

Investigating the organic and conventional production type of olive oil with target and suspect screening by LC-QTOF-MS, a novel semi-quantification method using chemical similarity and advanced chemometrics

Natasa P. Kalogiouri¹ · Reza Aalizadeh¹ · Nikolaos S. Thomaidis¹

Received: 31 December 2016 / Revised: 27 April 2017 / Accepted: 3 May 2017 / Published online: 25 May 2017
© Springer-Verlag Berlin Heidelberg 2017

Abstract The discrimination of organic and conventional production has been a critical topic of public discussion and constitutes a scientific issue. It remains a challenge to establish a correlation between the agronomical practices and their effects on the composition of olive oils, especially the phenolic composition, since it defines their organoleptic and nutritional value. Thus, a liquid chromatography-electrospray ionization-quadrupole time of flight tandem mass spectrometric method was developed, using target and suspect screening workflows, coupled with advanced chemometrics for the identification of phenolic compounds and the discrimination between organic and conventional extra virgin olive oils. The method was optimized by one-factor design and response surface methodology to derive the optimal conditions of extraction (methanol/water (80:20, v/v), pure methanol, or acetonitrile) and to select the most appropriate internal standard (caffeic acid or syringaldehyde). The results revealed that extraction with methanol/water (80:20, v/v) was the optimum solvent system and syringaldehyde 1.30 mg L⁻¹ was the appropriate internal standard. The proposed method demonstrated low limits of detection in the range of 0.002 (luteolin) to 0.028 (tyrosol) mg kg⁻¹. Then, it was successfully applied in 52 olive oils of *Kolovi* variety. In total, 13 target and 24 suspect phenolic

compounds were identified. Target compounds were quantified with commercially available standards. A novel semi-quantitation strategy, based on chemical similarity, was introduced for the semi-quantification of the identified suspects. Finally, ant colony optimization-random forest model selected luteolin as the only marker responsible for the discrimination, during a 2-year study.

Keywords High - resolution mass spectrometry · Semi-quantification method · Ant colony optimization · Random forest · Luteolin · EVOOs

Introduction

The term “organic food” denotes products that have been produced in accordance with the principles and practices of organic agriculture [1]. Organic farming is a production system which avoids the use of synthetically compounded fertilizers, pesticides, and growth regulators. Organic farming practices are based on the idea that each part of the farm operation augments the other parts to form an efficient and sustainable food production system, offering many advantages, such as minimizing all forms of pollution and producing food of high quality [1]. For these reasons, there is considerable increase in consumer’s demand for organic foods, in a global scale.

In respect to consumer’s needs, many attempts have been made in order to differentiate between organic and conventional products, but the results were controversial [2]. However, new results arising from foodomics and metabolomic studies have detected differences in minor food components, such as polyphenols and other bioactive compounds [3]. The term foodomics describes the discipline that studies the food and nutrition domain through the application of advanced omics technologies to improve consumer’s well-

Published in the topical collection *Advances in LC-MS/MS Analysis* with guest editors Damià Barceló and Mira Petrovic.

Electronic supplementary material The online version of this article (doi:10.1007/s00216-017-0395-6) contains supplementary material, which is available to authorized users.

✉ Nikolaos S. Thomaidis
ntho@chem.uoa.gr

¹ Laboratory of Analytical Chemistry, Department of Chemistry, National and Kapodistrian University of Athens, Panepistimiopolis Zografou, 15771 Athens, Greece

being, health, and confidence [4]. Among other applications, foodomics can explore the effect of the agronomic environment on the metabolite profile of food. In this field, a lot of studies have been conducted [5–12]. Koh et al. [5] compared the chemical composition of organic and conventional spinach using liquid chromatography-electrospray ionization-tandem mass spectrometry (LC-ESI-MS/MS) and analysis of variance (ANOVA), concluding that the content of flavonoids is higher in organic products. These findings reported to affect the carbon/nutrient balance theory and growth rate as well as growth differentiation balance hypothesis, indicating the allocation of plant metabolism toward higher carbon-containing components (like flavonoids) [5]. In addition, Ren et al. [6] as well as Vallverdu-Qurealt et al. [8] observed the same effects in vegetables by means of target screening in LC-MS. However, the studies on this subject for olive oil are scarce in the literature, trying to investigate the correlation between olive oil contents and production types, using metabolomics and chemometrics [8]. Nonetheless, the comparison between environmental factors and harvesting years has been neglected [11]. Anastasopoulos et al. [12] and Rosati et al. [9] measured total phenolic content with Folin Ciocalteu. They observed that the phenolic content was higher in organic production type. The abovementioned studies [9–12] denote that there is an existing connection between the phenolic content and the production type. Thus, the development of high resolution analytical methodologies, with higher identification confidence, that enable the identification of phenolic compounds in such cases is of high interest.

While dealing with analysis of the multi-class of compounds having different polarities (polar compounds such as alcohols and acids and less polar ones, like secoiridoids, lignans, and flavonoids), there is an emerging need to derive the optimum experimental conditions. Chemometric methods have been frequently applied to optimize analytical methods, introducing several advantages such as reduction in the number of experiments, reagent consumption, and less laboratory work [13]. Moreover, these methods can reveal the significance of the factors, their effects, and interaction effects. Factorial design, one factor design (OFD), central composite designs (CCD), and Box-Behnken design are some of the widely used methods for design of experiments [13–15]. These methods often couple to response surface methodology (RSM) to derive the optimal conditions for any property under study [16, 17]. After the optimization of the analytical methodology and the identification of the target and suspect compounds, the semi-quantification of the suspect remains a challenge since in most cases; there are no reference standards commercially available.

Usefulness and reliability of the semi-quantification can be more well-established using the most relevant standard [18]. To such an end, chemical similarity analysis can be applied to rank the standards for semi-quantification purposes. Chemical

similarity has been subject for nearly a decade, trying to find the correct and meaningful similarity assignment between compounds [19–26]. From chemical perspective, similar compounds should have similar functional groups or fragments [24]. The scoring function and a scale that can describe the chemical space edge are vital [24]. Such a score can be easily developed as chemical space is subjected to understand the correlation between a property and chemical descriptors [27]. Application of such chemical descriptors has made a breakthrough in terms of identification of similar chemical structures in a large-scale database [24]. Introduction of chemical fingerprints with a suitable similarity metrics (Euclidean or Tanimoto) [23, 24] could also help to assign an accurate chemical similarity score.

Following the optimization of analytical method and selection of appropriate standards for semi-quantification purposes, a robust model should be developed to discriminate between organic and conventional extra virgin olive oils (EVOOs). Although partial least squares-discriminant analysis (PLS-DA) can score the MS features and select markers [9], the interpretation of results might be complex when the explained variances are too low giving little discriminative power to the PLS-DA model. In such a case, models must be inspected by cross-validation analysis and an external test set to verify that the model is capable of correct class assignment [28]. Moreover, it is of great need to set a threshold for the suggested markers that define the olive oil production types. PLS-DA is not capable of setting such threshold. In order to have a discriminative method applicable and reliable, the threshold derived for each marker should be evaluated throughout the changes in environmental conditions between different harvesting years. Building a correlation between the interaction factors and phenolic content may help to understand whether there are significant differences between organic and conventional olive oils and if olive oils of different harvesting years are comparable or not. This reveals that the markers identified are extremely relevant to build the discriminative models.

The main aim of this study is to develop an optimized reversed-phase ultra-high performance liquid chromatography-electrospray ionization quadrupole time of flight tandem mass spectrometric method (RP-UHPLC-ESI-QTOF-MS), using target and suspect screening workflows combined with advanced chemometrics to reveal the correlation between the phenolic compounds and the production type. The second objective is to identify the markers responsible for the discrimination in a 2-year study. The method was optimized by OFD-RSM to derive the optimal conditions for the extraction of the phenolic compounds, the appropriate internal standard, and its concentration. The method was applied in 52 EVOOs of *Kolovi* variety from Lesvos, both organic and conventional that were harvested during the years 2014–2015 and 2015–2016, for the determination of 13 target phenolic compounds and suspect screening was followed for

the identification of 96 suspect phenolic compounds. The target phenolic compounds were quantified and a novel semi-quantitation strategy is introduced based on chemical similarity analysis. Then, ant colony optimization-random forest (ACO-RF) was employed to investigate alterations between organic and conventional olive oils and introduce one or more markers, suggesting a concentration threshold and discriminate between organic and conventional EVOOs.

Materials and methods

Chemicals and standards

All standards and reagents were of high-purity grade (>95%). Methanol (MeOH) as well as acetonitrile (ACN) of LC-MS grade and sodium hydroxide (>99%) were purchased from Merck (Darmstadt, Germany). Ammonium acetate ($\geq 99.0\%$) for HPLC and formic acid (LC-MS Ultra) were purchased from Fluka (Buchs, Switzerland). Isopropanol was purchased from Fisher Scientific (Geel, Belgium). Distilled water was provided by a Milli-Q purification apparatus (Millipore Direct-Q UV, Bedford, MA, USA). For the analytical method validation, the following reagents were used: syringic acid 95% was purchased from Extrasynthèse (Genay, France), gallic acid 98%, ferulic acid 98%, epicatechin 97%, *p*-coumaric (4-hydroxycinnamic acid) 98%, homovanillic acid 97%, as well as oleuropein 98% and pinoresinol 95% were obtained from Sigma-Aldrich (Steinheim, Germany), and hydroxytyrosol 98% and luteolin 98% were acquired from Santa Cruz Biotechnologies. Vanillin 99%, ethyl vanillin 98%, apigenin (4,5,7-trihydroxyflavone) 97%, and tyrosol (2-(4-hydroxyphenyl) ethanol) 98% were acquired from Alfa Aesar (Karlruhe, Germany). Caffeic acid 99% and syringaldehyde 98% (internal standards) were purchased from Sigma-Aldrich (Steinheim, Germany). Stock standard solutions of individual compounds (1000 mg L^{-1}) were solubilized in MeOH and stored at $-20 \text{ }^\circ\text{C}$ in dark brown glass. All intermediate standard solutions containing the analytes were prepared by dilution of the stock solutions in MeOH.

Olive oil samples

Overall, 52 monovarietal EVOOs were acquired from the Island of Lesvos for a 2-year study. Forty-one EVOOs of *Kolovi* variety were produced from olives cultivated over the harvesting period 2015–2016, consisting of 17 organic and 24 conventional olive oils. Moreover, 11 extra virgin olive oils of the same variety produced during the harvesting period 2014–2015 (two organic and nine conventional) were also included in the current research, as a test set to evaluate the successful applicability of the proposed discrimination models in previous harvesting periods. Figure 1 presents the geographical

distribution of the monovarietal organic and conventional extra virgin olive oils that were produced during the harvesting periods 2014–2015 and 2015–2016. In this figure, all samples that are in italic relate to the harvesting period 2014–2015 and all samples in bold relate to the harvesting period 2015–2016. Moreover, samples labeled as organic are marked with an “asterisk.” More information regarding the harvesting and production details of the EVOOs can be found in the Electronic Supplementary Material (ESM, Table S1). All samples were protected from light and humidity and were preserved as it has already been reported by Kalogiouri et al. [29].

Instrumental analysis

A UHPLC system with an HPG-3400 pump (Dionex UltiMate 3000 RSLC, Thermo Fisher Scientific, Germany) was used for RP analysis, interfaced to a QTOF mass spectrometer (Maxis Impact, Bruker Daltonics, Bremen, Germany), in negative electrospray ionization mode. Separation was carried out using an Acclaim RSLC C18 column ($2.1 \times 100 \text{ mm}$, $2.2 \text{ }\mu\text{m}$) purchased from Thermo Fisher Scientific (Driesch, Germany) with a pre-column of ACQUITY UPLC BEH C18 ($1.7 \text{ }\mu\text{m}$, VanGuard Pre-Column, Waters (Ireland)). Column temperature was set at $30 \text{ }^\circ\text{C}$. The solvents used consisted of (A) 90% H_2O , 10% MeOH, and 5 mM $\text{CH}_3\text{COONH}_4$ and (B) 100% MeOH and 5 mM $\text{CH}_3\text{COONH}_4$. The adopted elution gradient started with 1% of organic phase B with flow rate 0.2 mL min^{-1} during 1 min, gradually increasing to 39% for the next 2 min and then increasing to 99.9% and flow rate 0.4 mL min^{-1} for the following 11 min. These almost pure organic conditions were kept constant for 2 min (flow rate 0.48 mL min^{-1}) and then initial conditions (1% B–99% A) were restored within 0.1 min (flow rate decreased to 0.2 mL min^{-1}) to re-equilibrate the column for the next injection.

The QTOF-MS system was equipped with an electrospray ionization interface (ESI), operating in negative mode with the following settings: capillary voltage of 3500 V, end plate offset of 500 V, nebulizer pressure of 2 bar (N_2), drying gas of 8 L min^{-1} (N_2), and drying temperature of $200 \text{ }^\circ\text{C}$. A QTOF external calibration was daily performed with sodium formate (cluster solution), and a segment (0.1–0.25 min) in every chromatogram was used for internal calibration, using calibrant injection at the beginning of each run. The sodium formate calibration mixture consisted of 10 mM sodium formate in a mixture of H_2O /isopropanol (1:1). Full scan mass spectra were recorded over the range of 50–1000 m/z , with a scan rate of 2 Hz. MS/MS experiments were conducted using AutoMS data-dependent acquisition mode based on the fragmentation of the five most abundant precursor ions per scan. The instrument provided a typical resolving power (FWHM) between 36,000 and 40,000 at m/z 226.1593, 430.9137, and 702.8636.

Fig. 1 Geographical distribution of EVOOs selected from Lesvos Island



Screening methodology

Target and suspect screening methodologies were followed, as it has already been described by our group [29]. The identification workflow incorporated strict filtering steps, interpretation of MS/MS spectra, and retention time prediction. A target list was created including 13 phenolic compounds, including phenolic acids, secoiridoids, flavonoids, and lignans (gallic acid, *p*-coumaric acid, ferulic acid, syringic acid, homovanillic acid, tyrosol, hydroxytyrosol, pinoresinol, apigenin, oleuropein, vanillin, ethyl vanillin and epicatechin) that have already been identified in extra virgin olive oils of *Kolovi* variety in our previous study [29]. A suspect list of 96 bioactive constituents was generated from literature including all the bioactive constituents and mainly the phenolic compounds that have been identified in olive oils, drupes, and leaves. The initial suspect list is presented in the ESM (Table S2).

The software packages Target Analysis 1.3 and Data Analysis 4.1 (Bruker Daltonics, Bremen, Germany) along with the tools of these packages Bruker Compass Isotope Pattern and SmartFormula Manually were in the target screening workflow. Extracted ion chromatograms (EICs) were obtained using the function Find Compounds-Chromatogram in Target Analysis Software. Mass accuracy was set at 2 mDa, mSigma was below or equal to 50, signal-to-noise threshold of 3, minimum area threshold of 800, and minimum intensity threshold of 200. The relative tolerance of the retention time window was set lower than ± 0.2 min. The target compounds were identified on the basis of mass accuracy, isotope pattern, retention time (t_R), and MS/MS fragments [29].

In suspect screening, the EICs were created using Target Analysis Software 1.3 and the following parameters were set: mass accuracy threshold of 2 mDa, isotopic fit below or equal

to 50, ion intensity of more than 800, peak area threshold of 2000, and peak score (area/intensity ratio) between 4 and 38 [29]. The EICs were studied using Data Analysis 4.1 software to confirm that the peak represents the suspect compound. The MS/MS fragments were compared and interpreted with the use of Metfrag [30] and FooDB [31]. The retention time of each suspect compound was predicted and compared with the experimental retention time with the use of quantitative structure-retention relationship model (QSRR) [32].

As for the level of confidence achieved in the identification of the suspect compounds, compounds are identified at level 1 when the structures are confirmed with available reference standards. In the cases that there are no standards commercially available, level 2 corresponds to probable structures (level 2a, MS/MS fragments were verified with spectral libraries or literature; level 2b, diagnostic evidence where no other structure fits the experimental MS/MS information) and level 3 corresponds to tentative candidates [33].

Optimization of experimental conditions

The initial design consisted of three main factors (one numeric and two categorical variables) within one block. The design model was selected quadratic to cover the multilevel limits for parameters intended to be optimized. Extraction (which was a categorical factor) was set at three levels (MeOH, MeOH/H₂O (80:20, v/v), and acetonitrile). The second factor was the internal standard (caffeic acid and syringaldehyde) and the final factor was the concentration of the internal standard, set within the range of 0.5 up to 1.5 mg L⁻¹. In the case of quadric model, five levels (0.5, 0.75, 1.00, 1.25, and 1.5 mg L⁻¹) for one numeric factor (concentration) are required with some replicate points. This design could be duplicated for every combination of

categorical factor levels. The optimization task was performed to minimize the relative standard deviation (%RSD) values of the peak areas of each spiked standard. The combination of all these factors required a set of 42 experiments. These experimental plans based on OFD method coupled to RSM [15], along with the %RSD value for each spiked standard as response ($n = 3$), are presented in the ESM (Table S3). The design of experiments and all statistical assessments were calculated by Design-Expert software version 7 [34].

Method validation

The optimized RP-UHPLC-ESI-MS method was validated to ensure that it is suitable for identification and quantification purposes. Standard addition curves were constructed for all the analytes. All the compounds were spiked in real EVOO samples. Gallic acid, *p*-coumaric acid, ferulic acid, syringic acid, homovanillic acid, pinoresinol, apigenin, vanillin, ethyl vanillin, epicatechin, and luteolin were spiked at concentrations between 0.02 and 10 mg kg⁻¹ (14 calibration levels with 3 replicates at each level). Tyrosol, hydroxytyrosol, and oleuropein calibration curves were constructed over the range of 0.02–100 mg kg⁻¹ (20 calibration levels with 3 replicates at each level). Calibration curves were constructed with the use of the peak area of the spiked analyte subtracted by the peak area of a neat sample and divided by the peak area of the internal standard. Limits of detection (LODs) and limits of quantification (LOQs) were calculated at the lowest concentration range of the analytes (0.02–1 mg kg⁻¹), by the equations:

$LOD = \frac{3.3 \times S_a}{b}$ and $LOQ = \frac{10 \times S_a}{b}$ where S_a is the standard error of the intercept a and b is the slope of the calibration curve. The accuracy of the method was estimated using recoveries, at 2 mg kg⁻¹ concentration level, calculated as follows:

$$\%RE = \frac{\text{Response}_{\text{extracted sample}}}{\text{Response}_{\text{postextracted spiked sample}}} \times 100 \quad (1)$$

where $\text{Response}_{\text{extracted sample}}$ is the average area of the analyte in matrix, which has been through the extraction process, from three replicates, divided each time by the peak area of the internal standard. $\text{Response}_{\text{post extracted spiked sample}}$ is the average area of each analyte, spiked into extracted matrix after the extraction procedure. To evaluate the matrix effect, the matrix factor was calculated at 2 mg kg⁻¹ concentration level according to the following equation:

$$MF = \frac{\text{Response}_{\text{postextracted sample}}}{\text{Response}_{\text{standard solution}}} \quad (2)$$

where $\text{Response}_{\text{post extracted sample}}$ is the average area of the analyte, spiked into the extracted matrix after the extraction procedure, and $\text{Response}_{\text{standard solution}}$ is the average area

count for the same concentration of analyte in a standard solution. For the calculation of ME, 1 was subtracted by the quotient (2) and multiplied by 100, so that the negative result indicates suppression and the positive result indicates enhancement of the analyte signal. The precision of the method was demonstrated in terms of repeatability (intraday precision) and intralaboratory reproducibility (interday precision). Repeatability was expressed as the %RSD_r values of six replicate analyses ($n = 6$) in the same day. Reproducibility experiments were expressed as the %RSD_R value of 3 replicates of three consecutive days ($n \times k = 3 \times 3 = 9$). Finally, lack-of-fit F test was applied to ensure that the calibration curves can be used for quantification purposes. For this scope, all three replicates of each concentration level were used and the number of data points (concentration levels) was 20 for oleuropein, tyrosol, and hydroxytyrosol and 14 for the rest of the analytes.

Chemical similarity analysis

Three standards including tyrosol, hydroxytyrosol, and oleuropein were used as a main scheme for semi-quantification to define the chemical space boundaries (chemical space edge) and their similarity distance from 14 secoiridoids (10-hydroxy-10 methyl oleuropein aglycone, methyl oleuropein aglycone, 10-hydroxy oleuropein aglycone, oleoside, oleuropein aglycone, oleomissional, lingstroside aglycone, oleokoronal, 10-hydroxy decarboxymethyl oleuropein aglycone, decarboxymethyl oleuropein aglycone, decarboxymethyl lingstroside aglycone, hydroxylated form of elenolic acid, elenolic acid, and hydroxytyrosol acetate). All structures of chemicals used here were drawn and their geometries were constructed by searching between conformers with lowest energy using Balloon [35]. The chemical similarity matrix for these compounds was then built based on the molecular descriptors. These molecular descriptors consisted of logD (at pH = 6.2) (measure of hydrophobicity for ionizable compounds), constitutional descriptors, topological descriptors, walk and path counts, connectivity indices, information indices, 2D autocorrelation, edge adjacency indices, burden eigenvalues, topological charge indices, eigenvalue-based indices, Randic molecular profiles, geometrical descriptors, radial distribution function descriptors (RDF), 3D molecular representation of structure based on electron diffraction descriptors (3D-MoRSE), weighted holistic invariant molecular descriptors (WHIM), geometry, topology, and atoms-weighted assembly (GETAWAY) descriptors, functional group counts, atom-centered fragments, charge descriptors, and molecular properties [36–39].

Concerning the above descriptors, they encode the atomic or molecular properties, overall molecular connectivity, molecular geometry, and their size and shape [40]. These descriptors were calculated by E-dragon [41, 42]. LogD was

calculated using the ChemAxon package [43] (the calculated molecular descriptors can be found in the ESM, Table S4). Afterwards, the calculated molecular descriptors were pre-treated in order to remove the constant and near constant descriptors. Molecular descriptors with intercorrelation above 0.95 were also removed using variable reduction method adapted from space-filling designs (V-WSP) as an unsupervised variable reduction method [44] (the survived molecular descriptors can be found in the ESM, Table S5). Euclidean-based similarity metric was used to measure the chemical similarity between the compounds. To define the chemical space, tyrosol, hydroxytyrosol, and oleuropein were used as a main scheme and the other 14 compounds were measured against these three standards. The chemical space edge was also achieved by normalizing the mean distance score for the three standards (these values range from 0 to 1 where 0.0 is least diverse and 1.0 is the most diverse compound). Then, the normalized mean distance scores for the rest of compounds were calculated, and those test compounds, which were scored outside of 0.0 to 1.0 range, were defined to be outside of the chemical space edge. Therefore, this method could define the most appropriate standard for semi-quantification. Similarity analysis and V-WSP calculation were done in MATLAB 8.5 (MathWorks) program.

Prioritizing MS features and modeling strategies

Overall, a matrix containing quantified and semi-quantified results (expressed in mg kg^{-1}) for 30 compounds was generated for 52 extra virgin olive oil samples. These samples were split into a training and a test set based on their harvesting year (to evaluate whether the discrimination achieved is applicable to previous years or not) to build the discrimination models and then evaluate the accuracy of the discrimination model for the external set of samples. ACO [27, 45] was used to prioritize compounds and rank them by their importance and contribution in increasing the accuracy of discrimination model. Details about the ACO and its internal parameters can be found in the ESM (Section S1). The fitness function (a measure of error for the discrimination model) was set based on the error of miss-discrimination in cross-validation leave-one-out analysis. ACO was then coupled to discrimination modeling techniques to evaluate the internal and external accuracy of models every time by inclusion of new features. Using feature selection coupled with discrimination model such as Linear Discriminative Analysis (LDA) [46–48] or RF [49] can prevent over-fitting issues and can introduce more accuracy to a discrimination problem. In RF, variables and their contributions can be ranked based on a measure of variable importance and the modeling can be followed based on the highly important predictors [50]. Therefore, the introduction of a features prioritizing method (ACO) might not be so important. More details about RF can be found in the ESM (Section S2). The

following fitness function was used to measure the error of discrimination in leave one out cross-validation analysis and to decrease it using ACO:

$$F = \sum \frac{\text{Class} \sim \text{Pred. Class}}{n} \quad (3)$$

where F is the objective function (discrimination error measure), Class is the observed group for a case (here is each sample), Pred. Class is the predicted group by the modeling technique, and n is the number of samples used to build the discrimination model. The entire data processing step was done in a homemade program, called ChemoTrAMS, written in MATLAB environment.

Validation procedure of the models

The initial parameter used to evaluate the internal accuracy of the models was the error rate of miss-discrimination in training set and cross-validation analysis. Leaving-one-out, cross-validation was also performed during the training step to understand the error rate by excluding a certain sample from the rest of the training set. The predictive power of the proposed discrimination model was evaluated independently using a set of external samples that were not part of the initial training set and confusion matrix was calculated to derive error rate, class specificity, and sensitivity [28]. Moreover, Receiver Operating Characteristics (ROC) was calculated to check the discrimination capability of the models. ROC curves were calculated for each class by plotting the sensitivity versus 1-specificity for a binary case study (organic or conventional). A perfect discrimination model would yield a point in the upper left corner of the ROC area, representing maximum sensitivity and specificity, while a random discrimination causes points to be along the diagonal line from the left bottom to the top right corner [28].

Results and discussion

Optimization of the method

The evaluation of the best extraction conditions and the selection of the appropriate internal standard took place using OFD. The goal of OFD-RSM was to optimize these three factors at a point which low %RSD values of the peak areas of the spiked standard compounds would be achieved. It was found that ACN has the lowest desirability [51] for all compounds under study and it shows high error (high %RSD). The interaction map and effect of extractions are shown in Fig. 2a, b. As it can be derived from Fig. 2a, b, MeOH/H₂O (80:20, v/v) has the highest desirability (lowest %RSD) among other

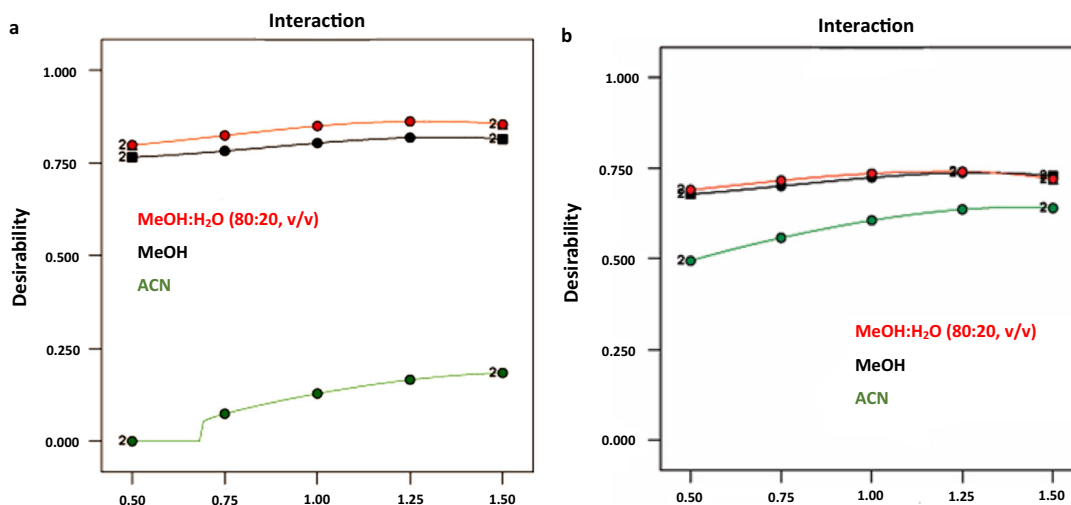


Fig. 2 Desirability of different extractions while using (a) syringaldehyde and (b) caffeic acid as internal standard

extractions. Since the desirability observed for MeOH and MeOH/H₂O (80:20, v/v) is close, the interaction maps were investigated.

Table 1 summarizes the results of the desirability plots for MeOH and MeOH/H₂O (80:20, v/v) using syringaldehyde as an internal standard as well as for MeOH/H₂O (80:20, v/v) using caffeic acid as internal standard, based on the set of 42 experiments (ESM, Table S3). For most of the spiked standard compounds, MeOH/H₂O (80:20, v/v) presented the lowest %RSD values and was selected as the optimum extractor. Comparing MeOH with MeOH/H₂O (80:20, v/v) when syringaldehyde was the as internal standard, both extraction solvents demonstrate close %RSD values for all the spiked standards, except for syringic acid which has highest %RSD while using solely MeOH. The good performance of MeOH/H₂O (80:20, v/v) is clearly demonstrated in Table 1 showing higher desirability values (especially for syringic acid) compared to pure MeOH.

Moreover, the comparison between the desirability plots where MeOH/H₂O (80:20, v/v) is the extracting solvent and the internal standard used is syringaldehyde in Fig. 3a and caffeic acid in Fig. 3b reveals that all the spiked standard compounds of the phenolic acid class presented higher desirability in the case that caffeic acid was used as an internal standard.

In a further step, OFD-RSM was applied in order to derive the optimum conditions and select the appropriate internal standards at the optimal concentration level, by providing prediction results of the %RSD values of the peak areas for the 14 spiked standard compounds ($n = 3$). It generated predicted %RSD values in the case that MeOH/H₂O (80:20, v/v) was the extractor and compared the desirability of both internal standards; syringaldehyde and caffeic acid at 1.30 and

1.20 mg L⁻¹, respectively (ESM, Table S6). The predicted results revealed that the optimum conditions are derived when syringaldehyde is the internal standard at 1.30 mg L⁻¹. The experimental factors suggested by OFD-RSM were applied, and the experimental %RSD values were in accordance with the predicted.

Moreover, the recoveries (RE) along with standard deviation (\pm SD) were calculated for all the spiked standard compounds in the three different extraction solvents (MeOH, MeOH/H₂O (80:20, v/v), ACN) with syringaldehyde at 1.30 mg L⁻¹ in order to further investigate the adequacy of MeOH/H₂O (80:20, v/v). The results are listed in Table 2. MeOH/H₂O is a better extracting media than pure MeOH, and syringaldehyde at 1.30 mg L⁻¹ presents higher desirability compared to caffeic acid 1.20 mg L⁻¹. Figure 4 illustrates these optimal experimental conditions.

These optimal conditions were implemented and a liquid-liquid microextraction (LLME) method was developed and validated in order to isolate all the phenolic compounds from the olive oil samples. For this, 0.5 g of each sample was weighted and spiked with 1.30 mg L⁻¹ syringaldehyde and in a further step 0.5 mL of MeOH/H₂O (80:20, v/v) was added to 2-mL Eppendorf tubes. Then, the mixture was vortexed for 2 min and centrifuged for 5 min at 13,400 rpm. Additionally, the upper phase was collected and filtered through membrane syringe filters of regenerated cellulose (CHROMAFIL® RC) (15-mm diameter, 0.22- μ m pore size, purchased by Macherey-Nagel, Düren, Germany). Finally, 5 μ L of this solution was injected into the chromatographic system. Procedural blanks were prepared and processed in the chromatographic system to detect any potential contamination. Quality control samples were prepared to confirm that the analytical system has been stabilized before the batch of

Table 1 Desirability values of all the spiked standard compounds

Extracting solvent Internal standard	MeOH Syringaldehyde Desirability values	MeOH/H ₂ O (80:20, v/v) Syringaldehyde	MeOH/H ₂ O (80:20, v/v) Caffeic acid
Vanillin	0.811019	0.893997	0.737738
Apigenin	0.849736	0.912608	0.377854
Epicatechin	0.926173	0.778114	0.541533
Ethyl vanillin	0.972953	0.962772	0.433490
Ferulic acid	0.899410	0.867664	0.926501
Gallic acid	0.769335	0.810612	1
Homovanillic acid	0.704925	0.770156	0.986127
Hydroxytyrosol	0.954376	1	0.925576
<i>p</i> -Coumaric acid	0.793496	0.833178	0.988094
Syringic acid	0.484745	0.701904	0.975974
Tyrosol	0.910728	0.933066	0.785103
Pinoselinol	0.895791	0.849412	0.632015
Oleuropein	0.825102	0.938799	0.758019
Combined	0.819218	0.861535	0.740338

Fig. 3 Desirability plots for MeOH/H₂O (80:20, v/v) using (a) syringaldehyde and (b) caffeic acid as internal standards

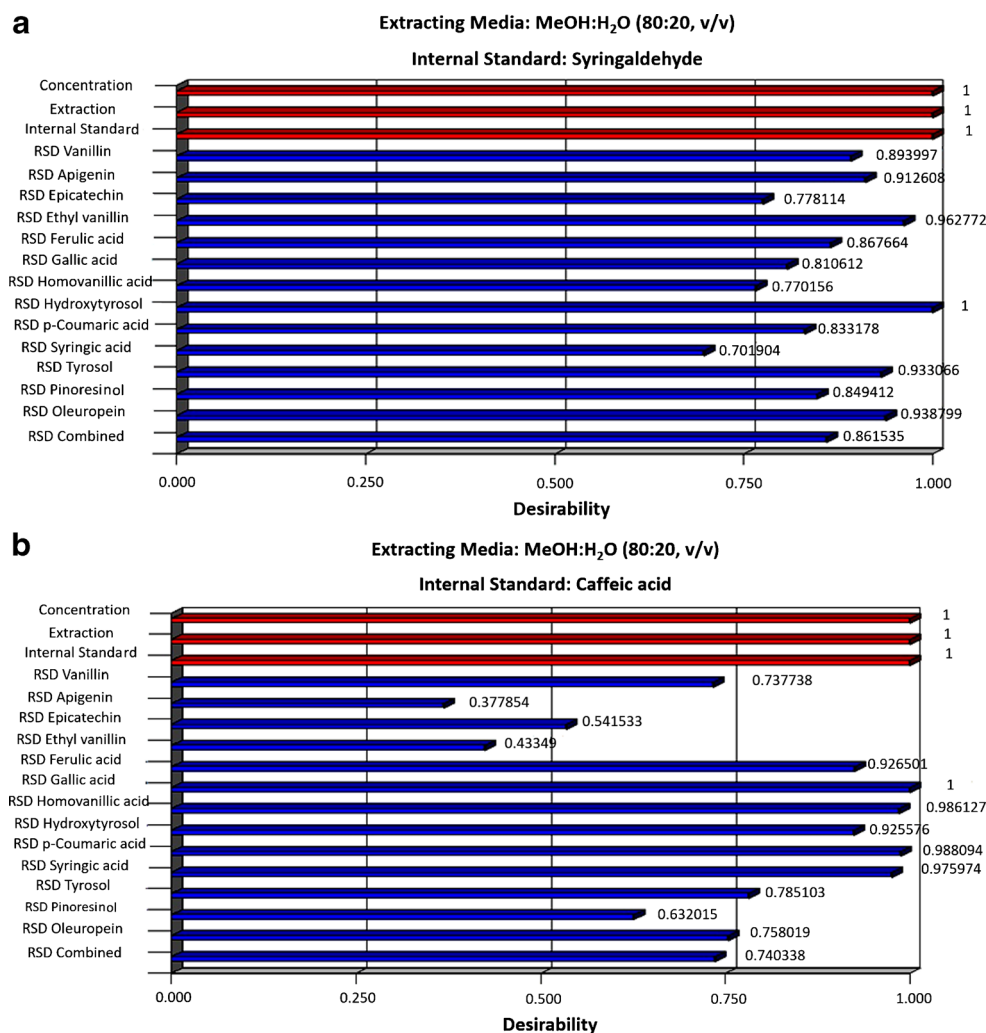


Table 2 Calculated recoveries (\pm standard deviation, $n = 3$) for all the spiked standard compounds in different extraction solvents (MeOH, MeOH/H₂O (80:20, v/v) ACN) and syringaldehyde at 1.30 mg L⁻¹ as an internal standard

Standard compound	MeOH	MeOH/H ₂ O (80:20, v/v)	ACN
Vanillin	93.2 \pm 0.30	95.7 \pm 0.25	92.9 \pm 0.39
Apigenin	94.1 \pm 1.40	93.5 \pm 0.24	91.7 \pm 0.43
Epicatechin	94.9 \pm 0.30	94.5 \pm 0.07	89.2 \pm 0.08
Ethyl vanillin	93.3 \pm 0.21	92.4 \pm 0.48	82.7 \pm 0.47
Ferulic acid	95.1 \pm 0.14	93.4 \pm 0.36	90.0 \pm 0.42
Gallic acid	92.2 \pm 0.14	96.1 \pm 0.11	84.6 \pm 0.42
Homovanillic acid	91.0 \pm 0.66	95.5 \pm 0.40	87.5 \pm 0.18
Hydroxytyrosol	95.5 \pm 0.10	95.5 \pm 0.34	84.2 \pm 0.33
<i>p</i> -Coumaric acid	91.4 \pm 0.16	95.6 \pm 0.28	95.8 \pm 0.48
Syringic acid	94.1 \pm 0.18	92.4 \pm 0.23	93.4 \pm 0.14
Tyrosol	95.2 \pm 0.41	96.5 \pm 0.28	89.1 \pm 0.21
Pinosresinol	92.9 \pm 0.12	94.3 \pm 0.23	86.2 \pm 0.33
Oleuropein	98.1 \pm 0.18	98.9 \pm 0.15	92.2 \pm 0.26

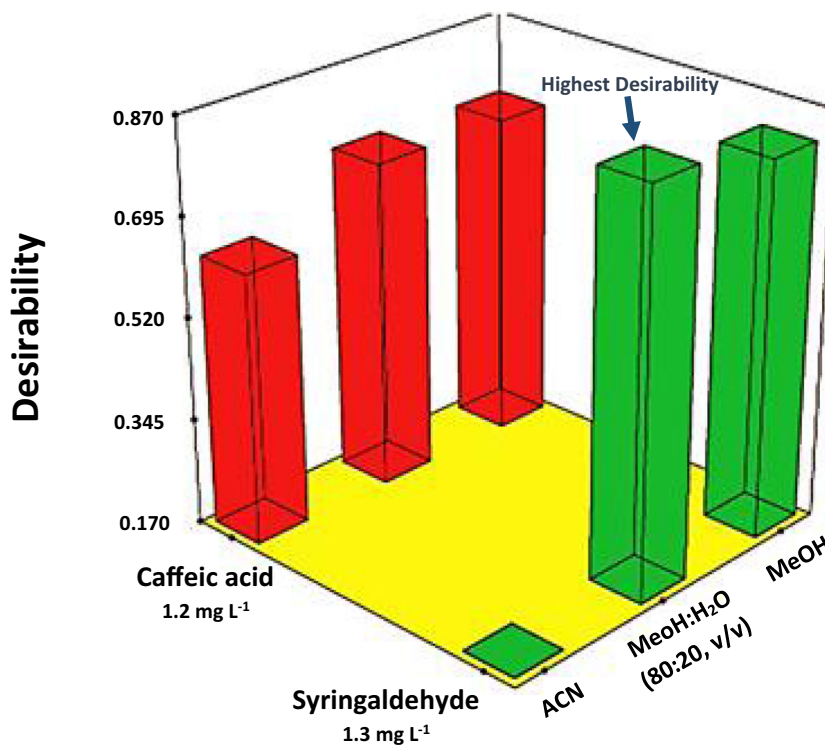
samples and to evaluate its performance. The quality control sample was prepared by mixing all aliquots of the samples. Then, it was spiked with 50 μ L of a standard solution mix (including all the target compounds: vanillin, apigenin, epicatechin, ethyl vanillin, ferulic acid, gallic acid, homovanillic acid, hydroxytyrosol, oleuropein, *p*-coumaric acid, pinosresinol, syringic acid, tyrosol, and luteolin, at a final concentration of 1 mg L⁻¹). It was injected at the beginning of the analysis (five times for conditioning), and afterward, it was injected at regular intervals (every ten sample injections). The calculated %RSDs for the retention time (t_R) and the peak areas as well as mass errors (Δm) are presented in the ESM

(Table S7) demonstrating the good performance of the analytical system ($n = 10$).

Target screening results

After the optimization of the experimental conditions, a data-dependent method was used to scan the presence of the target compounds in real olive oil samples. All the target phenolic compounds that belonged in the initial target list such as gallic acid, *p*-coumaric acid, ferulic acid, syringic acid, homovanillic acid, tyrosol, hydroxytyrosol, pinosresinol, apigenin, oleuropein, vanillin, ethyl vanillin, and epicatechin were

Fig. 4 Derived optimal experimental conditions



determined. The mass accuracies of the precursor ions and the qualifier ions of the detected compounds were less than 2 mDa compared with the standard solutions and the isotopic fit was calculated less than 50 mSigma in all cases. Moreover, the retention time shift was less than 0.05 min for all the detected target compounds. The most abundant fragments provided by the AutoMS spectra were verified with MS/MS records of a previous study by our group [29]. Target screening results are summarized in the ESM (Table S8).

All validation parameters including LODs and LOQs, calculated recoveries, regression equations, regression coefficient (r^2), the lack-of-fit test, method precision expressed as intraday and interday precision, as well as the matrix factor and matrix effect are summarized in Table 3.

The analytes presented satisfying recovery efficiency (92–99%). The precision limit ranged between 0.7 and 2.2% for intraday experiments and between 1.4 and 5.4% for interday experiments, demonstrating the good precision of the optimized method. The method demonstrated low LODs over the range of 0.002 mg kg⁻¹ (luteolin) and 0.028 (tyrosol) and adequate LOQs over the range of 0.007 mg kg⁻¹ (luteolin) and 0.086 mg kg⁻¹ (tyrosol). The analytical curves presented an adequate fit when

submitted to the lack-of-fit test ($F_{\text{calculated}}$ was less than $F_{\text{tabulated}}$ in all cases), with r^2 above 0.99, proving that they can be used for the quantification of the phenolic compounds. The matrix factor ranged between 0.92 and 0.96 and low matrix suppression was observed for all the analytes, up to 7.75%.

In a further step, the 13 target compounds detected were quantified in all EVOO samples on the basis of their reference standards, using syringaldehyde as the internal standard. Quantitative results for the target compounds were expressed as milligrams per kilogram and can be found in the ESM (Table S9).

Suspect screening

In suspect screening, 24 phenolic compounds were tentatively identified in real olive oil samples of *Kolovi* variety with ion intensities above 800 and peak areas of more than 2000, in all cases. The results presented high mass accuracy (less than 2 mDa) and acceptable isotopic fit values (less than 50 mSigma). The peak score (peak area/peak intensity ratio) ranged between 10 and 22 for all the suspect compounds. MS/MS spectra were examined with Metfrag [30], FooDB

Table 3 Validation results

Compound	LOD (mg kg ⁻¹)	LOQ (mg kg ⁻¹)	Equation	r^2	Lack of fit		Intraday precision RSD _r , (%) (n = 6)	Interday precision RSD _R , (%) (n = 3 × 3)	Recovery (%)	Matrix factor	Matrix effect (%)
					F_{calc}	F_{tab}					
Gallic acid	0.005	0.014	$y = (-0.79 \pm 0.68) + (7.73 \pm 0.16)x$	0.995	1.510	2.118	2.1	5.1	99	0.94	-5.89
<i>p</i> -Coumaric acid	0.024	0.073	$y = (0.06 \pm 0.05) + (0.51 \pm 0.01)x$	0.994	1.772	2.118	1.9	4.1	97	0.94	-6.13
Ferulic acid	0.011	0.032	$y = (0.008 \pm 0.017) + (0.687 \pm 0.004)x$	0.999	0.056	2.118	2.2	3.5	95	0.96	-4.31
Syringic acid	0.003	0.008	$y = (0.06 \pm 0.04) + (0.48 \pm 0.01)x$	0.995	1.740	2.118	1.8	3.6	98	0.95	-4.90
Homovanillic acid	0.017	0.051	$y = (0.02 \pm 0.09) + (1.49 \pm 0.02)x$	0.998	0.726	2.118	1.9	5.4	98	0.93	-7.27
Tyrosol	0.028	0.086	$y = (-4.34 \pm 2.83) + (4.99 \pm 0.08)x$	0.996	1.493	1.868	1.3	2.6	97	0.92	-7.78
Hydroxytyrosol	0.010	0.031	$y = (-0.43 \pm 1.07) + (3.77 \pm 0.03)x$	0.999	0.420	1.868	1.8	2.8	97	0.94	-5.80
Pinoselinol	0.023	0.068	$y = (0.01 \pm 0.02) + (0.297 \pm 0.005)x$	0.996	0.405	2.118	1.5	2.9	99	0.93	-7.12
Apigenin	0.024	0.074	$y = (0.08 \pm 1.28) + (12.81 \pm 0.305)x$	0.993	1.910	2.118	2.3	3.1	92	0.92	-7.60
Oleuropein	0.003	0.008	$y = (0.91 \pm 1.84) + (4.58 \pm 0.05)x$	0.998	1.762	1.868	0.7	1.4	98	0.96	-4.16
Vanillin	0.018	0.054	$y = (-0.46 \pm 0.36) + (4.09 \pm 0.09)x$	0.995	0.176	2.118	2.2	3.1	95	0.92	-7.59
Ethyl vanillin	0.013	0.039	$y = (0.51 \pm 0.23) + (2.04 \pm 0.06)x$	0.991	1.940	2.118	1.6	2.8	95	0.94	-6.30
Epicatechin	0.019	0.059	$y = (-0.18 \pm 0.21) + (3.67 \pm 0.05)x$	0.998	0.330	2.118	1.8	3.2	94	0.96	-4.47
Luteolin	0.002	0.007	$y = (0.53 \pm 0.40) + (3.64 \pm 0.10)x$	0.992	1.793	2.118	2.1	3.4	97	0.94	-6.34

LOD limit of detection, LOQ limit of quantification, r^2 regression coefficient, F_{tab} $F_{\text{tabulated}}$, F_{calc} $F_{\text{calculated}}$, RSD relative standard deviation

[31], and literature records. The lists of the fragments of all the identified phenolic compounds were compared and verified with those reported in a previous study of our group [29]. Table S10 in the ESM summarizes the suspect screening results, providing information about the identification criteria and the level of identification of each compound. The derivatives of oleuropein aglycone, oleuropein aglycone monoaldehydic form, oleuropein aglycone dialdehydic form, as well as the enol form of oleuropein aglycone, known as oleomissional [52], were identified at level 3. The qualifier ions of oleuropein aglycone monoaldehydic form ($t_R = 7.43$ min) were detected at m/z 69.0345, 99.0088, 121.0294, 127.0400, 135.0453, 151.0401, and 163.0400 correspond to C_4H_5O , $C_4H_3O_3$, $C_7H_5O_2$, $C_6H_7O_3$, $C_8H_7O_2$, $C_8H_7O_3$, and $C_9H_7O_3$, respectively. The MS/MS spectrum of the dialdehydic form of oleuropein aglycone ($t_R = 7.61$ min) shows peaks at m/z 59.0139, 67.0187, 95.0138, 123.0453, 128.0478, 153.0558, and 195.0662 that correspond to $C_2H_3O_2$, C_4H_3O , $C_5H_3O_2$, $C_7H_7O_2$, $C_6H_8O_3$, $C_8H_9O_3$, and $C_{10}H_{11}O_4$. Oleomissional elutes at 7.75 min and shows two qualifier ions at m/z 101.0245 and 163.0400, corresponding to $C_4H_5O_3$ and $C_9H_7O_3$, respectively. The Extracted Ion Chromatogram (EIC) at m/z 361.1291 presented four different peaks. It has been suggested in our previous study that lingstroside aglycone eluted at 6.63 min [29]. In the present study, it eluted at 6.65 min, the MS/MS spectra were compared and verified with previously reported fragments [29]. It is possible that the other three peaks with retention times (t_R) 7.81, 8.13, and 8.34 belong to lingstroside aglycone monoaldehydic form, the dialdehydic form of oleuropein aglycone, and the enol form of lingstroside aglycone, named oleokoronal [52], respectively. These three isomers of lingstroside aglycone were identified at level 3. The MS/MS spectrum of lingstroside aglycone monoaldehydic form shows two qualifier ions at m/z 137.0608 and 241.0718, corresponding to $C_8H_9O_2$ and $C_{11}H_{13}O_6$, respectively. Next, oleokoronal presents two characteristic fragments at m/z 195.0663 and 291.0874, corresponding to $C_{10}H_{11}O_4$ and $C_{15}H_{15}O_6$, respectively. In the MS/MS spectrum of the dialdehydic form of oleuropein aglycone, the fragments at m/z 69.0346, 101.0244, and 259.0976 correspond to C_4H_5O , $C_4H_5O_3$, and $C_{15}H_{15}O_4$, respectively.

The QSRR model [32] was used for the prediction of the possible retention time of oleuropein aglycone and lingstroside isomers, since there were no reference standards available. The difference of the experimental retention time and the predicted was less than 1 min for all the suspect isomers which were inside the applicability domain of the model. More information about the QSRR model can be found in the ESM (Section S3).

In a further step, all the suspect compounds were semi-quantified. The lignans syringaresinol, 1-hydroxypinoresinol, and the isomer of 1-hydroxypinoresinol as well as 1-acetoxypinoresinol were semi-quantified with the use of

pinoresinol calibration curve. The suspect compounds which belong to the class of secoiridoids were semi-quantified on the basis of target compounds having similar structure (oleuropein, tyrosol or hydroxytyrosol), after measuring similarity with chemometric tools, as it is described in the following section “Semi-quantification and similarity measurement.”

Semi-quantification and similarity measurement

Similarity indices were performed over 14 compounds so that they could be semi-quantified with the most appropriate standard (Fig. 5; see also ESM, Fig. S2). It was found that oleuropein is the most appropriate standard to be used to semi-quantify 10-hydroxy oleuropein aglycone, oleuropein aglycone, lingstroside aglycone, methyl oleuropein aglycone, 10-hydroxy-10-methyl oleuropein aglycone, oleomissional, and oleoside. Moreover, hydroxylated form of elenolic acid, 10-hydroxydecarboxymethyl oleuropein aglycone, decarboxymethyl oleuropein aglycone, decarboxymethyl lingstroside aglycone, elenolic acid, and hydroxytyrosol acetate can be semi-quantified with both tyrosol and hydroxytyrosol. However, the degree of similarity indices is closer to tyrosol. Oleokoronal can also be quantified based on tyrosol as its similarity indices is in the middle of oleuropein and tyrosol.

The suspect secoiridoids were semi-quantified, as suggested above, and the semi-quantification concentrations of all the suspect compounds are presented in the ESM (Table S11; the concentrations are expressed in $mg\ kg^{-1}$).

ACO-LDA

The discrimination between organic and conventional EVOOs was based on the quantification and semi-quantification results of the target and suspect compounds, respectively. LDA tries to separate the samples by increasing the variance between the classes and decreasing the variance within class. For an optimal discrimination, it is essential to know the class posterior probability. The probability of each sample belonging to the corresponding classes along with the predicted classes (organic and conventional production type of olive oil) is given in the ESM, in Table S12. The LDA model was built only based on luteolin as ACO selected it as the most important feature, causing discrimination between classes. The validation criteria were met for ACO-LDA and the miss-discrimination error for training, cross-validation analysis, as well as external samples was zero.

ACO-RF/RF

Using ACO-LDA selects the appropriate compound to discriminate between two classes, but it cannot justify at which threshold this discrimination is achieved. Therefore, it is

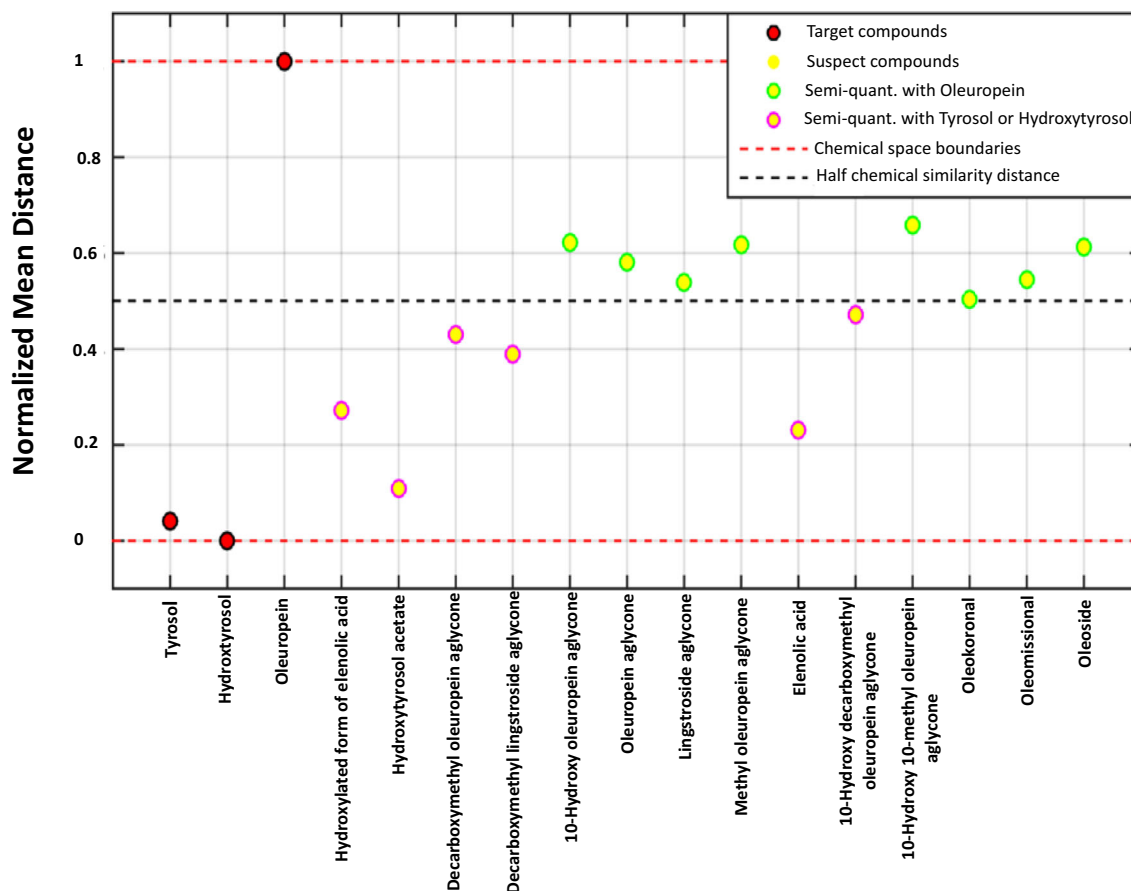


Fig. 5 Similarity indices calculated between compounds to be semi-quantified

needed to create a tree with defined threshold. After using ACO as a variable selection tool to select the most appropriate compounds (independent variable), luteolin was selected due to the higher contribution to the discrimination problem in RF. Here, using ACO coupled with RF did not affect the outcome, comparing to using RF alone. This was expected since RF has a capability to neglect inclusion of extra variables if the miss-discrimination rate achieved to a minimum value with a single variable. RF generated a simple tree to justify how production type (organic and conventional) of olive oil can be predicted, using the concentration (mg kg^{-1}) of luteolin in a sample. An EVOO is organic if the concentration (mg kg^{-1}) of luteolin is more than 4.16 mg kg^{-1} ; otherwise, the EVOO is conventional (Fig. 6).

The validation results of the proposed decision tree suggested that the ACO-RF/RF model shows both internal and external accuracy. The miss-discrimination rate was obtained zero for both training and the external samples. Leave-one-out cross-validation analysis was also indicated that there is not any sample that its removal could affect the outcome of model substantially (the miss-discrimination error was zero). ROC curves were also calculated for both classes, and the results indicated that the discrimination model was well-established with specificity and sensitivity equals to 1.

A well-established discrimination model should show sensitivity toward the changes in the environmental conditions for a 2-year study. This could be studied by setting EVOOs produced in different years into different evaluation sets. Here, we applied the discrimination models trained on the olive oils produced during the harvesting period of 2015–2016 and tested with those that were produced during the harvesting period 2014–2015. The results show that the marker which is responsible for the discrimination between organic and conventional

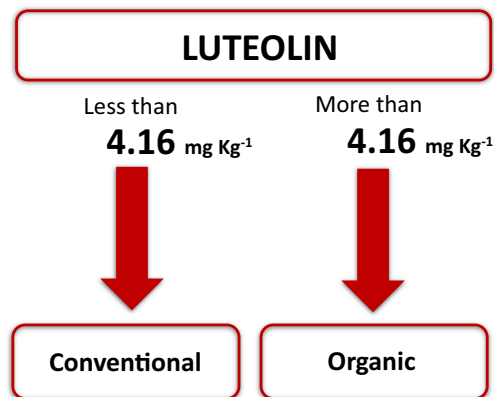


Fig. 6 Discrimination results of organic and conventional EVOOs using ACO-RF/RF

EVOOs is the same for those that were produced in the previous harvesting period as well.

Luteolin is a predominant flavonoid in olive oil, originating from the glucoside that is present in the drupe, and its concentration highly depends on the geographical area, season, and environmental conditions [52]. High luteolin content is crucial due to its antioxidant and other health-related activities [53]. Kessen et al. [52] have reported that the concentration of luteolin ranged between 1.51 and 7.57 mg Kg⁻¹, showing variations among olive oils of different geographical regions and harvest years. In this research, luteolin exhibited higher values for organic EVOOs compared to conventional but showed no difference in the content range between the previous harvesting year. Luteolin ranged between 4.16 and 7.03 mg Kg⁻¹ for organic EVOOs, both harvested in 2014–2015 and 2015–2016. Therefore, it can be a good indicator for the discrimination of organic and conventional EVOOs in different harvesting years, as it is not affected by climate changes. The feature selection algorithm found luteolin as the top compound to make this discrimination possible and predict if a sample is organic or conventional. ACO-RF/RF found that there is a threshold for luteolin in EVOOs of different production types and calculated this threshold at 4.16 mg Kg⁻¹. EVOOs with higher concentrations of luteolin are organic, and those with lower than 4.16 mg Kg⁻¹ are predicted as conventional.

Conclusions

This study contributes to the field of food authenticity by discriminating the organic and conventional EVOOs using an optimized LLME-UHPLC-QTOF-MS method. Target and suspect screening quantification results together with ACO-RF established a discrimination model that could reveal markers responsible for the discrimination of production type in EVOOs.

The optimum extraction condition and the selection of the appropriate internal standard were achieved using OFD-RSM, as an experimental design and optimization technique. The results showed that the extraction with MeOH/H₂O (80:20, v/v) presents the lowest %RSD values and showed that syringaldehyde 1.30 mg L⁻¹ is the most appropriate internal standard.

The proposed method was successfully applied in 52 EVOOs of *Kolovi* variety produced during the harvesting periods of 2014–2015 and 2015–2016. Totally, 13 target and 24 suspect phenolic compounds were identified. All target compounds were quantified based on their commercially available reference standards, while the identified suspect compounds were semi-quantified according to a novel strategy that incorporates the chemical structure similarity.

A robust discrimination model was established by RF and it further coupled to RF to prioritize the target and suspect phenolic compounds quantification and semi-quantification results, respectively, according to their importance in

discrimination task. However, coupling ACO to RF did not change the initial results of RF. Eventually, the flavonoid luteolin was found to be responsible for the discrimination, and if its content is higher than 4.16 mg kg⁻¹, the EVOO is organic; while if it is less than 4.16 mg kg⁻¹, the EVOO should be characterized conventional. The proposed discrimination model, based on 52 samples of *Kolovi* variety from Lesvos Island within a 2-year study, is robust showing high internal and external accuracy, and thus, it could be sufficiently employed for the discrimination between organic and conventional EVOOs and further guarantee quality and authenticity.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Research involving human participants and/or animals This work does not contain any studied with human participants or animals.

References

1. Bourn D, Prescott J. A comparison of the nutritional value, sensory qualities, and food safety of organically and conventionally produced foods. *Crit Rev Food Sci Nutr*. 2002;42(1):1–34.
2. Vallverdú-Queralt A, Lamuela-Raventós RM. Foodomics: a new tool to differentiate between organic and conventional foods. *Electrophoresis*. 2016;37:1784–94.
3. García-González DL, Aparicio R. Research in olive oil: challenges for the near future. *J Agric Food Chem*. 2010;58:12569–77.
4. Cifuentes A. Food analysis: present, future, and Foodomics. *ISRN Anal Chem*. 2012;2012:1–16.
5. Koh E, Charoenprasert S, Mitchell AE. Effect of organic and conventional cropping systems on ascorbic acid, vitamin C, flavonoids, nitrate, and oxalate in 27 varieties of spinach (*Spinacia oleracea* L.). *J Agric Food Chem*. 2012;60(12):3144–50.
6. Ren H, Endo H, Hayashi T. Antioxidative and antimutagenic activities and polyphenol content of pesticide-free and organically cultivated green vegetables using water-soluble chitosan as a soil modifier and leaf surface spray. *J Sci Food Agr*. 2001;81(15):1426–32.
7. Shepherd LVT, Hackett CA, Alexander CJ, Sungurtas JA, Pont SDA, Stewart D, et al. Effect of agricultural production systems on the potato metabolome. *Metabolomics*. 2014;10(2):212–24.
8. Vallverdú-Queralt A, Medina-Remón A, Casals-Ribes I, Lamuela-Raventós RM. Is there any difference between the phenolic content of organic and conventional tomato juices? *Food Chem*. 2012;130(1):222–7.
9. Rosati A, Cafiero C, Paoletti A, Alfei B, Caporali S, Casciani L, et al. Effect of agronomical practices on carpology, fruit and oil composition, and oil sensory properties, in olive (*Olea europaea* L.). *Food Chem*. 2014;159:236–43.
10. Ranalli A, Contento S. Analytical assessment of destoned and organic-destoned extra-virgin olive oil. *Eur Food Res Technol*. 2010;230(6):965–71.
11. Ninfali P, Bacchiocca M, Biagiotti E, Esposito S, Servili M, Rosati A, et al. A 3-year study on quality, nutritional and organoleptic evaluation of organic and conventional extra-virgin olive oils. *J Am Oil Chem Soc*. 2008;85(2):151–8.

12. Anastasopoulos E, Kalogeropoulos N, Kaliora AC, Kountouri A, Andrikopoulos NK. The influence of ripening and crop year on quality indices, polyphenols, terpenic acids, squalene, fatty acid profile, and sterols in virgin olive oil (Koroneiki cv.) produced by organic versus non-organic cultivation method. *Int J Food Sci Tech*. 2011;46(1):170–8.
13. Ferreira SL, Bruns RE, Ferreira HS, Matos GD, David JM, Brandao GC, et al. Box-Behnken design: an alternative for the optimization of analytical methods. *Anal Chim Acta*. 2007;597(2):179–86.
14. Maragou NC, Thomaidis NS, Koupparis MA. Optimization and comparison of ESI and APCI LC-MS/MS methods: a case study of Irgarol 1051, Diuron, and their degradation products in environmental samples. *J Am Soc Mass Spectrom*. 2011;22(10):1826–38.
15. Anderson MJ, Whitcomb PJ. RSM simplified: optimizing processes using response surface methods for design of experiments. 2nd ed. New York: Productivity Press; 2016.
16. Baş D, Boyacı İH. Modeling and optimization I: usability of response surface methodology. *J Food Eng*. 2007;78(3):836–45.
17. Bezerra MA, Santelli RE, Oliveira EP, Villar LS, Escalera LA. Response surface methodology (RSM) as a tool for optimization in analytical chemistry. *Talanta*. 2008;76(5):965–77.
18. Yang J, Wang AQ, Li XJ, Fan X, Yin SS, Lan K. A chemical profiling strategy for semi-quantitative analysis of flavonoids in ginkgo extracts. *J Pharm Biomed Anal*. 2016;23:47–154.
19. Holliday JD, Salim N, Willett P. On the magnitudes of coefficient values in the calculation of chemical similarity and dissimilarity. *ACS Symp Ser*. 2005;894:77–95.
20. Cincilla G, Thormann M, Pons M. Structuring chemical space: similarity-based characterization of the PubChem database. *Mol Inform*. 2010;29(1–2):37–49.
21. Andersson PL, Fick J, Rannar S. A multivariate chemical similarity approach to search for drugs of potential environmental concern. *J Chem Inf Model*. 2011;51(8):1788–94.
22. Hariharan R, Janakiraman A, Nilakantan R, Singh B, Varghese S, Landrum G, et al. MultiMCS: a fast algorithm for the maximum common substructure problem on multiple molecules. *J Chem Inf Model*. 2011;51(4):788–806.
23. Ma C, Wang L, Xie XQ. GPU accelerated chemical similarity calculation for compound library comparison. *J Chem Inf Model*. 2011;51(7):1521–7.
24. Maggiora G, Vogt M, Stumpfe D, Bajorath J. Molecular similarity in medicinal chemistry. *J Med Chem*. 2014;57(8):3186–204.
25. Singer HP, Wössner AE, McArdell CS, Fenner K. Rapid screening for exposure to "non-target" pharmaceuticals from wastewater effluents by combining HRMS-based suspect screening and exposure modeling. *Environ Sci Technol*. 2016;50(13):6698–707.
26. Bu Q, Luo Q, Wang D, Rao K, Wang Z, Yu G. Screening for over 1000 organic micropollutants in surface water and sediments in the Liaohe River watershed. *Chemosphere*. 2015;138:519–25.
27. Dorigo M, Birattari M, Stützle T. Ant colony optimization artificial ants as a computational intelligence technique. *IEEE Comput Intell Mag*. 2006;1(4):28–39.
28. Ballabio D, Consonni V. Classification tools in chemistry. Part 1: linear models. PLS-DA. *Anal Methods*. 2013;5(16):3790–8.
29. Kalogiouri NP, Alygizakis NA, Aalizadeh R, Thomaidis NS. Olive oil authenticity studies by target and nontarget LC-QTOF-MS combined with advanced chemometric techniques. *Anal Bioanal Chem*. 2016;408(28):7955–70.
30. Wolf S, Schmidt S, Muller-Hannemann M, Neumann S. In silico fragmentation for computer assisted identification of metabolite mass spectra. *BMC Bioinf*. 2010;11:148. <http://foodb.ca/>. Accessed 22 Nov 2016.
31. Aalizadeh R, Thomaidis NS, Bletsou A, Gago-Ferrero P. Quantitative structure-retention relationship models to support suspect high resolution mass spectrometric screening of emerging contaminants in environmental samples. *J Chem Inf Model*. 2016;56(7):1384–98.
32. Schymanski EL, Jeon J, Gulde R, Fenner K, Ruff M, Singer HP, et al. Identifying small molecules via high resolution mass spectrometry: communicating confidence. *Environ Sci Technol*. 2014;48(4):2097–8.
33. Stat-Ease I: Design-Expert. In., 7.0.0 edn. USA; 2005.
34. Liu P, Agrafiotis DK, Theobald DL. Fast determination of the optimal rotational matrix for macromolecular superpositions. *J Comput Chem*. 2010;31(7):1561–3.
35. Todeschini R, Lasagni M, Marengo E. New molecular descriptors for 2D and 3D structures. *Theory J Chemom*. 1994;8(4):263–72.
36. Todeschini R, Moro G, Boggia R, Bonati L, Cosentino U, Lasagni M, et al. Modeling and prediction of molecular properties. Theory of grid-weighted holistic invariant molecular (G-WHIM) descriptors. *Chemom Intel Lab Syst*. 1997;36(1):65–73.
37. Todeschini R, Cazar R, Collina E. The chemical meaning of topological indices. *Chemom Intell Lab Syst*. 1992;15(1):51–9.
38. Todeschini R, Gramatica P, Provenzani R, Marengo E. Weighted holistic invariant molecular descriptors. Part 2. Theory development and applications on modeling physicochemical properties of polyaromatic hydrocarbons. *Chemom Intell Lab Syst*. 1995;27(2):221–9.
39. Todeschini R, Consonni V. Frontmatter. Handbook of molecular descriptors. In: Mannhojd R, Kubinyi H, Timmerman H, editors. Methods and principles in medicinal chemistry, vol. 11. Weinheim: Wiley-VCH; 2000. doi:10.1002/9783527613106.fmatter.
40. Tetko IV, Gasteiger J, Todeschini R, Mauri A, Livingstone D, Ertl P, et al. Virtual computational chemistry laboratory—design and description. *J Comput Aided Mol Design*. 2005;19(6):453–63.
41. Virtual Computational Chemistry Laboratory. <http://www.vcclab.org>. Accessed 10 Sept 2016.
42. Partitioning(logD) plugin was used for the calculation of logD, ChemAxon. <http://www.chemaxon.com>. Accessed 10 Sept 2016.
43. Ballabio D, Consonni V, Mauri A, Claeys-Bruno M, Sergeant M, Todeschini R. A novel variable reduction method adapted from space-filling designs. *Chemometr Intell Lab Syst*. 2014;136:147–54.
44. Dorigo M, Blum C. Ant colony optimization theory: a survey. *Theor Comput Sci*. 2005;344(2–3):243–78.
45. Kim KS, Choi HH, Moon CS, Mun CW. Comparison of k-nearest neighbor, quadratic discriminant and linear discriminant analysis in classification of electromyogram signals based on the wrist-motion directions. *Curr Appl Phys*. 2011;11(3):740–5.
46. McLachlan G. Discriminant analysis and statistical pattern recognition. New York: Wiley; 2004.
47. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction. 2nd ed. United States: Springer; 2009.
48. Breiman L. Random forests. *Mach Learn*. 2001;45(1):5–32.
49. Genuer R, Poggi J-M, Tuleau-Malot C. Variable selection using random forests. *Pattern Recogn Lett*. 2010;31(14):2225–36.
50. Candiotti LV, De Zanb MM, Cámarab MS, Goicoechea HS. Experimental design and multiple response optimization. Using the desirability function in analytical methods development. *Talanta*. 2014;124:123–38.
51. Kesen S, Kelebek H, Selli S. LC-ESI-MS characterization of phenolic profiles Turkish olive oils as influenced by geographic origin and harvest year. *J Am Oil Chem Soc*. 2014;91(3):385–94.
52. Diamantakos P, Velkou A, Killday KB, Gimisis T, Melliou E, Magiatis P. Oleokoronal and oleomissional: new major phenolic ingredients of extra virgin olive oil. *Olivae*. 2015;122:22–33.