Microchemical Journal 156 (2020) 104830

Contents lists available at ScienceDirect

Microchemical Journal

journal homepage: www.elsevier.com/locate/microc

RGCxGC toolbox: An R-package for data processing in comprehensive two-dimensional gas chromatography-mass spectrometry

Cristian Quiroz-Moreno[a], Mayra Fontes Furlan[b], João Raul Belinato[b], Fabio Augusto[b], Guilherme L. Alexandrino[c], Noroska Gabriela Salazar Mogollón[a,*]

[a] Biomolecules discovery group, Universidad Regional Amazónica Ikiam, Km 7 Via Muyuna, Tena, Napo, Ecuador
[b] Institute of Chemistry, University of Campinas (Unicamp) and National Institute of Bioanalytical Science and Technology (INCTBio), CP 6154, 13083-970 Campinas, – São Paulo, Brazil
[c] Department of Plant and Environmental Sciences, Faculty of Science, University of Copenhagen, Thorvaldsensvej 40, DK-1871, Copenhagen, Denmark

ARTICLE INFO

Keywords:
Metabolomics
Exploratory analysis
Comprehensive two-dimensional gas chromatography
Toolbox
Signal processing

ABSTRACT

Comprehensive two-dimensional gas chromatography (GC × GC) offers detailed chemical information about volatile and semivolatile analytes from complex samples. However, the high complexity of the data structure encourages the development of new tools for a more efficient data handling and analysis. Although some tools have already been presented to overcome this challenge, there is still need for improvement. In this manuscript, we present a toolbox containing a pipeline for end-to-end basic GC × GC data processing which can be used for both, signal pre-processing and multivariate data analysis. The pre-processing algorithms perform signal smoothing, baseline correction, and peak alignment, while the multivariate analysis is done through Multiway Principal Component Analysis (MPCA). The software is capable to prepare the chromatographic data for further applications with other chemometric tools, e.g.: cluster analysis, regression, discriminant analysis, etc. The performance of this new software was tested on in-house experimental dataset and on two other published datasets.

1. Introduction

Gas chromatography (GC), and particularly comprehensive two-dimensional gas chromatography, have come practically a mandatory technique for the analysis of the volatile and semivolatile compounds in matrices of high chemical complexity [1]. In GC × GC, the enhanced separation power is achieved by two capillary columns with preferably orthogonal separation capabilities, connected by the modulator. The modulator periodically concentrates a (coeluting) fraction of the eluate coming from the first column (first dimension, 1D) and next, reinjects this fraction as a narrower band into the second column (second dimension, 2D). Therefore, compounds that would coelute in conventional GC can be potentially separated in GC × GC system [2,3]. Because of the benefits provided by GC × GC, it has been widely applied in forensic [4], environmental [5], fuel [6], and metabolomics [7] analysis.

Apart from the enhanced peak capacity of the GC × GC and the better elucidation of the chemical fingerprint from complex samples, isomers and homologous series are usually identified more easily in the two-dimensional chromatograms, e.g.: the roof-lite effect of hydrocarbons, and therefore the non-ambiguous identifications of unknowns is also improved [8,9]. However, GC × GC can be coupled to multichannel detectors, such as mass spectrometers, which results in a large amount of data, e.g.: up to 2 GB raw files per sample in GC × GC–MS, and therefore efficient data handing tools are necessary [10–12].

Chemometrics use mathematical and statistical methods to analyze multivariate chemical data [13,14]. However, the efficient application of chemometrics in chromatography usually requires a previous pre-processing of the signals to reduce or mitigate undesirable artifacts, such as instrumental noise (e.g.: detectors' signal fluctuation and retention time shifts across multiple runs) and chemical noise (e.g.: column bleeding and peak saturation) [15]. The common pre-processing algorithms used in GC/GC × GC to correct for column bleeding and signal fluctuation are baseline correction and signal smoothing [15]. For GC × GC, the algorithms for correction of retention time shifts across samples should handle shifts in both ^{1}D and ^{2}D, such as the two-dimension correlation optimized warping (2D-COW) [16], parametric time warping (DTW) [17], distance and spectrum correlation optimization (DISCO and DISCO2) [18,19], graph-based multiple alignment

* Corresponding author.
E-mail address: noroska.salazar@ikiam.edu.ec (N.G.S. Mogollón).

https://doi.org/10.1016/j.microc.2020.104830
Received 17 February 2020; Received in revised form 11 March 2020; Accepted 11 March 2020
Available online 14 March 2020
0026-265X/

Abbreviations

GC	Gas chromatography
GC × GC	Comprehensive two-dimensional gas chromatography
TOF	Time of flying
NetCDF	Network common data form
2DCOW	Two-dimensional correlation optimized warping
TIC	Total ion current
PLS	Partial least squares
PLS-DA	Partial least squares – discriminant analysis
PCA	Principal component analysis
MPCA	Multiway principal component analysis
PC	Principal component
S/N	Singal to noise ratio
MYL	*Myrothecium sp.*
CUI	*Curvularia sp.*
PDA	Potato-dextrose-agar
SPME	Solid-phase microextraction
GUI	Graphical user interface
CRAN	Comprehensive R archive network

(BIPACE 2D) [20].

Data handling in GC × GC, especially the pre-processing step, is usually performed by commercial software, and a minor percentage is done using open source tools [21]. For instance, over 85% of data analysis has been performed with the vendor or commercial software in metabolomics research [22]. Although the open-source toolbox Guineu [23] and R2DGC [23] have been developed for data handling and pre-processing of GC × GC using single-and/or multichannel detection, there is a still lack for extended workflow pipeline that also performs data analysis using chemometrics. In this stream, the RMet toolbox has been proposed as an end-to-end pipeline to process GC/GC × GC–MS data [24]. One of the main advantages of RMet is that presents a Graphical User Interface, which makes the analysis easier for inexperienced users. In contrast with command line toolboxes, GUI based ones do not allow an easy way to run multiple sets of parameter combinations until reach convergence, as usually needed in order to tune parameters. Moreover, in the RMet toolbox, discriminant analysis is performed only by building a partial least squares-discriminant analysis model. Supervised models, such as PLS-DA, has to be built carefully, since one of the main disadvantages is model overfitting [25].

Data analysis in GC/GC × GC has been commonly performed by pixel-based level, or by peak picking approaches [11,26,27]. The pixel-based approach performs the data analysis directly in the two-dimensional chromatogram which can be considered as a pixel image, obtained at the retention times in ^{1}D and ^{2}D, respectively [28]. Conversely, the peak picking approach picks, integrates, and then organizes the individual peaks of the chromatograms in a peak table where the variances of the areas can be analyzed between multiple samples. There are advantages and drawbacks for both, pixel-based and peak picking approaches, and the decision about which approach to choose depends on the goals of the study. For instance, data analysis based on peak picking handles significantly fewer variables than in the pixel-based approach, which is an advantage for data analysis using the conventional univariate statistic. However, the quality of the data depends on the efficient selection and integration of the peaks, which can be problematic for highly coeluting peaks. Furthermore, the two-dimensional structure of the chromatograms is not seen straightforwardly as with a pixel-based approach. Therefore, interpreting the chromatographic differences between samples can be less intuitive than the pixel-based approach.

In this work, we present the toolbox RGCxGC which was developed for data processing in GC × GC–MS, based on the open-source R environment. This toolbox contains an end-to-end pipeline for the most common processing techniques that are required for GC × GC, such as baseline correction, signal smoothing, and two-dimensional peak alignment. Moreover, the pixel-based analysis can be performed using MPCA. On the other hand, the data can be exported in a more compatible format to be used with external toolboxes for chemometrics. The performance of this open-source is demonstrated with three datasets in total. Two datasets are from microbial antagonism interaction and one of the was homemade created, while the third dataset was retrieved from literature and is related to typhoid carriage diagnosis.

2. Materials and methods

2.1. Methods

2.1.1. Biological material and its maintenance

All fungal strains were obtained from "Collection of Bahia Microorganisms " (CCMB) and were kept in Petri dishes containing 20 mL of carrot-maize-agar (CMA) culture media at (25,0 ± 1,0) °C in a growth chamber (Eletrolab, model EL202) with 12 h of photo-period.

2.1.2. Fungal inoculation and headspace extraction

PDA culture media (25 mL) was placed in 50 mL polypropylene centrifuge tubes using angulation (elevation of 1,5 cm). The tube cap was modified with a 15 mm diameter hole and PTFE septa held by the aluminum ring.

Inoculation was made from Petri dishes with the fully grown fungal cells, and sterile distilled water was used to wash the surface of the plate. The plate was scraped with a sterile glass handle to obtain the spore suspension. The suspension was liquated and the concentration of 2.4 × 105 spores / mL was determined using a Neubauer chamber and an optical microscope. Then, 50 µL of the suspension was inoculated in a flow chamber into the tubes containing the culture medium. Tubes were kept at (25,0 ± 1,0) °C in a growth chamber (Eletrolab, model EL202) with 12 h of photoperiod.

A solid-phase microextraction (SPME) assay containing a DVB / CAR / PDMS (Divinylbenzene / Carboxene / Polymethylsiloxane 50/30 mm, Supelco) fiber was placed into the tube headspace for 35 min at (25,0 ± 1,0) °C.

2.1.3. GCxGC-QMS

A set of columns consisting of HP-5MS 30 *m* × 0,25 mm × 0,25 µm (Supelco) connected to a Supelcowax 1 *m* × 0.10 mm × 0.10 µm (Supelco) with a 1 *m* × 0.25 mm deactivated silica capillary being used as the loop. The modulation period was set to 5.0 s. For GC × GC-QMS experiments was used a temperature program were 60 °C - 165 °C @3 °C/min; 165 °C - 260 °C @20 °C/min; 260 °C (5 min); flow rate 0,6 mL/min (Helium 5.0 carrier gas); splitless injection mode, ion source temperature 200 °C, interface temperature 260 °C; voltage 0,9 kV; mass range 50–380 *m/z*; acquisition rate 25 Hz and electron ionization (70 eV). For GC × GC-QMS data acquisition, GCMSsolution version 5.3 software (Shimadzu, Tokyo, Japan) and GCImage version 2.0 software (Zoex - Houston, TX, USA) was used for the analysis of two-dimensional chromatograms.

2.1.4. Tentative identification

For tentative identification, a two-dimensional chromatographic run was performed before the experiment with standards of homologs n-alkanes (C_8–C_{20}, Sigma Aldrich). 2 µL of the standard solution was transferred to a vial and the SPME fiber was exposed during 15 min after chromatographic desorption. The NIST 2008 spectra library (NIST – Gaithersburg, MD, EUA) was used (considering 80% of similarity) and comparison was made with the van den Don and Kratz retention index [29].

2.2. Software implementation

The basic workflow of the RGCxGC package is composed of three main steps; data importing, pre-processing and multivariate analysis. First, the raw Network Common Data Form (NetCDF) chromatogram is imported with the "read_chrom" function, in which the user needs to set the modulation time in which the GC × GC data was acquired. Next, you can perform smoothing and baseline correction using the function "wsmooth" and "baseline_corr", respectively. Then, peak alignment from a single sample can be done using the "twod_cow" function, based on the two-dimensional correlation optimized warping (2DCOW) algorithm. Alternatively, multiple sample alignments can be performed with the "batch_2DCOW" routine, where the first chromatogram will be considered as the reference while aligning the remaining chromatograms. After pre-processing, MPCA can be performed on the dataset using the "m_prcomp" function, which provides the scores and loadings matrices and the summary with the explained and cumulative variance per Principal Component (PC). In the case of the loading matrix, they can be plotted using "plot_loading" and retrieved with the "scores" functions, while the "print" function you can access to the MPCA summary. The toolbox pipeline is summarized in Fig. 1. Library methods, and their arguments and comments, are summarized in Table 1.

2.1.1. Importing data

This initial step is an adaptation of the Skov routine [30]. The procedure about how to extract and handle chromatographic signals was based also on Skov's routine. In the importing function, an option to import specific retention times ranges in both dimensions were included (*see x_cut and y_cut*) see Table 1. First, the chromatogram has to be exported from vendor software into a NetCDF file. This file extension is commonly used in scientific approaches. The exported chromatogram contains the retention time and the Total Ion Current (TIC). Due to NetCDF architecture, data is stored into one-dimensional arrays, therefore, the signal is accessed through the arrays named scan_acquisition_time and total_intensity. Since the acquired data by both mass analyzers, time of flying and quadrupole, converge in a NetCDF file, they can be imported and analyzed with the proposed toolbox. Thus, the retention time is divided by sixty to transform the signal from seconds to minutes. Before the one-dimensional array is folded into a more familiar two-dimensional chromatogram, the sampling rate is evaluated to be homogenous. In other words, the software ensures that the entire run has the same sampling rate since non-integer sampling rates leads to unpaired data points over the chromatographic run.

Once the one-dimensional vector is stored in memory, the routine proceeds to fold it into a two-dimensional chromatogram. Each modulation period creates a matrix $C_{(I,\ J)}$, where I is the mass spectra scans, acquired in a given modulation and J is the modulation index. The number of mass spectra scans is related to the sampling rate (Hz) of the mass analyzer. The sampling rate exhibit by TOF mass analyzer is greater than quadrupole mass analyzers. Thus, a chromatogram acquired with a TOF will have more scans per seconds than a chromatogram acquired with a quadrupole. Therefore, in order to calculate the

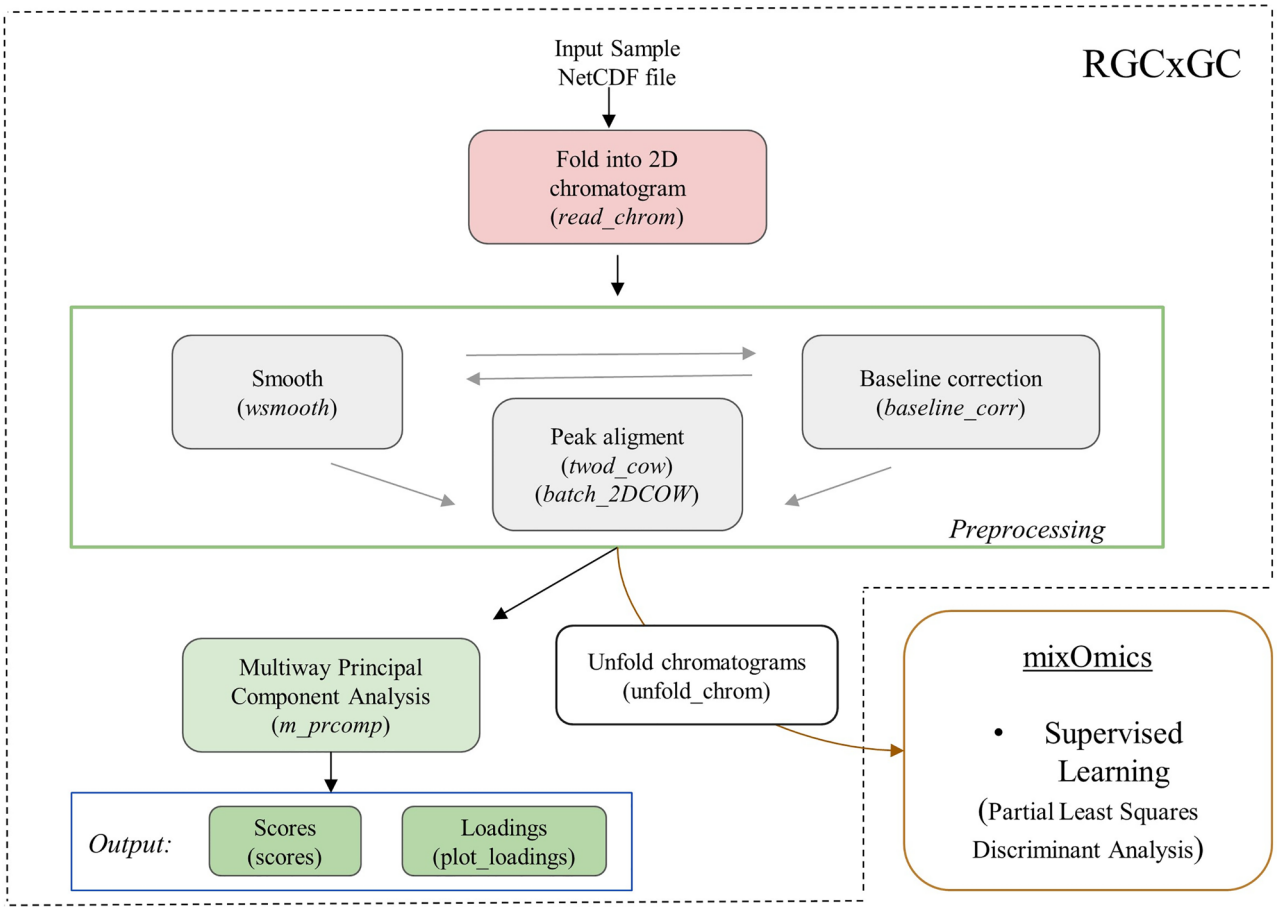

Fig. 1. The proposed pipeline of non-targeted GC × GC-MS data analysis workflow in the RGCxGC toolbox. First, chromatograms are imported with the "read_chrom" function. Next, the user can pre-process them by smoothing, baseline correction, and peak alignment with the "wsmooth", "baseline_corr" and "twod_cow" functions, respectively. Then, chromatograms from multiple cohorts are gathering in a single object, before to be subjected to multiway principal component analysis. On the other hand, the user is able to export all chromatograms with the "unfold_chrom", in order to perform different statistical analysis. While the dashed lines enclose the functionalities implemented in the RGCxGC toolbox, the external parts show functionalities of external R toolboxes.

Table 1
Description of the main functions presented in the RGCxGC toolbox, classified by task aim. Each function has a short description, and the full argument list and default argument value. **The reference is included in the functions that are adapted from other, in cases when the function is built from scratch, the reference is left in blank.*

Aim	Task	Function name	Function description	Arguments	Argument description	Reference*
Importing data and visualization	Read raw chromatogram	read_chrom	This function reads the netCDF file and retrieves the values in the scan_acquisition_time and total_intensity variables. Then, with the provided sampling rate and modulation time, the chromatogram is folded into a numerical matrix (two-dimensional chromatogram)	name	The name of the netCDF file which the data will be retrieved	[30]
				mod_time	The modulation time of the chromatographic run	
				sam_rate	The sampling rate of the equipment. If sam_rate is missing, the sampling rate is calculated by the dividing one by the difference of two adjacent scan time	
				per_eval	An integer with the percentage of the run time to be evaluated, if the sampling rate is homogeneous	
				x_cut	The retention time range in the first dimension to be maintained while importing data.	–
				y_cut	The retention time range in the first second to be maintained while importing data.	
	Plot two-dimensional chromatogram	plot	This function receives a two-dimensional matrix with TIC signals and plots them into a contour or filled contour plot	type	A single character indicating the type of chromatogram representation. By default, type= "f" for a filled contour, if type = "c" only contours or isolines will be displayed.	–
Preprocessing	Smoothing	wsmooth	This function takes a raw two-dimensional chromatogram and performs the weighted Wittaker smoother routine. It smooths with linear or quadratic penalty alongside the first dimension, based on Whittaker smoother	chromatogram	A two-dimensional chromatogram	
				penalty	The penalty order. Only penalty of first (penalty = 1) and second-order (penalty = 2) are allowed. By default, it is performed with the first penalty order.	[37]
				lambda	smoothing parameter: larger values lead to more smoothing	
	Baseline correction	baseline_corr	It corrects the baseline of the chromatogram based on asymmetric least squares	chromatogram	A two-dimensional chromatogram	[39]
	Peak alignment	twod_cow	It aligns a sample chromatograms against a reference chromatogram by two-dimensional correlation optimized warping algorithm	sample_chrom	A two-dimensional chromatogram which will be aligned	[40]
				ref_chrom	A reference two-dimensional chromatogram	
				segments	Two integers with the number of segments in which the first and second dimension will be subdivided, respectively.	
				max_warp	A two integer vector with the maximum warping parameter	
Multivariate Analysis	Multiway Principal Component Analysis	m_prcomp	It performs a multiway principal component analysis on the given two-dimensional chromatograms. Before to perform the calculation, each given chromatograms are unfolded to a one-dimensional vetor	chrom	Multiple chromatograms or batch ones aligned	
				center	A logical value indicating whether the variables should be shifted to be zero centered. True is set by default and is strongly suggested not to change to False.	
				scale	a logical value indicating whether the variables should be scaled to have unit variance before the analysis takes place. The default is True to give the same variable importance in chemometrics.	[44]
				npcs	An integer indicating how many principals components are desired to maintain.The default is 3 principal component	
	Retrieve scores from MPCA	scores	exports the scores matrix of the previously MPCA performed	Object	The result of m_prcomp function	–
	Plot MPCA loadings	plot_loading	This function takes the loadings of MPCA and eval if a certain variable was removed previous to compute the MPCA and fill the removed variables with cero (zero variance variables). Then, it plots the loadings of a single principal component in two dimensions	Object	The result of m_prcomp function	–
				type	The value type of loadings, “p” for positive, “n” for negative, and “b” for negative and positive loading values	
				pc	The number of the principal component to the plot	

total number of modulations, the total number of scans is divided by the number of scans per modulation. Finally, once the number of scans per modulation (I) and the total number of modulations (J) are calculated, the one-dimensional array is folded into a two-dimensional matrix. Usually, a chromatographic run contains incomplete modulation scans at the end of the run, these scans are removed previous to create the two-dimensional chromatogram while printing a message.

In GC, solvent effect and column bleeding are almost unavoidable. Therefore, in order to remove parts of the chromatogram that does not contain significant information, the user can provide the retention time range that they would like to keep. Another example is large chromatographic runs, where a cleaning ramp at high temperatures is included at the end of the run, producing large values at ^{1}D. Here, the user can avoid importing the last part of the chromatogram by changing the "x_cut" argument.

2.1.2. Visualization

As described by Reichenbach [31], the GC × GC image visualization consists of colored pixels layers. In this case, interpolation techniques are used for a two-dimensional image view. The interpolation function evaluates the collection of TIC at the $C_{(I)}$ and $C_{(J)}$ coordinates, and then approximate the contours with the computed interpolation. On the other hand, it is common in chromatography that the signals contain highly similar intensities, or a certain group of chemical entities produces signals several times higher than the rest of the molecules. Therefore, in this library, the user has two options to display two-dimensional chromatograms, filled contour, and contour plot. While contour plot displays low and high-intensity TIC signals as isolines with a white background, filled contour assigns a different color to the background, which may obscure low-intensity signals. In other words, analyte concentrations are not usually evenly distributed in the sample, they may produce a large range of signal intensities. Thus, analytes with greater concentrations my opaque the visualization of the analytes with lower concentrations. To overcome this issue, the user can choose the contour plot to display low signal intensities.

Moreover, the color palette must be taken into consideration in this type of graphical representations, since they play the main role by capturing reader attention [32]. Building an effective color ramp may be difficult for inexperienced users. We encourage users to employ the colorRmps package [33]. On the other hand, users with more programming skills can create a color palette from scratch, as explained in [32].

2.1.3. Pre-processing

High throughput chemical equipment achieves a great level of detail, in which external artifacts are also included, such as instrument variability or sample matrix effect. Therefore, the analysist has to eliminate undesirable information to convert the raw data into useful information, because it has a huge effect on the downstream analysis [11,28,34]. Consequently, several pre-processing techniques have been developed in order to remove chemical and instrument noise. In general, pre-processing techniques include three basic modules: smoothing, baseline correction, and peak alignment.

2.1.3.1. Smoothing. Denoising signals enhance the signal to noise ratio (S/N) in the chromatogram, increasing both accuracy and precision analytical results. Although the pioneers in smoothing were Savitzky and Golay with the local likelihood approach [35,36], Whittaker smoother was introduced as a general-purpose algorithm in chemistry, showing better performance [37]. Whittaker smoothing works in the time domain by discrete penalized least squares taking into consideration data fidelity of original data and roughness of the fitted data. This algorithm starts with a supposition of a noisy signal y and a smoothed signal z that fits y. The roughness (R) of the smoother can be described as the sum square of z $R = \Sigma_i (z_i - z_{i-1})^2$. Moreover, the lack of fitting can be expressed as the sum of squares of differences $S = \Sigma_i (y_i - z_i)^2$. Finally, the governing equation of Whittaker smoother can be stated as:

$$Q = S + \lambda R$$

Where λ is a user given multiplicative factor to the roughness. The aim of Whittaker smoother is to find the combination of z that minimizes Q. While the user gives λ larger values, z will be more smoothed. Different results by varying λ are showed in [38]. These advantages account for automatically boundaries adaptation, missing values and sparse handling, and good computational efficiency in a desktop computer. The Whittaker routing is available through the "wsmooth" function.

2.1.3.2. Baseline correction. The baseline drift in GC/GC × GC is mostly caused by column bleeding or complex mixtures that cannot be separated [11]. In order to remove this type of noise, baseline correction removes the baseline noise and centering the signal around zero. The proposed library implements baseline correction by asymmetric least squares algorithm [39]. Eilers proposed the baseline correction by adapting Whittaker smoothing to calculate the trend of the baseline. In this extension, weights (ω) are introduced, stated as.

$$Q = \sum_i \omega_i (y_i - z_i)^2 + \lambda \sum_i (\Delta^2 z_i)^2$$

While ordinary least squares obtain the residual based on the difference of the raw and fitted signal (y - z), and the sign of the residuals has the same effects over the penalties, the asymmetric least squares give more weight to negative residuals than positive residuals. This approach is considered based on the positive residuals are obtained when a peak is detected. Therefore, the analytical peak signal has not had to be distorted. In contrast, negative residuals are more penalized. As stated above, weights are assigned based on the sign of the residual as follows: $\omega_i = p$ if $y_i > z_i$ and $\omega_i = 1 - p$ otherwise. Here, p is introduced as a user parameter. In cases when the fitted signal is greater than the raw signal, weights are the difference of $\omega_i = 1 - p$.

Although asymmetric least squares offer advantages, such as short computational times, parameter flexibility, the parameters have to optimize by hand. This function is available in the library through the "baseline_corr" command.

2.1.3.3. Peak alignment. The retention shift in chromatography is an unavoidable source of experimental variation [34]. Retention shift can be caused by stationary phase decomposition, column change during usage, or different modulation temperatures over the experiment. As described above, this software follows the stream of pixel-based analysis. Therefore, 2DCOW algorithm was implemented [40]. Basically, the 2DCOW works by splitting the sample (A) and reference (B) chromatogram of $X_{(I, J)}$ dimensions into m segments for both dimensions (^{1}D and ^{2}D), respectively. The new partitioned matrix can be stated as n_i and n_j with column nodes in the first $\{e_0, e_1, \bullet\bullet\bullet, e_{n_i}\}$, and the second $\{\{f_0, f_1, \bullet\bullet\bullet, f_{nj}\}\}$ column, which are a segment subset of {1, 2, •••, I} and {1, 2, •••, J}, respectively. Each grid from the partitioned matrix, for the sample and the reference chromatogram, can be expressed as $\{(e_k, f_l)\colon k = 0, 1, \bullet\bullet\bullet, n_i; l = 0, 1, \bullet\bullet\bullet, n_j\}$. At each row e_k of A, a new row vector $\sim A_{e_k} = (\sim A_{e_k 1}, \sim A_{e_k 2}, \cdots, \sim A_{e_k I})$ is obtained, with the jth component $\sim A_{e_k f}$ through.

$$\sim A_{i_k,j} = \sum_{e=1}^{n} A_{ef} W\ e - e_k/h / \sum_{e=1}^{n} We - e_k/h$$

This routine requires one pair of arguments for each of the first and second dimensions that are called segment length, which is the number of sections to split the chromatogram and slack, which is the maximum warping level. In the proposed library, there are two functions to perform two-dimensional peak alignment, "twod_cow" and "batch_2DCOW". While the first command can align a single chromatogram against a reference, the second routine can align a set of

chromatograms to a reference. Regarding the reference chromatogram, it has to keep high peak similarity between sample chromatograms since the signals should match to perform the alignment, a more detailed explanation about criteria to choose the target chromatogram is provided in [11,15]. One of the main advantages that provides 2DCOW is the interpolative warping of the warped region and the reference in order to maximize the correlation between them, correcting the retention time shifts.

One strategy to select the reference chromatogram is to create an artificial chromatogram by summing or averaging pixels of a set of chromatograms from different groups. In the proposed toolbox, we provide the "ref_chrom" function, which receives multiple chromatograms and computes a new temporal chromatogram to be employed as the reference.

2.1.4. Multivariate analysis

Multivariate methods are capable to analyze multiple variables at a time in order to expose group-wise variation. There are two flavors for multivariate analysis focused on statistical learning, supervised and unsupervised analysis. Supervised analysis requires prior information about the sample space composition, a predicted variable, which commonly is the sample class, to train the model. In contrast, the unsupervised approach does not need extra information about sample composition to compute its routines and creates a discrimination model. In consequence, supervised approaches receive more information about group arrangement, and usually show better results than unsupervised techniques [41]. Concerning to multivariate analysis, the toolbox presents multiway principal component analysis as an unsupervised method. Also, RGCxGC presents capabilities to export chromatographic data in a compatible format to be used in external toolboxes. For example, the most popular supervised and unsupervised routines, such as partial least squares-discriminant analysis such as mixOmics [42].

2.1.4.1. Multiway principal component analysis. Principal Component Analysis is probably the most unsupervised analysis in many areas with multiple purposes such as clustering, classification, dimensional reduction. It was introduced in 1901 by Karl Pearson [43]. In chemometrics, PCA has been widely applied for pattern recognition. An adaptation of PCA, is MPCA explained by Wold [44], which can deal with higher-order data. Although there are methods that can analyze high-order data, the authors conclude that the same results can be obtained by unfolding the data into two-way matrices [45,46]. In the case of GC × GC, as stated above, each chromatogram consists of a two-way matrix of dimensions $A_{(I,\ J)}$, being I the number of modulations per run, and J the number of scans per modulation. the unfolding procedure is carried out as follows: the modulation $I + 1$ is concatenated after the modulation I, and so on for all modulations. As a result, the two-dimensional chromatogram has been unfolded into a one-dimensional row-wise vector. All chromatograms are subjected to this procedure in order to obtain a two-way matrix where the columns are the retention times and rows are samples. Then an ordinary PCA can be performed.

The PCA then decomposes $X_{(q,r)}$ into a score (S) and loading (L) matrices, so that $X = SL^T$. Score matrix is the projection in the reduced multivariate space spanned by principal components, and it is related to the (chromatographic) differences among the samples. On another hand, the loading matrix explains the relationship between variables, where positive values refer to similarities between variables and negative values denote differences in variables across samples [47]. The MPCA can be performed with the "m_prcomp" command. Prior MPCA, the user must center the signal in order to adjust flunctations around the metabolite consentration [48]. In the proposed library, chromatograms are mean centering by setting to *TRUE* the center argument (see Table 1).

Once data was subjected to MPCA, the user can access the chromatogram projection into the principal component space (scores matrix) by the "scores" command. Differently in the loading matrix, each eigenvector contains all input variables. Then, each principal component should be interpreted as a two-dimensional chromatogram. Thus, each eigenvector is retrieved from MPCA and folded again into a typical GC × GC chromatogram. This task is carried out by the "plot_loading" command. In order to improve the loading inspection, the "plot_loading" function has a threshold argument (thresh) which filter the loading value over the given value. Finally, one extra parameter in MPCA is the explained variance, which can be retrieved through the "print" command.

2.1.4.2. External discriminant analysis. Although PCA based techniques are one of the most unsupervised algorithms, GC × GC data can also be subjected to supervised algorithms, such as the Partial Least Squares-Discriminant Analysis (PLS-DA) which is one of the most common classification model widely applied in chemometrics [49]. Therefore, in order to ensure a set of discriminant analysis that can be performed once the chromatograms are exported, we also include the PLS-DA analysis. The RGCxGC library can export chromatograms in a friendly structure to communicate with external libraries. In this work, we extend our analysis by testing our datasets with PLS-DA available in mixOmics [42]. Even though, more external libraries can be used to employ any desired classification algorithm such as hierarchical clustering, artificial neural networks or supporting vector machine.

2.3. Benchmarking

In order to evaluate the performance with large GC × GC-TOF/MS data, we perform a benchmarking of every important algorithm in RGCxGC package with chromatograms from the *Salmonella* dataset, in order to simulate a large scale experiment. The *Salmonella* dataset was chosen since the mass analyzer used in this study was a time of flying. In consecuence of TOF characteristics (i.e. resolution and acquisition rate), cromatograms are dimensional higher than chromatograms analyzed with quadrupole mass analyzers. Thus, the computational efficiency was monitored with the *Salmonella* dataset. We measured the time elapsed to perform the desired routine from 1 sample, with an upper limit of 100 samples with increments of 10 samples per step, except in the first increment where it was 9 samples. The benchmarking was performed in a computer with an Intel Core i7 2.7 GHz with a Linux based operating system.

2.4. Application

The proposed software was fully tested with a real laboratory experiment data based on microbial antagonism and two published

Table 2
Description of the dataset analyzed with the proposed RGCxGC toolbox. While the dimensions of the single chromatogram are given by $A_{(I,\ J)}$, the dimensions of the entire datasets are given by $A_{(I,\ J,\ K)}$ where K represents the number of samples in each experiment.

Datasets	Samples	Number of categories	Chromatogram dimensions $A_{(I,J)}$	Mass analyzer	Reference
Salmonella	30	3	$A_{(500,\ 709)}$	TOF	[52]
Penicillium	18	2	$A_{(150,\ 368)}$	Quadrupole	[53]
Myrothecium	38	5	$A_{(125,\ 381)}$	Quadrupole	in-house

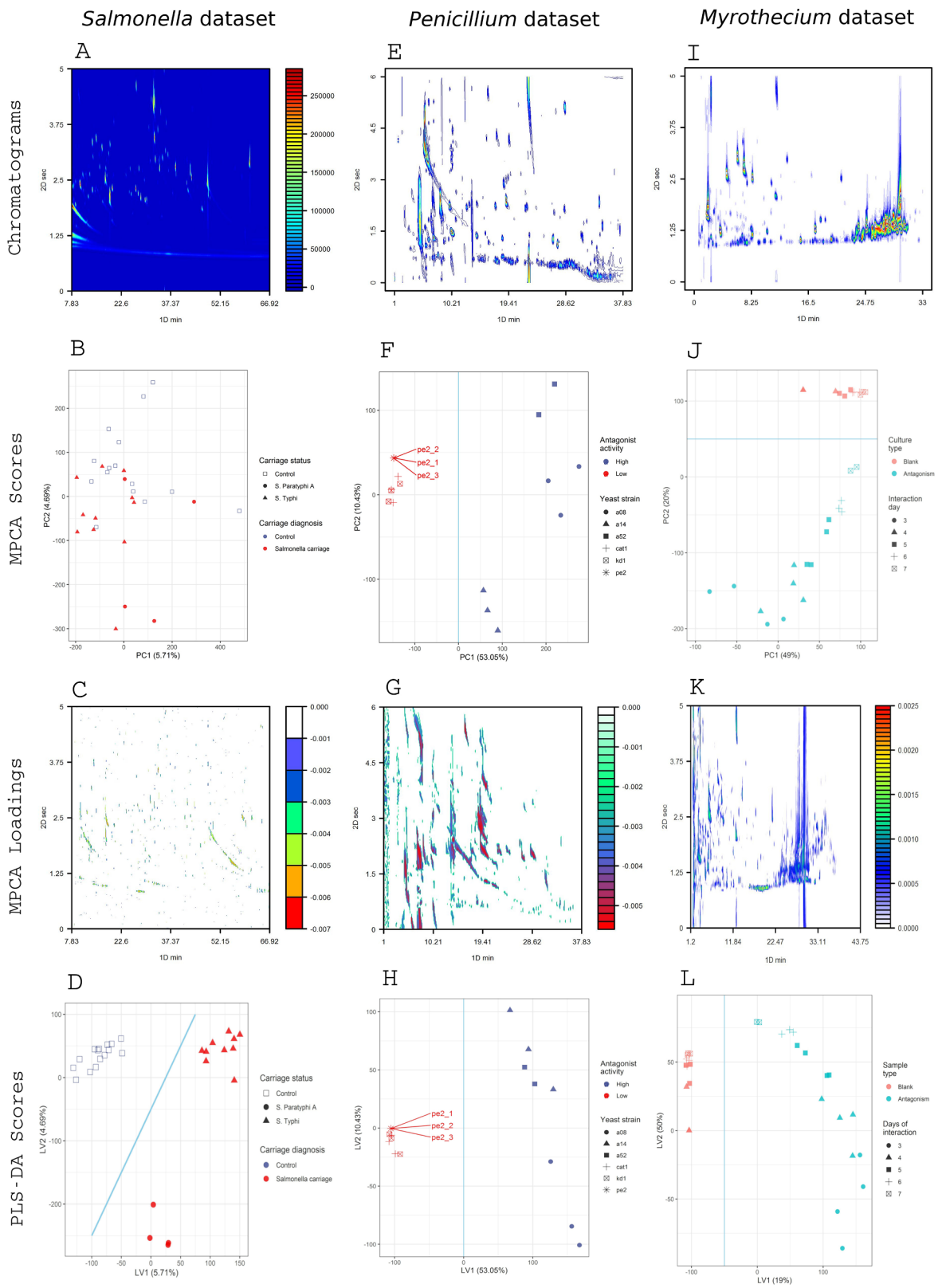

Fig. 2. The grid shows the results of the three datasets analyzed with the proposed toolbox; *Salmonella* (A, B, C, and D), *Penicillium* (E, F, G, and H) and Myrothecium (I, J, K, L). The first row shows the representative two-dimensional TIC chromatograms, the second row presents the MPCA scores, the third row displays the MPCA loadings, and the fourth row displays the PLS-DA scores. The results of unsupervised discriminant analysis by MPCA, of the *Penicillium* (F) and *Myrothecium* (G) datasets, present a clear separation between categories (edges in blue) in the two first PC's, with an explained variance greater than 50%. On the other hand, the MPCA cannot explain the differences between the *Salmonella* (B) dataset categories. In the case of supervised classification with PLS-DA, the model can discriminate between categories for all datasets (D, H, and L).

datasets. The first dataset was aforementioned in the methods section (see 2.1). The second datasets come from research focused on the discriminant analysis of chronic typhoid carriages acquired with a TOF mass analyzer [50]. The third datasets is related to antiphytopathogenic interaction of different yeast strains against *P. digitatum* acquired with quadrupole mass analyzer [51]. One principal advantage to work with the published dataset is the verification of the software performance through the comparison with real datasets, and being able to show the capability to work with different mass analyzers (Table 2).

3. Results and discussion

In the following section, we present the performance of the RGCxGC package with three different datasets collected with different mass analyzers. We have also performed a supervised analysis (PLS-DA) to confirm the connection of RGCxGC with external tools.

3.1. Salmonella dataset

The *Salmonella* dataset was downloaded from MetaboLights database with the MTBLS579 identifier [52]. Chromatograms were downloaded and manually checked for consistency; those with different dimensions specified in Table 2, were removed. The entire data comprises 30 blood plasma samples with two main categories related to carriage diagnosis, control samples, and *Salmonella sp.* carriage. The acquisition rate used in this study is 100 spectra/second. Moreover, in the second category, authors collected samples from two different etiological agents; *S. typhi* and *S. Paratyphi A*. A representative chromatogram, based on the number of peaks detected, is presented in the Fig. 2A.

All chromatograms were smoothed with a quadratic penalty and a λ equal to 10. Then, the baseline correction was performed with a correction factor equal to 1000. For the two-dimensional alignment, a sample of confirmed *S. Paratyphi* sample carriage (07_GB, MetaboLight identifier) was chosen to be used as a template and the remaining chromatogram was aligned against it. First and the second dimension was divided into 20 and 40 segments, respectively. The maximum warping values for the first and second dimensions were 2 and 8, respectively. Prior to principal component analysis, chromatograms were mean-centered.

Samples groups do not present a clear classification in the projected principal component space. Also, the explained variance for this dataset is the lowest (< 15%) obtained by MPCA in the first two principal components (Fig. 2B and C) and there is not a clear difference between the two etiological agents. This poor differentiation between *S. typhi* and *S. Paratyphi A* carriage patients maybe for the chromatogram similarity and the performance of the statistical learning algorithm. In other words, MPCA is an unsupervised approach, in which the classification is performed with no prior knowledge about the sample space composition is not suitable to discriminate between control samples and *Salmonella* carriage.

Subsequently, chromatograms were subjected to a supervised learning algorithm, PLS-DA. For this analysis, the aligned chromatograms were exported by unfolding them into a matrix. This matrix represents the explanatory variables, while the carriage status represents the predicted variable. The PLS-DA was performed according to the mixOmics procedure [42]. In contrast with the MPCA results, the PLS-DA model defines clusters between control and carriage patients (Fig. 2D). Furthermore, the model also clusters the two different etiological agents (*S. paratyphi A* and *S. typhi*). The classification improvement can be explained for the type of statistical learning employed. In the case of PLS-DA, it is a supervised algorithm, which is trained with the correct sample category. It is not surprising since supervised algorithms show better performance than unsupervised algorithms. Even though, the explained variance is not higher than the 15%.

3.2. Penicillium dataset

Raw chromatograms were provided by the authors of the antiphytopathogenic yeast strains experiment [53]. *Penicillium digitatum* is a pathogen that infects citrus fruits, causing product degradation within the product's storage, transportation, and market. For this reason, microbial activity was tested against different yeast strains by a coculture experiment in triplicates. Yeast strains, with validated high and low antagonist activity, was selected for downstream analysis.

All chromatograms were baseline corrected with a correction factor equal to 0.5. Then, chromatograms were smoothed with a linear penalty and a λ equal to 2 (Fig. 2E). For the peak alignment process, a consensus chromatogram was computed by averaging the pixel values for multiple chromatograms, as explained above, and the remaining chromatogram was aligned against it. The maximum warping value for the first and second dimension were 4 and 10, respectively. Prior to principal component analysis, chromatograms were mean-centered.

The separation achieved by MPCA was clearly notorious. In this context, the total explained variance between the first two PC was 60.48% (Fig. 2F and G). While all low activity yeast strains are clustering at the negative PC1 scores, all high active antagonist yeast strains clustered in positives values. Furthermore, triplicates of each high active yeast strains were clustered together, expressing replicate and strain similarities. For example, the strain pe2 (*Saccharomyces cerevisiae* ACB-PE2) has different profiles than cat1 and kd1, (*S. cerevisiae* ACB-CAT1 and *S. cerevisiae* ACB-KD1) together.

Furthermore, a PLS-DA classification we also conducted, as a supervised discriminant algorithm. For this analysis, the aligned chromatograms were exported by unfolding them into a matrix. This matrix represents the explanatory variables, with antiphytopathogenic activity being the predicted variable. In this case, the MPCA and PLS-DA results show high similarities, moth models can differentiate high and low strain activities in the first projected dimension. The PLS-DA was not able to separate yeast strains with low antagonist activity, it classifies all strains into a single cluster (Fig. 2H).

3.3. Myrothecium dataset

The *Myrothecium* dataset was locally made with the aforementioned methodology in the method section (2.2). The entire data comprise 38 samples with two main categories, control (Bco) and Myrothecium (Myl) antagonism interaction. Within the Myl category, there were five subcategories in concordance with the interaction days that the fungal antagonism was in the culture.

For the peak alignment process, the chromatogram from the Myl antagonism on the fourth day was selected as a representative chromatogram, and the remaining chromatograms were aligned against it (Fig. 2I). Chromatograms were baseline corrected with λ equal to 10.The maximum warping value for the first and second dimension were 4 and 10, respectively. Prior MPCA, chromatograms were mean-centered. The effect of pre-processing procedure over the chromatograms of the *Myrothecium* dataset can be found in the supplemental material.

In the MPCA score plot, the separation between control and antagonism samples was clearly appreciated (Fig. 2J and 2K). Thus, the accumulated explained variance in the two first principal components of this model was about 70%. In contrast with the two previously discussed datasets, in this case, groups were separated by the second PC. Therefore, for metabolite annotation, compounds with the highest eigenvector values were annotated, resulting in 48 metabolites with higher values of similarity than 80% in the database (Supp. Tab. 1). Between the annotated metabolites, several chemical species were found that were previously described to have fungal biocontrol activity. For example, the presence of phenylethyl alcohol, α-curcumeno and α-terpineno that were described in *Trichoderma* genus control [54]. In the same manner, these compounds had also been reported to be produced

by *Memnoniella* genus, which can induce pathogen resistance in plants through its volatilome [55]. Moreover, phenylethyl alcohol was also found in antimicrobial extracts from the *Glicladium* genus [56].

Finally, we also performed a PLS-DA model on the in-house *Myrothecium* dataset. As stated earlier, the pre-processed chromatograms were exported by unfolding them into a two-way matrix. The PLS-DA, also present consistency with the MPCA, with a rotation in the latent variables (Fig. 2L). In other words, in contrast with the MPCA that captures the variance in the second projected variable (PC2), the PLS-DA captured the experiment variance in the first projected variable (latent variable). Moreover, both models can also explain intra-day antagonism variability.

3.4. Computational efficiency

In order to simulate a large scale experiment, we performed the benchmarking of the main functions of the proposed toolbox. Then, chromatograms of the *Salmonella* dataset were chosen due to the higher dimensions, as described in Table 1. All available functions showed to have a linear increment in the elapsed time to be performed (Fig. 3). Moreover, the longest time required to perform a chemometric routine with 100 samples was 82 min was the "twod_cow" routine. This was expected since peak alignment is the most time-consuming task and the major bottleneck in this type of analysis, since different algorithms and parameters may be tested before useful information can be extracted from the raw signals. In the case of the MPCA routine, the benchmarking showed to have the second longest elapsed time, requiring 17.5 min to analyze 100 samples. Furthermore, functions that manage graphical components ("plot" and "plot_loading") require over 3 min to display 100 samples. On the other hand, the rest of the functions did not need half of a minute to work with the maximum number of samples tested. In comparison with similar alignment methodologies for GC × GC, this procedure is the most time-consuming routine [57]. For example, in the case of a MATLAB pixel-by-pixel alignment, between 10 to 20 min were required to align a 1 sample [57]. In the case of RGCxGC package, with the same time range, 10 to 20 samples can be processed.

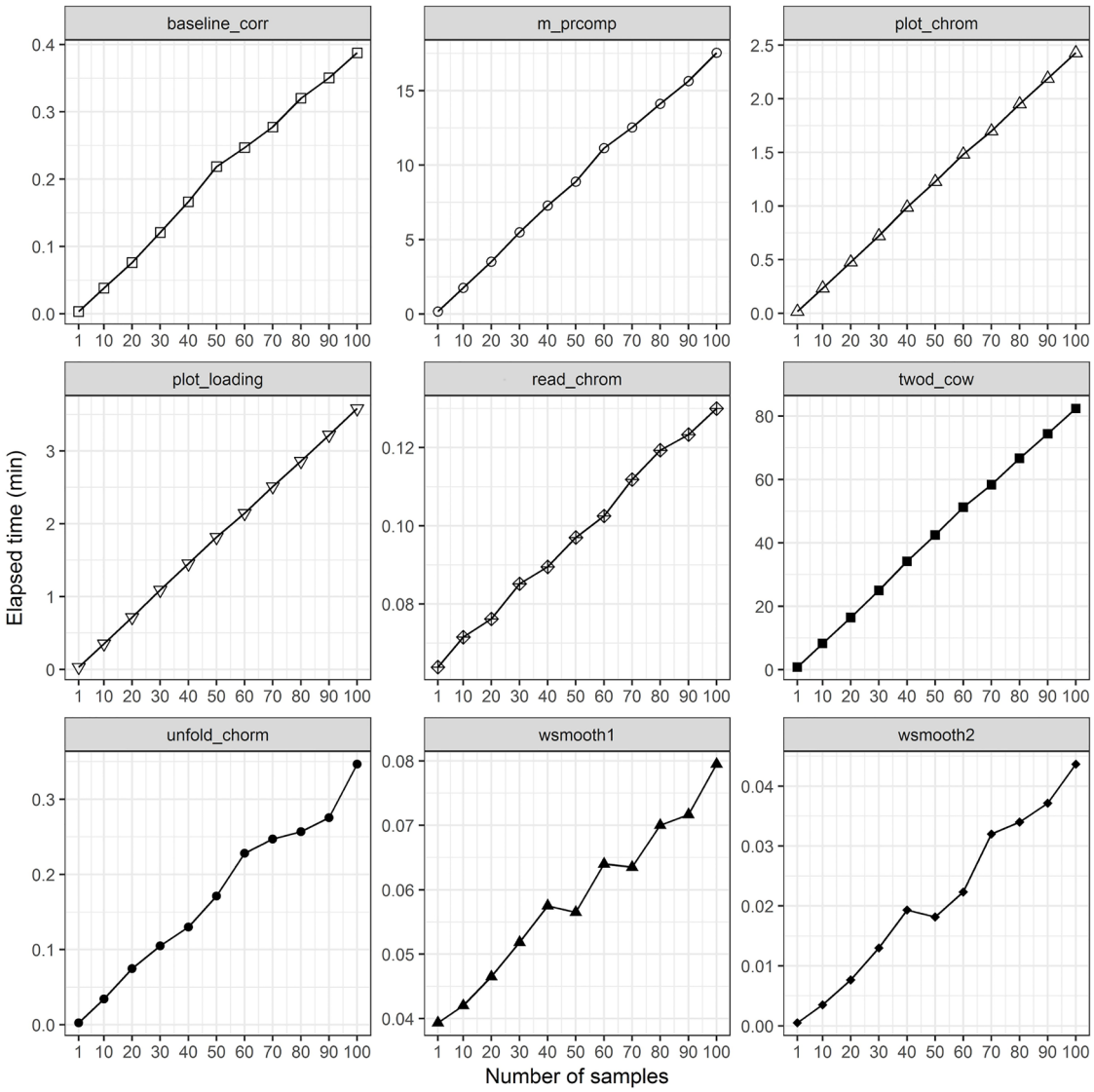

Fig. 3. Benchmarking results of the main functionalities of the RGCxGC toolbox. The head of each figure represents the function name that was tested. In order to create a similar situation of a big scale experiment, we tested the elapsed time for each function from 1 to 100 chromatograms samples in increments of 10 samples, except for the first step where the increment was 9 samples. All functions present a linear increment while the number of samples increases. Meanwhile "m_prcom" and "twod_cow" require 20 and 82 min to process 100 samples, the rest of the functions requires less than 5 min to process 100 samples.

Conclusions

In this manuscript, we present a novel end-to-end workflow for non-targeted GC × GC–MS exploratory data analysis by signal processing and with statistical learning algorithms. The currently available functions for signal processing in the RGCxGC package are compiled with baseline correction, smoothing, two-dimensional peak alignment. While for statistical learning, the multiway principal component analysis was implemented for unsupervised classification and partial least squares discriminant analysis was tested. Furthermore, we provide a generic manner to connect with other libraries in order to provide a wide spectrum of possible classification algorithms, such as linear discriminant analysis, artificial neuronal networks.

The presented software showed to be capable to process a considerable amount of data (> 1 GB). Also, the longest required time for the desired routine to be performed was less than 2 h. This is an advantage for large-scale experiments, such as metabolomics studies, to overcome the bottleneck of data analysis. On the other hand, the characteristic of free open source implementation could help with research reproducibility and reduce the dependence of private license depending software. Our approach was successfully tested in two published datasets and one in-house dataset. The key benefit of this implementation is to avoid multiple software in non-target studies. In addition, the software is continuously checked and maintained by package developers and CRAN team, in order to ensure long term user availability and avoid obsolescence. The proposed library, as well as a detailed user manual and a complete tutorial, is freely available and can be found at https://cran.r-project.org/web/packages/RGCxGC/index.html. mmc1.docx mmc2.xlsx mmc3.xlsx mmc4.xlsx

CRediT authorship contribution statement

Cristian Quiroz-Moreno: Software. **Mayra Fontes Furlan:** Resources. **João Raul Belinato:** Data curation. **Fabio Augusto:** Writing - review & editing. **Guilherme L. Alexandrino:** Validation, Writing - original draft. **Noroska Gabriela Salazar Mogollón:** Supervision, Project administration.

Declaration of Competing 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.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.microc.2020.104830.

References

[1] C. Poole, Gas Chromatography, 1st ed, Elsevier, 2012.
[2] S.E. Prebihalo, K.L. Berrier, C.E. Freye, H.D. Bahaghighat, N.R. Moore, D.K. Pinkerton, R.E. Synovec, Multidimensional Gas Chromatography, Advances in instrumentation, chemometrics, and applications, Anal. Chem 90 (2018) 505–532, https://doi.org/10.1021/acs.analchem.7b04226.
[3] P.Q. Tranchida, Comprehensive two-dimensional gas chromatography: a perspective on processes of modulation, J. Chromatogr. A. 1536 (2018) 2–5, https://doi.org/10.1016/j.chroma.2017.04.039.
[4] B. Gruber, B.A. Weggler, R. Jaramillo, K.A. Murrell, P.K. Piotrowski, F.L. Dorman, Comprehensive two-dimensional gas chromatography in forensic science: a critical review of recent trends, TrAC - Trends Anal. Chem 105 (2018) 292–301, https://doi.org/10.1016/j.trac.2018.05.017.
[5] A.M. Muscalu, T. Górecki, Comprehensive two-dimensional gas chromatography in environmental analysis, TrAC - Trends Anal. Chem 106 (2018) 225–245, https://doi.org/10.1016/j.trac.2018.07.001.
[6] B.J. Pollo, G.L. Alexandrino, F. Augusto, L.W. Hantao, The impact of comprehensive two-dimensional gas chromatography on oil & gas analysis: recent advances and applications in petroleum industry, TrAC - Trends Anal. Chem 105 (2018) 202–217, https://doi.org/10.1016/j.trac.2018.05.007.
[7] T. Miyazaki, K. Okada, T. Yamashita, M. Miyazaki, Two-dimensional gas chromatography time-of-flight mass spectrometry-based serum metabolic fingerprints of neonatal calves before and after first colostrum ingestion, J. Dairy Sci 100 (2017) 4354–4364, https://doi.org/10.3168/jds.2017-12557.
[8] J.C. Giddings, Sample dimensionality: a predictor of order-disorder in component peak distribution in multidimensional separation, J. Chromatogr. A. 703 (1995) 3–15, https://doi.org/10.1016/0021-9673(95)00249-M.
[9] R.B. Wilson, W.C. Siegler, J.C. Hoggard, B.D. Fitz, J.S. Nadeau, R.E. Synovec, Achieving high peak capacity production for gas chromatography and comprehensive two-dimensional gas chromatography by minimizing off-column peak broadening, J. Chromatogr. A. 1218 (2011) 3130–3139, https://doi.org/10.1016/j.chroma.2010.12.108.
[10] L. Yi, N. Dong, Y. Yun, B. Deng, D. Ren, S. Liu, Y. Liang, Chemometric methods in data processing of mass spectrometry-based metabolomics: a review, Anal. Chim. Acta 914 (2016) 17–34, https://doi.org/10.1016/j.aca.2016.02.001.
[11] K.M. Pierce, B. Kehimkar, L.C. Marney, J.C. Hoggard, R.E. Synovec, Review of chemometric analysis techniques for comprehensive two dimensional separations data, J. Chromatogr. A. 1255 (2012) 3–11, https://doi.org/10.1016/j.chroma.2012.05.050.
[12] E. Szymańska, Modern data science for analytical chemical data – A comprehensive review, Anal. Chim. Acta (2018) 1028, https://doi.org/10.1016/j.aca.2018.05.038.
[13] R. Brereton, Chemometrics For Pattern Recognition, Wiley, 2009.
[14] J. Krumsiek, F.J. Theis, Statistical methods for the analysis of high-throughput metabolomics data abstract : metabolomics is a relatively new high-throughput technology that aims at measuring all endogenous metabolites within a biological sample in an unbiased fashion, Resu. Comput. Struct. Biotechnol. J. (2013) 4.
[15] L. Komsta, Y. Vander Heyde, J. Sherman, Chemometrics in Chromatography, 1st ed., CRC Press, 2018.
[16] D. Zhang, X. Huang, F.E. Regnier, M. Zhang, Two-dimensional correlation optimized warping algorithm for aligning GCxGC-MS data, Anal. Chem 80 (2008) 2664–2671, https://doi.org/10.1021/ac7024317.
[17] P.H.C. Eilers, Parametric time warping, Anal. Chem 76 (2004) 404–411, https://doi.org/10.1021/ac034800e.
[18] B. Wang, A. Fang, J. Heim, B. Bogdanov, S. Pugh, M. Libardoni, X. Zhang, DISCO: distance and spectrum correlation optimization alignment for two dimensional gas chromatography Time-of- Flight Mass spectrometry-based metabolomics, Anal. Chem. 82 (2011) 5069–5081, https://doi.org/10.1021/ac100064b.DISCO.
[19] B. Wang, A. Fang, X. Shi, S.H. Kim, X. Zhang, DISCO2?: A Comprehensive Peak Alignment Algorithm for Two-Dimensional Gas Chromatography Time-of-Flight Mass Spectrometry, Int. Conf. on Intell. Compu. (2012) 486–491.
[20] N. Hoffmann, M. Wilhelm, A. Doebbe, K. Niehaus, J. Stoye, BiPACE 2D-Graph-based multiple alignment for comprehensive 2D gas chromatography-mass spectrometry, Bioinformatics 30 (2014) 988–995, https://doi.org/10.1093/bioinformatics/btt738.
[21] S. Castillo, I. Mattila, J. Miettinen, M. Orešič, T. Hyötyläinen, Data analysis tool for comprehensive two-dimensional gas chromatography/time-of-flight mass spectrometry, Anal. Chem 83 (2011) 3058–3067, https://doi.org/10.1021/ac103308x.
[22] E.A. Higgins Keppler, C.L. Jenkins, T.J. Davis, H.D. Bean, Advances in the application of comprehensive two-dimensional gas chromatography in metabolomics, TrAC Trends Anal. Chem 109 (2018) 275–286, https://doi.org/10.1016/J.TRAC.2018.10.015.
[23] R.C. Ramaker, E. Gordon, S.J. Cooper, R2DGC : threshold-free peak alignment and identifica- tion for 2D gas chromatography mass spectrometry in R, Bioinformatics 34 (2017) 1789–1791.
[24] S. Moayedpour, H. Parastar, RMet: an automated R based software for analyzing GC-MS and GC × GC-MS untargeted metabolomic data, Chemom. Intell. Lab. Syst 194 (2019) 103866, , https://doi.org/10.1016/j.chemolab.2019.103866.
[25] R.G. Brereton, G.R. Lloyd, Partial least squares discriminant analysis: taking the magic away, J. Chemom 28 (2014) 213–225, https://doi.org/10.1002/cem.2609.
[26] Y. Zushi, S. Hashimoto, Direct classification of GC × GC-Analyzed complex mixtures using non-negative matrix factorization-based feature extraction, Anal. Chem 90 (2018) 3819–3825, https://doi.org/10.1021/acs.analchem.7b04313.
[27] P.E. Sudol, D.V. Gough, S.E. Prebihalo, R.E. Synovec, Impact of data bin size on the classification of diesel fuels using comprehensive two-dimensional gas chromatography with principal component analysis, Talanta 206 (2020) 120239, , https://doi.org/10.1016/J.TALANTA.2019.120239.
[28] Y. Izadmanesh, E. Garreta-Lara, J.B. Ghasemi, S. Lacorte, V. Matamoros, R. Tauler, Chemometric analysis of comprehensive two dimensional gas chromatography–mass spectrometry metabolomics data, J. Chromatogr. A. 1488 (2017) 113–125, https://doi.org/10.1016/j.chroma.2017.01.052.
[29] V. Den Dool, P.D. Kratz, A generalization of the retention index system including linear temperature programmed gas-liquid, J. Chrmatography A. (1962) 463–471, https://doi.org/10.1007/978-3-319-70262-9_7.
[30] T. Skov, R. Bro, Solving fundamental problems in chromatographic analysis, Anal. Bioanal. Chem 390 (2008) 281–285, https://doi.org/10.1007/s00216-007-1618-z.
[31] S.E. Reichenbach, M. Ni, V. Kottapalli, A. Visvanathan, Information technologies for comprehensive two-dimensional gas chromatography, Chemom. Intell. Lab. Syst 71 (2004) 107–120, https://doi.org/10.1016/j.chemolab.2003.12.009.
[32] E. Pante, B. Simon-Bouhet, marmap: a package for importing, plotting and analyzing bathymetric and topographic data in R, PLoS ONE 8 (2013) 6–9, https://doi.org/10.1371/journal.pone.0073051.
[33] K. T, colorRapms, R Packag (2012).
[34] L. Yi, N. Dong, Y. Yun, B. Deng, D. Ren, S. Liu, Y. Liang, Chemometric methods in data processing of mass spectrometry-based metabolomics: a review, Anal. Chim. Acta 914 (2016) 17–34, https://doi.org/10.1016/j.aca.2016.02.001.
[35] T. Hastie, R. Tibshirani, Generalized additive models, Monogr. Stat. Appl. Probab.

15 (1990).
[36] J. Fan, I. Gijbels, Local Polynomial Modelling and Its Applications, Chapman and Hall, New York, 1996.
[37] P.H.C. Eilers, A perfect smoother, Anal. Chem. 75 (2003) 3631–3636, https://doi.org/10.1021/ac034173t.
[38] S. Brown, L. Sarabia, T. Johan, Comprehensive chemometrics: Chemical and Biochemical Data Anlysis, Elsevier, 2009.
[39] J. Peng, S. Peng, A. Jiang, J. Wei, C. Li, J. Tan, Asymmetric least squares for multiple spectra baseline correction, Anal. Chim. Acta 683 (2010) 63–68, https://doi.org/10.1016/j.aca.2010.08.033.
[40] Dabao Zhang, Xiaodong Huang, E Fred, Regnier, min zhang, two-dimensional correlation optimized warping algorithm for aligning GCXGC-MS data, Anal. Chem 80 (2008) 2664–2671, https://doi.org/10.1021/ac7024317.
[41] G. James, D. Witten, T. Hastie, R. Tibshirani, An Introduction to Statistical Learning: with Applications in R, Springer Science & Business Media, 2013.
[42] F. Rohart, B. Gautier, A. Singh, K.-A. Lê Cao, mixOmics: an r package for 'omics feature selection and multiple data integration, PLoS Comput. Biol. 13 (2017) e1005752, , https://doi.org/10.1371/journal.pcbi.1005752.
[43] K. Pearson, On lines and planes of closest fit to systems of points in space, london, edinburgh, dublin philos, Mag. J. Sci. 2 (1901) 559–572, https://doi.org/10.1080/14786440109462720.
[44] S. Wold, P. Geladi, K. Esbensen, J. Öhman, Multi-way principal components- and PLS-analysis, J. Chemom 1 (1987) 41–56, https://doi.org/10.1002/cem.1180010107.
[45] L.S. Ramos, K.R. Beebe, W.P. Carey, S.M. Eugenio, B.C. Erickson, B.E. Wilson, B.R. Kowalski, L.E. Wangen, Chemometrics, Anal. Chem 58 (1986) 294–315, https://doi.org/10.1021/ac00296a020.
[46] P. Geladi, Analysis of multi-way (multi-mode) data, Chemom. Intell. Lab. Syst. 7 (1989) 11–30, https://doi.org/10.1016/0169-7439(89)80108-X.
[47] M. Castro, Quimiometria: Conceitos, Métodos e Aplicações, 1st ed., UNICAMP, Brazil, 2009.
[48] R.A. van den Berg, H.C.J. Hoefsloot, J.A. Westerhuis, A.K. Smilde, M.J.van der Werf, Centering, scaling, and transformations: improving the biological information content of metabolomics data, BMC Genomics 7 (2006) 1–15, https://doi.org/10.1186/1471-2164-7-142.
[49] J. Bartel, J. Krumsiek, F.J. Theis, Statistical methods for the analysis of high-throughput metabolomics data, Comput. Struct. Biotechnol. J. 4 (2013) e201301009, , https://doi.org/10.5936/csbj.201301009.
[50] E. Nässtrom, P. Jonsson, A. Johansson, S. Dongol, A. Karkey, B. Basnyat, N. Tran Vu Thieu, T. Trinh Van, G.E. Thwaites, H. Antti, S. Baker, Diagnostic metabolite biomarkers of chronic typhoid carriage, PLoS Negl. Trop. Dis. 12 (2018) 1–15, https://doi.org/10.1371/journal.pntd.0006215.
[51] J.R.B. de Souza, K.C. Kupper, F. Augusto, In vivo investigation of the volatile metabolome of antiphytopathogenic yeast strains active against penicillium digitatum using comprehensive two-dimensional gas chromatography and multivariate data analysis, Microchem. J. 141 (2018) 204–209, https://doi.org/10.1016/j.microc.2018.05.036.
[52] E. Nässtrom, P. Jonsson, A. Johansson, S. Dongol, A. Karkey, B. Basnyat, N. Tran Vu Thieu, T. Trinh Van, G.E. Thwaites, H. Antti, S. Baker, Diagnostic metabolite biomarkers of chronic typhoid carriage, PLoS Negl. Trop. Dis. 12 (2018) 1–15, https://doi.org/10.1371/journal.pntd.0006215.
[53] J.R. Belinato, K.C. Kupper, F. Augusto, In vivo investigation of the volatile metabolome of antiphytopathogenic yeast strains active against penicillium digitatum using comprehensive two-dimensional gas chromatography and multivariate data analysis, Microchem. J. 141 (2018) 362–368, https://doi.org/10.1016/J.MICROC.2018.05.047.
[54] N. Stoppacher, B. Kluger, S. Zeilinger, R. Krska, R. Schuhmacher, Identification and profiling of volatile metabolites of the biocontrol fungus Trichoderma atroviride by HS-SPME-GC-MS, Jour. of Microbio. Met. 81 (2010) 187–193.
[55] P.F. De Lima, M.F. Furlan, F.A. De Lima Ribeiro, S.F. Paschollati, F. Augusto, In vivo determination of the volatile metabolites of saprotroph fungi by comprehensive two-dimensional gas chromatography, J. Sep. Sci. 38 (2015) 1924–1932, https://doi.org/10.1002/jssc.201401404.
[56] K. Liouane, D. Saïdana, H. Edziri, S. Ammar, J. Chriaa, M.A. Mahjoub, K. Said, Z. Mighri, Chemical composition and antimicrobial activity of extracts from Gliocladium sp. growing wild in Tunisia Kaouthar Liouane, Med. Chem. Res. 19 (2010) 743–756, https://doi.org/10.1007/s00044-009-9227-3.
[57] Y. Zushi, J. Gros, Q. Tao, S.E. Reichenbach, S. Hashimoto, J.S. Arey, Pixel-by-pixel correction of retention time shifts in chromatograms from comprehensive two-dimensional gas chromatography coupled to high resolution time-of-flight mass spectrometry, J. Chromatogr. A. 1508 (2017) 121–129, https://doi.org/10.1016/j.chroma.2017.05.065.