Secondary-metabolites fingerprinting of Argania spinosa kernels using liquid chromatography-mass spectrometry and chemometrics, for metabolite identification and quantification as well as for geographic classification

Kharbach, Mourad; Viaene, Johan; Yu, Huiwen; Kamal, Rabie; Marmouzi, Ilias; Bouklouze, Abdelaziz; Heyden, Yvan Vander

Published in:
Journal of Chromatography. A

DOI:
10.1016/j.chroma.2022.462972

Publication date:
2022

Secondary-metabolites fingerprinting of *Argania spinosa* kernels using liquid chromatography–mass spectrometry and chemometrics, for metabolite identification and quantification as well as for geographic classification

Mourad Kharbach[1,2,3*], Johan Viaene[3], Huiwen Yu[4], Rabie Kamal[5], Ilias Marmouzi[6], Abdelaziz Bouklouze[6] and Yvan Vander Heyden[3**]

[1] Department of Food and Nutrition, P.O. Box 66, 00014, University of Helsinki, Finland

[2] Department of Computer Science, University of Helsinki, 00014 Helsinki, Finland

[3] Department of Analytical Chemistry, Applied Chemometrics and Molecular Modelling, Vrije Universiteit Brussel (VUB), Laarbeeklaan 103, B-1090 Brussels, Belgium

[4] Chemometrics and Analytical Technology, Faculty of Science, University of Copenhagen, Rolighedsvej 26, 1958 Frederiksberg, Denmark

[5] Pharmacodynamy Research Team. Laboratory of Pharmacology and Toxicology, Faculty of Medicine and Pharmacy, University Mohammed V, Rabat, Morocco

[5] Biopharmaceutical and Toxicological Analysis Research Team, Laboratory of Pharmacology and Toxicology, Faculty of Medicine and Pharmacy, University Mohammed V, Rabat, Morocco

*Corresponding author: Mourad Kharbach, E-mail address mourad.kharbach@hotmail.fr

**Corresponding author: Yvan Vander Heyden, E-mail address yvanvdh@vub.be

Highlights

- Geographical origin of Argan kernels based on secondary-metabolite profiles.
- 36 secondary metabolites (33 polyphenolic and 3 non-phenolic) were quantified.
- Untargeted UPLC-MS fingerprints were decomposed by metabolomic data handling tools.
- MCR–ALS and XCMS were compared to extract the features from UPLC-MS data.
- PCA, PLS-DA, SIMCA and data fusion (low- and mid-) were used to handle the data.

Abstract

Argan (*Argania spinosa* L.) fruit kernels' composition has been poorly studied and received less research intensity than the resulting Argan oil. The Moroccan Argan kernels contain a wealth of metabolites and can be investigated for nutritional and health aspects as well as for economic benefits. Ultra-Performance Liquid Chromatography Mass Spectrometry (UPLC-MS) was employed to trace the geographical origin of Argan kernels based on secondary-metabolite profiles. One-hundred and twenty Argan fruit kernels from five regions ('Agadir', 'Ait-Baha' 'Essaouira', 'Tiznit' and 'Taroudant') were studied. Characterization and quantification of 36 secondary metabolites (33 polyphenolic and 3 non-phenolic) were achieved. Those metabolites are highly influenced by the geographic origin. Then, the untargeted UPLC-MS fingerprint was decomposed by metabolomic data handling tools, such as multivariate curve resolution alternating least squares (MCR–ALS) and XCMS. The two resulting data matrices were pretreated and prepared separately by chemometric tools and then two data fusion strategies (low- and mid-levels) were applied on them. The four data sets were comparatively investigated. Principal component analysis (PCA), Partial Least Squares Discriminant Analysis (PLS-DA), and Soft Independent Modeling of Class Analogies (SIMCA) were used to classify samples. The exploration or classification models demonstrated a good ability to discriminate and classify the samples in the geographical-origin based classes.

Summarized, the developed fingerprints and their metabolomics-based data handling successfully allowed geographical traceability evaluation of Moroccan Argan kernels.

Keywords: Argan fruit kernels; *Argania spinosa* L.; UPLC-MS; Multivariate Classification; Untargeted fingerprints; Metabolomic profiles.

1. Introduction

The last decades, in relation to human nutrition, the search for bioactive secondary metabolites molecules, including polyphenoliccompounds (e.g. phenolic acids, flavonoids, hydroxycinnamic acids, hydroxybenzoic acids and stilbenes) received increased consideration [1]. Daily consumption of polyphenol sources (i.e. marines, plants, fruits, vegetables, juices and beverages) leads to a balanced diet and health benefits, such as antioxidant, cardiovascular, antithrombotic, anti-inflammatory and antitumor ones [2]. Food materials are characterized by their nutritional and health properties, which are directly linked to metabolites composition. Additionally, particular chemical metabolites, such as polyphenolic compounds, influence food properties, color, taste, health and nutritional quality [3]. Nowadays, one of the major challenges in the agricultural–food industry is to develop objective tools to determine the geographical origin of raw materials as well as of finished food products, certifying their traceability [4]. Argan kernels/oils are exported worldwide, and the geographical origin is very important for quality insurance and traceability.

Argan fruits (kernels, pulp or seeds) are a rich source of bioactive compounds, including polyphenols (flavonoids, phenolic acids and aminophenols), tocopherols, fatty acids, phytosterols, saponins, and triacylglycerols [5]. The secondary-metabolites, explicitly polyphenols, were evidenced to be functional and to have health benefits, as mentioned above. Argan kernels are harvested from the Argan tree (*Argania spinosa*. L), which is an endemic tree growing in arid and semiarid Moroccan regions [6]. The Argan tree plays an ecologic, economic, and social role, and maintains soil against erosion and desertification. The Argan kernels production ranges from 1.3 to 42.1 kg per tree, while Argan oil occurs in an amount from 2 to 7% of the Argan kernels [7]. Moreover, a high quantity of the Argan kernels byproducts (after oil extraction) is wasted, though they offer a source of secondary metabolites and should be investigated and valorized for economic benefits.

Nowadays, the *in vitro* antioxidant potential and the total phenolic content measurements, based on colorimetric or spectrophotometric methods, are no longer considered appropriate [8]. The explanations given are that no single standard method assessing the global antioxidant potential is available (i.e. different methods provide different results) and only techniques identifying antioxidant molecules are considered efficient [9]. Phenolic compounds are plant secondary-metabolites possessing an aromatic ring bearing one or more hydroxyl groups [10]. It are substances derived from phenylalanine, which is transformed to tyrosine, cinnamic acid, or some combinations between hydroxyl groups and phenyl rings to form different phenolic

classes. Phenolic compounds can be classified in two major groups, *flavonoids* (anthocyanidins, chalcones, flavanols, flavanones, flavones, flavonols, iso-flavonoids, neoflavonoids and proanthocyanidins) and *non-flavonoids* (polyphenolic amides, coumarins, hydroxybenzoic acids, hydroxycinnamic acids, lignoids, stilbenes and tannins) [11]. In this respect, liquid chromatography coupled to a mass-spectrometric detector is highly recommended to either identify, characterize or quantify metabolites in different matrices [2].

The determination and certification of product origin represent a form of quality, safety and credibility for the food industry. The polyphenolic compounds in Argan kernels have received low interest and only few studies are reported [12-14]. Those studies are based on a limited number of samples while the geographic origin effect is neglected. The phenolic compounds content in plant matrices depends on several factors, including agronomic practice, geographic origin, postharvest conditions, environmental and bioclimatic factors, and genetic traits [11]. The geographic origin influences on polyphenolic contents in, for instance, almonds [15], coffees [16], olive oils [17], chocolate [18], fruits [19], vegetables [20], and wines [20], were evaluated by liquid chromatography mass spectrometry (LC-MS) associated to chemometric tools. Metabolomics is a scientific field involving the study and identification of small metabolites and is commonly applied in many domains, including plants [21], food [22], environmental analysis [23], diseases [24], and pharmaceutical and clinical research [25]. Metabolomics is highly related to the type of analytical technique (NMR, GC-MS, CE-MS, LC-MS) used and to the data-analysis efficiency. The analytical techniques generate complex data, which may need preprocessing (e.g., binning, isotoping, noise filtering, alignment, peak-picking, peak resolution and feature identification) to extract useful information. Several statistical and data handling tools, including MCR-ALS [26], XCMS [27], MetaboAnalyst [28], MetAlign [29], and MZmine [30], were developed to interpret the untargeted metabolomics data and to identify features (metabolites). The outcome of those tools highly depends on parameter settings that affect the data quality and that could generate different numbers of features determined [31].

The Argan kernels (AK) present a wealth source of secondary metabolites, specifically of polyphenols. To the best of our knowledge, no extensive study concerning the geographic origin effect on the secondary metabolites (polyphenolic and non-phenolic compounds) content in AK has been reported yet. The aim of this study is to survey the geographic origin effect on the secondary-metabolites (polyphenols and other metabolites) distribution of AK from five regions, based on their untargeted UPLC-MS profiles. The obtained data set was treated by

metabolomics tools (XCMS and MCR-ALS), data fusion and multivariate methods (e.g., XCMS, MCR-ALS, PCA, SIMCA and PLS-DA). On the other hand, the study also aimed at the development of an analytical UPLC-MS fingerprint allowing the separation and accurate quantification of many (here 36) secondary-metabolites in AK.

2. Material and methods

2.1. Argan kernels collection and extraction

Mature and healthy Argan fruits were harvested from the same Argan trees and from five Moroccan regions, namely Agadir, Essaouira, Ait-Baha, Taroudant and Tiznit, in two successive harvests 2016 and 2017 (between August and October). The sampling was concentrated and distributed, as much as possible, on the three main climate zones (arid, semi-arid and sub-Saharan) that characterize all the areas of Argan tree forest in Morocco. More detailed information about geographical parameters, specific provenances and climatic characteristics (average annual temperature and rainfall) is given in Table S1 (Supplementary material). One-hundred-twenty samples were collected (twenty-four from each region), and 500g of each sample is weighted. A fruit sample is a mixture of three fruit-form varieties (fusiform, oval, round) without considering the genotypic profile. Then, in the traditional method (Women cooperatives), the fruits were carefully crushed using two stones. Only intact kernels (not crushed) were gathered in dark glass bottles and stored in the freezer (-5°C) for one month until transport to the Belgium laboratory for analysis. Before analysis, the kernels were placed at room temperature for a few hours, then ground manually using a porcelain mortar and pestle until a smooth and homogeneous powder mixture is obtained. A homogenous 500 mg of each AK powder was treated separately as follows, targeting mainly polyphenols, 4 mL methanol (solid/liquid extraction) was added, then vortexed for 10 min and stored in the dark for 12 h. The same procedure was applied three times, and then the methanolic extracts (∽12 mL) were combined and washed with 5 mL hexane (liquid/liquid extraction) twice to eliminate the lipid traces. The methanolic solvent was eliminated by using a rotatory evaporator (45 °C). A dilution of 1:5 with methanol (w/v) was prepared for each sample, which was filtered through a membrane filter (0.20-µm, PVC) in a glass vial.

The Argan tree genetic variability, tree elite genotypes, morphological characterization, and tree phenological traits (flowers, leaves, fruits, and kernels) were not taken in consideration and

studying their specific effects on the polyphenols or metabolic profiles is considered out of the scope of this study.

2.2. UPLC-PAD-QTOF/MS separation conditions

The secondary-metabolites profiling was carried out on an Ultra-Performance Liquid Chromatography (UPLC, Acquity system, Waters, Milford, MA, USA) equipped with a photo diode array detector (PDA) and an electrospray ionization quadrupole time-of-flight tandem mass spectrometer (ESI–QTOF/MS; XevoG2-S). Separation was achieved on a BEH phenyl C18 column (100 mm × 2.1 mm, 1.7 µm, ACQUITY UPLC®, Waters), at a temperature of 40°C. A binary mobile phase was used, solvent A (0.1% formic acid in water) and solvent B (0.1% formic acid in acetonitrile), in a gradient separation of 30 min: starting with 90% A, 0 min; 30% A, 0–18 min; 0% A, 18–20 min; 0% A, 20–23 min; 90% A, 23–25 min; and reconditioning the system with 90% A, 25–30 min. The flow rate was maintained at 0.5 ml/min, and 10 µL was the injection volume. PDA absorbance was recorded from 210 to 400 nm. All solvents are UPLC-MS grade from Biosolve (Valkenswaard, the Netherlands).

The mass spectrometry system was applied in negative mode. The operating conditions were as follows: mass range (50–1200 m/z, full mode scan); desolvation temperature, 500 °C; source temperature, 150 °C; capillary voltage, 1 kV; extractor voltage, 2 V; and cone voltage, 30 V. Nitrogen (N_2) was used as desolvation and cone gas. The cone gas flow and desolvation gas flow were 0 and 1000 L/h, respectively. The mass-spectra data acquisition was recorded in two continuous modes, a no collision energy mode for precursor ion information (MS, 0 eV), and a high collision energy mode for fragment information (MS^E, 15-45 eV). The ESI source system was calibrated using leucine-enkephalin (Sigma-Aldrich, Steinheim, Germany) as an internal lock-mass reference (LockSpray™). The MassLynx™ 4.1 software (Waters) was used to acquire the data. During the method optimization and development both positive and negative modes were tested but only the negative mode showed a stable and reproducible MS profile. Therefore, only the negative mode was applied for the secondary-metabolites fingerprinting.

2.3. Secondary-metabolites identification & quantification

A tentative identification and quantification of secondary metabolites was performed. Thirty-seven phenolic standards (Sigma Aldrich, St Quentin Fallavier, France) were used: catechin, epicatechin, epigallocatechin 3-o-gallate; kaempferol; quercetin; quercitrin (quercetin 3-o-rhamnoside); rutin (quercetin 3-o-rutinoside); hesperidin (hesperetin 7-o-rutinoside);

hesperetin; naringin; naringenin; luteolin; benzoic acid; gallic acid; protocatechuic acid (3,4-dihydroxybenzoic acid); salicylic acid (2-hydroxybenzoic acid); syringic acid; vanillic acid (4-hydroxy-3-methoxybenzoic acid); m-hydroxybenzoic acid (3-hydroxybenzoic acid); p-hydroxybenzoic acid (4-hydroxybenzoic acid); caffeic acid; chlorogenic acid (5-caffeoylquinic acid); ferulic acid (3-methoxy-4-hydroxycinnamic acid); m-coumaric acid (3-hydroxycinnamic acid); o-coumaric acid (2-hydroxycinnamic acid); p-coumaric acid (4-hydroxycinnamic acid); rosmarinic acid; sinapic acid; resveratrol; esculetin (6,7-dihydroxycoumarin); esculin; pyrocatechol (1,2-dihydroxybenzene); pyrogallol (1,2,3-trihydroxybenzene); 4-hydroxycoumarin; citric acid; quinic acid and succinic acid. Moreover, avicularin (quercetin 3-O-arabinoside), hyperin (quercetin 3-O-galactoside), isoquercitrin (quercetin 3-O-glucoside), quercetin glycocoumarate, quercetin glycoferulate, quercetin glycohydroxybenzoate, quercetin glycogallate, and quercetin glycosinapate were identified using information from the literature on phenolic compounds in AK [13, 14], and were quantified relative to a quercetin standard curve.

2.4. Reagents

A stock solution (50 mL) of each pure standard (a concentration of 1 mg/mL) was prepared in methanol/water (60:40, v: v). To ensure dissolution, the solution was vortexed for 10 min and put on an ultrasonic bath for 15 min. Then 6 dilutions were prepared (1, 5, 10, 25, 50, 100 µg/mL), which were stored in the freezer (-20 °C). Validation parameters were acquired following the recommendations and regulations from internationally recognized guidelines [32]-[33]. The following parameters were reported: the method precision, i.e., intra- and inter-day precision (expressed as relative standard deviation (RSD)), linearity range, limit of detection (LOD) and the lowest limit of quantitation (LOQ). The intra-day precision was achieved from 5 replicates and the inter-day precision from 5 replicates on three days, then same replicates are under intra-day conditions.

The calibration curves were constructed (peak area vs concentration), and the quantitative analysis was carried out using the regression equations ($y = ax + b$), y is the peak area, x is the concentration, and the determination coefficients (r^2). Quantification results were determined in triplicates and expressed as $mg.kg^{-1}$ of AK ($\mu g.kg^{-1}$ were used to express the LOD and LOQ). Linear analysis by means of least-squares regression was applied to link the mass-spectra detector signal of each compound to the concentration. The method linearity was determined using a standards concentration range from 1 to 100 $\mu g.mL^{-1}$.

The identification of the phenolic compounds in the samples was achieved based on the standards information, i.e. retention times, mass spectra (accurate mass [M-H^-] precursor ion (molecular ion), mass-to-charge (m/z), and its fragmentation), on literature results [12-14], and online database (PubChem®; ChemSpider®; Phenol-Explorer; MassBank; and Spectral Database for Organic Compounds) information. The developed untargeted secondary-metabolites fingerprint served for geographical origin classification, while the metabolites identification and quantification were given as a complementary information.

2.5. Chemometric strategy approaches

The UPLC-MS raw data files were converted to an open format file (NetCDF) with the Databridge software (Waters). However, the UPLC-MS fingerprint is complex with thousands of variables (m/z ratio, retention times, noise, and features) for a rather high number of samples (120). Therefore, two approaches of metabolomics data handling, i.e. XCMS and MCR-ALS, were considered to process and reduce the data complexity. Those two strategies were used in order to extract useful information from the profiles which could serve for the study purpose.

2.5.1 Metabolomic tools

XCMS is an R-based package for mass spectrometry data processing visualization and metabolite detection [27]. The UPLC-MS (NetCDF) data was transferred directly to the R platform (version "4.0") (https://www.r-project.org/), and the XCMS script was applied using the centWave algorithm XCMS 3.10.1 package (https://bioconductor.org/packages/release/bioc/html/xcms.html) [27]. The XCMS processing parameters for our data were optimized with the Isotopologue Parameter Optimization (IPO) package [34]. The parameters optimized by IPO were method="centWave"; ppm = 10 (maximal tolerated m/z deviation in consecutive scans in ppm for peak picking); peakwidth = c(20, 50) (chromatographic peak width, specified as a range (min, max) in seconds) ; snthresh = 6 (signal-to-noise threshold); prefilter = c(3, 100) (3 is the scan points minimum per peak in prefilter, and 100 was the peak intensity minimum in prefilter); integrate = 1 (peak limits are found through descent on the mexican hat filtered data); mzdiff = 0.01 (minimum difference in m/z for peaks with overlapping retention times). Other parameters were set by default within the XCMS algorithm. The XCMS generated a new matrix of dimensions 120 samples vs 1934 features.

Multivariate curve resolution - alternating least squares (MCR-ALS) is a chemometric technique for decomposing chromatographic data profiles and extracting features [35, 36]. MCR-ALS decomposes the data in peak concentration- and pure mass spectra profiles. This technique is based on an alternating least squares (ALS) iterative process, while the singular value decomposition (SVD) algorithm was used to determine a minimum number of mixture components [37].

Before proceeding to a MCR-ALS data decomposition, the UPLC-MS data was compressed through a regions of interest (ROI) algorithm [38]. The ROI search is a good alternative for the classical data binning [39], it compresses the data without losing the mass accuracy. This strategy is already included in the centWave algorithm of XCMS [27]. The parameter settings for the ROI search were as follows, mass accuracy 0.05 Da, migration time for a peak 10s, signal threshold 1000 (10% of maximum MS intensity signal) and 800 time points [26]. After ROI selection, the data (MS-ROI) is arranged in augmented column-wise data matrix (D_{aug}). In the present study, 134 components were selected considering their variance explanation and lack-of-fit error, while non-negativity constraints were applied for both elution and spectral profiles. MCR-ALS decomposes the D_{aug} matrix (MS-ROI) into two factor matrices using a bilinear decomposition model,

$$\mathbf{D_{aug}} = \mathbf{C_{aug}S^T} + \mathbf{E_{aug}} \qquad (1)$$

where $\mathbf{C_{aug}}$ is the augmented matrix of the chromatographic elution profiles (concentration of features), $\mathbf{S^T}$ is mass-spectra matrix, and $\mathbf{E_{aug}}$ is the augmented residuals matrix [38]. The $\mathbf{C_{aug}}$ matrix will be used to investigate the geographical origin. This matrix has dimensions of 120 samples vs 134 features and is named the MCR-ALS dataset throughout this paper.

2.5.2 Data fusion

The data fusion (DF) strategy consists of combining features from several datasets. Nowadays, DF is widely applied to combine information from several techniques (e.g. from spectroscopy, chromatography and/ or sensors) for food and beverage classification and/or authentication [40]. Both “Low-level” and “Mid-level” data fusion strategies were addressed to handle the UPLC-MS fingerprints.

The “Low-level” data fusion (LL-DF) is an approach based on the concatenation of original data variables (a pretreatment is recommended). The XCMS and MCR-ALS datasets were

pretreated separately (area normalization between 0 and 1, followed by autoscaling), then LL-DF was applied to concatenate them. A new X dataset (LL-DF) with dimensions 120 samples vs 2088 features is obtained.

The "Mid-level" data fusion (ML-DF) is based on concatenation of Principal component analysis (PCA) features from each dataset (XCMS and MCR-ALS). The 10 first PCs were selected from each dataset to extract most of the data variance. The ML-DF dataset has dimensions of 120 samples vs 20 features. The two new datasets (LL-DF and ML-DF) were used for classification purposes.

2.5.3 Exploration and classification tools

PCA was applied to visualize and identify possible AK groups according to their geographical origin.

PLS-DA is a supervised classification technique, which uses a PLS algorithm to construct a regression model between the **X** matrix data and **y**-block (classes) [41]. A response between 0 and 1 is predicted, with a value above 0.5 indicating a class membership for a given sample.
Soft independent modeling of class analogy (SIMCA) is a class modeling technique that uses the PCA data decomposition [42]. A PCA analysis was fitted on each separate class, then a SIMCA model was constructed on the selected PCs. PLS-DA and SIMCA classified the AK samples according to their geographical origin, applying the UPLC-MS fingerprints.
Before chemometric analysis, all data were preprocessed (features or peak areas) by normalization (area normalization between 0 and 1) and autoscaling. The classification models were optimized using 'Venetian-blinds' cross-validation procedures [43]. The models were evaluated considering their sensitivity, specificity and accuracy (% correct classification). Sensitivity (or true positive rate) expresses the percentage of actual positive samples correctly classified in their proper class.

$$\textbf{Sensitivity}\ (\%) = 100\ x\ \frac{\text{Number of true positive samples (TP)}}{\text{Number of true positive samples (TP)} + \text{Number of false negative samples (FN)}} \quad (2)$$

Specificity (or true negative rate) expresses the percentage of actual negatives samples who are correctly classified in their proper class.

$$\textbf{Specificity}\ (\%) = 100\ x\ \frac{\text{Number of true negative samples (TN)}}{\text{Number of true negative samples (TN) + Number of false positive samples (FP)}} \quad (3)$$

Accuracy (or correct classification) represents the percentage of correct classified samples divided by the total number of assessments (equation 4).

$$\textbf{Accuracy}\ (\%) = 100\ x\ \frac{\text{Total number of corrected classified samples}}{\text{Total number of assessments}} = 100\ x\ \frac{\text{TN+TP}}{\text{TN+TP+FN+FP}} \quad (4)$$

2.6. Software

The geographical origin impact on the secondary metabolites was determined using one-way analysis of variance (ANOVA) and a Tukey's post hoc test ($\alpha = 5\%$).

Chemometric data analysis and processing was performed using MATLAB software 18.a (The Math-Works, Natick MA, USA), MCR-ALS toolbox (www.mcrals.info), Bioinformatics Toolbox® (The Mathworks) and PLS Toolbox® 8.6.1 (Eigenvector Research, Wenatchee, WA, USA).

3. Results and discussion

3.1. Analytical procedures (UPLC/ESI-QTOF-MS method)

Typical UPLC/ESI-QTOF-MS chromatograms of AK extracts from each studied region are exposed in Fig. 1. Different information was used to identify and quantify the compounds found in the AK fingerprints. Table S2 (Supplementary material) shows the analytical performance parameters of the optimized method.

The % RSD was used to express the precision (repeatability and time-different intermediate precision). LOD and LOQ were estimated using the signal-to-noise ratio, i.e., S/N=3 and 10, respectively.

Regarding the quantitative results from the prepared standards, good results were achieved. The calibration curves (peak area vs concentration) issued for the chemical standards were reported in Table S2 (Supplementary material). The determination coefficients were very high (r^2 >

0.99), except for salicylic acid ($r^2 > 0.96$) and chlorogenic acid ($r^2 > 0.95$). The detection limits ranged between 0.280 $\mu g.kg^{-1}$ for salicylic acid and 0.647 $\mu g.kg^{-1}$ for protocatechuic acid and the quantification limits between 0.720 $\mu g.kg^{-1}$ for succinic acid and 1.195 $\mu g.kg^{-1}$ for resveratrol. The inter-day RSD varied between 1.13% (epigallocatechin 3-O-gallate) and 4.16% (quinic acid), while intra-day RSD was between 2.15% (catechin) and 5.12% (quinic acid). All analytical parameters are acceptable to proceed a quantification analysis.

3.2. Characterization, identification and quantification of secondary metabolites

Table S3 (Supplementary material) summarizes the class, sub-class, name, chemical formula, molecular weight, PubChem identity (ID), retention time (RT), calculated mass - to - charge signal [M—H^-] and MS^2 ion fragments (m/z) of the secondary metabolites identified in the kernels. From the chromatographic profiles, four main polyphenolic classes were distinguished among which flavonoids and phenolic acids and stilbenes, and other polyphenols. The compounds in the flavonoids class belong to the flavanols (catechin, epicatechin, and epigallocatechin 3-O-gallate), flavonols (avicularin (quercetin 3-O-arabinoside), hyperin (quercetin 3-O-galactoside), isoquercitrin (quercetin 3-O-glucoside), quercetin, quercetin glycocoumarate, quercetin glycoferulate, quercetin glycohydroxybenzoate, quercetin glycogallate, and quercetin glycosinapate, quercitrin, and rutin), the flavanones (hesperidin, hesperetin, and naringin) and flavones (luteolin). Among the phenolic acids, the following hydroxybenzoic acids were identified: gallic acid, protocatechuic acid, salicylic acid, vanillic acid, m- hydroxybenzoic acid, and p-hydroxybenzoic acid; further also the following hydroxycinnamic acids: chlorogenic acid, ferulic acid, m-coumaric acid, o-coumaric acid, and p-coumaric acid. Resveratrol was identified as main compound in the stilbene class. Other polyphenols were identified, such as, pyrocatechol, pyrogallol and 4-hydroxycoumarin. Three non-phenolic compounds were also identified, i.e. citric acid, quinic acid, and succinic acid. Quinic acid may be issued from a natural degradation of chlorogenic acid or other polyphenols, while succinic acid is a component of citric acid. In total, 36 compounds were identified and quantified in AK samples, whereas 7 known polyphenolic compounds were not detected, i.e. kaempferol, naringenin, caffeic acid, rosmarinic acid, sinapic acid, esculetin and esculin. The compounds' identification and characterization were achieved using parameters from the pure standards (retention-time, mass-spectra (accurate mass [M—H^-] precursor ion (molecular peak), mass-to-charge (m/z), and fragmentation), and from comparison with mass databases (PubChem®; ChemSpider®; Phenol-Explorer; MassBank; and Spectral Database for Organic

Compounds). Avicularin, hyperin, isoquercitrin, quercetin glycocoumarate, quercetin glycoferulate, quercetin glycohydroxybenzoate, quercetin glycogallate, and quercetin glycosinapate were identified using parameters from the literature [13, 14] and their quantification was made relative to the quercitin standard curve. The identified secondary-metabolites and specifically the polyphenolic compounds are in agreement with previous reports of the AK analysis [5].

3.3. Secondary-metabolites content distribution

The secondary-metabolites profile was examined by UPLC-MS, and a simultaneous quantification was performed. This is the first extensive study of secondary-metabolites distribution and involved the determination of 33 polyphenolic and 3 non-phenolic compounds. The results were inspected by ANOVA analysis followed by a Tukey's post-hoc test to elucidate the geographic origin influence on the secondary-metabolite distribution (flavonoids, phenolic acids, stilbenes, other polyphenols and other compounds). Table 1 summarizes the quantitative results of 36 secondary-metabolites (33 polyphenolic and 3 non-phenolic compounds) from the five regions, expressed as mean ± standard deviation. The individual compounds and their total contents exhibited a clear statistical difference according to the geographical provenance ($p < 0.05$) (Table 1). Therefore, the quantitative difference in the secondary-metabolites profile could be exploited for geographical classification by multivariate data analysis tools.

The most abundant class are the non-phenolic compounds comprehending quinic acid followed by citric acid and succinic acid. Those compounds were influenced by the geographical provenance. Although, the flavonoids class was reported as the second abundant fraction, while 18 compounds were quantified. The flavanols sub-class exhibited highest concentrations as follows: epigallocatechin 3-O-gallate, followed by catechin, and epicatechin. Flavonol compounds were found as follows: hyperin, avicularin, isoquercitrin, quercetin glycohydroxybenzoate, quercetin glycocoumarate, quercetin glycosinapate, quercetin glycoferulate, quercetin, rutin, and quercitrin was detected in trace amount. In addition, the geographical origin was affecting the amount of those 18 compounds. The flavanones were varied from hesperidin, naringin, and trace amount of hesperetin. The luteolin (flavones) concentration was also varied depending to the five regions. The phenolic acids were reported as the third abundant class metabolites. The hydroxybenzoic acids were changed from region

to region as follows: protocatechuic acid, salicylic acid, vanillic acid, gallic acid, p-hydroxybenzoic acid, and m-hydroxybenzoic acid. The hydroxycinnamic acids varied according to provenance as follows: chlorogenic acid, followed by p-coumaric acid, ferulic acid, o-coumaric acid, and m-coumaric acid. The resveratrol content (stilbene class) was showed a high variation from Agadir samples to Ait-Baha samples. Some other polyphenolic compounds were also quantified in AK extracts (pyrocatechol, pyrogallol and 4-hydroxycoumarin). Earlier, amino-phenolics, catechins, flavonoids, procyanidins, phenolic acids and their glycolysated derivatives were quantified in the Argan fruits [5, 13, 14].

The geographical origin effect in the total fractions of secondary-metabolites was illustrated in Figure 2. The Ait-Baha samples exhibited the highest total content of flavonoids (2800 $mg.kg^{-1}$), phenolic acids (1430 $mg.kg^{-1}$), stilbenes (3 $mg.kg^{-1}$), other polyphenols (56 $mg.kg^{-1}$) and non-phenolic compounds (7000 $mg.kg^{-1}$) from the five provenances. The Ait-Baha region is characterized by a desert nature, an arid Mediterranean bioclimate, slight annual rainfall (average below 300 mm) and warm temperature (annual average 19 °C). The lowest content of those metabolites was reported in the Essaouira and Agadir regions (two oceanic coastal regions with high humidity, warm and humid arid bioclimate). The secondary metabolites amount was diverse and fluctuated considerably confirming the geographical origin impacts. Comprehensive information about geographical parameters, sample provenances and climatic characteristics, was recapitulated in Table S1 (Supplementary material). Result interpretation from literature studies of secondary metabolites in AK was not possible because of insufficient information, limited samples and usually unknown geographical origin. However, the comparison might be in the number of identified secondary metabolites and specifically polyphenols. Previously, from the Essaouira region sixteen polyphenolic compounds were identified in Argan kernels [12], sixteen in the Argan press-cake (AK residual after Argan oil extraction) [44] and thirty-two in immature AK [13]. Only nineteen compounds were identified in [14], in mature AK samples from both Agadir and Essaouira. The aminophenol fraction (e.g. arganimide A and argaminolics A-C) was isolated from AK by Klika et al. [45, 46], and their chemical structures were elucidated by NMR. It is worthwhile to mention that the cited studies focused on a limited number of samples (one or two) and from either Essaouira or Agadir regions.

In the literature, polyphenolic compounds distributions in food materials, confirming the effect of geographical origin, were described from almonds [15], coffee beans [16], olive oils [17], chocolate [18], fruits [19], vegetables [20], and wines [20]. On the other hand, the polyphenol content in foods is highly variable and depending on diverse factors, e.g. genetic characteristics,

maturity of composition, bioclimatic conditions, postharvest and storage time and conditions [47]. Moreover, this variability in polyphenols is influencing the food function and nutritional aspects.

In the next sections, the entire UPLC-MS secondary-metabolites fingerprint profiles (considering all metabolites) will be handled as a matrix for geographic origin discrimination purposes. The quantified compounds represent less than 10% of the data fingerprints. Therefore, two metabolomics (XCMS and MCR-ALS) tools were compared to extract characteristic features.

3.4. Chemometric data analysis

3.4.1. Data pretreatment

The overall secondary-metabolites fingerprints were considered as an untargeted profile matrix to construct the models. Since the UPLC-MS profile is complex to use at its raw state, it needs pre-processing steps (e.g., binning, isotoping, noise filtering, alignment, peak-picking, peak resolution and feature identification) to extract characteristic information. Therefore, two statistical data handling tools, MCR-ALS and XCMS, which include all these steps, helped to reduce the complexity of the untargeted UPLC-MS data and to target characteristic features. Further, two data fusion strategies were applied, i.e. low-level (concatenation of XCMS and MCR-ALS datasets) and mid-level (concatenation of PCA scores decomposition of XCMS and MCR-ALS datasets). Fig. 3, shows the used metabolomic tools and chemometric data handling strategies. Data fusion was implemented to enhance the classification models and combine metabolites information for geographical origin identification. Subsequently chemometric approaches (PCA exploration, SIMCA and PLS-DA classification) were investigated using features from four datasets, XCMS, MCR-ALS and data fusions (LL and ML), in order to evaluate the geographical origin.

The MCR-ALS dataset profile was composed of 120 samples and 134 features, the XCMS profile are of 120 samples and 1954 features, LL-DF of 120 samples and 2088 features (concatenation of 1954 and 134 features), and ML-DF of 120 samples and 20 features (10 first PCs was selected from XCMS and MCR-ALS datasets).

Each dataset was split in two, a calibration (87.5%; 105 samples) and a validation set (12.5%; 15 samples) based on the Kennard and Stone algorithm. The calibration set was composed of 21 samples from each region and the validation set of 3 per region.

The data normalization (mentioned above) followed by autoscaling preprocessing was applied on the four individual datasets in order to provide the same contribution for each feature in a given dataset.

3.4.2 PCA data exploration

As an initial step, to provide an overview of the dataset structure, a PCA exploration was applied to locate graphically the classes. PCA models were built from the four datasets of the AK extracts. This unsupervised technique exposed potential of grouping the samples according to their provenance. For instance, for the XCMS dataset, three PCs explained 55% of the total variance (Fig. 4 A). Four groups were clearly discriminated such as Agadir, Essaouira, Ait-Baha and Tiznit, while the Taroudant samples are situated between the two last groups (AB and Tiz).

For PCA applied on the MCR-ALS dataset, the first three PCs explained 59% of the total variance (Fig. 4 B). In the score plot, the discrimination of four groups is achieved but one group is composed of two types of samples (T and Tiz). Some samples from the Essaouira and Taroudant regions are located far away from the original group. This was also the case for the previous dataset.

The PCA applied on the LL-DF dataset explained 54% of the total variance using the 3 first PCs (Fig. 4 C). Three clusters were clearly observed, i.e. the AB samples and two groups composed of mainly two samples types (A and E) and (T and Tiz), respectively.

The PCA applied on ML-DF dataset exposed only 29% of the total variance when considering the first three PCs (Fig. 4 D). Four groups are distinguished, E, A, AB and T and Tiz samples which are gathered close. However, the data complexity is needed more than 3 PCs to expose its variability.

As a conclusion, the PCA models for the four datasets (XCMS, MCR-ALS, DF-LL and DF-ML) were displayed clear group distinction (less distinction between T and Tiz samples), and some samples from Essaouira regions could considered outliers.

Indeed, after a PCA data visualization, the capacity of secondary metabolites fingerprints to discriminate AK samples according to their origins should be possible.

3.4.3 Geographical origin classification

The UPLC-MS fingerprints were investigated to relate them to the geographic origin of the samples. Evaluation of the chromatographic profiles of AK revealed that the samples could be classified according to their geographic origin. Different classification models were constructed with the four datasets to assess the geographical origin effect on the distribution of secondary metabolites in the samples.

3.4.3.1 XCMS dataset

The XCMS dataset encompasses 1954 features. The classification results of the PLS-DA and SIMCA methods results are summarized in Table 2 (**A**). For the calibration, both PLS-DA and SIMCA models disclosed comparable results, 100% of sensitivity, specificity, and accuracy (correct classification) for the five classes (A, AB, E, Tiz and T). For the validation, PLS-DA displayed a perfect performance for all classes, and SIMCA 81% of correct classifications (sensitivity 60 % and specificity of 90%). The predictions for the PLS-DA models are shown in Figure 5 (a-b), while the Q residual vs T^2 Hoteling used for outliers detection.

3.4.3.2 MCR-ALS dataset

The MCR-ALS dataset (134 features) was also used to construct PLS-DA and SIMCA models Table 2 (**B**) outlines the resulting classification results. Both PLS-DA and SIMCA calibration models achieved perfect classification, with 100% sensitivity, specificity and correct classification for the five classes. For the prediction samples, a perfect prediction was achieved for both types of models. The prediction performances for calibration and validation of the PLS-DA models are shown in Figure 5 (c-d).

3.4.3.3 LL-DF dataset

The data fusion combines the XC-MS and MCR-ALS datasets (2088 features) to classify the samples according to their geographic origin. The classification results of the LL-DF dataset are displayed in Table 2 (**C**). Both PLS-DA and SIMCA calibration models obtained 100% accurate classification in the five classes, with perfect sensitivity, specificity and

accuracy. The prediction performance is better for the PLS-DA models (100%). SIMCA had some issues with only 91% overall correct classification (79% for A, 100% for AB and Tiz, 83% for E, and 92% for Tiz) with 2 E-class samples one Tiznit sample were wrongly predicted as A class. The results for the PLS-DA models are shown Figure 5 (e-f).

3.4.3.4 ML-DF dataset

The data fusion based on the mid-level strategy (PCA features extraction (20 features)), was also applied for the geographic origin classification. Table 2 (**D**) summarizes the classification results of the ML-DF dataset. Both PLS-DA and SIMCA calibration models attained a perfect performance with 100% of sensitivity, specificity and accuracy and all validation samples (15) were correctly classified. The PLS-DA graphs of the classification prediction results are given in Figure 5 (j-h).

Summarized, regarding the classification results from the four datasets and applying the two linear models (PLS-DA or SIMCA), the best modeling performance was consequently obtained from the ML-DF, MCR-ALS, LL-DF, and XCMS datasets. The PLS-DA models showed a better performance than the SIMCA models, and thus are preferred to perform the classification. The data fusion strategy enabled a good predictive ability for the geographic origin classification. ML-DL or MCR-ALS allowed perfect classification both for PLS-DA and SIMCA. Our study reveals that the untargeted UPLC-MS approach combined with appropriate chemometric data handling is an appropriate methodology for the geographical origin indication. The knowledge about the secondary-metabolites distribution and their geographical origin understanding will provide the necessary information for economic valorization of the AK (either whole fruits or by-products after oil extraction).

4. Conclusion

In this study, a workflow for identification, characterization and quantification of secondary metabolites in AK was demonstrated. Preliminary analysis exhibited thirty-six secondary metabolites (33 polyphenolic and 3 non-phenolic compounds) contents significantly varied according to their geographic provenance. The untargeted UPLC-MS data was handled

by metabolomic tools (XC-MS and MCR-ALS), then data fusion and chemometric tools (PCA, PLS-DA and SIMCA) were applied. Results revealed that PCA exploration besides PLS-DA and SIMCA classification models enabled the discrimination of AK according to their geographical origin.

The approaches applying either XCMS or MCR-ALS data as well as their data fusion demonstrated the feasibility of UPLC-MS metabolomic profiling for Argan kernels' geographic origin traceability.

Further research should be performed to identify given secondary metabolites as potential biomarkers responsible for the geographical origin characterization. It is worthwhile noticing that these fairly promising results encourage further studies, extending the study to a larger number of samples, and taking also into account several factors, for instance, in relation to agronomic practice, postharvest conditions, environmental and climatic factors, and tree genetic diversity and their combinations in order to generate a global and robust model.

Author contributions

M.K.: Conceptualization; Methodology; Formal analysis; Investigation; Carried out the experiment; Data curation, Writing - original draft. J.V.: Supervision; UPLC-MS experiments, Data curation. H.Y.: Investigation; Formal analysis, Resources. R.K.: Investigation; Resources. I.M.: Investigation; Resources. A.B.: Conceptualization; Validation; Writing - review & editing, Project administration. Y.V.H.: Conceptualization; Validation; Writing - review & editing, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that no known competing financial interests or personal relationships that could influence the reported work.

Acknowledgements

The authors appreciate the support of the "Mohammed VI Foundation for Research and Protection of the Argan Tree". The authors also acknowledge the financial support of Fonds Wetenschappelijk Onderzoek - Vlaanderen (FWO-Vlaanderen) (GO38816N), VLIR-UOS (Team project-VLIR 345 MA2017), the Vrije Universiteit Brussel (VUB), and the Faculty of Medicine and Pharmacy-Rabat (FMPR).

References

[1] D. Granato, A. Mocan, J.S. Câmara, Is a higher ingestion of phenolic compounds the best dietary strategy? A scientific opinion on the deleterious effects of polyphenols in vivo, Trends in Food Science & Technology, 98 (2020) 162-166.

[2] D. Granato, J.S. Santos, L.G. Maciel, D.S. Nunes, Chemical perspective and criticism on selected analytical methods used to estimate the total content of phenolic compounds in food matrices, TrAC Trends in Analytical Chemistry, 80 (2016) 266-279.

[3] H. Cory, S. Passarelli, J. Szeto, M. Tamez, J. Mattei, The role of polyphenols in human health and food systems: A mini-review, Frontiers in Nutrition, 5 (2018) 87.

[4] M. Garaus, H. Treiblmaier, The influence of blockchain-based food traceability on retailer choice: The mediating role of trust, Food Control, 129 (2021) 108082.

[5] F. Khallouki, M. Eddouks, A. Mourad, A. Breuer, R. Owen, Ethnobotanic, Ethnopharmacologic Aspects and New Phytochemical Insights into Moroccan Argan Fruits, International Journal of Molecular Sciences, 18 (2017) 2277.

[6] D. Guillaume, D. Pioch, Z. Charrouf, Argan [Argania spinosa (L.) Skeels] Oil, in: Fruit Oils: Chemistry and Functionality, Springer International Publishing, 2019, pp. 317-352.

[7] M. Zunzunegui, F. Ain-Lhout, J. Jáuregui, M.C. Díaz Barradas, S. Boutaleb, L. Álvarez-Cansino, M.P. Esquivias, Fruit production under different environmental and management conditions of argan, Argania spinosa (L.), Journal of Arid Environments, 74 (2010) 1138-1145.

[8] J. Harnly, Antioxidant methods, Journal of Food Composition and Analysis, 64 (2017) 145-146.

[9] D. Granato, F. Shahidi, R. Wrolstad, P. Kilmartin, L.D. Melton, F.J. Hidalgo, K. Miyashita, J.v. Camp, C. Alasalvar, A.B. Ismail, S. Elmore, G.G. Birch, D. Charalampopoulos, S.B. Astley, R. Pegg, P. Zhou, P. Finglas, Antioxidant activity, total phenolics and flavonoids contents: Should we ban in vitro screening methods?, Food Chemistry, 264 (2018) 471-475.

[10] F. Shahidi, M. Naczk, Phenolics in Food and Nutraceuticals, CRC Press, 2003.

[11] R. Tsao, Chemistry and Biochemistry of Dietary Polyphenols, Nutrients, 2 (2010) 1231-1246.

[12] Z. Charrouf, M. Hilali, O. Jauregui, M. Soufiaoui, D. Guillaume, Separation and characterization of phenolic compounds in argan fruit pulp using liquid chromatography–negative electrospray ionization tandem mass spectroscopy, Food Chemistry, 100 (2007) 1398-1401.

[13] F. Khallouki, R. Haubner, I. Ricarte, G. Erben, K. Klika, C.M. Ulrich, R.W. Owen, Identification of polyphenolic compounds in the flesh of Argan (Morocco) fruits, Food Chemistry, 179 (2015) 191-198.

[14] F. Khallouki, J. Voggel, A. Breuer, K.D. Klika, C.M. Ulrich, R.W. Owen, Comparison of the major polyphenols in mature Argan fruits from two regions of Morocco, Food Chemistry, 221 (2017) 1034-1040.

[15] S.D. Čolić, M.M.F. Akšić, K.B. Lazarević, G.N. Zec, U.M. Gašić, D.Č.D. Zagorac, M.M. Natić, Fatty acid and phenolic profiles of almond grown in Serbia, Food chemistry, 234 (2017) 455-463.

[16] B. Mehari, M. Redi-Abshiro, B.S. Chandravanshi, S. Combrinck, M. Atlabachew, R. McCrindle, Profiling of phenolic compounds using UPLC–MS for determining the geographical origin of green coffee beans from Ethiopia, Journal of Food Composition and Analysis, 45 (2016) 16-25.

[17] A. Bajoub, A. Carrasco-Pancorbo, N. Ouazzani, A. Fernández-Gutiérrez, Potential of LC–MS phenolic profiling combined with multivariate analysis as an approach for the determination of the geographical origin of north Moroccan virgin olive oils, Food chemistry, 166 (2015) 292-300.

[18] R.N. D'Souza, S. Grimbs, B. Behrends, H. Bernaert, M.S. Ullrich, N. Kuhnert, Origin-based polyphenolic fingerprinting of Theobroma cacao in unfermented and fermented beans, Food Research International, 99 (2017) 550-559.

[19] J. Guo, Y. Yuan, P. Dou, T. Yue, Multivariate statistical analysis of the polyphenolic constituents in kiwifruit juices to trace fruit varieties and geographical origins, Food chemistry, 232 (2017) 552-559.

[20] L. Jaitz, K. Siegl, R. Eder, G. Rak, L. Abranko, G. Koellensperger, S. Hann, LC–MS/MS analysis of phenols for classification of red wine according to geographic origin, grape variety and vintage, Food Chemistry, 122 (2010) 366-372.

667 [21] J.-L. Wolfender, S. Rudaz, Y. Hae Choi, H. Kyong Kim, Plant metabolomics: from holistic data to
668 relevant biomarkers, Current Medicinal Chemistry, 20 (2013) 1056-1090.
669 [22] E. Cubero-Leon, R. Penalver, A. Maquet, Review on metabolomics for food authentication, Food
670 Research International, 60 (2014) 95-107.
671 [23] J.G. Bundy, M.P. Davey, M.R. Viant, Environmental metabolomics: a critical review and future
672 perspectives, Metabolomics, 5 (2009) 3.
673 [24] N. Vinayavekhin, E.A. Homan, A. Saghatelian, Exploring disease through metabolomics, ACS
674 Chemical Biology, 5 (2010) 91-103.
675 [25] E.Y. Xu, W.H. Schaefer, Q. Xu, Metabolomics in pharmaceutical research and development:
676 metabolites, mechanisms and pathways, Current Opinion in Drug Discovery Development, 12 (2009)
677 40-52.
678 [26] E. Gorrochategui, J. Jaumot, R. Tauler, A protocol for LC-MS metabolomic data processing using
679 chemometric tools, Protocol Exchange, (2015).
680 [27] C.A. Smith, E.J. Want, G. O'Maille, R. Abagyan, G. Siuzdak, XCMS: processing mass spectrometry
681 data for metabolite profiling using nonlinear peak alignment, matching, and identification, Analytical
682 chemistry, 78 (2006) 779-787.
683 [28] J. Xia, D.S. Wishart, Using MetaboAnalyst 3.0 for comprehensive metabolomics data analysis,
684 Current protocols in bioinformatics, 55 (2016) 14.10. 11-14.10. 91.
685 [29] A. Lommen, MetAlign: interface-driven, versatile metabolomics tool for hyphenated full-scan
686 mass spectrometry data preprocessing, Analytical Chemistry, 81 (2009) 3079-3086.
687 [30] F. Olivon, G. Grelier, F. Roussi, M. Litaudon, D. Touboul, MZmine 2 data-preprocessing to enhance
688 molecular networking reliability, Analytical Chemistry, 89 (2017) 7836-7840.
689 [31] H. Zheng, M.R. Clausen, T.K. Dalsgaard, G. Mortensen, H.C. Bertram, Time-saving design of
690 experiment protocol for optimization of LC-MS data processing in metabolomic approaches, Analytical
691 Chemistry, 85 (2013) 7109-7116.
692 [32] M. Thompson, S.L. Ellison, R. Wood, Harmonized guidelines for single-laboratory validation of
693 methods of analysis (IUPAC Technical Report), Pure Applied Chemistry, 74 (2002) 835-855.
694 [33] E. Guide, The Fitness for Purpose of Analytical Methods–A Laboratory Guide to Method Validation
695 and Related Topics, ed. B. Magnusson and U. Örnemark, Eurachem, 2nd edn, 2014.
696 [34] G. Libiseller, M. Dvorzak, U. Kleb, E. Gander, T. Eisenberg, F. Madeo, S. Neumann, G. Trausinger,
697 F. Sinner, T. Pieber, IPO: a tool for automated optimization of XCMS parameters, BMC Bioinformatics,
698 16 (2015) 118.
699 [35] M. Farrés, B. Piña, R. Tauler, Chemometric evaluation of Saccharomyces cerevisiae metabolic
700 profiles using LC–MS, Metabolomics, 11 (2015) 210-224.
701 [36] R.J.C. Tauler, i.l. systems, Multivariate curve resolution applied to second order data,
702 Chemometrics Intelligent Laboratory Systems 30 (1995) 133-146.
703 [37] J. Jaumot, A. de Juan, R. Tauler, MCR-ALS GUI 2.0: new features and applications, Chemometrics
704 Intelligent Laboratory Systems 140 (2015) 1-12.
705 [38] E. Gorrochategui, J. Jaumot, R. Tauler, ROIMCR: a powerful analysis strategy for LC-MS
706 metabolomic datasets, BMC Bioinformatics, 20 (2019) 256.
707 [39] E. Gorrochategui, J. Jaumot, S. Lacorte, R. Tauler, Data analysis strategies for targeted and
708 untargeted LC-MS metabolomic studies: overview and workflow, TrAC Trends in Analytical Chemistry,
709 82 (2016) 425-442.
710 [40] E. Borràs, J. Ferré, R. Boqué, M. Mestres, L. Aceña, O. Busto, Data fusion methodologies for food
711 and beverage authentication and quality assessment–A review, Analytica Chimica Acta, 891 (2015) 1-
712 14.
713 [41] D. Ballabio, V. Consonni, Classification tools in chemistry. Part 1: linear models. PLS-DA, Analytical
714 Methods, 5 (2013) 3790-3798.
715 [42] M. Kharbach, R. Kamal, M.A. Mansouri, I. Marmouzi, J. Viaene, Y. Cherrah, K. Alaoui, J.
716 Vercammen, A. Bouklouze, Y. Vander Heyden, Selected-ion flow-tube mass-spectrometry (SIFT-MS)
717 fingerprinting versus chemical profiling for geographic traceability of Moroccan Argan oils, Food
718 chemistry, 263 (2018) 8-17.

[43] F. Westad, F. Marini, Validation of chemometric models–a tutorial, Analytica Chimica Acta, 893 (2015) 14-24.
[44] L.B. Rojas, S. Quideau, P. Pardon, Z. Charrouf, Colorimetric evaluation of phenolic content and GC-MS characterization of phenolic composition of alimentary and cosmetic argan oil and press cake, Journal of Agricultural Food Chemistry 53 (2005) 9122-9127.
[45] K.D. Klika, F. Khallouki, R.W. Owen, Amino phenolics from the fruit of the argan tree Argania spinosa (Skeels L.), Zeitschrift für Naturforschung C, 69 (2014) 363-367.
[46] K.D. Klika, F. Khallouki, R.W. Owen, Carboxy methyl and carboxy analogs argaminolics B and C, Rec. Nat. Prod, 9 (2015) 597-602.
[47] F.Fernandes de Araújo, D. de Paulo Farias, I.A. Neri-Numa, G.M. Pastore, Polyphenols and their applications: An approach in food chemistry and innovation potential, Food chemistry, 338 (2021) 127535.

Figure legends

Fig.1. Typical UPLC–ESI–QTOF/MS chromatograms (Total Ion Current (TIC)) for AK samples from (A) Agadir, (B) Ait-Baha, (C) Essaouira, (D) Taroudant, (E) Tiznit. Experimental conditions: see text.

Fig. 2. Comparison of the secondary-metabolites content (mg/kg) distribution in Argan kernels from different geographical origins (Agadir, Ait-Baha, Essaouira, Taroudant and Tiznit).

Fig.3. Untargeted metabolomics approaches and chemometric data handling strategies.

Fig.4. 3D-PCA score plots for the (A) XCMS, (B) MCR-ALS, (C) LL-DF, and (D) ML-DF datasets.

Fig.5. PLS-DA plots, y-predicted versus the sample number for the models where one class is discriminated relative to the others, while the Q residual vs T^2 Hoteling used for data homogeneity, from (A) XCMS, (B) MCR-ALS, (C) LL-DF, and (D) ML-DF datasets, left column (a, c, e, j): calibration; right column (b, d, f, h): validation.

Figure 1

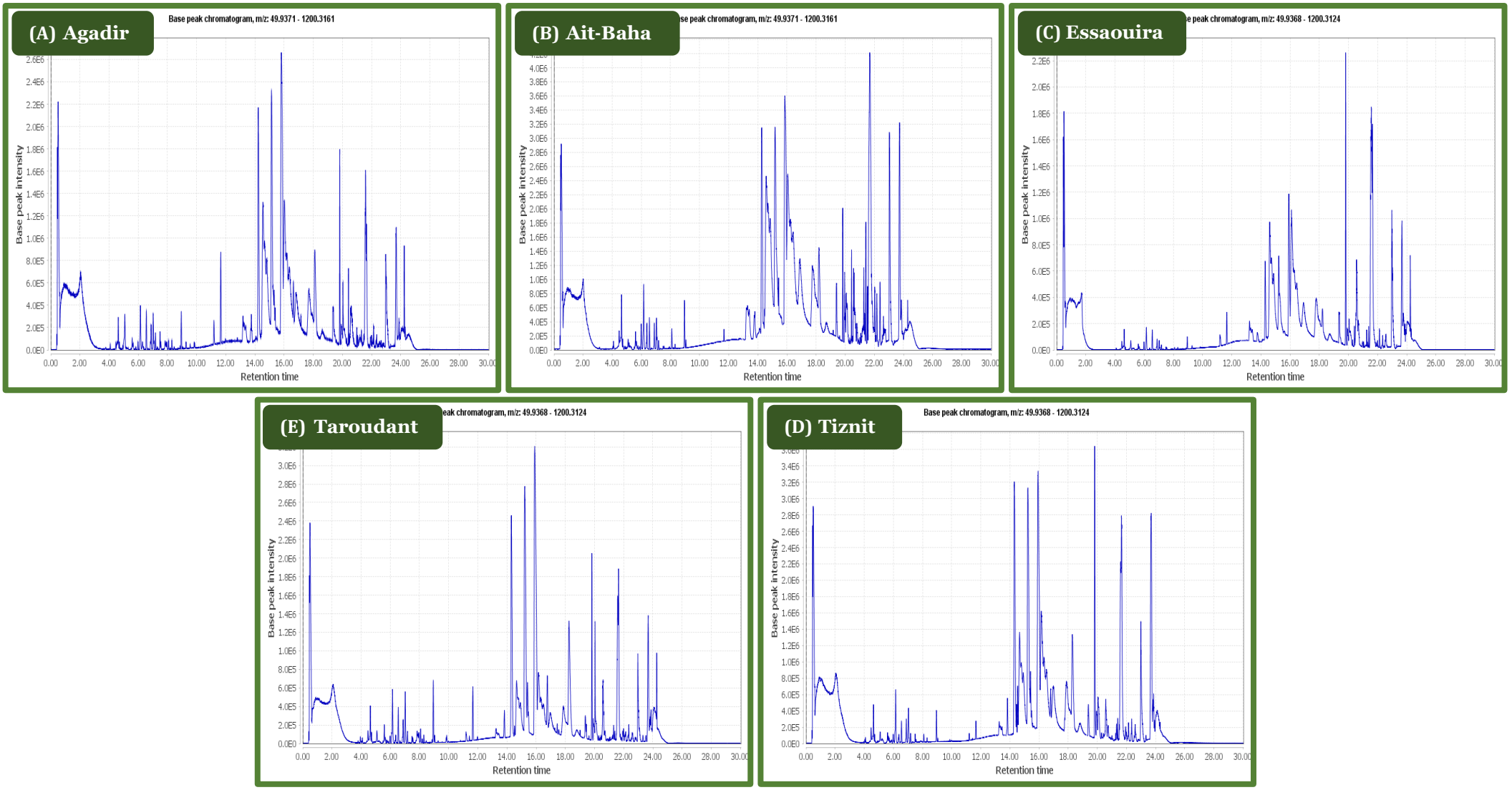
(A) Agadir
Base peak chromatogram, m/z: 49.9371 - 1200.3161
(B) Ait-Baha
Base peak chromatogram, m/z: 49.9371 - 1200.3161
(C) Essaouira
Base peak chromatogram, m/z: 49.9368 - 1200.3124
(E) Taroudant
Base peak chromatogram, m/z: 49.9368 - 1200.3124
(D) Tiznit
Base peak chromatogram, m/z: 49.9368 - 1200.3124
Base peak intensity
Retention time

Figure 2

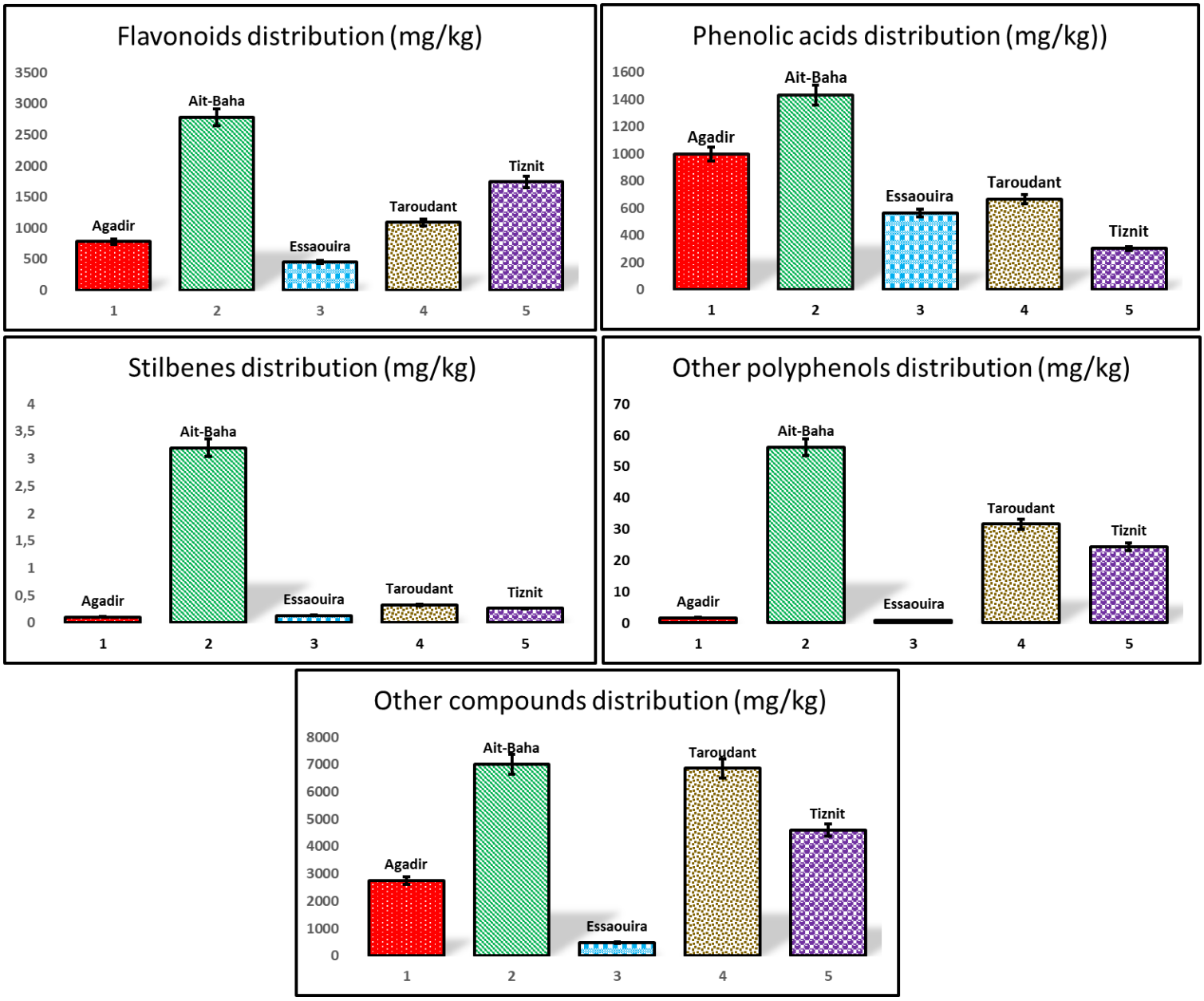
Flavonoids distribution (mg/kg)
3500
3000
2500
2000
1500
1000
500
0
Agadir
Ait-Baha
Essaouira
Taroudant
Tiznit
1
2
3
4
5
Phenolic acids distribution (mg/kg))
1600
1400
1200
1000
800
600
400
200
0
Agadir
Ait-Baha
Essaouira
Taroudant
Tiznit
1
2
3
4
5
Stilbenes distribution (mg/kg)
4
3,5
3
2,5
2
1,5
1
0,5
0
Agadir
Ait-Baha
Essaouira
Taroudant
Tiznit
1
2
3
4
5
Other polyphenols distribution (mg/kg)
70
60
50
40
30
20
10
0
Agadir
Ait-Baha
Essaouira
Taroudant
Tiznit
1
2
3
4
5
Other compounds distribution (mg/kg)
8000
7000
6000
5000
4000
3000
2000
1000
0
Agadir
Ait-Baha
Essaouira
Taroudant
Tiznit
1
2
3
4
5

Figure 3

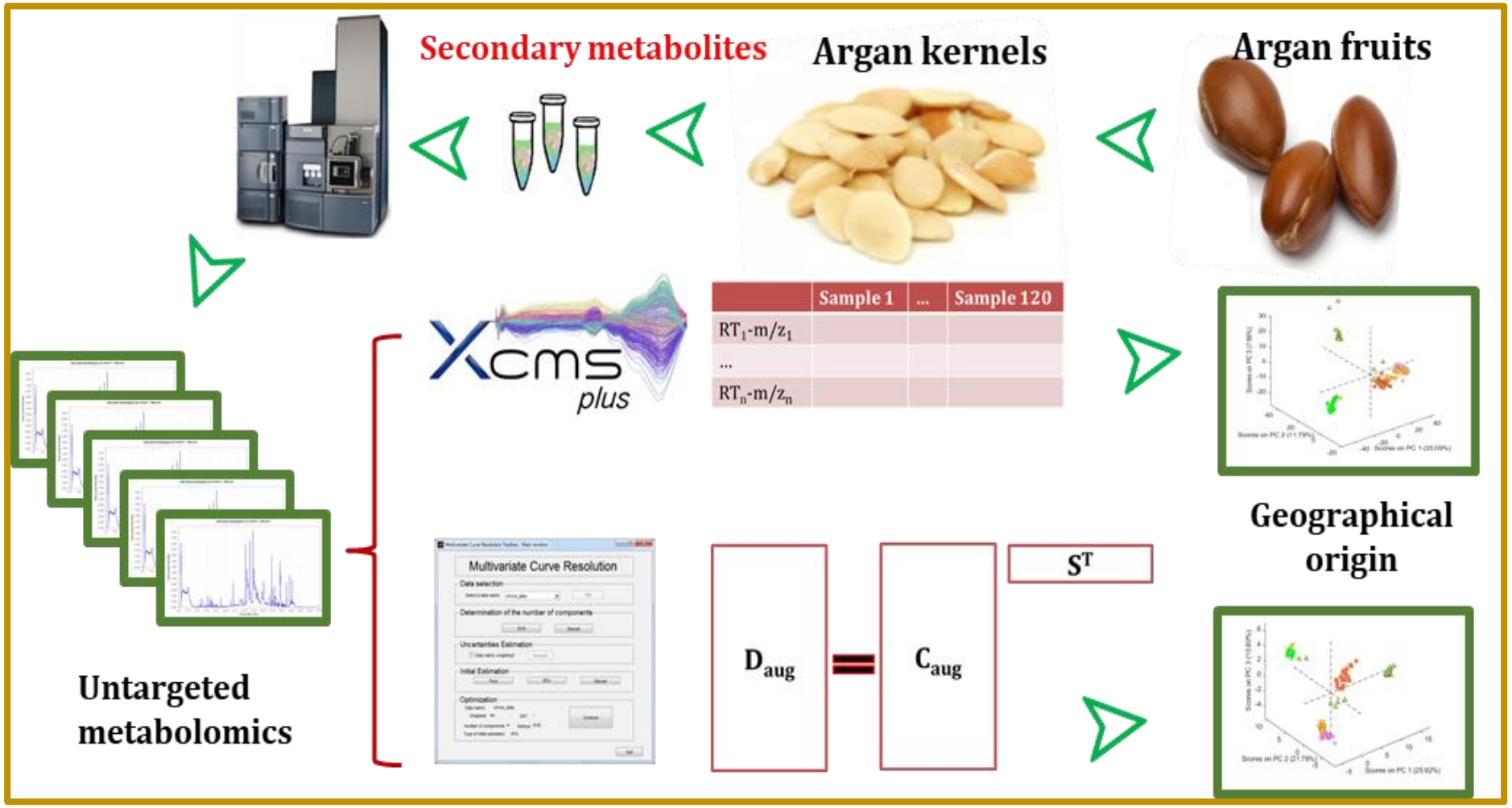
Secondary metabolites
Argan kernels
Argan fruits
Sample 1
...
Sample 120
RT1-m/z1
...
RTn-m/zn
Xcms plus
Multivariate Curve Resolution
Data selection
Determination of the number of components
Uncertainties Estimation
Initial Estimation
Optimization
Daug
Caug
ST
Untargeted metabolomics
Geographical origin

Figure 4

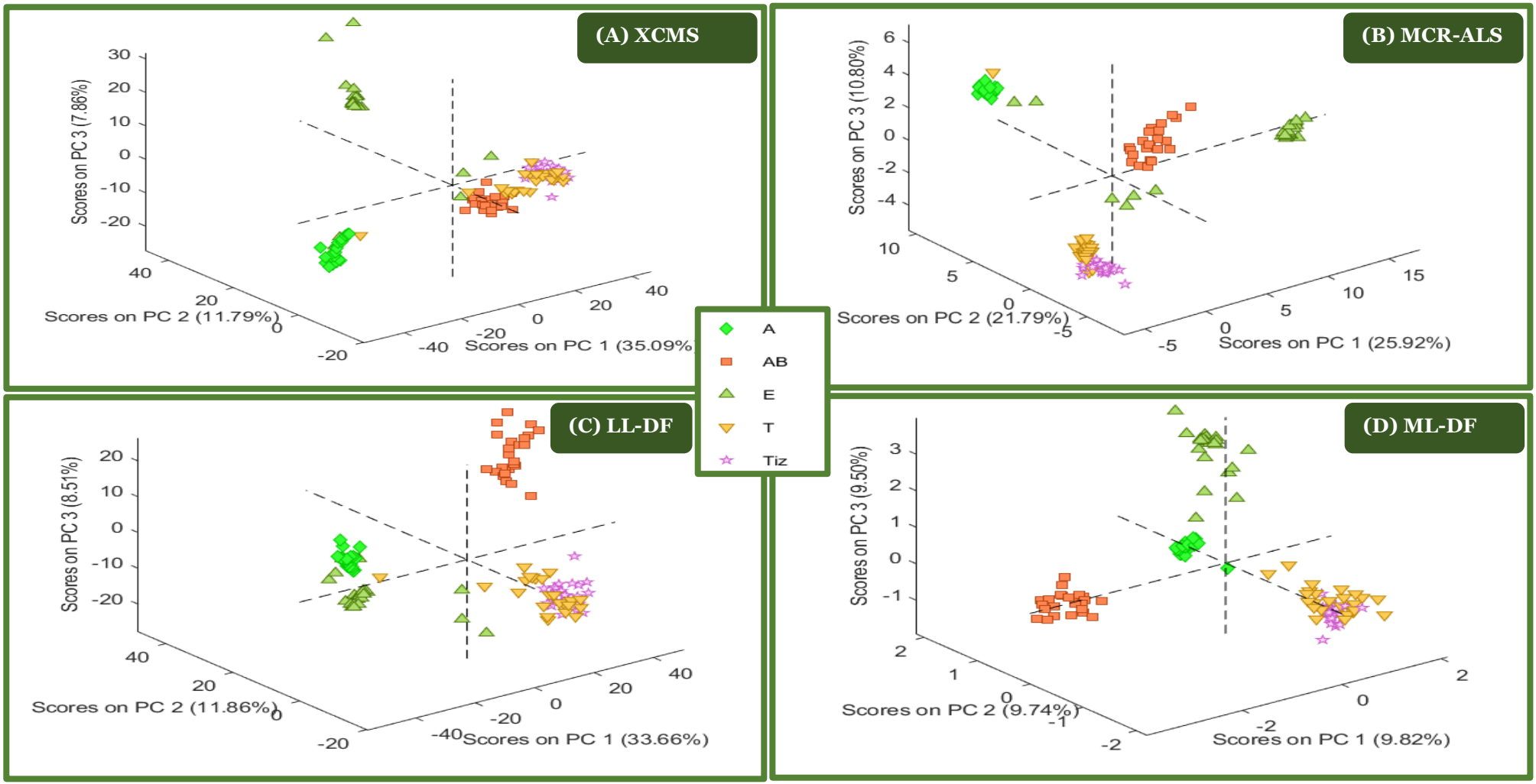
(A) XCMS
Scores on PC 3 (7.86%)
Scores on PC 2 (11.79%)
Scores on PC 1 (35.09%)
(B) MCR-ALS
Scores on PC 3 (10.80%)
Scores on PC 2 (21.79%)
Scores on PC 1 (25.92%)
(C) LL-DF
Scores on PC 3 (8.51%)
Scores on PC 2 (11.86%)
Scores on PC 1 (33.66%)
(D) ML-DF
Scores on PC 3 (9.50%)
Scores on PC 2 (9.74%)
Scores on PC 1 (9.82%)
A
AB
E
T
Tiz

Figure 5

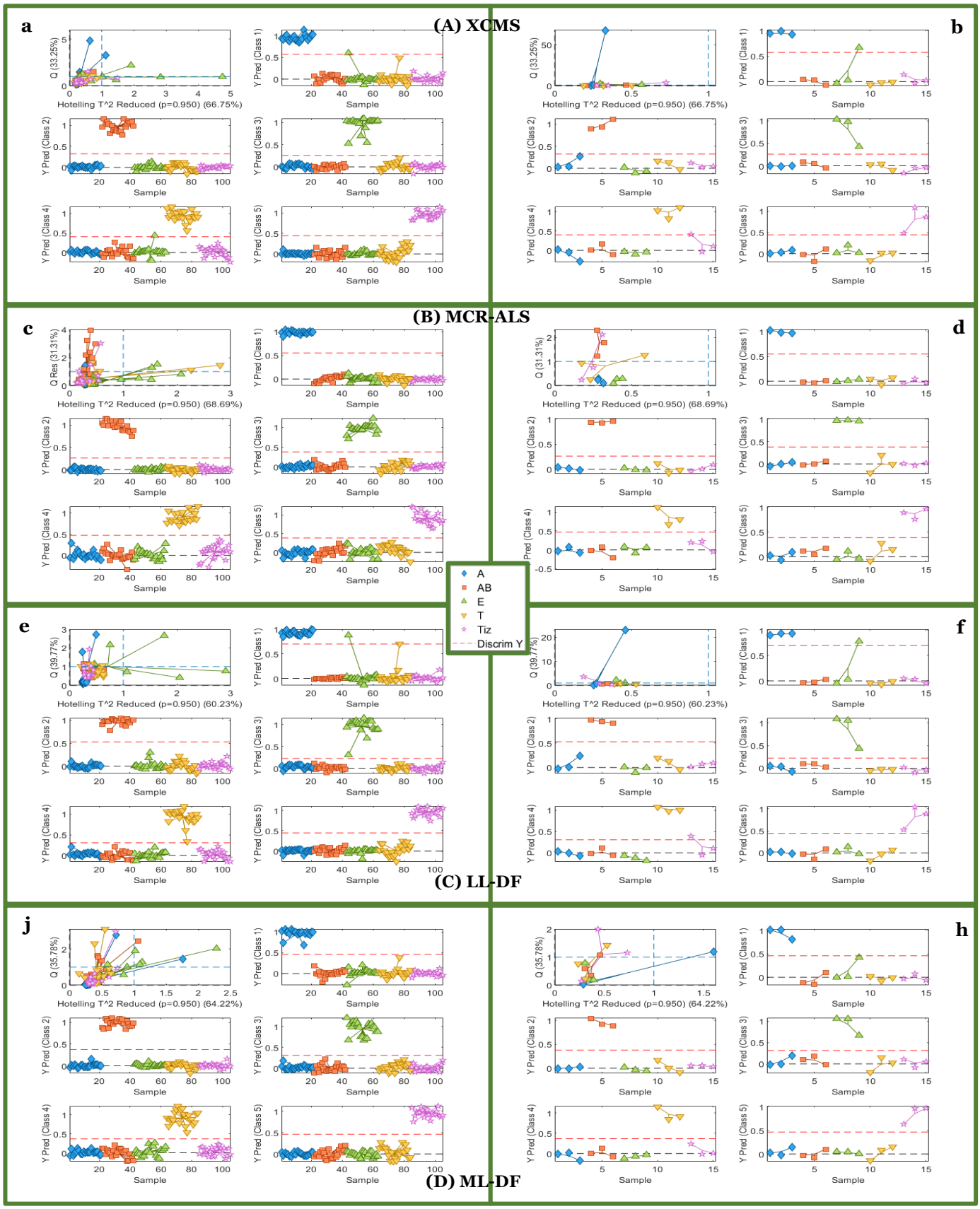
a
(A) XCMS
b
Q (33.25%)
Hotelling T^2 Reduced (p=0.950) (66.75%)
Y Pred (Class 1)
Y Pred (Class 2)
Y Pred (Class 3)
Y Pred (Class 4)
Y Pred (Class 5)
Sample
c
(B) MCR-ALS
d
Q Res (31.31%)
Q (31.31%)
Hotelling T^2 Reduced (p=0.950) (68.69%)
A
AB
E
T
Tiz
Discrim Y
e
(C) LL-DF
f
Q (39.77%)
Hotelling T^2 Reduced (p=0.950) (60.23%)
j
(D) ML-DF
h
Q (35.78%)
Hotelling T^2 Reduced (p=0.950) (64.22%)

Table 1. Secondary-metabolites composition (mean ± standard of deviation) in mg/kg, quantified by UPLC-ESI-QTOF/MS for 120 Moroccan Argan kernels. N = 24 per class.

	Geographic class				
Compounds	**Agadir**	**Ait-Baha**	**Essaouira**	**Taroudant**	**Tiznit**
Quinic acid	2600 ± 100 [b]	6700 ± 300 [d]	450 ± 20 [a]	**6700 ± 300** [d]	4400 ± 200 [c]
Citric acid	85 ± 4 [b]	**260 ± 10** [e]	19 ± 1 [a]	97 ± 5 [c]	138 ± 7 [d]
Pyrogallol	0.35 ± 0.02 [b]	**5.2 ± 0.3** [e]	0.15 ± 0.01 [a]	4.1 ± 0.2 [d]	2.6 ± 0.1 [c]
Succinic acid	**25 ± 1** [e]	7.7 ± 0.4 [c]	17.7 ± 0.9 [d]	7.2 ± 0.4 [b]	4.4 ± 0.2 [a]
Gallic acid	**80 ± 2** [e]	44 ± 2 [b]	75 ± 2 [d]	61 ± 3 [c]	38 ± 2 [a]
Chlorogenic acid	0.20 ± 0.01 [a]	**2.8 ± 0.1** [e]	0.34 ± 0.02 [b]	1.3 ± 0.1 [c]	1.6 ± 0.1 [d]
Pyrocatechol	1.03 ± 0.05 [b]	**48 ± 2** [e]	0.52 ± 0.03 [a]	27 ± 1 [d]	21 ± 1 [c]
Protocatechuic acid	910 ± 20 [d]	**1020 ± 20** [e]	480 ± 10 [b]	590 ± 20 [c]	250 ± 10 [a]
p-Hydroxybenzoic acid	1.66 ± 0.08 [b]	**16.9 ± 0.8** [e]	0.49 ± 0.02 [a]	3.1 ± 0.2 [c]	4.0 ± 0.2 [d]
Catechin	5.8 ± 0.3 [d]	**7.0 ± 0.4** [e]	0.76 ± 0.04 [a]	0.97 ± 0.05 [b]	4.0 ± 0.2 [c]
Epicatechin	**5.4 ± 0.3** [e]	5.0 ± 0.2 [d]	0.59 ± 0.03 [a]	0.92 ± 0.05 [b]	3.6 ± 0.12 [c]
m-Hydroxybenzoic acid	0.08 ± 0.01 [b]	0.20 ± 0.01 [d]	0.06 ± 0.01 [a]	0.13 ± 0.01 [c]	**0.22 ± 0.01** [e]
Epigallocatechin 3-O-gallate	60 ± 3 [a]	**1520 ± 80** [e]	147 ± 7 [b]	260 ± 10 [c]	340 ± 20 [d]
Vanillic acid	1.48 ± 0.07 [c]	**156.32 ± 7.82** [e]	0.660 ± 0.030 [a]	2.79 ± 0.14 [d]	1.32 ± 0.06 [b]
p-Coumaric acid	0.31 ± 0.02 [b]	**1.55 ± 0.08** [e]	0.15 ± 0.01 [a]	1.0 ± 0.05 [d]	0.55 ± 0.03 [c]
Rutin	82 ± 4 [b]	179 ± 8 [d]	15 ± 1 [a]	132 ± 7 [c]	**280 ± 10** [e]
m-Coumaric acid	0.011 ± 0.001 [b]	0.021 ± 0.001 [d]	0.008 ± <0.0005 [a]	0.017 ± <0.0005 [c]	**0.035 ± <0.0005** [e]
Ferulic acid	0.072 ± 0.001 [a]	**0.33 ± 0.02** [d]	0.22 ± 0.01 [c]	0.14 ± 0.01 [b]	0.23 ± 0.01 [c]
Quercitrin	0.004 ± <0.0005 [b]	**0.014 ± <0.0005** [e]	0.002 ± <0.0005 [a]	0.007 ± <0.0005 [c]	0.008 ± <0.0005 [d]
o-Coumaric acid	0.008 ± 0.001 [b]	0.032 ± <0.0005 [d]	0.007 ± <0.0005 [a]	0.020 ± 0.00 [c]	**0.035 ± <0.0005** [e]
Salicylic acid	0.60 ± 0.03 [a]	**188 ± 9** [e]	1.47 ± 0.07 [b]	6.1 ± 0.3 [d]	4.3 ± 0.2 [c]

Naringin	0.04 ± 0.001 [a]	**0.69 ± 0.04** [e]	0.25 ± 0.01 [b]	0.48 ± 0.02 [c]	0.62 ± 0.03 [d]
Hesperidin	0.670 ± 0.03 [b]	**22 ± 1** [e]	0.49 ± 0.02 [a]	1.78 ± 0.09 [d]	0.77 ± 0.04 [c]
Luteolin	0.002 ± <0.0005 [a]	**0.030 ±** <0.0005 [c]	0.002 ± <0.0005 [a]	0.005 ± <0.0005 [b]	0.002 ± <0.0005 [a]
Resveratrol	0.10 ± 0.001 [a]	**3.2 ± 0.2** [e]	0.13 ± 0.001 [b]	0.33 ± 0.02 [d]	0.264 ± 0.010 [c]
Quercitin	15.12 ± 0.70 [c]	**50.545 ± 2.80** [e]	16.15 ± 0.71 [d]	6.78 ± 0.30 [a]	9.01 ± 0.80 [b]
Hesperetin	0.004 ± <0.0005 [a]	**0.019 ± 0.001** [e]	0.006 ± <0.0005 [b]	0.012 ± 0.001 [d]	0.011 ± 0.001 [c]
4-Hydroxycoumarin	0.27 ± 0.01 [b]	**3.0 ± 0.2** [e]	0.18 ± 0.01 [a]	0.43 ± 0.02 [c]	0.54 ± 0.03 [d]
Quercetin glycogallate	9.88 ± 0.78 [d]	**15.43 ± 0.87** [e]	5.25 ± 0.41 [c]	2.21 ± 0.14 [b]	1.10 ± 0.12 [a]
Hyperin	102 ± 5 [b]	270 ± 20 [d]	37 ± 2 [a]	167 ± 9 [c]	**330 ± 10** [e]
Isoquercitrin	121 ± 6 [d]	115 ± 6 [c]	49 ± 3 [a]	98 ± 5 [b]	**230 ± 10** [e]
Avicularin	290 ± 10 [c]	285 ± 10 [c]	65 ± 3 [a]	260 ± 10 [b]	**310 ± 10** [e]
Quercetin glycohydroxybenzoate	20 ± 1 [a]	**113 ± 6** [e]	33 ± 2 [b]	85 ± 4 [c]	101 ± 5 [d]
Quercetin glycosinapate	28 ± 1 [d]	**69 ± 3** [e]	25 ± 1 [c]	14.6 ± 0.9 [a]	22 ± 1 [b]
Quercetin glycoferulate	32 ± 1 [c]	47 ± 3 [d]	19.2 ± 0.7 [b]	12.3 ± 0.7 [a]	**56 ± 1** [e]
Quercetin glycocoumarate	7.1 ± 0.6 [a]	**77 ± 4** [e]	34 ± 1 [b]	43 ± 2 [c]	60 ± 3 [d]
Σ Flavonoids (mg/kg)	780 ± 40 [b]	**2800 ± 100** [e]	450 ± 22 [a]	1090 ± 50 [c]	1740 ± 90 [d]
Σ Phenolic acids (mg/kg)	1000 ± 50 [d]	**1430 ± 70** [e]	560 ± 30 [b]	670 ± 30 [c]	300 ± 20 [a]
Σ Stilbenes (mg/kg)	0.10 ± 0.01 [a]	**3.2 ± 0.2** [e]	0.13 ± 0.01 [b]	0.33 ± 0.02 [d]	0.26 ± 0.01 [c]
Σ Other polyphenols (mg/kg)	1.65 ± 0.08 [b]	**56 ± 3** [e]	0.85 ± 0.04 [a]	32 ± 2 [d]	24 ± 1 [c]
Σ Other compounds (mg/kg)	2800 ± 100 [b]	**7000 ± 400** [d]	484 ± 24 [a]	6900 ± 300 [d]	4600 ± 200 [c]

The lowercase letters indicate significant differences in the same line (comparison between the classes, Tukey's test). (a: lowest → e: highest).

The bold numbers represent the highest content of a compound in the five classes.

Table 2. Classification parameters obtained by PLS-DA and SIMCA models using data matrices from (A) XCMS, (B) MCR-ALS, (C) low-level data fusion, and (D) mid-level data fusion.

A) XCMS

	PLS-DA					SIMCA						Sensitivity		Specifity		Accuracy	
Category	1	2	3	4	5	1	2	3	4	5		PLS-DA	SIMCA	PLS-DA	SIMCA	PLS-DA	SIMCA
Training set																	
1 Agadir (21)	21	0	0	0	0	21	0	0	0	0		100	100	100	100	100	100
2 Ait-Baha (21)	0	21	0	0	0	0	21	0	0	0		100	100	100	100	100	100
3 Essaouira (21)	0	0	21	0	0	0	0	21	0	0		100	100	100	100	100	100
4 Taroudant (21)	0	0	0	21	0	0	0	0	21	0		100	100	100	100	100	100
5 Tiznit (21)	0	0	0	0	21	0	0	0	0	21		100	100	100	100	100	100
											Overall rate	100	100	100	100	100	100
Prediction set																	
1 Agadir (3)	3	0	0	0	0	3	2	2	0	2		100	100	100	50	100	55
2 Ait-Baha (3)	0	3	0	0	0	0	1	0	0	0		100	33	100	100	100	83
3 Essaouira (3)	0	0	3	0	0	0	0	1	0	0		100	33	100	100	100	83
4 Taroudant (3)	0	0	0	3	0	0	0	0	3	0		100	100	100	100	100	100
5 Tiznit (3)	0	0	0	0	3	0	0	0	0	1		100	33	100	100	100	83
											Overall rate	**100**	60	**100**	90	**100**	81

B) MCR-ALS

	PLS-DA					SIMCA						Sensitivity		Specifity		Accuracy	
Category	1	2	3	4	5	1	2	3	4	5		PLS-DA	SIMCA	PLS-DA	SIMCA	PLS-DA	SIMCA
Training set																	
1 Agadir (21)	21	0	0	0	0	21	0	0	0	0		100	100	100	100	100	100
2 Ait-Baha (21)	0	21	0	0	0	0	21	0	0	0		100	100	100	100	100	100
3 Essaouira (21)	0	0	21	0	0	0	0	21	0	0		100	100	100	100	100	100
4 Taroudant (21)	0	0	0	21	0	0	0	0	21	0		100	100	100	100	100	100
5 Tiznit (21)	0	0	0	0	21	0	0	0	0	21		100	100	100	100	100	100
											Overall rate	**100**	**100**	**100**	**100**	**100**	**100**
Prediction set																	
1 Agadir (3)	3	0	0	0	0	3	0	0	0	0		100	100	100	100	100	100
2 Ait-Baha (3)	0	3	0	0	0	0	3	0	0	0		100	100	100	100	100	100
3 Essaouira (3)	0	0	3	0	0	0	0	3	0	0		100	100	100	100	100	100
4 Taroudant (3)	0	0	0	3	0	0	0	0	3	0		100	100	100	100	100	100
5 Tiznit (3)	0	0	0	0	3	0	0	0	0	3		100	100	100	100	100	100
											Overall rate	**100**	**100**	**100**	**100**	**100**	**100**

C) Data fusion (low-level)

	PLS-DA					SIMCA						Sensitivity		Specificity		Accuracy	
Category	1	2	3	4	5	1	2	3	4	5		PLS-DA	SIMCA	PLS-DA	SIMCA	PLS-DA	SIMCA
Training set																	
1 Agadir (21)	21	0	1	0	0	21	0	0	0	0		100	100	99	100	100	99
2 Ait-Baha (21)	0	21	0	0	0	0	21	0	0	0		100	100	100	100	100	100
3 Essaouira (21)	0	0	20	0	0	0	0	21	0	0		95	100	100	100	100	100
4 Taroudant (21)	0	0	0	21	0	0	0	0	21	0		100	100	100	100	100	100
5 Tiznit (21)	0	0	0	0	21	0	0	0	0	21		100	100	100	100	100	100
											Overall rate	99	**100**	**100**	**100**	**100**	**100**
Prediction set																	
1 Agadir (3)	3	0	0	0	0	3	0	2	0	1		100	100	100	75	100	79
2 Ait-Baha (3)	0	3	0	0	0	0	3	0	0	0		100	100	100	100	100	100
3 Essaouira (3)	0	0	3	0	0	0	0	1	0	0		100	33	100	100	100	83
4 Taroudant (3)	0	0	0	3	0	0	0	0	3	0		100	100	100	100	100	100
5 Tiznit (3)	0	0	0	0	3	0	0	0	0	2		100	67	100	100	100	92
											Overall rate	**100**	80	**100**	95	**100**	91

D) Data fusion (mid-level)

	PLS-DA					SIMCA						Sensitivity		Specifity		Accuracy	
Category	1	2	3	4	5	1	2	3	4	5		PLS-DA	SIMCA	PLS-DA	SIMCA	PLS-DA	SIMCA
Training set																	
1 Agadir (21)	21	0	0	0	0	21	0	0	0	0		100	100	100	100	100	100
2 Ait-Baha (21)	0	21	0	0	0	0	21	0	0	0		100	100	100	100	100	100
3 Essaouira (21)	0	0	21	0	0	0	0	21	0	0		100	100	100	100	100	100
4 Taroudant (21)	0	0	0	21	0	0	0	0	21	0		100	100	100	100	100	100
5 Tiznit (21)	0	0	0	0	21	0	0	0	0	21		100	100	100	100	100	100
											Overall rate	**100**	**100**	**100**	**100**	**100**	**100**
Prediction set																	
1 Agadir (3)	3	0	0	0	0	3	0	0	0	0		100	100	100	100	100	100
2 Ait-Baha (3)	0	3	0	0	0	0	3	0	0	0		100	100	100	100	100	100
3 Essaouira (3)	0	0	3	0	0	0	0	3	0	0		100	100	100	100	100	100
4 Taroudant (3)	0	0	0	3	0	0	0	0	3	0		100	100	100	100	100	100
5 Tiznit (3)	0	0	0	0	3	0	0	0	0	3		100	100	100	100	100	100
											Overall rate	**100**	**100**	**100**	**100**	**100**	**100**

Secondary-metabolites fingerprinting of *Argania spinosa* kernels using liquid chromatography–mass spectrometry and chemometrics, for metabolite identification and quantification as well as for geographic classification

Mourad Kharbach[1,2,3*], Johan Viaene[3], Huiwen Yu[4], Rabie Kamal[5], Ilias Marmouzi[6], Abdelaziz Bouklouze[6] and Yvan Vander Heyden[3**]

[1] Department of Food and Nutrition, P.O. Box 66, 00014, University of Helsinki, Finland

[2] Department of Computer Science, University of Helsinki, 00014 Helsinki, Finland

[3] Department of Analytical Chemistry, Applied Chemometrics and Molecular Modelling, Vrije Universiteit Brussel (VUB), Laarbeeklaan 103, B-1090 Brussels, Belgium

[4] Chemometrics and Analytical Technology, Faculty of Science, University of Copenhagen, Rolighedsvej 26, 1958 Frederiksberg, Denmark

[5] Pharmacodynamy Research Team. Laboratory of Pharmacology and Toxicology, Faculty of Medicine and Pharmacy, University Mohammed V, Rabat, Morocco

[5] Biopharmaceutical and Toxicological Analysis Research Team, Laboratory of Pharmacology and Toxicology, Faculty of Medicine and Pharmacy, University Mohammed V, Rabat, Morocco

*Corresponding author: Mourad Kharbach, E-mail address mourad.kharbach@hotmail.fr

**Corresponding author: Yvan Vander Heyden, E-mail address yvanvdh@vub.be

Supplementary Material

Table S1. Geographical parameters and provenances of the *Argania spinosa* trees with climatic characteristics.

Geographical origin	**Specific region**	**Sample size**	Average temperature (°C)		Annual rainfall (mm)	Bioclimate
			Max	Min		
Agadir	Drarga	12	25	13	230	Hot humid and arid bioclimate with Mediterranean oceanic influence
	Tighanimine	12				
Ait-Baha	Aday	8	23	16	300	Arid Mediterranean bioclimate
	Biougra	8				
	Tafraoute	8				
Essaouira	Sidi Kawki	8	23	14	315	Hot humid and arid bioclimate with Mediterranean oceanic influence
	Smimou	8				
	Tidzi	8				
Taroudant	Aoulouz	8	26	14	220	Arid Mediterranean bioclimate
	Iforire	8				
	Arazane	8				
Tiznit	Lakhsas	12	24	14	175	Hot semi-arid bioclimate
	Tagzou	12				

Table S2. Analytical results with the developed UPLC/TOF-MS method: relative standard deviation (RSD %) inter- and intra-day, limit of detection (LOD) and quantification (LOQ), linearity, calibration curves and coefficient of determination (r^2). Compound arranged according to their retention time. y= peak area, x=concentration.

Compound	RSD_{inter} (%)	RSD_{intra} (%)	LOD (µg/kg)	LOQ (µg/kg)	Linearity (µg/kg)	Calibration curves	r^2
Quinic acid	4.16	5.12	0.580	0.840	LOQ-100	y = 42.45x - 0.47	0.9939
Citric acid	1.17	2.50	0.570	0.772	LOQ-50	y = 1637x - 815	0.9991
Pyrogallol	1.42	4.23	0.452	0.785	LOQ-100	y = 13513x - 516	0.9999
Succinic acid	1.28	2.52	0.532	0.720	LOQ-50	y = 187.8x - 1.9	0.9910
Gallic acid	2.15	3.49	0.420	0.880	LOQ-100	y = 1789x + 6132	0.9925
Chlorogenic acid	4.15	4.30	0.524	0.945	LOQ-100	y = 1817x + 15508	0.9547
Pyrocatechol	4.09	4.17	0.534	0.730	LOQ-50	y = 6817x + 3231	0.9938
Protocatechuic acid	2.24	3.73	0.647	0.957	LOQ-100	y = 6995.6x - 1.1	0.9998
p-Hydroxybenzoic acid	2.19	2.58	0.502	0.894	LOQ-50	y = 10870x + 2354.6	0.9989
Catechin	1.26	2.15	0.495	0.732	LOQ-50	y = 18269x + 7310.4	0.9954
Epicatechin	3.62	4.13	0.380	0.944	LOQ-50	y = 13527x + 1668.5	0.9996
m-Hydroxybenzoic acid	1.32	2.40	0.315	0.791	LOQ-50	y = 1191x - 593.16	0.9991
Epigallocatechin 3-O-gallate	1.13	4.18	0.517	0.985	LOQ-100	y = 2.7782x - 0.18	0.9992
Vanillic acid	1.29	2.76	0.488	0.865	LOQ-100	y = 215.22x + 228.83	0.9993
p-Coumaric acid	2.15	2.32	0.396	0.835	LOQ-50	y = 22759x + 4504.5	0.9992
Rutin	3.85	4.04	0.496	0.840	LOQ-50	y = 10823x + 946	0.9998
m-Coumaric acid	1.53	2.89	0.395	0.882	LOQ-50	y = 7006.6x - 0.8	0.9995
Ferulic acid	2.28	2.77	0.308	0.897	LOQ-50	y = 7772x + 2013	0.9984
Quercitrin	1.94	3.19	0.488	1.125	LOQ-50	y = 15591x - 0.8	0.9993
o-Coumaric acid	1.14	3.73	0.350	0.816	LOQ-50	y = 5988.1x - 3.7	0.9992
Salicylic acid	2.75	3.21	0.280	0.945	LOQ-25	y = 21692x + 22910	0.9697
Naringin	2.95	3.32	0.366	0.786	LOQ-25	y = 22087x - 0.8	0.9991
Hesperidin	3.15	4.51	0.459	0.858	LOQ-25	y = 72.701x - 0.2	0.9952
Luteolin	3.64	3.86	0.350	0.975	LOQ-25	y = 39587x - 0.6	0.9985

Resveratrol acid	2.59	2.61	0.410	1.195	LOQ-50	y = 17613x + 7282	0.9949
Quercitin	3.82	4.90	0.408	0.976	LOQ-100	y = 46001x + 14695	0.9969
Hesperetin	3.19	4.05	0.553	0.915	LOQ-25	y = 7168x - 0.9	0.9995
4-Hydroxycoumarin	2.81	2.88	0.354	0.864	LOQ-50	y = 9793x – 1.0	0.9987

Table S3. Parameters of the secondary-metabolites compounds identified in the studied AK samples by using UPLC–ESI–QTOF/MS. ND= not detected.

Polyphenol Class	Polyphenol Sub-Class	Name (Synonym)	Chemical Formula	Molecular Weight	PubChem ID	UPLC-ESI-QTOF/MS		
						RT (min)	$[M-H]^-$	MS^2 ion fragments (m/z) [c]
Flavonoids	Flavanols	(+)-Catechin [a]	$C_{15}H_{14}O_6$	290.268	9064	1.63	289.064	290, 357, 174, 245
		(-)-Epicatechin [a]	$C_{15}H_{14}O_6$	290.268	72276	1.65	289.064	290, 287, 174
		(-)-Epigallocatechin 3-O-gallate [a]	$C_{22}H_{18}O_{11}$	458.372	65064	2.01	457.078	445, 305, 169, 125
	Flavonols	Avicularin (Quercetin 3-O-arabinoside) [b]	$C_{20}H_{18}O_{11}$	434.350	5481224	14.35	433.252	301, 300
		Hyperin (Quercetin 3-O-galactoside) [b]	$C_{21}H_{20}O_{12}$	464.376	5281643	13.80	463.178	303, 129
		Isoquercitrin (Quercetin 3-O-glucoside) [b]	$C_{21}H_{20}O_{12}$	464.376	5280804	14.01	463.178	303
		Kaempferol [a]	$C_{15}H_{10}O_6$	286.236	5280863	ND	285.040	257, 151, 169, 241
		Quercetin [a]	$C_{15}H_{10}O_7$	302.236	5280343	5.99	301.000	302, 299, 169
		Quercitrin (Quercetin 3-O-rhamnoside) [a]	$C_{21}H_{20}O_{11}$	448.377	5359430	3.52	447.120	301, 300, 255, 179
		Quercetin glycocoumarate [b]	$C_{30}H_{26}O_{14}$	610.132	—	19.75	609.024	463, 301, 300, 271
		Quercetin glycoferulate [b]	$C_{31}H_{28}O_{15}$	640.143	—	19.21	639.056	463, 301, 300, 271
		Quercetin glycohydroxybenzoate [b]	$C_{28}H_{24}O_{14}$	584.117	—	15.83	583.024	463, 301, 300
		Quercetin glycogallate [b]	$C_{28}H_{24}O_{16}$	616.106	—	11.35	615.010	463, 301, 300, 270, 169
		Quercetin glycosinapate [b]	$C_{32}H_{30}O_{16}$	670.153	—	17.53	669.067	463, 301, 300
		Rutin (Quercetin 3-O-rutinoside) [a]	$C_{27}H_{30}O_{16}$	610.518	5280805	2.63	609.100	610, 174, 235
	Flavanones	Hesperidin (Hesperetin 7-O-rutinoside) [a]	$C_{28}H_{34}O_{15}$	610.561	10621	3.95	609.172	672, 301
		Hesperetin [a]	$C_{16}H_{14}O_6$	302.282	72281	6.77	301.015	301, 258, 143
		Naringin [a]	$C_{27}H_{32}O_{14}$	580.535	442428	3.95	579.173	459, 271, 235
		Naringenin [a]	$C_{15}H_{12}O_5$	272.253	439246	ND	271.061	151
	Flavones	Luteolin [a]	$C_{15}H_{10}O_6$	286.236	5280445	5.17	285.040	217, 199
Phenolic acids	Hydroxybenzoic acids	Benzoic acid [a]	$C_7H_6O_2$	122.121	243	ND	121.031	77
		Gallic acid [a]	$C_7H_6O_5$	170.12	370	0.66	168.90	125, 391, 170
		Protocatechuic acid (3,4-Dihydroxybenzoic acid) [a]	$C_7H_6O_4$	154.12	36062	0.95	153.010	153, 109

		Salicylic acid (2-Hydroxybenzoic acid) [a]	$C_7H_6O_3$	138.121	338	3.74	137.025	327, 297, 138
		Syringic acid [a]	$C_9H_{10}O_5$	198.173	10742	ND	197.045	174, 327, 235, 265
		Vanillic acid (4-Hydroxy-3-methoxybenzoic acid) [a]	$C_8H_8O_4$	168.147	8468	2.12	167.036	174, 235, 357, 296
		m-Hydroxybenzoic acid (3-Hydroxybenzoic acid) [a]	$C_7H_6O_3$	138.121	7420	1.78	137.025	174, 235, 138
		p-Hydroxybenzoic acid (4-Hydroxybenzoic acid) [a]	C7H6O3	138.121	135	1.30	137.050	174, 235, 138
	Hydroxycinnamic acids	Caffeic acid[a]	$C_9H_8O_4$	180.157	689043	ND	179.035	135
		Chlorogenic acid (5-Caffeoylquinic acid) [a]	$C_{16}H_{18}O_9$	354.309	12310830	0.83	353.202	174, 191, 235, 207
		Ferulic acid (3-Methoxy-4-Hydroxycinnamic acid) [a]	$C_{10}H_{10}O_4$	194.184	709	3.12	193.050	116, 194, 235, 178
		m-Coumaric acid (3-Hydroxycinnamic acid) [a]	$C_9H_8O_3$	164.158	637541	2.98	163.042	164, 119, 174
		o-Coumaric acid (2-Hydroxycinnamic acid) [a]	$C_9H_8O_3$	164.158	11968	3.65	163.042	164, 119, 174
		p-Coumaric acid (4-Hydroxycinnamic acid) [a]	$C_9H_8O_3$	164.158	322	2.21	163.042	164, 119, 174
		Rosmarinic acid [a]	$C_{18}H_{16}O_8$	360.315	5281792	ND	359.054	360, 313, 179
		Sinapic acid [a]	$C_{11}H_{12}O_5$	224.21	10743	ND	223.061	174, 235, 208
Stilbenes	Stilbenes	Resveratrol[a]	$C_{14}H_{12}O_3$	228.243	445154	5.83	227.072	228, 326, 273
Other polyphenols	Hydroxycoumarins	Esculetin (6,7-Dihydroxycoumarin) [a]	$C_9H_6O_4$	178.141	5281416	ND	177.018	133, 105, 89
		Esculin (esculetin-6-O-glucoside) [a]	$C_{15}H_{16}O_9$	340.282	5281417	ND	339.072	177, 133
	Other polyphenols	Pyrocatechol (1,2-Dihydroxybenzene) [a]	$C_6H_6O_2$	110.111	289	0.94	109.028	271, 233, 210
		Pyrogallol (1,2,3-Trihydroxybenzene) [a]	$C_6H_6O_3$	126.111	1057	0.60	125.024	303, 126
		4-Hydroxycoumarin [a]	$C_9H_6O_3$	162.142	54682930	7.11	161.039	133, 117
Other compounds		Citric acid [a]	$C_6H_8O_7$	192.123	311	0.63	191.102	111, 173
		Quinic acid [a]	$C_7H_{12}O_6$	192.167	6508	0.51	191.120	192, 405, 365
		Succinic acid [a]	$C_4H_6O_4$	118.088	1110	0.65	117.018	73

[a] Compounds identified by comparing MS data and retention times with those of references (standard compounds).
[b] Compounds (tentatively) identified by comparing MS data with literature and online database.
[c] The MS^2 ion fragments were arranged based on relative ionic abundances.