

Unsupervised methods in LC-MS data treatment

Turova, Polina; Styles, Iain; Timashev, Vladimir; Kravets, Konstantin ; Grechnikov, Alexander ; Lyskov, Dmitry ; Samigullin, Tahir ; Podolskiy, Ilya ; Shpigun, Oleg ; Stavrianidi, Andrey

DOI:

[10.1016/j.jpba.2021.114382](https://doi.org/10.1016/j.jpba.2021.114382)

License:

Creative Commons: Attribution-NonCommercial-NoDerivs (CC BY-NC-ND)

Document Version

Peer reviewed version

Citation for published version (Harvard):

Turova, P, Styles, I, Timashev, V, Kravets, K, Grechnikov, A, Lyskov, D, Samigullin, T, Podolskiy, I, Shpigun, O & Stavrianidi, A 2021, 'Unsupervised methods in LC-MS data treatment: application for potential chemotaxonomic markers search', *Journal of Pharmaceutical and Biomedical Analysis*, vol. 206, 114382. <https://doi.org/10.1016/j.jpba.2021.114382>

[Link to publication on Research at Birmingham portal](#)

General rights

Unless a licence is specified above, all rights (including copyright and moral rights) in this document are retained by the authors and/or the copyright holders. The express permission of the copyright holder must be obtained for any use of this material other than for purposes permitted by law.

- Users may freely distribute the URL that is used to identify this publication.
- Users may download and/or print one copy of the publication from the University of Birmingham research portal for the purpose of private study or non-commercial research.
- User may use extracts from the document in line with the concept of 'fair dealing' under the Copyright, Designs and Patents Act 1988 (?)
- Users may not further distribute the material nor use it for the purposes of commercial gain.

Where a licence is displayed above, please note the terms and conditions of the licence govern your use of this document.

When citing, please reference the published version.

Take down policy

While the University of Birmingham exercises care and attention in making items available there are rare occasions when an item has been uploaded in error or has been deemed to be commercially or otherwise sensitive.

If you believe that this is the case for this document, please contact UBIRA@lists.bham.ac.uk providing details and we will remove access to the work immediately and investigate.

1 **Unsupervised methods in LC-MS data treatment:**
2 **application for potential chemotaxonomic markers search**

3
4 Polina Turova^{*a}, Iain Styles^b, Vladimir Timashev^a, Konstantin Kravets^c,
5 Alexander Grechnikov^c, Dmitry Lyskov^d, Tahir Samigullin^e, Ilya Podolskiy^f,
6 Oleg Shpigun^a, Andrey Stavrianidi^a

7
8 ^a*M.V. Lomonosov Moscow State University, Faculty of Chemistry, 1-3 Leninskie Gory,*
9 *Moscow, 119991, Russia*

10 ^b*University of Birmingham, School of Computer Science, Edgbaston, Birmingham B15 2TT,*
11 *United Kingdom*

12 ^c*Vernadsky Institute of Geochemistry and Analytical Chemistry of the Russian Academy of*
13 *Sciences, Kosygina 19, Moscow, 119991, Russia*

14 ^d*M.V. Lomonosov Moscow State University, Faculty of Biology, 1-12 Leninskie Gory, Moscow,*
15 *119234, Russia*

16 ^e*M.V. Lomonosov Moscow State University, Belozersky Institute of Physico-Chemical Biology,*
17 *1-40 Leninskie Gory, 119234, Moscow, Russia*

18 ^f*Bruker Ltd., Pyatnitskaya 50/2 build. 1, Moscow, 119017, Russia*

19
20
21 *Corresponding author. Address: Lomonosov Moscow State University, Faculty of Chemistry, 1-3
22 Leninskie Gory, Moscow, 119991, Russia; email: turova.polina@gmail.com; phone: +79854722433

Abstract

The combination of Liquid Chromatography and Mass Spectrometry (LC-MS) is commonly used to determine and characterize biologically active compounds because of its high resolution and sensitivity. In this work we explore the interpretation of LC-MS data using multivariate statistical analysis algorithms to extract useful chemical information and identify clusters of similar samples. Samples of leaves from 19 plants belonging to the Apiaceae family were analyzed in unified LC conditions by high- and low-resolution mass spectrometry in a wide range scan mode. LC-MS data preprocessing was performed followed by statistical analysis using tensor decomposition in the form of Parallel Factor Analysis (PARAFAC); matrix factorization following tensor unfolding with principal component analysis (PCA), independent component analysis (ICA), non-negative matrix factorization (NMF); or unsupervised feature selection (UFS). The optimal number of components for each of these methods were found and results were compared using four different metrics: silhouette score, Davies-Bouldin index, computational time, number of noisy components. It was found that PCA, ICA and UFS give the best results across the majority of the criteria for both low- and high-resolution data. An algorithm for biomarker signal selection is suggested and 23 potential chemotaxonomic markers were tentatively identified using MS² data. Dendrograms constructed by the methods were compared to the molecular phylogenetic tree by calculating pixel-wise mean square error (MSE). Therefore, the suggested approach can support chemotaxonomic studies and yield valuable chemical information for biomarker discovery.

Keywords

Liquid chromatography, mass spectrometry, machine learning, Apiaceae, multi-way data

1. Introduction

In recent years many approaches for the investigation of plants' taxonomy have been developed. These includes morphological, anatomical and chemotaxonomic classification. Chemotaxonomy is used for the classification of plants on the basis of their chemical composition[1]. The main task of this approach is to search for primary and secondary metabolites and on the basis of their presence or concentration create new classifications and reveal their relation to the molecular phylogeny classification. In previous works it was shown that for the Rutaceae family such markers are coumarins[2]. Coumarins are secondary metabolites which are considered as a chemical defense against predators and their content depends heavily on the growing conditions. In previous works[3] it was shown that some coumarins can appear or disappear from the chemical composition depending on the variety of conditions: geographical origin of the plants; environmental conditions (climate, pollution, light irradiation, etc); physiological variations (stage of development of the plant organ, plant part used, etc.); sample storage conditions and many others. Plants from the Apiaceae family are also rich sources of biologically active compounds such as coumarins and they are useful as food and flavoring and possess diverse pharmacological activities[4]. For the chemotaxonomic markers search it is necessary to use highly precise analytical methods as chromatography, mass spectrometry, nuclear magnetic resonance and complex statistical algorithms.

Liquid chromatography coupled with mass spectrometry (LC-MS) provides rich information about biological samples and is widely used in plant extract analysis. One of the major difficulties in the LC-MS method is that raw data, which is naturally structured as a 3D array, is difficult to interpret manually and automated analysis methods are needed to extract

72 the most important information. In popular metabolomics approaches, “peak picking” software
73 (e.g. MZmine, XCMS) and peak alignment algorithms [5,6] are widely used to reduce the 3D
74 dataset to a set of peaks determined, by some means, to be the most informative. The first
75 general problem in this approach is that some information will inevitably be lost because many
76 peaks are discarded. The second problem is that other methods of analysis which may be more
77 informative have not been fully investigated. They typically involve decompositions of the data
78 into a set of factors which may be more easily interpretable.

79 Data decomposition methods typically belong to two classes[7]. In the first, the 3d array
80 of LC-MS data is treated “as is” and tensor decomposition methods are applied. The most
81 widely used 3D tensor decompositions are Parallel factor analysis (PARAFAC) and Tucker
82 decomposition, both of which decompose a tensor into a set of matrices. These methods have
83 previously been applied to different types of mass spectrometry data [8]. The alternative
84 approach is to unfold the 3D data into a 2D array by reshaping a tensor of size $X \times Y \times Z$ into a
85 $X \times N$ matrix (where $N = Y \times Z$) which can then be factorized using a wide range of techniques
86 for matrix factorization. In LC-MS the dimensions that are combined in reshaping are retention
87 time and m/z values. Tensor unfolding for LC-MS data is not widely described in the literature,
88 but it has been successfully used for mass spectrometry imaging (MSI) data[9]. Unfolding data
89 in this way opens up a much wider range of potential factorization methods, but it has the
90 disadvantage of combining two orthogonal dimensions, which may remove some of the data’s
91 structure and information content.

92 A second choice that must be made is whether subsequent chemometric analysis is
93 supervised or unsupervised. In general, unsupervised techniques are applied (following any
94 necessary preprocessing) when there is no or little prior knowledge about samples; or
95 unobvious patterns are expected to be revealed; or when the goal is to identify which intrinsic
96 (latent) factors are responsible for the greatest variability in the data. Results of unsupervised
97 approaches are therefore typically most suitable for the discovery of markers present in
98 significant concentrations. Lower abundance analytes can most reliably be identified using a
99 supervised knowledge-based approach or by informed selection of specific areas/windows of
100 the dataset[10]. A wide range of unsupervised approaches have been considered in the literature
101 for applications including dimensionality reduction; resolution of samples; biomarker
102 discovery; outlier identification; interference identification[9]. Among the most common
103 unsupervised methods applied to mass spectrometry data are principal component analysis
104 (PCA), independent component analysis (ICA), and non-negative matrix factorization (NMF)
105 [11].

106 There have been only a few attempts to directly compare unsupervised treatment of
107 mass spectrometry data. Different approaches to chemometric analysis of LC-MS data have
108 been compared by classification accuracy, computational time and F1 score [12], but this work
109 used a specific preprocessing protocol in which only the data points with the highest intensity
110 within each peak were retained. In other work[9] different unsupervised treatment methods
111 applied to MALDI imaging MS data were compared. It was shown that NMF and ICA
112 produced components which mapped the spatial distribution of molecules and for which the
113 associated spectra featured lower noise.

114 The aim of the present study was to compare the possibilities of different unsupervised
115 factorization approaches in LC-MS data treatment, and to discover potential chemotaxonomic
116 markers for 19 plant species from Apiaceae family. Data was acquired from the samples on
117 two instruments and two types of data were investigated: LC-MS with low resolution MS
118 (LRMS) and with high resolution MS (HRMS). In both cases raw LC-MS data was recorded
119 in tensor form with three dimensions corresponding to samples, retention time, and mass to
120 charge ratio. Different data factorization techniques were applied to the data using two general
121 approaches: direct decomposition of the 3d tensor, and decomposition of the unfolded tensor

122 (Fig. 1). For direct tensor factorization we used non-negative PARAFAC decomposition. On
123 unfolded tensors, we applied a range of dimensionality reduction and feature selection
124 methods. For dimensionality reduction PCA, ICA and NMF were used; for feature selection
125 variance-based feature filtering was employed.

126 For this research we used a dataset consisting of 57 samples from 19 plants belonging
127 to the Apiaceae family and representing 7 genera: Prangos, Ferulago, Cachrys, Bilacunaria,
128 Diplotaenia, Azilia and Seseli (Table 1). Application of unsupervised methods to such a diverse
129 dataset can reveal the most variable chemotaxonomic markers of this family.

130

131 **2. Experimental**

132 **2.1. Instrumentation**

133 The LC-LRMS apparatus consisted of a HPLC Thermo Scientific Dionex Ultimate
134 3000 (MA, USA) system with a binary analytical pump, an automatic sample injector coupled
135 on-line with AB Sciex Qtrap 3200 (ON, Canada) mass spectrometer with an electrospray
136 ionization interface. The column effluent was analyzed by ESI-MS in positive ion mode and
137 the mass spectra were acquired and processed using the Analyst software (version 1.5)
138 provided by AB Sciex. For the MS, the following conditions were used: ion spray voltage:
139 5500V; ion source heater temperature: 350°C; entrance potential: 10 V; declustering potential:
140 40 V; mass range 100-1200 Da.

141 The LC-HRMS apparatus consisted of a Thermo Scientific Accela HPLC system (CA,
142 USA) coupled on-line with Orbitrap Exactive mass spectrometer (Dreieich, Germany). The
143 column effluent was analyzed by HESI-MS in positive ion mode and the mass spectra were
144 acquired and processed using the Xcalibur™ Software (version 2.2) provided by Thermo
145 Scientific™. For the MS, the following conditions were used: spray voltage: 3.90 kV, capillary
146 temperature: 300 °C, capillary voltage: 50.0 V, tube lens voltage: 100.0 V, skimmer voltage:
147 20 V, heater temperature: 350 °C, resolution 35 000, mass range 100-1200 Da.

148 In both LC-LRMS and LC-HRMS experiments the HPLC separation was conducted on
149 a C18 column (Acclaim RSLC 2.1×150 mm, 2.2 μm) at a flow rate of 0.35 mL/min and oven
150 temperature 35 °C. Two solvents were used: (A) 0.5% HCOOH aqueous solution and (B)
151 MeCN. The gradient was as follows: 0 – 3 min 10 % B; 3 – 20 min linear gradient from 10 to
152 95 % B; 20 – 22 min 95 % B; 22 – 22.2 min linear gradient from 95 to 10 %; 22.2 – 27 min 10
153 % B.

154 Biomarker identification was performed on a Bruker Elute LC system coupled on-line
155 with a Bruker Impact II high-resolution Quadrupole Time-of-Flight Instrument. HPLC
156 separation was conducted on a C18 column (Intensity Solo 1.8 2.1× 100 mm) at a gradient flow
157 rate (from 0.200 to 0.480 mL/min). Two solvents were used: (A) 5 mM Ammonium Formate
158 and 0.01 % FA in MeOH:H₂O 1:99 mixture with and (B) 5 mM Ammonium Formate and 0.01
159 % FA in MeOH. The gradient was as follows: 0 – 0.1 min 4 % B; 0.1 – 1 min linear gradient
160 from 4 to 18.3 %; 1 – 2.5 min linear gradient from 18.3 to 50 % B; 2.5 – 14 min linear gradient
161 from 50 to 99.9 % B; 14 – 16 min 99.9 % B; 16 – 16.1 min linear gradient from 99.9 to 4 % B;
162 16.1 – 20 min 4 % B.

163 **2.2. Materials and reagents**

164 Deionized water was from a Milli-Q system from Millipore (MA, USA); HPLC-grade
165 acetonitrile was purchased from Panreac (Barcelona, Spain) and >99.8% pure ethanol was from
166 Sigma-Aldrich (Steinheim, Germany); Formic acid >99.9% purity was purchased from Acros
167 (Geel, Belgium); MeOH >99.9% purity was purchased from Burdick & Jackson (Seelze,
168 Germany) and Ammonium Formate ≥99.0% purity was purchased from Sigma-Aldrich

169 (Steinheim, Germany). Plant material was collected by botanists from Lomonosov Moscow
170 State University.

171 **2.3. Sample preparation**

172 Plant material was collected in Iran, Portugal, Kyrgyzstan, and Uzbekistan in 2013 –
173 2019 and housed in Moscow University Herbarium (MW) or in the private collection of Dmitry
174 Lyskov (information about herbarium specimens is available at <https://plant.depo.msu.ru/>). All
175 plant specimens used for the analysis are listed in Table 1. Material was dried; extracts were
176 prepared by weighting 0.01 g of a plant sample, adding 1 mL of methanol:water (3:1, v/v)
177 mixture and extracting in an ultrasonic bath for 30 minutes, all extracts were prepared in three
178 replicates. Extracts were centrifuged and diluted by a factor of ten with 10% aqueous
179 acetonitrile. 2 mg/mL solution of eleutheroside B was used as an internal standard (IS) and 20
180 μ L of this solution was added to each sample. For most of the plants, leaf samples were used,
181 but for some samples leaves were missing and stems were used instead.

182 **2.4. Software and packages**

183 All LC-MS files were converted into mzXML format using MSConvert from
184 ProteoWizard Tools. Data analysis was performed in Python 3 using the following modules:
185 pymzML for mzML data files parsing[13]; scipy.signal for signal smoothing; pandas for arrays
186 pretreatment; tensorly for PARAFAC decomposition; scikit-learn for PCA, ICA, NMF, UFS
187 algorithms and performance metrics; matplotlib for data visualization, biopython for
188 hierarchical and molecular phylogenetic trees visualization. Corcondia criteria and explained
189 variance for PARAFAC models were calculated in MATLAB using the N-way toolbox. For
190 dendrogram construction unweighted pair group method with arithmetic mean was used to
191 cluster objects. Minkowski distance was used as a metric to evaluate object similarity. For
192 phylogenetic tree ‘identity’ model for distance calculation was employed. All files from LC-
193 MS analysis in mzML format and implemented algorithms are available at the github
194 repository (<https://github.com/turovapolina/unsupervised-LC-MS-data-treatment>).
195

196 **3. Results and discussion**

197 **3.1. Data acquisition and preparation**

198 The model dataset consisted of 57 samples which are three replicates of 19 plant species
199 which represents 7 genera from Apiaceae family. As a preliminary step, four extraction systems
200 of methanol, water and dichloromethane mixtures were tested to maximize the signal[14]. The
201 methanol:water (75:25, v/v) systems provided maximum peak capacity and the highest
202 intensities in the chromatograms of all samples. Composition of the mobile phase was varied
203 in a wide range during the gradient program in order to elute both polar and non-polar
204 compounds and resolve as many distinct peaks as possible. MS data was collected in scan mode
205 in the range 100–1200 m/z. All samples were analyzed in the same chromatographic conditions
206 by LC-LRMS and LC-HRMS.

207 *LC-LRMS data treatment.* For mass chromatogram smoothing continuous wavelet
208 transform, Baseline Estimation and Denoising With Sparsity (BEADS) approach and Savitzki-
209 Golay filter were assessed with different parameters[15]. The goal was to choose a smoothing
210 algorithm and associated parameters which will work successfully, i.e. smooth as much noise
211 as possible but preserve peak shapes, on all mass chromatograms. In particular, the percentage
212 of acetonitrile in the mobile phase was found to increase noise, and the smoothing algorithm
213 should be capable of removing this noise. The optimal method was found to be a Savitzki-
214 Golay filter with window size 13 and polynomial order 1, results of its implementation for both
215 noisy and informative chromatograms (with or without distinct peaks) are shown in Fig. 2
216 (A,B). The step between time points varies across samples between 0.02 and 0.04 min and the

217 time axis was linearly interpolated with a step size of 0.05 min in order to unify the time axis.
218 A final time scale with a range from 2 to 22 minutes was chosen in order to disregard unretained
219 compounds at the beginning of the chromatogram, very noisy signals at high percentages of
220 acetonitrile and reequilibration time at the end of the chromatogram. For the mass axis
221 unification, intensities for signals with residual masses in the range from -0,35 to +0,65 were
222 summed and assigned to a cell with the corresponding integer m/z value. Data from all samples
223 were combined into one tensor with dimensions $57 \times 380 \times 1200$ corresponding to number of
224 samples, number of retention time points and number of m/z values respectively.

225 *LC-HRMS data treatment.* A significant challenge when dealing with HRMS data is
226 that the instrument is able to separate ion peaks with an m/z difference of 0.00001 which means
227 that theoretically the spectrum may contain up to 100000×1200 (mass range) components. In
228 reality, each time point of each sample had about 15000 ion peaks in the spectrum, but only a
229 small portion of them had significant intensities. Thus, to reduce computational costs only
230 signals with intensities higher than 5 % of the highest peak in the spectrum were selected
231 (shown in Fig. 2 (C,D)). After the elimination of weak and noisy signals 40-60 important peaks
232 were left in each spectrum. A dataframe containing the first timepoint of the first sample was
233 created and was filled sequentially by all subsequent time points from all samples. When an
234 m/z was found that had not been seen in previous time points, a new column was created and
235 filled with zeros for all preceding rows. This procedure constructs a unified mass scale across
236 all time points and all samples. To assess both environment-dependent and instrument-
237 dependent fluctuations in measured masses and retention times, an internal standard (IS) was
238 added to each sample. The mean absolute error (MAE) of the IS measured mass (m/z
239 395.13180) across all samples calculated for inter-day measurements was less than 0.005 Da.
240 The MAE is greater for bigger masses, therefore it was decided to set the m/z window size
241 equal to 0.01 Da. At the next step intensities of m/z signals in the dataframe which have mass
242 differences lower than 0.01 Da were considered to results from for one m/z and summed. Cells
243 with missing m/z signals were replaced by zero values. Finally, due to low reproducibility of
244 retention times probably caused by the unstable performance of the LC pump, time periods of
245 length 0.5 minutes were used instead of a continuous time axis. Unlike the LRMS dataset
246 where the m/z scale interval is constant, the size of the final dataframe for HRMS data will
247 depend on the number of unique m/z values observed in the particular dataset. However, the
248 approach adopted here can be applied to data produced by any HRMS system. The final array
249 was reshaped into a tensor with dimensions $51 \times 45 \times 2580$ with the same axes as the LRMS
250 data tensor.

251 **3.2. Chemometric analysis**

252 The obtained tensors were either directly subjected to PARAFAC decomposition or
253 unfolded into a 2D array. The unfolding procedure takes a tensor of dimensions $I \times J \times K$ and
254 rearranges it in such a way that the number of samples I remains unchanged and two other
255 dimensions (m/z and retention time (RT)) are combined into a single new dimension with size
256 $J \times K$. Therefore, the new feature space consists of the concatenation of the mass spectra and
257 retention time pairs for each sample. PCA, ICA, NMF and UFS methods were applied on data
258 organised by this approach and compared with the direct tensor decomposition. For
259 PARAFAC, PCA, ICA and NMF a critical parameter is the number of components, which was
260 chosen based on statistical analysis of each method without using any prior information about
261 dataset.

262 For the PARAFAC one- to fifteen-component PARAFAC models with non-negative
263 constrains were fitted to each dataset; the explained variance, corcondia criteria, error and
264 number of iterations for all models were compared and finally the optimal number of
265 components was selected to find the best balance across the criteria as shown in Fig. S1 (A,

266 B)[16]. The choice of the number of components was validated by half-split analysis (Fig. S1
267 C,D). Results of hierarchical clustering analysis following PARAFAC and all other methods
268 are presented in the Supplementary Information (Fig. S3 – S12). For PCA we selected the
269 number of components that was sufficient to explain 95% of the variance in the data, which
270 was 13 components (for LRMS) and 10 components (for HRMS). For the determination of the
271 number of ICA components, ICA-by-blocks method was used[17]. In Fig. S2 (A,B) signal-
272 correlation plots for LRMS and HRMS datasets are presented. It can be seen that for both
273 LRMS and HRMS data after extracting more than 4 components, the curves decrease
274 progressively which means that the correlations between the components of the different blocks
275 are much lower. Thus, the optimal number of components in those datasets is 4. To identify
276 the optimal number of components (rank of the matrix factors) for NMF, the residual sum of
277 squares (RSS) was calculated and its correlation with the number of components was
278 visualized, as shown in Fig. S2 (C,D). The optimal number was decided using a previously
279 suggested method[18] of identifying where the graph of RSS against the number of
280 components shows an inflection point (8 for LRMS and 9 for HRMS). Among different feature
281 selection methods variance-based UFS was chosen as the most suitable approach. It eliminates
282 features with variances below a predefined threshold which in this case was the mean of all
283 variances[19]. Using this threshold 97 % and 99 % of features from LRMS and HRMS datasets
284 respectively were excluded.

285 At the next step all five methods in the optimized conditions were applied to the
286 obtained datasets.

287 The results of the applied algorithms were compared using multiple criteria:
288 computational time, number of noisy components, silhouette score, and Davies-Bouldin index.
289 All results are presented in Table 2.

290 The Silhouette score was also used to understand how close the sample is to its parent
291 cluster compared with its neighboring cluster[20]. Silhouette coefficients close to +1 suggest
292 the sample is near to its true parent but distant from the neighboring clusters. A value of 0
293 means that the sample is between two adjacent clusters on or very close to the decision
294 boundary and negative values indicate incorrect cluster assignment for that sample. Values for
295 all samples are calculated and average among all of them is considered as silhouette score.
296 PCA showed the best performance with 0.71 and 0.48 scores for LRMS and HRMS data
297 respectively.

298 Another metric used for clustering performance evaluation was Davies-Bouldin
299 index[20]. This criterion is based on an averaged ratio “within-cluster” and “between-cluster”
300 distances. If two clusters are close together and have a large spread then this ratio will be large,
301 indicating that clusters are not very distinct. PCA and UFS produced the best results for LRMS
302 and HRMS respectively.

303 Computational costs for PCA, ICA, UFS were relatively the same and less than for
304 NMF and PARAFAC. It was shown that PCA and ICA produced fewer noisy components.

305 Based on the discussed criteria PCA, ICA and UFS methods demonstrated similar
306 performance in LRMS and HRMS data treatment. Therefore, the next stage was to compare
307 them by ability to discover biomarkers and by closeness of their clustering results to biological
308 molecular phylogenetic tree. However, all of these unsupervised techniques allow the most
309 variable markers in the composition of investigated samples to be identified.

310 **3.3. Biomarker identification**

311 The final stage of the data analysis was identification of the markers which were the
312 most important for clustering. In the metabolomic approach each feature ultimately represents
313 a single compound, because redundancy related to isotopic peaks and adduct ions is removed,
314 and only one time point of the peak vertex is taken into account for each feature. In our study
315 extra information about isotopologues and peak shapes is preserved, however the related

316 signals from one peak should be regarded as one compound for the purpose of biomarker
317 discovery. To extract such signals from LC-LRMS data treated by either PCA or ICA, a
318 retention time window of 0.4 min was established to group signals with the highest weights in
319 each component as well as m/z values of signals attributable to one isotopic pattern (A, A+1,
320 A+2). Each group of signals, therefore, could be regarded as one compound. Although the
321 number of such components was different for these three methods (see section 3.2), it was
322 decided to extract 50 compounds for each method by evenly extracting them from all
323 components. In the same manner signals corresponding to the first 50 most significant
324 compounds were selected after the UFS procedure. Further, an intersection of all these lists of
325 signals was obtained. Approximately 50 % of signals from each method's list were captured in
326 the intersection list. Among them 23 compounds were interpreted and remaining signals known
327 to be noisy (from high retention time) were not considered.

328 For LC-HRMS data same strategy was employed. The results of three methods (PCA,
329 ICA, and UFS) were intersected and the same potential chemotaxonomic markers were
330 observed. They correlate with most of the features from the intersection list generated using
331 LC-LRMS data.

332 Finally, it was tried to perform dereplication of these compounds based on the literature
333 data and available databases. Representative samples which contain compounds of interest
334 were reanalyzed on the qTOF instrument in auto-MS² mode. The results of the annotation are
335 presented in Table 3. Spectra for all compounds are presented in Supplementary (Fig. S14 –
336 S62)

337 **Compound 1** possessed a molecular weight of 328 deduced from the protonated
338 molecule ($[M + H]^+$) peak at m/z 329.1596 (C₁₆O₇H₂₅, eluted at 4.7 min), which produced
339 predominant fragment ion m/z 167.1061 corresponding to the cleavage of glucose molecule.
340 The exact position of the substituent could not be assigned and this compound was tentatively
341 assigned as verbenone glycoside or one of its isomers previously isolated from Prangos species
342 along with γ -pyrone glucosides[21], which are structurally similar to **compound 2**.

343 **Compound 2** had a molecular weight 432 deduced from the sodium adduct ($[M + Na]^+$)
344 signal at m/z 455.1162 and $[M + H]^+$ ion peak at m/z 433.1342 (C₁₈H₂₅O₅₂, eluted at 6.2). An
345 $[M + H]^+$ precursor ion produced the predominant fragment ion m/z 127.0387, which allowed
346 to suspect a structure similar to maltol glucoside. The fragment ion at m/z 329.0839 observed
347 in the ESI/MS² spectrum of the sodium adduct may be interpreted as a fragment of hydroxy-3-
348 methylglutaric acid (HMG) substituted glucopyranosyl side chain. Thus, **compound 2** can be
349 tentatively identified as previously reported licoagroside B or its isomer[22].

350 **Compound 3** was detected as the precursor ion $[M+H]^+$ at m/z 425.1451 (C₂₀O₁₀H₂₅,
351 eluted at 8.0 min). The observed fragment ions at m/z 263 and m/z 245 in its MS² spectrum
352 can be produced by loss of sugar moiety with a successive loss of a neutral fragment (H₂O).
353 Presence of the most intensive ion peak at m/z 191 corresponding to the additional loss of C₄H₆
354 allowed preliminary identification of this compound as rutarin or its positional isomer. Other
355 candidates were rejected after manual comparison with the spectra from GNPS library.

356 **Compound 4** showed a protonated molecule ($[M+H]^+$) ion peak at m/z 479.0821
357 (C₂₁H₁₈O₁₃, eluted at 8.3 min) and a predominant fragment at m/z 303.0501 in its MS²
358 spectrum, which should be attributed to the mild elimination of a glucuronic acid as a neutral
359 loss. The ions produced by the m/z 303.0501 precursor are in accordance with typical
360 fragmentation pattern of quercetin[23]. Therefore, this compound was tentatively identified as
361 quercetin glucuronide.

362 **Compounds 5 and 6** are a pair of isomers with molecular weight of 217 determined by
363 the protonated molecule ($[M+H]^+$) peak at m/z 217.0495 (C₁₂H₈O₄, eluted at 12.3 and 13.1
364 min). In the ESI/MS² spectra these precursor ions displayed the ion peak at m/z 202.0261
365 (C₁₁H₅O₄) corresponding to demethylation together with a signal at m/z 174 produced via

366 additional drop of CO. The ion peaks detected at m/z 189 and 161 resulted from the successive
367 losses of two CO molecules. Although these compounds could not be distinguished by their
368 ESI/MS spectra, the comparison of their elution order on a RP-C18 column with reported in
369 literature[23] allows a tentative identification of compound **5** as xanthotoxin and thus
370 compound **6** as bergapten.

371 **Compound 7** possessed a molecular weight of 246 deduced from the protonated
372 molecule ($[M+H]^+$) ion peak at m/z 247.0603 ($C_{13}H_{11}O_5$, eluted at 13.1 min). The molecular
373 weight of compound 7 is 30 Da larger than xanthotoxin (**5**), which corresponds to the additional
374 $-OCH_3$ substituent. A similar fragmentation pattern to other linear furanocoumarins (Table 3)
375 allows tentative identification of **compound 7** as isopimpinellin[24].

376 **Compound 8** had a molecular weight 260 determined by the protonated molecule
377 ($[M+H]^+$) signal at m/z 261.1123 ($C_{15}H_{16}O_4$, eluted at 13.3 min). Among the observed peaks
378 of its isomers, this one is the most retained. Moreover, the ion peak corresponding to the loss
379 of H_2O was not observed in its ESI/MS spectra, while the predominant fragment ion was
380 detected at 189 m/z . Therefore, compound **8** was tentatively characterized as isomeranzin[25].

381 **Compounds 9 and 10** both had a molecular weight 286 Da determined by the presence
382 of a protonated molecule peak at m/z 287.0918 ($C_{16}H_{15}O_5$, eluted at 13.4 and 14.1 min). These
383 two compounds showed the typical fragmentation patterns of monosubstituted
384 furanocoumarins, with the presence of m/z 203, 175, 159 and 145 in their ESI/MS² spectra
385 (Table 3). With respect to the presence of the fragment ion at m/z 269 $[M + H - 18]^+$ in the MS²
386 spectrum of **compound 9**, it was tentatively identified as pabulenol, and thus **compound 10**
387 would be oxypeucedanin[24].

388 **Compound 11** was detected by the presence of a the protonated molecule ($[M+H]^+$)
389 ion peak at m/z 323.0679 ($C_{16}H_{16}O_5Cl$, eluted at 14.6 min) with a specific isotopic distribution
390 corresponding to a monochlorinated compound. In the MS² spectrum this precursor produced
391 the predominant fragment ions corresponding to the loss of HCl and side chain cleavage at m/z
392 287 and 203, respectively. Other observed fragment ions were the same as for other
393 monosubstituted fumarocoumarins. Therefore, this compound was tentatively identified as
394 saxalin[21].

395 **Compounds 12 and 22** both had a molecular weight 270 Da determined by the
396 presence of protonated molecule ($[M+H]^+$) signal at m/z 271.0961 ($C_{16}H_{15}O_4$, eluted at 15.8
397 and 16.8 min) in their ESI/MS spectra. These precursors have shown ion peaks at m/z 203,
398 175, and 159 common for monosubstituted furanocoumarins (Table 3). However, efforts to
399 distinguish the paired isomers **12** and **22** by ESI/MS² analysis were unsuccessful, and these
400 compounds were differentiated by comparison of their elution order on a RP-C18 column with
401 reported in literature[24]. Thus, **compounds 12** and **22** were tentatively identified as
402 imperatorin and isoimperatorin, respectively.

403 **Compounds 13** had a molecular weight 316 Da determined by the presence of
404 protonated molecule ($[M+H]^+$) ion peaks at m/z 317.1384 ($C_{18}H_{21}O_5$, eluted at 16.0 min), it
405 exhibited the same fragmentation pattern as compounds **14**, **17**, **18**, **20**. Thus, **compound 13**
406 was tentatively assigned as linear isomer of cnidiadin[26]. It should be noted, that its pyrano-
407 analogue might also be found in some of the samples at almost the same retention time.

408 **Compounds 16 and 23** were a pair of isomers with molecular weight of 244 determined
409 by the protonated molecule ($[M+H]^+$) ion peak at m/z 245.1177 ($C_{16}H_{15}O_4$, eluted at 16.5 and
410 17.2 min), but their ESI/MS fragmentation patterns were quite different. Compound **23** showed
411 the presence of characteristic fragment ion at m/z 187 $[245-C_4H_{10}]^+$, while compound **16**
412 exhibits the predominant ion at m/z 189 $[245-C_4H_8]^+$, which was probably caused by different
413 $\pi-\pi$ conjugation extensions. Accordingly, compounds **16** and **23** were assigned as osthol[23]
414 and suberosin[21], which could be also confirmed by their retention in RP HPLC[27].

415 For **compounds 14,15,17-21**, all showed protonated molecule ($[M+H]^+$) peak at m/z
416 329,1387 ($C_{19}H_{20}O_5$, eluted between 16.3 and 17.0 min). These isomers may belong to the
417 classes of furanocoumarins and pyranocoumarins. Their ESI/MS² spectra demonstrated two
418 distinguishing patterns. One of them includes predominant ion peaks at m/z 229, 247 and 213,
419 while the second exhibits the most intensive ion peaks at m/z 229, 187 and 159. It was found
420 from literature that linear monosubstituted furanocoumarins and pyranocoumarins exhibit the
421 first fragmentation pattern while the angular ones show predominant ion peak at m/z 187[28].
422 Tentative assignments of angular and linear structures could be also confirmed by comparison
423 of their relative retention time in a RP HPLC column. It is known that angular coumarins are
424 more strongly retained compared to their linear isomers[29], and angelate isomer is eluted after
425 its senecioic acid analogue[30]. Moreover, pyranocoumarins tend to be eluted before
426 furanocoumarins[31]. Therefore compound **14** and **17** were tentatively identified as decursin
427 and decursinol angelate[32], and thus **compounds 18** and **20** would be prantschimgin and
428 deltoin (syn. sprengelianin)[33]. Similarly, compound **15** was tentatively assigned as
429 jatamansin, thus compounds **19** and **21** would be libanorin and columbianadin[34].

430 Many more chromatographic peaks of structurally similar compounds were observed
431 in the chromatograms. However, it is nearly impossible to differentiate all of them, because the
432 corresponding mass spectra are sometimes missing or not well described in the available
433 literature. Thus, however, the application of the suggested data treatment techniques allowed
434 identification of the most plausible chemotaxonomic marker candidates. Moreover, these
435 markers are expected to be significant due to the fact that they were found in the high-ranking
436 components of all three selected methods.

437 **3.4. Application to chemotaxonomic purposes**

438 Classification of plants on the basis of their secondary metabolites and their
439 biosynthetic pathways is called chemotaxonomy[1]. The main purposes of chemotaxonomy
440 are to improve the existing system of plant differentiation and to incorporate the modern
441 knowledge of the natural relationship of plants. One example of compounds which might be
442 used as chemotaxonomic markers are coumarins[2,4]. In the work [4] coumarin-containing
443 species, namely, *Angelica Sinensis*, *Angelica Dahurica*, *Angelica Decursiva*, *Peucedanum*
444 *Praeruptorium*, *Peucedanum Pubescens* were analyzed by direct injection MS in positive
445 multiple ion monitoring mode and the results showed that only several sample classes could be
446 separated from the main cluster in the PCA score plot. The variables responsible for this
447 classification were structurally described as angular-type pyranocoumarins, linear-type
448 pyranocoumarins, angular-type furanocoumarins, and ligustilide derivatives. In the present
449 work coumarins profiles were for the first time compared in the range of genera and species.
450 The distribution of biomarkers identified in this work is shown in Fig. 3. After careful
451 consideration of the identified biomarkers, it was concluded that there are no unique
452 compounds for any of the genera. In order to find out which markers depend on the growing
453 conditions and what are markers of each genus, a more extensive research with larger number
454 of biological replicates of each species and more representatives of each genus should be
455 conducted in future. We also note that for this particular task supervised techniques might show
456 significantly improved classification performance.

457 Another way to visualize the results of unsupervised learning is hierarchical tree
458 construction (Fig. S3-S12). Comparison by closeness of dendrograms created by each method
459 to the molecular phylogenetic tree (Fig. S13) was done. It should be noted that trees generated
460 from LC-MS data show differences in chemical composition which is not correlated with plant
461 molecular phylogenetic analysis results. For the evaluation of these differences, an approach
462 that involves computing pairwise distances between all data items and showing the distances
463 in a matrix form was employed[35]. As a quantitative characteristic pixel-wise mean square

464 error (MSE) can be calculated by Eq. (1) (where $I_{i,j}^1$ and $I_{i,j}^2$ are the i,j elements of the first and
465 second distance matrix respectively), in the form where instead of pixel values the original
466 distance values in the matrices are considered.

467

$$468 \quad MSE(I^1, I^2) = \frac{1}{n} \sum_{i,j} (I_{i,j}^1 - I_{i,j}^2)^2 \quad (1)$$

469

470 Errors calculated by this method were compared and the lowest value were obtained by
471 the UFS method for both LRMS and HRMS data: 0.105 and 0.144 respectively. Although plant
472 tissue chemical composition is highly variable, it may be beneficial to use the combination of
473 LC-MS-based methods and unsupervised machine learning algorithms along with molecular
474 phylogeny data in chemotaxonomic studies.

475

476

4. Conclusions

477 Two types of data analysis were considered for LC-LRMS and -HRMS data: tensor
478 decomposition by PARAFAC and decomposition following tensor unfolding into two
479 dimensions. For unfolded tensors, four approaches to data reduction and factorization were
480 considered: PCA, ICA, NMF and UFS. Results obtained by these methods from both datasets
481 were compared by several criteria. Applied to LC-LRMS and LC-HRMS data treated by
482 suggested approaches, PCA showed the best results according to silhouette coefficient, Davies-
483 Bouldin index, computational time and number of noisy components. However, PCA, ICA and
484 UFS demonstrated comparable performance and similar lists of biomarkers were revealed from
485 their results. A list of 23 compounds, most of which belong to the coumarin class were
486 extracted from the intersection of the results from all employed methods. These compounds
487 were tentatively identified by comparing their ESI/MS spectra with published data. The
488 distribution of these biomarkers in different species from the Apiaceae family was shown. The
489 identified compounds can potentially serve as chemotaxonomic markers because they were
490 chosen by the algorithms as features with highest dispersion across the samples.

491 Although the methodology allowed successful separation of each sample with its
492 replicates from the rest of the dataset, it has demonstrated some limitations in application to
493 biological classification. It was shown that dendrograms constructed by the employed methods
494 differ from the molecular phylogenetic tree, which may be caused by changes in chemical
495 composition of the studied extracts related to different environmental factors. Due to the high
496 chemical diversity of coumarins and other plant constituents, future studies should use a larger
497 number of biological replicates for each species.

498

Conflicts of interest

499 There are no conflicts to declare.

500

501

Acknowledgements

502 Process of plant material collection for this research was supported by the Russian
503 Foundation for Basic Research (project no. 19-04-00496 a). All other parts of the reported
504 study including experiment and data analysis were funded by RFBR, project number 19-33-
505 90036. HPLC-MS-TOF analysis was performed using the equipment of the demo laboratory
506 of Bruker Ltd., Moscow, Russia.

507

508

References

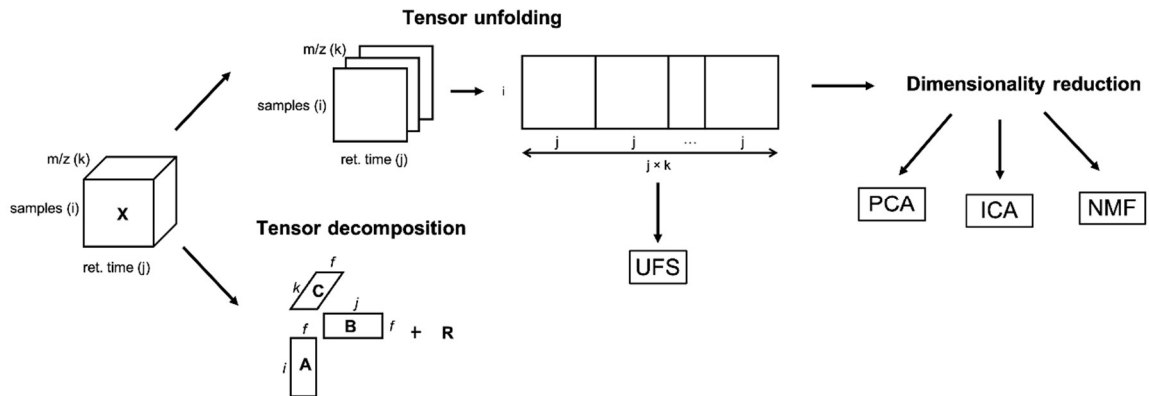
- 509 [1] R. Singh, Chemotaxonomy: A Tool for Plant Classification, Journal of Medicinal
510 Plants Studies. 4 (2016) 90–93.

511

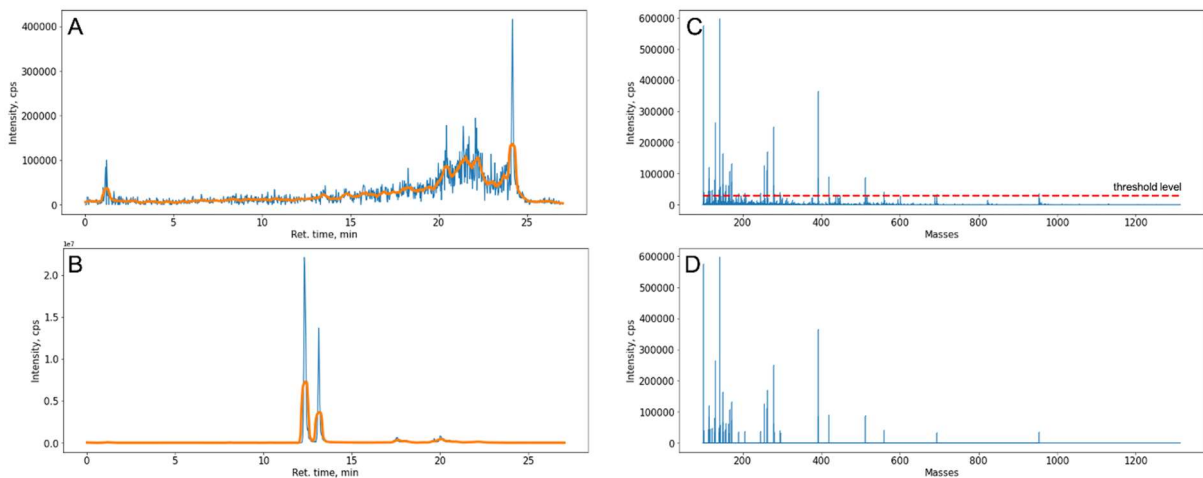
- 512 [2] A. Dugrand-Judek, A. Olry, A. Hehn, G. Costantino, P. Ollitrault, Y. Froelicher, F.
513 Bourgaud, The distribution of coumarins and furanocoumarins in Citrus species
514 closely matches Citrus phylogeny and reflects the organization of biosynthetic
515 pathways, PLoS ONE. 10 (2015). <https://doi.org/10.1371/journal.pone.0142757>.
- 516 [3] A. Forycka, W. Buchwald, Variability of composition of essential oil and coumarin
517 compounds of *Angelica archangelica* L., *Herba Polonica*. 65 (2019) 62–75.
518 <https://doi.org/10.2478/hepo-2019-0027>.
- 519 [4] X. Xu, W. Li, T. Li, K. Zhang, Q. Song, L. Liu, P. Tu, Y. Wang, Y. Song, J. Li, Direct
520 Infusion-Three-Dimensional-Mass Spectrometry Enables Rapid Chemome
521 Comparison among Herbal Medicines, *Analytical Chemistry*. 92 (2020).
522 <https://doi.org/10.1021/acs.analchem.0c00483>.
- 523 [5] K. Kumar, Introducing an integral optimised warping (IOW) approach for achieving
524 swift alignment of drifted chromatographic peaks: an optimisation of the correlation
525 optimised warping (COW) technique, *Analytical Methods*. 10 (2018).
526 <https://doi.org/10.1039/C8AY00963E>.
- 527 [6] K. Kumar, Optimizing the process of reference selection for correlation optimised
528 warping (COW) and interval correlation shifting (icoshift) analysis: automating the
529 chromatographic alignment procedure, *Analytical Methods*. 10 (2018).
530 <https://doi.org/10.1039/C7AY02340E>.
- 531 [7] W. Sun, R.D. Braatz, Opportunities in tensorial data analytics for chemical and
532 biological manufacturing processes, *Computers & Chemical Engineering*. 143 (2020)
533 107099. <https://doi.org/10.1016/j.compchemeng.2020.107099>.
- 534 [8] P. Turova, I. Rodin, O. Shpigun, A. Stavrianidi, A new PARAFAC-based algorithm
535 for HPLC–MS data treatment: herbal extracts identification, *Phytochemical Analysis*.
536 31 (2020) 948–956. <https://doi.org/10.1002/pca.2967>.
- 537 [9] P.W. Siy, R.A. Moffitt, R.M. Parry, Y. Chen, Y. Liu, M.C. Sullards, A.H. Merrill,
538 M.D. Wang, Matrix factorization techniques for analysis of imaging mass
539 spectrometry data, in: 2008 8th IEEE International Conference on BioInformatics and
540 BioEngineering, IEEE, 2008: pp. 1–6. <https://doi.org/10.1109/BIBE.2008.4696797>.
- 541 [10] N. Kuhnert, R. Jaiswal, P. Eravuchira, R.M. El-Abassy, B. von der Kammer, A.
542 Materny, Scope and limitations of principal component analysis of high resolution LC-
543 TOF-MS data: The analysis of the chlorogenic acid fraction in green coffee beans as a
544 case study, *Analytical Methods*. 3 (2011) 144–155.
545 <https://doi.org/10.1039/c0ay00512f>.
- 546 [11] Y. Gut, M. Boiret, L. Bultel, T. Renaud, A. Chetouani, A. Hafiane, Y.M. Ginot, R.
547 Jennane, Application of chemometric algorithms to MALDI mass spectrometry
548 imaging of pharmaceutical tablets, *Journal of Pharmaceutical and Biomedical*
549 *Analysis*. 105 (2015) 91–100. <https://doi.org/10.1016/j.jpba.2014.11.047>.
- 550 [12] P. Kharyuk, D. Nazarenko, I. Oseledets, I. Rodin, O. Shpigun, A. Tsitsilin, M.
551 Lavrentyev, Employing fingerprinting of medicinal plants by means of LC-MS and
552 machine learning for species identification task, *Scientific Reports*. 8 (2018).
553 <https://doi.org/10.1038/s41598-018-35399-z>.
- 554 [13] T. Bald, J. Barth, A. Niehues, M. Specht, M. Hippler, C. Fufezan, pymzML--Python
555 module for high-throughput bioinformatics on mass spectrometry data, *Bioinformatics*.
556 28 (2012) 1052–1053. <https://doi.org/10.1093/bioinformatics/bts066>.
- 557 [14] M.B. Gholivand, Y. Yamini, M. Dayeni, Y. Shokoohinia, The influence of the
558 extraction mode on three coumarin compounds yield from *Prangos ferulacea* (L.) Lindl
559 roots, *Journal of the Iranian Chemical Society*. 12 (2015) 707–714.
560 <https://doi.org/10.1007/s13738-014-0529-0>.

- 561 [15] K. Kumar, Standardising the chromatographic denoising procedure, *Analytical*
562 *Methods*. 10 (2018). <https://doi.org/10.1039/C8AY01606B>.
- 563 [16] R. Bro, H.A.L. Kiers, A new efficient method for determining the number of
564 components in PARAFAC models, *Journal of Chemometrics*. 17 (2003) 274–286.
565 <https://doi.org/10.1002/cem.801>.
- 566 [17] D. Jouan-Rimbaud Bouveresse, A. Moya-González, F. Ammari, D.N. Rutledge, Two
567 novel methods for the determination of the number of components in independent
568 components analysis models, *Chemometrics and Intelligent Laboratory Systems*. 112
569 (2012) 24–32. <https://doi.org/10.1016/j.chemolab.2011.12.005>.
- 570 [18] L.N. Hutchins, S.M. Murphy, P. Singh, J.H. Graber, Position-dependent motif
571 characterization using non-negative matrix factorization, *Bioinformatics*. 24 (2008)
572 2684–2690. <https://doi.org/10.1093/bioinformatics/btn526>.
- 573 [19] D. Weigt, D.A. Sammour, T. Ulrich, B. Munteanu, C. Hopf, Automated analysis of
574 lipid drug-response markers by combined fast and high-resolution whole cell MALDI
575 mass spectrometry biotyping, *Scientific Reports*. 8 (2018).
576 <https://doi.org/10.1038/s41598-018-29677-z>.
- 577 [20] L. Vendramin, R.J.G.B. Campello, E.R. Hruschka, Relative clustering validity criteria:
578 A comparative overview, *Statistical Analysis and Data Mining: The ASA Data*
579 *Science Journal*. 3 (2010). <https://doi.org/10.1002/sam.10080>.
- 580 [21] Y. Shikishima, Y. Takaishi, G. Honda, M. Ito, Y. Takeda, O.K. Kodzhimatov, O.
581 Ashurmetov, Terpenoids and γ -pyrone derivatives from *Prangos tschimganica*,
582 *Phytochemistry*. 57 (2001) 135–141. [https://doi.org/10.1016/S0031-9422\(00\)00407-6](https://doi.org/10.1016/S0031-9422(00)00407-6).
- 583 [22] S. Kozachok, Ł. Pecio, J. Kolodziejczyk-Czepas, S. Marchyshyn, P. Nowak, J.
584 Mołdoch, W. Oleszek, γ -Pyrone compounds: flavonoids and maltol glucoside
585 derivatives from *Herniaria glabra* L. collected in the Ternopil region of the Ukraine,
586 *Phytochemistry*. 152 (2018) 213–222.
587 <https://doi.org/10.1016/j.phytochem.2018.05.009>.
- 588 [23] X. Zheng, X. Zhang, X. Sheng, Z. Yuan, W. Yang, Q. Wang, L. Zhang, Simultaneous
589 characterization and quantitation of 11 coumarins in *Radix Angelicae Dahuricae* by
590 high performance liquid chromatography with electrospray tandem mass spectrometry,
591 *Journal of Pharmaceutical and Biomedical Analysis*. 51 (2010) 599–605.
592 <https://doi.org/10.1016/j.jpba.2009.09.030>.
- 593 [24] B. Li, X. Zhang, J. Wang, L. Zhang, B. Gao, S. Shi, X. Wang, J. Li, P. Tu,
594 Simultaneous characterisation of fifty coumarins from the roots of *angelica dahurica*
595 by off-line two-dimensional high-performance liquid chromatography coupled with
596 electrospray ionisation tandem mass spectrometry, *Phytochemical Analysis*. 25 (2014)
597 229–240. <https://doi.org/10.1002/pca.2496>.
- 598 [25] L. Duan, L. Guo, K. Liu, E.H. Liu, P. Li, Characterization and classification of seven
599 Citrus herbs by liquid chromatography-quadrupole time-of-flight mass spectrometry
600 and genetic algorithm optimized support vector machines, *Journal of Chromatography*
601 *A*. 1339 (2014) 118–127. <https://doi.org/10.1016/j.chroma.2014.02.091>.
- 602 [26] Y. Chen, G. Fan, Q. Zhang, H. Wu, Y. Wu, Fingerprint analysis of the fruits of
603 *Cnidium monnieri* extract by high-performance liquid chromatography-diode array
604 detection-electrospray ionization tandem mass spectrometry, *Journal of*
605 *Pharmaceutical and Biomedical Analysis*. 43 (2007) 926–936.
606 <https://doi.org/10.1016/j.jpba.2006.09.015>.
- 607 [27] M. Figueroa, I. Rivero-Cruz, B. Rivero-Cruz, R. Bye, A. Navarrete, R. Mata,
608 Constituents, biological activities and quality control parameters of the crude extract
609 and essential oil from *Arracacia toluensis* var. *multifida*, *Journal of*
610 *Ethnopharmacology*. 113 (2007) 125–131. <https://doi.org/10.1016/j.jep.2007.05.015>.

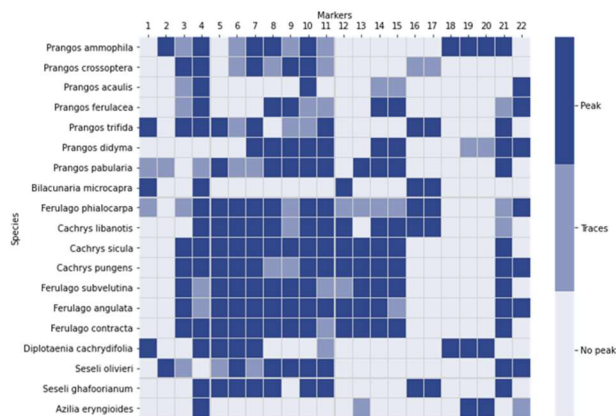
- 611 [28] V. Shukla, P. Singh, D. kumar, R. Konwar, B. Singh, B. Kumar, Phytochemical
612 analysis of high value medicinal plant *Valeriana jatamansi* using LC-MS and it's in-
613 vitro anti-proliferative screening, *Phytomedicine Plus*. 1 (2021) 100025.
614 <https://doi.org/10.1016/j.phyplu.2021.100025>.
- 615 [29] G.F. Spencer, L.W. Tjarks, R.G. Powell, Analysis of linear and angular
616 furanocoumarins by dual-column high-performance liquid chromatography, *Journal of*
617 *Agricultural and Food Chemistry*. 35 (1987) 803–805.
618 <https://doi.org/10.1021/jf00077a040>.
- 619 [30] J. Zhang, L. Li, T.W. Hale, W. Chee, C. Xing, C. Jiang, J. Lü, Single oral dose
620 pharmacokinetics of decursin and decursinol angelate in healthy adult men and
621 women, *PLoS ONE*. 10 (2015). <https://doi.org/10.1371/journal.pone.0114992>.
- 622 [31] S.Y. Kang, K.Y. Lee, S.H. Sung, M.J. Park, Y.C. Kim, Coumarins isolated from
623 *Angelica gigas* inhibit acetylcholinesterase: Structure-activity relationships, *Journal of*
624 *Natural Products*. 64 (2001) 683–685. <https://doi.org/10.1021/np000441w>.
- 625 [32] M.J. Ahn, M.K. Lee, Y.C. Kim, S.H. Sung, The simultaneous determination of
626 coumarins in *Angelica gigas* root by high performance liquid chromatography-diode
627 array detector coupled with electrospray ionization/mass spectrometry, *Journal of*
628 *Pharmaceutical and Biomedical Analysis*. 46 (2008) 258–266.
629 <https://doi.org/10.1016/j.jpba.2007.09.020>.
- 630 [33] Y. Xu, H. Cai, G. Cao, Y. Duan, K. Pei, S. Tu, J. Zhou, L. Xie, D. Sun, J. Zhao, J. Liu,
631 X. Wang, L. Shen, Profiling and analysis of multiple constituents in Baizhu Shaoyao
632 San before and after processing by stir-frying using UHPLC/Q-TOF-MS/MS coupled
633 with multivariate statistical analysis, *Journal of Chromatography B: Analytical*
634 *Technologies in the Biomedical and Life Sciences*. 1083 (2018) 110–123.
635 <https://doi.org/10.1016/j.jchromb.2018.03.003>.
- 636 [34] B. Wang, X. Liu, A. Zhou, M. Meng, Q. Li, Simultaneous analysis of coumarin
637 derivatives in extracts of *Radix Angelicae pubescentis* (Duhuo) by HPLC-DAD-ESI-
638 MS technique, *Analytical Methods*. 6 (2014) 7996–8002.
639 <https://doi.org/10.1039/c4ay01468e>.
- 640 [35] J. Wang, X. Liu, H.W. Shen, High-dimensional data analysis with subspace
641 comparison using matrix visualization, *Information Visualization*. 18 (2019) 94–109.
642 <https://doi.org/10.1177/1473871617733996>.
- 643



644
645 Fig. 1. Data organization and chemometric treatment workflow.
646



647
648 Fig. 2. An example of noisy raw LC-LRMS mass chromatogram smoothing by Savitzki-Golay
649 filter (A). An example of informative raw LC-LRMS mass chromatogram smoothing by
650 Savitzki-Golay filter (B). A representative raw mass spectrum from LC-HRMS before (C) and
651 after (D) noise subtraction below the threshold line.
652



653
654 Fig. 3. Distribution of revealed biomarkers in studied species.
655

656

Table 1. List of specimens used in the experiments

#	Plant species	Part	Specimen's voucher
1	<i>Prangos pabularia</i>	Leaves	MW0858238
2	<i>Cachrys libanotis</i>	Leaves	MW0798144
3	<i>Prangos acaulis</i>	Leaves	MW0744005
4	<i>Prangos ferulacea</i>	Stems	MW0751912
5	<i>Prangos didyma</i>	Stems	MW0857912
6	<i>Ferulago subvelutina</i>	Leaves	098-IR-19
7	<i>Prangos ammophila</i>	Leaves	MW0857867
8	<i>Prangos trifida</i>	Leaves	MW0798580
9	<i>Ferulago angulata</i>	Leaves	085-IR-19
10	<i>Cachrys sicula</i>	Leaves	MW0798143
11	<i>Ferulago contracta</i>	Leaves	053-IR-19
12	<i>Cachrys pungens</i>	Leaves	MW0784701
13	<i>Diplotaenia cachrydifolia</i>	Leaves	164-IR-19
14	<i>Ferulago phialocarpa</i>	Leaves	169-IR-19
15	<i>Azilia eryngioides</i>	Leaves	167-IR-19
16	<i>Seseli olivieri</i>	Leaves	173-IR-19
17	<i>Prangos crossoptera</i>	Leaves	MW0753036
18	<i>Bilacunaria microcapra</i>	Leaves	028-IR-19
19	<i>Seseli ghafoorianum</i>	Leaves	124-IR-19

657

658

Table 2. Comparison of data treatment techniques.

Method	Davies-Bouldin index	Silhouette score	Computational time, sec	Noisy components
LRMS data				
PCA	0.33	0.71	4.26	1
ICA	0.48	0.64	5.69	1
NMF	0.50	0.59	108.87	3
PARAFAC	0.43	0.47	140.59	3
UFS	0.52	0.53	1.26	–
HRMS data				
PCA	0.83	0.48	2.48	0
ICA	1.25	0.44	1.80	0
NMF	1.90	0.25	78.42	1
PARAFAC	1.05	0.38	122.86	1
UFS	0.75	0.40	0.26	–

659

Table 3. Chromatographic and mass-spectral data for compounds defined as biomarkers.

Number	RT (min), m/z	Components	[M+H] ⁺ , m/z (formula, error (ppm))	Adduct ions, m/z	Key MS/MS fragmentation	Identity	Reference
1	4.7 125	LRMS: PCA 2, ICA 2, NMF 2, FS	329.1595 (C ₁₆ H ₂₄ O ₇ , -0.1)	351.1419 [M+Na] ⁺	329-167 = Glc 167-125 = C ₂ H ₂ O	Verbenone glycoside	[21]
2	6.2 127	LRMS: PCA 4, ICA 4, NMF 5, FS HRMS: PARAFAC 9	433.1342 C ₁₈ H ₂₄ O ₁₂ , -0.4)	455.1162 [M+Na] ⁺	433-127 = C ₁₂ H ₁₈ O ₉ 127-85 = C ₂ H ₂ O 455-329 = C ₆ H ₆ O ₃	Licoagroside B	[22]
3	8.0 263	LRMS: PCA 9, ICA 3, NMF 3, PARAFAC 9 FS HRMS: PCA 4, ICA 1, NMF 3, PARAFAC 3	425.1451 (C ₂₀ H ₂₄ O ₁₀ , -1.9)	442.1716 [M+NH ₄] ⁺ 447.1271 [M+Na] ⁺	263-245 = H ₂ O 245-191 = C ₄ H ₆	Rutarin	GNPS library
4	8.3 303	LRMS: PCA 8, ICA 2, NMF 1 PARAFAC 9, FS HRMS: PARAFAC 3	479,0821 (C ₂₁ H ₁₈ O ₁₃ , -0.3)	501.0639 [M+Na] ⁺	479-303 = GluA 303-229 = C ₂ H ₂ O ₃ 303-153 = C ₈ H ₆ O ₃ 303-137 = C ₈ H ₆ O ₄	Quercetin glucuronide	[23]
5	12.3 217	LRMS: PCA 1, ICA 3, NMF 3, PARAFAC 2, FS HRMS:	217.0495 (C ₁₂ H ₈ O ₄ , 0.1)	234.0760 [M+NH ₄] ⁺	217-202 = CH ₃ 217-189 = CO 202-174 = CO 217-161 = 2CO 161-146 = CH ₃	Xanthotoxin	[23]

		PCA 4, ICA 2, NMF 3, PARAFAC 2			146-118 = CO		
6	13.1 217	LRMS: PCA 9, ICA 3, NMF 6, PARAFAC 2, FS HRMS: PCA 8, ICA 3, NMF 3, PARAFAC 1	217.0495 (C ₁₂ H ₈ O ₄ , 0.3)	234.0760 [M+NH ₄] ⁺	217-202 = CH ₃ 202-174 = CO 217-161 = 2CO 161-146 = CH ₃ 146-118 = CO	Bergapten	[23]
7	13.1 247	LRMS: PCA 10, ICA 3, NMF 3, PARAFAC 6, FS HRMS: PARAFAC 2	247.0605 (C ₁₃ H ₁₀ O ₅ , -1.4)	269.0429 [M+Na] ⁺ 264.0872 [M+NH ₄] ⁺	247-232 = CH ₃ 232-217 = CH ₃ 217-189 = CO 189-161 = CO	Isopimpinellin	[23]
8	13.3 261	LRMS: PCA 5, ICA 3, NMF 3, PARAFAC 6, FS HRMS: PARAFAC 1	261.1123 (C ₁₅ H ₁₆ O ₄ , -0.5)	283.0944 [M+Na] ⁺	261-217 = CO ₂ 261-189 = C ₄ H ₇ O 189-161 = CO	Isomeranzin	[25]
9	13.4 203, 269	LRMS: PCA 5, ICA 3, NMF 6, PARAFAC 2, FS HRMS: PCA 1, ICA 4, NMF 1, PARAFAC 2	287.0916 (C ₁₆ H ₁₄ O ₅ , -0.7)	304.1182 [M+NH ₄] ⁺	287-203 = C ₅ H ₉ O 203-175 = CO 203-159 = CO ₂ 175-147 = CO	Pabulenol	[24]
10	14.1 287	LRMS: PCA 4, ICA 4, NMF 5, PARAFAC 6, FS	287.0918 (C ₁₆ H ₁₄ O ₅ , -1.3)	309.0739 [M+Na] ⁺	287-203 = C ₅ H ₉ O 203-159 = CO ₂ 203-147 = 2CO 159-131 = CO	Oxypeucedanin	[24]

		HRMS: PCA 2, ICA 1, NMF 2, PARAFAC 8					
11	14.6 323	LRMS: PCA 5, ICA 1, NMF 4, PARAFAC 6, FS HRMS: PCA 6, ICA 2, NMF 7, PARAFAC 1	323.0679 (C ₁₆ H ₁₅ ClO ₅ , -0.6)	345.0501 [M+Na] ⁺	323-287 = HCl 287-203 = C ₅ H ₉ O 203-159 = CO ₂ 203-147 = 2CO	Saxalin	[21]
12	15.8 203	LRMS: PCA 4, ICA 4, NMF 5, PARAFAC 4, FS HRMS: PARAFAC 7	271.0961 (C ₁₆ H ₁₄ O ₄ , 1.5)	288.1224 [M+NH ₄] ⁺	271-215 = 2CO 271-203 = C ₅ H ₈ 203-175 = CO 203-159 = CO ₂ 175-147 = CO 175-131 = CO ₂	Imperatorin	[24]
13	15.9 317	LRMS: PCA 2, ICA 2, NMF 2, PARAFAC 4, FS HRMS: PARAFAC 5	317.1384 (C ₁₈ H ₂₀ O ₅ , -1.1)	339.1208 [M+Na] ⁺	317-247 = C ₄ H ₆ O 247-229 = H ₂ O	Cnididin (linear isomer)	[26]
14	16.3 329	LRMS: PCA 10, ICA 1, NMF 6, PARAFAC 8, FS HRMS: PCA 3, ICA 4, NMF 5, PARAFAC 5	- 329.1387 (C ₁₉ H ₂₀ O ₅ , -1.2)	351.1208 [M+Na] ⁺	329-247 = C ₅ H ₆ O 247-229 = H ₂ O	Decursin	[32]
15	16.3 329	LRMS: PCA 10, ICA 2, NMF 2, PARAFAC 8, FS	329.1387 (C ₁₉ H ₂₀ O ₅ , -1.2)	351.1207 [M+Na] ⁺	329-229 = C ₅ H ₈ O ₂ 229-187 = C ₃ H ₆ 187-159 = CO	Jatamansin	[34]

					159-131 = CO		
16	16.5 189 245	LRMS: PCA 5, ICA 3, NMF 3, PARAFAC 8, FS HRMS: PCA 8, ICA 3, NMF 3, PARAFAC 5	245.1177 (C ₁₅ H ₁₆ O ₃ , -2.0)	262.1350 [M+NH ₄] ⁺ 267.0990 [M+Na] ⁺	245-189 = C ₄ H ₈ 189-159 = CH ₂ O 159-131 = CO	Osthol	[23]
17	16.6 329	LRMS: PCA 2, ICA 2, NMF 2, PARAFAC 8, FS	329.1387 (C ₁₉ H ₂₀ O ₅ , -1.2)	351.1211 [M+Na] ⁺	329-247 = C ₅ H ₆ O 247-229 = H ₂ O	Decursinol Angelate	[32]
18	16.7 329	LRMS: PCA 2, ICA 2, NMF 2, PARAFAC 3, FS	329.1387 (C ₁₉ H ₂₀ O ₅ , -1.1)	351.1209 [M+Na] ⁺	329-247 = C ₅ H ₆ O 247-229 = H ₂ O	Prantschimgin	[33]
19	16.7 329	LRMS: PCA 2, ICA 2, NMF 2, PARAFAC 3, FS	329.1385 (C ₁₉ H ₂₀ O ₅ , -0.5)	351.1206 [M+Na] ⁺	329-229 = C ₅ H ₈ O ₂ 229-187 = C ₃ H ₆ 187-159 = CO 159-131 = CO	Libanorin	[34]
20	16.8 329	LRMS: PCA 10, ICA 2, NMF 2, PARAFAC 3, FS HRMS: PCA 3, ICA 4, NMF 5, PARAFAC 5	329,1387 (C ₁₉ H ₂₀ O ₅ , -1.1)	351.1208 [M+Na] ⁺	329-247 = C ₅ H ₆ O 247-229 = H ₂ O	Sprengelianin (Deltoin)	[33]
21	17.0 329	LRMS: PCA 2, ICA 2, NMF 2, PARAFAC 3, FS	329.1386 (C ₁₉ H ₂₀ O ₅ , -0.8)	351.1204 [M+Na] ⁺	329-229 = C ₅ H ₈ O ₂ 229-187 = C ₃ H ₆ 187-159 = CO 159-131 = CO	Columbianadin	[34]

22	16.8 229 203	LRMS: PCA 10, ICA 4, NMF 5, PARAFAC 8, FS HRMS: PARAFAC 7	271.0961 (C ₁₆ H ₁₄ O ₄ , 1.6)	293.0782 [M+Na] ⁺	271-215 = 2CO 271-203 = C ₅ H ₈ 203-175 = CO 203-159 = CO ₂ 175-147 = CO 175-131 = CO ₂	Iso-imperatorin	[24]
23	17.2 245	LRMS: PCA 4, ICA 4, NMF 5 PARAFAC 4, FS HRMS: PARAFAC 5	245.1172 (C ₁₅ H ₁₆ O ₃ , -0.1)	267.0996 [M+Na] ⁺	245-215 = C ₂ H ₆ 245-187 = C ₄ H ₁₀ 245-131 = C ₆ H ₁₀ O ₂	Suberosin	[21]