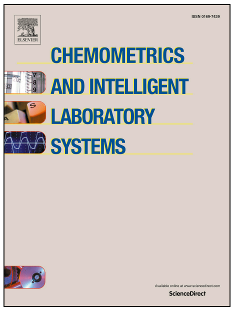


A data mining tool for untargeted biomarkers analysis: Grapes ripening application

Sandia Machado, Luisa Barreiros, António R. Graça, Ricardo N.M.J. Páscoa, Marcela A. Segundo, João A. Lopes

PII: S0169-7439(22)00256-8

DOI: https://doi.org/10.1016/j.chemolab.2022.104745

Reference: CHEMOM 104745

To appear in: *Chemometrics and Intelligent Laboratory Systems*

Received Date: 19 October 2022

Revised Date: 5 December 2022

Accepted Date: 22 December 2022

A data mining tool for untargeted biomarkers analysis: grapes ripening application

Sandia Machado[1], Luisa Barreiros[1,2], António R. Graça[3], Ricardo N. M. J. Páscoa[1], Marcela A. Segundo[1], João A. Lopes[4]

[1] LAQV, REQUIMTE, Departamento de Ciências Químicas, Faculdade de Farmácia, Universidade do Porto, Rua de Jorge Viterbo Ferreira 228, 4050-313 Porto, Portugal.

[2] Escola Superior de Saúde, Instituto Politécnico do Porto, Rua Dr. António Bernardino de Almeida 400, 4200-072 Porto, Portugal

[3] Departamento de Investigação e Desenvolvimento, SOGRAPE Vinhos S.A., Aldeia Nova, 4430-852 Avintes, Portugal

[4] Research Institute for Medicines (iMed-ULisboa) Faculdade de Farmácia, Universidade de Lisboa, Av. Prof. Gama Pinto, 1649-003 Lisboa, Portugal.

* corresponding author

E-mail: jlopes@ff.ulisboa.pt

Tel: +351 217946434

Fax: +351 217946470

Abstract

In metabolomics, data generated by untargeted approaches can be very complex due to the typically extensive number of features in raw data (with and without chemical relevance), dependence on raw data preprocessing methods, and lack of selective data mining tools to appropriately interpret these data. Extraction of meaningful information from these data is still a significant challenge in metabolomics. Moreover, currently available tools may overprocess the data, eliminating useful information. This work aims at proposing a data mining tool capable of dealing with metabolomics data, specifically liquid chromatography-mass spectrometry (LC-MS) to enhance the extraction of meaningful chemical information. The algorithm construction intended to be as general as possible in highlighting chemically relevant features, discarding non-informative signals specially background features.

The proposed algorithm was applied to an LC-MS data set generated from the analysis of grapes collected over a developmental period encompassing a 4-month period. The algorithm outcome is a short list of features from metabolites that are worth to be further investigated, for example by HRMS fragmentation for subsequent identification. The performance of the algorithm in estimating potentially interesting features was compared with the commercial MZmine software. For this case study, the MZmine output yielded a final set of 37 features (out of 1543 initially identified) with noise features while the proposed algorithm identified 99 systematic features without noise. Also, the algorithm required 2 times less user-defined parameters when compared to MZmine. Globally, the proposed algorithm demonstrated a higher ability to pin-point features that may be associated with grapes developmental and maturation processes requiring minimal parameters definition, thus preventing user uncertainty and the compromise of experimental information.

Keywords: Metabolomics; data mining; untargeted analysis; LC-MS; grape ripening

1. Introduction

Metabolomics is a scientific approach that aims to study the set of metabolites present in a sample, the metabolome [1, 2]. The metabolome is the result of the interaction between gene expression and environmental factors and its study may give rise to the characterization of phenotypes. Due to their participation in metabolic reactions, it is also possible to associate metabolites with physiological responses and use them as biomarkers [3, 4]. Mass spectrometry, used both for direct analysis or coupled to chromatographic techniques, has emerged as an extremely useful tool in metabolomic studies [5, 6].

In specific, liquid chromatography coupled to mass spectrometry (LC-MS) is considered a golden method because it allows efficient and selective separation and provides structural information of metabolites [5, 7]. In metabolomics, the metabolome profile can be assessed according to a targeted or untargeted analysis [2, 5]. Targeted analysis allows a sensitive detection of known metabolites and their quantification. This approach requires the use of standards, though its widespread use has increased the possibility of analyzing a wide range of metabolites, allowing, nowadays, the metabolome characterization on a large-scale [5]. The untargeted analysis performs a full scan of the sample giving access to the entire metabolome, such as in a footprint [8]. This approach allows global metabolomic profile analysis without a reasonable understanding of the sample composition and utilization of standards [5]. Nonetheless, the extraction of useful information from untargeted analysis can be a complex task due to the presence of signals without chemical relevance that may mask relevant features and introduce apparent complexity [9]. Signals generated by the ionization of metabolites such as ions-fragment, adducts and isotopes, and background signals, namely chemical and white noise are examples of information that should not be considered in the analyses [2]. In the absence of adequate data processing tools, it is very difficult to interpret, and discriminate signals associated with the metabolites of interest [5, 10]. Another factor that increases the complexity of the analysis is the dataset size and number of dimensions.

Each scan obtained at a specific retention time generates a mass spectrum featuring mass/charge values (m/z) and for each one a specific intensity [2, 5]. Additionally, shifts can occur in terms of m/z due to mass errors that are inevitable, even in the most sophisticated spectrometers, as well as in terms of retention time, which can cause misalignments, overlapping and even swapping in the elution order when comparing samples [11]. Also, missing values, defined here as signals that are not detected, may represent around 20% of all values in MS-based data sets. This means that, depending on the performance of the spectrometer, some metabolites may not be detected despite their presence in the sample. Due to these factors, intrinsic variability between runs of the same sample is often observed [2, 5, 12, 13].

There are several tools described in the literature designed to handle data from untargeted analyses that attempt at the extraction of chemically relevant features [2, 14]. However, most tools include extensive processing, such as noise removal, baseline subtraction, peak detecting, isotope removal, signal alignment or matching, identification, and normalization, that greatly reduce the complexity of the data but increase the risk of removing meaningful information from data [5, 15]. For instance, baseline subtraction can affect peaks shape due to the application of derivatives that smooth the signals. Similarly, noise removal requires setting a threshold, which may cause the elimination of low intensity signals, though with potential chemical relevance [16]. In addition, many tools require user-defined parameters that often cause uncertainty for the user in the definition of the most suitable parameters. [12, 17]. Some tools provide default values for these parameters, but users may also find it difficult to assess their suitability to specific data sets and how they impact the results [2, 12]. For example, many tools incorrectly assume that the order of elution is maintained, in the step of matching the signals [5, 12]. This step should not be performed through systematic alignment, but through a mapping of identical characteristics because the variation factor is not stable across the chromatogram [12]. There are few comparative assessments among tools because many of them are not described with sufficient detail for their comprehension and implementation [12]. Also, some of the comparative studies have reported different results with the same data but using different tools [18]. These challenges still make the topic of mass spectrometry-based metabolomics a current subject of critical review [1-3, 10, 13, 14, 18-21].

In this work, we present a data mining tool developed to reduce the complexity of data from untargeted LC-MS analyses. It distinguishes from other available tools in the way it overcomes some of the concerns mentioned above. This data mining tool is based on an algorithm which processes LC-MS data and transforms them into a simple and intelligible data matrix that highlights the chemically relevant features requiring minimal parameters and easily defined by the user. The proposed algorithm was applied to the study of the grape metabolome. The idea is to understand the timeline evolution of the grapes metabolomic profile and assess the typical metabolites at each developmental stage. The algorithm, in specific, has the purpose of showing the features that may be associated with these typical metabolites and allowing the optimization of the subsequent fragmentation process by MS^n and use of high-resolution mass spectrometry for the identification of metabolites. The efficiency of the developed algorithm was evaluated submitting the same data to the open source MZmine software version 2.53 and comparing the approaches.

2. Material and methods

2.1. Chemicals

Acetonitrile (LiChrosolv LC-MS grade), methanol (LiChrosolv LC-MS grade) and formic acid were acquired from Merck (Darmstadt, Germany). Chloroform stabilized with c.a. 0.6% ethanol (AnalaR NORMAPUR) was obtained from VWR International S.A.S. (Fontenay-sous-Bois, France). Water from arium water purification system (resistivity >18 MΩ cm, Sartorius, Göttingen, Germany) was used for the preparation of all solutions. 5-Carbamimidamido-2-(2,2-diphenylacetamido)-N-[(4-hydroxyphenyl)methyl] pentanamide (BIBP-3226), used as internal standard, were purchased from Tocris (Bristol, UK). BIBP-3226 stock solution was prepared in a water:methanol mixture (33:67, v/v) to achieve a final concentration of 3 µg mL^{-1}. Stock solution was stored at -20 °C.

2.2. Sample preparation

Grape samples (bunches) of Touriga National variety were collected from eight different locations of a vineyard in a Portuguese wine region (Dão region) during different developmental stages, including green grapes (collected in June and July) and mature grapes (collected in August and September) in order to assess the metabolomic changes during grape evolution. Samples were prepared using a protocol previously described [22]. Briefly, the samples were grounded into powder with liquid nitrogen and 2 g were extracted using a mixture of water:methanol:chloroform (20:40:40, v/v/v). The chloroform fraction was discarded and the aqueous methanolic fraction was filtered through a 0.2 µm PTFE filter. The aqueous methanolic fraction was used for direct analysis of polar metabolites with liquid chromatography coupled to mass spectrometry using an untargeted approach. Prior to analysis, the internal standard BIBP-3226 was added to every 32 grape samples. A blank sample was prepared using only the grape sample medium (water:methanol mixture) in the ratio used in the extraction (33:67, v/v), and a standard solution was prepared by adding BIBP-3226 to the water:methanol mixture (33:67, v/v).

2.3. Instrumental analysis

Grape samples, blanks and standard solutions were analyzed by liquid chromatography coupled to mass spectrometry (LC-MS). Chromatographic analysis was performed in a Nexera X2 UHPLC system comprising two LC-30AD pumps, a DGU-20A5R degassing unit, a SIL-30AC autosampler and a CTO-20AC oven (Shimadzu Corporation, Kyoto, Japan). The MS system was a triple quadrupole LCMS-8040 mass spectrometer equipped with an electrospray ionization source (ESI) (Shimadzu Corporation). Chromatographic separation was performed using a reversed-phase RRHD Eclipse Plus C18 column (1.8 µm, 2.1 × 100 mm; Agilent, California, USA) at 40°C and using elution in gradient mode with a flow rate of 0.3 mL min^{-1}. The mobile phase was constituted by water as solvent A and acetonitrile as solvent B, both containing 0.1% (v/v) formic acid. The gradient was established as follows (min/A%): 0.0/95, 4.0/80, 5.7/80, 12/55, 14.7/0, 17.3/0, 20/95, 22/95. The injection volume was 2 µL and the samples were kept at 6°C during analysis. The analysis by mass spectrometry was performed in scanning mode (untargeted approach) to maximize the number of detected instrumental signal. The spectra were collected in positive ionization mode over a range of 100-1000 m/z and a scan speed of 13000 m/z s^{-1}. The analysis of the 32 grape samples (8 sampling locations in the same vineyard × 4 months) was performed once and sorted by 4 groups, corresponding to each month. An injection of blank was included in the middle of each group and at least two blanks before and after a new group, with a total of 16 blank injections. The following parameters were used for analysis: nebulizing gas (N_2) flow rate at 2.6 L min^{-1}, desolvation gas (N_2) flow rate at 15 L min^{-1}, desolvation line temperature at 300 °C, heat block temperature at 425 °C, capillary voltage: 3.5 kV in positive mode.

2.4. Data analysis

The developed algorithm was created and executed in MATLAB version R2021b (MathWorks Inc., Massachusetts, USA) in combination with PLS Toolbox Version 7 (Eigenvector Research Inc., Manson, WA), using the raw data files in a CDF format. Open source MZmine software, version 2.53, developed by MZmine Development Team and distributed by the website http://mzmine.github.io/download.html, was chosen to analyze the data processed by the algorithm. This procedure was performed to compare the results from both methodologies and to evaluate the effectiveness of the proposed algorithm [23].

2.5. Algorithm development

The algorithm was conceived to highlight only the features with potential chemical interest, ensuring minimal processing to avoid distortion of the original instrumental signals and minimal compromise of experimental information. The algorithm does not assume any threshold, allowing the inclusion of possible interesting low-intensity compounds that may be hidden in the noise, something that is normally excluded from many available software [10]. The output of the algorithm is a data matrix that was designed to provide an intuitive and simple overview of the features present across multiple samples.

The algorithm comprises different steps, such as the database preparation, the processing of information through a programming code created in MATLAB and the treatment of the final data matrix. Before starting the code development, the database was cleaned up. This step consists in organizing a database by indexing the variables to access them easily. A flowchart plan was drawn to code functions, methods, and variables in order to extract information from the database (**Fig.1**). The code was written in the MATLAB program and comprises two main stages: 1) signal pre-processing and 2) benchmarking. A treatment to the matrix generated by the code was performed removing possible noise that may have remained and removing isotopes in order to simplify the data.

2.5.1. Programming code

The code proceeds as follow (**Table 1**):

a) Reset the variables analyzed before (stored in samples and blanks).

b) User-defined samples based on the index of the previous organized databased. It is possible to add more samples in each set or more sets (according to the blanks).

c) User-defined blanks based on the index of the previous organized databased. It is possible to add more blanks in each set or more sets (according to the samples).

d) Samples and blanks are compiled. Three "for loops" were created to access the database information of the samples and blanks defined in the previous step.

e) All the samples and blanks are analyzed in a loop by the function "vid_peaks". This function allows to detect all the peaks from samples and blanks and to remove chemical noise from samples. This function results are stored in "R". "Result1" is the database containing all the information concerning the retention time in which the scans were performed, the m/z values that were detected in each scan and their respective intensity. "S{k,1}" and "S{k,2}" are the information compiled previously based on the user-defined samples and blanks (pre-processing stage). "Flag" is a user-defined parameter to display or not an image of the results, "1" or "0", respectively. "mz" is a user-defined parameter to set the range of m/z values. Peaks in samples and in blanks are respectively detected by “peakparamS” and “peakparamB”, when both of this user-defined parameters are set with a width, tolfac and w. To remove chemical noise, peaks from blanks are subtracted from the samples regarding the user-defined parameter “tol” comprised by the tolerance of m/z, retention time and intensity. In cases where more than one blank is used in the subtraction process, the “flagblank” user-defined parameter defines if the peaks to be considered are the total present in the blanks (sum = “1”) or only the common ones between blanks (intersection = “0”)

f) The function “vid_new” allows to compare and align the results (Benchmarking step) generated by the previous step (“R”) in a data matrix which are stored in “Result”. Whenever different samples have the same m/z value and have a retention time less than the defined tolerance (“RT tol”), they are aligned in a matrix line together with information on their intensity and number of peaks and points.

2.5.1.1. Pre-processing: peak detection and noise removal

The pre-processing stage is composed by two steps, specifically peak detection, and noise removal, which combined allows to highlight the chemically relevant features from each sample.

2.5.1.1.1 Peak detection

The first step is to make the detection of peaks from each individual spectrum to generate a matrix that encompasses all the features with the corresponding retention times and intensities. This step is performed using the *peakfind* function from the PLS Toolbox, a tool for use within the MATLAB environment, and allows the automatic identification of peaks through the definition of three parameters, namely, *width*, *tolfac* and *w*. *Width* is the number of points in the *Savitzky-Golay* filter. The *tolfac* parameter is defined by the tolerance of the residues. Peaks are selected when their intensity are higher than the product of *tolfac* and the residues. The *w* parameter is the width of the window that allows the detection of local maxima. This function defines the center of each peak, allocating the retention time and intensity at this point, thus reducing the size of data. For each detected peak, the output displays the number of points that surround the local maxima.

2.5.1.1.2. Noise removal by subtracting blanks

The noise removal step aims to discard mainly background signals, namely chemical noise (ionization of mobile phase ions) from samples by subtracting the blank signal [15, 17]. This strategy aims to compare samples with blanks to exclude common characteristics that do not provide relevant information about the sample composition [2, 17]. These common features are most likely associated with chemical noise considering the fact that blanks and samples share the same features generated by the mobile phase (water:acetonitrile:formic acid), by their common solvent (water:methanol), and by the interaction of both. Therefore, features with the same m/z value and within a specific retention time tolerance were considered common. However, by coincidence there may be the possibility that relevant features have the same m/z value as features from noise, particularly when using generated data by low resolution MS. In order not to erase these relevant features, a differentiation is made based on their intensity, since this must be higher when compared to the noise feature. The tolerance of the retention time was not considered very important, because noisy instrumental signals may be present over a longer period of the chromatographic run and usually with an erratic distribution. Therefore, the tolerance to be used must be large and can even have the value of the entire chromatographic run. This is only possible because the algorithm also considers the intensities, which can prevent an instrumental signal from a chemical species present in the sample to be eliminated by a noise instrumental signal from the blank with the same m/z value but present in a completely different retention time.

2.5.1.2. Benchmarking

The next stage, designated benchmarking, allows the alignment of the previous preprocessed data. To be considered as the same feature in different samples, the m/z value must be equal and the retention time must be similar within a given tolerance. This procedure generates a data matrix in which each row represents a feature with indication of the m/z value in the first column and the information of retention time and intensity annotated in the columns of the sample where the feature appears. This alignment allows to examine several samples simultaneously (without a limit number), to make a comparative study among the metabolomic profiles. The recognition of certain metabolites within a specific group of samples (e.g., pertaining to a month of development) allows to suggest them as a response to the conditions of that group, which can be useful for biomarkers search.

2.5.2. Data treatment

2.5.2.1. Noise removal regarding peaks and points

This strategy allows to reduce the noise that may have remained from the previous noise removal process, improving the elimination of all sample noise. Besides removing chemical noise, this procedure also removes white noise, which are random signals generated by interferences from the instrumental system [15, 17]. This step is performed considering the number of peaks and the number of points that constitute a peak, based on the empirical observation that is possible to used them to assess different types of signals. The peak of an analytical signal, for example, is formed by several continuous points with a relationship in terms of intensities (Gaussian peak) [2, 16, 24]. Therefore, peaks with higher areas are defined by a higher number of points. White noise is characterized by a random manifestation of a single point with no width or peak shape. Chemical noise, as a representative signal of the mobile phase, can be present continuously during the chromatographic run, without sudden variations in intensity and fitting into the baseline structure. Like an analytical signal, this type of signal can be translated in Gaussian peaks, as small variations in intensity can also lead to the detection of local maxima. But, unlike an analytical signal, these peaks end up being small, defined by a low number of points and, due to the constant presence in the chromatographic run, this signal is characterized by a higher number of peaks (**Fig. 2**). Therefore, an appropriate definition of the number of peaks and points can help to discriminate noise features, so that they can be excluded from samples. This process is performed after the sample alignment step to minimize erratic elimination of relevant analytical signals with low number of points (possibly associated with low-intensity compounds), as these could be confused with noise signals, according to the *modus operandi* of this strategy. As in the alignment step, each feature can be common between different samples, it is enough to prove the relevance of the feature in one of the samples (meet the peaks and peak requirements), so that it is not excluded from the other samples, even if these do not meet the requirements. Therefore, this process allows keeping the relevant features present at low intensities. This step can be performed using a spreadsheet program and is performed defining, for a specific feature, the maximum number of points and the minimum number of peaks that was detected considering all samples where it appears. These values are later used to verify compliance with peak and point requirements and if not even one sample meets the requirements, the respective feature is deleted in all the samples.

To assess the typical number of peaks and points of noisy analytical signals, random blanks, composed predominantly by chemical noise, were processed by the code so that they could be further and easily studied.

2.5.2.2. Isotope removal

To simplify the data matrix generated by the code, isotopes were excluded as they do not provide useful information in this work. However, the information about its existence was maintained by marking the molecular ion in blue colors when isotopes were associated. This marking helps to easily visualize possible major compounds, as isotopes are normally associated to more intense features.

The parameters used to locate isotopes and remove them, was defined studying the feature profile of the internal standard.

2.6. Evaluation of the algorithm performance

To evaluate the effectiveness of the algorithm, the same data was processed by the MZmine software as similar as possible to the algorithm, to compare the results. As suggested by the MZmine instructions, several steps are required to process raw data. First, *Baseline correction* was applied. The option *RollingBall* was chosen as baseline corrector for allowing the descent of the baseline without visually changing the structure of the peaks (**Fig. 3**). The second step applied and considered one of the most critical is *Mass detection*, as it allows the detection of the peaks present in each scan, creating a mass list with the features present above a noise level defined by the user. Noise level defines the minimum intensity below which masses are excluded and should correspond to the noisy area. *Mass detection* was performed using the *Centroid mass detector* algorithm, as it is the only option for raw data that is already centroid. The other algorithms work only with continuous type data. Subsequently, *Chromatogram builder* was applied to connect the points of the mass list and build the chromatograms of each feature. To separate each extracted ion chromatogram into different peaks, the step of *Chromatogram deconvolution* was required. To finish, *Alignment* was applied to align peaks from different samples as performed in the algorithm. This alignment applies a match score calculated based on the mass and retention times tolerance defined by user. In all these steps, the user is asked to define different parameters. To compare the results, when possible, parameters of MZmine were defined as similar as possible to those of MATLAB, specifically the tolerances for the retention time and m/z value.

3. Results and discussion

A total of 8 grape samples collected in 4 different months (n=32), were analyzed by the proposed algorithm to identify typical features of each stage of grape development.

Before the code analysis, the m/z values were rounded to the nearest unit since uncertainty can be 0.10 in most abundant peaks (**Fig. 4**) and 0.25 in the less abundant peaks due to the equipment's scan accuracy. A highest precision for m/z values was not considered necessary, because, on one hand, the aligning in terms of retention time will already reduce the likelihood of matching equal m/z values that come from different metabolites, and on the other hand because the algorithm was not developed for identification purposes, but to give an overview of the areas with metabolic changes for further fragmentation and identification using MS/MS.

3.1. Definition of parameters

3.1.1. Peak detection (pre-processing)

Peakfind function was applied to six samples for the detection of all peaks through the definition of *width*, *tolfac* and *w*. Blanks were also submitted to this process to allow the comparison with samples and to exclude chemical noise. Width values were tested between 0 and 15. As can be seen in **Fig. 5a** a lower *width* (5 points) provided the definition of smaller peaks, maintaining its originality, and led to a higher number of features (77) contrary to what happens with a larger *width* (15 points) that lead to an inferior number of features (63) (**Fig. S1**). Therefore, an intermediate *width* (10 points), as shown in **Fig. 5b**, was chosen because a lower *width* can also cause the separation of the largest chromatographic peaks, being one of the reasons for the highest number of features, as can be seen by the duplicates in the **Fig. S1**. As the result of this function represents only the center of the peak, the smallest *width* was chosen for blanks (3 points) in order to segment each peak, as much as possible, and reflect all the typical mobile phase signals in the maximum of retention times. This is especially useful in the next procedure, when samples are compared with blanks to exclude features from noise. A value of 3 was chosen for *tolfac* and *w* parameters as it is the recommended default of the *peakfind* function.

3.1.2. Noise removal by subtracting blanks (pre-processing)

Regarding the first strategy to remove noise by subtracting the common features between samples and blanks, parameters related to m/z value, intensity and retention time tolerance were chosen for the present data. For the m/z value, no tolerance was used, which means that if a feature from a sample is found in a blank with the same m/z value, this feature is excluded. Nevertheless, if the intensity is 3 times higher in the sample compared to the blank (signal-to-noise ratio of 3:1, defined as limit of detection - LOD), the feature should not be excluded as it can be an analytical signal. In the present case, 5 minutes was chosen as the retention time tolerance. If in the current data the entire runtime were used as tolerance, the intensity parameter might not be effective, since the gradient mode used in this analysis is normally associated with a constant increase in baseline intensity.

3.1.3. Benchmarking

Although the samples, in the present case, are from 4 different months of development, the alignment was performed for the total of samples since a feature can be present for more than a month and it is important to have this information aligned. Subsequently, the groups of each month were highlighted for a better interpretation of the information. A feature is aligned whenever it has the same m/z value across different samples and the same retention time, within a specific tolerance. This tolerance was chosen examining the retention time variation of a specific feature among different samples. The assessment of this variation was performed using the internal standard (m/z=474) as it corresponds to a compound added to all samples (**Table S1**). A tolerance value of 0.10 min was set, corresponding to the maximum value observed for the retention time (RT) variation. Nevertheless, this value can be increased when the intensity of a peak is lower (e.g., m/z=476), probably due to the greater difficulty in defining the center of the peak.

3.1.4. Noise removal regarding peaks and points (data treatment)

For the proper choice of parameters for this strategy of noise removal, more specifically, to eliminate the noise features from samples just regarding the number of peaks and points, blanks, predominantly composed by noise, were analyzed to understand the profile of this type of signals. Results were obtained considering the data from 4 blanks, randomly chosen, and which combined contain 14246 features.

Regarding white noise, the parameters were defined to remove peaks with just 1 point, as it cannot be assigned as a random signal or generated by a metabolite.

Regarding the number of points that define a peak, the results showed that, of the 14246 features present in blanks, 14197 have 6 or less points, which is equivalent to a percentage of 99.7% of total data and there were no noise peaks with 9 or more points. Regarding the peaks with between 6 and 9 points, 75.7% of the peaks with 7 points and 33.3% of the peaks with 8 points were considered noise. This indicates that a smaller number of points in a peak increases the possibility of a m/z value being noise. This is predictable, as the noise constituted the baseline structure whose signal is composed by low intensity peaks and, therefore, by a low number of points. As features of interest are normally more intense, having more points can help to distinguish them from noisy features when both have the same m/z value.

Regarding the number of peaks, the results showed that, in the total run, 98.7% of the features from blanks have more than 10 peaks. This indicates that a greater number of peaks increases the possibility of a m/z value being noise. This is predictable, as the noise is part of the mobile phase and therefore appears repeatedly in the chromatographic run.

It is important to consider the number of peaks and points strategies together, as one increases the confidence of the other, resulting in a more efficiently noise exclusion, without compromising the relevant information. However, to define a compromise solution between peaks and points it is important to keep in mind that the probability of including noise in the results increases as the number of points decreases and the number of peaks increases.

As in the present study, it is intended to obtain a general idea of the typical compounds of each stage of development, we believe that the use of 9 points is sufficient and ensured the exclusion of noise more safely and without the need to defining a number of peaks.

To increase the amount of data, the procedure was also tested including features with 8 points. However, given the higher probability of this features being noise (33%), we decided to set a low number of peaks, specifically 3 (included), to increase the likelihood of not including noise.

Lower point numbers, for example 7, were not tested because the probability of it being noise is at least double compared to the 8 points, creating a higher risk of including noise in the data.

3.2.2. Isotope removal (data treatment)

For the proper choice of parameters to locate isotopes and remove them, the feature profile of the internal standard was studied. When isotopes are involved, an isotopic pattern can be observed, in which the isotopes are less intense than the molecular ion and continue to decrease as their mass/charge value increases, as can be verify by the molecular ion (m/z=474.35) and respective isotopes (m/z=475.35 and 476.40) of BIBP-3226 compound used as IS (**Fig. 6**). Also, it is possible to verify that the difference between neighboring isotopes is one neutron. The approximate mass of a neutron is 1.008665 Da, however this small difference only become significant with high-resolution MS data, which is not the case. Therefore, features present at the same retention time with a difference of 1 m/z value between them are considered part of an isotopic pattern. Among these, the one with the highest intensity is considered the molecular ion. This strategy allowed the recognition of 24 isotopes.

3.2. Application to a dataset

The application of the algorithm generated a data matrix that summarize the features present in the 32 samples. Since 2 different parameters were used for noise removal using the peaks and points (**section 3.1.4.**), 2 different results were obtained. Through this step, noise was removed eliminating all features that have less than 9 points provided a data matrix with 91 features. Removing features with less than 8 points and more than 3 peaks, increases by 8 features the matrix obtained previously (99 features) (**Table S2**). This last matrix was used for further studies as it contains more information. After removing the isotopes, the data matrix with 99 features was reduced to 75 features (**Table 2**). As the objective of this study was to focus on the metabolomic profile changes during the developmental stages, we chose to consider the result of the 8 samples together for each month. As the grapes were collected in different locations, features in common can be more confidently associated to the grape stage than features that only appear in one site, as they can be an artifact or caused by other reasons besides the ripening of the grape (for example, response to the terroir). Thus, Table 2 is summarized, providing, for each month, the number of samples that presented a specific feature (frequency). When present, this information is accompanied by the RT average and the maximum intensity found for each month. The formation and transformation of specific metabolites during grape development can occur continuously and for this reason a feature can appear in more than one month with decreasing or increasing intensity. When the intensity of a feature decreases to lower levels, the equipment may not be able to detect it causing the feature to be found in less samples (lower frequency). This can help to verify the month in which a particular feature is mostly found, allowing its use as a biomarker of this development phase. Therefore, data were ordered and marked in two shades of green according to features frequency in more or less than 50% of the samples. Features present in at least 5 samples were marked in dark green and in light green those that appear at most in 4 samples (**Table 2**). Considering that the June and July grape samples were green and that the August and September grape samples were already ripe, the data were ordered considering the predominant features in each of these two developmental stages. As shown in **Table 2**, this ordering allows a clear distinction of the typical features of each developmental stage. Each of these two groups were then organized in terms of retention time to highlight groups of features that can be associated with the same compound.

3.2. Evaluation of the algorithm performance

As the composition of samples is unknown, one of the biggest challenges in the analysis of untargeted data is the discrimination between features of interest and those originated from noise. The idea of creating this tool came precisely from the need to work around this problem and minimize the amount of noise in the data without an extensive processing.

The performance of the algorithm was evaluated comparing the results with those obtained by the MZmine processing and verifying the amount of noise and chemically relevant features presented in each one of the tools.

Blanks were initially investigated to define the typical noisy features. Since there is a large number of features in blanks, we chose to rely on the features present in at least 2 of the 4 blanks randomly chosen. **Table S3** summarizes the most evident noise features from the blanks. This table present the m/z value of the feature, the frequency (number of samples) which the feature appears, and the number of peaks found for each feature and the duration of the feature appearance (initial and final RT).

3.2.3. Results comparison

For the MZmine processing, two strategies were tested to remove the maximum of noise from the data, specifically using the noise level parameter and applying the *Alignment* step to align samples with blanks and to exclude the common values between both (possibly belonging to the chemical noise). Regarding the noise level parameter, two values were tested, specifically a null noise level to verify the possibility of avoiding the use of this processing and a noise level based on the maximum intensity of the noise line before the void volume (no analytical peak present). For the null noise level, a matrix with 1543 features was generated of which at least 75% of the features belonged to noise. In the case of the noise level defined by the maximum intensity before the void volume, the result gave 143 features, and the same percentage of noise was found. The application of different noise levels showed a strong influence over the number of detected features but proved to be ineffective in excluding noise. The presence of noise in the last case may be due to the fact that the baseline is not constant in terms of intensity during the chromatographic run, especially since it was operated in a gradient mode. The baseline correction was expected to minimize this problem, but according to the results it was not enough since several noise features remained in the sample. The possibility of a highest noise level was not considered because typical noise features were found with intensities in the same order of magnitude as the intensity of some chromatographic peaks, with the feature with m/z value 143 (pertaining to noise) being the base peak almost in the entire run (**Fig. S2**). Therefore, the increase of the noise level would not only exclude noise but could also exclude features of interest.

The strategy of aligning the samples with blanks to exclude the common features was left to last, because, contrary to what our algorithm allows, the exclusion of common features would be done without considering the intensities. This means that if the sample has a feature equal to a feature from blank it would be excluded even if it has an intensity three times greater, which could be indicative of a relevant feature. However, without this step, features from noise would remain in results and it would be impossible to discriminate them the relevant features. To consider the intensities, this process would have to be carried out manually, which would become unfeasible mainly for large amounts of data. The alignment of the samples with the blanks and the subtraction of the common features was tested in both noise levels previously defined. The purpose of performing this procedure on data with the null noise level was to understand whether this, by itself, would be sufficient to exclude noise, in order to avoid extensive data processing. This gave rise to a matrix with 477 features, of which approximately 21% belonged to noise. As this option still contains features from noise, the step of *Alignment* was performed in the data already processed with the noise level to verify the effective of both strategies in excluding the noise in its entirety. However, one feature from the noise remained in a total of 37 features (**Table 3**). Regarding that both of June and July samples were green grapes and that samples from August and September were already mature grapes, the data were sorted firstly by the predominant features in each of these two phases and subsequently by the retention time to highlight groups of features that can be associated with the same compound.

Regarding the 99 features obtained by the proposed algorithm (without excluding isotopes), none corresponded to the typical noisy features. Comparing the results with those that present less noise in MZmine, it is possible to verify that the algorithm was able to find almost three times more features, some of which were also detected in more samples than in MZmine. This means that, in the present case, the algorithm allowed the extraction of more information when compared to MZmine, and it integrates less noisy features in the results. This probably happens because the algorithm does not exclude signals based on the intensity and so it has the ability to bypass the noisy features including the more intense ones and accessing a larger number of peaks that would be excluded if a noise level was defined.

Regarding the implementation part of the two tools, MZmine includes 6 processing steps and requires 21 user-defined parameters, and the proposed algorithm includes 4 processing steps and requires 9 user-defined parameters (**Table 4**). Due to the highest number of processing steps and user-defined parameters, it was possible to ascertain that the processing performed by the MZmine may increase the possibility of compromising the experimental information. Also, some of the parameters may be more difficult to set by the user in the case of MZmine. This is a very important aspect because, as previously mentioned, if the user does not know which inputs are most suitable for the data, one can skew the results. In the case of algorithm, the definition of parameters does not require the same effort from the user, as they can be easily estimated from a simple observation of the data.

Additionally, some frameworks have been proposed for denoising and feature extraction concerning mass spectrometry data. For instance, an alternating direction minimization based denoising framework for extracted ion chromatograms has been proposed and successfully applied to proteomic analysis, with enhanced suitability for quantitative tasks [25]. Nevertheless, this is only suitable for data preprocessing, as feature selection is not included in the algorithm. Recently, a suite of R language-based software enabled feature extraction and importance ranking, following an untargeted metabolomics approach for wine analysis [26]. However, different software packages must be used for preprocessing and peak detection. Also using MATLAB, the Finnee toolbox allows the plotting of mono-dimensional representations and the profiling of spectra along the separation from X-MS data (where X represent any separative technique such as LC, capillary electrophoresis or GC) [27]. Additionally, it can convert the original continuous profile to a discrete spectrum (centroidization) or to a chromatographic based dataset but it is not tailored for feature extraction as in the algorithm proposed here.

4. Conclusions

The aim of this study was to create a tool to turn complex untargeted LC-MS data into an intelligible matrix minimizing noise and highlighting relevant features. Algorithm was applied to LC-MS data of grape samples to highlight the metabolic differences among developmental stages and to find possible biomarkers associated to grape ripening, which can be further submitted to fragmentation for identification purposes. The strategy of alignment of the algorithm allows the comparison of different samples and resuming the information by groups related to the month of development allow to surpass missing values. Missing values are features that exist in a sample but are not detected for reasons inherent to the equipment. When only one sample is studied, a missing value is naturally excluded, but considering it as part of a group, it is possible to detect its absence in a sample if it exists in the other samples of the same group. Although the algorithm was applied to data with a significantly high baseline and no chromatographic analysis replicates, it was nonetheless more efficient generating a matrix with relevant information and free from noise when comparing to MZmine. The proposed algorithm has the advantage of quickly obtaining results without the need of a detailed study of the data to define parameters and it does not work by excluding features based on intensity, having the ability to maintain relevant features present below the noise level. This algorithm shows to be useful handling data from untargeted analyses without an extensive preprocessing that could compromise experimental information. Although it remains a method that, like the others, cannot guarantee that no important feature is lost, it has been shown to identify a higher number of relevant features, basically by increasing its ability to keep minor compounds, contributing with a new routine for the metabolome toolkit.

CRediT author statement

Sandia Machado: Software, Formal analysis, Investigation, Data Curation, Writing - Original Draft Preparation, Visualization; Luisa Barreiros: Investigation, Data Curation, Writing - Review & Editing; António R. Graça: Resources, Writing - Review & Editing; Ricardo N. M. J. Páscoa: Validation, Investigation, Writing - Review & Editing; Marcela A. Segundo: Conceptualization, Validation, Writing - Review & Editing, Visualization, Supervision, Project administration, Funding acquisition; João A. Lopes: Conceptualization, Methodology, Software, Writing - Review & Editing, Supervision, Project administration, Funding acquisition.

Acknowledgements

This work received financial support from the European Union (FEDER funds through COMPETE POCI-01-0145-FEDER-016735) and National Funds (FCT, Fundação para a Ciência e a Tecnologia) through project PTDC/AGR-PRO/6817/2014 and also under UIDB/50006/2020. This work was also supported by AgriFood XXI I&D&I project (NORTE-01-0145-FEDER-000041) cofinanced by European Regional Development Fund (ERDF), through the NORTE 2020 (Programa Operacional Regional do Norte 2014/2020). S. Machado thanks FCT and POCH (Programa Operacional Capital Humano) for her PhD grant (SFRH/BD/122730/2016). J. Lopes received additional support from FCT-MCTES (UIDB/04138/2020). L. Barreiros and R.N.M.J. Páscoa acknowledge funding from FCT through program DL 57/2016 – Norma transitória.

References

[1] M. Creydt, M. Fischer, Food Phenotyping: Recording and Processing of Non-Targeted Liquid Chromatography Mass Spectrometry Data for Verifying Food Authenticity, Molecules, 25 (2020) 3972. https://doi.org/10.3390/molecules25173972.

[2] C. Schiffman, L. Petrick, K. Perttula, Y. Yano, H. Carlsson, T. Whitehead, C. Metayer, J. Hayes, S. Rappaport, S. Dudoit, Filtering procedures for untargeted LC-MS metabolomics data, BMC Bioinform., 20 (2019) 334. https://doi.org/10.1186/s12859-019-2871-9.

[3] O.C. Zeki, C.C. Eylem, T. Recber, S. Kir, E. Nemutlu, Integration of GC-MS and LC-MS for untargeted metabolomics profiling, J. Pharm. Biomed. Anal., 190 (2020) 113509. https://doi.org/10.1016/j.jpba.2020.113509.

[4] A. Raza, Metabolomics: a systems biology approach for enhancing heat stress tolerance in plants, Plant Cell Rep., 41 (2020) 741-763. https://doi.org/10.1007/s00299-020-02635-8.

[5] E. Gorrochategui, J. Jaumot, S. Lacorte, R. Tauler, Data analysis strategies for targeted and untargeted LC-MS metabolomic studies: Overview and workflow, TrAC - Trends Anal. Chem., 82 (2016) 425-442. https://doi.org/10.1016/j.trac.2016.07.004.

[6] M. Doppler, C. Bueschl, F. Ertl, J. Woischitzschlaeger, A. Parich, R. Schuhmacher, Towards a broader view of the metabolome: untargeted profiling of soluble and bound polyphenols in plants, Anal. Bioanal. Chem., 414 (2022) 7421–7433. https://doi.org/10.1007/s00216-022-04134-z.

[7] S. Heiles, Advanced tandem mass spectrometry in metabolomics and lipidomics-methods and applications, Anal. Bioanal. Chem., 413 (2021) 5927-5948. https://doi.org/10.1007/s00216-021-03425-1.

[8] A. Razzaq, B. Sadia, A. Raza, M. Khalid Hameed, F. Saleem, Metabolomics: A Way Forward for Crop Improvement, Metabolites, 9 (2019) 303. https://doi.org/10.3390/metabo9120303.

[9] R. Wang, H.C. Ji, P. Ma, H.T. Zeng, Y.M. Xu, Z.M. Zhang, H.M. Lu, Fast pure ion chromatograms extraction method for LC-MS, Chemom. Intell. Lab. Syst., 170 (2017) 68-74. https://doi.org/10.1016/j.chemolab.2017.10.001.

[10] E. Gorrochategui, J. Jaumot, R. Tauler, ROIMCR: a powerful analysis strategy for LC-MS metabolomic datasets, BMC Bioinform., 20 (2019) 256. https://doi.org/10.1186/s12859-019-2848-8.

[11] C. Brunius, L. Shi, R. Landberg, Large-scale untargeted LC-MS metabolomics data correction using between-batch feature alignment and cluster-based within-batch signal intensity drift correction, Metabolomics, 12 (2016) 173. https://doi.org/10.1007/s11306-016-1124-4.

[12] R. Smith, D. Ventura, J.T. Prince, LC-MS alignment in theory and practice: a comprehensive algorithmic review, Brief. Bioinformatics, 16 (2015) 104-117. https://doi.org/10.1093/bib/bbt080.

[13] G. Delaporte, M. Cladiere, V. Camel, Missing value imputation and data cleaning in untargeted food chemical safety assessment by LC-HRMS, Chemom. Intell. Lab. Syst., 188 (2019) 54-62. https://doi.org/10.1016/j.chemolab.2019.03.005.

[14] M. Pérez-Cova, C. Bedia, D.R. Stoll, R. Tauler, J. Jaumot, MSroi: A pre-processing tool for mass spectrometry-based studies, Chemom. Intell. Lab. Syst., 215 (2021) 104333. https://doi.org/10.1016/j.chemolab.2021.104333.

[15] S. Castillo, P. Gopalacharyulu, L. Yetukuri, M. Oresic, Algorithms and tools for the preprocessing of LC-MS metabolomics data, Chemom. Intell. Lab. Syst., 108 (2011) 23-32. https://doi.org/10.1016/j.chemolab.2011.03.010.

[16] W. Windig, The use of the Durbin-Watson criterion for noise and background reduction of complex liquid chromatography/mass spectrometry data and a new algorithm to determine sample differences, Chemom. Intell. Lab. Syst., 77 (2005) 206-214. https://doi.org/10.1016/j.chemolab.2004.10.008.

[17] T.J. Ho, C.H. Kuo, S.Y. Wang, G.Y. Chen, Y.F.J. Tseng, True ion pick (TIPick): a denoising and peak picking algorithm to extract ion signals from liquid chromatography/mass spectrometry data, J. Mass Spectrom., 48 (2013) 234-242. https://doi.org/10.1002/jms.3154.

[18] Z. Li, Y. Lu, Y. Guo, H. Cao, Q. Wang, W. Shui, Comprehensive evaluation of untargeted metabolomics data processing software in feature detection, quantification and discriminating marker selection, Anal. Chim. Acta, 1029 (2018) 50-57. https://doi.org/10.1016/j.aca.2018.05.001.

[19] E. Rampler, Y.E. Abiead, H. Schoeny, M. Rusz, F. Hildebrand, V. Fitz, G. Koellensperger, Recurrent Topics in Mass Spectrometry-Based Metabolomics and Lipidomics-Standardization, Coverage, and Throughput, Anal. Chem., 93 (2021) 519-545. https://doi.org/10.1021/acs.analchem.0c04698.

[20] M. Perez-Cova, R. Tauler, J. Jaumot, Chemometrics in comprehensive two-dimensional liquid chromatography: A study of the data structure and its multilinear behavior, Chemom. Intell. Lab. Syst., 201 (2020) 104009. https://doi.org/10.1016/j.chemolab.2020.104009.

[21] E. Dubin, M. Spiteri, A.S. Dumas, J. Ginet, M. Lees, D.N. Rutledge, Common components and specific weights analysis: A tool for metabolomic data pre-processing, Chemom. Intell. Lab. Syst., 150 (2016) 41-50. https://doi.org/10.1016/j.chemolab.2015.11.005.

[22] G. Theodoridis, H. Gika, P. Franceschi, L. Caputi, P. Arapitsas, M. Scholz, D. Masuero, R. Wehrens, U. Vrhovsek, F. Mattivi, LC-MS based global metabolite profiling of grapes: solvent extraction protocol optimisation, Metabolomics, 8 (2012) 175-185. https://doi.org/10.1007/s11306-011-0298-z.

[23] T. Pluskal, S. Castillo, A. Villar-Briones, M. Oresic, MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data, BMC Bioinform., 11 (2010) 395. https://doi.org/10.1186/1471-2105-11-395.

[24] R. Stolt, R.J.O. Torgrip, J. Lindberg, L. Csenki, J. Kolmert, I. Schuppe-Koistinen, S.P. Jacobsson, Second-Order Peak Detection for Multicomponent High-Resolution LC/MS Data, Anal. Chem., 78 (2006) 975-983. https://doi.org/10.1021/ac050980b.

[25] T.J. Li, L. Chen, X.L. Lu, An Alternating Direction Minimization based denoising method for extracted ion chromatogram, Chemometr Intell Lab, 206 (2020). ARTN 104138 10.1016/j.chemolab.2020.104138.

[26] S. Li, J.W. Blackman, L.M. Schmidtke, Exploring the regional typicality of Australian Shiraz wines using untargeted metabolomics, Aust J Grape Wine R, 27 (2021) 378-391. 10.1111/ajgw.12493.

[27] G.L. Erny, T. Acunha, C. Simo, A. Cifuentes, A. Alves, Finnee - A Matlab toolbox for separation techniques hyphenated high resolution mass spectrometry dataset, Chemometr Intell Lab, 155 (2016) 138-144. 10.1016/j.chemolab.2016.04.013.

Figure captions

Fig. 1. MATLAB code flowchart encompassing the pre-processing and benchmarking stages.

Fig. 2. Profile of a noisy feature originated by the mobile phase (feature with m/z value 105 from a blank).

Fig. 3. RollingBall baseline corrector.

Fig. 4. Natural variation of the m/z value of the internal standard due to the equipment's scan accuracy.

Fig. 5. The influence of Savitzky-Golay width filter on peak detection. a: width = 5 points; b: width = 10 points.

Fig. 6. Isotopic pattern of Internal Standard (BIBP-3226).

Table 1. MATLAB programming code (example).

a)	clear R samples blanks S
b)	%Define Samples
	samples{1}=[5,6,7,8,10,11,12,13];
	samples{2}=[38,39,40,41,43,44,45,46];
c)	%add more if needed
	%Define blanks
	blanks{1}=[3,4];
d)	blanks{2}=[36,37];
	%add more if needed
	%Compile samples and blanks
	i=0;
	for k=1:length(samples)
	for j=1:length(samples{k})
	i=i+1;
	S{i,1}=samples{k}(j);
	S{i,2}=blanks{k};
e)	end
	end
	%Analyse all samples and blanks m/z peakparamB flagblank
	for k=1:size(S,1)
f)	R{k}=vid_peaks(Result1,S{k,1},S{k,2},1,[100,1000],[10,3,3],[3,3,3],[0,5,3],1)
	end RT tol
	%Compare samples flag peakparamS tol
	Result=vid New(R,0.12);

Table 2. List of features generated by the algorithm. Features present in at least five samples were marked in dark green and in at most in four samples in light green.

	JUNE			JULY			AUGUST			SEPTEMBER		
m/z value	**Frequency**	**RT (min)**	**Intensity**	**Frequency**	**RT (min)**	**Intensity**	**Frequency**	**RT (min)**	**Intensity**	**Frequency**	**RT (min)**	**Intensity**
479	6	3.49	0.02	1	3.52	0.02						
351	8	3.49	0.09	8	3.49	0.04						
376	8	3.50	0.18	8	3.50	0.06						
704	8	3.51	0.06	6	3.50	0.03						
363	8	3.51	0.02	6	3.51	0.02						
373	8	3.51	0.09	8	3.51	0.07						
488	8	3.51	0.08	8	3.51	0.03						
222	8	3.51	0.13	8	3.51	0.07						
258	8	3.51	0.13	8	3.51	0.11	2	3.51	0.04			
433	8	3.51	0.05	8	3.51	0.05	1	3.50	0.02			
163	8	3.50	0.08	8	3.51	0.05	4	3.50	0.02	4	3.48	0.02
474	8	3.53	0.03	8	3.51	0.03						
238	8	3.53	0.04	8	3.51	0.04						
205	8	3.58	0.67	8	3.59	0.83	8	3.58	1.51	8	3.58	0.98
246	8	3.58	0.38	8	3.58	0.54	8	3.58	0.66	8	3.58	0.46
188	8	3.58	0.09	8	3.58	0.14	8	3.58	0.21	8	3.58	0.14
209	8	3.58	0.06	8	3.61	0.05						
209	2	3.80	0.03	8	3.81	0.04						
579	8	3.86	0.19	8	3.87	0.18	8	3.87	0.08	6	3.87	0.04
579	8	4.12	0.14	8	4.11	0.13	2	4.12	0.04			
250	8	4.29	0.09	8	4.27	0.04						
455	7	4.28	0.05	5	4.28	0.04						
417	8	4.29	0.10	8	4.31	0.07						
332	8	4.32	0.06	8	4.32	0.11						
291	8	4.32	0.57	8	4.32	0.98	3	4.32	0.08	2	4.33	0.06
247	7	4.33	0.04	8	4.33	0.07						
867	8	4.34	0.06	8	4.38	0.05						
439	8	4.45	0.18	8	4.45	0.17	4	4.45	0.03			
441	1	4.63	0.03	5	4.63	0.05	6	4.62	0.04	6	4.62	0.04
579	2	4.68	0.04	8	4.66	0.12	8	4.66	0.11	8	4.66	0.09
351	1	5.05	0.03	8	5.04	0.20						
247	8	5.05	0.11	8	5.04	0.21	2	5.05	0.09	1	5.05	0.06
292	3	5.04	0.05	8	5.04	0.19	2	5.07	0.03			
239	1	5.02	0.02	8	5.08	0.04						
332	2	5.06	0.03	8	5.04	0.17						
268				8	5.06	0.05						
579	4	5.40	0.05	4	5.40	0.04						
731	8	5.41	0.22	8	5.41	0.23	2	5.39	0.06			
443	6	6.28	0.08	8	6.29	0.26						
323	6	6.29	0.09	8	6.28	0.19						
344				6	6.29	0.04						
479	8	6.30	0.18	7	6.29	0.14	7	6.30	0.11	6	6.30	0.09
867				4	5.17	0.03	1	5.18	0.02			
465							8	3.84	0.24	8	3.85	0.21
449							5	4.24	0.05	4	4.24	0.04
479							8	4.37	0.46	8	4.36	0.32
463							8	4.75	0.41	8	4.76	0.50
493							8	4.85	1.80	8	4.88	1.88
507							6	5.14	0.09	6	5.16	0.06
535							5	5.52	0.04	3	5.53	0.04
481							8	5.59	0.07	8	5.60	0.09
319							6	5.59	0.04	5	5.60	0.05
322							6	5.59	0.04	4	5.62	0.04
521							8	5.63	0.25	8	5.63	0.23
493							6	5.63	0.05	5	5.63	0.05
349							8	5.65	0.09	8	5.65	0.10
511							8	5.65	0.08	8	5.65	0.09
337							8	5.66	0.06	8	5.65	0.05
505							8	6.06	0.21	8	6.07	0.24
535							3	6.08	0.91	4	6.08	0.99
535							5	6.11	1.47	5	6.12	1.57
611							8	6.27	0.19	8	6.28	0.16
465				4	6.34	0.04	7	6.35	0.11	6	6.34	0.12
303				2	6.33	0.03	5	6.34	0.07	3	6.35	0.06
314							6	6.34	0.07	4	6.35	0.06
655							5	6.77	0.06	5	6.76	0.05
595							5	6.91	0.04	3	6.92	0.03
625							8	7.09	0.19	8	7.08	0.18
639							1	7.68	0.06			
509							4	7.70	0.04	6	7.69	0.06
639							4	7.71	0.11	7	7.71	0.11
639							1	8.12	1.75			
609							8	8.13	0.36	8	8.13	0.58
639							8	8.18	1.88	8	8.17	1.94
288				1	13.50	0.03	5	13.49	0.04	6	13.50	0.05

Table 3. Results after MZmine processing ordered by the retention time in each month.

	JUNE			JULY			AUGUST			SEPTEMBER		
m/z value	**Frequency**	**RT (min)**	**Intensity**	**Frequency**	**RT (min)**	**Intensity**	**Frequency**	**RT (min)**	**Intensity**	**Frequency**	**RT (min)**	**Intensity**
373	8	3.49	1.0E+06									
376	7	3.47	1.6E+06									
258	8	3.51	1.2E+06	4	3.53	5.4E+05						
206	4	3.56	1.0E+06	1	3.55	6.1E+05	6	3.56	1.4E+06	5	3.57	7.3E+05
188	6	3.56	1.0E+06	3	3.56	8.2E+05	5	3.57	1.6E+06	5	3.56	8.0E+05
205	8	3.56	7.4E+06	8	3.56	5.1E+06	8	3.56	1.0E+07	8	3.56	7.1E+06
246	8	3.56	4.4E+06	8	3.56	3.3E+06	7	3.57	5.3E+06	8	3.57	3.4E+06
332	1	4.30	5.9E+05	3	4.30	5.9E+05						
291	8	4.30	5.9E+06	8	4.30	4.9E+06						
292	2	4.30	1.1E+06	5	4.30	8.2E+05						
292	2	4.30	1.1E+06	5	5.02	1.2E+06						
291	8	4.30	5.9E+06	8	5.02	7.4E+06						
247	7	5.01	1.2E+06	8	5.01	1.3E+06						
351				6	5.01	1.2E+06						
332				6	5.02	1.0E+06						
731	8	5.39	2.1E+06	2	5.40	5.3E+05						
479	7	6.28	1.6E+06	1	6.27	5.1E+05						
323	4	6.27	8.3E+05	3	6.27	7.3E+05						
443	3	6.27	7.0E+05	3	6.28	9.6E+05						
143	4	13.44	1.2E+06	3	13.48	1.2E+06						
247							2	3.57	8.0E+05	4	3.57	5.6E+05
479							4	4.37	1.4E+06	3	4.35	8.2E+05
464							1	4.73	4.4E+05	3	4.80	5.1E+05
463							6	4.77	1.8E+06	8	4.77	2.1E+06
495							5	4.82	6.2E+05	7	4.82	6.0E+05
493							8	4.83	8.2E+06	8	4.83	8.3E+06
494							8	4.83	2.9E+06	8	4.83	2.3E+06
521							4	5.60	9.7E+05	2	5.64	9.5E+05
349										1	5.65	4.1E+05
505							3	6.06	5.7E+05	4	6.06	7.9E+05
536							8	6.07	1.5E+06	8	6.07	1.5E+06
535							8	6.07	5.3E+06	8	6.07	5.6E+06
609							8	8.12	2.1E+06	8	8.12	2.9E+06
610							6	8.13	6.6E+05	6	8.13	1.0E+06
639							8	8.15	9.3E+06	8	8.15	9.8E+06
640							8	8.15	4.1E+06	8	8.15	4.7E+06
641							8	8.16	1.0E+06	8	8.16	1.0E+06

Table 4. Differences in operating mode and user-defined parameters between MZmine and the MATLAB algorithm.

MZMINE	
PROCESSING STEPS (6)	**USER-DEFINED PARAMETERS (21)**
Baseline correction:	- Choose type of baseline correction - m/z bin size - Width of local window for minimization/maximization - Width of local window for smoothing
Mass detection:	- Choose type of mass detection - Noise level
Chromatogram builder:	- Minimum time span - Minimum peak height - m/z tolerance
Chromatogram deconvolution:	- Choose type of algorithm - Minimum peak height - Peak duration range - Noise amplitude
Normalization:	- m/z tolerance - Retention time tolerance - Minimum standard intensity
Alignment (Join aligner):	- Choose type of alignment - m/z tolerance - Weight for m/z - Retention time tolerance - Weight for retention time

ALGORITHM	
PROCESSING STEPS (4)	**USER-DEFINED PARAMETERS (9)**
Peak detection	- width - tolfac - w
Noise removal subtracting blanks	- Choose blanks average or intersection - m/z tolerance (corresponding to the variation among scans) - Retention time tolerance (corresponding to the variation among samples) - Intensity tolerance (advisable to use a minimum of 3 as it is the signal-to-noise ratio of 3:1)
Benchmarking	- Retention time tolerance
Noise removal considering peaks and points	- Typical number of points and peaks that define different signals

Figura 1

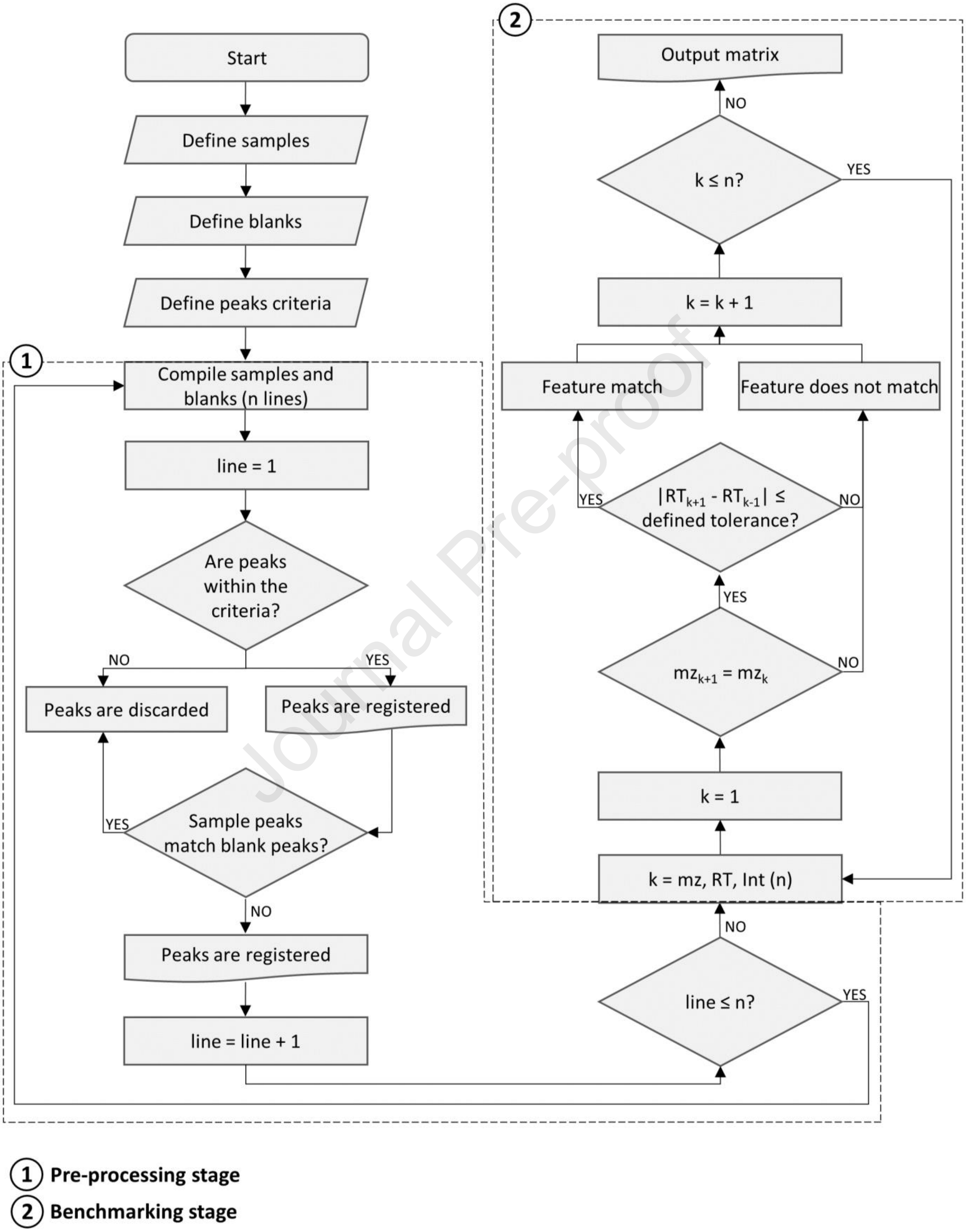

Start
Define samples
Define blanks
Define peaks criteria
1
Compile samples and blanks (n lines)
line = 1
Are peaks within the criteria?
NO
YES
Peaks are discarded
Peaks are registered
YES
Sample peaks match blank peaks?
NO
Peaks are registered
line = line + 1
line ≤ n?
YES
NO
2
k = mz, RT, Int (n)
k = 1
mz$_{k+1}$ = mz$_k$
NO
YES
|RT$_{k+1}$ - RT$_{k-1}$| ≤ defined tolerance?
YES
NO
Feature match
Feature does not match
k = k + 1
k ≤ n?
YES
NO
Output matrix
1 Pre-processing stage
2 Benchmarking stage

Figura 2

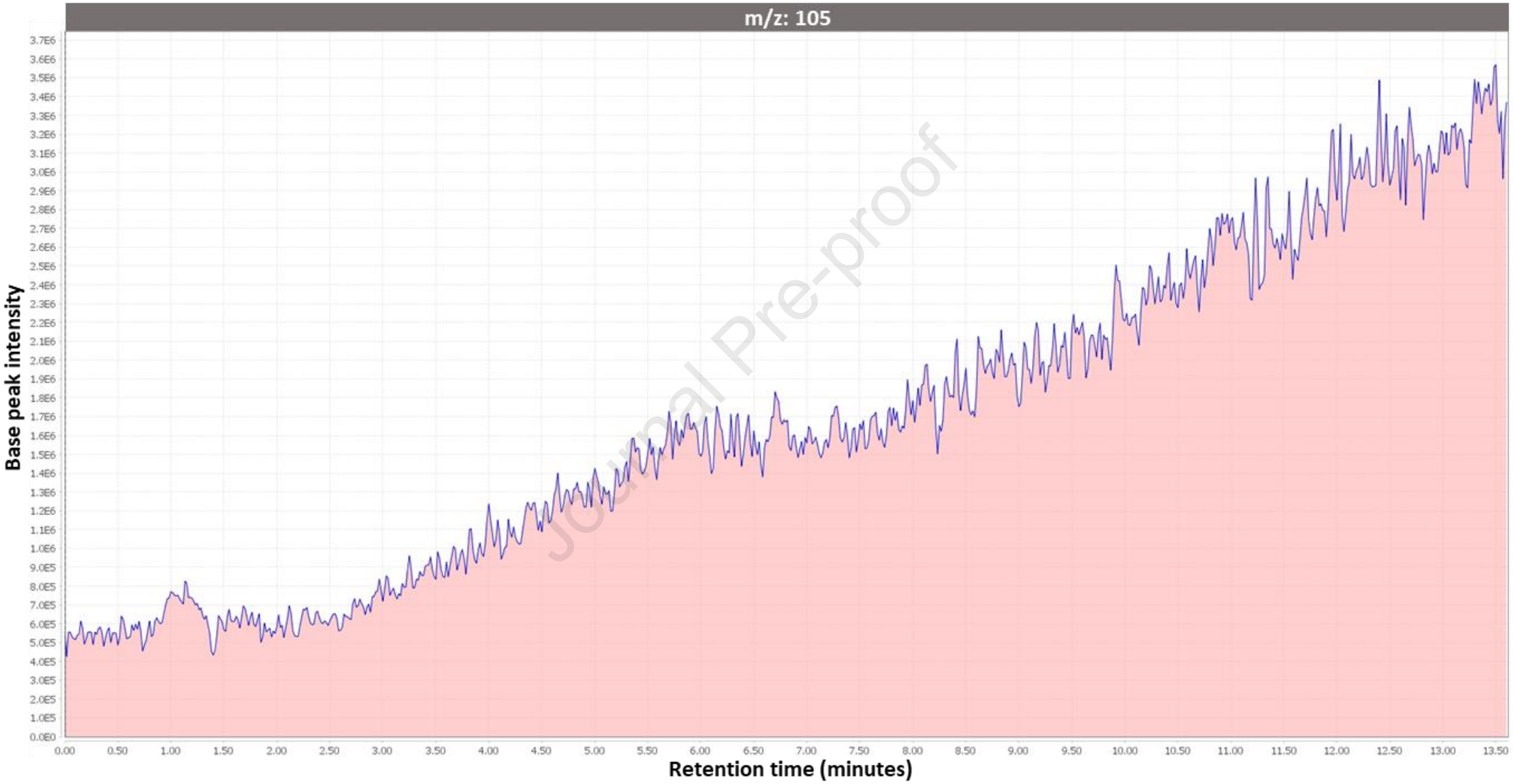

m/z: 105
Base peak intensity
3.7E6
3.6E6
3.5E6
3.4E6
3.3E6
3.2E6
3.1E6
3.0E6
2.9E6
2.8E6
2.7E6
2.6E6
2.5E6
2.4E6
2.3E6
2.2E6
2.1E6
2.0E6
1.9E6
1.8E6
1.7E6
1.6E6
1.5E6
1.4E6
1.3E6
1.2E6
1.1E6
1.0E6
9.0E5
8.0E5
7.0E5
6.0E5
5.0E5
4.0E5
3.0E5
2.0E5
1.0E5
0.0E0
0.00
0.50
1.00
1.50
2.00
2.50
3.00
3.50
4.00
4.50
5.00
5.50
6.00
6.50
7.00
7.50
8.00
8.50
9.00
9.50
10.00
10.50
11.00
11.50
12.00
12.50
13.00
13.50
Retention time (minutes)

Figura 3

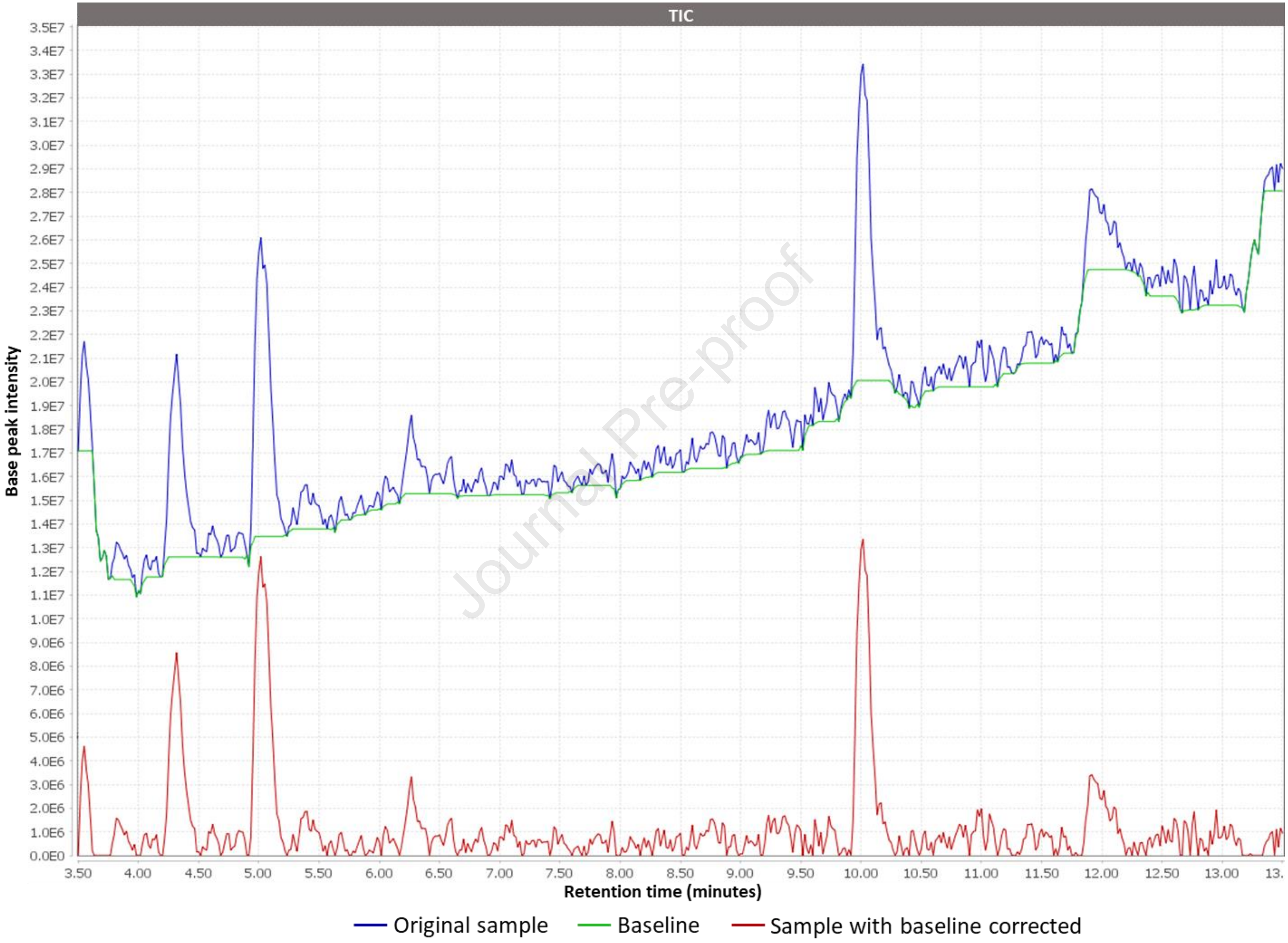

TIC
Base peak intensity
3.5E7
3.4E7
3.3E7
3.2E7
3.1E7
3.0E7
2.9E7
2.8E7
2.7E7
2.6E7
2.5E7
2.4E7
2.3E7
2.2E7
2.1E7
2.0E7
1.9E7
1.8E7
1.7E7
1.6E7
1.5E7
1.4E7
1.3E7
1.2E7
1.1E7
1.0E7
9.0E6
8.0E6
7.0E6
6.0E6
5.0E6
4.0E6
3.0E6
2.0E6
1.0E6
0.0E0
3.50
4.00
4.50
5.00
5.50
6.00
6.50
7.00
7.50
8.00
8.50
9.00
9.50
10.00
10.50
11.00
11.50
12.00
12.50
13.00
13.
Retention time (minutes)
Original sample
Baseline
Sample with baseline corrected

Figura 4

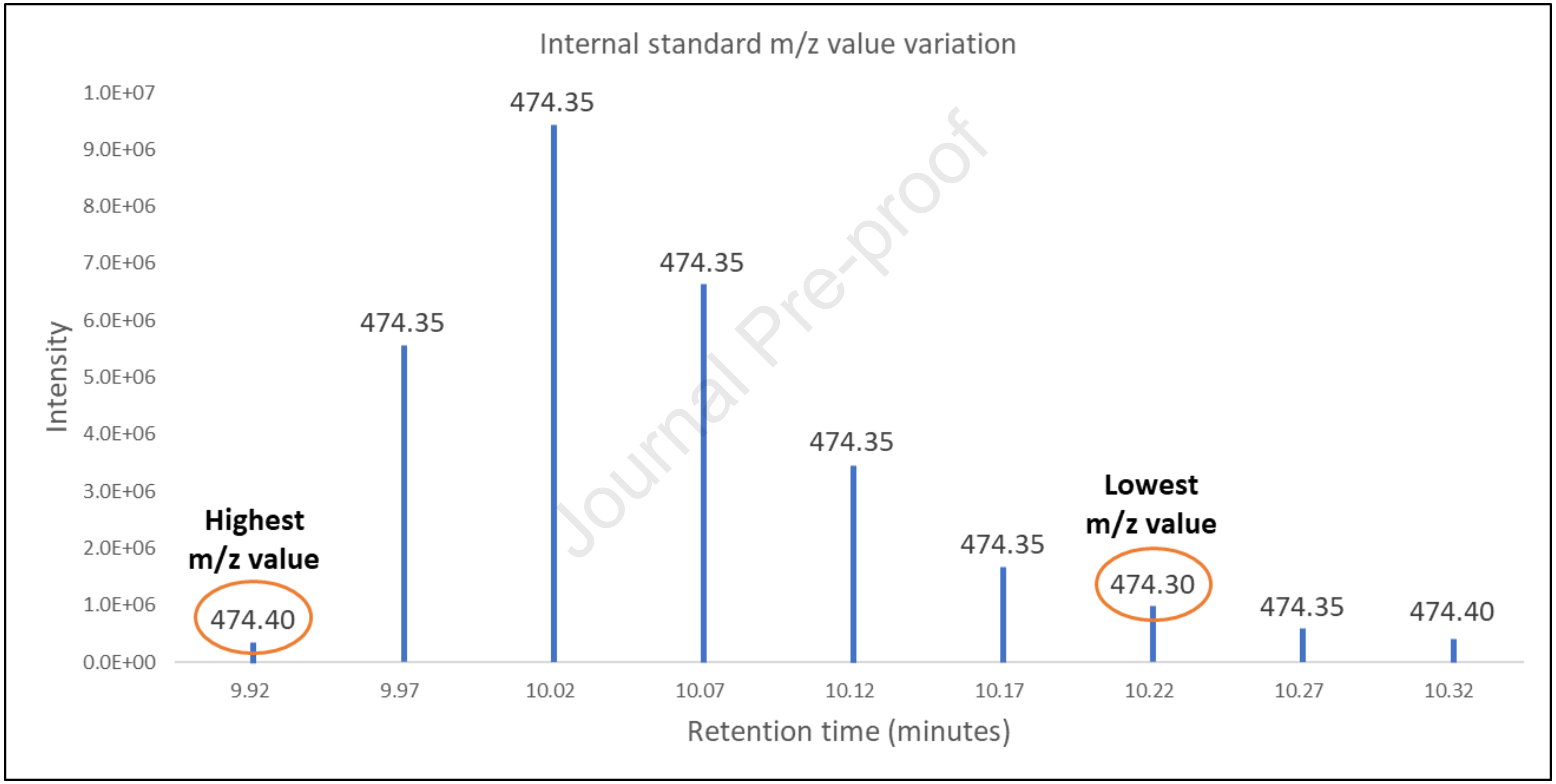

Internal standard m/z value variation
Intensity
1.0E+07
9.0E+06
8.0E+06
7.0E+06
6.0E+06
5.0E+06
4.0E+06
3.0E+06
2.0E+06
1.0E+06
0.0E+00
Highest
m/z value
474.40
474.35
474.35
474.35
474.35
474.35
Lowest
m/z value
474.30
474.35
474.40
9.92
9.97
10.02
10.07
10.12
10.17
10.22
10.27
10.32
Retention time (minutes)

Figura 5

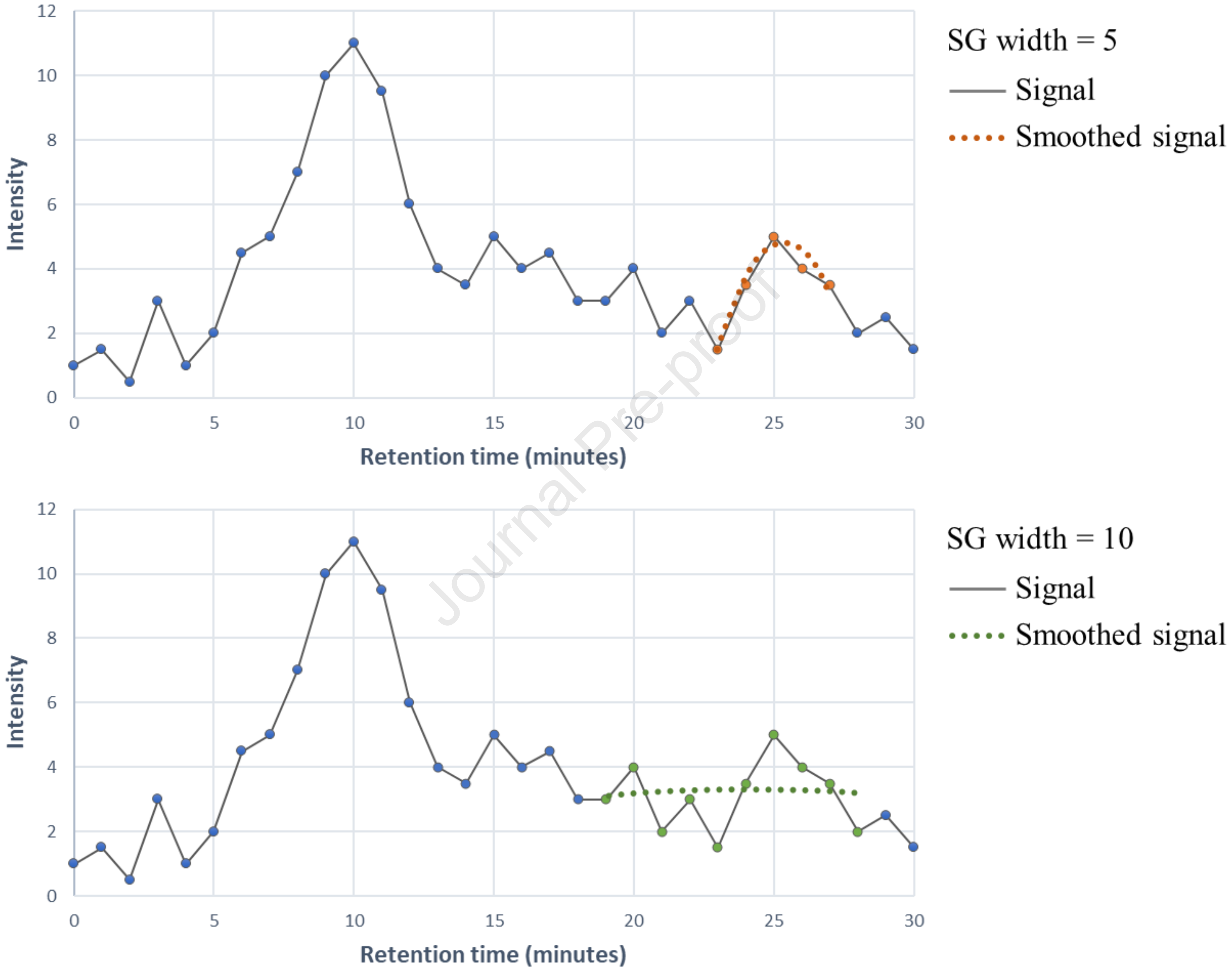
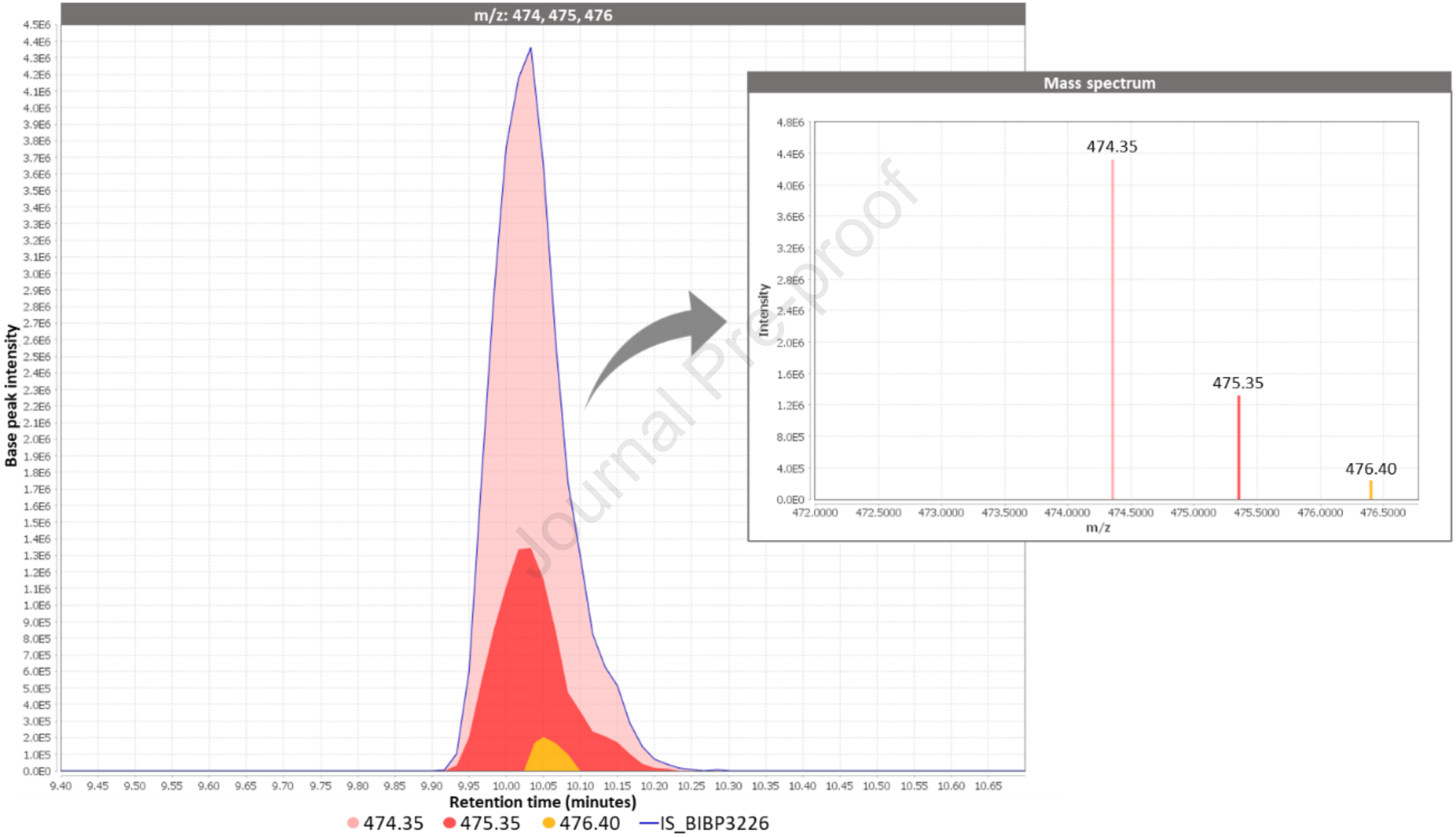

12
10
8
6
4
2
0
Intensity
0
5
10
15
20
25
30
Retention time (minutes)
SG width = 5
Signal
Smoothed signal
12
10
8
6
4
2
0
Intensity
0
5
10
15
20
25
30
Retention time (minutes)
SG width = 10
Signal
Smoothed signal

Figura 6


m/z: 474, 475, 476
Base peak intensity
4.5E6
4.4E6
4.3E6
4.2E6
4.1E6
4.0E6
3.9E6
3.8E6
3.7E6
3.6E6
3.5E6
3.4E6
3.3E6
3.2E6
3.1E6
3.0E6
2.9E6
2.8E6
2.7E6
2.6E6
2.5E6
2.4E6
2.3E6
2.2E6
2.1E6
2.0E6
1.9E6
1.8E6
1.7E6
1.6E6
1.5E6
1.4E6
1.3E6
1.2E6
1.1E6
1.0E6
9.0E5
8.0E5
7.0E5
6.0E5
5.0E5
4.0E5
3.0E5
2.0E5
1.0E5
0.0E0
9.40
9.45
9.50
9.55
9.60
9.65
9.70
9.75
9.80
9.85
9.90
9.95
10.00
10.05
10.10
10.15
10.20
10.25
10.30
10.35
10.40
10.45
10.50
10.55
10.60
10.65
Retention time (minutes)
● 474.35
● 475.35
● 476.40
—IS_BIBP3226
Mass spectrum
Intensity
4.8E6
4.4E6
4.0E6
3.6E6
3.2E6
2.8E6
2.4E6
2.0E6
1.6E6
1.2E6
8.0E5
4.0E5
0.0E0
474.35
475.35
476.40
472.0000
472.5000
473.0000
473.5000
474.0000
474.5000
475.0000
475.5000
476.0000
476.5000
m/z

Highlights

- An algorithm for complex untargeted LC-MS data analysis was developed.
- The algorithm highlights features that are worth to be further investigated.
- Application to grape metabolomic profile originated 99 features.
- The algorithm provided almost 3 times more features than MZmine using fewer inputs.

CRediT author statement

Sandia Machado: Software, Formal analysis, Investigation, Data Curation, Writing - Original Draft Preparation, Visualization; Luisa Barreiros: Investigation, Data Curation, Writing - Review & Editing; António R. Graça: Resources, Writing - Review & Editing; Ricardo N. M. J. Páscoa: Validation, Investigation, Writing - Review & Editing; Marcela A. Segundo: Conceptualization, Validation, Writing - Review & Editing, Visualization, Supervision, Project administration, Funding acquisition; João A. Lopes: Conceptualization, Methodology, Software, Writing - Review & Editing, Supervision, Project administration, Funding acquisition

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: