



HAL
open science

Untargeted metabolomics-based approach using UHPLC-HRMS to authenticate carrots (*Daucus carota* L.) based on geographical origin and production mode

Katy Dinis, Lucie Tsamba, Eric Jamin, Valérie Camel

► To cite this version:

Katy Dinis, Lucie Tsamba, Eric Jamin, Valérie Camel. Untargeted metabolomics-based approach using UHPLC-HRMS to authenticate carrots (*Daucus carota* L.) based on geographical origin and production mode. *Food Chemistry*, 2023, 423, pp.136273. 10.1016/j.foodchem.2023.136273. hal-04096837

HAL Id: hal-04096837

<https://hal.science/hal-04096837v1>

Submitted on 16 Jul 2024

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.



Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License

1 **Untargeted metabolomics-based approach using UHPLC-HRMS to authenticate carrots (*Daucus***
2 ***carota* L.) based on geographical origin and production mode**

3 Katy Dinis ^{a,b}, Lucie Tsamba ^a, Eric Jamin ^a, Valérie Camel ^{b*}

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

6 ^b Université Paris-Saclay, INRAE, AgroParisTech, UMR SayFood, 91120 Palaiseau, France

7 * Corresponding author: Université Paris-Saclay, INRAE, AgroParisTech, UMR SayFood, 22 place de
8 l'Agronomie, F-91120 PALAISEAU, France. E-mail address: valerie.camel@agroparistech.fr

9
10 **Abstract**

11 Carrots produced in different agricultural regions with organic or conventional mode were analyzed
12 by untargeted UHPLC-HRMS using reversed-phase and HILIC modes. Data were first treated
13 separately, and further combined to possibly improve results. An in-house data processing workflow
14 was applied to identify relevant features after peak detection. Based on these features,
15 discrimination models were built using chemometrics. A tentative annotation of chemical markers
16 was performed using online databases and UHPLC-HRMS/MS analyses. An independent set of
17 samples was analyzed to assess the discrimination potential of these markers. Carrots produced in
18 the New Aquitaine region could be successfully discriminated from carrots originating from the
19 Normandy region by an OLPS-DA model. Arginine and 6-methoxymellein could be identified as
20 potential markers with the C18-silica column. Additional markers (N-acetylputrescine, L-carnitine)
21 could be identified thanks to the polar column. Discrimination based on production mode was more
22 challenging: some trend was observed but model metrics remained unsatisfactory.

23 **Keywords**

24 Food authenticity, High resolution mass spectrometry, Liquid chromatography, Metabolomics, PCA,
25 PLS-DA, OPLS-DA

26

27 **1. Introduction**

28 Carrot (*Daucus carota* L., a member of the *Apiaceae* family) is a root vegetable produced and
29 consumed worldwide. Its production underwent a major increase in the 1980s (30% increase
30 between 1980 and 1990), and is still increasing [Arscott and Tanumihardjo, 2010]. Worldwide 13.7
31 million tons were produced in 1990 [Arscott and Tanumihardjo, 2010] and 27 million tons in 2008
32 [Stolarczyk and Janick, 2011]. Carrot offers interesting nutritional benefits, mainly due to the
33 presence of carotenoids especially provitamin A (i.e. beta-carotene converted to vitamin A in the
34 body) and numerous phenolic compounds [Arscott and Tanumihardjo, 2010; Stolarczyk and Janick,
35 2011; Ahmad et al., 2019]. Several positive impacts on consumers health are assumed for this
36 vegetable, such as anticarcinogenic and antioxidant effects [Akhtar et al., 2017]. Another striking
37 feature of carrots is their possibility to be consumed under various forms: fresh, processed (i.e.
38 juice), dried, boiled or fried. This vegetable has interesting technological properties for the food
39 industry and the cosmetic industry as well [Stolarczyk and Janick, 2011], and a growing demand is
40 foreseen in the near future since a recent study has suggested carrot could be a valuable ingredient
41 for several processed foods [Rocchetti et al., 2020].

42 China is the major producing country, with other top producers Uzbekistan, Russia and the USA,
43 contributing together to nearly 50% of the world carrot production; among other leading producers
44 are Ukraine, Poland, the United Kingdom, Germany, Indonesia, Turkey and France [Arscott and
45 Tanumihardjo, 2010]. Besides physical quality attributes (such as size, shape, uniformity, color and
46 texture), sensory and nutritional quality attributes are also important for carrots. As with other
47 vegetables, these attributes are dependent on different factors: the cultivar, the geographical origin,
48 the production method and post-harvest conditions [Pereira et al., 2016; Koudela et al., 2021; Ahmad
49 et al., 2019]. Since carrots are among the most popular agricultural commodities in the world and
50 among the world's most economically important vegetables, food control is an important issue to
51 guaranty the quality of carrots put on the market, as well as the fair labelling of their geographical
52 origin and production mode.

53 Plant metabolites deserve extensive research in the food control field, especially secondary
54 metabolites that are responsible for several bioactive properties of plants and may serve as
55 biomarkers for plant characterization as recently reviewed [Pedrosa et al., 2021]. For vegetables,
56 several interesting applications have been reported in food authentication, mainly based on LC
57 methods targeted on several molecular markers and combined with chemometrics. As an illustration,
58 UHPLC-HRMS or UHPLC-HRMS/MS were applied to the determination of phenolic compounds:
59 phenolic profiles and concentration levels were good chemical descriptors when using chemometric
60 tools (principal component analysis (PCA) and partial least squares - discriminant analysis (PLS-DA))
61 to separate paprika samples based on the production region and flavor varieties [Barbosa et al.,
62 2020]. For carrots, phenolic acids (such as 5-O-caffeoylquinic acid) were investigated as possible
63 biomarkers of the production mode, but no statistical differences between organic and conventional
64 growth systems were found [Soltoft et al., 2010].

65 In recent years, metabolomics-based untargeted analytical approaches have raised a growing
66 interest in the food control field. They seem to be promising in this field as they allow to obtain a
67 global view of the sample, opening the way to the classification of food samples based on several
68 descriptors referring to different quality attributes. Several applications to plant-based foods (fruits,
69 vegetables, spices) have been reported. The phenolic compounds fingerprinting of paprika samples
70 acquired by LC and fluorescence detection enabled their classification according to the production
71 region after data treatment by PLS-DA [Campmajo et al., 2021]. Mie et al. (2014) discriminated
72 conventional and organic white cabbage during a long-term study (2 years) thanks to the untargeted
73 analysis of 1,600 compounds of the plant metabolome by UHPLC-HRMS combined with
74 chemometrics (PCA separated samples by production year, and orthogonal partial least squares -
75 discriminant analysis (OPLS-DA) models discriminated 83% samples based on the production mode);
76 in their work, white cabbage samples were produced in only three farms in Denmark, covering a
77 narrow geographical zone. Using a similar methodology (UHPLC-HRMS, PCA and linear discriminant
78 analysis (LDA)), discrimination between organic and conventional tomato crops was achieved by
79 other authors with 73% of samples being correctly classified [Martinez Bueno et al., 2018]. Other
80 authors successfully authenticated organic oranges using a similar approach (LC-HRMS and OPLS-DA
81 model), with approximately 90% of samples correctly classified [Cuevas et al., 2017]. Another work
82 on goldenberry samples applying UHPLC-HRMS combined to PCA permitted the separation of sample
83 groups based on organic and conventional productions [Llano et al., 2018].

84 Yet, applications of LC-MS-based untargeted analytical approaches to carrot samples are scarce. To
85 our knowledge, only one recent study in the area of organic food authenticity has been reported
86 [Cubero-Leon et al., 2018]. In this work, carrot samples of two varieties (Nerac and Namur) were
87 collected in two Belgian Walloon regions over four consecutive years – each time both organic and

88 conventional production modes were considered. Combining LC-HRMS with OPLS-DA models
89 correctly classified 100% of unknown carrot samples if models were refined to exclude variables
90 contributing to the production year [Cubero-Leon et al., 2018]. These promising results need to be
91 confirmed by other studies that take into account a greater diversity of geographical origins.
92 Interestingly, Cubero-Leon et al. (2018) identified several markers related to carbohydrate
93 metabolism and plant defense mechanism as responsible for the differences between organic and
94 conventional growth systems.

95 Carrots are interesting vegetables since they are produced during almost all seasons, which
96 maximizes the possibility to find samples and makes sample collection easier. For carrots, sample
97 discrimination based on both geographical origin and production mode would be of great value in
98 practice for the laboratories and control authorities due to their worldwide production. In France,
99 carrots are the organic vegetables with the biggest market size (125 M€ in 2019) and the 5th biggest
100 market for organic food products. Although the additional cost of organic carrots compared to
101 conventional carrots is limited (25%, whereas organic peaches are 200% more expensive than
102 conventional peaches), the economic impact of fraud on organic carrots is substantial because of the
103 market size of this product.

104 In this study, we have investigated the capability of a metabolomics-based untargeted analytical
105 method combined with chemometrics tools for carrots discrimination based on these two factors.
106 Carrot samples produced in several French agricultural production regions with either organic or
107 conventional methods were analyzed by untargeted UHPLC-HRMS. The novelty of our work lies in
108 the realization of a representative sampling of the carrot production in France. In addition, all carrot
109 samples were systematically analyzed using two chromatographic modes: reversed-phase (on a non-
110 polar C18-silica column, classically used in applications reported in literature) and HILIC (on a
111 dedicated polar column, to preferentially retain polar to highly polar compounds). Indeed, as amino
112 acids were identified as possible marker compounds for authentication of organic foods [Dinis et al.,
113 2022; Cuevas et al., 2017; Mihailova et al., 2021], polar chromatographic columns might be
114 interesting. One previous study reported the targeted analysis of polar phospholipids, small peptides
115 and amino acids using HILIC mode for the assessment of garlic authenticity [Hrbek et al. 2018]. In our
116 work, results obtained from both chromatographic columns were compared to assess which family of
117 compounds (polar or non-polar compounds) provides better discrimination between carrot samples.

118 **2. Materials and Method**

119 *2.1. Reagents and chemicals*

120 Acetonitrile (ACN), methanol, water and formic acid (FA), all LC-MS grade, were purchased from
121 Fisher Scientific. Ammonium formate (LC-MS grade) was purchased from Sigma-Aldrich.

122 Different internal standards listed below were used to assess the analytical performance and stability
123 by monitoring their *m/z* and retention time. Carbaryl-d7 (purity: 98%) and a mix containing 9
124 pesticides (namely: boscalid (purity: 98%), captan (purity: 99%), chlorantraniliprole (purity: 98%),
125 daminozide (purity: 99%), dimethomorph (purity: 99%), fenhexamid (purity: 99%), flonicamid (purity:
126 99%), hexythiazox (purity: 99%) and pyridaben (purity: 99%)) were both purchased from Restek. The
127 carbaryl-d7 standard (20 µg/mL) as well as the mix standard (each pesticide at 100 µg/mL) were in
128 acetonitrile. The pesticide mix was chosen to cover a wide range of retention times and *m/z* values,
129 with pesticides unexpected on carrot samples; it was used as an internal standard to control sample
130 preparation.

131 The Pierce™ positive and negative ion calibration solution (purchased from Thermo Fisher Scientific)
132 was used to weekly calibrate the MS detector.

133 *2.2. Samples description and preparation*

134 Forty-four carrot samples were collected in the frame of the TOFoo (True Organic Food) collaborative
135 project (see <https://www.tofoo-project.com/en/>). Carrots were sent by their producers from several
136 regions of France to the laboratory. Thus, the samples collection depended on the producers'
137 possibilities and on the harvest time. Most of the carrots were collected between January, 25th and
138 February, 24th of year 2021.

139 After collection, the samples were crushed, homogenized, and placed in a 50 mL flask in the
140 laboratory and stored at -20 °C before analysis. After defrosting, aliquots (5 g) of samples were
141 extracted in 5 ml of water and 15 ml of methanol. Next, 10 µl of the pesticide mix standard was
142 added to control the sample preparation. Samples were then homogenized using an Ultra-Turrax for
143 30 sec. Samples were next agitated for 10 min using a mechanical stirring and further centrifuged for
144 5 min at 4,000 rpm. The supernatant was collected and introduced directly in a vial before analysis.
145 Then, 20 µl of the carbaryl-d7 standard was also added in the vial as an internal standard for
146 injection. Two replicates per sample (i.e. all extractions were performed in duplicate) were prepared
147 for each column analysis.

148 Samples were analyzed in two different analytical batches, each composed of 30 samples. The first
149 batch was dedicated to models building. Data from the second batch were used for models'
150 validation and markers tentative identification, since fragmentations were performed during this
151 experiment. MS/MS data acquisition was performed in a second step to obtain more information on
152 the discriminant markers after evaluating the ability of the untargeted UHPLC-HRMS method to
153 authenticate carrots samples.

154 The first batch contained 19 organic and 11 conventional carrot samples, and the second batch 14
155 organic and 16 conventional carrot samples. A detailed list of the samples is presented in **Table A.1**
156 and **Table A.2** (Supplementary material). It can be observed that our study samples came mainly
157 from two French regions, namely Normandy and New Aquitaine. Two samples in the first batch were
158 from two additional geographical origins (Brittany and Hauts-de-France).

159 Within each batch, all the samples were pooled to prepare a Quality Control (QC) sample, as well as
160 diluted QC samples (2-fold and 5-fold in water). Each analytical sequence was then built based on
161 randomized sample vials, QC vials every 10 sample vials as well as analytical blanks every 20 sample
162 vials. MS/MS acquisitions were performed during the analysis of the second batch.

163 *2.3. Analytical conditions*

164 The UHPLC-HRMS system used was from ThermoFisher® composed of a Vanquish Flex UHPLC system
165 (equipped with a binary pump, a refrigerated sampler and a column oven) and an Orbitrap® mass
166 spectrometer QExactive Plus with a heated electrospray ion source (HESI). Chromatographic
167 separation was performed using either a hydrophobic C18-silica (reversed-phase mode) or a
168 hydrophilic one (HILIC mode). In both cases, the column temperature was set at 30°C, and the
169 injection volume was 1 µL.

170 *2.3.1. Reversed-phase mode*

171 The UHPLC separation was achieved using the C18 Hypersil Gold column detailed in our previous
172 work with the same injection volume, flow-rate and linear gradient elution (Dinis et al., 2022). Mobile

173 phases compositions were slightly different to improve the separation: water acidified with 0.05% FA
174 and 5 mM ammonium formate (A), and ACN acidified with 0.05% FA (B). Also, a 0.1 ml/min of ACN
175 was added before the HRMS acquisition to improve compounds ionization. An example of
176 chromatogram for a QC is given in **Fig. A.1**.

177 *2.3.2. HILIC mode*

178 The UHPLC separation was achieved using an amide-HILIC column (150 x 2.1 mm, 2.6 μ m) at a 0.6
179 ml/min flow-rate. The mobile phases were water acidified with 0.05% FA and 5 mM ammonium
180 formate (A), and ACN (B), with the following linear gradient: 0-2 min, B: 98%; 2-8 min, B: 98-75%; 8-
181 8.1 min: B: 75-40%; 8.1-10 min, B: 40%; 10-10.1 min, B: 40-98%; 10.1-15 min, B: 98%. An example of
182 chromatogram for a QC is given in **Fig. A.1**.

183 *2.3.3. MS and MS/MS data acquisition*

184 MS data were acquired in centroid mode using TraceFinder software (version 3.1, ThermoFisher®),
185 with a mass range from m/z 100 to 1000. The electrospray ion source was operated in the positive
186 ion mode (ESI+) (detailed conditions are presented in **Table A.3** of Supplementary material) and the
187 resolution fixed at 140,000.

188 MS/MS data were obtained by performing full scan data-dependent analyses using an inclusion list
189 dedicated for each chromatographic mode (the list was composed of 25 or 12 features for the
190 reversed-phase or HILIC mode, respectively). The inclusion lists were built after the data processing
191 of the first sample batch where several features were identified as discriminant. The resolution was
192 set at 17,500. An isolation window of ± 1 uma was used and three normalized collision energies were
193 applied (10; 30 and 60 eV) for MS/MS acquisition.

194 *2.4. Data processing*

195 After data files conversion to mzXML format using ProteoWizard, raw data were processed following
196 a workflow originally developed on the online W4M platform [Dinis et al., 2022], and adapted here
197 for its implementation in-house on the R software (see **Fig. A.2** of the Supplementary material). Raw
198 data obtained from the two different analytical columns were processed separately. The “features”,
199 defined by their m/z and retention time (RT), and their intensities in different samples were used for
200 the statistical analysis as commonly reported. Those found to be the most discriminant between the
201 sample groups were tentatively identified using online databases, as described below.

202 *2.4.1. Data treatment using the XCMS R-package*

203 The first steps of the workflow were performed using several XCMS functions: (1) peak detection and
204 extraction using the “centWave” method of the “findChromPeaks” function (this first step
205 transformed the chromatograms into a 2D-matrix where each peak was described by its feature); (2)
206 peaks grouping according to their RT using the “groupChromPeaks” function (peaks were grouped
207 across the samples according to their RT values, and values of m/z and RT were then averaged in the
208 data matrix); (3) peak alignment using the “adjustRtime” function (RT deviation was corrected); (4)
209 peaks grouping again; (5) missing intensity values were completed using the “fillChromPeaks”
210 function. All XCMS parameters used for these steps are presented in **Table A.4** (Supplementary
211 material). A data matrix was then generated, giving the integration of each peak made by XCMS for
212 each feature and for each sample.

213 *2.4.2. Data filtration and analytical drift correction*

214 Several successive filtration steps were used to remove irrelevant information present on the
215 previously obtained data matrix, as the number of features detected by XCMS was very high (about
216 20,000 for the C18-silica analysis and about 10,000 for the HILIC analysis). Firstly, all features
217 detected in the dead volume and during the column equilibration were removed (for C18-silica
218 analysis: before 1.7 min and after 26 min; for HILIC analysis: before 0.75 min and after 12 min).
219 Secondly, features that were primarily detected in blank analyses were removed based on their fold
220 change ($FC(\text{samples})/FC(\text{blanks}) < 4$). Thirdly, features with low stability in the QC analyses were
221 removed based on their relative standard deviation ($RSD > 30\%$). Also, features for which the ratio of
222 RSD pool over RSD sample was higher than 1.25 were deleted. Next, the analytical signal drift was
223 corrected using a LOESS regression method from the replicate injections of the QC sample. Finally,
224 dilution effects were removed using the probabilistic quotient normalization (PQN). Before
225 chemometrics, the data matrix was Pareto scaled.

226 2.4.3. Chemometrics

227 PCA was considered for exploratory purpose of the data sets and to detect outliers; no outliers were
228 detected. Then classification of samples was investigated by building PLS-DA and OPLS-DA models
229 using a 7-fold cross validation [Dinis et al., 2022]. This cross validation permitted to determine the
230 optimal number of latent variables (LV) to build the models. Several metrics were considered to
231 assess the quality of our models: the goodness of fit (R^2X), the proportion of the response matrix
232 variance explained (R^2Y), and the predictive performance (Q^2Y). Their values are in the range 0-1,
233 and the model performance is increasing with their values. The Q^2Y metric and the root mean square
234 error of prediction (RMSEP) were particularly considered here, as they represent the prediction
235 efficiency of the model.

236 To evaluate further the classification performance of the PLS-DA and OPLS-DA models, the data set
237 was divided into four different subsets between the calibration set and validation set. Four PLS-DA
238 models and four OPLS-DA models were thus obtained using the four subsets. The confusion matrix of
239 each model was then extracted, containing the number of true positive samples (TP, samples
240 correctly classified in class A), the number of true negative samples (TN, samples correctly classified
241 in class B), the number of false positive samples (FP, samples incorrectly classified in class A) and
242 false negative samples (FN, samples incorrectly classified in class B). The resulting confusion matrices
243 for these four subsets were used to calculate the class sensitivity, the class specificity, the model
244 accuracy, and the number of misclassifications (NMC) as detailed below [Riedl et al., 2015]:

$$245 \text{ Sensitivity} = \frac{TP}{TP + FN}$$

$$246 \text{ Specificity} = \frac{TN}{TN + FP}$$

$$247 \text{ Accuracy} = \frac{TP + TN}{TP + FN + TN + FP}$$

$$248 \text{ NMC} = FP + FN$$

249 As the number of detected features was quite high, feature selection was necessary to identify the
250 most discriminant features that could be considered as discriminant markers. An ANOVA was then
251 performed to select significant features between the two studied groups (the two regions of interest
252 or organic vs. conventional carrots respectively); a maximal accepted p-value of 0.01 was chosen. The
253 Bonferroni correction method was applied to limit the number of false positive results. New PLS-DA

254 and OPLS-DA models were then built based on the features selected by ANOVA to improve the
255 performance of the models.

256 The chemometrics were applied on the data matrix obtained for each chromatographic column
257 separately, as well as after fusion of both data matrices. We aimed to investigate if the discrimination
258 models could be improved by combining the information from both chromatographic modes. Some
259 factors such as geographical origin and sample variety are described as the most important sources
260 of sample variability [Cubero-Leon et al., 2018; Mihailova et al., 2021]. Thus, PLS-DA and OPLS-DA
261 models were first built based on the geographical origin of the samples. The features with a high VIP
262 (variable importance on projection) were then removed from the data matrix to remove variability
263 related to this factor; a VIP value filtration criterion of 1 was applied as previously tested in other
264 study [Dinis et al., 2022]. Then, new PLS-DA and OPLS-DA models were built for the discrimination
265 between organic and conventional samples.

266 *2.4.4.Annotation*

267 The number of features was quite high as detailed above. It was reduced by filtering the features
268 according to their VIP value (calculated during the construction of the PLS-DA and OPLS-DA models)
269 as well as their p-value (from the ANOVA results) as this strategy was found efficient as already
270 reported [Dinis et al., 2022].

271 The MS spectra of the remaining features were first studied, which allowed the determination of the
272 adduct types of the observed m/z and thus the exact mass of the compounds and, consequently,
273 molecular formulas were suggested for each feature. An online database (FooDB) was then used to
274 tentatively annotate these discriminant features. Moreover, the MS/MS experiments were used to
275 confirm the tentative annotation by the help of spectral databases (MassBank and mzCloud).

276 **3. Results and Discussion**

277 The results obtained with the two analytical columns were quite similar. It was therefore decided to
278 describe mainly the results obtained with the reversed-phase column because it is a widely used
279 column. Some results obtained with the HILIC column are presented in the Supplementary material.
280 Combining the data provided by both chromatographic modes was tested but this did not improve
281 our results.

282 *3.1. Authentication of the geographical origin of carrots*

283 The dataset used for the reversed-phase mode contained 30 samples (in columns) and 4,652 features
284 (in rows). The dataset used for the HILIC mode contained 30 samples (in columns) and 1,426 features
285 (in rows). After fusion of both data matrices, the dataset used contained 30 samples (in columns) and
286 6,078 features (in rows).

287 The PCA score plot from the C18-silica column permitted to observe a cluster of the QC samples (see
288 **Fig. 1**), highlighting a good system stability during the analysis. Near 50% of variance was explained
289 by the first three principal components (PC1: 22%, PC2: 19%; PC3: 8%). Similar results were observed
290 for data obtained with the HILIC column (**Fig. A.3** of supplementary material). The two replicates per
291 samples were grouped, showing good reproducibility. Thus, mean of the two replicates was used for
292 the following chemometric tools.

293 Despite the fact that no clear separation could be obtained between the sample groups based on
294 their geographical origin, a trend seemed to emerge on the PC1 axis regarding the separation of the
295 New Aquitaine and the Normandy sample groups. It is also interesting to note that Brittany and New

296 Aquitaine samples appear in proximity possibly due to their geographic proximity, with the same
 297 trend observed for Hauts-de-France and Normandy samples. Since Brittany and Hauts-de-France
 298 concerned by our samples were regions represented by only one sample each, it was decided to
 299 discard them as these two geographical origins were not sufficiently represented to be modeled. The
 300 discriminant and classification models built were therefore only based on the Normandy and New
 301 Aquitaine samples.

302 The PLS-DA model was built using three latent variables and showed a good predictive ability (Q2Y:
 303 0.632) as shown in **Fig. 2**. The OPLS-DA model was built using three orthogonal latent variables and
 304 showed a similar predictive ability (Q2Y: 0.697). The PLS-DA and OPLS-DA models obtained with the
 305 four different subsets showed very similar results, with acceptable sensitivity and specificity as
 306 indicated in **Table 1**, showing good capacity of the models to correctly classify the samples in their
 307 belonging class. Overall, about 85% of model accuracy was obtained. In all the models, at least one
 308 sample was misclassified. The observed Q2Y values (about 0.7) showed good predictive ability for the
 309 obtained models, but the quite high RMSEP values (around 0.4), indicated that an error on the
 310 prediction of new samples may occur. The HILIC mode analysis showed very close results (**Fig. A.4**
 311 and **Table A.5** of supplementary materials), with slightly better Q2Y and RMSEP values (about 0.75
 312 and 0.35 respectively).

313

314 **Table 1:** Metrics obtained from the confusion matrices on four different subsets to assess the
 315 performance of PLS-DA and OPLS-DA models for carrot samples discrimination based on their
 316 geographical origin using the reversed-phase mode analysis

# Subset	Number of samples (calibration set)	Number of samples (test set)	Sensitivity	Specificity	Accuracy	Q2Y	RMSEP	NMC
<i>PLS-DA models</i>								
1	20	8	100.0%	75.0%	87.5%	0.707	0.349	1
2	20	8	100.0%	75.0%	87.5%	0.699	0.396	1
3	19	9	60.0%	100.0%	77.8%	0.733	0.441	2
4	19	9	100.0%	80.0%	88.9%	0.546	0.402	1
<i>OPLS-DA models</i>								
1	20	8	100.0%	75.0%	87.5%	0.854	0.348	1
2	20	8	100.0%	75.0%	87.5%	0.784	0.368	1
3	19	9	60.0%	100.0%	77.8%	0.739	0.450	2
4	19	9	100.0%	80.0%	88.9%	0.659	0.402	1

317

318 Geographical origin of carrots has already been assessed by other methods in previous studies.
 319 Jandric et al. used isotope ratios of strontium combined with the content of 12 elements to classify
 320 carrots from different Austrian regions with OPLS-DA [Jandric et al., 2021]. Targeted LC-MS also
 321 showed good classification rates on the same dataset. Magdas et al. reported the potential of rare
 322 earth elements (REEs) measurement by ICP-MS to discriminate geographical origin of several food
 323 matrices, among which carrots [Magdas et al., 2019]. Models built on REEs values showed a good
 324 stability with harvest year, which could not be tested in this study. However, isotopic ratios and REEs
 325 are more likely to fail to discriminate geographical origins on processed or mixed food due to
 326 fractionation.

327 To our knowledge, our study reports for the first time the successful discrimination of carrot samples
328 based on geographical origin using untargeted LC-HRMS analysis. It would be an interesting
329 perspective to test the validity of our models presented here on processed or mixed carrots, and
330 compare their performances with other reported methods such as NMR. Indeed, NMR could
331 discriminate carrot juices from three locations in Italy [Tomassini et al., 2016]. The authors showed
332 that the different profiles were linked to the pedoclimatic conditions, which could be the case in our
333 study as well: there were indeed around 1,500 hours of sunlight in Normandy in 2020, whereas in
334 New Aquitaine, this number rises to 2,100 hours.

335 *3.2. Authentication of organic carrots*

336 For the sample discrimination based on their production mode, all 30 samples of the first batch were
337 considered: 19 organic and 11 conventional carrot samples (see **Table A.1**). Similar to the
338 methodology applied by Cubero-Leon et al. (2018), features having a VIP higher than 1 for the
339 geographical origin discrimination were removed from the data matrix to discard variability related
340 to this factor. As previously stated, the mean value of the two replicates was used to build the
341 models. After this filtration step, 4,268 features were left in the dataset for the reversed-phase mode
342 and 1,314 features for the HILIC mode dataset.

343 PCA was unsuccessful to distinguish between organic and conventional samples as illustrated in **Fig.**
344 **A.5** (Supplementary material). Similar results were reported for carrot samples by Cubero-Leon et al.
345 (2018). They suggested that this may be due to the fact that PCA is an unsupervised technique and
346 also to other factors (such as variety) that may further influence differences in metabolites.

347 The PLS-DA model was built using three latent variables and the OPLS-DA model was built using three
348 orthogonal latent variables. The Q²Y obtained with these models were not as good as those obtained
349 for the discrimination based on the geographical origin as presented in **Fig. 3** (Q²Y: 0.304 for PLS-DA
350 and 0.154 for OPLS-DA). Globally, for all the distributions tested, about 65% of model accuracy was
351 obtained (see **Table A.6**). In particular, conventional samples were rarely recognized as being
352 conventional samples by the different models as described by the specificity values obtained. On the
353 contrary, organic samples were mostly correctly recognized, as described by the sensitivity values.
354 The NMC obtained were higher than when dealing with the discrimination of the geographical origin
355 (between 3 and 5 samples). The RMSEP values were also higher (around 0.5) and the Q²Y values
356 lower (about 0.35), showing that the obtained models had a worse classification performance. The
357 results from the HILIC mode gave very similar results (**Fig. A.6** and **Table A.7** of supplementary
358 material), with even higher NMC values (between 5 and 8 samples). We may attribute this
359 unsatisfactory result to the unbalanced composition of our sample batch, and subsequently the low
360 number of conventional samples present in the data set (11 in total). As already reported,
361 classification models describe better the majority class, leading to poor prediction accuracy for the
362 minority class. Unfortunately, the dataset of this study was too small to allow the use of most data
363 balancing techniques.

364 Several explanations may be proposed to explain the poor performances of these discrimination
365 models. Firstly, these poor performances may be linked to the variety of the studied samples. Ten
366 different varieties were present in our sample set, most of them having only one sample per variety
367 (see **Table A.1** in Supplementary material). Variety has already been identified as a strong source of
368 variability for this type of study as the composition of carrots may vary with the variety [Arscott and
369 Tanumihardjo, 2010]. In their long-term study of carrots, Cubero-Leon et al. (2018) considered only
370 two varieties which could explain the good metrics of their model. It also has to be noticed that their
371 carrot samples came from the same geographical area (all fields were within 20 km). Another

372 explanation for the poor performances of the model presented here is that removing the features
373 which contribute the most to the geographical discrimination is not sufficient to reduce the
374 variability caused by the geographical origin. However, the number of samples from the same area
375 was too low to test this hypothesis here. Finally, the poor performances of these models may be due
376 to the limited number of samples in our study (30 samples), since it is generally advised to consider
377 over 40 to 60 samples (Mialon et al., 2023).

378 Globally, few studies could successfully differentiate conventional from organic carrots. In particular,
379 the nitrogen isotope ratios failed to discriminate organic and conventional carrot crops, whereas the
380 same method was successfully applied on tomatoes and lettuce [Bateman et al., 2007]. An
381 explanation suggested by the authors is the need for nitrogen during growth which is lower for
382 carrots than for tomatoes and lettuces. Similarly, no discrimination model using NMR has been
383 published for carrots, although this technique has proven its ability to discriminate organic and
384 conventional plants for many other different fruits and vegetables [Pacifico et al., 2013]. To our
385 knowledge, only one study reported successful discrimination between organic and conventional
386 carrots by analyzing oxygen isotope ratio of sulphate [Novak et al., 2019]; however, the built model
387 was not tested on commercial samples.

388 *3.3. Annotation and tentative identification of discriminant features for the geographical* 389 *origin*

390 A VIP value filtration criterion of 1 remained approximately 100 features left after filtration. Using
391 further the ANOVA results and focusing on features with the lowest p-value, 18 features were
392 selected for the reverse-phase mode dataset as indicated in **Table 2**, and 10 features for the HILIC
393 mode dataset (see **Table A.8** in Supplementary material). No additional discriminant feature could be
394 obtained by combining the information from both chromatographic modes.

395 Arginine was tentatively identified and confirmed by its MS/MS spectrum (see **Fig. 4**), in line with
396 previous work reporting this amino acid as a potential marker of the agricultural practice for carrots
397 [Cubero-Leon et al., 2018]. In particular, these authors linked the difference observed in the carrot
398 metabolome to the impact of fertilization on the soil metabolome. If this hypothesis is valid, the
399 difference observed in our study could result from a difference in soil metabolome between
400 Normandy and New Aquitaine. Arginine was also reported in garlic as a possible biomarker of the
401 geographical origin [Hrbek et al., 2018].

402 Other amino acid derivatives were suspected as markers, although they could not be confirmed
403 based on their MS/MS spectra (see **Table 2**). Our results are in line with previous ones showing that
404 amino acids are strong markers of the geographical origin for carrots because their concentrations
405 are linked to acclimation processes and thus depend on pedoclimatic conditions [Tomassini et al.,
406 2016]. Sciubba et al. observed a correlation between some amino acid concentrations in carrots and
407 the harvest time [Sciubba et al., 2020]. However, this hypothesis could not be confirmed in our study,
408 as carrots from geographical areas were regularly collected during the harvest period of time.

409 Another discriminant marker (feature 16) could also be tentatively identified and confirmed by
410 MS/MS data. Several suspected compounds were discarded based on their MS/MS spectra and also
411 since they had never been reported in carrots (neither from FoodDB nor from the literature).
412 Sinapaldehyde was a suspected compound for this feature, already reported in foods according to
413 FoodDB. However we discarded this molecule due to some differences observed between its
414 experimental MS/MS spectrum with the spectra reported in FoodDB as shown in **Fig. A.7**
415 (Supplementary material). We finally tentatively identified feature 16 as being 6-methoxymellein (or

416 phytoalexin 8-hydroxy-3-methyl-6-methoxy-3,4-dihydroisocoumarin) as indicated in **Table 2** and
417 shown in **Fig. 4**.

418 This compound has already been reported in fresh carrots, with varying levels (0.02 to 76 µg/g)
419 depending on carrot genotype, maturity as well as environmental and growing conditions [De
420 Girolamo et al., 2004]. It is generally considered of being associated with the bitterness in strained
421 carrots, and partly responsible for the sensory quality of carrots. A threshold concentration of 94
422 µg/g has been reported to bitter flavor in carrots [De Girolamo et al. 2004]. Levels of 6-
423 methoxymellein in carrots were found stable at low temperature for quite long periods.

424 Examples of chromatograms for the discriminant features identified according to the MS/MS
425 acquisitions for each analytical column are presented in **Fig. A.8** and **Fig. A.9** (Supplementary
426 material).

427 Two additional biomarkers could be tentatively identified based on their MS/MS spectra thanks to
428 the results from the HILIC polar column as indicated in **Table A.8**, **Fig. A.9** and **Fig. A.10**
429 (Supplementary material): N-acetylputrescine and L-carnitine.

430 N-Acetylputrescine was reported in plants, especially in *Daucus carota*, being linked to plant defense
431 mechanism [Lou et al., 2016]. This molecule comes from the acetylation of putrescine, a polyamine
432 biosynthesised from arginine that may accumulate in plants in response to several biotic and abiotic
433 stresses. Interestingly, in our work both arginine avec N-acetylputrescine are found as potential
434 biomarkers of the New Aquitaine geographical origin.

435 L-Carnitine (or 2-hydroxy-4-trimethylammonium butyrate) is an amino acid already reported in
436 plants, playing a role in their lipid metabolism [Bourdin et al., 2007]. Concentrations of 37 µg/g have
437 been recently reported in carrots [Kepka et al., 2021], and nutritional sources of carnitine are
438 considered as important for the human body. It is suspected to be biosynthesised in the plant from
439 gamma-butyrobetaine (or 4-trimethylammonium butyrate) as a precursor [Rippa et al., 2012]. In our
440 work, we found both L-carnitine as a potential marker of the Normandy geographical origin (see
441 **Table A.8** and **Fig A.9**), and gamma-butyrobetaine was considered as a possible compound for
442 feature 7. However, 3-dehydroxycarnitine was also a possible compound for this feature, even
443 though to our knowledge it has never been reported in plants.

444 During the MS/MS experiments, a full scan analysis was also acquired. It was thus possible to use
445 these independent acquisitions to evaluate the potential of the discriminant features to serve as
446 marker compounds. It is noteworthy that the majority of the selected features were successfully
447 observed on this second sample batch (only three features for the reversed-phase mode and one
448 feature for the HILIC mode showed a very low intensity). No significant difference in intensity was
449 observed between the two batches of samples, but a trend was noted for some features (12 features
450 for the reversed-phase mode and 6 features for the HILIC mode). It is thus very likely that the
451 combination of several markers would be necessary to discriminate geographical origins of carrots.
452 Future work should be conducted in that field. In addition, implementation of data processing
453 workflow may allow to standardize signals obtained in different working sessions.

454
455
456

457 **Table 2:** Discriminant features for the authentication of geographical origin (compounds confirmed based on MS/MS data are indicated in bold characters)
 458 with the C18-silica column.

# Features	Detected <i>m/z</i>	RT (min)	<i>p</i> -value	VIP	Characteristic	Adduct type	Monoisotopic mass	Proposed molecular formulas	Tentative identification
1	256.1291	1.78	1,67E-06	6,5	Normandy	[M+H] ⁺	255.1219	C ₁₁ H ₁₇ N ₃ O ₄	<i>N</i> -alpha-(<i>tert</i> -Butoxycarbonyl)-L-histidine*
2	175.1190	2.02	7,07E-03	4,4	New Aquitaine	[M+H] ⁺	174.1117	C ₆ H ₁₄ N ₄ O ₂	Arginine
3	229.1548	2.64	5,11E-03	3,1	Normandy	[M+H] ⁺	228.1474	C ₁₁ H ₂₀ N ₂ O ₃	<i>L</i> -Isoleucyl-L-proline*, <i>L</i> -Leucyl-L-proline*
4	146.0812	2.85	3,32E-04	2,2	New Aquitaine	[M+H] ⁺	145.0739	C ₆ H ₁₁ NO ₃	<i>Allysine</i> *, <i>L</i> -cis-4-(Hydroxymethyl)-2-pyrrolidinecarboxylic acid*, 4-Acetamidobutanoic acid*
5	188.0706 205.0971	6.5	1,22E-04	7,6	New Aquitaine	[M+H] ⁺ [M+NH ₄] ⁺	187.0633	C ₁₁ H ₉ NO ₂	3-(1 <i>H</i> -Indol-3-yl)-2-propenoic acid*
6	383.1312	6.64	9,66E-04	1,4	Normandy	[M+H] ⁺	382.1237 382.1232 382.1250 382.1250 382.1224	C ₁₄ H ₁₈ N ₆ O ₇ C ₂₈ H ₁₆ NO C ₁₅ H ₁₄ N ₁₀ O ₃ C ₁₆ H ₂₀ N ₃ O ₈ C ₁₃ H ₂₂ N ₂ O ₁₁	n.a.
7	378.1757	6.65	5,70E-04	1,8	Normandy	[M+H] ⁺	377.1686 377.1672 377.1672 377.1699	C ₁₆ H ₂₇ NO ₉ C ₁₄ H ₂₅ N ₄ O ₈ C ₁₃ H ₁₉ N ₁₁ O ₃ C ₁₇ H ₂₃ N ₅ O ₅	n.a.
8	342.1758	7.15	1,76E-08	1,1	Normandy	[M+H] ⁺	341.1686 341.1672 341.1699	C ₁₃ H ₂₇ NO ₉ C ₁₀ H ₁₉ N ₁₁ O ₃ C ₁₄ H ₂₃ N ₅ O ₅	n.a.
9	390.1393	8.31	3,83E-05	1,8	Normandy	[M+H] ⁺	389.1322 389.1309 389.1308 389.1335	C ₁₆ H ₂₃ NO ₁₀ C ₁₄ H ₂₁ N ₄ O ₉ C ₁₃ H ₁₅ N ₁₁ O ₄ C ₁₇ H ₁₉ N ₅ O ₆	n.a.
10	490.2282	8.45	3,37E-07	1,2	Normandy	[M+H] ⁺	489.2210 489.2205 489.2218 489.2197 489.2223	C ₂₂ H ₃₅ NO ₁₁ C ₃₅ H ₂₇ N ₃ C ₃₇ H ₂₉ O C ₂₀ H ₃₃ N ₄ O ₁₀ C ₂₃ H ₃₁ N ₅ O ₇	n.a.

11	432.1864	9.08	1,60E-04	2,0	Normandy	[M+H] ⁺	431.1791 431.1791 431.1791 431.1800 431.1778 431.1778 431.1805	C ₁₇ H ₁₇ N ₁₅ C ₁₈ H ₂₃ N ₈ O ₅ C ₁₉ H ₂₉ NO ₁₀ C ₃₄ H ₂₃ C ₁₇ H ₂₇ N ₄ O ₉ C ₁₆ H ₂₁ N ₁₁ O ₄ C ₂₀ H ₂₅ N ₅ O ₆	n.a.
12	354.1757	9.16	8,72E-03	1,2	New Aquitaine	[M+H] ⁺	353.1699	C ₁₅ H ₂₃ N ₅ O ₅	<i>Dihydrozeatin riboside*</i> , 2-[[1-(2-amino-3-hydroxybutanoyl)pyrrolidine-2-carbonyl]amino]-3-(1H-imidazol-5-yl)propanoic acid*
13	348.2744 313.2371	12.27	2,36E-03	16,5	Normandy	[M+H] ⁺ Fragment	347.2672 347.2672 347.2658 347.2685	C ₁₇ H ₃₁ N ₈ C ₁₈ H ₃₇ NO ₅ C ₁₆ H ₃₅ N ₄ O ₄ C ₁₉ H ₃₃ N ₅ O	n.a.
14	683.4699 353.2296	12.27	7,09E-03	4,6	Normandy	[M+H] ⁺ Fragment	682.4629 682.4616 682.4611 682.4597 682.4656	C ₃₄ H ₆₂ N ₆ O ₈ C ₃₃ H ₆₆ N ₂ O ₁₂ C ₄₆ H ₅₈ N ₄ O C ₄₅ H ₆₂ O ₅ C ₃₈ H ₆₆ O ₁₀	n.a.
15	699.4347	12.27	2,94E-03	2,7	Normandy	[M+H] ⁺	698.4241	C ₃₇ H ₆₂ O ₁₂	<i>Cyclopassifloside I*</i> , <i>Cyclopassifloside IV*</i> , <i>Cyclopassifloside X*</i>
16	209.0809 191.0703	13.46	6,06E-05	14,5	New Aquitaine	[M+H] ⁺ Fragment	208.0736	C ₁₁ H ₁₂ O ₄	6-Methoxymellein , <i>Sinapaldehyde</i> , <i>Caffeic acid ethyl ester*</i> , <i>Methyl ferulate*</i> , 3-(3,4-Dimethoxyphenyl)-2-propenoic acid*, <i>Furapiole*</i> , 1-(2-Methoxy-3,4-methylenedioxyphenyl)-1-propanone*, 3-(3-Methoxy-4,5-methylenedioxyphenyl)-2-propen-1-ol*, <i>Anthriscinol*</i> , (R)-2-Benzylsuccinate*, 5-(3',4'-Dihydroxyphenyl)-gamma-valerolactone*, 5-(3',5'-Dihydroxyphenyl)-gamma-valerolactone*
17	374.2901	14.42	4,69E-03	2,3	New Aquitaine	[M+H] ⁺	373.2842	C ₂₁ H ₃₅ N ₅ O	1-[[[(2R,4S,5R)-5-(5-cyclohexyl-2-methylpyrazol-3-yl)-1-azabicyclo[2.2.2]octan-2-yl]methyl]-3-ethylurea*
18	305.2474	19.31	8,86E-04	1,7	New Aquitaine	[M+H] ⁺	304.2402	C ₂₀ H ₃₂ O ₂	<i>Oryzalexin D*</i> , <i>Oryzalexin F*</i> , <i>Arachidonic acid*</i> , <i>Sideridiol*</i> , <i>ent-17-Hydroxy-16b-kauran-19-al*</i> , <i>Yucalexin P21*</i> , <i>Copalic acid*</i> , <i>Junicedral*</i> , 7,13-Eperudien-15-oic acid*, <i>Oryzalexin S*</i> , <i>Oryzalexin E*</i> , <i>Mesterolone*</i> , <i>Abietadiene-diol*</i> , <i>Sodium oleate*</i>

459 n.a.: not applicable; * not confirmed based on MS/MS data

460 **4. Conclusion**

461 Forty-four samples of carrots were analysed by UHPLC-HRMS using two different chromatographic
462 phases (C18-silica and HILIC). The samples were collected from various production sites at the
463 beginning of 2021 and originated from different regions of France. PLS-DA and OPLS-DA models were
464 constructed to discriminate production modes (conventional *versus* organic) and geographical origin
465 (New Aquitaine *versus* Normandy). Despite some trend in sample discrimination based on the
466 production mode, the models metrics were not satisfying (Q2Y below 0.5). On the contrary, good
467 results were obtained for the geographical discrimination. An additional batch was analysed to
468 validate the discrimination performance and tentatively identify discriminant markers. Arginine and
469 6-methoxymellein were proposed as potential biomarkers of the geographical origin based on the
470 C18-silica analyses and their MS/MS spectra. The HILIC analyses provided the tentative identification
471 of two additional biomarkers confirmed by their MS/MS spectra: N-acetylputrescine and L-carnitine.

472

473 **Appendix A. Supplementary material**

474

475 **Acknowledgements**

476 The authors warmly thank GRAB, Carottes de France, Larrère, Legum'land, GIE de l'Ombrière, SICA
477 Altus, Valorex, Agrial and SILEBAN for kindly providing them with several samples of carrots from
478 their fields or from suppliers of their network. These samples were collected through the TOFoo
479 (True Organic Food) project, supported by the French Government in the framework of the
480 "Investissements d'avenir" program. The authors are also thankful for the financial support provided
481 through this program, as well as for the financial support provided by the Association Nationale
482 Recherche et Technologie (ANRT, France) through the CIFRE program (CIFRE n°2018/0937).

483

484 **Conflict of interest**

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

487

488 **Credit author statement**

489 **Katy Dinis:** Conceptualization, Methodology, Data curation, Software, Validation, Formal analysis,
490 Writing- Original draft preparation; **Lucie Tsamba.:** Supervision, Writing- Reviewing and Editing; **Eric**
491 **Jamin:** Resources, Funding acquisition, Conceptualization, Project administration; **Valérie Camel:**
492 Supervision, Conceptualization, Writing- Reviewing and Editing.

493

494 **References**

495 Ahmad, T., Cawwod, M., Iqbal, Q., Arino, A., Batool, A., Tariq, R.M.S., Azam, M., & Akhtar, S. (2019).
496 Phytochemicals in *Daucus carota* and their health benefits – review article. *Foods*, 8(9), 424.
497 <https://doi.org/10.3390/foods8090424>

498 Akhtar, S., Rauf, A., Imran, M., Qamar, M., Riaz, M., & Mubarak, M. S. (2017). Black carrot (*Daucus*
499 *carota* L.), dietary and health promoting perspectives of its polyphenols: A review. *Trends in Food*
500 *Science & Technology*, 66, 36–47. <https://doi.org/10.1016/j.tifs.2017.05.004>

501 Arscott, S. A., & Tanumihardjo, S. A. (2010). Carrots of many colors provide basic nutrition and
502 bioavailable phytochemicals acting as a functional food. *Comprehensive Reviews in Food Science and*
503 *Food Safety*, 9(2), 223–239. <https://doi.org/10.1111/j.1541-4337.2009.00103.x>

504 Barbosa, S., Campmajo, G., Saurina, J., Puignou, L., & Nunez, O. (2020). Classification and
505 authentication of paprika by UHPLC-HRMS fingerprinting and multivariate calibration methods (PCA
506 and PLS-DA). *Foods*, 9(4), 486. <https://dx.doi.org/10.3390/foods9040486>

507 Bateman, A. S., Kelly, S. D., & Woolfe, M. (2007). Nitrogen Isotope Composition of Organically and
508 Conventionally Grown Crops. *Journal of Agricultural and Food Chemistry*, 55(7), 2664–2670.
509 <https://doi.org/10.1021/jf0627726>

510 Bourdin, B., Adenier, H., & Perrin, Y. (2007). Carnitine is associated with fatty acid metabolism in
511 plants. *Plant Physiology and Biochemistry*, 45(12), 926-931.
512 <http://dx.doi.org/10.1016/j.plaphy.2007.09.009>

513 Campmajó, G., Rodríguez-Javier, L. R., Saurina, J., & Núñez, O. (2021). Assessment of paprika
514 geographical origin fraud by high-performance liquid chromatography with fluorescence detection
515 (HPLC-FLD) fingerprinting. *Food Chemistry*, 352, 129397.
516 <https://doi.org/10.1016/j.foodchem.2021.129397>

517 Cubero-Leon, E., De Rudder, O., & Maquet, A. (2018). Metabolomics for organic food authentication:
518 Results from a long-term field study in carrots. *Food Chemistry*, 239, 760–770.
519 <https://doi.org/10.1016/j.foodchem.2017.06.161>

520 Cuevas, F. J., Pereira-Caro, G., Moreno-Rojas, J. M., Muñoz-Redondo, J. M., & Ruiz-Moreno, M. J.
521 (2017). Assessment of premium organic orange juices authenticity using HPLC-HR-MS and HS-SPME-
522 GC-MS combining data fusion and chemometrics. *Food Control*, 82, 203–211.
523 <https://doi.org/10.1016/j.foodcont.2017.06.031>

524 De Girolamo A., Solfrizzo M., Vitti C., & Visconti, A. (2004). Occurrence of 6-methoxymellein in fresh
525 and processed carrots and relevant effect of storage and processing. *Journal of Agricultural and Food*
526 *Chemistry*, 52(21), 6478-6484, <https://doi.org/10.1021/jf0491660>

527 Dinis, K., Tsamba, L., Thomas, F., Jamin, E., & Camel, V. (2022). Preliminary authentication of apple
528 juices using untargeted UHPLC-HRMS analysis combined to chemometrics, *Food Control*, 139,
529 109098. <https://doi.org/10.1016/j.foodcont.2022.109098>

530 Hrbek V., Rektorisova M., Chmelarova H., Ovesna J., & Hajslova J. (2018). Authenticity assessment of
531 garlic using a metabolomic approach based on high resolution mass spectrometry, *Journal of Food*
532 *Composition and Analysis*, 67, 19-28. <https://doi.org/10.1016/j.jfca.2017.12.020>

533 Jandric Z., Tchaikovsky A., Zitek A, Causon, T, Stursa V, Prohaska T, & Hann, S. (2021). Multivariate
534 modelling techniques applied to metabolomic, elemental and isotopic fingerprints for the verification
535 of regional geographical origin of Austrian carrots. *Food Chemistry*, 338, 127924.
536 <https://doi.org/10.1016/j.foodchem.2020.127924>.

537 Kepka, A., Ochocinska, A., Chojnowska, S., Borzym-Kluczyk, M., Skorupa, E., Knas, M., & Waszkiewicz,
538 N. (2021). Potential Role of L-Carnitine in Autism Spectrum Disorder. *Journal of Clinical Medicine*,
539 10(6), 1202. <https://doi.org/10.3390/jcm10061202>

540 Koudela, M., Schulzova, V., Krmela, A., Chmelarova, H., Hajslova, J., & Novotny, C. (2021). Effect of
541 agroecological conditions on biologically active compounds and metabolome in carrot. *Cells*, 10(4),
542 784. <https://doi.org/10.3390/cells10040784>

543 Llano, S.M., Muñoz-Jiménez, A.M., Jiménez-Cartagena, C., Londoño-Londoño, J., & Medina, S. (2018).
544 Untargeted metabolomics reveals specific withanolides and fatty acyl glycoside as tentative
545 metabolites to differentiate organic and conventional *Physalis peruviana* fruits. *Food Chemistry*, 244,
546 120–127. <https://doi.org/10.1016/j.foodchem.2017.10.026>

547 Lou, Y-R., Bor, M., Yan, J., Preuss, A.S., & Lander, G. (2016). Arabidopsis NATA1 acetylates putrescine
548 and decreases defense-related hydrogen peroxide accumulation. *Plant Physiology*, 171, 1443–1455.
549 <https://doi.org/10.1104/pp.16.00446>

550 Magdas, D. A., Marincas, O., Cristea, G., Feher, I., & Vedeanu, N. (2020). REEs – a possible tool for
551 geographical origin assessment? *Environmental Chemistry*, 17(2), 148.
552 <https://doi.org/10.1071/EN19163>

553 Martínez Bueno, M.J., Díaz-Galiano, F.J., Rajska, Ł., Cutillas, V., & Fernández-Alba, A.R. (2018). A non-
554 targeted metabolomic approach to identify food markers to support discrimination between organic
555 and conventional tomato crops. *Journal of Chromatography A*, 1546, 66–76.
556 <https://doi.org/10.1016/j.chroma.2018.03.002>

557 Mie, A., Laursen K.H., Aberg, K.M., Forshed, J., Lindahl, A., Thorup-Kristensen, K., Olsson, M.,
558 Knuthsen, P., Larsen, E.H., & Husted, S. (2014). Discrimination of conventional and organic white
559 cabbage from a long-term field trial study using untargeted LC-MS-based metabolomics. *Analytical
560 and Bioanalytical Chemistry*, 406(12), 2885–2897. <https://doi.org/10.1007/s00216-014-7704-0>

561 Mialon, N., Roig, B., Capodanno, E., & Cadiere, A. (2023). Untargeted metabolomic approaches in
562 food authenticity: a review that showcases biomarkers. *Food Chemistry*, 398, 133856.
563 <https://doi.org/10.1016/j.foodchem.2022.133856>

564 Mihailova, A., Kelly, S.D., Chevallier, O.P., Elliott, C.T., Maestroni, B.M., & Cannavan, A. (2021). High-
565 resolution mass spectrometry-based metabolomics for the discrimination between organic and
566 conventional crops: A review. *Trends in Food Science & Technology*, 110, 142–154.
567 <https://doi.org/10.1016/j.tifs.2021.01.071>

568 Novak, V., Adler, J., Husted, S., Fromberg, A., & Laursen, K. H. (2019). Authenticity testing of
569 organically grown vegetables by stable isotope ratio analysis of oxygen in plant-derived sulphate.
570 *Food Chemistry*, 291, 59–67. <https://doi.org/10.1016/j.foodchem.2019.03.125>

571 Pacifico, D., Casciani, L., Ritota, M., Mandolino, G., Onofri, C., Moschella, A., Parisi, B., Cafiero, C., &
572 Valentini, M. (2013). NMR-Based Metabolomics for Organic Farming Traceability of Early Potatoes.
573 *Journal of Agricultural and Food Chemistry*, 61(46), 11201–11211.
574 <https://doi.org/10.1021/jf402961m>

575 Pedrosa, M.C., Lima, L., Heleno, S., Caroch, M., Ferreira, I.C.F.R., & Barros, L. (2021). Food
576 metabolites as tools for authentication, processing, and nutritive value assessment. *Foods*, 10(9),
577 2213. <https://doi.org/10.3390/foods10092213>

578 Pereira, F.d.O., Pereira, R.d.S., Rosa, L.d.S., & Teodoro, A.J. (2016). Organic and conventional
579 vegetables: comparison of the physical and chemical characteristics and antioxidant activity. *African*
580 *Journal of Biotechnology*, 15(33), 1746-1755. <https://doi.org/10.5897/1JB2016.15386>

581 Riedl, J., Esslinger, S., & Fauhl-Hassek, C. (2015). Review of validation and reporting of non-targeted
582 fingerprinting approaches for food authentication. *Analytica Chimica Acta*, 885, 17–32.
583 <https://doi.org/10.1016/j.aca.2015.06.003>

584 Rippa, S., Zhao, Y., Merlier, F., Charrier, A., & Perrin, Y. (2012). The carnitine biosynthetic pathway in
585 *Arabidopsis thaliana* shares similar features with the pathway of mammals and fungi. *Plant*
586 *Physiology and Biochemistry*, 60, 109–114. <https://dx.doi.org/10.1016/j.plaphy.2012.08.001>

587 Rocchetti, G., Pateiro, M., Campagnol, P.C.B., Barba, F.J., Tomasevic, I., Montesano, D., Lucini, L., &
588 Lorenzo, J.M. (2020). Effect of partial replacement of meat by carrot on physicochemical properties
589 and fatty acid profile of fresh turkey sausages: a chemometric approach. *Journal of the Science of*
590 *Food and Agriculture*, 100(13), 4968-4977. <https://doi.org/10.1002/jsfa.10560>

591 Sciubba, F., Tomassini, A., Giorgi, G., Brasili, E., Pasqua, G., Capuani, G., Aureli, W., & Miccheli, A.
592 (2020). NMR-Based Metabolomic Study of Purple Carrot Optimal Harvest Time for Utilization as a
593 Source of Bioactive Compounds. *Applied Sciences*, 10(23), 8493.
594 <https://doi.org/10.3390/app10238493>

595 Søltoft, M., Nielsen, J., Holst Laursen, K., Husted, S., Halekoh, U., & Knuthsen, P. (2010). Effects of
596 Organic and Conventional Growth Systems on the Content of Flavonoids in Onions and Phenolic
597 Acids in Carrots and Potatoes. *Journal of Agricultural and Food Chemistry*, 58(19), 10323–10329.
598 <https://doi.org/10.1021/jf101091c>

599 Stolarczyk, J., & Janick, J. (2011). Carrot: history and iconography. *Chronia Horticulturae*, 51(2).

600 Tomassini, A., Sciubba, F., Di Cocco, M. E., Capuani, G., Delfini, M., Aureli, W., & Miccheli, A. (2016).
601 ¹H NMR-Based Metabolomics Reveals a Pedoclimatic Metabolic Imprinting in Ready-to-Drink Carrot
602 Juices. *Journal of Agricultural and Food Chemistry*, 64(25), 5284–5291.
603 <https://doi.org/10.1021/acs.jafc.6b01555>

604

605

606 **Figure caption**

607 **Fig. 1:** PCA score plots of PC1 vs. PC2 and PC1 vs. PC3, obtained using the reversed-phase mode
608 analysis for the first batch of samples (red squares: Normandy (No) samples, blue crosses: New
609 Aquitaine (NA) samples, orange circles: Brittany (B) sample, magenta triangles: Hauts-de-France (H)
610 sample, and green filled circles: quality control (QC) samples)

611 **Fig. 2:** PLS-DA and OPLS-DA score plots of LV1 vs. LV2 obtained using the reversed-phase mode
612 analysis for the geographical origin discrimination (blue: New Aquitaine (NA) samples, red: Normandy
613 (No) samples)

614 **Fig. 3:** PLS-DA and OPLS-DA score plots of LV1 vs. LV2 obtained using the reverse phase mode
615 analysis for the production mode discrimination (blue crosses: conventional samples, red squares:
616 organic samples)

617 **Fig. 4:** Experimental MS/MS spectrum of features 2 and 16 for authentication of the geographical
618 origin using the reversed-phase mode analysis, and their corresponding database MS/MS spectrum
619 (arginine and 6-methoxymellein respectively)

620

621