Trends in Analytical Chemistry 190 (2025) 118286

Contents lists available at ScienceDirect

Trends in Analytical Chemistry

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

On the road to automation: a comparative review on chemometric strategies for GC-MS data analysis

Agnieszka Smolinska [a], Samuele Pellacani [b], Michal Skawinski [a], Caterina Durante [b,*]

[a] Department of Pharmacology & Toxicology, NUTRIM Institute of Nutrition and Translational Research in Metabolism, Maastricht University, Maastricht, the Netherlands
[b] Department of Chemical and Geological Sciences, University of Modena and Reggio Emilia, Modena, Italy

ARTICLE INFO

Keywords:
MCR-ALS
PARADISe
GC-MS
Chemometrics
Resolution
Compounds extraction
Automatic data pre-processing

ABSTRACT

It is widely proved in the literature the pivotal role of GC-MS technique in obtaining information across diverse research contexts. Given the great importance of GC-MS analysis and the undeniable role of chemometric techniques in signal processing, this review provides perspectives on the latest automatic procedures for performing peak detection, resolution, baseline, and time-shift correction. Herein, insights into the principles behind the most common chemometric methods in scientific literature, PARAFAC2 (PARAllel FACtor analysis 2) and MCR-ALS (Multivariate Curve Resolution - Alternating Least Squares), are provided, along with their respective strengths and limitations. In particular, regarding PARAFAC2, the PARADISe (PARAFAC2-based Deconvolution and Identification System) has been further described, while for MCR-ALS, the latest automated procedures have been explored. The aim of this review is to shed light on the evolving field of automatic GC-MS data analysis and to facilitate the advancement of this field.

1. Introduction

Gas chromatography-mass spectrometry (GC-MS) has emerged as a cornerstone analytical tool, providing unique capabilities for characterizing semi-volatile and volatile compounds within complex samples [1]. Its versatility makes GC-MS a powerful tool used in various research fields, including metabolomics [2], food analysis [3], environmental analysis [3], medical analysis [4], and beyond [5,6]. This ability comes from its high sensitivity, precision, and accuracy [6], making it indispensable for researchers seeking to find the compositions of several sample matrices. However, despite its widespread adoption and efficacy, GC-MS data analysis can be challenging, due to the inherent complexities and voluminous data generated [7] as well as coeluting peaks, noise, and baseline fluctuation. These aspects further complicate the extraction of meaningful chemical information from GC-MS datasets [6] and compromise accurate quantification and structural elucidation of compounds. Furthermore, when performing direct comparison of multiple samples to identify discriminatory compounds, time shifts between samples present a significant bottleneck [8]. Mass spectra are undeniably a valuable tool for aligning peaks, particularly in targeted approaches. However, for untargeted GC-MS analysis, it is common to encounter compounds with similar chemical structures. This resemblance can generate similar mass spectra, potentially leading to inaccuracies in peak alignment, especially when these compounds elute closely. Consequently, the reliability of qualitative and quantitative information derived from GC-MS analyses directly influences the accuracy and validity of downstream analyses, statistical analyses, biomarker screening, pathway analyses, and interpretations.

Some of the most used methods in the processing and analysis of GC-MS in the last ten years are reported in Table 1 together with some key features (if open-access, type of supported format, manual preprocessing, presence of statistical tools and graphical interface). All software described here is freely available or downloadable under a free academic license. Reader is referred to the original article for more in-depth information.

Among the main software used for GC-MS data processing and analysis, XCMS and Mzmine2 are the most widely cited algorithms. Both rely on the continuous wavelet transform (CWT)-based *centWave* algorithm [25] for peaks detection. However, they require manual set up of several parameters (*m/z* tolerance, peak width, S/N threshold etc.), to construct extracted ion chromatograms (EICs) and select peaks. Furthermore, two key aspects of the *centWave* algorithm, namely

This article is part of a special issue entitled: Chemometrics published in Trends in Analytical Chemistry.

* Corresponding author.
E-mail address: caterina.durante@unimore.it (C. Durante).

https://doi.org/10.1016/j.trac.2025.118286
Received 3 May 2024; Received in revised form 13 March 2025; Accepted 2 May 2025
Available online 9 May 2025

Table 1
Main software used in the processing and analyzing of GC-MS data.

Tools	Main publication, year	Overall citations[a]	Citations from 2014[a]	Key features
XCMS	XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching and, identification, 2006 [9]	3691	2957	Open-source package, based on the R programming language. Freely available under an open-source license https://xcmsonline.scripps.edu/landing_page.php?pgcontent=mainPage **Data formats:**[a] mzXML,[a] mzML,[a].NetCDF, etc. **Functions:** peak detection, integration, alignment, peak identification **Statistical tools:** yes **Preprocessing setting:** manual **Interface:** online version of XCMS has been released in 2012 with a dedicated GUI
Mzmine2	Mzmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data, 2010 [10]	2726	2540	Open source software https://github.com/mzmine/mzmine2/releases **Data formats:**[a] mzXML,[a] mzML,[a].NetCDF,[a] cdf, Thermo Fisher raw format **Functions:** peak detection, integration, alignment, peak identification **Statistical tools:** yes **Preprocessing setting:** manual **Interface:** graphical
MS-DIAL	MS-DIAL: Data-independent MS/MS deconvolution for comprehensive metabolome analysis, 2015 [11]	1759	1759	Open source software https://systemsomicslab.github.io/compms/msdial/main.html **Data formats:** netCDF (AIA), mzML, ([a] mzML, .[a]NetCDF, etc.) **Functions:** peak detection, integration, alignment, peak identification **Statistical tools:** yes **Preprocessing setting:** manual **Interface:** graphical
MCR-ALS	A graphical user-friendly interface for MCR-ALS: a new tool for multivariate curve resolution in MATLAB, 2005 [12]	995	719	Developed in Matlab, available free of charge https://mcrals.wordpress.com/download/ Baseline correction and alignments are not necessary. Constraints could be necessary to help and improve the resolution results (unimodality, no-negativity etc.) **Functions:** peak detection and integration, no peak identification **Interface:** graphical
MetAlign	Metalign: Interface- driven, versatile metabolomics tool for hyphenated full-scan mass spectrometry data preprocessing, 2009 [13]	603	417	Open-access software https://zenodo.org/records/7273832 **Data formats:** .[a]mzXML,[a] mzData, .[a]cdf, MassLynx, Xcalibur, Agilent **Functions:** peak detection, integration, alignment, peak identification **Preprocessing setting:** automatic/manual **Interface:** graphical
AMDIS	An integrated method for spectrum extraction and compound identification from gas chromatography-mass spectrometry data, 1999 [14]	696	387	Freely available stand-alone software http://www.amdis.net/ **Data formats:** *.netCDF, *.raw; *.dat, etc. **Functions:** peak detection and identification, no peak alignment **Preprocessing setting:** automatic/manual **Interface:** graphical
TagFinder	TagFinder for the quantitative analysis of gas chromatography – mass spectrometry (GC-MS)-based metabolite profiling experiments, 2008 [15]	460	305	TagFinder is made freely available for academic use from http://www-en. Mpimp-golm.mpg.de/03-research/researchGroups/01-dept1/Root_ Metabolism/smp/TagFinder/index.html TagFinder facilitates the analysis of all fragment ions, which are observed in GC-(EI-TOF)-MS profiling experiments **Data fromats:** The recommended data import uses the commonly accepted chromatography interchange format[a] netCDF **Interface:** graphical
Metabolite Detector	MetaboliteDetector: comprehensive analysis tool for targeted and nontargeted GC/MS based metabolome analysis, 2009 [16]	329	274	MetaboliteDetector is available for 32 bit and 64 bit Linux operating systems https://md.tu-bs.de/node/10 **Data formats:** raw GC-MS data can be imported in[a] netCDF or FastFlight2 format **Functions:** peak detection, integration and identification **Statistical tools:** yes **Preprocessing setting:** automatic/manual **Interface:** graphical
MET-IDEA	MET-IDEA: Data Extraction Tool for Mass Spectrometry- Based Metabolomics, 2006 [17]	232	110	It is not an open-source software https://sumnerlab.missouri.edu/download/- MET-IDEA imports raw data in net.cdf format It requires an input list composed of a series of ion/retention time pairs, which guides the selected ion extraction process **Functions:** peak detection, integration and identification **Statistical tools:** yes **Preprocessing setting:** automatic/manual **Interface:** graphical

(*continued on next page*)

signal-to-noise ratio estimation and CWT ridgeline detection [26], could allow to obtain a considerable number of false-positive peaks and significant differences in the peak lists of the two packages, making automation more complicated.

As previously highlighted, preprocessing is another important aspect to be considered in GC-MS datasets analysis, before peak detection, since it allows to eliminate artefacts and reduce irrelevant variations, such as noise, retention time shift, skewing in mass spectra. In particular, the latter can occur due to several factors, such as detector response, ionization efficiency variations, or chromatographic anomalies. Skewed peaks can complicate deconvolution, leading to inaccurate quantification and identification [27]. Proper instrument calibration, consistent maintenance, and software correction algorithms are essential to mitigate skewing before data analysis. Although methods such as AMDIS, MS-DIAL, MetAlign have achieved successful applications handling skewing through preprocessing, they still require manual intervention for parameter setup (Table 1). On the road to automation, chemometrics and deep learning surely represent valid support to tackle preprocessing, coelution problems, and to address other challenges associated with extracting meaningful information from GC-MS datasets [6]. In particular, MCR-ALS and PARAFAC2 over the last decade, have been considered, a valid choice to effectively resolve GC-MS signals (n° of citations: more than 700 and 500, respectively), due to their advanced capabilities in handling complex GC-MS datasets, particularly in resolving preprocessing challenges, baseline correction, and co-elution issues. Both algorithms can automatically separate baseline contributions from informative signals by assigning them to distinct components. This ensures cleaner and more accurate signal interpretation. In case of co-eluted peak resolution, each technique handles it differently. In particular, PARAFAC2 does not require strict peak alignment, allowing variability in retention times and focusing on covariance rather than exact chromatographic profiles. This makes it ideal for datasets with retention time shifts. MCR-ALS uses a least-squares approach to estimate pure components from mixed signals, enhancing its ability to deconvolute overlapping peaks effectively. Both methods allow simultaneous resolution of multiple chromatograms, significantly reducing data analysis time and improving efficiency, which is crucial in untargeted analysis. Unlike targeted methods, these approaches exploit the full information in the signal, allowing the detection of unknown compounds, which is particularly advantageous in fields such as metabolomics and food authentication where unknown adulterants might be present. Their output can be directly used for advanced data analysis techniques like SIMCA [28] and PLS-DA [29], facilitating sample classification and discrimination. Both methods support automated workflows by extracting peak areas, contributing to both qualitative and quantitative analyses seamlessly.

Furthermore, these chemometrics tools are compatible with various MS analyzers (Quadrupole, Time-of-Flight (TOF), Orbitrap and Fourier Transform Ion Cyclotron Resonance) but their effectiveness can vary based on the type. The best fit for PARAFAC2 and MCR-ALS is TOF and Orbitrap analyzers due to their high resolution and fast acquisition rates, which are crucial for detailed and accurate deconvolution. For effective application of deconvolution tools like PARAFAC2 and MCR-ALS, a minimum acquisition rate of 10–20 Hz (scans per second) is recommended. This rate provides sufficient temporal resolution to accurately define peak shapes and separate closely eluting compounds. In addition, to acquisition rate, good signal to noise is necessary.

From an applicative point of view, despite the effectiveness of the

Table 1 (*continued*)

Tools	Main publication, year	Overall citations[a]	Citations from 2014[a]	Key features
PARADISe	Gas chromatography– mass spectrometry data processing made easy, 2017 [18]	218	218	Open-source software; trial versions or academic licenses at https://ucphchemometrics.com/paradise/; **Data formats:**[a] cdf,[a] mat **Functions:** peak detection, by means PARAFAC2 algorithm [19] and integration, alignment and peak identification **Preprocessing setting:** automatic/manual **Interface:** graphical
eRah	eRah: a computational tool integrating spectral deconvolution and alignment with quantification and identification of metabolites in GC-MS based metabolomics, 2016 [20]	92	92	Free computational tool written in the open language R http://CRAN.R-project.org/package=erah **Data formats:**[a] mzXML **Functions:** peak detection, integration, alignment and peak identification **Preprocessing setting:** manual
PyMS	PyMS: a Python tool kit for processing of gas chromatography-mass spectrometry (GC-MS) data. Application and comparative study of selected tools, 2012 [21]	63	56	It is a Python software package particularly suitable for scripting of customized processing pipelines and for data processing in batch mode https://code.google.com/archive/p/pyms/ **Data formats:** ANDI- MS/Net.CDF and JCAMP-DX **Functions:** peak detection, alignment and peak identification **Preprocessing setting:** manual
ADAP-GC 4.0	ADAP-GC 4.0: application of clustering-assisted multivariate curve resolution to spectral, 2019 [22]	46	46	It is implemented in Java and it is open-source. http://www.du-lab.org It has been incorporated into Mzmine 2. **Functions:** peak detection, integration, alignment and peak identification **Preprocessing setting:** manual
MARS	Feature extraction from resolution perspective for gas chromatography-mass spectrometry datasets, 2016 [23]	16	16	MATLAB toolbox open-source https://github.com/zmzhang/MARS **Functions:** peak detection and integration **Preprocessing setting:** automatic/manual It is time-consuming for large-scale GC-MS datasets and sensitive to the component number estimation.
MARS 2	MARS 2: a computational tool to resolve and extract features from large- scale GC-MS datasets, 2019 [24]	9	9	It is implemented in Python programming language and open-sourced at https://github.com/mapancsu/MARS2 **Data formats:**[a] cdf **Functions:** peak detection, integration and identification **Preprocessing setting:** automatic/manual **Interface:** graphical

[a] by Scopus, last accessed April 30, 2024.

MCR-ALS and PARAFAC2 model to resolve signals, they require a great expertise in model development. Indeed, there are some aspects that are worth noting. The first one is that MCR-ALS can be applied only on bi-dimensional datasets (usually, mass fragmentation in one dimension and elution profile in the second one). When more than one sample is analyzed with GC-MS, the resulting data is a three-way array (cube with the recorded mass fragment as first mode, elution profiles in the second mode and samples in the third mode). Therefore, before applying MCR-ALS, the obtained array must be unfolded into a two-dimensional matrix without losing any information. This process significantly increases the matrix size, and consequently, extends the analysis time. Additionally, when applying MCR-ALS, the solution is not unique, due to the so-called rotation ambiguity, as explained in Session 3 and supplementary file. In short, this means that the MCR-ALS can find the 'true' solution in a set of infinite combinations that describe the data equally well. This problem can be addressed by imposing constraints on the model (unimodality, non-negativity etc.), thus reducing the number of feasible solutions.

PARAFAC2 model can handle three-way array, time shift and baseline drifts in the data. Therefore, GC-MS dataset does not need to be preprocessed before the development of the model. Furthermore, uniqueness can be obtained when the right number of components is included. In this context, visual inspection of different parameters, such as the explained variance by the model, the entity of residuals, the model's fit (%), the number of iterations, etc. (as explained in Session 2), is one of the main ways to assess the right number of components. From a practical point of view, choosing the right solutions requires the best compromise between these criteria as well as a great deal of experience on the part of the operator.

The introduction of a semi-automated tool, namely PARADISe (PARAFAC2- based Deconvolution and Identification System) [30], with user-friendly graphical interface, represents great breakthrough for the use of PARAFAC2.

Although, both PARADISe and MCR-ALS are considered semi-automated tools, the user must make critical decisions regarding the selection of some parameters, namely the number of components (for both MCR-ALS and PARADISe) and the intervals used for peak detection (for PARADISe). The number of components determines how many underlying chemical signals will be resolved by the model, while the intervals define the chromatographic region where the peaks will be identified and PARAFAC2 models calculated. These interventions are of extreme importance, as they can introduce bias, variability into the results as well as potentially generate false components. Bias is of utmost importance in quantitative analysis, and it is a meaning of systematic errors (given by the sum of the differences between known and calculated analyte concentration divided by the total number of the considered samples). Indeed, in both MCR-ALS and PARADISe, choosing too few components can miss important details, while too many might introduce noise. Moreover, incorrect selection of a PARADISe interval could lead to a systematic error in the calculation of the concentration of an analyte, resulting in constant shift of the regression function represented in the graph of predicted values as a function of known values.

Based on these considerations, it is important to have a good understanding of the principles behind the development of the MCR-ALS and PARAFAC2 models and the choice of parameter configuration, for obtaining robust models and consistent results. Therefore, both the methods deserve a detailed discussion in the next section of this review.

Furthermore, this review examines the latest advances in the development of emerging automated approaches based on the publications over the past decade. We aim to illuminate the changing landscape of GC-MS data analysis and pave the way for future breakthroughs in this critical field. For more detailed information on a certain strategy, readers are encouraged to refer to the original article.

2. PARAFAC2-based approach

2.1. Theoretical background

The simplest way to handle higher-dimensional arrays is to unfold the original data by combining some parts of it and subsequently applying conventional methods, like standard principal component analysis (PCA) [36]. This, however, has several limitations, therefore, methods used to analyze multi-way data, particularly three-way data, are based on a tensor multidimensional array decomposition into a set of simpler component arrays (known as factors) [31].

To perform tensor decomposition several techniques have been developed, including CANDECOMP/PARAFAC decomposition, proposed independently at the same time by Carroll and Chang [32] as CANDECOMP (canonical decomposition) and Harshman [33] as PARAFAC (PARAllel FACtor analysis). This technique is based on the parallel proportional profile principle, which factorizes a tensor into a sum of rank-one trilinear component arrays, although it can be extended to higher-order tensors as well. Mathematically, given a three-way chromatography data array $\mathbf{X}$ ($I \times J \times K$), I represents the number of scans or measurements taken over the elution time in a chromatographic system, J corresponds to the number of mass-to-charge ratios observed, and K is the number of samples, the PARAFAC will be modelled as:

$$\boldsymbol{X_k} = \boldsymbol{F D_k A^T} + \boldsymbol{R_k} \quad \text{Eq. 1}$$

here $\mathbf{F}$ ($I \times R$) is a scores matrix (for the row units), $\mathbf{A}$ ($\boldsymbol{J} \times \boldsymbol{R}$) is a loadings matrix (for the column units), $\mathbf{D}_k$ ($R \times R$) is a diagonal matrix with the loadings for the kth partition of $\mathbf{X}$, and $\mathbf{R}_k$ ($I \times J$) contains residuals matrix. R denotes the desired rank of the decomposition [34]. The above model is one generalization of bilinear PCA, capable of handling three-way data. Unlike standard PCA, where each component is represented by one scores vector and one loadings vector, PARAFAC involves each component array consisting of one score vector and two loadings vectors, which are often treated equally [35]. Under certain mild conditions and given the right rank, there is a unique solution for the PARAFAC model resolving the components to the individual underlying latent variables without any additional constraints [34]. However, the above model assumes the identity of the observations of individual components in all samples and fixed dimensionality across all modes (e.g., samples, retention times, and mass spectra in GC-MS data). Therefore, it applies the same score matrix, which is often too restrictive [36]. For GC-MS data, PARAFAC model assumes consistent retention times across all samples, consistent peak shapes and noise. However, those assumptions are often violated by shifts in retention time, variability in peak shapes and shifts in noise and baseline. In the presence of those, the observation units are not directly comparable.

To address those limitations, PARAFAC2 was developed, making it more suitable for GC-MS data [18]. Unlike PARAFAC, which imposes an unimodality constraint, PARAFAC2 allows variability in one mode [36]. This flexibility enables retention times to vary across samples, accommodating variations in dimensionality and alignment issues. It also handles co-eluted peaks by allowing changes in the loadings along one mode. Additionally, PARAFAC2 can easily distinguish signal (i.e., peaks) from noise due to its unique modeling approach that accommodates variability in one mode while preserving the core trilinear structure of the data. PARAFAC2 allows for variation in the retention time dimension across samples. In GC-MS data, where retention times can naturally shift, this flexibility ensures that the signal (true peaks) can align across samples even when the chromatographic conditions vary. Noise, on the other hand, typically does not exhibit consistent shifts and therefore fails to align in the same structured way. The model captures the consistent chemical profile of the analytes (e.g., mass spectral patterns or chromatographic peaks) through shared loadings across samples. True signals contribute consistently to the model's components, while random noise lacks this consistency and is not captured as a significant

component. PARAFAC2 models structured and random variations by capturing components with consistent profiles in two modes (e.g., mass spectrum and intensity) while allowing variability in the third mode (e.g., retention time). Noise, being random and unstructured, does not fit well into the model's structured framework and is instead relegated to the residuals. During the decomposition process, the model essentially 'filters out' components that do not contribute meaningfully to the overall explained variance. Since noise typically presents as low-variance, high-frequency signals, it does not form strong components and is thus not retained as a significant factor in the model. PARAFAC2 is able to differentiate between the structured, consistent patterns of true signals and the inconsistent nature of noise. By modeling only the structured components and treating the rest as noise, it effectively enhances the signal-to-noise ratio in complex GC-MS datasets.

Each score matrix for a sample in PARAFAC2 is replaced with:

$$\boldsymbol{F_k} = \boldsymbol{P_k F}, \quad \boldsymbol{k} = \mathbf{1}, \ldots, \boldsymbol{K} \qquad \text{Eq. 2}$$

where $\mathbf{P}_k$ ($I \times R$) is a column-wise orthonormal matrix and $\mathbf{F}$ ($\boldsymbol{R} \times \boldsymbol{R}$) is a matrix with a common part of observations from different samples. This way, the elution profiles for a sample are unrelated to the profiles of other samples, which allows for retention time shifts. Importantly, the PARAFAC2 model holds the structural uniqueness property, similar to PARAFAC [34]. PARAFAC2 as a stand-along algorithm is implemented in PLS toolbox (Eigenvector Research Inc., Manson, WA, USA) and N-way Toolbox in MATLAB [37]. However, the ability of PARAFAC2 to handle retention time shift and co-elution has found its application in the processing of untargeted GC-MS data, utilized in a semi-automatic tool, PARADISe [30]. PARADISe applies three-way tensor decomposition on chromatographic data considering all samples to find co-eluting peak areas and identify underlying compounds. In short, PARADISe performs the following steps: 1) visualization of raw chromatographic data, 2) user-supported selection of retention time intervals, 3) PARAFAC2 deconvolution and identification of peaks with a deep learning algorithm, 4) user-supported selection and extraction of peaks, 5) (optional) identification of analytes using NIST mass spectra library, 6) generation of output spreadsheet with peak areas table. Each selected interval with peaks is decomposed with multiple PARAFAC2 models, with a different number of components (in default from 1 to 8). The resulting models are not nested, meaning that each model represents a different orientation of components within the subspace, and the choice of the final model depends on the user [34]. PARADISe tries to resolve chemical peaks and separate them from the baseline and co-eluting compounds that are represented by different PARAFAC2 factors. In addition, overlapping peaks can be identified even with a small signal-to-noise ratio. However, PARADISe identifies compounds that have similar mass spectra as one peak, regardless of the number of components in the model [18].

In the updated version of PARADISe, a deep learning algorithm, utilizing a convolutional neural network, automatically evaluates chromatographic data elution profiles, discerning peak characteristics from background noise and accommodating overlaps and retention time shifts [38]. PARADISe Community Edition is available to download for free from the webpage at kromath.com as stand-alone program in Windows (*.exe) with MATLAB runtime or as a MATLAB app. Alternatively, a Professional Edition can be purchased, which additionally allows for automatic model fitting and evaluation, selection of a deep learning tool, reading retention indices from library and an access to the Kromath help desk. General recommended hardware requirements are between 1 and 512 CPU cores and 8 GB of RAM, however, PARADISe can be quite computationally expensive [30]. It is also possible to use parallel computing to deconvolute peaks, where each PARAFAC2 model is built simultaneously by a different core to speed up the computations.

2.2. Practical application of PARADISe

PARADISe effectively addresses the co-elution issues of interferents with similar mass spectra ratios, making it a robust method for untargeted analyses. Additional visualizations, semi-automation, deep learning algorithm, and extensive tutorial make PARADISe a valuable tool for data processing in various fields, as shown in Table 2.

Before running any analysis, data should be exported into.CDF (Computable Document Format) or MATLAB-formatted files (*.mat), structured in the same way as included with the installation template file. Alternatively, it is possible to import previous sessions or previously selected intervals.

Although PARADISe can handle time shift, it is beneficial to align the data beforehand using retention time adjustments (such as build-in alignment with initial coshifting [53], and correlation optimized warping [54]). Indeed, when analyzing several samples affected by significant time shifts, defining interval limits can become intricate. PARAFAC2 deconvolution requires the user to manually select retention time intervals for peak detection, ideally isolating a single peak with its start and tail. Focusing on one peak per interval minimizes the risk of misinterpretations and inaccurate deconvolution. However, accurately defining these intervals can be difficult, especially with overlapping or ambiguous peaks. In such cases, extending or dividing intervals, as shown in Fig. 1, is advised. Consequently, PARAFAC2 models will be fitted for each chosen interval.

Adjusting the PARAFAC2 parameters within the PARADISe offers some flexibility for the analysis process, however, the default values should be appropriate for most datasets [30]. By default, PARADISe applies flexible coupling PARAFAC2 with non-negativity and fits models up to 8 components, which can be further increased/decreased depending on the complexity of the data and intervals. Moreover, the maximal interactions per model (5000) or the least squares error convergence threshold (10) can be further adjusted for complex coeluted peaks. For example, if complex regions with multiple overlapping peaks are expected, the number of components could be gradually increased, e.g. by 2, and the resulting models inspected, as too complex models might start to explain the noise. Alternatively, for simple and known peaks, that number could be decreased. Furthermore, increasing the number of iterations can help in achieving a better convergence, but will demand more computational time.

The manual user-assisted assessment of fitted models and the optimization process to choose the 'right' number of components and peaks for each interval is the most troublesome part of the whole analysis in PARADISe. This step will determine the number of identified compounds and has a substantial influence on the final results. Furthermore, the graphical examination of resulting models, even with the help of a deep learning algorithm in the classification of peaks, leaves a risk of incorrect or poor results. Therefore, PARADISe, notwithstanding its efficacy in peak deconvolution and identification, persists as a semi-automated approach, wherein user interventions can significantly impact the outcomes. The manual procedure, after fitting the models for each interval is demonstrated in Fig. 2 and further explained below.

After fitting the models for each selected interval, the user needs to determine which model out of all models with a different number of components is the best, based on the provided metrics and visual inspection of the plots (Fig. 2 **steps I-III**). The **Fit** (from 0 to 100 %) indicates how well the model explains the information within the interval, and the **core consistency** (from 0 to 100 %) represents the adequacy of the model. The higher the Fit and core consistency, the better. However, in practice, the Fit reaches a plateau when increasing the number of components and the core consistency reaches zero. Therefore, the balance between the number of compounds and the Fit should be achieved in order to avoid overfitting, while maintaining a reasonably well fit (of at least 95 %) **(step I)**. Another metric is the number of peaks and types of components in the interval, identified automatically by a deep learning algorithm [38]. The types of peaks classified by PARADISe are

Table 2
Application of PARADISe across different fields. Q, Quadrupole mass detector. MS, Mass Selective detector. PTR, Proton Transfer Reaction.

	Sample	Aim of the study	Analyte	Platform	PARADISe usage	Reference
Plant	Dwarf birch	Evaluation of VOC emission responses to climate change	VOCs	GC-MS	Analysis of chromatograms	[39]
	Wetland plants	Identification of metabolic traits of selected plant species from *Ranunculus* genus	Plant leaves metabolites	GC-MS	Processing of chromatograms along with the NIST MS Search	[40]
	Cucumber	Investigation of cucumber metabolic responses to fullerenol treatments	Plant metabolites	GC-MS	Analysis of chromatograms along with the NIST MS Search	[41]
	Bulbs	Identification of anti- inflammatory alkaloids from H. elegans	Alkaloids	GC-Q-MS	Untargeted deconvolution and identification of peaks along with the NIST MS Search resulting in a peak area table	[42]
Animal	Mouse	Novel sampling strategy for animal volatolome	VOCs	GC-Q-MS	Processing of data, deconvolution of peaks along with the NIST MS Search	[43]
	Fly	Profiling of fatty acids in long-legged fly	Fatty acids	GC-MS	Processing of data, deconvolution and identification of peaks along with the NIST MS Search	[44]
Food and nutrition	Minced bovine meat	Flavor characterization of animal hydrolysates	VOCs	GC-MS with Triple-Axis detector	Processing of data resulting in a peak area table	[45]
	Fermented milk	Investigation of the impact of different prebiotics on the biological metabolites	Milk metabolites	GC-Q-MS	Processing of data	[46]
	Beer	Investigation of the impact of brewing yeast on the flavor stability during storage	VOCs	GC-MS	Processing of data resulting in a peak area table	[47]
	Whisky	Profiling of whisky aroma and dilution with water	VOCs	GC-MS	Untargeted deconvolution of peaks along with NIST MS Search	[48]
	Sourdough bread	Characterization of flavor profiles in sourdough cultures	VOCs	GC-PTR-MS	Deconvolution of peaks along with the NIST MS Search resulting in a peak area table	[49]
Human	Urine	Examining the effect of chemotherapy on urinary volatile biomarkers for lung cancer	VOCs	GC-Q-MS	Processing of data, deconvolution and identification of peaks, along with the NIST MS Search	[50]
	Blood	Metabolomic exploration of the endothelium in pulmonary hypertension	Amino and non-amino organic acids	GC-Q-MS	Processing of data	[51]
	Vaginal swabs and urine	Detection of potential biomarkers for sexually transmitted infections	VOCs	GC-Q-MS	Processing of data, deconvolution and identification of peaks along with the NIST MS Search resulting in a peak area table	[52]

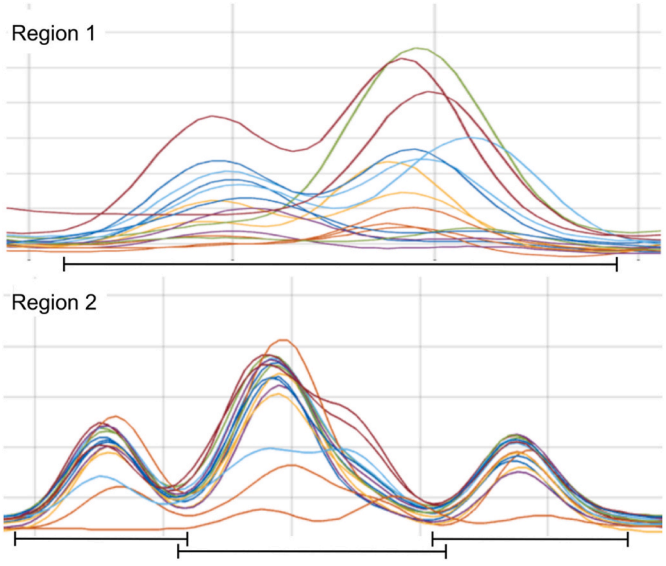

Fig. 1. Example of two regions in PARADISe software. The boundaries of the chosen intervals are marked with black lines. In the given example, the first region presents a challenge in determining the start and end points of the peak, making it difficult to precisely define its boundaries. As a result, the entire region containing this ambiguous peak could be selected as a single interval. In contrast, the second region clearly displays three distinct peaks. For accurate analysis, each of these peaks should be separately selected, resulting in three intervals for that region.

peak, baseline, offset, and unknown. Each model can be further assessed by visual inspection of TIC/model residuals (**steps II a-b**), retention time (min)/normalized profiles (**step IIc**), and MZ/Samples plots (**steps II d-e**). An example assessment of the plots from the **step II** is visualized in Fig. 3. TIC can be a reference for the number of peaks to be expected in the current interval. Co-eluted and overlapping peaks might be however not visible on this plot, therefore a precise analysis of the components of each model, based on the retention time and normalized profiles should be carried out. The retention time plot displays the elution profiles for each model component, colored by the component number or type of the peak classified from the deep learning algorithm (Fig. 3B1 and 3C1). Each component represents a different information. Moreover, for each model, this information will likely change. Larger peaks or baseline might cover other peaks, therefore playing with different visualizations (displaying and hiding some components) is advised to better understand what each component represents. A balance between the number of peaks in the interval, baseline, and additional components of unknown type should be achieved to avoid overfitting while extracting the relevant information. Alternatively, normalized elution profiles can be displayed and analyzed, where each component is normalized for all samples so that the Euclidean norm equals one. This visualization can help to detect Gaussian peaks but might scale up noise.

An important model evaluation is a plot of model residuals, which represents the unexplained part of the model (Fig. 3B2 and 3C2). Ideally, residuals should be 'flat' and not demonstrate any underlying structures. On the contrary, peak structures can imply the presence of additional, unexplained information and a choice of model with more components can be considered. Data alignment, peak structure on a TIC, and expected number and structure of peaks should be taken into consideration, as especially the right or left parts of the interval might overlap

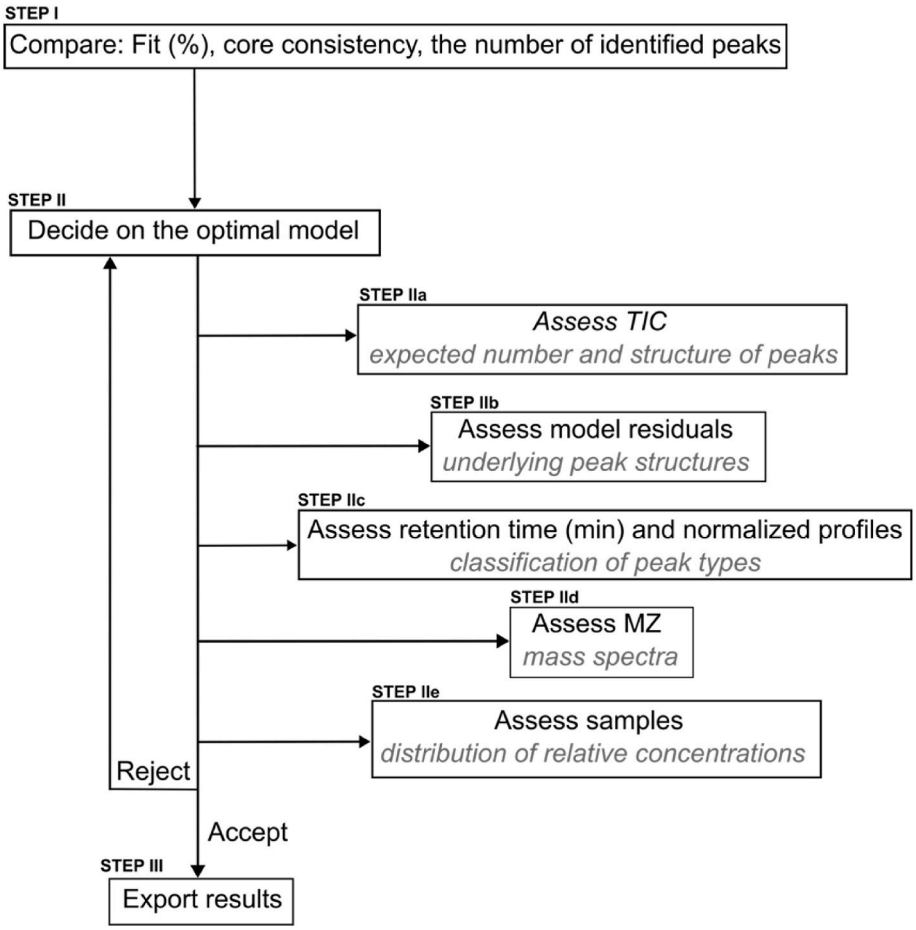

Fig. 2. Standard manual procedure after fitting the models, performed for each selected interval. Step I requires the user to compare model Fit (%), core consistency and the number of identified peaks for all created models with different numbers of components in order to choose one for further inspection. Step II consists of graphical inspection of (a) TIC, (b) model residuals, (c) retention time and normalized profiles, (d) mass spectra, and (e) distribution of relative concentrations across sample plots. Step III requires the user to select peaks and export them to the report.

with neighboring peaks that are of no interest for the selected interval and therefore could be ignored. In case of unsatisfactory results due to the structure interferences with neighboring regions or too long/short intervals, it is still possible to reselect the intervals and fit additional models.

The last panel of Fig. 3 (**B3 and C3**) shows mass spectra for each component of the model and the distribution of relative concentrations between samples for each component. Peaks identified through PARADISe should represent different chemical compounds if their mass spectra differ. On the opposite, similar mass spectra for peaks can imply model overfitting and too many selected components.

After the evaluation and a choice of the model, the user needs to select which peaks to export to the report (Fig. 2 **steps II-III**), that contains the estimated peak area table of compounds per sample, top NIST hits (if the NIST MS SEARCH was available) for each component from all intervals, estimated resolved mass spectra, and the details on intervals. The peak area table can be used for further statistical and chemometrics analysis beyond PARADISe.

A detailed walkthrough and more explanation of PARADISe can be found in the YouTube channel at http://www.youtube.com/@Kromath, and in the PARADISe tutorial [38]. Additionally, any issues encountered while using the tool are displayed in the PARADISe logbook.

2.3. Advanced topics in PARADISe and PARAFAC2

PARAFAC2 is an iterative optimization algorithm that ends if the maximum number of iterations is reached or the error is below the convergence threshold. With complex datasets, convergence issues can arise, including slow convergence and local minima is reached. PARADISe itself offers few opportunities for this problem other than increasing the number of iterations or changing the convergence threshold. However, applying the PARAFC2 algorithm with non-negativity constraints (default option in PARADISe) significantly decreases this problem [54]. The model fitting process might issue several warning about matrix singularity. This can happen when the data bears high levels of collinearity and correlation between factors, as in GC-MS data. This results from compounds coming from the same source, being involved in the same biochemical pathway, being used together, or being chemically similar with similar retention times and mass spectra.

PARAFAC2 models are fitted with different number of components (ranks), therefore, especially for the models with a higher number of components, the rank can be inappropriate, leading to overfitting. Therefore, the balance between model fit and the number of identified peaks and components, and peak structure should be achieved. Another important issue with PARAFAC2 is the inability to apply constraints to a specific component, which is not the case for MCR-ALS. In PARAFAC2, constraints such as non-negativity, or orthogonality are typically applied globally to all components, due to the variations in one model [34]. If different constraints are required for different components, the iterative optimization process becomes too complex. Non-negativity constrain, for example, is implemented in PARADISe to ensure that all elements of the factor matrices are non-negative, which corresponds to chromatographic quantities that cannot be negative. An optimization process is modified to ensure that any negative values are set to zero or adjusted to

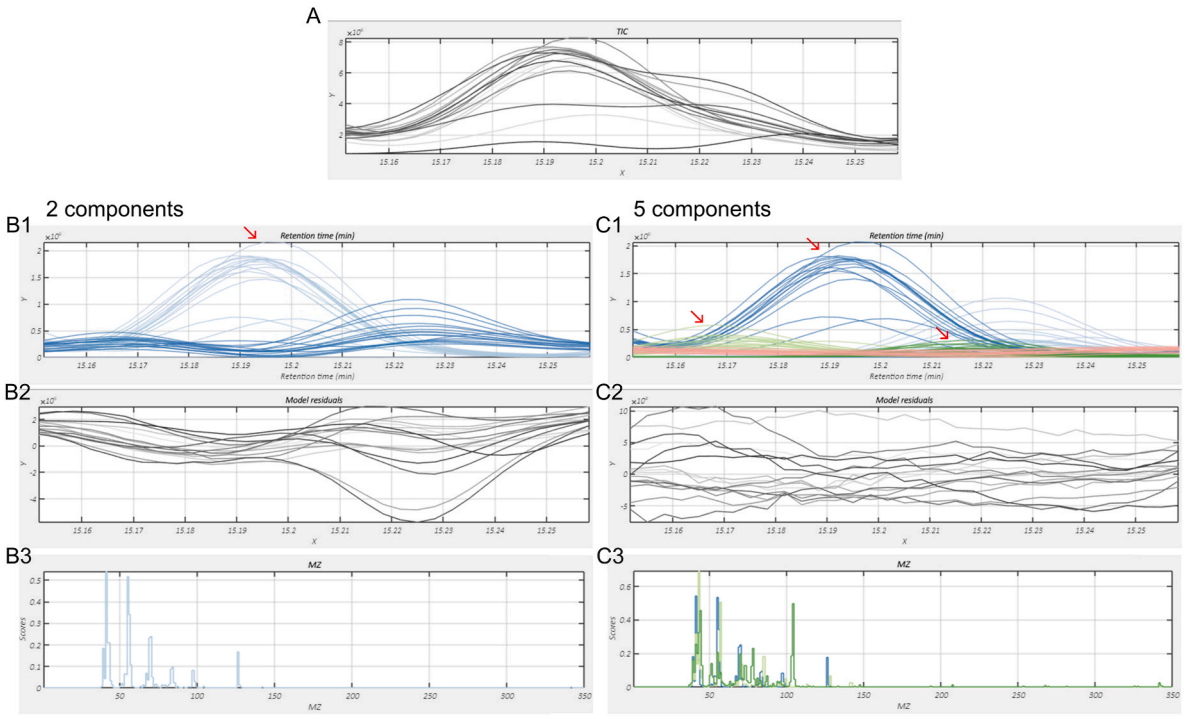

Fig. 3. Comparison of 2 components (99 % fit) and 5 components (~100 % fit) models. (A) TIC with 1 expected peak. PARADISe identified one peak for 2 components model (B1) and three peaks for 5 components model (C1). Based on the model residuals, the 2 components model exhibits an underlying peak structure (B2), opposite to the 5 components model, with flatter lined (C2). However, the 5 components model is overfitted, with similar mass spectra for peaks (C3) and incorrect classification of peaks/noise (C1). Moreover, it explains only one percent more of the variation in the data than the 2 components model. Therefore, the choice of 2 components model is better, despite the model residuals structure.

satisfy the constraint at any iteration. It should be noted, that zero values in some matrices can violate one of the uniqueness assumptions, however applying non-negativity can help reduce the rotational ambiguity by limiting the solution space to non-negative values, making the factors more stable and easier to interpret [34]. Despite allowing elution profiles to vary in shape across different samples, PARAFAC2 retains uniqueness properties, if the successful model is obtained (ensuring, similarly to PARAFAC, that the factors in the non-varying modes must be linearly independent). Consequently, PARAFAC2 can directly decompose mixture data into the individual contributions of the underlying analytes, including their concentrations, elution profiles, and mass spectra [55].

3. MCR-ALS-based approach

The main aim of using MCR-ALS [56,57] on GC-MS data is to resolve a mixed chemical data set into the pure component profiles using mathematical decomposition tools. All the performed GC-MS measurements can be organized in a data matrix, where one direction is related to the elution profile (e.g. chromatographic profile) and the other direction refers to mass spectra. MCR-ALS model can be written as reported in Eq. (3), where X is the experimental data matrix $\mathbf{X}$ ($I \times J$), namely GC-MS data, which is resolved into the product of a matrix of row profiles $\mathbf{S}^{\mathbf{T}}$ ($F \times J$) and a column matrix $\mathbf{C}$ ($I \times F$).

$$\mathbf{X}_{I\times J} = \mathbf{C}_{I\times F}\mathbf{S}_{F\times J}{}^{\mathrm{T}}\mathrm{E}_{I\times J} \quad \text{Eq. 3}$$

In the case of GC-MS data, matrix $\mathbf{C}$ expresses the abundance of the pure responses along the row direction of the data set (elution profiles, quantitative information), while matrix $\mathbf{S}^{\mathbf{T}}$ is related to the qualitative information of the components in the system analyzed (mass spectra). The scalar I and J are the total number of time and m/z points of the studied dataset, respectively, and F is the number of components to be resolved. Finally, $\mathbf{E}_{I\times J}$ is the matrix that represents variance that the bilinear model is unable to explain and should be small and random.

An alternating least-squares algorithm fits the $\mathbf{C}$ and $\mathbf{S}^{\mathbf{T}}$ matrices iteratively to the experimental data X in order to estimate the model parameters. Initial estimates of either the $\mathbf{C}$ or the $\mathbf{S}^{\mathbf{T}}$ matrix, with a predetermined number of components (F), are used to fit the model. Thus, starting with reasonable initial estimates is crucial. Concentration profiles or spectra can serve as first estimations in MCR-ALS. Best practice entails beginning with initial predictions and adhering to the constraints applied during optimization. Random initial values are ineffective if specific profiles are desired. Techniques such as the Orthogonal Projection Approach (OPA) [56] and Simple-to-use self--modeling analysis (SIMPLISMA) [56] are commonly utilized for initial estimation in GC-MS data analysis.

Despite the simplicity on the application of this approach, existence of (i) permutation, (ii) intensity and (iii) rotational ambiguities represents a limitation of MCR-ALS approach. The idea of ambiguity states that there is not a unique solution, and the original data set may be reproduced with the same fit quality using various combinations of concentration profile and spectrum sets [58]. As regard permutation ambiguity, it is worth to note that MCR components are not arranged in a particular order. As a result, they can be rearranged in different way in the concentration and spectra matrix while maintaining the proper relationship and producing the same outcomes. In the second case, namely intensity ambiguity, the original data set is as well reproduced by the product of profiles with the same shape but there is a difference in the relative scales between the concentration profile and spectrum. In the case of rotational ambiguities, differently shaped sets of concentration profiles and spectra can replicate the original data set with a same fit quality.

To reduce the extent of ambiguity, different constraints can be used [16,23]. Some restrictions include mathematical criteria or models (i.e. non-negativity, normalization, unimodality, local rank and hard modelling), while others are related to the assessment of the data set's

inherent features (i.e. closure and selectivity) [16,18,23].

Non-negativity is one of the most common constraints [6,59] in GC-MS data analysis; it can be applied to both chromatographic and mass spectrometric directions, since negative values in an elution profile or mass spectrum are not physically interpretable. Non-negativity (Fig. 4a), which compels positive values to form profiles, can be carried out by using softer algorithms, like non-negative least squares or fast non-negative least squares, or by substituting zeros for negative values. However, other constraints can be used in GC-MS data analysis, and they can be related to the peak shape, i.e. unimodality or to sequential elution pattern of compounds, i.e. selectively. As regards unimodality (Fig. 4b), it permits each profile to have a single maximum, while a selectively constrain (Fig. 4c) may fix elution profile of a component to zero, if it is known a priori that a component is not present in a given dataset.

Finally, normalization or closure constraints, which are discussed in more detail elsewhere [59], can help avoid problems associated with intensity ambiguities [59]. The normalization constraint is generally applied to the spectra and may constrain the Euclidian norm or the maximum of $\mathbf{s}_i^T$ (pure spectra profile) to be one. This is recommended to avoid scale instabilities during the evolution of the ALS optimization. A closure constraint [60] is usually implemented on the rows of matrix $\mathbf{C}_j$ (individual time/concentration profile) and it constrains the sum of the elements of each row to be equal to a known constant (e.g. for mass balance in concentrations).

In the case of quantitative studies, it is possible to exploit the correlation constraint. During the ALS optimization, calibration samples provide concentration values (x areas in the case of GC-MS data) which may correlate with known analyte concentrations (k). Unknown samples (u) are those with analyte concentrations that need to be determined. A local linear model is constructed between x and k values, expressed as:

$$k = xb + b_0 \quad \text{Eq. 4}$$

Once b and b_0 are estimated via linear regression, these parameters can be used to predict the analyte concentrations in the unknown samples (u) using the equation:

$$\underline{u}b + b_0 = p \quad \text{Eq. 5}$$

The complete profile, now updated with the new concentration estimates (both known (k) and predicted (p)), is then used to proceed to the next ALS optimization step.

As previously stated, GC-MS datasets suffer from baseline and retention time shift. Both of them can be overcome with the MCR-ALS approach [61]. Baseline/background contributions can be modelled by extra MCR-ALS components and the final correct number of components would contain only chemical information (i.e. chromatographic peaks) [62]. However, this situation is not always possible to achieve due to the complexity of the dataset, i.e. baseline could be not reproducible for all samples. In some cases, it is better to pretreat the raw two-way chromatographic data before applying MCR-ALS [63]. It is possible to remove the baseline drift and the spectral background using congruence analysis and least square fitting [64]. Retention time shift can be correctly modelled with MCR-ALS. In fact, if the spectra of a specific analyte among the different samples remain the same, it will be detected as a feature even if the chromatographic peaks are not aligned [61]. A practical overview of the MCR-ALS approach can be found in the supplementary file.

3.1. Practical application of MCR-ALS

The application of MCR-ALS to the analysis of GC-MS data has seen significant development over the past decade. For the sake of brevity, studies in the last 10 years on environmental, food, botanical, and essential oil samples are reported in Table 3.

3.2. Implementation of MCR-ALS in different environments and some advancements

One of the key factors in data analysis is selecting the appropriate computational environment. When it comes to analysing GC-MS data, numerous options are available. Beyond the commercial software mentioned in the introduction (Paragraph 1), researchers can utilize both free routines on commercial platforms, such as MATLAB (The MathWorks, MA, USA, 2024), and free routines on open-source platforms like RStudio, Python, and Java.

In the MATLAB environment, the MCR-ALS graphical user interface (GUI), developed by Tauler et al., in 2015, is widely used [57]. Additionally, a more advanced tool for preprocessing chromatograms, known

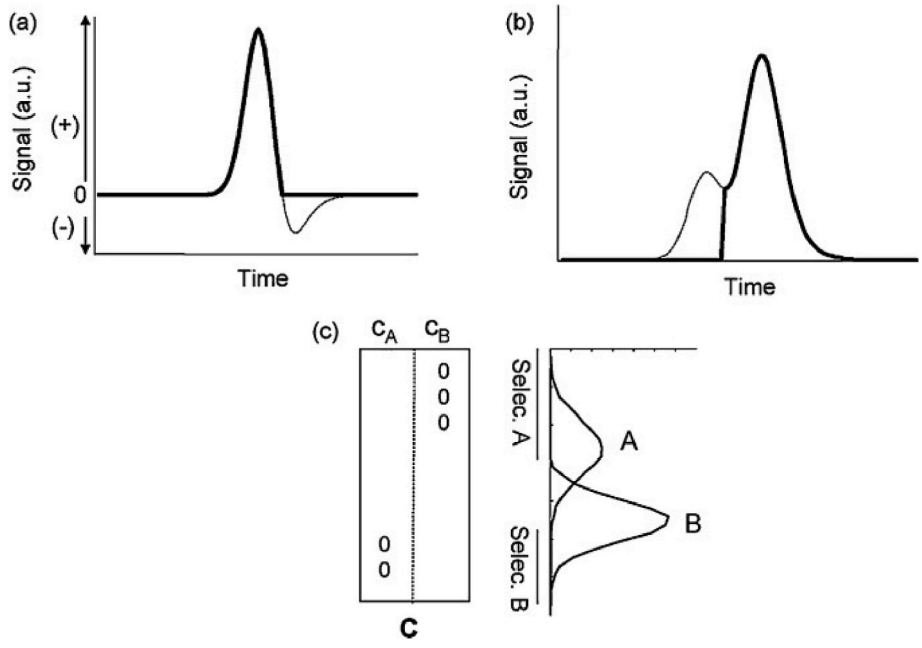

Fig. 4. Graphical representation of the most common constraints used during GC-MS data analysis. (a) Non-negativity, (b) unimodality e (c) selectivity. Reproduced with permission. Copyright Elsevier [6].

Table 3
Applications of MCR-ALS for the analysis of GC-MS data in the last ten years.

	Sample	Aim of the study	Analyte/technique	Platform	MCR-ALS implementation	Reference
Plant	Citrus	Investigation of the chemical variability of three varieties of the C.aurantium L. family	VOCs by static headspace sampling (SHS)	GC-Q-MS	COW-MCR-COW	[65]
	Saffron flowers	Identification of volatile metabolites	Volatile metabolites. Ultrasound-assisted solvent extraction (UASE) and dispersive liquid–liquid microextraction (DLLME)	GC-Q-MS	MCRC software for preprocessing and chemical rank determination	[66]
	Thymus	Comprehensive and rapid analysis of complex natural extracts	Solid-Phase Microextraction (SPME)	GC-Q-MS	MCRC software for preprocessing and chemical rank determination	[67]
Food	Avocado seeds and pulp	Fatty acids methyl esters (FAME) profiling and quantification	FAME. Esterification and direct injection in GC-MS	GC-Q-MS	Peak resolution	[68]
	Rice	Aroma characterization of rice samples	VOCs by SHS-GC-MS	GC-Q-MS	MCRC software for preprocessing and chemical rank determination	[69]
Animal	Daphnia magna	Metabolic analysis to study changes due to environmental stressors	Volatile metabolites. Extraction and derivatization of metabolites	GC-Q-MS	Peak resolution	[70]
Environment	Daphnia magna	Metabolic analysis to study changes due to salinity, temperature and hypoxia	Volatile metabolites. Extraction and derivatization of metabolites	GC-Q-MS	Peak resolution	[71]
	Water samples	Determination of polycyclic aromatic hydrocarbons	Polycyclic aromatic hydrocarbons (PAHs), Extraction and injection in GC-MS	GC-Q-MS	MCRC software for preprocessing and chemical rank determination	[72]
Essenti al oils	Aerosol samples	Determination and quantification of PAHs	Polycyclic aromatic hydrocarbons (PAHs), Extraction and injection in GC-MS	GC-Q-MS	Alternative decomposition-assisted (ATLD)- MCR-ALS	[73]
	Perfumes salmes	Determination of essential oils in complex perfume mix	Direct injection	GC-Q-MS	MCR-ALS on total chromatogram average mass spectra	[74]
Other	Myrtus communis L.	Characterization of essential oils from iranian Mirtus	Essential oils. Simultaneous distillation–extraction (SDE). Direct injection of the extracts	GC-Q-MS	MCRC software for preprocessing and chemical rank determination	[75]
	Ocimun Basilucum L.	Characterization of essential oils from Basilucum and assessing their antioxidant proprieties	Essential oils. Hydrodistillation and then direct injection	GC-Q-MS	Peak resolution	[76]
	Dracocephal um moldavica	Characterization of essential oils from Dracocephalum moldavica	Essential oils. Hydrodistillation and then direct injection	GC-Q-MS	MCRC software for preprocessing and chemical rank determination	[77]
	Elettaria Cardamomu m	Characterization of essential oils from Elettaria Cardamomum	Essential oils. Essential oils. Hydrodistillation and then direct injection	GC-Q-MS	MCRC software for preprocessing and chemical rank determination	[63]
	Herbal slimming pills	Characterization of volatile constituents of nine slimming pills to discover probable adulterations.	VOCs. Extraction and direct injection	GC-Q-MS	MCRC software for preprocessing and chemical rank determination	[78]

as Multivariate Curve Resolution of two-way Chromatographic data (MCRC) [79], is also available within MATLAB. MCRC offers several features: (a) a range of preprocessing techniques, (b) methods for determining chemical rank, (c) both iterative and non-iterative techniques for resolving chromatographic data, and (d) a GUI that provides various graphical outputs.

RMet [80], developed using RStudio version 1.1.383 and R core version 3.4.3, is another software designed for comprehensive untargeted metabolic data analysis. It encompasses all steps of the analysis process, including preprocessing, segmentation, data compression, multivariate curve resolution (MCR), metabolite identification, and classification. Specifically, RMet utilizes MCR-ALS as its resolution algorithm, effectively addressing challenges in GC-MS and GC × GC-MS analyses, such as elution time shifts, baseline interference, peak overlap, and changes in peak shape [3]. It then classifies metabolites using partial least squares-discriminant analysis (PLS-DA) [18,19] and highlights significant metabolites through variable importance in projection (VIP) [20] scores. Following this, metabolic pathway analysis should be conducted using pathway databases to identify affected pathways.

In Python, the MARS 2 (MS-Assisted Resolution of Signal) [24] tool can be employed. The MARS 2 workflow is structured into five key steps: (1) extracting mass spectrum and retention time (MSRT) pairs; (2) identifying peak regions (PR) of interest; (3) initially estimating the number of components and chromatographic peaks; (4) extracting pure chromatographic peaks using MCR-ALS; and (5) identifying compounds via a NIST library search.

The spectral deconvolution algorithm in ADAP-GC 4.0 [22], developed in Java environment, marks a significant advancement, adapting the MCR approach and departing from the traditional two-step deconvolution methods used in earlier versions. This algorithm incorporates ideas from both the traditional and modern approaches, such as dividing the retention time range into deconvolution windows, refining the apex retention time of EIC peaks, clustering EIC peaks to estimate component numbers, and applying unimodality constraints to model peaks. Moreover, ADAP-GC 4.0 addresses the issue of missing low-intensity peaks, a common problem in previous peak detection algorithms.

Many efforts have focused on automating MCR-ALS analysis of GC-MS dataset [5,81], including two strategies to overcome the problem of interval selection [82] and rotational ambiguity [83].

Zhang et al. [5] proposed a new automatic strategy for performing TIC peak detection and resolution: **autoGCMSDataAnal**. The authors developed a user-friendly MATLAB GUI which can be freely downloaded at http://software.tobaccodb.org/software/autogcmsdataanal [5]. A brief workflow of autoGCMSDataAnal is shown in Fig. 5. It can be performed as single sample analysis or batch analysis. In both cases, originally acquired GC-MS data files are used as input to automatically perform TIC peak detection, component resolution, time-shift correction, statistical analysis, and compound identification.

Single sample analysis involves two main components: component resolution and time-shift correction. Batch analysis includes component registration and peak filling, aimed at facilitating statistical analysis for metabolite screening. Peak detection is performed for TIC and EIC

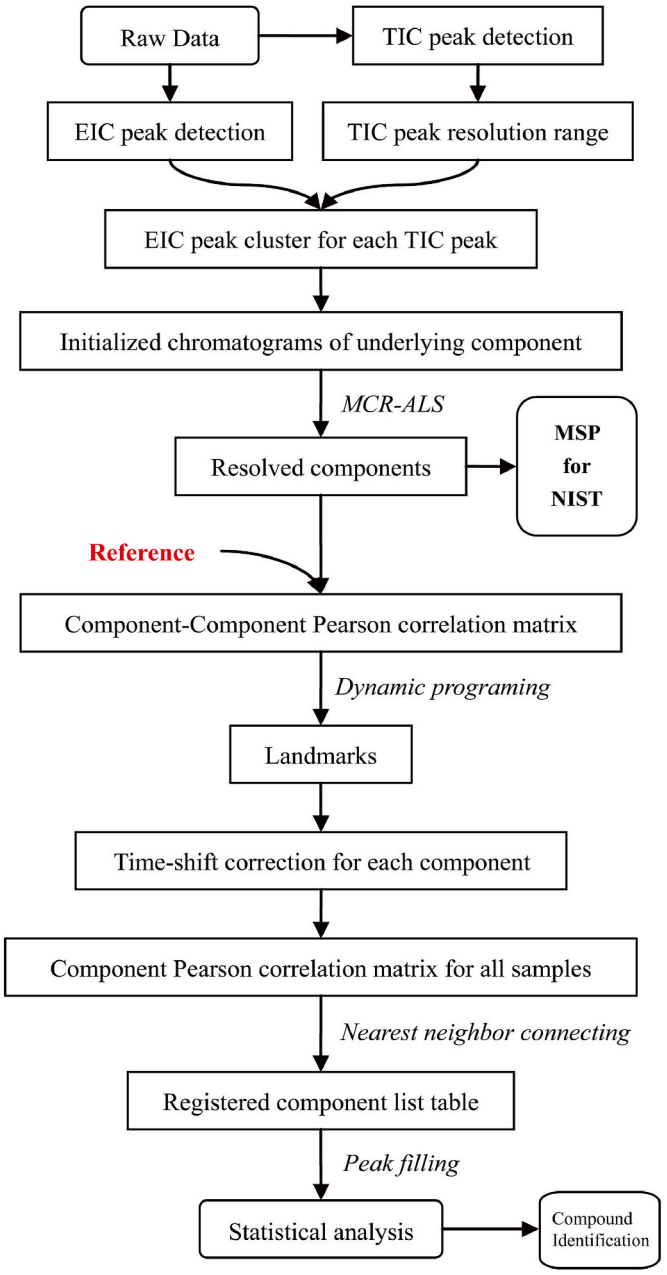

Fig. 5. Workflow of the developed autoGCMSDataAnal. Reproduced with permission. Copyright Elsevier 2019 [5].

profiles, independently. After having performed a Gaussian smoothing-based peak detection algorithm [84], the peak resolution range of each TIC peak is determined considering its respective starting and ending points. Furthermore, MCR-ALS is automatically performed on the selected interval for the peak resolution. To calculate initialized parameters, the authors proposed a modified hierarchical clustering method, where EIC profiles under the TIC peak are clustered based on their Pearson correlation coefficients of peak shapes and retention time difference, with threshold values of 0.95 and 0.02 min, respectively. The corrected retention time and resolved mass spectrum of each component are used for component registration. AutoGCMSDataAnal provides a file for each sample containing all registered components that can be directly imported into the NIST for searching for candidates. The performance of autoGCMSDataAnal was tested for the analysis of untargeted GC-MS metabolomics datasets coming from tobacco plants of three different Chinese provinces [5]. The results revealed that autoGCMSDataAnal outperformed MET-IDEA, MS-DIAL, and eRah, since it effectively differentiated samples from different growth stages across all three provinces. However, a critical aspect is selecting the intervals for dividing the chromatogram to apply MCR. Garrata-Laura et al. [70] proposed to split the chromatograms at sufficient retention times to cover entire peak clusters and to avoid peak halving. Furthermore, they recommend using a sufficient number of components to describe the background noise and possible impurities. On the other hand, Domingo-Almenara et al. [82] proposed to analyze the entire chromatogram by applying the moving window-based MCR-ALS. Successive windows must overlap in order to guarantee that a compound that may split on the window borders is completely covered by the following window. However, this causes compounds to be resolved in multiple windows, or obtaining partial eluting profiles. This results in numerous duplicates, which makes it difficult to choose the well-resolved -chromatographic profiles. In order to guarantee that each analyte has a single chromatographic profile and spectrum, the authors suggested a duplicity filter that minimizes the residual sum of squares (RSS). The scenario with the least RSS is the combination that best describes the data, and the chromatographic profiles, that are not included in this combination, are removed. The main advantage of this approach is the possibility of analyzing all the chromatograms in an automatic way. The main disadvantage is the computational time since the same data is analyzed twice due to the overlap. Applying a trilinearity constraint can resolve the rotational ambiguity in MCR-ALS models [85]. This approach generates distinct profiles for each component, providing a straightforward way to explain data variance [86]. However, this is applicable only if the data conform to the trilinear model for effective resolution. The application of a flexible trilinearity constraint in MCR-ALS was described by Zhang et al. [83].

In 2022, Fan et al. [81] proposed a novel pipeline named Deep-learning-assisted Multivariate Curve Resolution 2 (DeepResolution2) to solve the co-eluting problem for GC-MS targeted and untargeted data. DeepResolution2 is implemented in Python and is available as an open-source package at https://github.com/XiaqiongFan/DeepResolution2.

The flowchart of DeepResolution2 is depicted in Fig. 6.

Briefly, the GC-MS chromatogram is automatically divided into small intervals for subsequent analysis, leveraging a deep neural network based on U-Net. These segments are then classified into either pure or co-eluted peaks. Co-eluted peaks are further resolved into chromatographic peaks and spectra of pure components for subsequent analysis by a convolutional neural network known as k-CNN, which automatically determines the number of components (k) within a segment. For pure segments, the TIC of the sub-matrix is utilized to calculate the peak area, while the mass spectrum at its apex is employed for identification purposes. Finally, resolution methods such as FRR, MCR-ALS, and ITTFA can be adaptively chosen based on the overlap degree of the corresponding segment. This selection process is aided by the elution region predicted by the U-Net model for elution region detection (U- Net4R). The proposed method was evaluated to assess its effectiveness and performance using three real GC-MS datasets: 10 amino acid standard mixtures at five different concentrations, 37 fatty acid methyl ester mixtures at five different concentrations, and human male infertility plasma samples. Notably, the analysis of the latter dataset, despite its complexity (136 samples x 8775 scans x 766 m/z), was completed within 3 h without any manual interventions, prior knowledge and parameter optimization. DeepResolution2 automatically resolved the chromatograms and mass spectra identifying 53 metabolites utilized in statistical analysis, namely partial least squares-discriminant analysis, to differentiate between healthy control and patients with semen abnormalities. The authors additionally conducted a comparison of DeepResolution2 with AMDIS, ADAP-GC, MS-DIAL, DeepResolution, and PARADISe. This comparison considered various characteristics such as the number of parameters requiring optimization, the presence of graphical user interfaces (GUIs), processing speed, and the necessity of manual

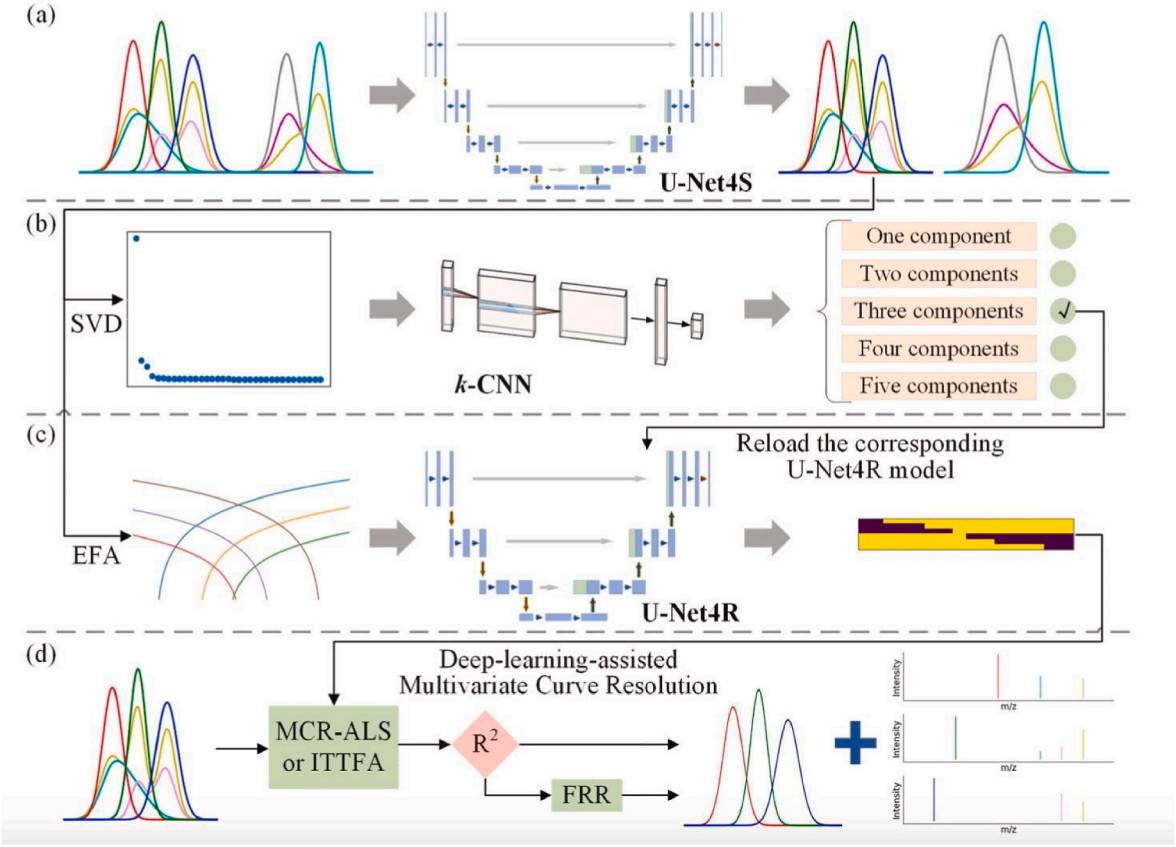

Fig. 6. Flowchart of DeepResolution2. (a) The U-Net4S model for automatic segmentation of chromatographic peaks. (b) The k-CNN model for determining the number of components k in a segment. (c) The U-Net4R model for elution region detection. (d) Deep-learning-assisted multivariate curve resolution. Reproduced with permission. Copyright Elsevier 2022 [81].

intervention or model updates with new components. According to these characteristics, the investigated methods obtained different scores with DeepResolution2 outperforming other techniques. DeepResolution2 scored highest as it only requires specifying a data directory. MS-DIAL and AMDIS, needing parameter optimization, scored slightly lower. ADAP-GC, requiring optimization and manual input, and PARADISe and DeepResolution, needing manual segmentation and checks scored lower. AMDIS, limited by single-file processing, scored lowest in batch capabilities, whereas other methods that handle multiple files scored higher. DeepResolution ranked lowest for model updates due to the needs to update the model. Although DeepResolution2, according to the authors, could be considered a highly automatic strategy, it might misjudge component numbers if peak intensity is low or baseline correction is inadequate, potentially leading to inaccurate resolutions.

4. AutoRes and GC-former automatic deep-learning approaches

In addition to PARAFAC2 and MCR-ALS, various automatic procedures that integrate advanced computational techniques, including deep learning, have been developed to enhance the processing of complex GC-MS data with less human intervention. In 2023, Fan et al. [87] introduced an innovative deep learning-based **Automatic Resolution method (AutoRes)** tailored for addressing overlapped peaks in complex GC-MS data. This method leverages a pseudo-Siamese convolutional neural network (pSCNN) architecture and is implemented in Python, available as an open-source package accessible at https://github.com/dyjfan/AutoRes.

The schematic overview of AutoRes is depicted in Fig. 7. AutoRes comprises three key components: data augmentation (Fig. 7A), pSCNN (Fig. 7B), and automatic resolution (Fig. 7C).

pSCNN consists of two distinct models, pSCNN1 and pSCNN2, sharing a unified architecture but trained on different datasets. Specifically, pSCNN1 operates on a spectral pair comprising two mass spectra of pure compounds, while pSCNN2 utilizes a spectral pair comprising a mass spectrum of a compound and a mass spectrum of a mixture. Subsequently, pSCNN1 and pSCNN2 predict the selective region and the elution region of each compound within the overlapped peak, respectively. These predicted regions serve as inputs for FRR method [87], facilitating the automated resolution of overlapped peaks. AutoRes streamlines the batch processing of untargeted GC-MS data, offering a parameter-free resolution process. The authors assessed its ability to resolve mass spectra, chromatograms, and peak areas against AMDIS and MZmine using simulated and essential oil datasets from GC-MS analysis of four plants. AutoRes and AMDIS identified 17 and 27 compounds respectively, while MZmine missed lower concentration compounds. Despite the automation AutoRes has limitations, such as not resolving embedded peaks and requiring two models, which increases analysis time.

Considering these concerns, in 2024 Guo et al. [88] proposed a Transformer-based automatic resolution method, GCFormer, to resolve mass spectra from GC-MS peaks in an end-to-end manner by predicting the mass spectra of compounds directly from raw overlapping peaks data. GCMSFormer is implemented in Python and is available as an open-source package at https://github.com/zxguocsu/GCMSFormer. GC-MS data are regarded as sequence data with a series of mass spectra, allowing the resolution of GC-MS peaks following the classic encoder and decoder architecture of the Transformer method [6]. The model output is the mass spectra with the highest probability. Each peak is decomposed by Singular Value Decomposition [89] and the obtained values are successively fed into a k-CNN model [81] to automatically

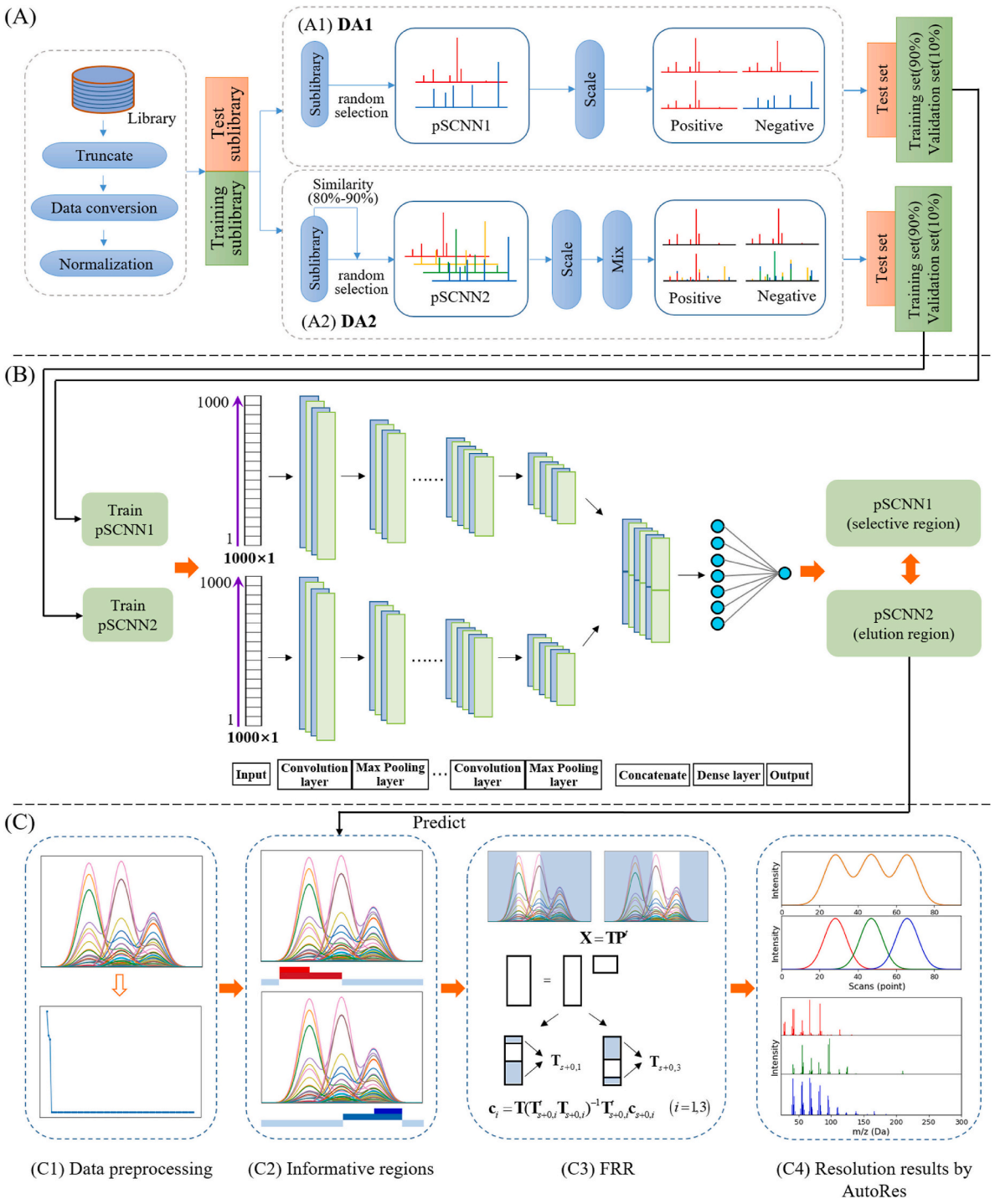

Fig. 7. Overview of AutoRes. (A) Data augmentation process. (A1) Data augmentation method 1, (DA1) for the pSCNN1 model. (A2) Data augmentation method 2, (DA2) for the pSCNN2 model. (B) The architecture of pseudo-Siamese convolutional neural network in AutoRes. (C) Automatic resolution process. (C1) Data preprocessing. (C2) Prediction of informative regions (dark: selective region and elution region, light: zero concentration region). (C3) The FRR method. (C4) Results obtained by AutoRes. Reproduced with permission. Copyright Elsevier 2022 [87].

determine the number of components. Additionally, OPR is integrated into GCMSFormer to resolve minor components. To demonstrate the performance of GCMSFormer in real-life applications, it was used to resolve five GC-MS data files coming from the GC-MS analysis of plant essential oils. Furthermore, a comparative analysis of the performance of GCMSFormer alongside other resolution tools such as MZmine, AMDIS, PARAFAC2, and MSHub [90]. The comparison was conducted based on various evaluation metrics, including the number of resolved compounds, mass spectral match score (s), correlation coefficient (r), and explained variance (R^2). Results indicated that GCMSFormer, PARAFAC2, and AMDIS successfully resolved 17 compounds each, while MZmine and MSHub detected only 14 and 13 compounds, respectively. Moreover, GCMSFormer demonstrated advantages in terms of the values of R^2, r, and s. Nonetheless, further testing on larger datasets, encompassing a broader range of samples and chemical compounds, could provide additional insights into the performance of these strategies and the timing of the analysis.

5. Conclusion and future outlook

The GC-MS technique is extensively utilized across various research domains, including metabolomics, food analysis, and environmental assessment. Despite its widespread use and profound potential, extracting meaningful information, such as the identification and quantification of compounds, from experimental data remains a significant challenge. This bottleneck primarily arises from issues related to huge dimension of dataset, data preprocessing (e.g., baseline correction, signal-to-noise ratio, and alignment) and coelution.

To address these challenges, numerous tools have been developed to assist researchers in GC-MS data analysis. However, several issues persist, particularly in defining optimal settings for data analysis and automating these processes to enhance efficiency and reliability. Chemometric tools, such as PARADISe and MCR-ALS, have shown considerable promise in this area. As shown the different published papers, reported in the present review, the application of multivariate modeling approaches can enhance peak detection and identification, allowing for the simultaneous analysis of multiple chromatographic profiles more effectively and efficiently. This capability is particularly crucial for managing the vast data generated in untargeted metabolomics, where careful selection of components and chromatographic regions significantly impacts the accuracy and usefulness of the analysis.

Despite these advancements, challenges remain, particularly in optimizing parameters such as the number of intervals in PARADISe, which can substantially increase model calculation time and accuracy of the results, and in selecting the appropriate number of components in MCR-ALS, especially in the context of untargeted analyses.

Future efforts should focus on developing expert systems for the automatic handling of model calculation, screening, and selection processes in PARADISe or MCR-ALS. Leveraging techniques such as experimental design and artificial neural networks, made more feasible by the increasing availability of data and advancements in computing power, could further streamline these processes. Additionally, conducting more comprehensive comparison studies is crucial for enhancing GC-MS analysis strategies across diverse and complex datasets. Such studies should focus on establishing robust, automated procedures that minimize user intervention.

These advancements are expected to play a pivotal role in modernizing data extraction methods and significantly advancing the field of untargeted metabolomics through GC-MS.

CRediT authorship contribution statement

Agnieszka Smolinska: Writing – original draft, Visualization, Supervision, Project administration, Methodology, Conceptualization. **Samuele Pellacani:** Writing – original draft, Visualization, Resources, Methodology, Investigation. **Michal Skawinski:** Writing – original draft, Visualization, Resources, Methodology, Investigation. **Caterina Durante:** Writing – original draft, Supervision, Project administration, Investigation, Conceptualization.

Declaration of competing interest

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.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.trac.2025.118286.

Data availability

Data will be made available on request.

References

[1] B. Gruber, F. David, P. Sandra, Capillary gas chromatography-mass spectrometry: current trends and perspectives, TrAC Trends Anal. Chem. 124 (2020) 115475, https://doi.org/10.1016/j.trac.2019.04.007.

[2] J. Ljujić, L. Vujisić, V. Tešević, I. Sofrenić, S. Ivanović, K. Simić, B. Anđelković, Critical review of selected analytical platforms for GC-MS metabolomics profiling—case study: HS-SPME/GC-MS analysis of blackberry's aroma, Foods 13 (2024), https://doi.org/10.3390/foods13081222.

[3] S.R. Maji, C. Roy, S.K. Sinha, Gas chromatography–mass spectrometry (GC-MS): a comprehensive review of synergistic combinations and their applications in the past two decades, J. Anal. Sci. Appl. Biotechnol. 5 (2023) 72–85, https://doi.org/10.48402/IMIST.PRSM/jasab-v5i2.40209.

[4] D.I.H. Hameed, F.M.S. Najm, A.H.A. Weiss, M.S. Kamel, Gas chromatography mass spectrometry (GC-MS): a tool for the health care, environmental, and pharmaceutical industries, Curr. Clin. Med. Educ. 2 (2024) 43–48.

[5] Y.-Y. Zhang, Q. Zhang, Y.-M. Zhang, W.-W. Wang, L. Zhang, Y.-J. Yu, C.-C. Bai, J.-Z. Guo, H.-Y. Fu, Y. She, A comprehensive automatic data analysis strategy for gas chromatography-mass spectrometry based untargeted metabolomics, J. Chromatogr. A 1616 (2020) 460787, https://doi.org/10.1016/j.chroma.2019.460787.

[6] N. Feizi, F.S. Hashemi-Nasab, F. Golpelichi, N. Saburouh, H. Parastar, Recent trends in application of chemometric methods for GC-MS and GC×GC-MS-based metabolomic studies, TrAC Trends Anal. Chem. 138 (2021) 116239, https://doi.org/10.1016/j.trac.2021.116239.

[7] H. Parastar, R. Tauler, Multivariate curve resolution of hyphenated and multidimensional chromatographic measurements: a new insight to address current chromatographic challenges, Anal. Chem. 86 (2014) 286–297, https://doi.org/10.1021/ac402377d.

[8] G. Tomasi, F. van den Berg, C. Andersson, Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data, J. Chemom. 18 (2004) 231–241, https://doi.org/10.1002/cem.859.

[9] C.A. Smith, E.J. Want, G. O'Maille, R. Abagyan, G. Siuzdak, XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification, Anal. Chem. 78 (2006) 779–787, https://doi.org/10.1021/ac051437y.

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

[11] H. Tsugawa, T. Cajka, T. Kind, Y. Ma, B. Higgins, K. Ikeda, M. Kanazawa, J. VanderGheynst, O. Fiehn, M. Arita, MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis, Nat. Methods 12 (2015) 523–526, https://doi.org/10.1038/nmeth.3393.

[12] J. Jaumot, R. Gargallo, A. de Juan, R. Tauler, A graphical user-friendly interface for MCR-ALS: a new tool for multivariate curve resolution in MATLAB, Chemom. Intell. Lab. Syst. 76 (2005) 101–110, https://doi.org/10.1016/j.chemolab.2004.12.007.

[13] A. Lommen, MetAlign: interface-driven, versatile metabolomics tool for hyphenated full-scan mass spectrometry data preprocessing, Anal. Chem. 81 (2009) 3079–3086, https://doi.org/10.1021/ac900036d.

[14] S.E. Stein, An integrated method for spectrum extraction and compound identification from gas chromatography/mass spectrometry data, J. Am. Soc. Mass Spectrom. 10 (1999) 770–781, https://doi.org/10.1016/S1044-0305(99)00047-1.

[15] A. Luedemann, K. Strassburg, A. Erban, J. Kopka, TagFinder for the quantitative analysis of gas chromatography—mass spectrometry (GC-MS)-based metabolite profiling experiments, Bioinformatics 24 (2008) 732–737, https://doi.org/10.1093/bioinformatics/btn023.

[16] K. Hiller, J. Hangebrauk, C. Jäger, J. Spura, K. Schreiber, D. Schomburg, MetaboliteDetector: comprehensive analysis tool for targeted and nontargeted GC/MS based metabolome analysis, Anal. Chem. 81 (2009) 3429–3439, https://doi.org/10.1021/ac802689c.

[17] C.D. Broeckling, I.R. Reddy, A.L. Duran, X. Zhao, L.W. Sumner, MET-IDEA: data extraction tool for Mass spectrometry-based metabolomics, Anal. Chem. 78 (2006) 4334–4341, https://doi.org/10.1021/ac0521596.

[18] L.G. Johnsen, P.B. Skou, B. Khakimov, R. Bro, Gas chromatography – mass spectrometry data processing made easy, J. Chromatogr. A 1503 (2017) 57–64, https://doi.org/10.1016/j.chroma.2017.04.052.

[19] R.A. Harshman, PARAFAC2: mathematical and technical notes, UCLA work, Pap. Phon. 22 (1972) 122215.

[20] X. Domingo-Almenara, J. Brezmes, M. Vinaixa, S. Samino, N. Ramirez, M. Ramon-Krauel, C. Lerin, M. Díaz, L. Ibáñez, X. Correig, A. Perera-Lluna, O. Yanes, eRah: a computational tool integrating spectral deconvolution and alignment with quantification and identification of metabolites in GC/MS-Based metabolomics, Anal. Chem. 88 (2016) 9821–9829, https://doi.org/10.1021/acs.analchem.6b02927.

[21] S. O'Callaghan, D.P. De Souza, A. Isaac, Q. Wang, L. Hodkinson, M. Olshansky, T. Erwin, B. Appelbe, D.L. Tull, U. Roessner, A. Bacic, M.J. McConville, V.A. Likić, PyMS: a python toolkit for processing of gas chromatography-mass spectrometry (GC-MS) data. Application and comparative study of selected tools, BMC Bioinf. 13 (2012) 115, https://doi.org/10.1186/1471-2105-13-115.

[22] A. Smirnov, Y. Qiu, W. Jia, D.I. Walker, D.P. Jones, X. Du, ADAP-GC 4.0: application of clustering-assisted multivariate curve resolution to spectral deconvolution of gas chromatography–mass spectrometry metabolomics data, Anal. Chem. 91 (2019) 9069–9077, https://doi.org/10.1021/acs.analchem.9b01424.

[23] P. Ma, Z. Zhang, X. Zhou, Y. Yun, Y. Liang, H. Lu, Feature extraction from resolution perspective for gas chromatography-mass spectrometry datasets, RSC Adv. 6 (2016) 113997–114004, https://doi.org/10.1039/C6RA17864B.
[24] P. Ma, M. Li, H. Lu, Z. Zhang, Mars 2: a computational tool to resolve and extract features from large-scale GC-MS datasets, Chemom. Intell. Lab. Syst. 191 (2019) 12–20, https://doi.org/10.1016/j.chemolab.2019.05.010.
[25] R. Tautenhahn, C. Böttcher, S. Neumann, Highly sensitive feature detection for high resolution LC/MS, BMC Bioinf. 9 (2008) 504, https://doi.org/10.1186/1471-2105-9-504.
[26] O.D. Myers, S.J. Sumner, S. Li, S. Barnes, X. Du, Detailed investigation and comparison of the XCMS and MZmine 2 chromatogram construction and chromatographic peak detection methods for preprocessing mass spectrometry metabolomics data, Anal. Chem. 89 (2017) 8689–8695, https://doi.org/10.1021/acs.analchem.7b01069.
[27] A. Samokhin, Spectral skewing in gas chromatography–mass spectrometry: misconceptions and realities, J. Chromatogr. A 1576 (2018) 113–119, https://doi.org/10.1016/j.chroma.2018.09.033.
[28] S. Wold, M. Sjostrom, SIMCA: a method for analysing chemical data in terms of similarity and analogy. Chemometrics: Theory and Application, American Chemical Society, 1977, pp. 243–282, https://doi.org/10.1021/bk-1977-0052.ch012. Chapter 12.
[29] M. Barker, W. Rayens, Partial least squares for discrimination, J. Chemom. 17 (2003) 166–173, https://doi.org/10.1002/cem.785.
[30] B. Quintanilla-Casas, R. Bro, J.L. Hinrich, C.L. Davie-Martin, Tutorial on Paradise: Parafac2-Based Deconvolution and Identification System for Processing GC–MS Data, PROTOCOL, 2023, https://doi.org/10.21203/rs.3.pex-2143/v1.
[31] Tamara G. Kolda, Brett W. Bader, Tensor decompositions and applications, SIAM Rev. 51 (2009) 455–500, https://doi.org/10.1137/07070111x.
[32] J.D. Carroll, J.J. Chang, Analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart-Young" decomposition, Psychometrika 35 (1970) 283–319.
[33] R.A. Harshman, Foundations of the PARAFAC procedure: models and conditions for an "explanatory" multi-modal factor analysis, UCLA Work. Pap. Phon. 16 (1970) 1–84.
[34] R. Bro, C.A. Andersson, H.A.L. Kiers, PARAFAC2—Part II. Modeling chromatographic data with retention time shifts, J. Chemom. 13 (1999) 295–309, https://doi.org/10.1002/(SICI)1099-128X(199905/08)13:3/4<295::AID-CEM547>3.0.CO;2-Y.
[35] R. Bro, PARAFAC. Tutorial and applications, Chemom. Intell. Lab. Syst. 38 (1997) 149–171, https://doi.org/10.1016/S0169-7439(97)00032-4.
[36] H. Kiers, J. Berge, R. Bro, PARAFAC2—Part I. A direct fitting algorithm for the PARAFAC2 model, J. Chemom. 13 (1999) 275–294, https://doi.org/10.1002/(SICI)1099-128X(199905/08)13:3/43.3.CO;2-2.
[37] C.A. Andersson, R. Bro, The *N*-way toolbox for MATLAB, Chemom. Intell. Lab. Syst. 52 (2000) 1–4, https://doi.org/10.1016/S0169-7439(00)00071-X.
[38] A.B. Risum, R. Bro, Using deep learning to evaluate peaks in chromatographic data, Talanta 204 (2019) 255–260, https://doi.org/10.1016/j.talanta.2019.05.053.
[39] J. Rieksta, T. Li, C.L. Davie-Martin, T. Simin, L.C. Brogaard Aeppli, T.T. Høye, M. A. Rasmussen, R. Rinnan, Seasonal and elevational variability in dwarf birch VOC emissions in Greenlandic tundra, J. Geophys. Res. Biogeosciences 128 (2023) e2023JG007475, https://doi.org/10.1029/2023JG007475.
[40] P.D. Smirnov, R.K. Puzanskiy, S.A. Vanisov, M.D. Dubrovskiy, A.L. Shavarda, M. F. Shishova, V.V. Yemelyanov, Metabolic profiling of leaves of four ranunculus species, Ecol. Genet. 21 (2024) 369–382, https://doi.org/10.17816/ecogen623592.
[41] N.P. Bityutskii, K.L. Yakkonen, R. Puzanskiy, K.A. Lukina, A.L. Shavarda, K. N. Semenov, Fullerenol changes metabolite responses differently depending on the iron status of cucumber plants, PLoS One 16 (2021) e0251396, https://doi.org/10.1371/journal.pone.0251396.
[42] A.S.Q. Souza, J.A.C. Sousa, C.S. Pinto, E.G. Alves Filho, R.C.A. Pereira, E.S. Brito, K.M. Canuto, L. Leal, Untargeted GC/MS-based approach for identification of anti-inflammatory alkaloids from Hippeastrum elegans (amaryllidaceae) using a human neutrophil model, J. Pharm. Biomed. Anal. 199 (2021) 114061, https://doi.org/10.1016/j.jpba.2021.114061.
[43] L. Lacalle-Bergeron, R. Goterris-Cerisuelo, T. Portoles, J. Beltran, J.V. Sancho, C. Navarro-Moreno, F. Martinez-Garcia, Novel sampling strategy for alive animal volatolome extraction combined with GC-MS based untargeted metabolomics: identifying mouse pup pheromones, Talanta 235 (2021) 122786, https://doi.org/10.1016/j.talanta.2021.122786.
[44] P.M. Duarte, E. Maciel, M. Pinho, M.R. Domingues, R. Calado, A.I. Lillebø, O.M.C. C. Ameixa, Omega-3 on the fly: long-legged fly Machaerium maritimae as a potential source of eicosapentaenoic acid for aquafeeds, J. Insects Food Feed 7 (2021) 1089–1100, https://doi.org/10.3920/jiff2020.0112.
[45] K.H. Bak, S.S. Waehrens, Y. Fu, C.Y. Chow, M.A. Petersen, J. Ruiz-Carrascal, W.L. P. Bredie, R. Lametsch, Flavor characterization of animal hydrolysates and potential of glucosamine in flavor modulation, Foods 10 (2021) 3008.
[46] F. Garavand, D.F.M. Daly, L.G. Gómez-Mascaraque, The consequence of supplementing with synbiotic systems on free amino acids, free fatty acids, organic acids, and some stability indexes of fermented milk, Int. Dairy J. 137 (2023) 105477, https://doi.org/10.1016/j.idairyj.2022.105477.
[47] A.N. Murmann, M. Bevilacqua, B.P. Danielsen, T. Jansson, K. Engholm-Keller, M. M. Poojary, N. Arneborg, M.N. Lund, The impact of lager brewing yeasts on flavor stability of pilot-scale beer during storage, Eur. Food Res. Technol. 250 (2024) 715–725, https://doi.org/10.1007/s00217-023-04421-8.
[48] P.L. Ashmore, A. DuBois, E. Tomasino, J.F. Harbertson, T.S. Collins, Impact of dilution on whisky aroma: a sensory and volatile composition analysis, Foods 12 (2023) 1276.
[49] A. Warburton, P. Silcock, G.T. Eyres, Impact of sourdough culture on the volatile compounds in wholemeal sourdough bread, Food Res. Int. 161 (2022) 111885, https://doi.org/10.1016/j.foodres.2022.111885.
[50] R. Rubio-Sánchez, R. Ríos-Reina, C. Ubeda, Effect of chemotherapy on urinary volatile biomarkers for lung cancer by HS-SPME-GC-MS and chemometrics, Thorac. Cancer 14 (2023) 3522–3529, https://doi.org/10.1111/1759-7714.15154.
[51] J. Carlsen, H.H. Henriksen, I. Marin de Mas, P.I. Johansson, An explorative metabolomic analysis of the endothelium in pulmonary hypertension, Sci. Rep. 12 (2022) 13284, https://doi.org/10.1038/s41598-022-17374-x.
[52] R. Rubio-Sánchez, C. Ubeda, R. Ríos-Reina, Feasibility of using volatile urine fingerprints for the differentiation of sexually transmitted infections, Appl. Microbiol. Biotechnol. 107 (2023) 6363–6376, https://doi.org/10.1007/s00253-023-12711-0.
[53] F.H. Larsen, F. van den Berg, S.B. Engelsen, An exploratory chemometric study of 1H NMR spectra of table wines, J. Chemom. 20 (2006) 198–208, https://doi.org/10.1002/cem.991.
[54] H. Yu, R. Bro, PARAFAC2 and local minima, Chemom. Intell. Lab. Syst. 219 (2021) 104446, https://doi.org/10.1016/j.chemolab.2021.104446.
[55] J.M. Amigo, T. Skov, R. Bro, J. Coello, S. Maspoch, Solving GC-MS problems with PARAFAC2, TrAC Trends Anal. Chem. 27 (2008) 714–725, https://doi.org/10.1016/j.trac.2008.05.011.
[56] A. de Juan, J. Jaumot, R. Tauler, Multivariate curve resolution (MCR). Solving the mixture analysis problem, Anal. Methods 6 (2014) 4964–4976, https://doi.org/10.1039/C4AY00571F.
[57] J. Jaumot, A. de Juan, R. Tauler, MCR-ALS GUI 2.0: new features and applications, Chemom. Intell. Lab. Syst. 140 (2015) 1–12, https://doi.org/10.1016/j.chemolab.2014.10.003.
[58] R. Tauler, M. Maeder, Two-way data analysis: multivariate curve resolution - error in curve resolution, Compr. Chemom. (2009) 345–363, https://doi.org/10.1016/B978-044452701-1.00051-X.
[59] I.H.M. Van Stokkum, K.M. Mullen, V.V. Mihaleva, Global analysis of multiple gas chromatography–mass spectrometry (GC/MS) data sets: a method for resolution of co-eluting components with comparison to MCR-ALS, Chemom. Intell. Lab. Syst. 95 (2009) 150–163, https://doi.org/10.1016/j.chemolab.2008.10.004.
[60] I. Santana, M. Breitkreitz, L. Pinto, Multivariate curve resolution alternating least squares applied to chromatographic data: from the basics to the recent advances, Braz. J. Anal. Chem. 8 (2021), https://doi.org/10.30744/brjac.2179-3425.RV-30-2021.
[61] R. Tauler, Multivariate curve resolution applied to second order data, Chemom. Intell. Lab. Syst. 30 (1995) 133–146, https://doi.org/10.1016/0169-7439(95)00047-X.
[62] E. Gorrochategui, J. Jaumot, R. Tauler, ROIMCR: a powerful analysis strategy for LC-MS metabolomic datasets, BMC Bioinf. 20 (2019), https://doi.org/10.1186/s12859-019-2848-8.
[63] M. Asadollahi-Baboli, A. Mani-Varnosfaderani, Chemometrics-assisted GC-MS analysis of volatile and semi-volatile constituents of Elettaria cardamomum, Food Anal. Methods 7 (2014) 1745–1754, https://doi.org/10.1007/s12161-014-9815-9.
[64] O.M. Kvalheim, Y.Zeng Liang, Heuristic evolving latent projections: resolving two-way multicomponent data. 1. Selectivity, latent-projective graph, datascope, local rank, and unique resolution, Anal. Chem. 64 (1992) 936–946, https://doi.org/10.1021/ac00032a019.
[65] F. Azimi, M.H. Fatemi, Multivariate curve resolution-correlation optimized warping applied to the complex GC-MS signals; toward comparative study of peel chemical variability of citrus aurantium L. varieties, Microchem. J. 143 (2018) 99–109, https://doi.org/10.1016/j.microc.2018.07.041.
[66] G. Aliakbarzadeh, H. Sereshti, H. Parastar, Pattern recognition analysis of chromatographic fingerprints of crocus sativus L. secondary metabolites towards source identification and quality control, Anal. Bioanal. Chem. 408 (2016) 3295–3307, https://doi.org/10.1007/s00216-016-9400-8.
[67] M. Asadollahi-Baboli, A. Aghakhani, V. Bikdeloo, Application of polyamide nanofibers, SPME/GC-MS, and chemometrics for comprehensive analysis of volatiles in Thymus vulgaris L. and Thymus serpyllum L, Food Anal. Methods 9 (2016) 528–536, https://doi.org/10.1007/s12161-015-0224-5.
[68] G. Aliakbarzadeh, H. Sereshti, H. Parastar, Fatty acids profiling of avocado seed and pulp using gas chromatography–mass spectrometry combined with multivariate chemometric techniques, J. Iran. Chem. Soc. 13 (2016) 1905–1913, https://doi.org/10.1007/s13738-016-0907-x.
[69] H. Malekzadeh, M.H. Fatemi, Application of multivariate curve resolution approaches to improve analytical separation of Iranian rice volatiles by GC-MS, Bull. Chem. Soc. Jpn. 88 (2015) 706–712, https://doi.org/10.1246/bcsj.20140362.
[70] E. Garreta-Lara, B. Campos, C. Barata, S. Lacorte, R. Tauler, Metabolic profiling of Daphnia magna exposed to environmental stressors by GC–MS and chemometric tools, Metabolomics 12 (2016) 86, https://doi.org/10.1007/s11306-016-1021-x.
[71] E. Garreta-Lara, B. Campos, C. Barata, S. Lacorte, R. Tauler, Combined effects of salinity, temperature and hypoxia on Daphnia magna metabolism, Sci. Total Environ. 610–611 (2018) 602–612, https://doi.org/10.1016/j.scitotenv.2017.05.190.
[72] M. Ahmadvand, H. Sereshti, H. Parastar, Chemometric-based determination of polycyclic aromatic hydrocarbons in aqueous samples using ultrasound-assisted emulsification microextraction combined to gas chromatography–mass spectrometry, J. Chromatogr. A 1413 (2015) 117–126, https://doi.org/10.1016/j.chroma.2015.08.026.

[73] X. Qing, X. Zhou, L. Xu, J. Zhang, Y. Huang, L. Lin, Z. Liu, X. Zhang, Application of alternating trilinear decomposition-assisted multivariate curve resolution to gas chromatography-mass spectrometric data for the quantification of polycyclic aromatic hydrocarbons in aerosols, R. Soc. Open Sci. 8 (2021) 210458, https://doi.org/10.1098/rsos.210458.
[74] L. Lebanov, L. Tedone, A. Ghiasvand, B. Paull, Characterisation of complex perfume and essential oil blends using multivariate curve resolution-alternating least squares algorithms on average mass spectrum from GC-MS, Talanta 219 (2020) 121208, https://doi.org/10.1016/j.talanta.2020.121208.
[75] Saadi Saaidpour, Armin Jahannamaie, Using gas chromatography–mass spectrometry in conjunction with chemometrics technique for characterization of Iranian essential oils Myrtus communis L, J. Anal. Chem. 74 (2019) 756–763, https://doi.org/10.1134/S1061934819080136.
[76] N. Vural, Chemometrics data analysis and controversial carcinogenic effect of Ocimum basilicum L. rich in methyl eugenol, J. Food Meas. Charact. 15 (2021) 4825–4837, https://doi.org/10.1007/s11694-021-01072-9.
[77] F. Tajabadi, F. Khalighi-Sigaroodi, S. Rezazadeh, Improving gas chromatography–mass spectrometry analysis of essential oils by multivariate curve resolution: full identification of Co-eluting compounds of Dracocephalum moldavica L, Chromatographia 80 (2017) 1069–1077, https://doi.org/10.1007/s10337-017-3322-2.
[78] N. Shekari, M. Vosough, K. Tabar Heidar, Chromatographic fingerprinting through chemometric techniques for herbal slimming pills: a way of adulterant identification, Forensic Sci. Int. 286 (2018) 213–222, https://doi.org/10.1016/j.forsciint.2018.03.022.
[79] M. Jalali-Heravi, H. Parastar, M. Kamalzadeh, R. Tauler, J. Jaumot, MCRC software: a tool for chemometric analysis of two-way chromatographic data, Chemom. Intell. Lab. Syst. 104 (2010) 155–171, https://doi.org/10.1016/j.chemolab.2010.08.002.
[80] 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.
[81] X. Fan, Z. Xu, H. Zhang, D. Liu, Q. Yang, Q. Tao, M. Wen, X. Kang, Z. Zhang, H. Lu, Fully automatic resolution of untargeted GC-MS data with deep learning assistance, Talanta 244 (2022) 123415, https://doi.org/10.1016/j.talanta.2022.123415.
[82] X. Domingo-Almenara, A. Perera, J. Brezmes, Avoiding hard chromatographic segmentation: a moving window approach for the automated resolution of gas chromatography–mass spectrometry-based metabolomics signals by multivariate methods, J. Chromatogr. A 1474 (2016) 145–151, https://doi.org/10.1016/j.chroma.2016.10.066.
[83] X. Zhang, R. Tauler, Flexible implementation of the trilinearity constraint in multivariate curve resolution alternating least squares (MCR-ALS) of chromatographic and other type of data, Molecules 27 (2022), https://doi.org/10.3390/molecules27072338.
[84] H.-Y. Fu, J.-W. Guo, Y.-J. Yu, H.-D. Li, H.-P. Cui, P.-P. Liu, B. Wang, S. Wang, P. Lu, A simple multi-scale gaussian smoothing-based strategy for automatic chromatographic peak extraction, J. Chromatogr. A 1452 (2016) 1–9, https://doi.org/10.1016/j.chroma.2016.05.018.
[85] A.C. Olivieri, K. Neymeyr, M. Sawall, R. Tauler, How noise affects the band boundaries in multivariate curve resolution, Chemom. Intell. Lab. Syst. 220 (2022) 104472, https://doi.org/10.1016/j.chemolab.2021.104472.
[86] R. Tauler, Multivariate curve resolution of multiway data using the multilinearity constraint, J. Chemom. 35 (2021) e3279, https://doi.org/10.1002/cem.3279.
[87] Y. Fan, C. Yu, H. Lu, Y. Chen, B. Hu, X. Zhang, J. Su, Z. Zhang, Deep learning-based method for automatic resolution of gas chromatography-mass spectrometry data from complex samples, J. Chromatogr. A 1690 (2023) 463768, https://doi.org/10.1016/j.chroma.2022.463768.
[88] Z. Guo, Y. Fan, C. Yu, H. Lu, Z. Zhang, GCMSFormer: a fully automatic method for the resolution of overlapping peaks in gas chromatography–mass spectrometry, Anal. Chem. 96 (2024) 5878–5886, https://doi.org/10.1021/acs.analchem.3c05772.
[89] G.H. Golub, C. Reinsch, Singular value decomposition and least squares solutions, Numer. Math. 14 (1970) 403–420, https://doi.org/10.1007/BF02163027.
[90] A.A. Aksenov, I. Laponogov, Z. Zhang, S.L.F. Doran, I. Belluomo, D. Veselkov, W. Bittremieux, L.F. Nothias, M. Nothias-Esposito, K.N. Maloney, B.B. Misra, A. V. Melnik, A. Smirnov, X. Du, K.L. Jones, K. Dorrestein, M. Panitchpakdi, M. Ernst, J.J.J. van der Hooft, M. Gonzalez, C. Carazzone, A. Amézquita, C. Callewaert, J. T. Morton, R.A. Quinn, A. Bouslimani, A.A. Orio, D. Petras, A.M. Smania, S. P. Couvillion, M.C. Burnet, C.D. Nicora, E. Zink, T.O. Metz, V. Artaev, E. Humston-Fulmer, R. Gregor, M.M. Meijler, I. Mizrahi, S. Eyal, B. Anderson, R. Dutton, R. Lugan, P.L. Boulch, Y. Guitton, S. Prevost, A. Poirier, G. Dervilly, B. Le Bizec, A. Fait, N.S. Persi, C. Song, K. Gashu, R. Coras, M. Guma, J. Manasson, J.U. Scher, D.K. Barupal, S. Alseekh, A.R. Fernie, R. Mirnezami, V. Vasiliou, R. Schmid, R. S. Borisov, L.N. Kulikova, R. Knight, M. Wang, G.B. Hanna, P.C. Dorrestein, K. Veselkov, Auto-deconvolution and molecular networking of gas chromatography-mass spectrometry data, Nat. Biotechnol. 39 (2021) 169–173, https://doi.org/10.1038/s41587-020-0700-3.