**Algorithm for comprehensive analysis of datasets
from hyphenated high resolution mass spectrometric
techniques using single ion profiles and cluster
analysis.**

Guillaume L. Erny,^{a,} Tanize Acunha,^b Carolina Simó,^b Alejandro Cifuentes^b and Arminda Alves^a*

^aLEPABE - Laboratory for Process Engineering, Environment, Biotechnology and Energy, Faculdade de Engenharia da Universidade do Porto, Rua Dr. Roberto Frias, 4200-465 Porto, Portugal..

^bLaboratory of Foodomics, CIAL, CSIC, Nicolas Cabrera 9, 28049 Madrid, Spain.

* Correspondence author Tel +351-22-508-4858, E-mail: guillaume@fe.up.pt. Fax: 1351-234-370084

13 **Abstract**

14 Various algorithms have been developed to improve the quantity and quality of information that can be
15 extracted from complex datasets obtained using hyphenated mass spectrometric techniques. While
16 different approaches are possible, the key step often consists in arranging the data into a large series of
17 profiles known as extracted ion profiles. Those profiles, similar to mono-dimensional separation
18 profiles, are then processed to detect potential chromatographic peaks. This allows extracting from the
19 dataset a large number of peaks that are characteristics of the compounds that have been separated.
20 However, with mass spectrometry (MS) detection, the response is usually a complex signal whose
21 pattern depends on the analyte, the MS instrument and the ionization method. When converted to ionic
22 profiles, a single separated analyte will have multiple images at different m/z range. In this manuscript
23 we present a hierarchical agglomerative clustering algorithm to group profiles with very similar feature.
24 Each group aims to contain all profiles that are due to the transport and monitoring of a single analyte.
25 Clustering results are then used to generate a 2 dimensional representation, called clusters plot, which
26 allows an in-depth analysis of the MS dataset including the visualization of poorly separated compounds
27 even when their intensity differs by more than two orders of magnitude. The usefulness of this new
28 approach has been validated with data from capillary electrophoresis- time of flight mass spectrometry
29 hyphenated via an electrospray ionization. Using a mixture of 17 low molecular endogenous compounds
30 it was verified that ionic profiles belonging to each compounds were correctly clustered even with very
31 low degree of separation (R below 0.03). The approach was also validated using a urine sample. While
32 with the total ion profile 15 peaks could be distinguished, 70 clusters were obtained allowing a much
33 thorough analysis. In this particular example, the total computing took less than 10 min.

34 **Keywords:**

35 Chemometrics; Hyphenated techniques; Data mining; Representation;

36 1. Introduction

37 Separation techniques coupled with high resolution mass spectrometry (X-HRMS, where X stands for
38 GC, LC, CE...) are key analytical hyphenated instruments in many areas of research [1-8]. In these
39 configurations, the mass spectrometer is placed at the outlet of the separation device, allowing recording
40 full mass spectra at regular intervals. Initially aimed as a technique to identify the main peaks in a
41 profile (following the recommendation of the IUPAC, profile can be an electropherogram,
42 chromatogram or any continuous trace recorded by an analytical instrument [9]), X-HRMS is now
43 essential for analysing complex matrices. MS instruments are considered almost universal detectors
44 while extremely selective and sensitive. As a result, a typical non-targeted analysis by X-HRMS
45 comprises a huge amount of information in a dataset practically impossible to handle in a manual way.
46 A typical approximation during any X-HRMS analysis is to narrow the mass range of the MS detector
47 to only detect those ions whose m/z are within the selected range. While this can be done via the
48 instruments setting, the easiest way is often done by filtering the resulting dataset to obtain the extracted
49 ion profile (EIP). Interestingly, filtering can be used multiple times using the same dataset, thereby,
50 obtaining profiles that are selective to potentially represent every analyte present in the sample.
51 However, knowledge of the target mass interval is a key requirement for minor components as they will
52 only be visible if the dataset is adequately filtered. This is the main limitation in non-targeted X-HRMS
53 analysis. Because it is important in many fields to obtain a complete analysis of the sample, especially
54 of unknown or unexpected components, many algorithms have been proposed to improve the quality
55 and quantity of information that can be obtained after a non-targeted X-HRMS analysis [10].

56 Automatic processing of hyphenated dataset often starts with the transformation of the raw MS data
57 into centroid data. This can be done by various algorithms that detect in each MS scan every peak and
58 calculate for each of them the centroid mass (also called the accurate mass [9]). The centroid data only
59 records the centroid m/z value and the peak amplitude. This allows a drastic reduction of the data size,
60 and as a consequence, faster computational speed. Centroid data can be sometimes obtained by the

61 proprietary acquisition or associated software, such as, compassXport (Bruker) or Masslynx (Waters),
62 or secondary software [11-13]. Those software also allow to export the data to formats such as mzML
63 [14], mzXML [15] or JCAMP [16] that can easily be read by users [16]. It should be emphasized that
64 the transformation to centroid data is an important transformation of the original data that can influence
65 a subsequent quantitative analysis [17]. The information in the resulting file can be seen as a collection
66 of 3 coordinates data point: the accurate mass m/z value (referred from now on as mass coordinate), the
67 scan number and the peak intensity. Such dataset can be processed in two ways, either each MS scan is
68 analyzed to detect peaks that are probably related to the same compounds (isotopic ions, common
69 adducts, possible fragmentations) [18-20] or profiles characteristic of the transport and separation of the
70 formed ions can be reconstituted and processed as chromatographic profiles. The latter is often the
71 method of choice in hyphenated MS datasets [21-23]

72 Two approaches are used to obtain chromatographic profiles, i.e., the binning approach and the pure
73 ion profile. In the binning approach, a series of EIPs are defined in such a way that each EIP will only
74 contain data within a very narrow mass range, so, all the information is distributed along the EIPs [11].
75 The mass range is the key parameter and should be carefully optimized to avoid peak splitting.
76 Algorithms have also been proposed to correct for this default [24]. The EIPs are then processed to
77 detect for the presence of chromatographic peaks. Those are then measured and figure of merits (peak,
78 center, amplitude, m/z , area...) recorded in a table [22, 24-26]. Another recent alternative is the use of
79 pure ion profile (PIP), which consists of scanning the dataset to find sequences of points that follow
80 each other in their scan coordinate and does not differ in their mass coordinate by more than a certain
81 threshold that corresponds to the variation in the accurate mass value determined for the same peak in
82 the mass profile (original data) at different scan [27-29]. The advantage of PIP over EIP (or binning
83 approach) is that in the PIP only the data points related to the target ions are obtained. This makes the
84 use of peak picking algorithms redundant. However, while EIP can be obtained from centroid or profile
85 dataset, PIP can only be calculated using centroid dataset. Automatic processing of EIP or PIP allows
86 obtaining a large number of chromatographic peaks that are characteristic of the analytes of interest

present in the sample. Most algorithms are able to retrieve peaks at very low intensity that would have never been detected manually without information on their mass.

Nevertheless, this is not sufficient for a comprehensive analysis of the whole X-HRMS dataset due to its intrinsic complexity. When processing the dataset, multiple EIE or PIP will be found for every analyte that has been separated. Thus, the remaining problem is to determine whether a minor peak is due to a new analyte or an image of an existing peak. Usually, this problem is bypassed by working in differential analysis [12, 30, 31]. However, PIP related to the same compounds are highly related with each other. All PIP due to isotopic ions, adducts or fragments are an image of the main profile where the only variation is the intensity of the peak. We have already used this concept in a previous work to propose a new representation [21]. The present work is a significant improvement from that previous approach. Here we use for the first time a hierarchical agglomerative cluster approach (HACA) [32] to classify the different profiles, based on similarity as measured by the correlation coefficient into clusters. Ideally, selected clusters should contain all profiles related to the same analyte. While peak aggregation as already been discussed [33-34], HACA has never been tested in this context. While previously published approaches [33, 34] used classical figures of merits such as time at peak apex and peak width, the HACA designed in this work is based on the value of the correlation coefficients between profiles. It is therefore less dependent on noise that can prevent the accurate determination of peak apex and peak width. In the present work, we demonstrate, using test and real complex samples, that HACA allows to rationally and systematically organize the various profiles in a limited amount of clusters. In this work, the clusters are used to provide a new two dimensional representation of the dataset, called clusters plot, which allows a comprehensive analysis of the results, making possible to distinguish analytes that are not well separated.

2. Materials and methods

2.1. Programming

112 Programming was done using Matlab 2013b and functions were run on a personal computer (Intel
113 Pentium CPU G3320@3.00GHz, 4.00GB RAM, 64-bit operating system) with Windows 7. All the
114 functions programmed and used during this work can be obtain by contacting the corresponding author.

115 2.2. *Chemicals and samples*

116 All reagents employed in the preparation of the CE buffer and sheath liquid (isopropanol, formic acid
117 and ammonium hydroxide, all of MS grade) were from Sigma-Aldrich (St. Louis, MO, USA). A
118 mixture of 17 small molecular mass compounds from Sigma-Aldrich was employed as a test mixture:
119 0.3 mM adenosine triphosphate (ATP), 7.8 mM nicotinic acid, 0.5 mM glutamic acid, 1.0 mM aspartic
120 acid, 0.3 mM glutathione oxidized, 0.6 mM glutathione reduced, 2.7 mM iminodiacetic acid, 0.2 mM
121 adenosine monophosphate (AMP), 0.2 mM panthotenic acid, 1.4 mM succinic acid, 0.1 mM gluconic
122 acid, 0.4 mM hippuric acid, 0.3 mM malic acid, 0.2 mM citric acid, 0.2 mM tartaric acid, and 2.8 mM
123 1,4 piperazinediethanesulfonic acid (PIPES) and 156.5 mM malonic acid. The urine was filtered through
124 0.2 μ m polyethersulfone filter before CE-MS analysis.

125 2.3. *CE-TOF MS analysis*

126 The capillary electrophoresis (CE) apparatus used was a P/ACE 5010 from Beckman (Fullerton, CA,
127 USA). The CE instrument was controlled by a PC running System GOLD software from Beckman. The
128 CE equipment was coupled to a time-of-flight (TOF) instrument “microTOF” from Bruker Daltonik.
129 CE-TOF coupling was carried out via an ESI interface model G1607A from Agilent Technologies.
130 Electrical contact at the electrospray needle tip was established via a flow of sheath liquid composed of
131 2-propanol-water (50:50, v/v) delivered by a 74900-00-05 Cole Palmer syringe pump (Vernon Hills, IL,
132 USA) at a flow rate of 4 μ L/min. Bare fused-silica capillary with 50 μ m i.d. and 85 cm of total length
133 was from Composite Metal Services (Worcester, England). The inner capillary wall was coated with a
134 cationic TEDETAMA-co-HPMA copolymer [35]. CE separation was performed at -20 kV in an acidic
135 BGE (1 M formic acid adjusted to pH 2.4 with ammonium hydroxide). Nebulizer pressure was 0.4 bar,
136 drying gas flow rate was 4 L/min, and ESI chamber temperature was set at 200 °C. TOF MS was

operated in the negative ion mode (capillary voltage was 4 kV) and spectra were acquired in the range of 50-700 m/z .

2.4. Datasets

Two datasets were used in this work, both obtained under the same CE-MS analytical conditions. The first one was obtained from the CE-MS analysis of a test mixture (17 small molecules). The second dataset was obtained from a urine sample injected under the same CE-MS analytical conditions.

3. Programming

In this work, a series of functions, programmed in Matlab, were designed in order to obtain a true two dimensional representation of the CE-TOF-MS datasets. The workflow is illustrated in Fig. 1 and consists of four main tasks: loading the data, extracting the single ion profiles, classification and generation of the clusters plot.

3.1. Data loading

This initial step consists of loading the centroid dataset. Centroid dataset can be obtained using various freeware as well as some proprietary software. In the case of Bruker instruments, for example, the software exports the rough data as CSV files and allows obtaining a centroid dataset. CompassXport is a software freely available from Bruker and also allows obtaining profile or centroid datasets from files in various format (text, JCAMP, mzXML, mzML...). In this work, Matlab functions were developed for mzML, JCAMP or CSV files that were processed from original Bruker files using CompassXport. Such formats are text based and can be easily read. During data loading, the only modification was to round off the m/z values to a fixed number of decimals; namely, five decimals were used for all TOF MS data since a higher precision is meaningless.

3.2. Pure ion profiles (PIPs)

The quality and quantity of the PIPs is paramount to this approach. Functions have been developed with a particular attention to increase both the computing speed and the ability to extract from the dataset PIPs with very low intensity. In this work, PIPs are obtained in three steps. First the dataset is

162 filtered to remove the random noise, this is done by rejecting all data points whose intensity is below a
163 set threshold. Value of this threshold is an important experimental parameter and should be accurately
164 determined. This can be done via the intensity frequency distribution of the dataset, *i.e.* the number of
165 data points in the dataset whose intensity is equal, or between the set values. Such a distribution for the
166 urine dataset is shown in Fig. 2A.

167 As it can be observed, while the intensities ranged from 1 to $\sim 110,000$ counts, a large fraction of the
168 data has an intensity below 20 counts. This very low value can be used confidently as threshold to filter
169 the dataset. Fig. 2B shows the comparison of the total ion current profiles (TICPs) obtained using all
170 data points (top) and the one obtained (bottom) after removal of the random noise (intensities below 20).
171 Fig. 2C shows the difference between the two TICPs. Using this threshold value, the data integrity is
172 maintained while the size of the dataset has been reduced by more than 91%. Such values are consistent
173 with other works and allow to drastically increasing the computing speed of the PIP algorithm.

174 In the work from Wang and co-workers [27], PIPs were obtained selecting a starting point and finding
175 data in the time coordinate that does not differ in their m/z coordinate by more than a fixed value that is
176 the uncertainty in the determination of the accurate mass. In the present work, we use a projection of the
177 data to a mass axis to obtain the total mass profile (TMP) as shown in Fig. 3. This profile can be used to
178 measure the accurate mass uncertainty as shown in inserts A to C. With the urine dataset this value can
179 be estimated at $0.002\ m/z$ for analyte peaks (insert A and C) and $0.005\ m/z$ for background ions (insert
180 B). The TMP can also be used to find mass intervals of interest (MII), determined as group of points in
181 the TMP where the m/z value of each point does not differ from its neighbours by more than $x/\sqrt{(y-1)}$,
182 where x is the mass uncertainty and y is the minimum number of consecutive points to define
183 a single ion profile. PIPs are then selected as any sequence of at least y points whose time coordinate
184 follow each other and belong to the same MII. Each PIP is recorded individually and figures of merit are
185 measured and recorded in a table. Those include the average and standard deviation of the accurate mass
186 measured with all the data points that build the PIP, the time and intensity of the most intense data

points, as well as the PIP zeroth and first statistical moment (area and centre). This is a very simple and robust approach to obtain single ion profiles with no assumption on their shape or intensity. A high number of PIPs are obtained in this way in a short time. For example with the urine sample 4779 PIPs ($x = 0.002$, $y = 5$) were obtained in less than 9 min. The computing speed depends on the size of the TMP that can be reduced by increasing the initial threshold. Using a threshold of 50 counts instead of 20, 1776 PIPs were obtained in 2 min. Interestingly, if only considering PIP with a maximum intensity higher than 100 (representing 0.05 % of the highest intensity), 1262 PIPs were obtained with a threshold of 20, while 1231 PIPs were obtained with a threshold of 50 showing the robustness of the algorithm. Hereinafter, profiles whose maximum intensity are below 100 counts will be discarded. This filtering step attempts to keep only profiles whose maximum intensity is higher than 5 times the background noise and is a simple step to only work with PIPs of chromatographic relevance.

3.3. Hierarchical agglomerative cluster analysis (HACA).

The main novelty of this work is the use of HACA to group PIPs based on similarities, giving as result an arrangement of all the data in individual clusters. HACA is a “bottom up” approach. Initially each profile starts in its own cluster and a measuring is used to assess the similarity between every cluster. The two most similar clusters are merged into a new one and this step is repeated until one cluster remains. The hierarchical structure is formed by the clusters obtained at each step and can be analyzed by visualizing how the data relate with each other at a particular hierarchical level (HL). For this work, clusters were compared by calculating the matrix of correlation between the most intense profiles in each cluster. The correlation between two profiles X and Y is defined as:

$$\rho_{X,Y} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} \quad (1)$$

210 $\text{Cov}(X,Y)$ is the covariance of X and Y, and σ_X and σ_Y are the variances of X and Y, respectively.
211 Correlation coefficients are normalized measurements with values between 1 and -1. This approach
212 allows comparing profiles not only based on their position but also on their global shape.

213 *3.4. Selection of the hierarchical level and cluster plot.*

214 While the HACA is systematic and does not require any input, the selection of the HL is critical for
215 the final representation. HL can be defined as the number of iterations but it is more pertinent to use
216 here the correlation between the two clusters merged at each iteration. Under these conditions, the HL
217 can vary from 1, where each profile will be in an independent cluster, till zero where each profile will be
218 in a single cluster.

219 An HL of 0.9 will indicate that within a cluster each PIP will be correlated by a value higher or equal
220 to 0.9. The task here, is to find the HL where all isotopes and adduct ion profiles will be regrouped into
221 the same cluster. Ideally, each cluster should only contain the profile related to only one component.
222 Obviously, this will depend on the degree of separation between compounds in the sample. Optimisation
223 of the HL is achieved using a two dimensional representation of the clusters, named clusters plot, which
224 is a two dimensional scatter plot of the main values representative of the clusters. An example of such
225 plots can be seen in Fig. 4 where each circle depict a cluster. Its time coordinate corresponds to the
226 average time at peak maxima of all PIPs in this cluster, its mass coordinate and size correspond to the
227 accurate mass and intensity of the most intense profile respectively. Because an analyte should have
228 more than one image profiles, the one corresponding to the molecular ions as well as the main isotopic
229 ions and adducts, clusters than only contain one profile are discarded in such a representation. The use
230 of the clusters plot to select the HL and to improve the visualization of key information from the dataset
231 will be discussed in the following section.

232 **4. Results and discussion**

233 *4.1. Test Mixtures*

234 The approach was first validated using the dataset obtained from the CE-TOF/MS analysis of the
235 standard sample composed of 17 metabolites. With an intensity threshold of 20, $x = 0.002$ and $y = 5$,
236 2347 PIPs were obtained from the dataset in less than 5 min; among them, only 440 had a maximum
237 intensity higher than 100. Those 440 PIPs were organized in the hierarchical structure in less than one
238 minute. Final results will depend on the HL selected. The only task done manually by the operator
239 consists of finding the right HL for which all image profiles of the same component fall within the same
240 clusters, while separates the different analytes in different clusters. As illustrated in Fig. 4, the clusters
241 plot allows the optimization of this parameter. In this figure two clusters plots are superposed, the plot
242 obtained at $HL = 0.8$ (stars) and the plot obtained at $HL = 0.95$ (circles). The clusters labeled with a
243 number were unambiguously assigned to one of the 17 compounds initially present in the test mixtures.
244 The clusters labeled with a letter are either false positive or impurities. The main differences between
245 both representations is in the clusters (circles) labeled as 7, 8 and 9 that were only obtained using a HL
246 value of 0.95 while with $HL = 0.8$, a single cluster (star near 9) containing 37 profile was obtained. The
247 distribution of the 20 most intense profiles collected in this single cluster at HL 8.0 in various clusters at
248 HL 0.95 is summarized in Table 1. As it can be seen, those PIPs are organized in at least 5 clusters (7, 8,
249 9, d and e in Fig. 4). It can also be observed that the profiles in the 3 main clusters (labeled as 7, 8 and 9
250 in Fig.4) are consistent with the presence of the molecular ion and at least one isotope ion (7: 178.0518
251 and 179.0554; 8: 195.0523, 196.0556 and 197.0547; 9: 132.0328 and 133.0342). Using the most
252 intense profiles in clusters 7 and 8, the resolution between two clusters was measured equal to 0.027.
253 This demonstrates that clusters plots allow to visually detect analytes that are poorly separated even
254 when their relative intensity differs significantly. For example the similarity between profiles from
255 clusters d and 9 in Fig. 4 can be assessed in Fig. 5. The resolution between those two profiles is equal to
256 0.59 and their relative intensity differ by one order of magnitude. Although the PIPs from these clusters
257 are very similar both in terms of m/z and migration time, they belong to different analytes and were
258 rightly assigned at $HL = 0.95$ to two different clusters. At $HL = 0.95$, all standards, numbered as 1 to 17
259 in Fig. 4, were successfully separated and with the exception of malonic acid (cluster 2), the most

intense PIP in each cluster corresponds to the molecular ions. Very good agreement were obtained between the exact mass and the accurate mass, this can be seen even with peaks poorly separated as hippuric acid (7), gluconic acid (8) and iminodiacetic acid (9). For malonic acid, the PIP corresponding to the molecular ions is the second most intense one.

The entire data processing necessary for obtaining the cluster plots (i.e., from the first step in which the original file was converted to mzML file till displaying the clusters plot) took less than 20 min. The list of the m/z values for all PIPs and each cluster can be found in supplementary material. For all tested datasets, HL = 0.95 always gave the best results and was used for all clusters plots. This value seems to be the highest value to accurately group all profiles related to the same analyte; however, co-clustering of more than one analyte was also observed with complex samples. This value may depend on the sampling frequency, peak efficiency and noise. The full listing of the m/z values of every PIP and cluster can be found in the supporting information.

4.2. Urine Sample.

The above data treatment was applied to the dataset obtained from the CE-TOF MS analysis of a urine sample. Under the same conditions used for the standard sample, 1262 PIPs were processed from the urine dataset with an intensity higher than 100. 177 of those PIPs were organised in 70 clusters with a maximum intensity ranging from 100000 to 1000. The clusters plot is shown in Fig. 6. Not surprisingly, this representation is more complex than that obtained with the standard mixtures. To validate the approach, the clusters within the square were analysed in detail below. Fig. 7 shows, for each cluster selected in Fig. 6, the most intense normalized PIPs. It can be observed that 6 out of 7 clusters clearly show pure profiles. Nevertheless, in all cases every cluster regroups at least one isotopic profile besides the main PIP (1. 135.0320 & 136.0347; 2. 191.0539 & 192.0603; 3. 165.0391 & 166.0439; 4. 195.0507 & 196.0544; 5. 263.1017 & 264.1037; 6. 149.0471 & 150.0516; 7. 179.0545 & 180.0573) allowing the identification of the component in each cluster unambiguously. The list of the m/z values for all PIPs and each cluster can be found in supplementary material. When working with real samples, numerous

clusters seem to have more than one compound that could easily be identified and separated using common rules of adducts formation.

While the relevance of all clusters should be further studied using multiple datasets, with a single dataset it is also easy to study the profile and m/z values in each cluster, helping to determine whether a cluster is an artifact or contain one or more analytes. This represents a systematic and robust approach for a comprehensive analysis of complex datasets from hyphenated X-HRMS techniques. While with small molecules, clusters are generally low populated, preliminary experiments with macromolecules (protein digests) showed that this approach could be particularly relevant for these type of analytes as clusters are more populated increasing the confidence of the representation. This will be the topic of a future work.

5. Conclusions

Hierarchical agglomerative cluster analysis is a logical and robust approach to order single ions profiles in clusters from complex X-HRMS datasets, as corroborated by the excellent results obtained in this paper for both a standard and a real sample analysed by CE-TOF MS. The cluster plot, a two dimensional representation of the data, allows to easily determine the best hierarchical level as well as it provides an excellent visualisation of the information contained in an X-HRMS dataset. This representation allows detecting poorly separated analytes even when one is at very low concentration. Another great advantage of the approach proposed in this work is that it is methodical and does not rely on any supervision from the operator. However, this approach relies on the consistency of the obtained PIPs. While the PIP algorithm presented here gave good results, it should also be compared with profiles obtained using other algorithms. In this work, because of the very narrow mass intervals obtained, no peak picking algorithms were used. However, with real sample, some of the PIPs clearly showed the presence of more than one ion. Peak picking algorithms may be introduced in a future work to resolve this issue. The approach developed here is particularly rapid allowing to test various

conditions in order to optimize the results in a fast and straightforward way. Matlab files can be obtained on demand.

Acknowledgments

This work was financially supported by: Project UID/EQU/00511/2013-LEPABE (Laboratory for Process Engineering, Environment, Biotechnology and Energy – EQU/00511) by FEDER funds through Programa Operacional Competitividade e Internacionalização – COMPETE2020 and by national funds through FCT - Fundação para a Ciência e a Tecnologia; by FEDER funds through the Operational Program for Human Potential and by National Funds through FCT-Foundation for Science and Technology under the project IF/00528/2013; and by AGL2014-53609-P project (Ministerio de Educación y Ciencia, Spain).

TEDETAMA-co-HPMA copolymer was gently provided by Dr. Alberto Gallardo from Institute of Polymer Science and Technology (ICTP-CSIC).

FIGURE CAPTIONS

Figure 1: Workflow used in this work

Figure 2: A: frequency distribution; B: TICP with full data (top) and filtered data (bottom) and C: residual between the full and filtered data.

Figure 3: Total mass profile obtained by projecting all data points of intensity higher than 20 counts on to a mass axis. Insert A, B and C are zooms of the encircled areas.

Figure 4: Two clusters plots of the standard mix obtained for merging correlation coefficients of 0.8 (stars) and 0.95 (circles). Clusters plot obtained at 0.95 are also labeled with numbers and correspond to the following compounds (values between parentheses correspond to the exact mass and the average

333 accurate mass respectively): 1. ATP ($505.9985 - 505.9891 \pm 0.0013$); 2. Malonic acid ($103.0037 -$
334 259.9258 ± 0.0014), 3. PIPES ($301.0522 - 301.0538 \pm 0.0004$), 4. Tartaric acid ($149.0092 -$
335 149.0118 ± 0.0005), 5. Citric acid ($191.0197 - 191.0207 \pm 0.0003$), 6. Malic acid ($133.0142 -$
336 133.0174 ± 0.0001), 7. Hippuric acid ($178.0510 - 178.0518 \pm 0.0013$), 8. Gluconic acid ($195.0520 -$
337 195.0523 ± 0.0009), 9. Iminodiacetic acid ($132.0291 - 132.0328 \pm 0.0008$), 10. Succinic acid ($117.0193 -$
338 117.0226 ± 0.0005), 11. Pantothenic acid ($218.1034 - 218.1034 \pm 0.0004$), 12. AMP ($346.0558 -$
339 346.0551 ± 0.0010), 13. Glutathione ($306.0765 - 306.0753 \pm 0.0010$), 14. Glutathione oxidized (611.1447
340 $- 611.1430 \pm 0.0006$), 15. Aspartic acid ($132.0302 - 132.0334 \pm 0.0009$), 16. Glutamic acid ($146.0459 -$
341 146.0482 ± 0.0006), 17. Nicotinic acid ($122.0248 - 122.0275 \pm 0.0008$). Clusters labeled with letters are
342 impurities (some of them are discussed in the text).

343 **Figure 5:** Comparison of the intensity obtained for the most intense profile from cluster d (left) and the
344 most intense profile from cluster 9 (right).

345 **Figure 6:** Cluster plots of the dataset obtained from the CE-TOF MS analysis of a urine sample at a
346 merging correlation coefficient of 0.95. The box highlights the clusters that are displayed in detail in
347 Fig. 7.

348 **Figure 7:** Normalized most intense profiles in the clusters marked in Figure 6 between 8.1 and 8.3 min.
349 List of m/z values 1. 135.0320 ± 0.0015 ; 2. 191.0539 ± 0.0022 ; 3. 165.0391 ± 0.0045 ; 4. 195.0507 ± 0.0031 ;
350 5. 178.0502 ± 0.0016 ; 6. 149.0471 ± 0.0015 ; 7. 179.0545 ± 0.0023 .

351

352

353 **References**

- 354 [1] A. Valdes, V. Garcia-Canas, C. Simo, C. Ibanez, V. Micol, J.A. Ferragut, A. Cifuentes,
355 Comprehensive Foodomics Study on the Mechanisms Operating at Various Molecular Levels in Cancer
356 Cells in Response to Individual Rosemary Polyphenols. *Anal. Chem.* 86 (2014) 9807-9815.
- 357 [2] M. Herrero, C. Simó, V. García-Cañas, E. Ibáñez, A. Cifuentes, Foodomics: MS-based strategies in
358 modern food science and nutrition. *Mass Spectrom. Rev.* 31 (2012) 49-69.
- 359 [3] C.C. Jacob, G. Dervilly-Pinel, G. Biancotto, F. Monteau, B. Le Bizec, Global urine fingerprinting by
360 LC-ESI(+)-HRMS for better characterization of metabolic pathway disruption upon anabolic practices
361 in bovine. *Metabolomics* 11 (2015) 184-197.
- 362 [4] S.D. Sarker, L. Nahar, Hyphenated Techniques and Their Applications in Natural Products Analysis,
363 In: *Methods in Molecular Biology*, 2012, pp 301-340.
- 364 [5] M. Petrović, M.D. Hernando, M.S. Díaz-Cruz, D. Barceló, Liquid chromatography–tandem mass
365 spectrometry for the analysis of pharmaceutical residues in environmental samples: a review. *J.*
366 *Chromatogr. A* 1067 (2005) 1-14.
- 367 [6] G.L. Erny, A. Cifuentes, Liquid separation techniques coupled with mass spectrometry for chiral
368 analysis of pharmaceuticals compounds and their metabolites in biological fluids. *J. Pharm. Biomed.*
369 *Anal.* 40 (2006) 509-515.
- 370 [7] W. Struck, M. Waszczuk-Jankowska, R. Kaliszan, M.J. Markuszewski, The state-of-the-art
371 determination of urinary nucleosides using chromatographic techniques "hyphenated" with advanced
372 bioinformatic methods. *Anal. Bioanal. Chem.* 401 (2011) 2039-2050.
- 373 [8] O.J. Pozo, J. Marcos, J. Segura, R. Ventura, Recent developments in MS for small molecules:
374 application to human doping control analysis. *Bioanalysis* 4 (2012) 197-212.
- 375 [9] K.K. Murray, R.K. Boyd, M.N. Eberlin, G.J. Langley, L. Li, Y. Naito, Definitions of terms relating
376 to mass spectrometry (IUPAC Recommendations 2013). *Pure Appl. Chem.* 85 (2013) 1515-1609.
- 377 [10] K. Kleparnik, Recent advances in combination of capillary electrophoresis with mass spectrometry:
378 Methodology and theory. *Electrophoresis* 36 (2015) 159-178.

379 [11] H. Gowda, J. Ivanisevic, C.H. Johnson, M.E. Kurczy, H.P. Benton, D. Rinehart, T. Nguyen, J. Ray,
 380 J. Kuehl, B. Arevalo, P.D. Westenskow, J.H. Wang, A.P. Arkin, A.M. Deutschbauer, G.J. Patti, G.
 381 Siuzdak, Interactive XCMS Online: Simplifying Advanced Metabolomic Data Processing and
 382 Subsequent Statistical Analyses. *Anal. Chem.* 96 (2014) 6931-6939.

383 [12] C.A. Smith, E.J. Want, G. O'Maille, R. Abagyan, G. Siuzdak, XCMS: Processing mass
 384 spectrometry data for metabolite profiling using Nonlinear peak alignment, matching, and identification.
 385 *Anal. Chem.* 78 (2006) 779-787.

386 [13] J.D. Holman, D.L. Tabb, P. Mallick, Employing ProteoWizard to Convert Raw Mass Spectrometry
 387 Data. *Curr. Protoc. Bioinform.* 46 (2014) 13.24.11-19.

388 [14] G. Mayer, L. Montecchi-Palazzi, D. Ovelheiro, A.R. Jones, P.-A. Binz, E.W. Deutsch, M.
 389 Chambers, M. Kallhardt, F. Levander, J. Shofstahl, S. Orchard, J.A. Vizcaino, H. Hermjakob, C.
 390 Stephan, H.E. Meyer, M. Eisenacher, H.-P. Group, The HUPO proteomics standards initiative- mass
 391 spectrometry controlled vocabulary. Database 2013:bat009 (doi:10.1093/database/bat009).

392 [15] S.M. Lin, L.H. Zhu, A.Q. Winter, M. Sasinowski, W.A. Kibbe, What is mzXML good for? *Expert*
 393 *Rev. Proteomics* 2 (2005) 839-845.

394 [16] P. Lampen, H. Hillig, A.N. Davies, M. Linscheid, Jcamp-Dx for Mass-Spectrometry. *Appl.*
 395 *Spectrosc.* 48 (1994) 1545-1552.

396 [17] L. Vergeynst, H. Van Langenhove, P. Joos, K. Demeestere, Accurate mass determination,
 397 quantification and determination of detection limits in liquid chromatography–high-resolution time-of-
 398 flight mass spectrometry: Challenges and practical solutions. *Anal. Chim. Acta* 789 (2013) 74-82.

399 [18] N. Huang, M. Siegel, G. Kruppa, F. Laukien, Automation of a Fourier transform ion cyclotron
 400 resonance mass spectrometer for acquisition, analysis, and E-mailing of high-resolution exact-mass
 401 electrospray ionization mass spectral data. *J. Am. Soc. Mass Spectrom.* 10 (1999) 1166-1173.

402 [19] D.M. Horn, R.A. Zubarev, F.W. McLafferty, Automated reduction and interpretation of high
 403 resolution electrospray mass spectra of large molecules. *J. Am. Soc. Mass Spectrom.* 2000, 11, 320-332.

404 [20] P. Kaur, P.B. O'Connor, Algorithms for Automatic Interpretation of High Resolution Mass
 405 Spectra. *J. Am. Soc. Mass Spectrom.* 17 (2006) 459-468.

406 [21] G.L. Erny, C. Simo, A. Cifuentes, V.I. Esteves, Introducing the concept of centergram. A new tool
 407 to squeeze data from separation techniques-mass spectrometry couplings. *J. Chromatogr. A* 1330 (2014)
 408 89-96.

409 [22] M. Katajamaa, M. Oresic, Data processing for mass spectrometry-based metabolomics. *J.*
 410 *Chromatogr. A* 1158 (2007) 318-328.

411 [23] G.L. Erny, A. Cifuentes, Simplified 2-D CE-MS mapping: Analysis of proteolytic digests.
 412 *Electrophoresis* 28 (2007) 1335-1344.

413 [24] R. Tautenhahn, C. Bottcher, S. Neumann, Highly sensitive feature detection for high resolution
 414 LC/MS. *BMC Bioinformatics* 8 (2008) 16.

415 [25] V. Andreev, T. Rejtar, H. Chen, E. Moskovets, A. Ivanov, B. Karger, A Universal Denoising and
 416 Peak Picking Algorithm for LC-MS Based on Matched Filtration in the Chromatographic Time
 417 Domain. *Anal. Chem.* 75 (2003) 6314 - 6326.

418 [26] R. Stolt, R. Torgrip, J. Lindberg, L. Csenki, J. Kolmert, I. Schuppe-Koistinen, S. Jacobsson,
 419 Second-Order Peak Detection for Multicomponent High-Resolution LC/MS Data. *Anal. Chem.* 78
 420 (2006) 975 - 983.

421 [27] S.-Y Wang, C.-H Kuo, Y.J. Tseng, Y. J. Ion Trace Detection Algorithm to Extract Pure Ion
 422 Chromatograms to Improve Untargeted Peak Detection Quality for Liquid Chromatography/Time-of-
 423 Flight Mass Spectrometry-Based Metabolomics Data. *Anal. Chem.* 87 (2015) 3048-3055.

424 [28] E. Tengstrand, J. Lindberg, K.M. Aberg, racMass 2-A Modular Suite of Tools for Processing
 425 Chromatography-Full Scan Mass Spectrometry Data. *Anal. Chem.* 86 (2014) 3435-3442.

426 [29] K. Aberg, R. Torgrip, J. Kolmert, O. Schuppe-Koistinen, J. Lindberg, Feature detection and
 427 alignment of hyphenated chromatographic-mass spectrometric data Extraction of pure ion
 428 chromatograms using Kalman tracking. *J. Chromatogr. A* 1192 (2008) 139 - 146.

429 [30] M. Katajamaa, M. Oresic, Processing methods for differential analysis of LC/MS profile data.
 430 BMC Bioinformatics 6 (2005) 179.

431 [31] S.C. Grace, S. Embry, H. Luo, Haystack, a web-based tool for metabolomics research. BMC
 432 Bioinformatics 15 (2014) 14.

433 [32] A.K. Jain, Data clustering: 50 years beyond K-means. Pattern Recogn. Lett. 31 (2010) 651-666.

434 [33] F. Fernandez-Albert, R. Llorach, C. Andres-Lacueva, A. Perera-Lluna, Peak Aggregation as an
 435 Innovative Strategy for Improving the Predictive Power of LC-MS Metabolomic Profiles. Anal. Chem.
 436 86 (2014) 2320-2325.

437 [34] B.D. Fitz, B.C. Reaser, D.K. Pinkerton, J.C. Hoggard, K.J. Skogerboe, E.R. Synovec, Enhancing
 438 Gas Chromatography–Time of Flight Mass Spectrometry Data Analysis Using Two-Dimensional Mass
 439 Channel Cluster Plots Anal. Chem. 86 (2014) 3973–3979.

440 [35] T. Acunha, C. Ibáñez, M.I. Pascual Reguera, M. Sarò, R. Navarro, J. Alfonso Redondo, H.
 441 Reinecke, A. Gallardo, C. Simó, A. Cifuentes, Potential of prodendronic polyamines with modulated
 442 segmental charge density as novel coating for fast and efficient analysis of peptides and basic proteins
 443 by CE and CE-MS. Electrophoresis 2015 (doi: 10.1002/elps.201400576).

Figure 1

Figure 2

Figure 3

Figure 4

Figure 5

Figure 6

Figure 7

TABLES

Table 1. Repartition of the 20 most intense profiles obtained in a single cluster at HL 0.8 in various cluster at HL 0.95. Each entry in this cluster correspond to the accurate mass and associated standard deviation of a PIP and the symbol of the cluster (Fig. 3, circle) where it was found at HL = 0.95.

<i>m/z</i>	@0.95	<i>m/z</i>	@0.95
132.0328±0.0008	9	318.9867±0.0007	NA
195.0523±0.0009	8	413.0939±0.0010	7
178.0518±0.0013	7	391.1103±0.0007	NA
317.9800±0.0013	e	472.9995±0.0025	e
134.0632±0.0005	d	273.9887±0.0010	NA
133.0342±0.0013	9	197.0547±0.0014	8
179.0554±0.0006	7	154.0145±0.0006	9
287.0498±0.0007	9	175.0258±0.0006	7
196.0556±0.0014	8	381.0000±0.0007	NA
303.0207±0.0017	NA	135.0669±0.0004	d