., 2012]. The main steps of data processing are: (1) peak detection; (2) retention time
173 alignment; (3) peaks grouping; (4) peak annotation; (5) data filtration and normalization, and (6)
174 chemometric analysis. The first four steps were performed using functions of the XCMS (an acronym
175 for various forms (X) of chromatography mass spectrometry) package [Smith et al., 2006] on the
176 W4M platform as illustrated in Fig. 1.

177 The features, defined by their m/z and retention time, and their intensities in different samples were
178 used for the statistical analysis as commonly reported [Cavanna et al., 2018]. The chemometric
179 methods used were principal component analysis (PCA) for exploratory purpose, as well as partial
180 least squares - discriminant analysis (PLS-DA), and orthogonal partial least squares - discriminant
181 analysis (OPLS-DA) in order to build models for discrimination and classification of samples groups.
182 Also, analysis of variance (ANOVA) and biosigner were used to reduce the number of features
183 selected for models building, which may improve models quality.

184 *2.4.1. Peak detection and alignment (XCMS)*

185 All the data from the first three analytical batches were processed simultaneously as illustrated in
186 Fig. 1. The XCMS phase includes the following steps. First, the peak detection and extraction is
187 achieved using the "findChromPeaks" function with the centWave method [Tautenhahn et al., 2008].

188 During this step, the chromatograms are described as a 2D-matrix where each peak is described by a
189 combination of its m/z value and retention time (RT), called "feature". The selected retention time
190 for each peak is the time corresponding to its apex of the intensity value. Then, the
191 "groupChromPeaks" function is used to group the extracted peaks across all the samples. This step is
192 applied to group ions with close RT between the samples. After this step, m/z and RT values are
193 averaged in the data matrix. The peaks are next aligned using the "adjustRtime" function to correct
194 the RT across the samples and then grouped again. Finally, the "fillChromPeaks" function is used to
195 identify features where there is no intensity value for some samples and the signal is integrated in
196 the region of the determined feature to avoid missing values. The XCMS parameters for each step
197 were optimized from the QC samples and are presented in Table A.3 (Supplementary material). A
198 data matrix is then generated, giving the area of each peak for each feature and for each sample.
199 Thus, the features are the variables of the models presented in this study.

200 2.4.2. Data filtration and batch correction

201 To perform the subsequent data filtration step, the data matrix was split in three distinct data
202 matrices, as shown in Fig. 1, corresponding to the three initial analytical batches, in order to perform
203 filtration steps within each analytical batch. These filtration steps were needed to remove irrelevant
204 information as the number of features detected by XCMS was very high (about 20,000).

205 First, all peaks corresponding to the dead volume and the column flush were excluded, which means
206 that all the features with a retention time lower than 1.7 min were removed from the data matrix.
207 Then, features that mainly result from blank analyses were removed by calculating the fold change in
208 blanks and samples analyses. For a feature, when the ratio of samples fold change over blanks fold
209 change is lower than 4, this feature is deleted. In this way, between 15% and 20% of the detected
210 features were removed. Finally, features showing a poor stability (relative standard deviation (RSD)
211 higher than 30%) according to QC analyses were also excluded. Similarly, features for which the ratio
212 of RSD pool over RSD sample is higher than 1.25 were deleted. At the end of this step, about 10,000
213 features remained. Analytical signal drift within the analytical batch was corrected using a LOESS
214 regression model using the QC sample injections, employing the *Batch Correction* module on the
215 W4M platform.

216 Then, the three data matrices corresponding to the three sample batches were merged (see Fig. 1)
217 and a second batch correction was applied to correct analytical signal drift between analytical
218 batches by the use of the QC sample injections.

219 The data matrix was then normalized using the Probabilistic Quotient Normalization method (PQN)
220 [Dieterle et al., 2006] using the QC samples. Its purpose is to limit potential dilution effects that can
221 affect restricted regions of the data. First, the median of each feature in QC samples is calculated,
222 providing a reference vector. Then, the values for each ion in samples are divided by this reference
223 vector. A median of the ratios for each sample is generated. Finally, initial values of each sample are
224 divided by the ratios median.

225 Prior to chemometrics analysis, the data matrix was Pareto scaled. Then, the data matrix was split to
226 create two distinct authentication studies: the first one contained samples from batches 2 and 3 to
227 evaluate the discrimination between pure and concentrated juice samples (58 samples in the data
228 set); the second study contained samples from batches 1 and 3 to evaluate the discrimination
229 between organic and conventional juice samples (54 samples in the data set).

230 2.4.3. Chemometrics

231 Multivariate statistical analyses were performed on the W4M platform using unsupervised and
232 supervised techniques. PCA was first performed to have an initial visualization of the data sets and to
233 detect outliers. In order to evaluate the ability of this methodology to discriminate the apple juice
234 samples, PLS-DA and OPLS-DA were used. These models were built using a 7-fold cross validation; by
235 this way, each data set was divided into 7 different parts. Each model was next built using 6 parts
236 (train set) and tested using the 7th part (test set); this step was then iterated until all the parts were
237 used as test set. The cross validation procedure permitted to determine the optimal number of latent
238 variables (LV) to build the PLS-DA and OPLS-DA models [Ballabio and Consonni, 2013; Wold et al.,
239 2001]. A new LV was added if the Q²_Y obtained with this LV was greater than 0.01. Indeed, the Q²_Y
240 was calculated from the ratio of PRESS (predictive residual sum of squares) including the new LV over
241 RSS (residual sum of squares) calculated from the model with the previous LV [Wold et al., 2001].
242 The quality of the built models was assessed by the goodness of fit (R²_X), the proportion of the
243 response matrix variance explained by the model (R²_Y) and the predictive performance of the model
244 (Q²_Y). These three metrics have values between 0 and 1. The higher they are, the better the
245 performance of the model. The Q²_Y metric is particularly important here, as it represents the
246 prediction efficiency of the model. An empirical value of 0.4 for Q²_Y has been previously established
247 to judge the quality of the model [Worley and Powers, 2012].

248 An analysis of variance (ANOVA) was also performed to select significant features between the two
249 studied groups (pure vs concentrated juices and organic vs conventional juices); a maximum
250 accepted p-value of 0.01 was chosen in order to select significant features. The features identified by
251 the ANOVA were used to build new PLS-DA and OPLS-DA models to compare models quality with
252 lower features.

253 The biosigner tool [Rinaudo et al., 2016] present on the W4M platform was also used for feature
254 selection. Briefly, this algorithm allows to obtain the smallest number of features which have the
255 most significant contribution in models performance (this module performed PLS-DA, Random Forest
256 (RF) and Support Vector Machine (SVM) models) after performing several iterations. The iterations
257 stop when the number of significant features remains equal to that of the previous iteration. Again,
258 the features selected by biosigner were used to build new PLS-DA and OPLS-DA models.

259 2.4.4. Annotation

260 Significant features were selected based on their results after the chemometric tools used,
261 particularly OPLS-DA and ANOVA results. After having investigated the MS spectra of those
262 discriminant features, the adduct type of the observed m/z was identified which permitted to
263 determine the exact mass of the compound and consequently to suggest molecular formulas. In
264 order to tentatively annotate these features that discriminate the samples, the online databases
265 HMDB [Wishart et al., 2018] and FooDB [FooDB, 2021] (as we are studying apple juice samples) were
266 used. Moreover, the obtained MS/MS spectra were used to confirm the annotation by comparing
267 them to two spectral databases: mzCloud and MassBank. In addition, some commercial compounds
268 known to be present in apple juices were analyzed thanks to available standards as detailed in
269 section 2.1, enabling to build an in-house database.

270 3. Results and discussion

271 3.1. Study 1: Authentication of pure apple juices

272 3.1.1. Principal component analysis

273 PCA is the most common unsupervised multivariate statistical technique [Medina et al., 2019b;
274 Oliveri and Simonetti, 2016] used for exploratory purposes. It was used here to evaluate the

275 reproducibility of three replicates of the same sample. For this study, 58 samples were considered
276 with three replicates per sample (resulting in a total of 174 samples). PCA was applied on two distinct
277 data matrices, containing either all values (i.e., including separate triplicate values) or only a single
278 value (being the mean of the three replicates) for each sample. In both cases, no outlier was
279 observed on the PCA score plots, so that we considered the three replicates to be reproducible.
280 Consequently, only the average of sample triplicates was considered for the following statistical
281 analyses.

282 As shown in Fig. A.1 (Supplementary material), the first three principal components explained about
283 50% of the variance (PC1: 27%; PC2: 14%; PC3: 8%). The replicates of the QC samples were fairly
284 close on the PCA scores plot, showing a good system stability during the analysis. A slight dispersion
285 was noticed in Fig A.1a, with two subsequent groups for the QC samples, in line with the two distinct
286 analytical batches; this observation highlights an analytical drift not completely corrected.
287 Interestingly, a trend seemed to appear for the discrimination between the two groups of samples
288 (single strength vs. both concentrated juices and juices from concentrate) on the PC3 axis, even
289 though no clear separation could be achieved.

290 Conversely, group separation of fruit juices based on the type of fruit were already reported using
291 PCA on UHPLC-HRMS data, with a distinct cluster for apple juices [Vaclavik et al., 2012]. Guo et al.
292 also reported group separation of fresh squeezed apple juices based on varieties by performing a
293 PCA on their concentrations in 23 polyphenols [Guo et al., 2013]. Therefore, it can be assumed that
294 the apple juice production method has fewer differences in the UHPLC-HRMS fingerprint, which
295 explains why no group separation was observed in our PCA.

296 3.1.2. Classification and prediction models: PLS-DA and OPLS-DA

297 PLS-DA and OPLS-DA models were built using the features left after the between batch correction
298 (9,234 features) and from the 58 samples.

299 PLS-DA models were already reported for classification purpose of orange juices, with a satisfactory
300 classification rate of samples regarding geographical origin (after cross-validation, the model showed
301 a 100% classification capacity) [Diaz et al., 2014]; unfortunately, in our study PLS-DA model remained
302 unsatisfactory (data not shown). The obtained model was built using two latent variables and had a
303 goodness of prediction of 0.37 and a goodness of fit of 0.27. OPLS-DA models were previously found
304 interesting for discrimination of Saffron sample origins [Rubert et al., 2016]; results from our data
305 also showed samples discrimination between single strength and both concentrated juices and juices
306 from concentrate, as presented in Fig. 2.

307 The model metrics indicated that the OPLS-DA model was quite satisfactory (as shown in Fig. 2) with
308 a goodness of fit of about 50% and a prediction capacity of about 60%. This OPLS-DA model was built
309 using 9,236 features, so that these metrics might be improved with fewer features. In their study on
310 Saffron, Rubert and co-workers reported that the best OPLS-DA model was obtained using 8 features
311 (of about 5,000 features detected) with 85% of prediction capacity and 97% of goodness of fit
312 [Rubert et al., 2016]. Nevertheless, on Fig. 2 it can be observed that one concentrated juice sample
313 was really close to the direct juice samples group. This observation can lead to incorrect classification
314 or prediction of samples. Consequently, further data processing was tested to improve the modeling.
315 Moreover, the high number of features used for building our OPLS-DA model may have induced
316 overfitting. It was thus important to reduce the number of features used for this model.

317 3.1.3. Feature selection using ANOVA before PLS-DA and OPLS-DA

318 Selection was needed to reduce as much as possible the number of features to be compare with
319 other analytical batches. This is necessary if we want to implement this methodology as a routine
320 analysis method for apple juice authentication assessment. ANOVA and similar t-test have already
321 been used to identify and select significantly different features between groups of samples [Llano et
322 al., 2018; Bat et al., 2018].

323 Performing an ANOVA proved to be greatly helpful: from the about 10,000 features obtained at the
324 end of the filtration steps, the ANOVA identified almost 2,000 significantly different features
325 between the two sample groups. Again, the PLS-DA model gave unsatisfactory results (data not
326 shown). The model was built using two latent variables and had a predictive ability of 0.58. In the
327 score plot, the two sample groups were not differentiated. Conversely, the OPLS-DA model obtained
328 with these identified features was again quite satisfactory (based on the metrics values), with more
329 variance being explained by the first latent variable (30% instead of 13% previously). However, the
330 separation of the two groups was not improved, being even quite worse (Fig. A.2a of Supplementary
331 material).

332 *3.1.4. Feature selection using biosigner before PLS-DA and OPLS-DA*

333 The biosigner module present on the W4M platform was also tested for features selection since it
334 allows selecting the fewest number of features to build discrimination models. Accordingly, only 20
335 features were identified by this tool here. Unfortunately, with these 20 features, the resulting OPLS-
336 DA model showed a worse discrimination of the two groups, even though the direct juice samples
337 stayed close together (Fig. A.2b of Supplementary material). The metrics of the model clearly
338 decreased, confirming the low quality of this model. The PLS-DA model obtained was also
339 unsatisfactory (data not shown). One latent variable was used to build this model and a predictive
340 ability of 0.27 was obtained.

341 Almost all features selected by the biosigner tool were also selected by the ANOVA (90%). The 20
342 features seemed thus to be discriminant, but it can be emphasized that as the number of samples
343 was quite low (58 samples), the features selected were not sufficiently discriminant to improve the
344 OPLS-DA model quality. A larger set of reference samples would be required to establish a robust
345 routine model.

346 *3.2. Study 2: authentication of organic apple juices*

347 *3.2.1. Principal component analysis*

348 In this study, 54 samples were used to build the PCA. As in the previous study, the reproducibility of
349 the triplicates was evaluated using the PCA scores plots. As the replicates showed to be reproducible,
350 the average of the three replicates per samples was used for the next chemometric analysis. PCA
351 scores plots of the filtered data using the mean of the triplicates are shown in Fig. A.3 of
352 Supplementary material. A good system stability was also observed for this study since the replicates
353 of the QC sample were clustered on the PCA scores plot. As in the previous study, two groups of QC
354 samples could be distinguished, showing that the analytical drift was not completely corrected.
355 However, the correction seemed to be better than in the first study because the QC replicates were
356 less dispersed.

357 Group separation between organic and conventional juice samples was not achieved by PCA. Using
358 the first three principal components, about 60% of the variance was explained (PC1: 32%; PC2: 15%
359 and PC3: 8%). Cuevas and coworkers also reported previously that PCA did not allow to separate
360 organic from conventional orange juices using UHPLC-HRMS analysis [Cuevas et al., 2017].

361 3.2.2. Classification and prediction models: PLS-DA and OPLS-DA

362 In order to build models for sample classification, PLS-DA and OPLS-DA analysis were performed (Fig.
363 A.4. of Supplementary material). PLS-DA and OPLS-DA models were both built using a 7-fold cross
364 validation on the 54 samples of the study. OPLS-DA enabled a clear separation between the two
365 groups. This is in line with another study where organic and conventional juices were discriminated
366 using an OPLS-DA model, with a specificity and sensitivity of nearly 90% for both sample classes after
367 cross-validation [Cuevas et al., 2017]. OPLS-DA models satisfactorily discriminated organic and
368 conventional carrot samples analyzed by UHPLC-HRMS, with a classification rate of about 80% using
369 a validation data set [Cubero-Leon et al., 2018].

370 The predictive ability of the OPLS-DA model was good (Q2Y: 0.746). As this model was obtained with
371 a high number of features (near 8400 features), it could be hypothesized that it could be improved
372 with a reduced number of features. Further works should focus on the external assessment of the
373 models performance, which was not allowed by the number of samples in this study.

374 3.2.3. Feature selection using ANOVA before PLS-DA and OPLS-DA

375 In this study, the ANOVA found 1,422 significantly different features between the organic and
376 conventional juice samples (from almost 10,000 features detected). PLS-DA and OPLS-DA models
377 obtained from these features are presented in Fig. 3.

378 The new models obtained with a reduced number of features showed similar metrics compared to
379 the previously obtained models (Fig. A.4 of Supplementary material). The discrimination between the
380 two sample groups was still not observed using PLS-DA model. On the contrary, OPLS-DA model
381 showed a clear separation. The predictive ability of this model indicated that it had a good
382 performance (Q2Y: 0.785) and it was slightly better than the OPLS-DA obtained with all the features.
383 The percentage of variance explained by the first LV had increased to 24% with the feature selection.

384 3.2.4. Feature selection using biosigner before PLS-DA and OPLS-DA

385 Again, the biosigner tool was used to find the smallest number of most significant features. This
386 module found 48 features. To evaluate whether these selected features were the most significant,
387 PLS-DA and OPLS-DA models were built using a 7-fold cross-validation. In contrast to the results
388 obtained from the feature selection using ANOVA, the obtained PLS-DA and OPLS-DA models were
389 not improved. The metrics showed that models were quite worse than with all the features (Fig. A.5
390 of Supplementary material).

391 Only 14 of the 48 features selected by the biosigner tool were also selected by the ANOVA. Most of
392 the features selected by this module were chosen based on their performance using SVM models.
393 SVM models can perform very well but they require lots of data (ideally thousands of samples). In
394 this study, there were only 54 samples, so the features selected using SVM models were not
395 discriminant enough to increase the PLS-DA and OPLS-DA model metrics.

396 3.3. Tentative identification of discriminant features in both studies

397 The number of features remained high, even after selection with ANOVA (about 1,500 features). To
398 reduce this number while keeping the most discriminant features, it was decided to filter them
399 according to their VIP (Variable Importance on Projection) value calculated during the construction of
400 the PLS-DA and OPLS-DA models. The VIP value of a feature indicates its importance on the model
401 building: the higher VIP value, the more discriminating the feature is [Wold et al., 2001]. In a previous
402 study, filtration based on the VIP value successfully selected 8 features out of about 5,000 features

403 detected [Rubert et al., 2016]. Other authors reported the use of VIP values to select discriminant
404 features by retaining 25 features (out of about 5,000 features detected) that were further tentatively
405 identified [Cavanna et al., 2020]. Cubero-Leon and colleagues also used a similar criterion (VIP
406 greater than 1) applied to remove features contributing to other variability (year of harvest); they
407 were able to build successful OPLS-DA models to discriminate between organic and conventional
408 carrot samples [Cubero-Leon et al., 2018]. In the literature, different VIP values between 1 and 2
409 have been used as a filtration criterion. In this work, a filter was applied to keep features having a VIP
410 value greater than 1, as proposed in different articles [Gorrochategui et al., 2016; Pezzati et al.,
411 2020]. After this filtration, about 150 features remained for both authentication applications
412 considered in our work.

413 To reduce the number of features to be identified, both results from ANOVA and from OPLS-DA were
414 used. We focused on the features with the highest VIP and the lowest p-value, and attempted to
415 identify them using online databases (HMDB and FooDB). Based on this strategy, less than 15
416 features were selected in each study for further identification, as indicated in Table 1 and Table A.4
417 (Supplementary material). Few of these features were also selected by the biosigner tool. Examples
418 of chromatograms for one feature identified as discriminating for each study are shown in Fig. 4 and
419 Fig. A.6 (Supplementary material).

420

421 Some features had the same retention time, being either coeluted chromatographic peaks or
422 fragments and/or adducts of a unique compound. The observation of the MS spectra permitted to
423 identify features which correspond to a same molecule (Table 1); in particular, the presence of
424 certain adducts such as $[M+NH_4]^+$ and $[M+K]^+$ allowed to attribute the adduct type of the observed
425 m/z . This was mostly the case for the features identified in the second study. As presented in Table 1
426 and Table A.4, a majority of features were still unknown as no matches were obtained on the online
427 databases used. For some features, several compounds matched the exact mass defined. It is
428 interesting to observe that results from HMDB and FooDB were really close, being a good starting
429 point to annotate features but still insufficient: the compounds of interest may not be present on
430 these databases.

431 In order to improve the annotation of these discriminant features, MS/MS acquisitions were
432 performed on a new set of samples (30 samples containing concentrated juice samples, conventional
433 direct juice samples and organic direct juice samples). Only few results were obtained from the
434 databases search, either with the monoisotopic mass or with the proposed molecular formula.
435 Interestingly, two amino acids (methionine and isoleucine or norleucine) could be proposed as
436 discriminant markers between organic and conventional apple juice (Table 1); this result seems
437 realistic since those compounds were already reported in apple juice samples [Ma et al., 2018]. In
438 particular a biosynthesis pathway leading to isoleucine formation in ripening apple fruit has been
439 recently reported (Sugimoto et al., 2021). The same methodology was applied for the features
440 identified in the authentication of pure apple juices with N-(1-deoxy-1-fructosyl)phenylalanine
441 proposed as a marker (Table A.4. of Supplementary material). Xu et al. also reported an amino-acid
442 (L-glutamine) to discriminate from concentrate and not from concentrate orange juices [Xu et al.,
443 2020]. Further investigation is needed to improve the annotation of the identified discriminant
444 features, probably by using other online databases or building in-house database.

445 During the MS/MS experiments, a full scan analysis was also acquired. It was thus possible to use
446 these independent acquisitions to evaluate the potential of the discriminant features to serve as
447 marker compounds. It is noteworthy that the previously selected features were successfully

448 observed on this fourth analytical batch (only one feature was missing because it has a low intensity);
449 nevertheless, only a few of them still showed a trend in the discrimination of sample groups. In
450 particular, for the authentication of direct apple juices, 5 features still showed a difference in
451 intensity between the two sample groups; these features might be used as markers compounds for
452 the concentrated juice characteristic. On the other hand, for the authentication of organic apple
453 juices, no trend was observed for the discrimination of the two samples groups by observing the
454 intensity of the features between the two sample groups.

455 It can be emphasized that these independent acquisitions permitted to highlight some marker
456 compounds as they were characteristic of the process type used (juice concentration). For the
457 organic juice characteristic, these new samples came from a different harvest year, which may
458 explain that the discriminant features found previously may fail to discriminate these new
459 acquisitions. Cubero-Leon and co-workers reported that the harvest year was one of the most
460 important variabilities in their studied samples [Cubero-Leon et al., 2018]. On the contrary, Diaz et al.
461 identified a biomarker for orange origin which seems to be independent from the harvest year [Diaz
462 et al., 2014]. Further investigation is thus needed to find reliable features for the authentication of
463 organic apple juice samples and to confirm the use of the 5 features for the authentication of direct
464 apple juice samples.

465 Based on the analysis of standards, two detected features might be assigned to phloridzin (p-value:
466 5.83 E-03) and alpha-terpineol (p-value: 1.08 E-04) according to their m/z and retention time;
467 unfortunately, these two features were outside the list of Table 1. It is not surprising that phloridzin
468 was not a discriminant compound as it is a naturally present molecule in apples, with varying
469 concentrations depending on different factors such as variety or processing technology used [Spinelli
470 et al., 2016]. The remaining standards were not detected in our samples, possibly because they were
471 not concentrated enough to be observed, while the other one (hydroxymethylfurfural) routinely
472 analyzed by LC-UV may not be present in the juice samples analyzed.

473 4. Conclusions

474 This work presents a methodology combining untargeted LC-HRMS analysis and chemometric tools
475 to authenticate apple juice samples. The OPLS-DA models showed good performance in sample
476 classification, especially for the discrimination between organic and conventional sample juices
477 (nearly 80% of predictive ability). To confirm their classification and prediction performance, further
478 validation of these models using an external data set is required.

479 Coupling the results of ANOVA and OPLS-DA seems to be an interesting methodology to determine
480 the discriminant features as it permitted to reduce the number of detected features (from almost
481 10,000 features detected to about 150 features) while keeping significant and discriminant features.
482 According to the chemometric tools used (OPLS-DA and ANOVA) about 20 features have been
483 identified as significantly discriminant and tentatively identified for the first time. Some compounds
484 were tentatively annotated as amino-acids and derivatives, and a few markers were confirmed by
485 MS/MS experiments. Interestingly, application of our analytical method to a new set of samples
486 showed that some features retained a tendency to discriminate between the two groups of samples,
487 mainly for authentication of direct apple juices.

488 The main additional research concerns the annotation workflow, which is the most time-consuming
489 part of this methodology. By building an in-house database, the identification of marker compounds
490 can be faster as the obtained mass spectra will be better compared than by using an online database,
491 the same instrument being used. By identifying the compounds responsible for the discrimination,

492 they could be analyzed in a routine analysis for apple juice authentication. Further investigation is
493 needed to correctly identify the compounds by the analysis of standards.

494 The proposed analytical methodology enabled, for the very first time, the authentication of apple
495 juice samples in two distinct scenarios using a single analysis (organic vs. conventional samples and
496 single strength juice vs. both concentrated juice and juice from concentrate samples). Other
497 chemometric models could be developed to implement juice discrimination based on variety and/or
498 geographical origin, in addition to the scenarios presented here.

499

500 **Acknowledgement**

501 The authors warmly thank Dr. Peter Rinke from SGF for kindly providing them with several samples of
502 apple juices used in this study. They are thankful for the financial support provided by the
503 Association Nationale Recherche et Technologie (ANRT) through the CIFRE program (CIFRE
504 n°2018/0937).

505

506 **Conflict of interest**

507 The authors declare that they have no commercial or financial relationships that could have influence
508 the research conducted in this paper.

509

510 **Appendix A. Supplementary Data**

511

512 References

- 513 AIJN Code of Practice (2020), AIJN European Fruit Juice Association.
- 514 Ballabio, D., Consonni, V., 2013. Classification tools in chemistry. Part 1: linear models. PLS-DA.
515 Analytical Methods 5, 3790. <https://doi.org/10.1039/c3ay40582f>
- 516 Bat, K.B., Vodopivec, B.M., Eler, K., Ogrinc, N., Mulič, I., Masuero, D., Vrhovšek, U., (2018). Primary
517 and secondary metabolites as a tool for differentiation of apple juice according to cultivar and
518 geographical origin. *LWT – Food Science and Technology*, 90, 238–245.
519 <https://doi.org/10.1016/j.lwt.2017.12.026>
- 520 Brooks, S., Elliott, C.T., Spence, M., Walsh, C., Dean, M., (2017). Four years post-horsegate: an update
521 of measures and actions put in place following the horsemeat incident of 2013. *npj Science of*
522 *Food*, 1. <https://doi.org/10.1038/s41538-017-0007-z>
- 523 Cavanna, D., Righetti, L., Elliott, C., & Suman, M. (2018). The scientific challenges in moving from
524 targeted to non-targeted mass spectrometric methods for food fraud analysis: A proposed
525 validation workflow to bring about a harmonized approach. *Trends in Food Science and*
526 *Technology*, 80, 223-241. <https://doi.org/10.1016/j.tifs.2018.08.007>
- 527 Cavanna, D., Loffi, C., Dall’Asta, C., Suman, M., (2020). A non-targeted high-resolution mass
528 spectrometry approach for the assessment of the geographical origin of durum wheat. *Food*
529 *Chemistry*, 317, 126366. <https://doi.org/10.1016/j.foodchem.2020.126366>
- 530 Chaleckis, R., Meister, I., Zhang, P., Wheelock, C.E., (2019). Challenges, progress and promises of
531 metabolite annotation for LC–MS-based metabolomics. *Current Opinion in Biotechnology*, 55, 44-
532 50. <https://doi.org/10.1016/j.copbio.2018.07.010>
- 533 Chambers, M.C., MacLean, B., Burke, R., Amodè, D., Ruderman, D. L., Neumann, S., ... Mallick, P.,
534 (2012). A cross-platform toolkit for mass spectrometry and proteomics. *Nature Biotechnology*,
535 30(10), 918–920. <https://doi.org/10.1038/nbt.2377>
- 536 Codex Alimentarius: Codex General Standard for Fruit Juices and Nectars (2005) CODEX STAN 247-
537 2005
- 538 Cuberon-Leon, E., Peñalver, R., Maquet, A., (2014). Review on metabolomics for food authentication.
539 *Food Research International*, 60, 95-107. <https://doi.org/10.1016/j.foodres.2013.11.041>
- 540 Cubero-Leon, E., De Rudder, O., Maquet, A., (2018). Metabolomics for organic food authentication:
541 Results from a long-term field study in carrots. *Food Chemistry*, 239, 760–770.
542 <https://doi.org/10.1016/j.foodchem.2017.06.161>
- 543 Cuevas, F.J., Pereira-Caro, G., Moreno-Rojas, J.M., Muñoz-Redondo, J.M., Ruiz-Moreno, M.J., (2017).
544 Assessment of premium organic orange juices authenticity using HPLC-HR-MS and HS-SPME-GC-
545 MS combining data fusion and chemometrics. *Food Control*, 82, 203–211.
546 <https://doi.org/10.1016/j.foodcont.2017.06.031>
- 547 Cuevas, F.J., Pereira-Caro, G., Muñoz-Redondo, J.M., Ruiz-Moreno, M.J., Montenegro, J.C., Moreno-
548 Rojas, J.M., (2019). A holistic approach to authenticate organic sweet oranges (*Citrus Sinensis* L. cv
549 Osbeck) using different techniques and data fusion. *Food Control*, 104, 63–73.
550 <https://doi.org/10.1016/j.foodcont.2019.04.012>

- 551 Danezis, G.P., Tsagkaris, A. S., Camin, F., Brusica, V., Georgiou, C.A., (2016). Food authentication:
552 Techniques, trends & emerging approaches. *Trends in Analytical Chemistry*, 85, 123–132.
553 <https://doi.org/10.1016/j.trac.2016.02.026>
- 554 Dasenaki, M.E. & Thomaidis, N.S. (2019). Quality and authenticity control of fruit juices - A review.
555 *Molecules*, 24, 1014. <https://doi.org/10.3390/molecules24061014>
- 556 Delaporte, G., Cladiere, M., Jouan-Rimbaud Bouveresse, D., Camel, V., (2019). Untargeted food
557 contaminant detection using UHPLC-HRMS combined with multivariate analysis: Feasibility study
558 on tea. *Food Chemistry*, 277, 54–62. <https://doi.org/10.1016/j.foodchem.2018.10.089>
- 559 Diaz, R., Pozo, O.J., Sancho, J.V., Hernandez, F., (2014). Metabolomic approaches for orange origin
560 discrimination by ultra-high performance liquid chromatography coupled to quadrupole time-of-
561 flight mass spectrometry. *Food Chemistry*, 157, 84–93.
562 <https://doi.org/10.1016/j.foodchem.2014.02.009>
- 563 Dieterle, F., Ross, A., Schlotterbeck, G., Senn, H., (2006). Probabilistic quotient normalization as
564 robust method to account for dilution of complex biological mixtures. Application in 1H NMR
565 metabonomics. *Analytical Chemistry*, 78(13), 4281–4290. <https://doi.org/10.1021/ac051632c>
- 566 Directive 2012/12/EC, (2012). Council Directive Relating to Fruit Juices and Certain Similar Products
567 Intended for Human Consumption of 19 April 2012
- 568 Dubin, E., Dumas, A.-S., Rebours, A., Jamin, E., Ginot, J., Lees, M., Rutledge, D.N., (2017). Detection of
569 Blackcurrant Adulteration by Aronia Berry Using High Resolution Mass Spectrometry, Variable
570 Selection and Combined PLS Regression Models. *Food Analytical Methods*, 10, 683–693.
571 <https://doi.org/10.1007/s12161-016-0638-8>
- 572 Esteki, M., Simal-Gandarab, J., Shahsavaria, Z., Zandbaafa, S., Dashtakia, E., Heydenc, Y.V., (2018). A
573 review on the application of chromatographic methods, coupled to chemometrics, for food
574 authentication. *Food Control*, 93, 195–182. <https://doi.org/10.1016/j.foodcont.2018.06.015>
- 575 FooDB, (2021). Food Database, <https://foodb.ca/> accessed on November 24th 2021
- 576 Giacomoni, F., Le Corguillé, G., Monsoor, M., Landi, M., Pericard, P., Pétéra, M., ... Caron, C., (2015).
577 Workflow4Metabolomics: A collaborative research infrastructure for computational
578 metabolomics. *Bioinformatics*, 31(9), 1493–1495. <https://doi.org/10.1093/bioinformatics/btu813>
- 579 Gorrochategui, E., Jaumot, J., Lacorte, S., Tauler, R, (2016). Data analysis strategies for targeted and
580 untargeted LC-MS metabolomic studies: overview and workflow. *Trends in Analytical Chemistry*,
581 82, 425–442. <https://doi.org/10.1016/j.trac.2016.07.004>
- 582 Guo, J., Yue, T., Yuan, Y., Wang, Y., (2013). Chemometric classification of apple juices according to
583 variety and geographical origin based on polyphenolic profiles. *Journal of Agricultural and Food
584 Chemistry*, 61, 6949–6963. <https://doi.org/10.1021/jf4011774>
- 585 IFU, (2021). International Fruit and Vegetable Juice Association, <https://ifu-fruitjuice.com/> accessed
586 on May 28th 2021
- 587 Jandric, Z., Roberts, D., Rathor, M.N., Abraham, A., Islam, M., Cannavan, A., (2014). Assessment of
588 fruit juice authenticity using UPLC-QToF MS: A metabolomics approach. *Food Chemistry*, 148, 7–
589 17. <https://doi.org/10.1016/j.foodcont.2018.06.015>

- 590 Jandric, Z., Islam, M., Singh, D.K., Cannavan, A., (2017). Authentication of Indian citrus fruit/fruit
591 juices by untargeted and targeted metabolomics. *Food Control*, 71, 181–188.
592 <https://doi.org/10.1016/j.foodcont.2015.10.044>
- 593 Knolhoff, A. M., Croley, T. R., (2016). Non-targeted screening approaches for contaminants and
594 adulterants in food using liquid chromatography hyphenated to high resolution mass
595 spectrometry. *Journal of Chromatography A*, 1428, 86–96.
596 <https://doi.org/10.1016/j.chroma.2015.08.059>
- 597 Llano, S.M., Muñoz-Jiménez, A.M., Jiménez-Cartagena, C., Londoño-Londoño, J., Medina, S., (2018).
598 Untargeted metabolomics reveals specific withanolides and fatty acyl glycoside as tentative
599 metabolites to differentiate organic and conventional *Physalis peruviana* fruits. *Food Chemistry*,
600 244, 120–127. <https://doi.org/10.1016/j.foodchem.2017.10.026>
- 601 Ma, S., Neilson, A.P., Lahne, J., Peck, G.M., O’Keefe, S.F., Stewart, A.C., (2018). Free amino acid
602 composition of apple juices with potential for cider making as determined by UPLC-PDA. *Journal*
603 *of the Institute of Brewing*, 124, 467–476. <https://doi.org/10.1002/jib.519>
- 604 Medina, S., Pereira, J.A., Silva P., Perestrelo, R., Câmara, J.S., (2019a). Food fingerprints – a valuable
605 tool to monitor food authenticity and safety. *Food Chemistry*, 278, 144-162.
606 <https://doi.org/10.1016/j.foodchem.2018.11.046>
- 607 Medina, S., Perestrelo, R., Silva, P., Pereira, J., Câmara, J.S., (2019b). Current trends and recent
608 advances on food authenticity technologies and chemometric approaches. *Trends in Food Science*
609 *& Technology*, 85, 163-176. <https://doi.org/10.1016/j.tifs.2019.01.017>
- 610 Mihailova, A., Kelly, S.D., Chevallier, O.P., Elliott, C.T. (2021). High-resolution mass spectrometry-
611 based metabolomics for the discrimination between organic and conventional crops: A review.
612 *Trends in Food Science & Technology*, 110, 142-154. <https://doi.org/10.1016/j.tifs.2021.01.071>
- 613 Moore, J., Spink, J., Lipp, M., (2012). Development and Application of a Database of Food Ingredient
614 Fraud and Economically Motivated Adulteration from 1980 to 2010. *Journal of Food Science*, 77,
615 R118-R126. <https://doi.org/10.1111/j.1750-3841.2012.02657.x>.
- 616 Oliveri, P., & Simonetti, R., (2016). Chemometrics for Food Authenticity Applications. In G. Downey
617 (Eds.), *Advances in Food Authenticity Testing* (pp.701-728). Woodhead Publishing is an imprint of
618 Elsevier
- 619 Pezzatti, J., Boccard, J., Codesido, S., Gagnebin, Y., Joshi, A., Picard, D., Gonzalez-Ruiz, V., Rudaz, S.,
620 (2020). Implementation of liquid chromatography - high resolution mass spectrometry methods
621 for untargeted metabolomic analyses of biological samples: a tutorial. *Analytical Chimica Acta*,
622 1105, 28e44. <https://doi.org/10.1016/j.aca.2019.12.062>.
- 623 Rinke, P., (2016). Tradition Meets High Tech for Authenticity Testing of Fruit Juices. In G. Downey
624 (Ed.), *Advances in Food Authenticity Testing* (pp.625-665). Woodhead Publishing is an imprint of
625 Elsevier
- 626 Rinke, P., & Jamin, E., (2018). Fruit juices. In Morin, J.-F., Lees, M. (Eds.), *FoodIntegrity Handbook: A*
627 *guide to food authenticity issues and analytical solutions*, 1st ed. Eurofins Analytics France.
628 <https://doi.org/10.32741/fihb>

- 629 Rinaudo, P., Boudah, S., Junot, C., Thévenot, E.A., (2016). biosigner: A New Method for the Discovery
630 of Significant Molecular Signatures from Omics Data. *Frontiers in Molecular Biosciences* 3.
631 <https://doi.org/10.3389/fmolb.2016.00026>
- 632 Rubert, J., Lacina, O., Zachariasova, M., Hajslova, J., (2016). Saffron authentication based on liquid
633 chromatography high resolution tandem mass spectrometry and multivariate data analysis. *Food*
634 *Chemistry*, 204, 201–209. <https://doi.org/10.1016/j.foodchem.2016.01.003>
- 635 Smith, C.A., Want, E.J., O’Maille, G., Abagyan, R., Siuzdak, G., (2006). XCMS: Processing Mass
636 Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and
637 Identification. *Analytical Chemistry*, 78, 779–787. <https://doi.org/10.1021/ac051437y>
- 638 Sobolev, A.P., Thomas, F., Donarski, J., Ingallina, C., Circi, S., Marincola, F.C., Capitani, D., Mannina, L.,
639 (2019). Use of NMR applications to tackle future food fraud issues. *Trends in Food Science &*
640 *Technology*, 91, 347–353. <https://doi.org/10.1016/j.tifs.2019.07.035>
- 641 Spinelli, F.R., Dutra, S.V., Carnieli, G., Leonardelli, S., Drehmer, A.P., Vanderlinde, R., (2016).
642 Detection of addition of apple juice in purple grape juice. *Food Control*, 69, 1–4.
643 <https://doi.org/10.1016/j.foodcont.2016.04.005>
- 644 Sugimoto, N., Engelgau, P., Jones, A.D., Song, J., Beaudry, R., (2021). Citramalate synthase yields a
645 biosynthetic pathway for isoleucine and straight- and branched-chain ester formation on ripening
646 apple fruit. *PNAS*, 118(3), e2009988118. <https://doi.org/10.1073/pnas.2009988118>
- 647 Tautenhahn, R., Bottcher, C., Neumann, S., (2008). Highly sensitive feature detection for high
648 resolution LC/MS. *BMC Bioinformatics*, 9. <https://doi.org/10.1186/1471-2105-9-504>
- 649 Vaclavik, L., Schreiber, A., Lacina, O., Cajka, T., (2012). Liquid chromatography-mass spectrometry-
650 based metabolomics for authenticity assessment of fruit juices. *Metabolomics*, 8, 793–803.
651 <https://doi.org/10.1007/s11306-011-0371-7>
- 652 Wishart, D.S., Feunang, Y.D., Marcu, A., Guo, A.C., Liang, K., Vázquez-Fresno, R., ... Scalbert, A.,
653 (2018). HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Research*, 46, D608–
654 D617. <https://doi.org/10.1093/nar/gkx1089>
- 655 Wold, S., Sjöström, M., Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics.
656 *Chemometrics and intelligent laboratory systems*, 58, 109–130.
- 657 Wolter, C., Gessler, A., Winterhalter, P., (2008). Aspects when evaluating apple-juice aroma. *Fruit*
658 *processing*, 64–80
- 659 Worley, B., Powers, R., 2012. Multivariate Analysis in Metabolomics. *Current Metabolomics* 1, 92–
660 107. <https://doi.org/10.2174/2213235X11301010092>
- 661 Xu, L., Xu, Z., Kelly, S., Liao, X., (2020). Integrating untargeted metabolomics and targeted analysis for
662 not from concentrate and from concentrate orange juices discrimination and authentication. *Food*
663 *Chemistry*, 329, 127130. <https://doi.org/10.1016/j.foodchem.2020.127130>
- 664
- 665
- 666
- 667

668 **Figure Captions**

669 **Fig. 1.** Workflow of the data treatment using W4M* (RSD: relative standard deviation) * *text in italic*
670 *refers to W4M functions.*

671 **Fig. 2.** Scores plot for OPLS-DA obtained with cross-validation (blue circles, concentrated juices and
672 juices from concentrate; red crosses, direct juices). The black ellipse represents 95% of the variability,
673 the blue and red ellipses represent 95% of the multivariate distributions of the sample groups.

674 **Fig. 3.** (a) Scores plot of PLS-DA and (b) scores plot of OPLS-DA obtained after features selection using
675 ANOVA (blue circles: organic juice samples; red crosses, conventional juice samples). The black
676 ellipse represents 95% of the variability, the blue and red ellipses represent 95% of the multivariate
677 distributions for each sample groups.

678 **Fig. 4.** Chromatogram of feature 13 for authentication of organic apple juices (black, organic juice
679 samples; red, conventional juice samples)

Table 1: Discriminant features for authentication of organic apple juices (compounds confirmed based on MS/MS data are indicated in bold characters).

# Compound	# Feature	Detected m/z	Adduct type	RT (min)	p-value	VIP	Characteristic	Monoisotopic mass	Proposed molecular formula	Proposed compounds (FoodB)	Proposed compounds (HMDB)
1	1*	132.1019	[M+H] ⁺	3.32	1.0E-05	10.7	Conv > Org	131.0947	C ₆ H ₁₃ NO ₂	Leucine, Isoleucine , 6-Deoxyfagomine, Alloseucine, Norleucine , Aminocaproic acid, Alanine betaine	Leucine, Isoleucine , 6-Deoxyfagomine, Alloseucine, Norleucine , Aminocaproic acid, Methylvaline, N-(2-Hydroxyethyl)-morpholine
	3	133.1052	M+1	3.32	1.0E-05	2.8					
2	2*	133.0317	[M+H] ⁺	1.95	4.4E-08	4.0	Conv > Org	132.0245	C ₃ H ₂ F ₂ N ₄	n.a.	n.a.
									C ₈ H ₃ FN	n.a.	n.a.
									C ₈ H ₄ O ₂	2,4,6-Octatriynoic acid	n.a.
3	4*	150.0583	[M+H] ⁺	1.95	3.8E-08	10.9	Conv > Org	149.0510	C ₅ H ₈ O ₂ S	n.a.	3-Methyl sulfolene
									C ₃ H ₅ F ₂ N ₅	n.a.	n.a.
									C ₈ H ₆ FN ₂	n.a.	n.a.
4	5*	151.0616	M+1	1.95	3.2E-08	2.5	Conv > Org	244.0693	C ₅ H ₁₁ NO ₂ S	Methionine	Methionine , Penicillamine
	6*	152.0541	M+2	1.95	3.4E-08	2.2					
5	7	245.0767	[M+H] ⁺	2.47	5.6E-05	3.6	Conv > Org	244.0693	C ₉ H ₁₂ N ₂ O ₆	<i>Pseudouridine, Uridine**</i>	<i>Pseudouridine, Uridine**</i>
									8	267.0585	[M+Na] ⁺
6	9	271.1149	[M+H] ⁺	6.93	4.5E-06	2.2	Conv > Org	270.1077	C ₉ H ₁₄ N ₆ O ₄	n.a.	n.a.
									C ₁₀ H ₁₀ N ₁₀	n.a.	n.a.
									C ₁₁ H ₁₆ N ₃ O ₅	n.a.	n.a.
7	10	331.1724	[M+H] ⁺	12.98	1.3E-05	3.8	Conv > Org	330.1652	C ₁₂ H ₂₂ N ₆ O ₅	n.a.	n.a.
									C ₁₃ H ₁₈ N ₁₀ O	n.a.	n.a.
									C ₁₄ H ₂₄ N ₃ O ₆	n.a.	n.a.
									C ₁₁ H ₂₆ N ₂ O ₉	n.a.	n.a.
8	11	433.2040	[M+H] ⁺	13.86	1.8E-05	3.4	Conv > Org	432.1968	C ₁₆ H ₂₈ N ₆ O ₈	n.a.	n.a.
									C ₁₇ H ₂₄ N ₁₀ O ₄	n.a.	n.a.
									C ₁₈ H ₃₀ N ₃ O ₉	n.a.	n.a.

									C ₂₄ H ₃₂ O ₅ S	S-Furanopetasitin	S-Furanopetasitin
8	12	495.1744	[M+H] ⁺	13.86	2.5E-06	2.2	Conv > Org	494.1671	C ₂₄ H ₂₄ N ₅ O ₇	n.a.	n.a.
									C ₂₃ H ₂₈ NO ₁₁	n.a.	n.a.
									C ₂₂ H ₂₂ N ₈ O ₆	n.a.	n.a.
									C ₂₁ H ₂₆ N ₄ O ₁₀	n.a.	n.a.
9	13	591.1677	[M+H] ⁺	9.58	9.8E-06	3.4	Org > Conv	590.1606	C ₂₄ H ₂₆ N ₆ O ₁₂	n.a.	n.a.
									C ₂₅ H ₂₂ N ₁₀ O ₈	n.a.	n.a.
									C ₂₈ H ₃₀ O ₁₄	Maysin 3'-methyl ether	n.a.

n.a.: not applicable

** These features were detected using biosigner*

*** invalid based on MS/MS data*

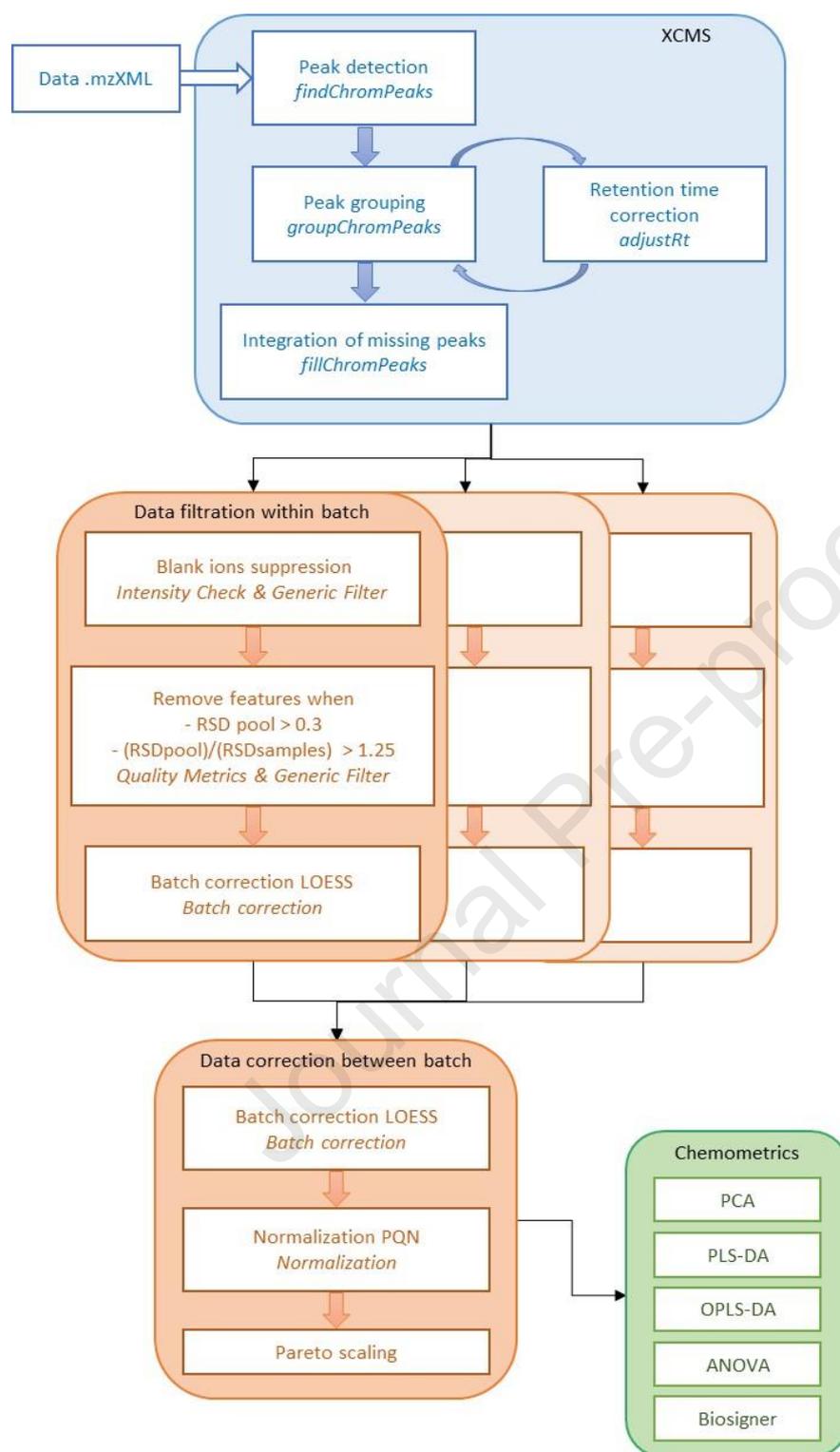


Fig. 1. Workflow of the data treatment using W4M* (RSD: relative standard deviation) * text in italic refers to W4M functions.

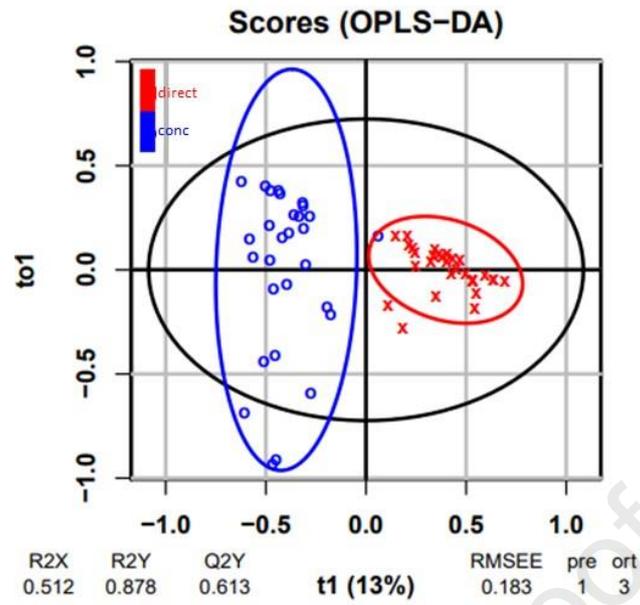


Fig. 2. Scores plot for OPLS-DA obtained with cross-validation (blue circles, both concentrated juices and juices from concentrate; red crosses, direct juices). The black ellipse represents 95% of the variability, the blue and red ellipse are the Mahalanobis ellipse of the sample groups.

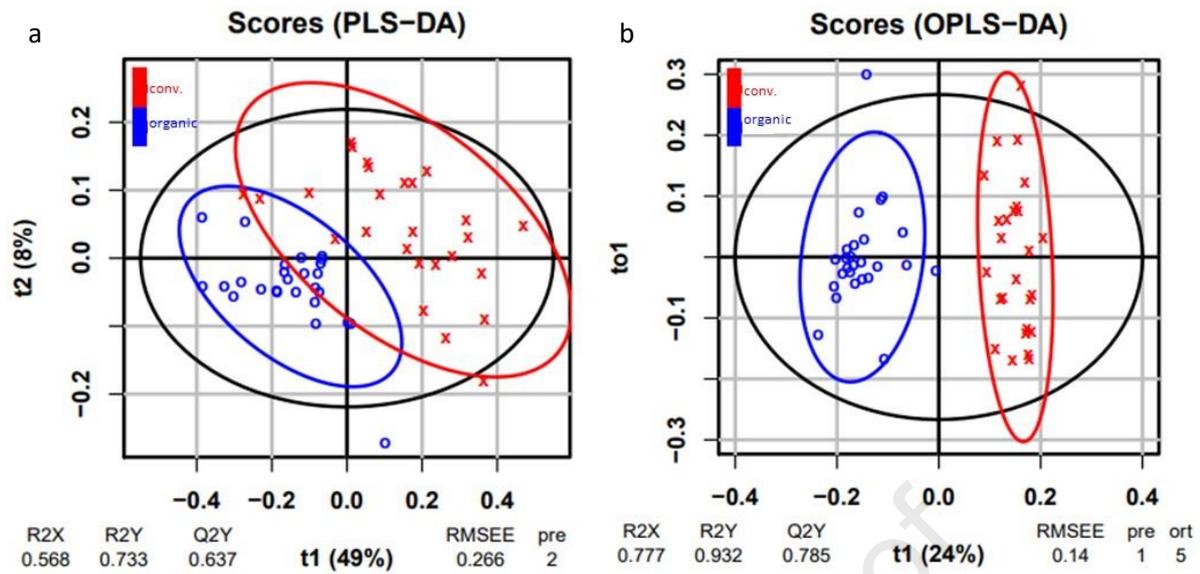


Fig. 3. (a) Scores plot of PLS-DA and (b) scores plot of OPLS-DA obtained after features selection using ANOVA (blue circles: organic juice samples; red crosses, conventional juice samples). The black ellipse represents 95% of the variability, the blue and red ellipses represent 95% of the multivariate distributions for each sample groups.

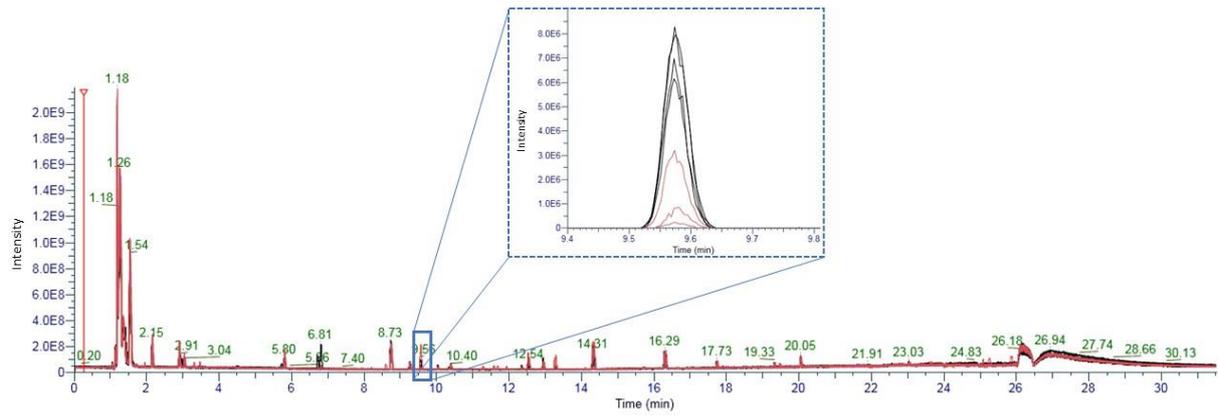


Fig. 4. Chromatogram of feature 13 for authentication of organic apple juices (black, organic juice samples; red, conventional juice samples)

Highlights

- Development of an UHPLC-HRMS metabolomics approach with great potential in juice authentication
- Discrimination between sample groups in two distinct authentication applications
- Relevant markers selected by OPLS-DA and ANOVA were tentatively identified
- Main discriminant compounds were identified as amino-acids and derivatives

Journal Pre-proof

Conflict of Interest and Authorship Conformation Form

Please check the following as appropriate:

- All authors have participated in (a) conception and design, or analysis and interpretation of the data; (b) drafting the article or revising it critically for important intellectual content; and (c) approval of the final version.
- This manuscript has not been submitted to, nor is under review at, another journal or other publishing venue.
- The authors have no affiliation with any organization with a direct or indirect financial interest in the subject matter discussed in the manuscript
- The following authors have affiliations with organizations with direct or indirect financial interest in the subject matter discussed in the manuscript:

Author's name

Affiliation

No affiliation
