Briefings in Bioinformatics, 2025, **26(2)**, bbaf080
https://doi.org/10.1093/bib/bbaf080
Problem Solving Protocol

GcDUO: an open-source software for GC × GC–MS data analysis

Maria Llambrich[1], Frans M. van der Kloet[2], Lluc Sementé[3], Anaïs Rodrigues[4], Saer Samanipour[5], Pierre-Hugues Stefanuto[4], Johan A. Westerhuis[2], Raquel Cumeras[1,3,*], Jesús Brezmes[1]

[1]Department of Electrical Electronic Engineering and Automation, Universitat Rovira i Virgili (URV), IISPV, C/Escorxador S/N, 43003, Tarragona, Spain
[2]Biosystems Data Analysis Group, Swammerdam Institute for Life Sciences, University of Amsterdam, P.O. Box 94215, 1090 GE Amsterdam, Netherlands
[3]Oncology Department, Hospital Universitari Sant Joan de Reus, Institut d'Investigació Sanitària Pere Virgili (IISPV), CERCA, Av. Joan Laporte 2, 43204, Reus, Spain
[4]Organic and Biological Analytical Chemistry Group, MolSys Research Unit, University of Liège, Allée du 6 aout, 11, B4000 Liège, Belgium
[5]Van't Hoff Institute for Molecular Sciences (HIMS), University of Amsterdam, PO Box 94157, 1090 GD Amsterdam, Netherlands

*Corresponding author. Oncology Department, Hospital Universitari Sant Joan de Reus, Institut d'Investigació Sanitària Pere Virgili (IISPV), CERCA, Reus, Spain. E-mail: raquel.cumeras@iispv.cat

Abstract

Comprehensive 2D gas chromatography coupled with mass spectrometry (GC × GC–MS) is a powerful analytical technique. However, the complexity and volume of data generated pose significant challenges for data processing and interpretation, limiting a broader adoption. Chemometric approaches, particularly multiway models like Parallel Factor Analysis (PARAFAC), have proven effective in addressing these challenges by enabling the extraction of meaningful chemical information from multi-dimensional datasets. However, traditional PARAFAC is constrained by its assumption of data tri-linearity, which may not be valid in all cases, leading to potential inaccuracies. To overcome these limitations, we present GcDUO, an open-source software implemented in R, designed specifically for the processing and analysis of GC × GC–MS data. GcDUO integrates advanced chemometric methods, including both PARAFAC and PARAFAC2, for a more accurate and comprehensive analysis. PARAFAC is particularly useful for deconvoluting overlapping peaks and extracting pure chemical signals, while PARAFAC2 relaxes de tri-linearity constraint, allowing the alignment between samples. The software is structured into six modules—data import, region of interest (ROI) selection, deconvolution, peak annotation, data integration, and visualization—facilitating comprehensive and flexible data processing. GcDUO was validated against the gold-standard software for comprehensive GC, demonstrating a high correlation ($R^2 = 0.9$) in peak area measurements, confirming its effectiveness and reliability. GcDUO provides a valuable, open-source platform for researchers in metabolomics and related fields, enabling more accessible and customizable GC × GC–MS data analysis.

Keywords: GC × GC–MS; multi-dimensional chromatography; metabolomics; chemometrics; PARAFAC; PARAFAC2; open-source software

Introduction

Comprehensive 2D gas chromatography (GC × GC–MS) has emerged as a powerful analytical tool for the untargeted analysis of complex samples [1–3]. However, the difficulties in the analysis of the data have prevented its widespread use [4, 5]. Together with the new commercial systems that increase resolution significantly, the data obtained has grown and its manipulation has become a serious computational issue [6]. In response to this, chemometrics—a multidisciplinary field integrating mathematics, statistics, and formal logic [7]—has been essential in maximizing the chemical information extracted from complex datasets, ensuring the appropriateness of the data, signal enhancement, noise minimization, and model building to get meaningful information about the sample measured.

Tensor methods, also known as multiway models, have proven to be a valuable option for extracting meaningful patterns from complex metabolomics datasets [4]. The tridimensionality inherent in GC × GC–MS data makes it an ideal input for chemometric tools such as Parallel Factor Analysis (PARAFAC) [8], although to preserve the tri-linearity of the data, careful experimental design is required [1, 4, 9, 10]. Despite the proven effectiveness of PARAFAC in GC–MS data analysis [11], its use in comprehensive analysis software using GC × GC–MS data in open-source platforms remains scarce. This is mainly caused by two reasons, one being the computational cost, and the second being the necessity to have prior knowledge of the components present in the samples. Furthermore, in untargeted analysis, the regions of interest (ROIs) are unknown whereas the selection of an accurate region for analysis helps in the deconvolution of peaks as less interferences is found [12]. Commercial software, however, often fails in providing comprehensive data interpretation, necessitating the development of advanced chemometric tools [10].

PARAFAC, widely employed in multi-way chromatography since the early 2000s, decomposes high dimensional data into a linear combination of second order tensors, enabling the differentiation between analyte signal and background noise [8, 13]. However, optimization of parameters such as the number of components and matrix dimensions are essential for accurate

Received: October 28, 2024. **Revised:** December 27, 2024. **Accepted:** February 17, 2025

analysis [10], making the automatization of the process difficult. Another major concern is the requirement of data tri-linearity for the use of the PARAFAC algorithm. Lately, an extension of PARAFAC, PARAFAC2, has been proposed, offering a solution as fewer restrictions on the data structure are required, allowing the loss of tri-linearity caused by misalignment between samples. In other words, the model does not assume the exact same shape for the same compound peak between samples [14]. To solve the tri-linearity requirement, multivariate curve resolution-alternative least squares (MCR-ALS) presents a different method but is limited by the lack of unique solutions and the requirement of bi-linearity in the data [1, 10]. This implies that their application to GC × GC–MS requires data unfolding, avoiding the simultaneous sample analysis and maintaining its original structure, which causes loss of information. Another proposal, PARADISe (PARAFAC2-based Deconvolution and Identification System) is a software tool designed to extract comprehensive chemical/metabolite information from raw GC–MS [11]. However, PARADISe ROIs are selected manually. Selecting the ROIs in GC × GC–MS data is a crucial step in the data analysis process since it determines the specific portion of the chromatogram to work in each iteration. Moreover, manual selection becomes problematic in untargeted analysis, where the exact number of compounds is unknown, and in large datasets, where the process is highly time-consuming.

Several software options are available for GC × GC–MS data analysis. Commercial solutions, which require a license, are typically designed to perform all steps through a graphical user interface [15]. The most used is ChromaTOF™ (LECO Corp., USA), which handles complex datasets by relying on peak tables. However, it requires a large computational capacity and can be challenging to use with large datasets [6]. Other software includes GC Image™ (Zoex Corp., USA) and GasPedal (DecoDon Software UG, Germany), which are specialized in visualization using an image-based approach [16, 17]; Chromspace® (SepSolve Analytical Ltd, UK) offering an intuitive interface for non-targeted analysis [1]; and AnalyzerPro® XD (SpectralWorks Ltd, UK), a GC–MS software adapted for GC × GC–MS. Additional commercial options include Canvas-2DGC® (J&X Technologies, China), ChromSquare®, and GasPedal®. Open-source alternatives, such as R packages like gcxgclab [18] or RGC × GC [19], are also available. The package gcxgclab focuses on pre-processing tasks such as peak detection and alignment, while RGC × GC transforms the 3D data into 2D for the application of classical GC–MS methods. Their main limitation is that these tools process samples individually and then combine results, which is computationally expensive and prone to peak alignment errors. Without batch processing, it becomes difficult to identify and correct systematic errors that may affect all samples in an analytical run. Therefore, there is a need for open-source software with batch processing features, something lacking in GC × GC.

Here, we present GcDUO—an open-source data processing software - that enables annotation, deconvolution, and analysis of batch GC × GC–MS data. GcDUO uses raw data to extract the features without any previous knowledge of the samples and then deconvolute and identify them. The GcDUO batch approach ensures consistency in peak detection, alignment, and annotation across all samples. We have thoroughly evaluated the differences in data deconvolution when using PARAFAC and PARAFAC2 algorithms. Through extensive testing against the gold-standard software for comprehensive GC–MS, ChromaTOF, we showcase GcDUO's potential in addressing the challenges of metabolomics data analysis with GC instruments. GcDUO is implemented in the R software platform and is available at GitHub: https://github.com/mariallr/GcDuo.

Materials and methods

Software architecture

GcDUO comprises six modules: data import, ROI selection, deconvolution, peak annotation, data integration, and visualization. GcDUO accepts Computable Document Format (CDF) files, a non-vendor-specific format, and therefore an open standard, which contains raw data as input (Fig. 1). GcDUO is implemented as a package in the R programming language.

Module I: data import

The data import module extracts individual vectors of intensities and mass-to-charge values from the CDF files, which are standardized formats that store the raw data captured by the mass spectrometry (MS), making data open and accessible [20]. These vectors are folded into a 4D tensor (I × J × K × L tensor) where the I dimension or mode is related to sample identifiers, J is for the mass to charge ions (m/z), K is for the acquisition values of the second-dimension time, and L is for the acquisition values of the first-dimension time. To construct the tensor, information on modulation (period between pulses) and mass-to-charge range needs to be provided.

The netCDF data is stored as independent vectors, including "scan_acquisition_time," "intensity_values," "mass_values," and "point_count." The "scan_acquisition_time" vector provides equidistant acquisition time points, and "point_count" is used to extract corresponding intensity and mass pairs. This information is organized into a list containing the time point and mass-intensity pairs. For non-nominal mass resolution, mass values are converted to integers, grouped, and padded with zeros to ensure equidistant mass-to-charge fragments. The matrix is constructed using the frequency and the modulation to fold the vectors. When second-dimension shifting is required, zero values are added at the beginning to create the necessary positions.

Module II: Region of interest (ROI) selection by inverse watershed algorithm

The inverse-watershed algorithm is used for ROI selection as it effectively identifies prominent objects against the background. The watershed algorithm is applied within a rolling window framework, where data is partitioned into smaller sections based on the selected number of modulations. This step ensures accurate detection of peaks that might otherwise be truncated at window boundaries. To decrease the processing time and increase efficiency, a rolling window spanning several modulations (between 2 and 4 are recommended) is applied to partition data prior to applying the inverse watershed algorithm [21]. Briefly, the algorithm evaluates the data to find prominent points and then delimits their region in a 2D space. The algorithm output is "blobs," which refer to 3D-peaks and their 2D coordinates [22]. Subsequently, the blobs undergo an evaluation using the minimum signal-to-noise (s/n) ratio determined by the user. For instance, an s/n of 10 means that the apex peak intensity must be 10 times higher than the noise, which is determined in low-frequency regions (beginning/end of the modulation) within each ROI. Furthermore, to qualify as a peak, we considered that blobs must contain a minimum of five points in the second column (K) dimension. The Gaussian shape of those chromatographic peaks is evaluated to distinguish between noise and chromatographic peaks. This evaluation ensures that the entire peak form is

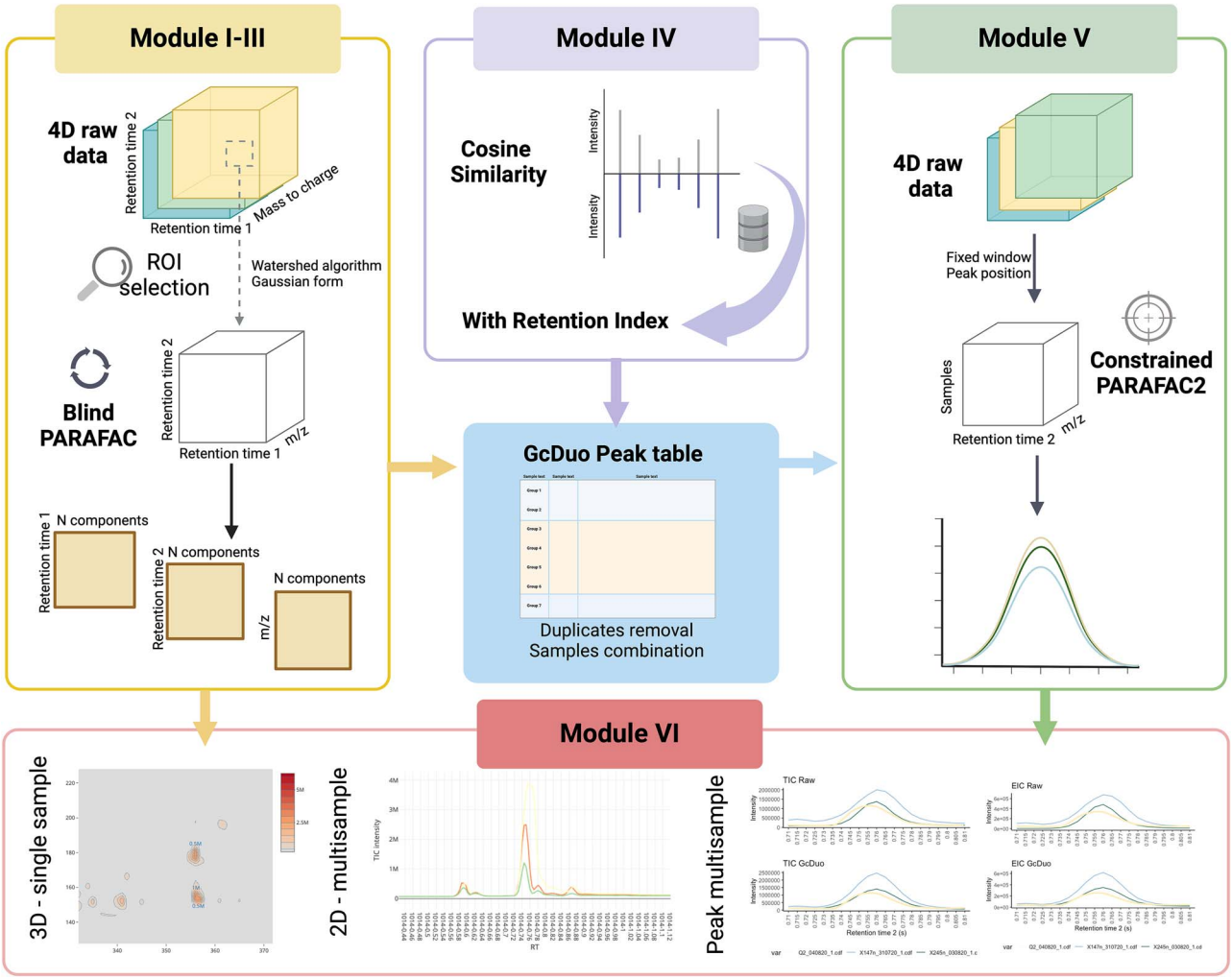

Figure 1. Workflow of GcDUO algorithm. It comprises six modules: Module I: (data import) from raw CDF files the vectors obtained are folded into a 4D matrix. Module II: (ROIs selection) ROIs are detected using the watershed algorithm. ROIs are evaluated by signal-to-noise ratio (s/n). Module III: (blind PARAFAC) deconvolution of peaks is performed with a PARAFAC algorithm applied sample-to-sample iteratively to determine the proper number of components. Consistency between results obtained between samples is checked and duplicates are removed using spectral similarity. GcDUO coordinates are kept for next modules. Module IV: annotation-obtained components are matched against a library using cosine similarity and (optionally) retention index. Module V: (constrained PARAFAC) a PARAFAC2 model is applied using the information obtained from previous steps: retention time windows, number of components, and the spectra identified, which is constrained. The obtained results are evaluated by their Gaussian form and signal-to-noise ratio. The area is calculated using the area under the curve. Module VI: data visualization different data visualizations are available using 3D, 2D, and the resolved peak plots. Created with Biorender.com.

encapsulated within the blob, serving as a quality control measure for the watershed algorithm.

This process iterates across each data window, with blob positions retained as ROIs for subsequent modules. Since the rolling window framework involves overlapping sections of data, some duplicates in blob detection may occur. These duplicates are systematically identified and removed in later processing stages to ensure unique and accurate ROI selection.

Module III: deconvolution by blind PARAFAC algorithm

To ensure the tri-linearity of the data, the PARAFAC algorithm is applied iteratively to each individual sample i and blob f [9, 23], so a 3D tensor ($J \times K \times L$ or $m/z \times RT_2 \times RT_1$) is used. As the tensor can become large, a computational strategy involving windowing L (first retention time) is applied. Therefore, the PARAFAC model $X_{ijk,l}$ per each sample, i, mass spectra j and second retention time k is applied iteratively across the running window l for the first retention time. The PARAFAC model used is represented by Eq. (1):

$$X_{ijk,l} = \sum_{f=1}^{F} A_{jf} B_{kf} D_{lf} + E_{ijk,l} \quad (1)$$

where A_{jf} are the loadings for the first mode (mass spectra j), B_{kf} are the loadings for the second mode (second retention time k), D_{lf} are the loadings for the first mode (*rentention time* l), F is the number of factors or components and $E_{ijk,l}$ is the residuals of the decomposition in the current window l. To streamline computational efficiency, the mass-to-charge dimension is reduced by selecting only the fragments with the highest percentage of variation. By default, the top 5% fragments are chosen, although users have the flexibility to adjust this parameter.

Prior to applying the PARAFAC algorithm, data must satisfy the prerequisites, a minimum of five points in the K dimension (second column) and a Gaussian chromatographic shape of the peak, therefore, the maximum point must exceed the noise threshold set by the user.

Subsequently, the algorithm is iteratively applied to determine the appropriate number of components. The number of components is increased effectively until the model's R^2 ceases to improve, and the Tucker congruence exceeds 0.9 [24]. Components derived from the PARAFAC algorithm undergo automatic scrutiny to eliminate any artifacts, with a minimum model score required. A maximum model score of less than 1 indicates a poor contribution to the component and is therefore excluded.

Since the algorithm is employed with a rolling window, it is probable that certain components describe identical chemical peaks. To avoid redundancy within each sample, duplicates of features exhibiting matching retention time, intensity and higher m/z fragments are removed. Moreover, overlap between samples is assessed by grouping features based on retention time in the first dimension and the highest m/z fragment. Subsequently, spectral similarity is evaluated using the dot product, with a minimum threshold of 60% required for features to be considered identical chemical peaks across samples. Features detected in only one sample are discarded. Then a consensus spectrum is created by the mean of all samples with that detected feature [25]. The iteration and the data size of GC × GC–MS samples make this process lengthy. With the goal to reduce the computational time, all the processes have been parallelized using the R package parallel.

Module IV: peak annotation

Prior to implementing the constrained PARAFAC2 model, the consensus spectra are matched against a library using cosine similarity and retention index, as described elsewhere [26]. The function will consider an identification as correct when the cosine similarity exceeds the user-determined threshold, optionally refined using the retention index. The matched spectra are then retained for use in module V.

Module V: data integration by the constrained PARAFAC2 algorithm

Once initial PARAFAC results are obtained and the model's robustness across samples is assessed, following the same requisites as in module III, this information is used to further refine the model. For each detected feature, a new PARAFAC2 model is calculated, incorporating parameters such as the number of components, chromatographic region for evaluation, and identified compound spectra. It uses the same equation as PARAFAC (Equation (1)) but the matrix B_{kf} can vary between samples, allowing for different elution profiles. Additionally, it does not require strict tri-linearity in the data. This approach facilitates the detection of peaks that might be missed in samples with low signal intensity. The library spectra are imposed in the PARAFAC2 model. By constraining one mode, improvements in chromatographic elution profile are observed, allowing the algorithm to effectively identify chemical peaks with minimal variation from noise. Upon completion, the area and intensity of each peak are obtained from the reconstructed matrix. Since the PARAFAC2 model uses all the samples (batch), the detection of peaks in low-concentration samples becomes possible.

Module VI: data visualization

For visual inspection of the data, we have created three visualization plots (Fig. 1). First a contour plot for each individual sample which is useful to visualize problems in its elution (i.e., chromatogram). Secondly, a chromatogram from the demodulated data for selected samples, with the purpose of observing misalignments between them. Finally, the resolved peaks plot to check the form of the peaks in order to discard non-correct peaks or shoulders, among others.

GC × GC–MS datasets

Training dataset

To develop the GcDUO software, a public dataset from Weggler *et al.* [27] and available at Harvard Dataverse (KA5BTU) was used. A commercial mix of fragrance and allergen standards (Restek, Bellefonte, USA) was diluted with methyl tert-butyl ether to the concentrations of 2, 1, 0.4, and 0.2 ppb. The internal standard used was 20 mg/ml of 1-fluoronaphtalene (Restek, Bellefonte, USA). The retention index was calculated by adding 50 mg/ml of a standard mixture of n-alkanes (Restek, Bellefonte, USA). Each dilution was measured in triplicate.

Validation datasets

We tested the results against ChromaTOF software v5.56.53 for Pegasus to evaluate the consistency of the results. For validation, we used two datasets.

First, we used a publicly available fruitybeer dataset [28]. Briefly, the aroma profile of five beers was investigated using purge-and-trap coupled with comprehensive GC and MS. Four replicates of each beer sample were analyzed by transferring 10 ml of degassed beer to a 20 ml headspace vial containing 1 g of NaCl. The trap-and-purge system was packed with Tenax TA before analysis using a Pegasus 4D GC × GC-TOF instrument coupled with an Agilent 7890 GC.

As a second test to evaluate our software, we used an in-house dataset obtained from an in-house QC solution, called Breath Mix, containing 12 molecules (see Supplementary Table S1 for composition) and an n-alkanes solution (13 molecules, C8-C20). Both solutions were prepared from pure analytical standards and diluted to 1 ppm in methanol and hexane, respectively. The solutions were injected into two distinct LECO GC × GC-ToFMS systems equipped, respectively, with an LN2 cryogenic and a cryo-free modulator. The standard compounds (2-hexanone, p-xylene, cyclohexanone, 6-methyl-5-hepten-2-one, acetophenone, decane, undecane, nonanal, 1-tetradecene, 6,10-dimethyl-5,9-undecandien-2-one, (Z), 6,10-dimethyl-5,9-undecandien-2-one, (E), 1-pentadecene) were purchased at Restek Corp. (Bellefonte, USA). The Alkane standard mixture C7-C30 was obtained from Sigma-Aldrich (Saint Louis, USA).

Results and discussion

The optimization of GcDUO is performed using the training dataset. In each module, we evaluated the performance of the selected parameters, and their efficiency compared to the expected results reported by the authors. To illustrate the software's functionalities, we selected three scenarios: the peak for Lilial, the most abundant standard with a well-resolved peak; the peak for limonene and 1,8-cineole, two partially overlapping peaks with differing concentrations; and the peak for 4-methoxybenzenyl alcohol, one of the smallest peaks coeluting with other components (Supplementary Fig. S1).

Data import

Module I reconstructs the 4D matrix from the raw data, obtaining the first GcDUO object. The main drawback of using CDF files is the limited metadata available, and we have to remember that the folding step is of high relevance. The incorrect folding of the 2D data to 3D can lead to the loss of tri-linearity or misalignments. A

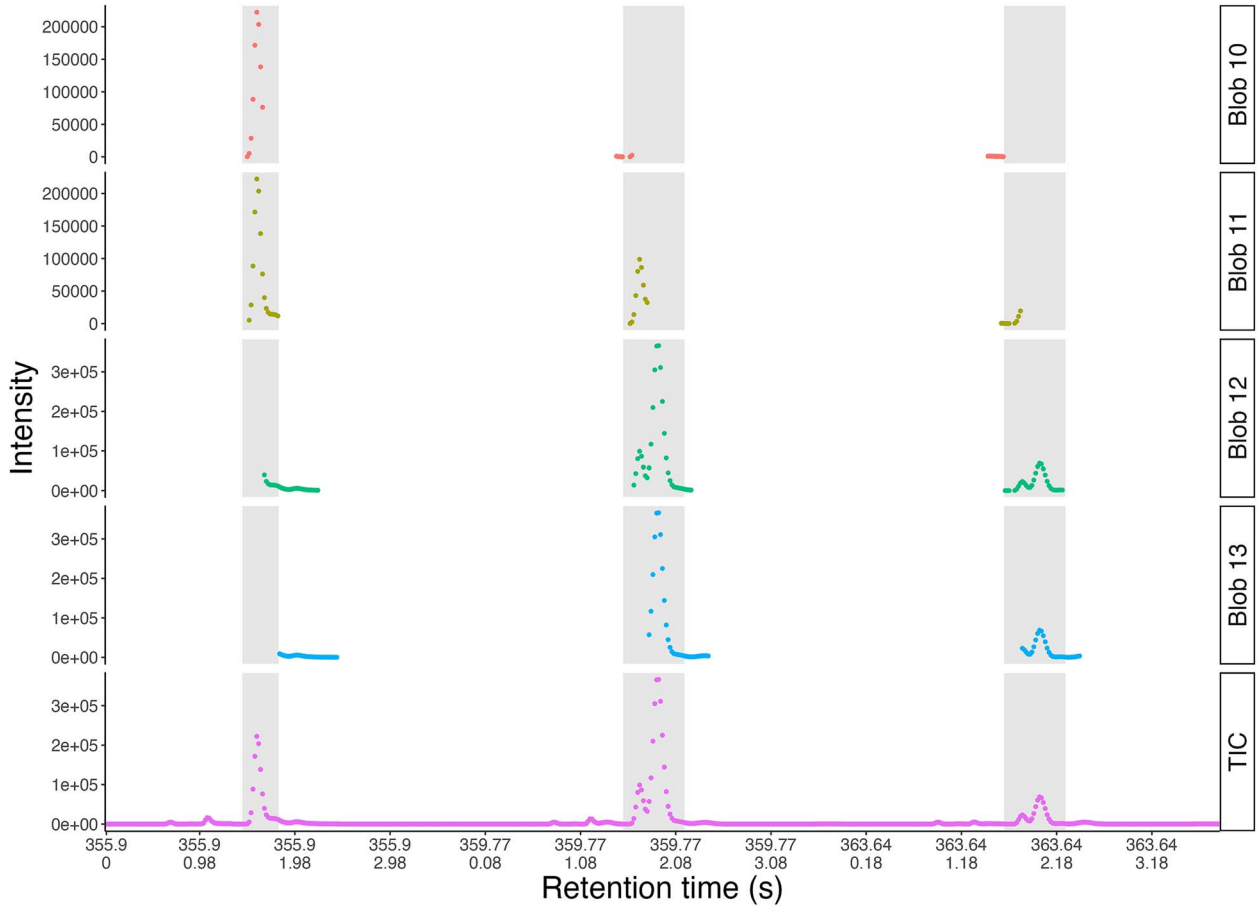

Figure 2. Data unfolding to evaluate the peak characteristics (Gaussian form + signal-to-noise) of the detected blobs inside one window. Gray areas show the chromatographic peaks detected in the sample's total ion chromatogram (TIC) from all samples of the standards training dataset. TIC in purple and other colors represent specific blobs.

visual check of the data is convenient before moving to the next module, to check if some big misalignments are present.

ROI selection

Module II uses the raw GC × GC–MS data matrix to delimitate the regions suitable for deconvolution. The training dataset is segmented into rolling windows, encompassing all samples while preserving the 4D structure. As a result, the prominent points of the matrix are identified by the inverse watershed algorithm within these windows to precisely delineate the ROI and segment it into individual peaks. However, the correction of false positives (noise peaks or artifacts) is necessarily based on peak characteristics. Subsequently, once the ROI is established, peak characteristics are assessed through data unfolding. Only peaks exhibiting adequate signal-to-noise ratios and conforming to a Gaussian form are retained (Fig. 2 and Supplementary Fig. S2). For instance, out of 17 initially detected blobs within the window, 4 blobs fulfilling the requirements are preserved (Fig. 2), while 13 blobs with a signal-to-noise ratio lower than 100 are not preserved (Supplementary Fig. S2). Lowering the signal-to-noise threshold to find small peaks, resulted in a notable increase in false positives, which increases considerably the computational processing time.

Moreover, matrix dimension plays a role in GcDUO. Despite that the PARAFAC algorithm can process the entire dataset, optimizing the matrix dimensions ensures an effective preprocessing, improving both computational efficiency and data representativeness. To achieve this, the watershed algorithm is set to include the minimum features possible within each ROI, reducing unnecessary complexity.

An ROI selection based on coordinates, similar to pixels, has the advantage of preserving the raw data [1]. It also has some disadvantages, like detecting spurious peaks generated by the detector noise and retention time misalignments. Thus, false positives can be corrected using some peak characteristics, such as the comparison to a Gaussian form. For example, in Blobs 15 or 9 (Supplementary Fig. S2) the entire shape of the peak was not properly selected, but it could be over the signal-to-noise threshold leading to false positives.

Blind PARAFAC

Module III applies the PARAFAC algorithm to the ROI regions selected from each sample individually. A key challenge in using PARAFAC for untargeted analysis is determining the correct number of components, a crucial parameter to obtain accurate results. To address this, GcDUO implements an iterative approach to component selection, increasing the number of components in each iteration and evaluating the results until no additional independent components are identified (Supplementary Fig. S3). However, even with the correct number of components, not all the information extracted necessarily corresponds to chemical peaks, as noise can be modeled as a component (Supplementary Fig. S4), particularly in cases of low-abundance peaks (Supplementary Fig. S5). To mitigate this, a further requisite is imposed in the blind PARAFAC:

all the extracted components should show a Gaussian form, a similar criterion to the one used during ROI selection.

To minimize noise modeling, only a fraction of the mass-to-charge channels is used during the component selection process. Specifically, the variation within each window is evaluated, and only the top 5% of m/z channels are selected. We observed that the mass fragments contributing to the chemical peak represent a small part of all fragments collected and that reducing the matrix size accelerates the iterative process (Supplementary Fig. S6). This approach prevents modeling a great part of the baseline noise, as it is characterized to have a low variation across the chromatogram. Once the number of components is obtained, the algorithm is applied to all mass fragments to ensure the complete deconvolution of the chemical peak spectra.

Annotation

Module IV ("Annotation") is performed with MS Search similarity [29] using an in-house library in MSP format. To increase the confidence level of identification, GcDUO incorporates the option of using retention index (RI) values. The training dataset contained 36 compounds (three where not available in our in-house library). Therefore, from the 33 compounds in our library, 16 where correctly annotated when no RI was used. When RI was used, the correct annotation of compounds increased to 22, representing a 37.5% increase in correct annotations compared to non-RI annotations (Supplementary Table S2).

Compound identification in metabolomics poses a considerable challenge. Relying on reference libraries or pure compounds often limits the identification to level 2 or putative annotation, rather than achieving full confirmation. In our case, two of the standard compounds reported were not present in our in-house library, resulting in alternative annotations. Despite advancements, comprehensive identification of all metabolites in samples remains incomplete, as matching relies on cosine similarity or the dot product between two vectors [29]. Therefore, additional information, such as retention times, co-elution patterns, and the context of the analysis, is often crucial for accurate identifications. Many public software tools use cosine similarity or dot product for annotation, including NIST MS Search [30], MS-DIAL [31], MassBank [32], GNPS [33], MetFrag [34], SIRIUS [34], or eRah [25]. In this context, the quality of the spectra being compared, and the reference library are critical factors. The successful identification of coeluting compounds, 1,8-cineole, and 4-methoxybenzenyl alcohol, demonstrated the high quality of the spectra obtained using GcDUO.

Constrained PARAFAC2

Module V focuses on refining the results obtained from previous modules, specifically accurate quantification of the peak. While PARAFAC yields optimal results with the appropriate number of components, iterative processes in module III can introduce spurious peaks. Additionally, results improve significantly when the ROIs are minimized to encompass only a single chromatographic peak.

This module utilizes all the information obtained to re-evaluate the raw data, similar to a targeted approach where the expected compound and the retention time are known, but with a few modifications compared to module III. The data now includes all samples in the dataset, and PARAFAC2 is applied. This approach allows for the detection of peaks with low signal-to-noise ratios (Supplementary Table S3) that were overlooked in module III. However, using multiple samples can potentially compromise tri-linearity due to misalignments. To address this, PARAFAC2 is employed, as it tolerates greater variation between samples while preserving the core data structure. When PARAFAC, which does not allow for variation between samples, is used the resulting chromatograms are perfectly aligned (see Fig. 3); however, some ghost peaks may appear due to the imposed shape (see example 4-methoxybenzenyl alcohol Supplementary Table S3). This peak alignment and impositions can sometimes lead to over-quantifications, which is why the PARAFAC2-constrained algorithm is preferred only to perform the quantification of peaks. Supplementary Fig. S7 provides a detailed comparison of the models, displaying the loadings, scores, and metrics for PARAFAC and PARAFAC2 from Fig. 3, highlighting the improved performance of PARAFAC2.

The performance of the model varies depending on characteristics of the peaks. For high-abundance peaks, such as Lilial (Supplementary Table S3), the constrained PARAFAC2 model tends to be more conservative in its quantifications compared to the raw area, whereas PARAFAC yields results more consistent with ChromaTOF. For small peaks (Supplementary Table S3), the constrained PARAFAC model successfully detects peaks at all concentrations, where other models fail, though this can occasionally lead to over-quantification. In case of overlapping peaks (Supplementary Table S3), the constraint model allows peak quantification in all samples, though it may result in slightly higher values than expected. When the models are unconstrained, both PARAFAC and PARAFAC2 preserve the original data's characteristics, yielding results closer to the raw data. Nevertheless, as overlapping peaks became more frequent, we selected PARAFAC2 for the quantification.

Validation

To assess the performance of the GcDUO software, two GC × GC-TOFMS datasets of two real metabolomics studies were used. To illustrate the complexity of the datasets analyzed during validation, we provide representative chromatograms from the training and validation datasets (Supplementary Fig. S8), alongside with a representative chromatogram from the training dataset. These datasets emphasize the software's ability to accurately process and analyze highly complex chemical matrices.

The annotation of compounds was tested using the publicly available fruitybeer dataset [28], as each flavor contains compounds unique to it. A total of 85% of the 27 peaks reported by the authors were successfully annotated with GcDUO, with the expected compound present in the matched lists (Supplementary Table S4). Only one compound was missing from our in-house library. Additionally, three peaks were detected (peak picking) but none of the compounds in the in-house library exhibited a cosine similarity higher than 0.7, therefore GcDUO did not return a result. However, the spectra obtained exhibited characteristic fragments of the expected compounds. The non-detected peaks correspond to small chromatographic signals with too few characteristic fragments to deconvolute. Positive annotations could be improved by incorporating retention index data. No further information was revealed by the authors.

To validate the information obtained using GcDUO, we analyzed an in-house set of five dilutions (0.1–1 ppm) of a breath mix solution containing 12 compounds (Table S1). We evaluated the annotation of expected compounds, the retention time, and the peak area comparing the results to ChromaTOF. GcDUO was applied to the raw datasets, generating a final peak list of 341 peaks. The system's performance was visually assessed to confirm the accuracy of the identified peaks. All the compounds were detected and positively annotated using the retention index

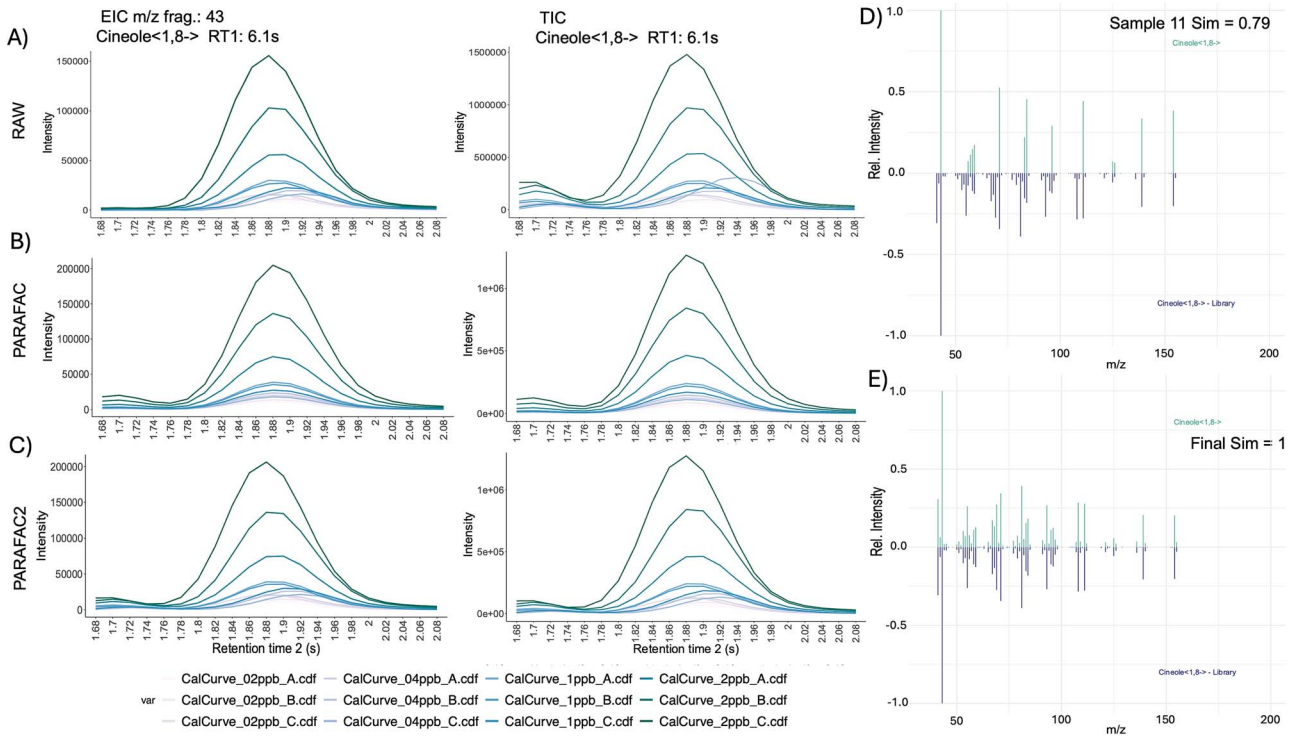

Figure 3. Results for 1,8-cineole peak from the training dataset, each color represents a sample. A) EIC/TIC chromatogram for raw data and the results of GcDUO. B) EIC/TIC chromatogram for blind PARAFAC results of GcDUO. C) EIC/TIC chromatogram for PARAFAC2 results of GcDUO (constrained PARAFAC). EIC in all cases is for m/z 43. D) Spectra obtained with module III (blind PARAFAC) for sample 11 compared to library spectra and E) the final spectra obtained with module IV (constrained PARAFAC) compared to library spectra.

information (Table 1). However, decane was only detected when the signal-to-noise parameter was reduced from 3 to 1. Retention times were consistent with those obtained using the vendors' software. Additionally, the peak areas for the 1 ppm sample showed a Pearson's correlation of 0.904 between the two software packages. The dilutions demonstrated strong linearity, with a mean $r^2 > 0.950$ (Supplementary Table S5). These results demonstrate the suitability of GcDUO for analyzing raw GC × GC–MS data.

Limitations

One of the major limitations of GC × GC–MS data analysis is the computational intensity required to handle large and complex datasets. This can result in longer processing times and may necessitate the use of high-performance computing environments to achieve optimal results. Additionally, GcDUO is implemented in R, an environment that is not optimized for handling large memory loads or parallel processing in multiple cores, which may further challenge its efficiency. Consequently, users need some level of expertise to effectively run the code and manage these computational demands. Despite efforts to minimize memory usage and processing time, relevant information remains accessible to the user if desired. For instance, features detected in only one sample are not further evaluated in module V, but their data is preserved in the results of module III.

Chemometrics tools, such as PARAFAC, also have inherent limitations. One of the primary issues is the tri-linearity assumption, which presumes that the data exhibits tri-linearity. However, this assumption is not always true in real datasets, potentially leading to inaccurate or misleading results. Therefore, it is crucial to visually inspect the raw chromatograms before performing peak picking and deconvolution to ensure correct folding of data. For datasets lacking trilinearity, models that do not assume trilinearity such as MCR, can be applied and yield comparable results (Supplementary Table S6) [35, 36]. Another characteristic is PARAFAC's sensitivity to noise, which can result in the detection of non-reproducible components or artifacts. In such cases, PARAFAC2 may offer advantages, as it is more robust in handling datasets that are noisy or contain unwanted variations [13]. Regarding model evaluation, new approaches with higher sensitivity are emerging, such as shape-sensitive congruence, which will be interesting to assess for component detection [37].

As with most software tools in metabolomics, careful validation and interpretation of the results are essential. Researchers must critically evaluate data, particularly when dealing with potential artifacts or a small number of components, to ensure the accuracy and reliability of their findings. Ultimately, the final interpretation of the data depends on the expertise of the researcher conducting the analysis.

Conclusion

GC × GC–MS is a powerful technique for analyzing complex mixtures, but its data complexity poses significant challenges for interpretation. GcDUO addresses these challenges by providing a comprehensive suite of tools for processing, deconvolution, and analyzing raw GC × GC–MS data. The software, implemented in R, operates through several key modules: data import, ROI selection, deconvolution, peak annotation, and data integration and visualization. It accepts non-vendor-specific, standardized CDF files and rearranges the data into 4D tensor structures, preserving the GC × GC–MS data structure in every dataset. GcDUO leverages advanced chemometric techniques, including PARAFAC and its variant PARAFAC2 to handle multi-dimensional data. The

Table 1. Breath dataset areas comparison between GcDUO and ChromaTOF for the 1 ppm standards mixture

	GcDUO		ChromaTOF	
Compound	**RT1, RT2 (s)**	**Area 1 ppm**	**RT1, RT2 (s)**	**Area 1 ppm**
2-Hexanone	259.98, 1.670	1.11E+07	259.99, 1.660	1.16E+07
Ethylbenzene	374.97, 1.820	1.95E+07	374.98, 1.807	2.44E+07
Cyclohexanone	417.47, 2.465	6.29E+06	417.48, 2.449	8.49E+06
6-Methyl-5-octen-2-one	579.96, 2.070	4.69E+06	579.97, 2.055	4.52E+06
Ethanone, 2-chloro-2,2-difluoro-1-phenyl-	727.45, 0.430	5.90E+06	727.46, 0.419	1.11E+07
Decane	604.80, 1.640	5.09E+05	604.96, 1.392	1.90E+05
Undecane	789.94, 1.430	1.29E+07	789.96, 1.423	9.34E+06
2-Undecene, (Z)-	797.44, 1.960	2.45E+06	797.46, 1.947	2.05E+06
1-Tetradecene	1279.91, 1.575	6.65E+06	1279.93, 1.561	5.17E+06
5,9-Undecadien-2-one, 6,10-dimethyl-, (E)-	1332.41, 2.205	3.27E+06	1332.42, 2.187	1.98E+06
5,9-Undecadien-2-one, 6,10-dimethyl-, (Z)-	1359.91, 2.230	3.71E+06	1359.92, 2.218	3.29E+06
3-Heptafluorobutyroxytetradecane	1429.90, 1.595	8.25E+06	1429.92, 1.580	5.21E+06

software includes features to optimize the analysis, such as noise reduction, signal enhancement, and peak alignment across multiple samples.

The effectiveness of GcDUO has been validated through the analysis of 2 datasets, identifying correctly in both datasets more than 89% of the present peaks. Comparison with ChromaTOF, the software considered the gold standard, showed a correlation of 0.909 between peak area measurements. This demonstrates GcDUO's capability, and it can be considered a robust tool for GC × GC–MS data analysis, offering an open-source alternative to black-box commercial software packages, that is accessible, transparent, and customizable, and therefore, potentially very useful for researchers in metabolomics and related fields. Overall, GcDUO provides a powerful and flexible platform for the comprehensive analysis of GC × GC–MS data, facilitating the extraction of meaningful insights from complex chemical datasets.

Key Points

- Comprehensive 2D GC coupled with MS (GC × GC–MS) generates complex data that complicates processing and interpretation.
- Chemometric approaches like Parallel Factor Analysis (PARAFAC) effectively extract information from multidimensional datasets but require data tri-linearity.
- GcDUO is an open-source software developed in R for processing GC × GC–MS data, integrating PARAFAC and PARAFAC2 for improved and automatic analysis.
- GcDUO has been validated against gold-standard software, showing a high correlation ($r^2 > 0.9$) in peak area measurements.
- GcDUO offers powerful capabilities making it ideal for researchers in metabolomics and related fields.

Acknowledgements

We thank Prof. H.T., Mr. Y.M., and Mr. Y.T. for their invaluable help in testing the GcDUO software. We also thank LECO Corporation for their collaboration on the instrumental side.

Supplementary data

Supplementary data is available at *Briefings in Bioinformatics* online.

Funding

This project received funding from the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No (798038). Grants PID2021-126543OB-C22 and RTI2018–098577-B-C21 funded by MICIU/AEI/10.13039/501100011033 and by ERDF/EU. MLL is thankful for her graduate fellowship from the URV PMF-PIPF program (ref. 2019PMF-PIPF-37) and Boehringer Ingelheim Fonds Travel Grant 2022. The Secretaria d'Universitats i Recerca del Departament d'Economia i Coneixement (2021 SGR 00818 and 2021 SGR 00842), CERCA program/Generalitat de Catalunya. AR is supported by the SRA-STEMA post-doctoral fellowship from Liège University.

Conflict of interest: None declared.

Data availability

The data underlying this article are available in Zenodo at 10.5281/zenodo.13947810 and can be accessed with identifier 13 947 810.

References

1. Trinklein TJ, Cain CN, Ochoa GS. *et al.* Recent advances in GC×GC and chemometrics to address emerging challenges in nontargeted analysis. *Anal Chem* 2023;**95**:264–86. https://doi.org/10.1021/acs.analchem.2c04235
2. Franchina FA, Purcaro G, Burklund A. *et al.* Evaluation of different adsorbent materials for the untargeted and targeted bacterial VOC analysis using GC×GC-MS. *Anal Chim Acta* 2019;**1066**:146–53. https://doi.org/10.1016/j.aca.2019.03.027
3. Morimoto J, Rosso MC, Kfoury N. *et al.* Untargeted/targeted 2D gas chromatography/mass spectrometry detection of the total volatile tea metabolome. *Molecules* 2019;**24**. https://doi.org/10.3390/molecules24203757
4. Guo L, Yu H, Li Y. *et al.* Tensor methods in data analysis of chromatography/mass spectroscopy-based plant metabolomics. *Plant Methods* 2023;**19**:1–13. 10.1186/s13007-023-01105-y
5. Berrier KL, Prebihalo SE, RE. *Synovec*. In: Snow NH. (ed), Separation Science and Technology (New York), Elsevier Inc., 2020, 229–68. https://doi.org/10.1016/B978-0-12-813745-1.00007-6
6. Stefanuto PH, Smolinska A, Focant JF. Advanced chemometric and data handling tools for GC×GC-TOF-MS: application of chemometrics and related advanced data handling

in chemical separations. *TrAC - Trends in Analytical Chemistry* 2021;**139**:116251. https://doi.org/10.1016/j.trac.2021.116251
7. Hopke PK, Spiegelman CH, K-S. Park. In: Balakrishnan N, Colton T, Everitt B, Piegorsch WW, Ruggeri F, Teugels JL. (eds), *Wiley StatsRef: Statistics Reference Online*. (New Jersey), John Wiley & Sons, Ltd, 2014.
8. Bro R. PARAFAC. Tutorial and applications. *Chemom Intel Lab Syst* 1997;**38**:149–71. https://doi.org/10.1016/S0169-7439(97)00032-4
9. Pinkerton DK, Parsons BA, Anderson TJ. *et al.* Trilinearity deviation ratio: a new metric for chemometric analysis of comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry data. *Anal Chim Acta* 2015;**871**:66–76. https://doi.org/10.1016/j.aca.2015.02.040
10. Prebihalo SE, Berrier KL, Freye CE. *et al.* Multidimensional gas chromatography: Advances in instrumentation, chemometrics, and applications. *Anal Chem* 2018;**90**:505–32. https://doi.org/10.1021/acs.analchem.7b04226
11. Johnsen LG, Skou PB, Khakimov B. *et al.* Gas chromatography - mass spectrometry data processing made easy. *J Chromatogr A* 2017;**1503**:57–64. https://doi.org/10.1016/j.chroma.2017.04.052
12. Mathema VB, Duangkumpha K, Wanichthanarak K. *et al.* CRISP: a deep learning architecture for GC × GC-TOFMS contour ROI identification, simulation and analysis in imaging metabolomics. *Brief Bioinform* 2022;**23**:1–17.
13. Amigo JM, Skov T, Bro R. *et al.* Solving GC-MS problems with PARAFAC2. *TrAC - Trends in Analytical Chemistry* 2008;**27**:714–25. https://doi.org/10.1016/j.trac.2008.05.011
14. Kronik OM, Liang X, Nielsen NJ. *et al.* Obtaining clean and informative mass spectra from complex chromatographic and high-resolution all-ions-fragmentation data by nonnegative parallel factor analysis 2. *J Chromatogr A* 2022;**1682**:463501. https://doi.org/10.1016/j.chroma.2022.463501
15. Wilde MJ, Zhao B, Cordell RL. *et al.* Automating and extending comprehensive two-dimensional gas chromatography data processing by interfacing open-source and commercial software. *Anal Chem* 2020;**92**:13953–60. https://doi.org/10.1021/acs.analchem.0c02844
16. Pollo BJ, Teixeira CA, Belinato JR. *et al. Trends Anal Chem* 2021;**134**:116111. https://doi.org/10.1016/j.trac.2020.116111
17. Decodon, (2008).
18. Gamble S, Chami P. Dental surgical emphysema following bridge sectioning. *Br Dent J* 2024;**237**:859–60. https://doi.org/10.1038/s41415-024-8150-9
19. Quiroz-Moreno C, Furlan MF, Belinato JR. *et al.* RGCxGC toolbox: an R-package for data processing in comprehensive two-dimensional gas chromatography-mass spectrometry. *Microchem J* 2020;**156**:104830. https://doi.org/10.1016/j.microc.2020.104830
20. Misra BB. Advances in high resolution GC-MS technology: a focus on the application of GC-Orbitrap-MS in metabolomics and exposomics for FAIR practices. *Anal Methods* 2021;**13**:2265–82. https://doi.org/10.1039/D1AY00173F
21. Beucher S. Watershed, hierarchical segmentation and waterfall algorithm. *Mathematical Morphology and Its Applications to Image Processing* 1994;69–76. https://doi.org/10.1007/978-94-011-1040-2_10
22. Samanipour S, Dimitriou-Christidis P, Gros J. *et al.* Analyte quantification with comprehensive two-dimensional gas chromatography: assessment of methods for baseline correction, peak delineation, and matrix effect elimination for real samples. *J Chromatogr A* 2015;**1375**:123–39. https://doi.org/10.1016/j.chroma.2014.11.049
23. Prebihalo SE, Pinkerton DK, Synovec RE. Impact of comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry experimental design on data trilinearity and parallel factor analysis deconvolution. *J Chromatogr A* 2019;**1605**:460368. https://doi.org/10.1016/j.chroma.2019.460368
24. Lorenzo-Seva U, ten Berge. Tucker's congruence coefficient as a meaningful index of factor similarity. *Methodology* 2006;**2**:57–64. https://doi.org/10.1027/1614-2241.2.2.57
25. Domingo-Almenara X, Brezmes J, Vinaixa M. *et al.* eRah: a computational tool integrating spectral deconvolution and alignment with quantification and identification of metabolites in GC/MS-based metabolomics. *Anal Chem* 2016;**88**:9821–9. https://doi.org/10.1021/acs.analchem.6b02927
26. Wan KX, Vidavsky I, Gross ML. Comparing similar spectra: from similarity index to spectral contrast angle. *J Am Soc Mass Spectrom* 2002;**13**:85–8. https://doi.org/10.1016/S1044-0305(01)00327-0
27. Weggler BA, Dubois LM, Gawlitta N. *et al.* A unique data analysis framework and open source benchmark data set for the analysis of comprehensive two-dimensional gas chromatography software. *J Chromatogr A* 2021;**1635**:461721. https://doi.org/10.1016/j.chroma.2020.461721
28. Franchina FA, Zanella D, Lazzari E. *et al.* Investigating aroma diversity combining purge-and-trap, comprehensive two-dimensional gas chromatography, and mass spectrometry. *J Sep Sci* 2020;**43**:1790–9. https://doi.org/10.1002/jssc.201900902
29. Stein SE, Scott DR. Optimization and testing of mass spectral library search algorithms for compound identification. *J Am Soc Mass Spectrom* 1994;**5**:859–66. https://doi.org/10.1016/1044-0305(94)87009-8
30. National Institute of Standards and Technology (2023) NIST Search Program (Software version 3.0). Available at https://chemdata.nist.gov/dokuwiki/doku.php?id=chemdata:nistlibs#nist_search_software.
31. Tsugawa H, Cajka T, Kind T. *et al.* MS-DIAL: Data-independent MS/MS deconvolution for comprehensive metabolome analysis. *Nat Methods* 2015;**12**:523–6. https://doi.org/10.1038/nmeth.3393
32. Horai H, Arita M, Kanaya S. *et al.* MassBank: a public repository for sharing mass spectral data for life sciences. *J Mass Spectrom* 2010;**45**:703–14. https://doi.org/10.1002/jms.1777
33. Ruttkies C, Schymanski EL, Wolf S. *et al.* MetFrag relaunched: incorporating strategies beyond in silico fragmentation. *J Chem* 2016;**8**:1–16. https://doi.org/10.1186/s13321-016-0115-9
34. Dührkop K, Fleischauer M, Ludwig M. *et al.* SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nat Methods* 2019;**16**:299–302. https://doi.org/10.1038/s41592-019-0344-8
35. Li M, Zhao Z, Zhang Y. *et al.* Chemometrics combined with comprehensive two-dimensional gas chromatography-mass spectrometry for the identification of baijiu vintage. *Food Chem* 2024;**444**:138690. https://doi.org/10.1016/j.foodchem.2024.138690
36. Pourasil RSM, Cristale J, Lacorte S. *et al.* Non-targeted gas chromatography orbitrap mass spectrometry qualitative and quantitative analysis of semi-volatile organic compounds in indoor dust using the regions of interest multivariate curve resolution chemometrics procedure. *J Chromatogr A* 2022;**1668**. https://doi.org/10.1016/j.chroma.2022.462907
37. Wünsch UJ, Bro R, Stedmon CA, Wenig P, Murphy KR. Emerging patterns in the global distribution of dissolved organic matter fluorescence. *Anal Methods* 2019;**11**:888–93. https://doi.org/10.1039/C8AY02422G

Briefings in Bioinformatics, 2025, **26(2)**, bbaf080
https://doi.org/10.1093/bib/bbaf080
Problem Solving Protocol