Journal of Chromatography A 1640 (2021) 461896

Contents lists available at ScienceDirect

Journal of Chromatography A

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

Review article

Data handling and data analysis in metabolomic studies of essential oils using GC-MS

Leo Lebanov[a,b], Alireza Ghiasvand[a], Brett Paull[a,b,*]

[a] Australian Centre for Research on Separation Science (ACROSS), School of Natural Sciences, University of Tasmania, Hobart, TAS, Australia
[b] ARC Industrial Transformation Research Hub for Processing Advanced Lignocellulosics (PALS), School of Natural Sciences, University of Tasmania, Hobart, TAS, Australia

ARTICLE INFO

Article history:
Received 10 November 2020
Accepted 8 January 2021
Available online 22 January 2021

Keywords:
Gas chromatography electron impact ionization mass spectrometry
Data handling
Multivariate statistical analysis: Essential oils

ABSTRACT

Gas chromatography electron impact ionization mass spectrometry (GC-EI-MS) has been, and remains, the most widely applied analytical technique for metabolomic studies of essential oils. GC-EI-MS analysis of complex samples, such as essential oils, creates a large volume of data. Creating predictive models for such samples and observing patterns within complex data sets presents a significant challenge and requires application of robust data handling and data analysis methods. Accordingly, a wide variety of software and algorithms has been investigated and developed for this purpose over the years. This review provides an overview and summary of that research effort, and attempts to classify and compare different data handling and data analysis procedures that have been reported to-date in the metabolomic study of essential oils using GC-EI-MS.

1. Introduction

Metabolomics is the systematic determination of the metabolite profile within biological samples, such cells, biofluids, tissues, organs and organisms, at any specific point in time. Collectively, these metabolites, and their functional interactions within biological system, are known as the metabolome. Differences in metabolomes reflect responses of the system to environmental or genetic variations, and their study provides insight into the precise nature of these specific physiological responses [1].

Essential oils (EOs) consist of volatile and semi-volatile compounds, which are products of the secondary metabolism of plants, where each species and chemotype produces EOs with unique chemical profiles [2,3]. Secondary metabolites are accumulated and stored in the plant organs at different rates, leading to the differences in the organ metabolomes. Abiotic factors coming from the geographical location also affect the physiological responses within the plant and consequently the chemical profile of EOs, as well as their physicochemical properties and biological potency [4–6].

Gas chromatography electron impact ionization mass spectrometry (GC-EI-MS) is the most widely used technique for the separation and chemical profiling of EO [7]. GC-EI-MS produces representative and reproducible chromatograms of volatile metabolites, together with the associated mass spectra for each separated component peak, which are then used to aid metabolite identification. EOs are complex samples which may contain up to several hundred compounds. For example, in the case of vetiver EO, a composition of in excess of 500 compounds has been identified, utilising comprehensive two-dimensional gas chromatography mass spectrometry system (GC × GC-MS) [8]. As a result of such sample complexity, GC-EI-MS metabolomic analysis of EOs inevitably produces large volumes of data. Therefore, to fully ascertain the effect of genetic and environmental variations on EO metabolome profiles typically requires application of multivariate statistical techniques [9]. Prior to any data analysis technique, the recorded raw data created via GC-EI-MS analysis needs to pass through data handling steps. Firstly, datasets need to be "cleaned" from any unwanted experimental and instrumental artefacts through data pre-processing. Also, the size of datasets needs to be reduced, as well as features aligned. If not performed appropriately, data handling procedures can introduce new unwanted variations, and hamper final metabolomic interpretations [10].

Data analysis techniques can be divided on pattern recognition and classification approaches, prediction methods, and multiway resolution techniques. An important aspect of pattern recognition techniques is the possibility to visualise multidimensional datasets into two- or three-dimensions, while preserving the variance among the samples [11]. This provides for observation of similarities/dissimilarities among samples. By determining the importance of each metabolite/feature in the created models, final con-

* Corresponding author.
E-mail addresses: leo.lebanov@utas.edu.au (L. Lebanov), alireza.ghiasvand@utas.edu.au (A. Ghiasvand), brett.paull@utas.edu.au (B. Paull).

https://doi.org/10.1016/j.chroma.2021.461896
0021-9673/

clusions regarding the effects of genetic and environmental variations on the EO chemical profiles can be made.

Over the years, there have been a number of reviews published on multivariate approaches in metabolomic studies, based on either GC-MS, liquid chromatography-mass spectrometry (LC-MS), or nuclear magnetic resonance spectroscopy (NMR) [11–21]. However, this review differs from those previously in its exclusive focus on novel approaches the handling of GC-EI-MS data, and specifically on feature alignment, data reduction, and feature selection in GC-EI-MS based metabolomic studies. Different statistical approaches to pattern recognition, prediction, and multiway resolution techniques, as well as validation models, applied within EO metabolomic studies, are also critically reviewed. This review also provides guidance in terms of advantages and potential pitfalls in GC-EI-MS data handling and validation procedures, in order to help practitioners in creating more robust models delivering greater accuracy.

2. Data handling and data analysis workflows

Fig. 1 depicts the typical workflow of a GC-EI-MS based metabolomic study. This process may consist of multiple steps, including data acquisition, data pre-processing, feature alignment, pre-treatment, data reduction, feature selection, data analysis, model validation and examination. Determination of the appropriate procedures for data handling and data analysis depends on several factors, including the nature of the particular phenomena under investigation, the defined metabolomic questions, and whether undertaking a targeted or untargeted approach within the metabolomic study.

2.1. Dimensionality and data acquisition in GC-EI-MS

In the GC-EI-MS analysis of EOs, each separated component provides a unique fragmentation pattern (*m/z* profile). The *m/z* intensities are recorded as a function of time and correspond to the chromatographic retention of the associated peak. Consequently, a three-way data array is created for each separated component, with retention time, *m/z* fragments, and intensities, a typical example of which can be seen as Fig. 2. In GC-EI-MS based metabolomic studies, where numerous samples of differing background are analysed, four-way data arrays are formed [22].

Data acquisition in GC-EI-MS can be performed either in profile or centroid mode. The real-time MS signal is always a continuous signal, which is sampled at a predefined time interval or displacement (distance), before being translated into the *m/z* profile. The profile is a convoluted signal between the theoretical isotope distribution and continuous function, called the peak shape function. The peak shape function provides a numerical representation of the arrival time distribution of a population of isotopically pure *m/z* ions, and it depends on the *m/z* ion energy and velocity spread inside the MS. The peak shape function is unique for each MS under given conditions, such as mass analyser geometries and applied voltages on various ion lenses [23].

The centroid data acquisition mode is much more common within GC-EI-MS. In this mode, from multiple *m/z* values corresponding to the same peak, only a single *m/z* is recorded with zero width, as the weighted centre of the *m/z* histogram. The intensity associated with the centroid *m/z* is the normalised area of the peak, calculated from the intensities of the associated data points [21]. Recording the data in centroid mode, detection of the noise level, linearity of ion signals, and designation of isotopes' fine features becomes more complicated [23]. Centroid data acquisition significantly decreases the data amount and size of the recorded files, although also sees a loss of information.

2.2. Signal pre-processing

Obviously, the majority of the signal and/or profile obtained in the instrumental analysis of a complex sample are specific to the sample itself, and components thereof. However, variations in the signal/profile can also originate from the applied instrumental and experimental conditions. To remove all unwanted signals coming from such instrumental and experimental conditions, it is necessary to perform data pre-processing steps. In this way, only relevant information related to the chemistry of the samples is obtained [24]. The variations due to instrumental and experimental conditions can emerge as baseline drift and spectral background. The latter portion is more significant for complex samples, which includes EOs. Fortunately, the negative effect of these inputs can be reduced using various denoising and smoothing methods, such as Fourier transformation (FT) and Savitzky-Golay derivatives, representing two of the most important procedures [25]. Noise can be removed by smoothing the signal, applying local medians, applying locally weighted scatterplot smoothers, or using wavelets [25]. If smoothing is performed by a weighted function, different parts of the spectra can be corrected independently [18]. Different baseline effects can lead to a vertical offset or a slope of the instrumental signal, which differs between different analytical techniques. The most common way for correcting a baseline is performed by subtraction of the estimated baseline from the measured spectrum [24]. Estimation of a baseline can be performed by detrending [26] or utilising an asymmetric least squares smoothing (AsLS) algorithm [27]. In the detrending technique, a fixed degree polynomial curve is fitted to the spectrum and subsequently subtracted from the instrumental signal. In the AsLS approach, the baseline is estimated using a smoother signal combined with asymmetric weighting of the deviations, observed from the smoothed peak signal. A derivative of the signal can also be used to correct for the baseline effects. A first-order derivative is applied to correct for a constant baseline (offset), while a second-order derivative will correct for a baseline slope, where the Savitzky-Golay algorithm is frequently used. However, derivatives can reduce the signal-to-noise ratio (S/N) [10] and subsequently increase the limits of detection and quantification of the instrumental method.

2.3. Data alignment

To observe differences among samples and create prediction models, there must be one-to-one correspondence between the variables being compared. This is done through the process of peak detection and deconvolution, followed by chromatogram alignment. For automatic peak detection, it is important to differentiate the signals originating from the sample, from that of the noise. Setting a single threshold to differentiate between signal and noise for a whole GC analysis is problematic, since there is typically an increase of the baseline during the chromatographic run, due to the application of a temperature gradient. As a result, procedures that can evaluate peak shape together with minimal intensity (peak area or S/N) are widely applied to distinguish signals from noise. Most of the applied algorithms for the peak detection use either derivative techniques or matched filter responses. The algorithms based on the first derivative of total ion current chromatogram (TICC) signal, consider the fact that true peaks will have signals going from positive to negative with zero-crossing (at peak maximum). The problem with the calculation of signal derivatives is that it magnifies noise. Therefore, an extensive and appropriate data pre-processing step is required. In contrast, the matched filter approach applies the assumption of the Gaussian shape of the peaks, without affecting the S/N [28].

Peak detection is followed by the peak/feature alignment. Proper alignment, in both *m/z* and retention time dimensions, is

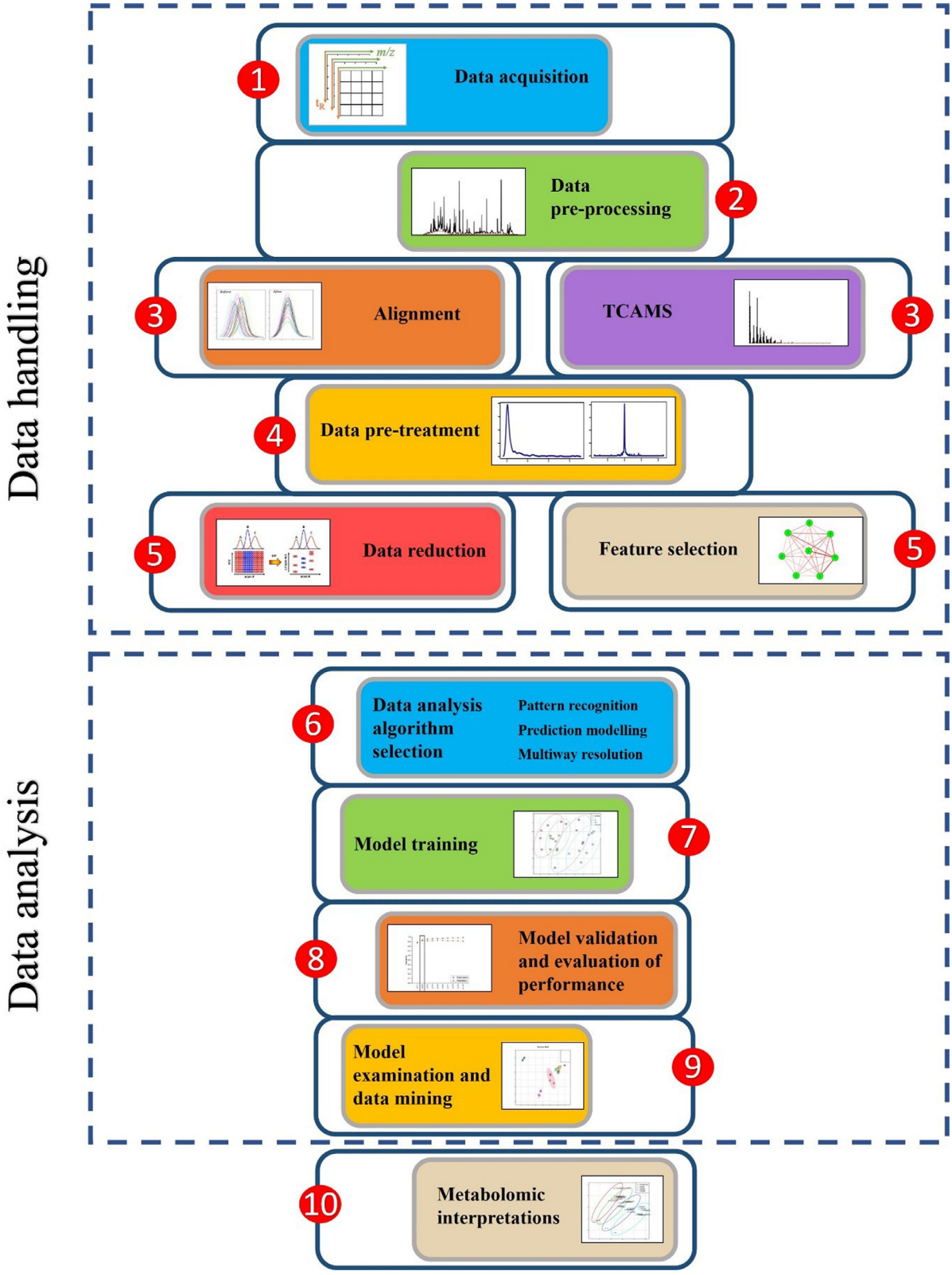

Fig. 1. Data handling and data analysis flowchart in GC-EI-MS metabolomic studies.

a mandatory requirement to create the input data matrix for subsequent data analysis procedures. While shifts along the *m/z* axis are usually easily corrected using calibration, alignment along the retention time axis is more complicated. Even though improvements in chromatographic instruments has reduced these shifts, they can originate from unforeseen factors, including temperature fluctuations or column deterioration [21]. Many algorithms have been developed for the chromatographic alignment along the retention time axis (see Table 1). The most common approaches for the chromatographic peak alignment, applied in com-

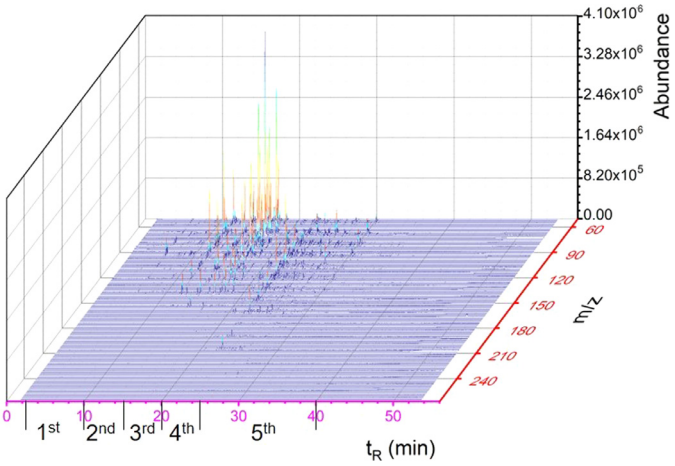

Fig. 2. A typical three-way data array recorded in GC-MS analysis.

mercially available software for LC and GC alignment, are based on the warping methods [29]. In this approach, the goal is to match specific (known) components between runs by mapping the shifts along the retention time axes. To achieve this, chromatograms are shifted, stretched, or squeezed monotonically until the differences between known component retention times are minimized. In the dynamic time warping (DTW) algorithm, peaks present across all chromatograms are used to minimise the retention time shifts during the analysis. In contrast, correlation-optimized time warping (COW) performs the alignment of the correlated *m/z* features. This is done through a piecewise linear alignment, where chromatograms are segmented, and each segment is adjusted (shrunken or stretched) to minimise retention time differences to the reference chromatogram (see Fig. 3). When applying such warping methods, due to the theoretical limitations of the algorithm, component-level shifts cannot be corrected. The warping function operates on a systematic level, where the scale is bounded by computational complexity. If the warping function is applied on more segments, more computation power is required [29]. With chromatographic separations, it is known that some sample components with similar selectivity can switch elution order. Therefore, most of the warping algorithms make the wrong assumption that elution order is preserved across runs. This change in elution order can only be corrected if the compounds in question have different masses and their *m/z* channels are warped independently [29]. In addition, application of warping algorithms can deform the original signal, which is detrimental when searching for patterns among the samples in the subsequent data analysis procedures, specifically where feature intensities are targeted in the discrimination/prediction models [30].

MetAlign [31] performs alignment by selecting the *m/z* features present across all samples, defined as landmarks. Through the iterative adjustments of landmarks versus the reference sample, all *m/z* features are corrected. The XCMS algorithm utilises a nonlinear retention time alignment method. In this approach, the observed deviations along the retention time axis are considered nonlinear and a local regression fitting method (LOESS) is applied. The LOESS approach segments the chromatogram, and retention time drifts within each segment are corrected by fitting low-order polynomials. This allows component-level shifts in the segment to be detected and corrected [32]. MZmine 2 [33] applies the same LOESS algorithm together with a random sample consensus (RANSAC) algorithm to estimate the optimal segment size for the alignment procedure.

Compared to warping techniques, the direct match approach aligns analytes based on established correspondence between mass spectra and retention indices. This approach is more common in software dedicated to the alignment of GC-EI-MS chromatograms. With the alignment procedure applied within TagFinder, retention indices are calculated based on the retention time of internal reference standards with known retention indices, such as n-alkanes or fatty acid methyl esters, through linear interpolation [34]. Based on these retention indices and mass spectra, mass fragments are sorted. The mass fragments from the same compounds across all samples are binned and aligned into mass tags and a Pearson/Spearman correlation is applied to identify correlated clusters of tags [34]. In the TargetSearch algorithm, peaks are aligned based on the retention indices [35]. With the ChromAlignNet software, a deep neural network is implemented to calculate similarity between the peaks based on their corresponding mass spectra, peak profiles and the segment of the chromatogram in which they elute [36]. Fragment-level analysis (flagme) [37] and Python toolkit for processing of GC-MS data (PyMS) [38] are based upon alignment procedures that apply dynamic programming approaches. In these methods, the similarities between peaks are calculated from their retention time and mass spectra. The alignment between the chromatograms is built progressively based upon similarity of their chemical profiles, starting with two chromatograms with the most similar chemical profiles. Recently developed QPMASS uses an advanced dynamic programming approach to align peaks, again based on retention time and mass spectra. Implemented parallel computing allows for fast processing of large datasets, while introduced missing value filtering and backfilling procedures reduce the number of potential misalignments [39]. eRah [40] is a free computational tool written in the open language R, which performs a peak deconvolution prior to peak alignment (Fig. 4). In this process, the user defines the maximum retention time drift and minimum spectral similarity for positive peak identification among chromatograms. The alignment is then performed through computation of Euclidean distances between the retention times and spectral similarity, where compounds appear across the maximum number of samples with the least retention time and spectral distance.

Another important step in peak alignment is defining the quantitative characteristics of the aligned peaks. Selection of the best quantitative *m/z* is a critical step in the GC-EI-MS data analysis, and therefore different strategies have been developed for selection of quantitative *m/z* within existing software. In the MetaboliteDetector algorithm the most intense and unique *m/z*, not shared with neighbouring peaks, are chosen. In PyMS, a single ion shared in all peaks in a certain time (from N abundant ions in the alignment results) is chosen, while in flagme, intensities of all ions corresponding to a peak are calculated. QPMASS uses a three-parameter strategy to select the optimal quantification *m/z*. If peak quantification is based on different mass fragments for the same peak in different samples, quantitative comparison between the samples becomes difficult and leads to wrongful conclusions [39].

Many algorithms, like XCMS and MetAlign, have been primarily developed for peak alignment in LC-MS. However, they have also been commonly applied to GC-EI-MS data. Being oriented toward the detection of single *m/z*, they can accurately detect single *m/z* traces. Therefore, in GC-EI-MS analysis, they do not assemble fragmentation patterns for each component, but rather report multiple *m/z* for each analyte [38].

Most of the alignment algorithms require the application of a reference sample, which is chosen by the user. This can cause different solutions based on the chosen reference sample. Critical assessment and comparison of GC-MS data alignment algorithms can be found elsewhere [41,42]. Vestner *et. al.* [43] proposed a mathematical method for the transformation of three-way data arrays from GC-MS through segmentation of the chromatograms using sums of squares and cross-product (SSCP) matrices. In this

Table 1
The reported alignment algorithms for the GC-MS data processing.

Alignment procedure	Software	Programming interface	Chromatographic data	Files	Platform	Ref.
Non-linear retention time alignment	XCMS and XCMS online	CLI, GUI	LC-MS/GC-MS	NetCDF, mzXML, MzData, mzML	R	[32,46]
Detection and connection of consecutive *m/z* values spanning over multiple scans, performing deconvolution followed by *m/z* alignment	MzMine 2	GUI	GC-MS/LC-MS	Most of GC-MS data format	Java	[33]
Two modes of rough and iterative alignment	MetAlign	CLI	GC-MS/LC-MS	Masslynx.raw, mzData, Xcalibur.raw, NetCDF, mzXML, Agilent.d	C, Visual C++	[31]
Alignment through the deconvolution process	AMDIS	GUI	GC-MS	Most of GC-MS data format	–	[47,48]
Linear interpolation between retention time anchors to calculate retention index; Pearson/ Spearman correlation is applied to find correlated clusters of tags	TagFinder	GUI	GC-EI-MS/GC-TOF-MS	NetCDF	Java	[34]
Ion extraction process from chromatograms based on ion – retention time pairs	MET-IDEA	CLI	GC-MS	NetCDF	Windows, .NET platform	[49]
Peak detection by first derivative of the signal, followed by peak deconvolution, determination of quantification ion; peak aligned based on retention index, shifts between analyses corrected using internal standard	Metabolite Detector	GUI or CLI	GC-MS	NetCDF, JEOL FastFlight2	C++	[50]
Peak detected using XCMS function, application of pseudospectra, a set of co-eluting *m/z*, for peak matching	metaMS	GUI	GC-MS	CDF, mzXML	R	[51]
Peaks identified as retention index – mass spectrum pairs followed by unsupervised clustering methods for the peak assignment based on MS similarities	MSeasy	GUI	GC-MS	netCDF, mzXML, ASCII	R	[52]
Dynamic time warping with an explicitly specified time shift function to align the data by minimising the dynamic programming score, followed by adjustments of signal to compensate for the compression or expansion of the peaks	MathDAMP	GUI	LC–MS, GC–MS, CE–MS	ChemStation.ms, mzXML, netCDF, Analyst.csv	Mathematica	[53]
Peak apex intensities are used for peak picking while retention time index is used for retention time alignment	TargetSearch	GUI or CLI	GC-MS	NetCDF	R	[35]
Retention indices used to minimize retention time shift	MetaQuant	GUI	GC-MS	NetCDF	Java	[54]
Performs deconvolution through identification of local maxima followed by comprising neighbouring apexing ions into the same signal peak. Dynamic programming used for peak alignment	PyMS	CLI	GC-MS	ANDI-MS/NetCDF, JCAMP-DX	Python	[38]
Deep learning neural network algorithm comparison of retention time, chromatogram segment, peak profile and mass spectra of peaks indicating how likely these peaks should be aligned together	ChromAlignNet	CLI	GC-MS	csv	Python	[36]
Dynamic programming-based alignment	flagme	CLI	GC-MS	AMDIS.elu, NetCDF	R	[37]
Parallel computing method for aligning peaks through calculation peak similarity score	QPMASS	CLI	GC-MS	mzXML, netCDF	C++	[39]
A two-step deconvolution based on compound match by local covariance and orthogonal signal deconvolution. eRah computes the Euclidean distance between retention time distance and spectral similarity	eRah	CLI	GC/MS	mzXML	R	[40]
Aligning each sample to reference sample based on individual aligning parameters	PYQUAN	CLI	Pyrolysis-GC/MS	ANDI-MS	Python	[55]
Pairwise alignment using dynamic time warping; matched peaks of already identified compounds used as anchors	ChromA	CLI	GC-MS/LC-MS	netCD, ASTM/AIA/ANDI-MS,xml, mzXML	Java	[56]
Alignment of the MCR-ALS resolved co-eluting peaks based on similarity of mass spectra profiles	MCR-ALS	GUI	GC-MS	cdf, csv	Matlab	[57]
Peaks are defined as mass spectrum and retention time pairs, where heuristic evolving latent projections (HELP) is used for resolving co-eluting peaks; moving sub-window factor analysis (MSWFA) used to locate and verify a component in the chromatogram	MARS	CLI	GC-MS	–	Matlab	[45]
Deconvolution of whole chromatograms based on PARAFAC2 algorithm	PARADISe	GUI	GC-MS	cdf, Matlab	Matlab	[44]

GUI: Graphical User Interface
CLI: Command Line Interface

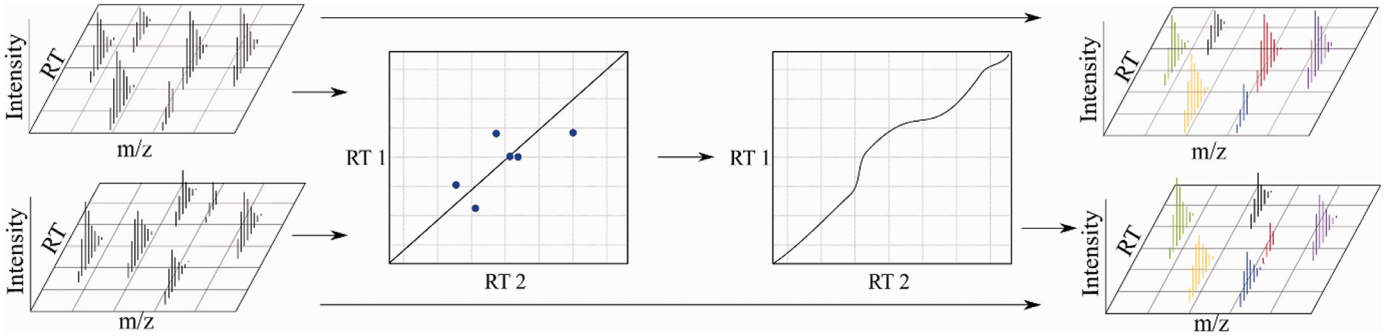

Fig. 3. Schematic representation of a typical warping alignment procedure in which a function is derived from comparison of two runs, where one is a reference analysis (e.g. bottom analysis). The alignment algorithm it is defined a set of global and piecewise adjustments (stretches and shrinks) in retention time axis. Followed by the alignment, corresponding analytes are matched (indicated by matched colours, while black ones do not have a match). Reproduced with permission from [29].

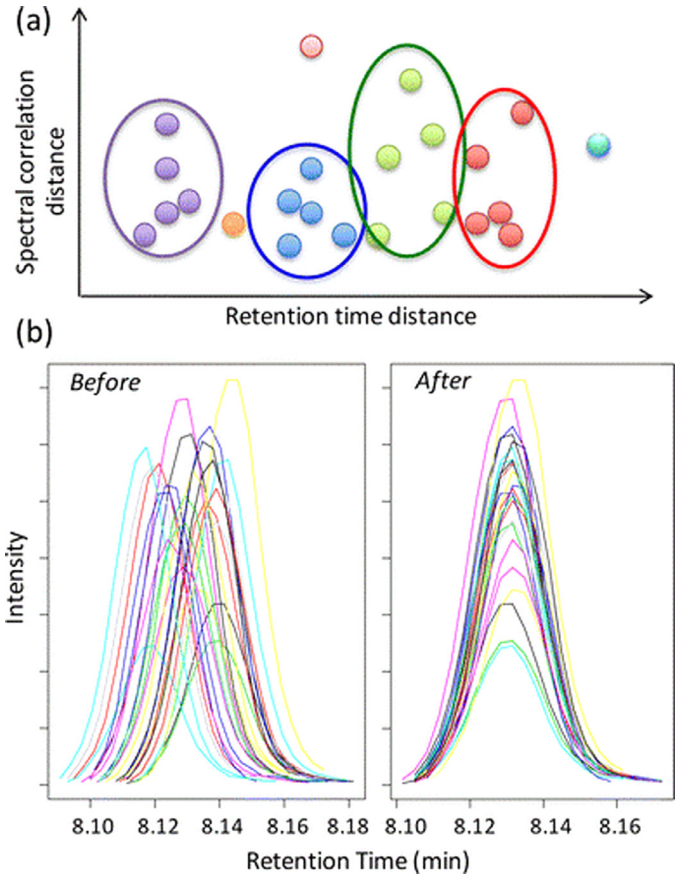

Fig. 4. (a) Representation of eRah alignment algorithm. The analytes are clustered based on the proximity in spectral similarity and retention time distance. In this example there were four different analytes which appear in five different sample, with three analytes found in only one sample and (b) elution profile of an analyte across samples before and after alignment. Reproduced with permission from [40].

way, the bilinear/trilinear data structure (due to the peak shifts) is restored and allows application of multiway resolution methods, which require a strict data trilinearity, e.g. parallel factor analysis (PARAFAC). Multiway resolution algorithms, such as multivariate curve resolution (MCR), PARAFAC, PARAFAC2, and PARAFAC2-based software PARADISe [44] and MARS [45], perform peak alignment and quantification through the deconvolution of the spectral information from the raw three-way data arrays generated during GC-EI-MS analysis. More information about these techniques can be found in Section 3.4.4 (Multiway resolution techniques).

2.4. Data pre-treatment

Variations in the measured response, which are unrelated to the chemical profile differences, are common when numerous samples are analysed. These variations arise from small differences in sample volumes during sample preparation and injection, as well as in instrument performance due to the fluctuations in ionisation, ion transfer and detector efficiency. They are typically the same for all sample components and can be corrected by a normalisation procedure [18]. The most common normalisation procedures apply the sum, mean or the median of the responses of all metabolites across a sample as a normalisation factor. Normalisation can be also performed based on the chosen internal standard [10,18].

Another consideration with the raw dataset (signal) for multivariate statistical analysis is that some important information may be within variables with small relative abundances. This information may be masked by the variance of other variables with higher abundances. Therefore, different scaling methods have been proposed to equalise the potential of each variable to contribute to the created model. The most commonly applied procedure to reduce this effect is autoscaling (standardisation), where all measured values for a variable are divided by its standard deviation. However, this procedure can lead to the inflation of importance of noisy variables with low standard deviations. Pareto and level scaling are techniques which try to equalise the variable contributions by assigning a different weight to each variable [10].

Another signal pre-treatment procedure is element-wise transformation. This differs from scaling as it is performed by modification of individual elements in the data matrix rather than entire variables. The best examples of such transformations are power and log transformations, which simply take the square root or log of each individual data element, respectively. An advantage of these transformations is that they lead to correction of heteroscedastic data, as well as data which do not follow a normal distribution. Also, power and log transformation cause pseudo scaling effect by reducing the difference between large and small values within the dataset. Another transformation method is based upon the rank, where the original values are replaced by their ranks, which consequently decreases the influence of outliers. However, this can hamper interpretation of the information in the created model [58]. The downside of such data scaling and transformation techniques is the inflation of the noisy variables that can diminish the performance of the statistical techniques [18].

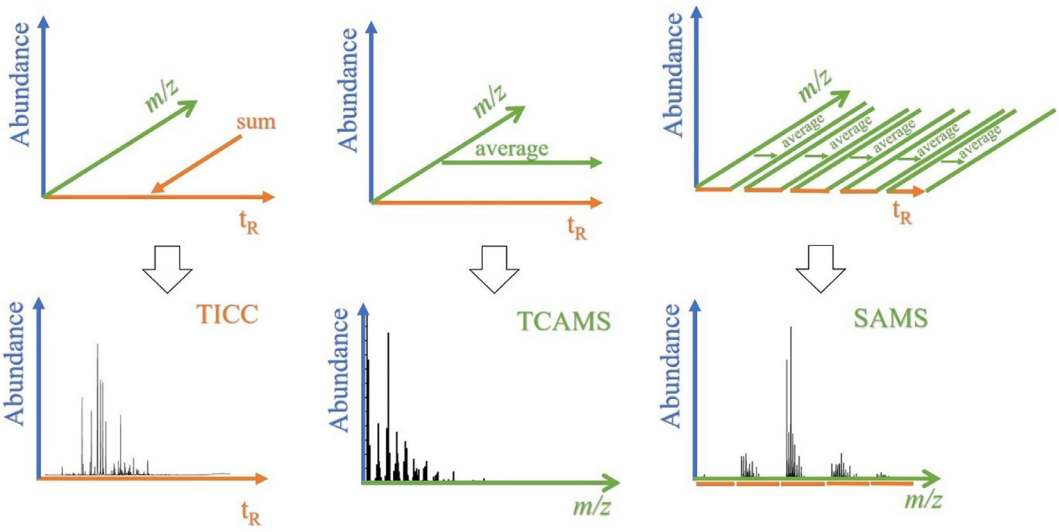

Fig. 5. Reduction of dimensionality of a three-way GC-MS data matrix for geranium EO to TICC, TCAMS, and SAMS datasets.

More about these pre-processing techniques can be found elsewhere [10,58].

2.5. Dimensionality reduction – data size reduction

Application of multivariate statistical approaches on raw full-spectral information recorded during GC-EI-MS analysis is prohibited for two main reasons. Firstly, the size of the dataset from a single EO metabolomic study that would require significant computational power. Secondly, the majority of variables/features do not contain any relevant information and their inclusion is detrimental to data analysis procedures. Reducing the size of such data matrices significantly simplifies their multivariate analysis. This is especially important in the GC-EI-MS analysis, where a 60 min analysis, with mass range of 200 and data acquisition frequency of 10 Hz, can produced 7.2 million data points in centroid mode. The most common approach for such chromatographic data reduction is the generation of chemical profiles, where all *m/z* generated in each scan are summed, creating a TICC (Fig. 5).

Multivariate statistical analysis, in GC-EI-MS based metabolomic studies, are most commonly applied to chemical profiles obtained from TICC, since discriminatory metabolites can be revealed through the data mining process. Due to heterogeneity among samples or the low concentration of some analytes, below detection limits, the compiled chemical composition dataset may contain missing analytes (missing values/features). This problem can be addressed in several ways. Firstly to disregard all the variables with missing data. This may be the optimal solution if the percentage of missing data is relatively low. If the percentage of missing values is higher, simply by disregarding the missing data, the dataset is simplified, with the important information being lost. Second approach is to impute the missing values. However, an arbitrarily chosen algorithm for the missing data imputation can introduce a large bias to the dataset and significantly affect further data analysis and the resulting conclusions [59]. Therefore, the first step in selecting an appropriate missing data imputation algorithm is estimating the nature of the missing values. If it is known that missing values are coming from the quantity of analytes present but below limit of detection, missing data can be replaced with the small arbitrarily chosen numbers. If the origin of the missing data is unknown, an algorithm based on searching local or global similarities, like K nearest neighbour (KNN), should be used. In any case, several algorithms should be compared to avoid introduction of bias into the dataset [59].

A GC-EI-MS data matrix can also be presented as a function of *m/z*, by averaging the intensity over the whole chromatogram, creating a total chromatogram average mass spectrum (TCAMS) data matrix, as depicted in Fig. 5 [60–62]. This allows straightforward compiling of the input data matrix for multivariate statistical analysis, without the necessity for application of peak detection and alignment algorithms, see Table 2, while the final data matrix does not include missing data points. However, the achieved chromatographic separation is not considered in creating such datasets. Application of TCAMS impairs data mining because attributing a *m/z* to a certain compound in the chemical profile can be troublesome, as compounds within EOs belong to only a few chemical classes and have similar fragmentation patterns. On the other hand, utilising TCAMS in data analysis allows different instrumental conditions and temperature gradients to be applied for the analysis [63].

Segmented average mass spectrum (SAMS) exploits the advantages of TCAMS and preserves the level of chromatographic separation. SAMS is created by dividing a GC-EI-MS three-way data array into time dependent sub-windows, where for each the AMS is calculated (Fig. 5). Criteria for the segment determination can be chemical class [63], specific targeted markers present in each sample [64], or alkane sequence [65]. Since this approach does not require strict peak alignment, the number of segments can depend on the complexity of the sample. Taking into account a certain level of chromatographic separation significantly increases the amount of information preserved in the input dataset for subsequent multivariate statistical analysis and increases the model accuracy [63]. Adutwum and Harynuk developed an unique ion filter (UIF) procedure to reduce the number of retained *m/z* for each peak to the most important few, enabling a decrease of variables by 99% and computing time for 98%, while preserving the model accuracy [66]. This algorithm is applied after the chromatogram alignment procedure. Prediction models, created on data obtained through data reduction procedures, have a low risk of model overfit, which may be a problem in feature selection procedures, see Table 2.

2.6. Feature selection in GC-MS studies

In most data reduction approaches three-way GC-EI-MS data arrays are reduced in an unsupervised manner. This can lead to loss

Table 2
Data reduction and feature selection procedures and their characteristics.

	Procedure	Alignment required	Model overfitting possibility	Interactions among variables	Computational time	Dependence on ranking algorithm or data analysis technique
Data reduction	Chemical profile	Yes	No	No	Low	No
	TCAMS	No	No	No	Low	No
	SAMS	Alignment of segments	No	No	Low	No
	Unique filter	Yes	No	No	Medium	No
Feature selection	Filter	Yes	No	No	Low	Yes, ranking algorithm
	Wrapper	Yes	Yes	Yes	High	Yes, data analysis procedure
	Algorithm embedded	Yes	Yes	No	Low	Yes, data analysis procedure
	Hybrid	Yes	No	Yes	Medium	Decreased dependence on ranking and data analysis techniques

of some important information while some of the uninformative features may be preserved. In contrast, different selection strategies enable selection of features containing the relevant information, by eliminating as many irrelevant features as possible. In this way, more robust models with greater accuracy and precision are created. However, the relevance of features is based on the applied data analysis algorithms, and different algorithms may preserve different features, which may lead to different metabolites/features identified as important in the data mining steps.

The feature selection methods that are applied to GC-EI-MS data, require the chromatograms to be aligned in both retention time and *m/z* dimensions. Feature selection algorithms falls into four main categories of 'filter', 'wrapper', 'embedded', and 'hybrid' methods [67]. Filter methods assess the merits of the feature itself, independent of the applied data analysis technique, and they can be based on data similarity, applied statistical methods, or response/class. Similarity-based methods assess the features' importance by their ability to preserve data structures from the original datasets. In the case of supervised feature selection, data similarity can be derived from the class/label information, while most of the unsupervised feature selection methods take advantage of different distance metric measures to obtain the data similarity. The most common similarity-based methods are Laplacian Score, Fisher Score, Trace Ratio Criterion, and ReliefF. Another class of the filter algorithms is based upon different statistical analysis. Most of the statistical-based algorithms analyse the features individually. Hence, feature redundancy is inevitably ignored during the selection phase [68]. The representative algorithms of this class are Low variance, T-score, Chi-square score, and GINI index. Response/class-based methods represent a large family of the existing filter algorithms. Algorithms in this family measure the importance of features, by maximising feature relevance, while minimising feature redundancy. Since the relevance of a feature is usually measured by its correlation with response/class labels, most feature selection approaches of this family are performed in a supervised way [68].

Filter methods are general selection procedures, in which the feature selection is independent of any learning algorithm, making the models less prone to overfitting. They are computationally inexpensive approaches that computes a ranking metric for each feature, representing the importance of the variable [69,70]. A problem with such filter methods is that different ranking metrics provide different feature selections. Also, similarity- and statistical-based methods evaluate the importance of features individually and hence cannot handle feature redundancy, where many correlated features can be present in the final selection. In contrast, most of the response/class-based feature selection algorithms can distinguish between feature relevance and feature redundancy [68]. Filter methods have been also used in hybrid approaches as a preselection step, prior to any other feature selection algorithm. Filter methods do not consider the interactions among the variables, as presented in Table 2. Therefore, in most cases, finding the optimal subset of features through the filter methods is not preferred unless the interactions and additive effects among variables are considered.

Wrapper techniques perform feature selection around an applied multivariate or a machine learning algorithm. Thus, the data analysis approach in the feature selection procedure is used as an evaluation function to help estimation of the worth (merit) of each feature [67]. The feature selection can be performed through stepwise forward selection (FS), or stepwise backward elimination (BE). In the BE procedure, the feature selection process starts with all variables of interest, followed by sequential removal of them. If the removed variable does not have negative impact on the model accuracy, it is removed permanently. In this way, the impact of each variable on the model accuracy is evaluated. Compared to the FS method, BE is computationally more "expensive". However, its significant advantage is that the influence of each variable is evaluated in the context of other variables [67,69]. Models created on features selected through such procedures show greater accuracy, compared to models created on features selected through filter methods. In wrapper algorithms, feature selection is performed to maximise the accuracy of the learning algorithm. However, this means that there is a significant risk of overfitting, since the feature selection is led by the learning algorithm, see Table 2. Therefore, it is of great importance that model validation is performed through double cross-validation (DCV). Wrapper methods require more computation power, compared to filter methods. Therefore, they are rarely applied as the only feature selection procedure in processing of big data.

In embedded methods, the feature selection algorithm is built into the multivariate algorithm. This means that during the learning process, weights are applied to features to achieve the optimal predictions. They use statistical criteria for the feature selection, as per filter methods, with the highest accuracy prediction accuracy, as with wrapper methods [71]. This leads to very good underlying learning algorithm performance, with good model interpretability as it provides explanations for the predictions made. As in filter methods, interactions among features are neglected. In the case of PCA, loading weights can be used to measure the importance of features. Also, in PLS and PLS-DA, variable importance in prediction (VIP) is used to identify the features' importance for the prediction [72].

As with wrapper methods, embedded feature selection approaches can lead to model overfit. Wrapper and embedded algorithms can also lead to an instability in the feature selection. This means that different sample subsets, used for training the model, apply different weights to a similar feature during the process of model creation. In this case, it is hard to obtain reliable importance for each of the variables. This can be overcome by using a more robust sample subset for the model training. Decision tree-based algorithms such as Random Forests create models on numerous dif-

ferent sample and feature subsets, decreasing the possibilities for feature selection instability or model overfit.

To overcome various shortcomings of different individual feature selection methods, hybrid feature selection methods were introduced. Hybrid selection methods aim to construct an aggregated classification of feature importance. This can be done based on different feature selection algorithms or by creating models on different sample and feature sub-datasets. In this way, the shortcomings of the individual feature selection algorithms can be mitigated, enhancing the subsequent learning tasks. As already mentioned, application of the wrapper methods on GC-EI-MS data can demand long feature selection processing. Therefore, to decrease the computation time, filter methods are applied to remove uninformative features, followed by the wrapper methods for further evaluation of the importance of the top-ranked features [69]. Even applying the wrapper methods solely on the top-ranked variables within GC-EI-MS analysis requires significant computation power and time. Sinkov and Harynuk used the SR approach to obtain the top 1000 variables from the aligned GC-EI-MS data unfolded along the time axes. Then, BE was performed on these top 1000 variables, while the variables ranked from 1001 to 3000 in the SR analysis were introduced to a FS algorithm and considered for inclusion [69]. The whole process lasted 36 h.

As already stated, decision tree-based algorithms create numerous prediction models on the random subset of samples and variables. They are far less sensitive to irrelevant features, since feature selection is essential to the tree growing process [74]. Based on this premise, several hybrid approaches have been developed for the evaluation of features' importance and interactions between the variables, leading to a greater model accuracy and sensitivity [75,76]. Abeel *et al.* applied a recursive feature elimination algorithm on 500 sub-datasets, each containing 90% of data, to create a support vector machine (SVM) model. Overall ranking of the features was created based on aggregated rankings of all 500 selected feature subsets and models' accuracy [77]. In a similar approach, based on the predictive capability of models created on many sample and feature sub-datasets, created using the Monte Carlo sampling (MCS) method, features were identified as informative, uninformative, or interfering. This approach was applied in a model population analysis (MPA) [75,76] and within a sub-window permutation analysis (SPA) [78]. In the variable complementary network (VCN) procedure, the complementary nature among features is assessed based on the model accuracy and visualized as a feature network [79]. The downside of such approaches is that importance of variables and their interactions are assessed based on the applied data analysis technique and consequently can be different when different data analysis algorithms are applied. Yun *et al.* [73] used nine different algorithms to rank the variables and create an aggregated rank list. The PLS-DA model created with the top variables from the aggregated list showed a superior sensitivity and specificity compared to the model created on the whole set of variables. Therefore, to obtain more comprehensive understanding of variable importance, several algorithms should be applied. More information about the feature selection and their statistical implications can be found elsewhere [67,68,71,72,80,81].

3. Data analysis

3.1. Pattern recognition and classification statistical techniques

One of the major goals in the EO metabolomic studies is to discover similarity/dissimilarity patterns among samples and the most relevant features (metabolites) across the samples, to perform classification and class comparison. This process can be performed based on variance between the samples, making the process unsupervised, or it can be used with *a priori* knowledge regarding the sample classes, making the pattern recognition and classification supervised [82]. Unsupervised pattern recognition algorithms are applied solely using variance among the samples, when the information regarding the relationship between the samples is not known, or it is desirable to observe the sample clustering without any class influences [82]. The most applied unsupervised pattern recognition algorithms in EO metabolomic studies are PCA and hierarchical cluster analysis (HCA), as can be seen in Table 3.

PCA is the most commonly used unsupervised pattern recognition statistical technique for classification and class comparison, with data mining through data decomposition into scores and loadings vectors. The detailed theory behind the PCA can be found elsewhere [158,159]. Besides being applied as an exploratory pattern recognition technique, PCA has also been applied as a feature reduction tool in the pre-process step and for rank determination [82]. PCA can be applied on the whole chemical profile, obtained from TICC, as well as on the user defined sections of the chromatogram, using all *m/z* fragments or a portion thereof. It also can be used on TCAMS and SAMS datasets [64]. Applying feature selection algorithms it is possible to obtain a set of variables (*m/z*) which will produce highest accuracy in the pattern recognition algorithm. Sinkov *et al.* applied a wrapper feature selection approach on top ranked variables, selected by the SR algorithm, to maximise cluster separation within PCA [69].

In the EO metabolomic studies, PCA has been most commonly applied on chemical composition datasets. In this way, the analytes that discriminate different classes of oils are easily identified from the loadings plots. PCA has been applied for discrimination of EO coming from different species, chemotype and cultivars, different organs and geographical locations, altitudes of growth, and influence of soil, and maturity of the trees (see Table 3). PCA was performed on the chemical composition dataset to differentiate between 7 different *Ocimum* species, using 3 PCs that captured 66% of variance. PC 1 was positively associated with 1,8-cineole and negatively with neral, geranial and nerol, while caryophyllene and β-elemene were highly correlated with the PC 2. The PC 3 was positively correlated with camphor [113]. Elaissi *et al.* [110] managed to differentiate *Eucalyptus* species into 9 chemotypes according to their chemical profiles, while according to the antimicrobial activity (against 4 different bacteria) they were clustered into 4 groups. Tohidi *et al.* [95] classified 14 different *Thymus* species based on their bioactive compounds using PCA. Application of TCAMS and SAMS datasets, instead of chemical profile, led to greater discrimination of different ylang-ylang distillation grades in the PCA. Using the TCAMS, distillation grades were discriminated using the PC 1 and 2, while in case of the SAMS distillation grades were discriminated using solely the PC 1 [64].

HCA is another pattern recognition technique commonly used in the EO metabolomic studies for exploring the similarities on the big datasets, by visualising the variance between the samples across a 2D space. The main aim of HCA is to connect samples with similar features into clusters, forming branches or dendrograms. Radulovic *et al.* applied HCA on the TCAMS datasets generated from GC-EI-MS for differentiation amongst 13 *Artemisia* species [61]. The TCAMS approach showed comparable or even better results compared to HCA applied to the chemical composition dataset.

3.2. Supervised pattern recognition and classification algorithms

Supervised pattern recognition techniques are applied when the information about samples is known *a priori*. Supervised pattern recognition approaches maximise the separation between the classes and have been widely used in EO metabolomic analyses (Table 3). The most applied supervised pattern recognition techniques in the EO metabolomic studies are linear discriminant anal-

Table 3
Application of analytical and pattern recognition techniques in the EO metabolomic studies.

Essential oil	Metabolomic study to differentiate between	Pattern recognition technique	Ref.
Artemisia spp	Species	HCA	[61]
Artemisia saharae Pomel	Altitude of growth	LDA	[83]
Citrus deliciosa Ten. x *C. paradisi* Macf.	Species and hybrids	LDA	[84]
Eucalyptus crebra F.Muell., *Eucalyptus melanophloia* F.Muell.	Species and their hybrids	PCA, LDA	[85]
Thymus algeriensis Boiss. and Reut	Geographical locations	LDA	[86]
Citrus sinensis L. Osbeck and *Citrus limon* L. Burm	Somatic hybrids	LDA	[87]
Somatic hybrids from *Citrus sinensis* L. and *Citrus limon* L. Burm	Somatic hybrids	LDA	[88]
Lavandula angustifolia Mill, *Lavandula x intermedia*	Lavender and lavandin	HCA, LDA	[89]
Rosmarinus officinalis L.	Geographical origin, genetic analysis	PCA	[90]
Rosmarinus officinalis var. *typicus* L. and R*osmarinus officinalis* var. *troglodytorum* L.	Different varieties	PCA	[91]
Houttuynia cordata Thunb.	Geographical locations	PCA	[92]
Chamaemelum nobile L. and *Matricaria recutita* L.	Roman and German chamomiles	PCA, PLS-DA	[93]
Ocimum spp	Different species	PCA	[94]
Thymus spp	Different species	HCA, PCA	[95]
Lippia origanoides Kunth	Chemotypes	PCA	[96]
Hypericum spp	Different species	HCA, PCA	[97]
Thymus serpylloides ssp. gadorensis (Pau) Jalas	Chemotypes	PCA	[98]
Myrtus communis var. *melanocarpa* DC and var. *leucocarpa* DC	Cultivated and wild species	PCA	[99]
Hypericum spp	Different species	PCA	[100]
Litsea cubeba (Lour.) Pers., lemongrass	Two species	PCA	[101]
Thymus capitatus (L.) Cav.	Chemotypes	PCA	[102]
Laurus nobilis L.	Plant organs	PCA	[103]
Mint oils	Geographic origin	HCA, PCA, LDA	[104]
Cinnamomum cassia L., *Cinnamomum verum*, and *cinnamomum loureirii* Nees	Species and habitats	HCA, PCA	[105]
Curcumae longa L.	Geographical origin	HCA, PCA, PLS-DA	[106]
Lavandula luisieri (Rozeira) Riv.-Mart.	Organs and origin	PCA	[107]
Aristolochia spp	Species	PCA	[108]
Centaurea spp	Species	HCA, PCA	[109]
Eucalyptus spp	Species according to antibacterial activity	HCA, PCA	[110]
Eucalyptus spp	Species	PCA	[111]
Cistus salvii folius L.	Chemotypes	PCA	[112]
Ocimum spp	Species according to antibacterial activity	PCA	[113]
Hyptis suaveolens (L.) Poit.	Fruiting stages and geographical variation	PCA	[114]
Chrysanthemum indicum L.	Geographical origin	PCA	[115]
Rosmarinus officinalis L.	Ecological groups	HCA, PCA	[116]
Valeriana jatamansi Jones	Chemotypes	PCA	[117]
Mentha suaveolens ssp. *insularis* (Req.) Greuter, *Mentha suaveolens* ssp. *suaveolens* (L.) Hudson.	Subspecies	PCA	[118]
Coriandrum sativum L.	Fruit age and geographical origin	PCA	[119]
Matricaria recutita L.	Chemotypes	PCA	[120]
Lavandula angustifolia and five Lavandula x hybrida	Chemotypes	PCA	[121]
Smyrnium olusatrum L.	Organs	PCA	[122]
Helichrysum italicum (Roth)G.Don ssp. *italicum*	Geographical origin	HCA, PCA	[123]
Rosmarinus officinalis L.	Geographical origin	HCA, PCA	[124]
Apioideae ssp	Geographical origin	PCA	[125]
Salvia officinalis L.	Geographical origin	HCA, PCA	[126]
Thymus algeriensis Boiss. & Reut.	Chemotypes	PCA	[127]
Satureja montana L. and *Satureja subspicata* Bartl. ex Vis.	Species	PCA	[128]
Humulus lupulus L.	Cultivars	PCA	[129]
Lavandula multifida L.	Arid areas	HCA, PCA	[130]
Juniperus oxycedrus ssp. *oxycedrus* L.	Organs	PCA	[131]
Copaifera multijuga Hayne	Influence of soil and dry land	HCA, PCA	[132]
Chrysopogon zizanioides (L.) Roberty	Geographical location	HCA, PCA	[133]
Eucalyptus globulus Labill. ssp. *globulus*	Different tree's maturity	HCA, PCA	[134]
Perilla frutescens L. (Britton)	Geographical origin	HCA, PCA	[135]
Humulus lupulus L.	Cultivars	HCA, PCA	[136]
Cuminum cyminum L.	Odour attributes identified by sensory profiling	PCA	[137]
Curcuma phaeocaulis Valeton, *C. kwangsiensis* S.G. Lee & C.L. Liang and *C. wenyujin* Y.H. Chen & C. Ling	Three Curcuma species	PCA, PLS-DA	[138]
Panax ginseng C. A. Mey.	Aroma profile at different ages	HCA, PCA	[139]
Citrus nagato-yuzukichi, Citrus junos, Citrus sphaerocarpa, Citrus sudachi	Japanese sour citrus fruits	PCA	[140]
Ocimum vulgare L. subsp. *hirtum* ietswaart (syn.), *O. heracleoticum* Sensu) lines	Oregano subsp. *hirtum* lines	PCA	[141]
Coriandrum sativum L.	Odours	PCA	[142]
Rosa × *damascena* Mill.	Three quality groups	PCA, LDA, SVM	[143]
Artemisia spp	Species	PCA, HCA	[144]
Juniperus rigida Sieb. et Zucc.	Chemotype	PCA, HCA	[145]
Acorus calamus L	Natural and cultivated conditions	PCA	[146]

(*continued on next page*)

Table 3 (*continued*)

Essential oil	Metabolomic study to differentiate between	Pattern recognition technique	Ref.
Borago officinalis L.	Geographical origin	PCA, HCA	[147]
Celastraceae spp	Species and plant organs	PCA	[148]
Vladimiriae Radix and *Aucklandiae Radix*	Two species	PCA, HCA, SIMCA, OPLS-DA	[149]
Achillea spp	Species	PCA	[150]
Premna odorata Blanco	Seasonal variations	PCA	[151]
Pinus sylverstris L.	Geographical origin	PLS-DA	[152]
Amomum villosum Lour., Amomum villosum Lour. var. xanthioides	Two variants	HCA, PCA, PLS-DA	[153]
Xylopia aethiopica (Dunal) A. Rich	Geographical origin	HCA, PCA, PLS-DA	[154]
Calamintha nepeta, Foeniculum vulgare and Ridolfia segetum	Low and high-active antiviral EOs	PLS-DA	[155]
Lavandula angustifolia Mill., *Lavandula* × *intermedia*	Lavender and lavandin	PCA, PLS-DA	[156]
Eucalyptus spp	Different species	PCA, HCA	[157]

ysis (LDA), partial least squares discriminant analysis (PLS-DA), soft independent modelling of class analogy (SIMCA), and the Random Forests (RF) machine learning algorithm. In LDA and PLS-DA algorithms the direction of pseudo-variables (latent variables, LVs) is selected to maximise the separation among the classes, while in SIMCA it is selected to maximise the variance among samples [160]. Supervised pattern recognition techniques can be applied to chemical composition datasets, as well as to TCAMS or SAMS based datasets.

LDA is a linear parametric discriminant technique which places emphasis on dissimilarities between the classes to find their optimal boundaries. Like PCA, LDA is also a dimension reduction approach. Compared to PCA, which takes the first pseudo component to retain maximum variance in the data, in LDA a direction is selected to maximize the separation among the classes [160]. A problem associated with LDA is its requirement that the number of predictor variables is smaller than the number of samples. This can be a challenge, especially in GC-EI-MS metabolomic studies which generate large data arrays. In this case, data reduction and feature selection procedure must be applied prior to the application of LDA.

PLS-DA is an extension of PLS used for the classification purposes [161]. As with PLS, in PLS-DA a response matrix must be defined. In the case of a two-class classification, the dummy variable for one class is 1, while for the other is 0 or -1. If the classification requires separating more than two classes, dummy variables are defined and the PLS2 algorithm is used [160,162,163]. Sabatino *et al*. applied PLS-DA for the discrimination of EOs from three species based on their activity against Herpes simplex virus (HSV-1), as well as their cytotoxicity [155]. Adulteration is a common issue in EO market. For instance, lavender oil (due to its high price) is often adulterated with lavandin oil, with much lower price and inferior fragrant character. Beale *et al.* developed a PLS-DA approach, based on a chemical profile obtained from the TICC from GC-MS analysis, for the discrimination of high- and low-quality lavender EOs [156]. Similarly, Lebanov *et al.* applied a PLS-DA strategy to TCAMS and SAMS datasets for discrimination of different ylang-ylang from various geographical origins [64].

RF is a regression and classification statistical approach for prediction or classification [164], which generates many decision trees, each of them constructed by using different sets of randomly selected input variables. The RF model uses a bootstrapping strategy by random sampling of the original dataset. Gromski *et al.* compared RF with some other discrimination statistical tools, including LDA, PLS-DA, and SVM, and demonstrated the advantages of RF for the classification of different *Bacillus* species based on GC-MS data [165]. Lebanov *et al.* applied a RF algorithm, based on TCAMS and SAMS datasets, for the classification of different EO classes [63]. Calculating the proximity between the samples it was possible to determine EOs with lower qualities. The RF model created on SAMS dataset showed superior performance by identifying more EOs with lower quality, compared to RF model created on TCAMS dataset.

3.3. Predictive modelling

The main objective of predictive modelling is to determine the relationship between variables obtained from instrumental analysis with one or more response variables [9,82]. Response variables that are used for modelling can be abundance of main metabolites, antioxidants, antimicrobial potency, or other characteristics of EOs. Commonly used predictive techniques are principal component regression (PCR) and partial least squares regression (PLS). Both approaches have the ability to deal with numerous variables which contain highly collinear features [166]. In the case of PCR, the direction of pseudo variables (PCs) is set to maximise variance among samples. A shortcoming of PCR is the fact that the maximum variance in the dataset has little or no relevance to the response [9]. In comparison with PCR, PLS reduces data dimensionality by calculating LVs, which are created as the covariance between the response and the matrix of the predictor variables [9]. Cross-validation is an acceptable way to determine the predictive power of created PLS regression models, which will be described in the following sections (Section 4. Model validation).

Prediction models has been rarely used in EO metabolomic studies based on GC-EI-MS analysis. However, by applying PLS on TCAMS and SAMS data, it was possible to determine distillation grade of ylang-ylang EOs and relative abundances of the main compounds. Relative abundances, obtained through PLS model, were used for the quality assessment of ylang-ylang EO based on the ISO standards [64]. In this way, the quantification of main analytes using either internal standards or GC-FID, was not necessary. Also, ylang-ylang grade PLS prediction models created using TCAMS and SAMS data were able to predict grades of EOs from geographical origins not represented in the calibration and validation sample sets, in contrast to models created on chemical profile which were unsuccessful in this task. Maietti *et al.* extracted and identified the EOs of six lavender cultivars *via* headspace GC-EI-MS analysis [121]. By correlating the list of metabolites with biological activities of the EOs (cytotoxicity, antifungal, and antioxidant activities) it was possible to predict biological activity based upon the chemical profile. Park *et al*. developed a PLS model for prediction of free radical scavenging activity (RSA) and reducing power of nine *Mentha* subspecies [167]. By evaluating VIP, shikimic acid, rutin, and succinic acid were identified as the greatest contributors to the RSA, while shikimic acid and succinic acid were the major contributors to the reducing power of these extracts.

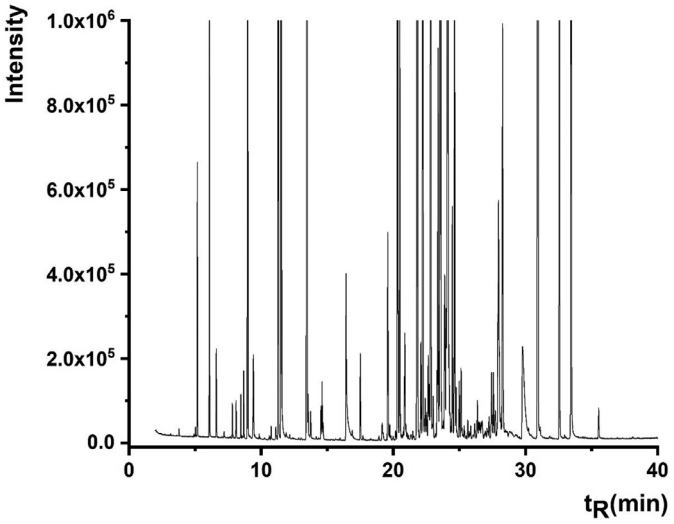

Fig. 6. TICC of ylang-ylang essential oil showing many co-eluting regions. Reproduced with permission and adapted from [179].

3.4. Multiway resolution techniques

The previously described statistical approaches cannot handle the full dimensionality of GC-EI-MS data and therefore require data reduction steps prior to the data analysis. Multiway resolution techniques are able to use the full dimensionality of GC-EI-MS data. One of the most beneficial features of MCR and PARAFAC-like methods is the second order advantage, in which the calibration can be built upon relatively few standards, instead of a large calibration set, with the additional potential to quantify the compounds of interest even in the presence of interferences, not originally present in the calibration standard set. This is performed through decomposition of experimental data array (**D**) to matrices related to mass spectra of pure compounds ($\mathbf{S}^T$), concentration profiles (**C**) and residual matrix (**E**) [18]. This process reveals the MS spectra of the pure analytes which co-elute in GC-EI-MS. Due to the complexity of EOs, with the potential presence of several-hundred compounds [5,168,169], there may be many co-eluting metabolites in the GC-MS separation, as shown in Fig. 6.

Therefore, the application of multiway resolution techniques like MCR, PARAFAC, and TUCKER deliver more accurate chemical profiles, through a deconvolution process, without using additional separations [170]. As the first step in the deconvolution process, it is critical to determine the rank of the examined convoluted system.

3.4.1. Methods for chemical rank determination

The rank of the data matrix is the minimum number of linear independent variables. For an EO sample, from the analytical point of view, it can be defined as the number of pure metabolites in the convoluted chromatogram. For the EO mixtures, such as case in EO blends and perfume mixtures, rank is the number of independent EOs used in the blending procedure. Chemical rank can be determined in different ways such as PCA, eigenvalues, orthogonal projection approach (OPA), morphological score, evolving factor analysis (EFA), singular value decomposition (SVD), etc. EFA has been applied for the determination of local rank in deconvolution of co-eluting analytes in the analysis of lavender [171] and spearmint EOs [172]. EFA and evolving latent projection graphs (ELPGs) were also applied for the rank determination of damask rose EO [173]. Morphological score has been applied to differentiate between noise-containing *m/z* channels from ones containing information relevant to chemistry of the samples [172]. Lebanov *et al.* showed that application of SVD can determine rank in analysis of complex EO blends and perfume mixtures based on TCAMS data [179]. More information regarding the determination of chemical and local rank can be found in the work of Jalali-Heravi and Parastar [24].

3.4.2. Multiway resolution methods

MCR methods were initially developed for dealing with the UV–Vis spectrophotometric data [170]. The basic premise behind MCR algorithms is the application of Beer's law on the multicomponent analysis system, in which the description of variations in the measurement is a linear model of the contributions of their pure components [174]. In order to extract the concentration and spectral profiles from the data matrix using MCR models, it is necessary to use three-way data arrays which contain the whole information regarding the profile and spectra [24]. MCR algorithms extract information of the pure components in a mixture using a bilinear model decomposition. The bilinear MCR model decomposes the experimental data matrix into the product of matrices containing concentrations and pure spectra of individual compounds. According to the Beer's law, MCR assumes that the individual responses (absorbances) of all components are additive. Most of the noise and background effects are retained in the residual **E** matrix. This leads to higher S/N in the recovered pure spectra and allow more accurate identification of co-eluting analytes [170]. The challenges associated with MCR applications are summarised by Ruckebusch and Blanchet [174].

Numerous compounds present in EOs lead to very complex TICCs. Therefore, in EO analysis using MCR algorithms the data matrix is divided into series of subwindows, dependent upon the complexity of the TICC. Each of these parts are analysed separately by determining their local rank and performing the MCR deconvolution, as shown in Fig. 7 [24]. MCR-ALS was applied for deconvolution of co-eluting peaks in the analysis of geranium EO and delivered an increase in the number of determined analytes from 61 to 85 [175]. In the case of damask rose EO, an increase from 70 to 95 was observed [173]. The application of MCR-ALS in EO analysis has also been applied to saffron [176], cumin and caraway EO [177], as well as for *Dracocephalum moldavica* L. EO [178]. In case of lavender EO analysis, the determined relative abundance of the main analyte (α-pinene) was 21.01%, while after application of MCR-ALS it was revealed as 16.51% [171]. MCR-ALS was applied to resolve pure spectra of EOs used in creating complex EO blends and perfume mixtures, based on TCAMS dataset. Resolved pure TCAMS spectra were further successfully applied for EO identification and quality assessment of ylang-ylang EOs used in creating perfume [179].

PARAFAC-based algorithms can be considered as a multiway extension of PCA, within which sets of loading and score tensors (multiway generalisation of pseudo variables) are produced., in place of scores and loadings vectors, produced in PCA. Scores and loadings vectors in PCA are orthogonal to each other, but tensors in PARAFAC are not [18]. In this way, PARAFAC-based approaches can be used as a pattern recognition technique, in a similar way to PCA. Examples include the use of PARAFAC-based algorithms for discrimination between the phytoplankton classes [180], as well as for distinguishing five different species of penicillium [181].

Like MCR algorithms, PARAFAC decomposes trilinear data using constraints to the concentration and spectral matrices [182]. PARAFAC enables resolving overlapped peaks in the datasets by maintaining second order advantage [183,184]. At the same time, it utilises the constraints which ensure data decomposition according to chemical and physical principles [18]. For example, Nikpour *et al.* used PARAFAC on a three-way GC-MS data array for the quantification of 6 compounds with known positive biological activities in *Teucrium polium* L. EO [183].

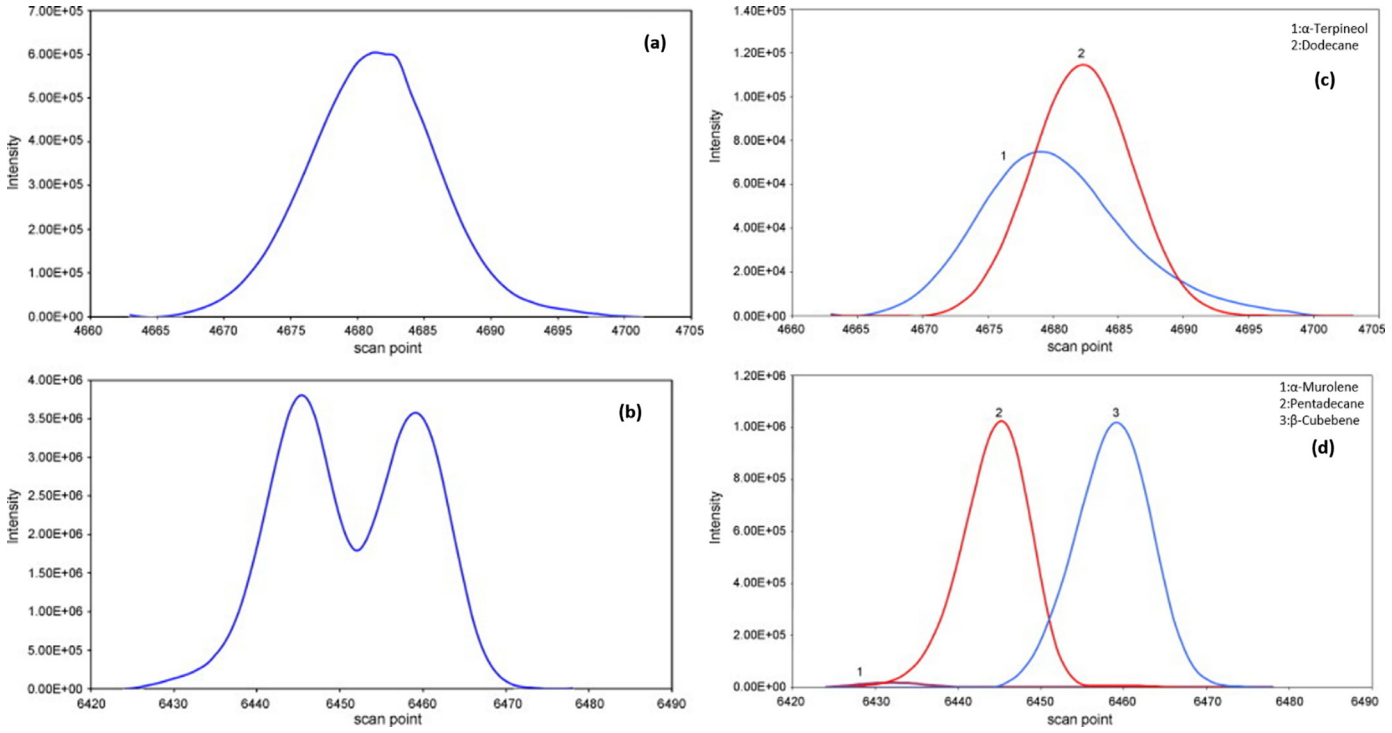

Fig. 7. Pure chromatographic profiles for two peak clusters (a and b) and the resolved peak clusters for three analytes (c and d), resulting from the GC-MS analysis of Iranian damask rose oil. Reproduced with permission from [173].

Application of the multiway approaches result in more accurate chemical profiles, with resolution of all co-eluting peaks. PARAFAC and TUCKER3 require trilinearity, three-way arrays to be aligned in both dimensions (retention time and *m/z*). In contrast, MCR and PARAFAC2 are able to handle small peak shifts between consecutive runs [43,44,184]. Due to the higher similarity of the MCR–ALS underlying model, to the real variations in the chromatographic data, resolved spectra of the co-eluting analytes have higher agreement with the true spectra, compared to the ones resolved with PARAFAC2 [185,186]. More information regarding the comparison of MCR-ALS, TUCKER3, PARAFAC, and PARAFAC2 for deconvolution of concentration and spectral profiles from chromatographic data can be found elsewhere [185]. PARADISe is a software for peak deconvolution based on the PARAFAC2 algorithm which allows the identification of resolved pure MS spectra for each of the peaks in a single run [44].

MARS is a fully automated peak deconvolution algorithm which extracts the chemical composition of the sample set through established mass spectrum and retention time pair, chosen from a reference chromatogram [45]. The eRah algorithm can also perform deconvolution of co-eluting compounds based on a covariance match filter, calculating the local covariance matrix. This allows the spectra of pure compounds to be resolved through an orthogonal signal deconvolution (OSD) procedure and quantification performed via a least absolute deviation (LAD) regression approach [40]. In this way, eRah resolves co-eluting compounds without the estimation of the number of pure compounds, which is a key step in MCR and PARAFAC-based techniques.

4. Model validation

Tuning of parameters, determining the optimal number of pseudo variables in PCA, LDA, PLS, is a crucial step in model building. In the case of PCA, a scree plot is used for the determination of the optimal number of PCs. A scree plot displays the eigenvalues (explained variance) for each PC in descending order versus the number of components. The optimal number of PCs can be obtained (estimated) by taking the last PC at which point little additional variance is captured by the PCs, which is often indicated by a so-called 'knee' in the plot. In this way, the model discriminates between the variance coming from the differences among the samples and the variance coming from the experimental and instrumental conditions (noise).

In case of PLS, PCR, SIMCA, LDA, etc., cross-validation (CV) is the most applied strategy in model tuning, due to its ability to select the model based on prediction capability on the validation dataset. In the CV procedure, the sample set is divided into the training and validation sub-datasets. The model is created on the training sub-dataset, while optimal number or pseudo variables are determined on the validation set, to obtain the highest accuracy. In the CV procedure, model tuning and model validation are performed in one step, on the validation sub-dataset. One of the simplest ways of CV is to divide sample set into two subsets (2/1: training/validation sets). The validation set is used to assess the goodness of the prediction (Q^2) of the model created on the training dataset, which performance is assessed through R^2. R^2 and Q^2 follow the same trend from 0 to 1, by adding each new pseudo variable. When the model starts to overfit, with addition of every next pseudo variable, R^2 keeps increasing towards 1, while Q^2 starts to decrease. At the point of divergence, the model reaches its optimal performance (accuracy). CV can also be performed in an n-fold model, where the dataset is randomly split into "n" mutually exclusive subsets (the folds) of the same size. A special case in CV is leave-one-out CV (LOO-CV), where one sample is held as the validation set and the model is created on the rest of the samples. This whole process is repeated until all samples are used in the validation set. The Q^2 values are averaged for all prediction models to calculate overall Q^2. This method has been used by different researchers, especially in the case of datasets with small sample size, since all samples are used in training and validation sets. Although, this approach often provides accurate model evaluation, there is a risk of model overfitting when the number of variables/features significantly ex-

ceeds the number of samples, which is very common in GC-EI-MS EO metabolomic studies [187].

In the DCV model, tuning and model validation are performed in two different steps. Firstly, model tuning is carried out using the validation sub-dataset, while the test set is used for model validation. The test set is an independent dataset neither used in training nor in tuning of the model. This means that the validation set is used only to optimize/tune the model and cannot be used for the assessment of the model accuracy [188]. By performing the DCV numerous times, every time resampling the training, validation and test sub-datasets, the calculated overall accuracy is adjusted for any potential bias occurred during the sampling process [63]. Sampling to validation and test sub-datasets can be performed either randomly, such as repeated random Monte Carlo sampling [189] or using the developed algorithms for stratified sampling like Duplex [190], Kennard-and-Stone [191], and SPXY algorithms [192].

In case of supervised classification models, such as PLS-DA, SIMCA, and Random Forests, several parameters have been developed for the determination of model performance. In a permutation test, class labels for each sample are randomly permuted, followed by calculating the statistics of such models. After numerous repetitions, differences between starting "right" classification and all other "wrong" models should be statistically different [162,193]. For the classification models, "*sensitivity*" and "*specificity*" should also be calculated, as the two parameters reflect the accuracy of the classification [162,193]. *Sensitivity* is the ratio of predicted true positives to total positives for modelled class, while *specificity* reflects the ratio of predicted true negatives to total negatives, as non-members of the modelled class. For the multiple class models, a confusion matrix can be created to visualize the performance of a prediction model. Each row in the confusion matrix presents the predicted class, while each column represents the actual class. Additional information regarding the method validation can be found elsewhere [162,188,193,194].

5. Workflow software

Over the years, several software packages have been developed, which cover all stages of statistical analysis including preprocessing, pre-treatment, and data analysis. The developed software are designed so that whole data processing is carried out using a single tool, leading to greater reproducibility and the possibility of comparison of different works. Although, most of the developed softwares are primarily developed for LC-MS, several of them have been successfully applied as workflow tools for GC-MS data handling. Workflow4metabolomics is a tool for processing LC-MS and GC-MS data. It comprises XCMS (data pre-processing), CAMERA (metabolite annotation), and ropls package (data analysis) tools. XCMS Online is an online implementation of XCMS [32,46], which provides predefined parameters sets for different instrumental setups, as well as limited number of data analysis techniques. It is linked to the METLIN database [195] and allows metabolite annotation. MetaboAnalyst 4.0 is another online software that implements the XCMS algorithms for raw data pre-processing, supporting only the most commonly used parameters [196]. MetaboAnalyst 4.0 is more focused on statistical, enrichment, and pathway analysis. It consists of eight data analysis modules set into three independent units (exploratory statistical analysis, functional analysis, and advanced methods for translational studies). In all of these software programs data reduction is based on creating the chemical profiles from the TICC. In MetaboAnalyst 4.0 the feature selection is based on the embedded algorithms, as well as on several filter statistical techniques.

6. Looking into the future

As it can be concluded from this review, there is no single platform which has included all elements of data handling and data analysis for GC-EI-MS metabolomic studies into a single system. Therefore, a challenging goal for researchers would be developing a single robust and reliable data handling and data analysis software for the processing of GC-EI-MS data. This would provide easier data handling, greater transparency, and better comparability among different metabolomic studies.

Application of multiple analytical platforms for the complementary analysis of EO metabolomic data has rarely been reported. Since 2005 there has been an increase in attention applied to the problems of data fusion and data analysis on multiblock systems [197–199]. Therefore, one of the greatest challenges in the future will be merging such big amounts of information, applying the most efficient data reduction and feature selection algorithms, as well as proper data mining and interpretations strategies.

Another challenging issue for the future will be dealing with applicability of prediction models. EOs are complex samples and their chemical composition depends on abiotic (environmental) variations. A great number of prediction and classification models have been created, and applied, on a limited number of EO samples, which incorporated limited variability. Therefore, the performance of such models for the evaluation of EOs on different aspects, such are different seasons and geographical origins, has only been tested in a few cases.

More effective applications of artificial intelligence, with machine learning algorithms as the automated systems, in the EO quality control is also expected. This will allow greater sample throughput, as well as greater accuracy in detecting fraudulent samples. With regard to the current extensive/progressive developments in metabolomic research, it is expected that the future will bring innovative solutions to these issues.

7. Conclusion

Metabolomic studies have demonstrated their crucial role in discovering the effects of different genetic predisposition and abiotic factors on the profiles of different EOs, where GC-EI-MS analysis combined with multivariate statistical techniques has found a primary role. This review has aimed to provide an overview of the complexity and the potential pitfalls in data handling processes of GC-EI-MS analysis to obtain unbiased datasets for data analysis and data mining steps. The emphasize was put on recently developed feature alignment algorithms, specifically those approaches designed for the GC-EI-MS data reduction and feature selection, as well as on the evaluation of their advantages and downsides when applied in EO metabolomic studies. In closing this review, a few concluding remarks can be stated as follows:

- Appropriate data pre-processing strategies can compensate the effects of instrumental artefacts in data analysis and data mining steps.
- Alignment of GC-MS data corresponding complex samples, like EOs and their mixture, is still very challenging.
- Application of inappropriate data pre-treatment can impair data analysis and data mining steps.
- Practitioners should be mindful regarding the advantages and drawbacks of each data reduction procedure.
- Application of multiple feature selection procedures allows for creating models with greater accuracy and robustness.
- Application of multivariate statistical techniques and artificial intelligence can simplify and increase robustness of essential oil characterisation.

- Application of double cross-validation can provide the most accurate prediction capacity of created models.
- Development of a single/comprehensive software for data handling and data analysis of GC-EI-MS data is greatly needed.

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.

Acknowledgment

Authors would like to thank the University of Tasmania and PALs Research Hub for their generous provision of co-funded PhD scholarship and project support for Leo Lebanov.

Funding

This work was supported (grant number: IH170100020) by the ARC Industrial Transformation Research Hub for Processing Advanced Lignocellulosics (PALS).

References

[1] M. Katajamaa, M. Orešič, Data processing for mass spectrometry-based metabolomics, J. Chromatogr. A 1158 (2007) 318–328, doi:10.1016/j.chroma.2007.04.021.
[2] C. Bicchi, E. Liberto, M. Matteodo, B. Sgorbini, L. Mondello, B.d.A. Zellner, R. Costa, P. Rubiolo, Quantitative analysis of essential oils: A complex task, Flavour Fragr. J. 23 (2008) 382–391, doi:10.1002/ffj.1905.
[3] C. Bicchi, C. Cagliero, P. Rubiolo, New trends in the analysis of the volatile fraction of matrices of vegetable origin: A short overview. A review, Flavour Fragr, J., 26 (2011) 321–325, doi:10.1002/ffj.2059.
[4] A.C. Figueiredo, J.G. Barroso, L.G. Pedro, J.J.C. Scheffer, Factors affecting secondary metabolite production in plants: Volatile components and essential oils, Flavour Fragr. J. 23 (2008) 213–226, doi:10.1002/ffj.1875.
[5] P. Marriott, R. Shellie, C. Cornwell, Gas chromatographic technologies for the analysis of essential oils, J. Chromatogr. A 936 (2001) 1–22, doi:10.1016/S0021-9673(01)01314-0.
[6] K.M. Namara, J. Howell, Y. Huang, A. Robbat Jr., Analysis of gin essential oil mixtures by multidimensional and one-dimensional gas chromatography/mass spectrometry with spectral deconvolution, J. Chromatogr. A 1164 (2007) 281–290, doi:10.1016/j.chroma.2007.07.042.
[7] P. Rubiolo, B. Sgorbini, E. Liberto, C. Cordero, C. Bicchi, Essential oils and volatiles: sample preparation and analysis. A review, Flavour Fragr, J., 25 (2010) 282–290, doi:10.1002/ffj.1984.
[8] C. Cordero, P. Rubiolo, L. Cobelli, G. Stani, A. Miliazza, M. Giardina, R. Firor, C. Bicchi, Potential of the reversed-inject differential flow modulator for comprehensive two-dimensional gas chromatography in the quantitative profiling and fingerprinting of essential oils of different complexity, J. Chromatogr. A 1417 (2015) 79–95, doi:10.1016/j.chroma.2015.09.027.
[9] T. Rajalahti, O.M. Kvalheim, Multivariate data analysis in pharmaceutics: A tutorial review, Int. J. Pharm. 417 (2011) 280–290, doi:10.1016/j.ijpharm.2011.02.019.
[10] J. Engel, J. Gerretzen, E. Szymańska, J.J. Jansen, G. Downey, L. Blanchet, L.M.C. Buydens, Breaking with trends in pre-processing? TrAC, Trends Anal. Chem. 50 (2013) 96–106, doi:10.1016/j.trac.2013.04.015.
[11] H.J. Issaq, Q.N. Van, T.J. Waybright, G.M. Muschik, T.D. Veenstra, Analytical and statistical approaches to metabolomics research, J. Sep. Sci. 32 (2009) 2183–2199, doi:10.1002/jssc.200900152.
[12] M. Brown, W.B. Dunn, D.I. Ellis, R. Goodacre, J. Handl, J.D. Knowles, S. O'Hagan, I. Spasić, D.B. Kell, A metabolome pipeline: From concept to data to knowledge, Metabolomics 1 (2005) 39–51, doi:10.1007/s11306-005-1106-4.
[13] K. Dettmer, P.A. Aronov, B.D. Hammock, Mass spectrometry-based metabolomics, Mass Spectrom. Rev. 26 (2007) 51–78, doi:10.1002/mas.20108.
[14] J. Trygg, E. Holmes, T. Lundstedt, Chemometrics in metabonomics, J. Proteome Res. 6 (2007) 469–479, doi:10.1021/pr060594q.
[15] O. Taketo, A. Farit Mochamad, A.-U.-A. Md, T. Hiroki, N. Kensuke, K. Shigehiko, Metabolomics of medicinal plants: The importance of multivariate analysis of analytical chemistry data, Curr. Comput. Aided Drug Des. 6 (2010) 179–196, doi:10.2174/157340910791760055.
[16] S. Lamichhane, P. Sen, A.M. Dickens, T. Hyötyläinen, M. Orešič, An overview of metabolomics data analysis: Current tools and future perspectives, Data analysis for omic sciences: Methods and applications (2018) 387–413, doi:10.1016/bs.coac.2018.07.001.
[17] L. Yi, N. Dong, Y. Yun, B. Deng, D. Ren, S. Liu, Y. Liang, Chemometric methods in data processing of mass spectrometry-based metabolomics: A review, Anal. Chim. Acta 914 (2016) 17–34, doi:10.1016/j.aca.2016.02.001.
[18] K.H. Liland, Multivariate methods in metabolomics – from pre-processing to dimension reduction and statistical analysis, TrAC, Trends Anal. Chem. 30 (2011) 827–841, doi:10.1016/j.trac.2011.02.007.
[19] H. Kanani, P.K. Chrysanthopoulos, M.I. Klapa, Standardizing GC-MS metabolomics, J. Chromatogr. B Analyt. Technol. Biomed. Life Sci. 871 (2008) 191–201, doi:10.1016/j.jchromb.2008.04.049.
[20] M.M.W.B. Hendriks, F.A.v. Eeuwijk, R.H. Jellema, J.A. Westerhuis, T.H. Reijmers, H.C.J. Hoefsloot, A.K. Smilde, Data-processing strategies for metabolomics studies, TrAC, Trends Anal. Chem. 30 (2011) 1685–1698, doi:10.1016/j.trac.2011.04.019.
[21] J. Boccard, J.L. Veuthey, S. Rudaz, Knowledge discovery in metabolomics: An overview of MS data handling, J. Sep. Sci. 33 (2010) 290–304, doi:10.1002/jssc.200900609.
[22] 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, doi:10.1016/j.trac.2008.05.011.
[23] Y. Wang, M. Gu, The concept of spectral accuracy for MS, Anal. Chem. 82 (2010) 7055–7062, doi:10.1021/ac100888b.
[24] M. Jalali-Heravi, H. Parastar, Recent trends in application of multivariate curve resolution approaches for improving gas chromatography-mass spectrometry analysis of essential oils, Talanta 85 (2011) 835–849, doi:10.1016/j.talanta.2011.05.045.
[25] V.J. Barclay, R.F. Bonner, I.P. Hamilton, Application of wavelet transforms to experimental spectra: smoothing, denoising, and data set compression, Anal. Chem. 69 (1997) 78–90, doi:10.1021/ac960638m.
[26] R.J. Barnes, M.S. Dhanoa, S.J. Lister, Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra, Appl. Spectrosc. 43 (1989) 772–777, doi:10.1366/0003702894202201.
[27] P.H.C. Eilers, Parametric time warping, Anal. Chem. 76 (2004) 404–411, doi:10.1021/ac034800e.
[28] R. Danielsson, D. Bylund, K.E. Markides, Matched filtering with background suppression for improved quality of base peak chromatograms and mass spectra in liquid chromatography–mass spectrometry, Anal. Chim. Acta 454 (2002) 167–184, doi:10.1016/S0003-2670(01)01574-4.
[29] R. Smith, D. Ventura, J.T. Prince, LC-MS alignment in theory and practice: A comprehensive algorithmic review, Brief Bioinform 16 (2015) 104–117, doi:10.1093/bib/bbt080.
[30] T.G. Bloemberg, J. Gerretzen, H.J.P. Wouters, J. Gloerich, M. van Dael, H.J.C.T. Wessels, L.P. van den Heuvel, P.H.C. Eilers, L.M.C. Buydens, R. Wehrens, Improved parametric time warping for proteomics, Chemometr. Intell. Lab. Syst. 104 (2010) 65–74, doi:10.1016/j.chemolab.2010.04.008.
[31] A. Lommen, MetAlign: Interface-driven, versatile metabolomics tool for hyphenated full-scan mass spectrometry data preprocessing, Anal. Chem. 81 (2009) 3079–3086, doi:10.1021/ac900036d.
[32] 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, doi:10.1021/ac051437y.
[33] 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 Bioinformatics 11 (2010) 395, doi:10.1186/1471-2105-11-395.
[34] 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 (Oxford, England) 24 (2008) 732–737, doi:10.1093/bioinformatics/btn023.
[35] Á. Cuadros-Inostroza, C. Caldana, H. Redestig, M. Kusano, J. Lisec, H. Peña-Cortés, L. Willmitzer, M.A. Hannah, TargetSearch - A Bioconductor package for the efficient preprocessing of GC-MS metabolite profiling data, BMC Bioinformatics 10 (2009) 428, doi:10.1186/1471-2105-10-428.
[36] M. Li, X.R. Wang, Peak alignment of gas chromatography-mass spectrometry data with deep learning, J. Chromatogr. A 1604 (2019) 460476, doi:10.1016/j.chroma.2019.460476.
[37] M. Robinson, flagme: Fragment-level analysis of GC-MS-based metabolomics data, (2020).
[38] 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 Bioinformatics 13 (2012) 115, doi:10.1186/1471-2105-13-115.
[39] L. Duan, A. Ma, X. Meng, G.A. Shen, X. Qi, QPMASS: A parallel peak alignment and quantification software for the analysis of large-scale gas chromatography-mass spectrometry (GC-MS)-based metabolomics datasets, J. Chromatogr. A 1620 (2020) 460999, doi:10.1016/j.chroma.2020.460999.
[40] X. Domingo-Almenara, J. Brezmes, M. Vinaixa, S. Samino, N. Ramirez, M. Ramon-Krauel, C. Lerin, M. Diaz, L. Ibanez, 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, doi:10.1021/acs.analchem.6b02927.
[41] Y. Koh, K.K. Pasikanti, C.W. Yap, E.C. Chan, Comparative evaluation of software for retention time alignment of gas chromatography/time-of-flight mass spectrometry-based metabonomic data, J. Chromatogr. A 1217 (2010) 8308–8316, doi:10.1016/j.chroma.2010.10.101.
[42] W. Niu, E. Knight, Q. Xia, B.D. McGarvey, Comparative evaluation of eight software programs for alignment of gas chromatography-mass spectrometry

chromatograms in metabolomics experiments, J. Chromatogr. A 1374 (2014) 199–206, doi:10.1016/j.chroma.2014.11.005.
[43] J. Vestner, G. de Revel, S. Krieger-Weber, D. Rauhut, M. du Toit, A. de Villiers, Toward automated chromatographic fingerprinting: A non-alignment approach to gas chromatography mass spectrometry data, Anal. Chim. Acta 911 (2016) 42–58, doi:10.1016/j.aca.2016.01.020.
[44] 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, doi:10.1016/j.chroma.2017.04.052.
[45] P. Ma, Z. Zhang, X. Zhou, Y. Yun, Y. Liang, H. Lu, Feature extraction from resolution perspective for gas chromatography-mass spectrometry datasets, RSC Advances 6 (2016) 113997–114004, doi:10.1039/c6ra17864b.
[46] R. Tautenhahn, G.J. Patti, D. Rinehart, G. Siuzdak, XCMS Online, A web-based platform to process untargeted metabolomic data, Anal. Chem. 84 (2012) 5035–5039, doi:10.1021/ac300698c.
[47] T. Davis, The new automated mass spectrometry deconvolution and identification system (AMDIS), Spectrosc. Eur. 10/3 (1998) 24–27.
[48] 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, doi:10.1021/jasms.8b01368.
[49] 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, doi:10.1021/ac0521596.
[50] 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, doi:10.1021/ac802689c.
[51] R. Wehrens, G. Weingart, F. Mattivi, metaMS: An open-source pipeline for GC-MS-based untargeted metabolomics, J. Chromatogr. B Analyt. Technol. Biomed. Life Sci. 966 (2014) 109–116, doi:10.1016/j.jchromb.2014.02.051.
[52] F. Nicole, Y. Guitton, E.A. Courtois, S. Moja, L. Legendre, M. Hossaert-McKey, MSeasy: Unsupervised and untargeted GC-MS data processing, Bioinformatics (Oxford, England) 28 (2012) 2278–2280, doi:10.1093/bioinformatics/bts427.
[53] R. Baran, H. Kochi, N. Saito, M. Suematsu, T. Soga, T. Nishioka, M. Robert, M. Tomita, MathDAMP: A package for differential analysis of metabolite profiles, BMC Bioinformatics 7 (2006) 530, doi:10.1186/1471-2105-7-530.
[54] B. Bunk, M. Kucklick, R. Jonas, R. Munch, M. Schobert, D. Jahn, K. Hiller, MetaQuant: A tool for the automatic quantification of GC/MS-based metabolome data, Bioinformatics (Oxford, England) 22 (2006) 2962–2965, doi:10.1093/bioinformatics/btl526.
[55] M.M. Smits, R. Carleer, J.V. Colpaert, PYQUAN: A rapid workflow around the AMDIS deconvolution software for high throughput analysis of pyrolysis GC/MS data, J. Anal. Appl. Pyrolysis 118 (2016) 335–342, doi:10.1016/j.jaap.2016.01.006.
[56] N. Hoffmann, J. Stoye, ChromA: signal-based retention time alignment for chromatography-mass spectrometry data, Bioinformatics (Oxford, England) 25 (2009) 2080–2081, doi:10.1093/bioinformatics/btp343.
[57] J. Jaumot, A. de Juan, R. Tauler, MCR-ALS GUI 2.0: New features and applications, Chemometr. Intell. Lab. Syst. 140 (2015) 1–12, doi:10.1016/j.chemolab.2014.10.003.
[58] R.A. van den Berg, H.C. Hoefsloot, J.A. Westerhuis, A.K. Smilde, M.J. van der Werf, Centering, scaling, and transformations: improving the biological information content of metabolomics data, BMC Genomics 7 (2006) 142, doi:10.1186/1471-2164-7-142.
[59] O. Hrydziuszko, M.R. Viant, Missing values in mass spectrometry based metabolomics: an undervalued step in the data processing pipeline, Metabolomics 8 (2011) 161–174, doi:10.1007/s11306-011-0366-4.
[60] M.E. Sigman, M.R. Williams, J.A. Castelbuono, J.G. Colca, C.D. Clark, Ignitable liquid classification and identification using the summed-ion mass spectrum, Instrum. Sci. Technol. 36 (2008) 375–393, doi:10.1080/10739140802151440.
[61] N.S. Radulović, P.D. Blagojević, Average mass scan of the total ion chromatograms: A new gas chromatography–mass spectrometry derived variable for fast and reliable multivariate statistical treatment of essential oil compositional data, J. Chromatogr. A 1301 (2013) 190–199, doi:10.1016/j.chroma.2013.05.039.
[62] L.A. Adutwum, R.J. Abel, J. Harynuk, Total ion spectra versus segmented total ion spectra as preprocessing tools for gas chromatography - mass spectrometry data, J. Forensic Sci. (2017), doi:10.1111/1556-4029.13657.
[63] L. Lebanov, L. Tedone, A. Ghiasvand, B. Paull, Random Forests machine learning applied to gas chromatography – mass spectrometry derived average mass spectrum data sets for classification and characterisation of essential oils, Talanta 208 (2020) 120471, doi:10.1016/j.talanta.2019.120471.
[64] L. Lebanov, S. Chatterjee, L. Tedone, S.C. Chapman, M.R. Linford, B. Paull, Comprehensive characterisation of ylang-ylang essential oils according to distillation time, origin, and chemical composition using a multivariate approach applied to average mass spectra and segmented average mass spectral data, J. Chromatogr. A 1618 (2020) 460853, doi:10.1016/j.chroma.2020.460853.
[65] N.A. Sinkov, B.M. Johnston, P.M. Sandercock, J.J. Harynuk, Automated optimization and construction of chemometric models based on highly variable raw chromatographic data, Anal. Chim. Acta 697 (2011) 8–15, doi:10.1016/j.aca.2011.04.029.
[66] L.A. Adutwum, J.J. Harynuk, Unique ion filter: a data reduction tool for GC/MS data preprocessing prior to chemometric analysis, Anal. Chem. 86 (2014) 7726–7733, doi:10.1021/ac501660a.
[67] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, J. Mach. Learn. Res. 3 (2003) 1157–1182.
[68] J. Li, K. Cheng, S. Wang, F. Morstatter, R.P. Trevino, J. Tang, H. Liu, Feature selection, ACM Comput. Surv. 50 (2018) 1–45, doi:10.1145/3136625.
[69] N.A. Sinkov, J.J. Harynuk, Three-dimensional cluster resolution for guiding automatic chemometric model optimization, Talanta 103 (2013) 252–259, doi:10.1016/j.talanta.2012.10.040.
[70] N.A. Sinkov, J.J. Harynuk, Cluster resolution: a metric for automated, objective and optimized feature selection in chemometric modeling, Talanta 83 (2011) 1079–1087, doi:10.1016/j.talanta.2010.10.025.
[71] J. Tang, S. Alelyani, H. Liu, Feature selection for classification: A review, in: Data classification: algorithms and applications, 2014, pp. 37–64, doi:10.1201/b17320.
[72] T. Mehmood, K.H. Liland, L. Snipen, S. Sæbø, A review of variable selection methods in partial least squares regression, Chemometr. Intell. Lab. Syst. 118 (2012) 62–69, doi:10.1016/j.chemolab.2012.07.010.
[73] Y.H. Yun, B.C. Deng, D.S. Cao, W.T. Wang, Y.Z. Liang, Variable importance analysis based on rank aggregation with applications in metabolomics for biomarker discovery, Anal. Chim. Acta 911 (2016) 27–34, doi:10.1016/j.aca.2015.12.043.
[74] V. Svetnik, A. Liaw, C. Tong, J.C. Culberson, R.P. Sheridan, B.P. Feuston, Random Forest: A classification and regression tool for compound classification and QSAR modeling, J. Chem. Inf. Comput. Sci. 43 (2003) 1947–1958, doi:10.1021/ci034160g.
[75] H. Li, Y. Liang, Q. Xu, D. Cao, Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration, Anal. Chim. Acta 648 (2009) 77–84, doi:10.1016/j.aca.2009.06.046.
[76] H.-D. Li, Y.-Z. Liang, Q.-S. Xu, D.-S. Cao, Model population analysis for variable selection, J. Chemom. 24 (2010) 418–423, doi:10.1002/cem.1300.
[77] T. Abeel, T. Helleputte, Y. Van de Peer, P. Dupont, Y. Saeys, Robust biomarker identification for cancer diagnosis with ensemble feature selection methods, Bioinformatics (Oxford, England) 26 (2010) 392–398, doi:10.1093/bioinformatics/btp630.
[78] H.-D. Li, M.-M. Zeng, B.-B. Tan, Y.-Z. Liang, Q.-S. Xu, D.-S. Cao, Recipe for revealing informative metabolites based on model population analysis, Metabolomics 6 (2010) 353–361, doi:10.1007/s11306-010-0213-z.
[79] H.-D. Li, Q.-S. Xu, W. Zhang, Y.-Z. Liang, Variable complementary network: a novel approach for identifying biomarkers and their mutual associations, Metabolomics 8 (2012) 1218–1226, doi:10.1007/s11306-012-0410-z.
[80] G. Chandrashekar, F. Sahin, A survey on feature selection methods, Comput. Elec. Eng. 40 (2014) 16–28, doi:10.1016/j.compeleceng.2013.11.024.
[81] S. Alelyani, J. Tang, H. Liu, Feature Selection for Clustering: A Review, in: Data clustering:Algorithms and Applications, 2018, pp. 29–60.
[82] S.E. Prebihalo, K.L. Berrier, C.E. Freye, H.D. Bahaghighat, N.R. Moore, D.K. Pinkerton, R.E. Synovec, Multidimensional gas chromatography: Advances in instrumentation, chemometrics, and applications, Anal. Chem. 90 (2018) 505–532, doi:10.1021/acs.analchem.7b04226.
[83] S. Zouari, I. Ayadi, N. Fakhfakh, H. Jdir, L. Aloui, M. Kossentini, A. Rebai, N. Zouari, Essential oil variation in wild populations of Artemisia saharae (Asteraceae) from Tunisia: chemical composition, antibacterial and antioxidant properties, Bot. Stud. 55 (2014), doi:10.1186/s40529-014-0076-0.
[84] G. Ruberto, D. Biondi, P. Rapisarda, A. Renda, A. Starrantino, Essential oil of Cami, a new Citrus hybrid, J. Agric. Food Chem. 45 (1997) 3206–3210, doi:10.1021/jf960768w.
[85] J.J. Brophy, P.I. Forster, R.J. Goldsack, D.B. Hibbert, A. Punruckvong, Essential oil variation in Eucalyptus crebra, E. melanophloia (Myrtaceae) and their hybrids, Aus. J. Bot. 57 (2009) 425–431, doi:10.1071/BT08171.
[86] N. Zouari, I. Ayadi, N. Fakhfakh, A. Rebai, S. Zouari, Variation of chemical composition of essential oils in wild populations of Thymus algeriensis Boiss. et Reut., a North African endemic Species, Lipids Health Dis 11 (2012), doi:10.1186/1476-511X-11-28.
[87] S. Fatta Del Bosco, L. Abbate, N. Tusa, T. Strano, A. Renda, G. Ruberto, Genetic improvement of Citrus fruits: The essential oil profiles in a Citrus limon backcross progeny derived from somatic hybridization, Food Res. Int. 50 (2013) 344–350, doi:10.1016/j.foodres.2012.10.041.
[88] L. Abbate, N. Tusa, S. Fatta Del Bosco, T. Strano, A. Renda, G. Ruberto, Genetic improvement of Citrus fruits: New somatic hybrids from Citrus sinensis (L.) Osb. and Citrus limon (L.) Burm. F, Food Res. Int. 48 (2012) 284–290, doi:10.1016/j.foodres.2012.04.007.
[89] O. Marincaş, I. Feher, A new cost-effective approach for lavender essential oils quality assessment, Ind. Crops Prod. 125 (2018) 241–247, doi:10.1016/j.indcrop.2018.09.010.
[90] A. Angioni, A. Barra, E. Cereti, D. Barile, J.D. Coïsson, M. Arlorio, S. Dessi, V. Coroneo, P. Cabras, Chemical composition, plant genetic differences, antimicrobial and antifungal activity investigation of the essential oil of Rosmarinus officinalis L, J. Agric. Food Chem. 52 (2004) 3530–3535, doi:10.1021/jf049913t.
[91] Y. Zaouali, T. Bouzaine, M. Boussaid, Essential oils composition in two Rosmarinus officinalis L. varieties and incidence for antimicrobial and antioxidant activities, Food Chem. Toxicol. 48 (2010) 3144–3152, doi:10.1016/j.fct.2010.08.010.
[92] C.J. Xu, Y.Z. Liang, F.T. Chau, Y.V. Heyden, Pretreatments of chromatographic fingerprints for quality control of herbal medicines, J. Chromatogr. A 1134 (2006) 253–259, doi:10.1016/j.chroma.2006.08.060.
[93] M. Wang, B. Avula, Y.H. Wang, J. Zhao, C. Avonto, J.F. Parcher, V. Raman, J.A. Zweigenbaum, P.L. Wylie, I.A. Khan, An integrated approach utilising chemometrics and GC/MS for classification of chamomile flowers, essential

oils and commercial products, Food Chem 152 (2014) 391–398, doi:10.1016/j.foodchem.2013.11.118.
[94] R.F. Vieira, J.E. Simon, Chemical characterization of basil (Ocimum spp.) based on volatile oils, Flavour Fragr. J. 21 (2006) 214–221, doi:10.1002/ffj.1513.
[95] B. Tohidi, M. Rahimmalek, A. Arzani, Essential oil composition, total phenolic, flavonoid contents, and antioxidant activity of Thymus species collected from different regions of Iran, Food Chem 220 (2017) 153–161, doi:10.1016/j.foodchem.2016.09.203.
[96] E.E. Stashenko, J.R. Martinez, C.A. Ruiz, G. Arias, C. Duran, W. Salgar, M. Cala, Lippia origanoides chemotype differentiation based on essential oil GC-MS and principal component analysis, J. Sep. Sci. 33 (2010) 93–103, doi:10.1002/jssc.200900452.
[97] A. Smelcerovic, M. Spiteller, A.P. Ligon, Z. Smelcerovic, N. Raabe, Essential oil composition of Hypericum L. species from Southeastern Serbia and their chemotaxonomy, Biochem. Syst. Ecol. 35 (2007) 99–113, doi:10.1016/j.bse.2006.09.012.
[98] F. Sáez, Volatile oil variability in Thymus serpylloides ssp. gadorensis growing wild in Southeastern Spain, Biochem. Syst. Ecol. 29 (2001) 189–198, doi:10.1016/S0305-1978(00)00040-5.
[99] G.L. Petretto, M. Maldini, R. Addis, M. Chessa, M. Foddai, J.P. Rourke, G. Pintore, Variability of chemical composition and antioxidant activity of essential oils between Myrtus communis var. Leucocarpa DC and var. Melanocarpa DC, Food Chem 197 (2016) 124–131, doi:10.1016/j.foodchem.2015.10.056.
[100] P.V. Petrakis, M. Couladis, V. Roussis, A method for detecting the biosystematic significance of the essential oil composition: The case of five Hellenic Hypericum L. species, Biochem. Syst. Ecol. 33 (2005) 873–898, doi:10.1016/j.bse.2005.02.002.
[101] T.T. Nhu-Trang, H. Casabianca, M.F. Grenier-Loustalot, Authenticity control of essential oils containing citronellal and citral by chiral and stable-isotope gas-chromatographic analysis, Anal. Bioanal. Chem. 386 (2006) 2141–2152, doi:10.1007/s00216-006-0842-2.
[102] A. Miceli, C. Negro, L. Tommasi, Essential oil variability in Thymbra capitata (L.) Cav. growing wild in Southern Apulia (Italy), Biochem. Syst. Ecol. 34 (2006) 528–535, doi:10.1016/j.bse.2005.12.010.
[103] H. Marzouki, A. Piras, K.B. Salah, H. Medini, T. Pivetta, S. Bouzid, B. Marongiu, D. Falconieri, Essential oil composition and variability of Laurus nobilis L. growing in Tunisia, comparison and chemometric investigation of different plant organs, Nat. Prod. Res. 23 (2009) 343–354, doi:10.1080/14786410802076200.
[104] E. Marengo, C. Baiocchi, M.C. Gennaro, P.L. Bertolo, S. Lanteri, W. Garrone, Classification of essential mint oils of different geographic origin by applying pattern recognition methods to gas chromatographic data, Chemometr. Intell. Lab. Syst. 11 (1991) 75–88, doi:10.1016/0169-7439(91)80007-D.
[105] Y.-Q. Li, D.-X. Kong, H. Wu, Analysis and evaluation of essential oil components of cinnamon barks using GC–MS and FTIR spectroscopy, Ind. Crops Prod. 41 (2013) 269–278, doi:10.1016/j.indcrop.2012.04.056.
[106] Y. Hu, W. Kong, X. Yang, L. Xie, J. Wen, M. Yang, GC-MS combined with chemometric techniques for the quality control and original discrimination of Curcumae longae rhizome: Analysis of essential oils, J. Sep. Sci. 37 (2014) 404–411, doi:10.1002/jssc.201301102.
[107] A. González-Coloma, D. Martín-Benito, N. Mohamed, M.C. García-Vallejo, A.C. Soria, Antifeedant effects and chemical composition of essential oils from different populations of Lavandula luisieri L, Biochem. Syst. Ecol. 34 (2006) 609–616, doi:10.1016/j.bse.2006.02.006.
[108] C.S. Francisco, G.B. Messiano, L.M.X. Lopes, A.G. Tininis, J.E. de Oliveira, L. Capellari, Classification of Aristolochia species based on GC–MS and chemometric analyses of essential oils, Phytochem 69 (2008) 168–175, doi:10.1016/j.phytochem.2007.07.007.
[109] G. Flamini, M. Tebano, P.L. Cioni, Y. Bagci, H. Dural, K. Ertugrul, T. Uysal, A. Savran, A multivariate statistical approach to Centaurea classification using essential oil composition data of some species from Turkey, Plant Syst. Evol. 261 (2006) 217–228, doi:10.1007/s00606-006-0448-3.
[110] A. Elaissi, K.H. Salah, S. Mabrouk, K.M. Larbi, R. Chemli, F. Harzallah-Skhiri, Antibacterial activity and chemical composition of 20 Eucalyptus species' essential oils, Food Chem 129 (2011) 1427–1434, doi:10.1016/j.foodchem.2011.05.100.
[111] A. Elaissi, H. Marzouki, H. Medini, M.L. Khouja, F. Farhat, F. Lynene, F. Harzallah-Skhiri, R. Chemli, Variation in volatile leaf oils of 13 Eucalyptus species harvested from souinet arboreta (tunisia), Chem. Biodivers. 7 (2010) 909–921, doi:10.1002/cbdv.200900229.
[112] C. Demetzos, D. Angelopoulou, D. Perdetzoglou, A comparative study of the essential oils of Cistus salvii folius in several populations of Crete (Greece), Biochem. Syst. Ecol. 30 (2002) 651–665, doi:10.1016/S0305-1978(01)00145-4.
[113] K. Carović-Stanko, S. Orlić, O. Politeo, F. Strikić, I. Kolak, M. Milos, Z. Satovic, Composition and antibacterial activities of essential oils of seven Ocimum taxa, Food Chem 119 (2010) 196–201, doi:10.1016/j.foodchem.2009.06.010.
[114] N.R. Azevedo, I.F.P. Campos, H.D. Ferreira, T.A. Portes, J.C. Seraphin, J. Realino de Paula, S.C. Santos, P.H. Ferri, Essential oil chemotypes in Hyptis suaveolens from Brazilian Cerrado, Biochem. Syst. Ecol. 30 (2002) 205–216, doi:10.1016/S0305-1978(01)00075-8.
[115] C. Zhang, M.J. Qin, P. Shu, J.L. Hong, L. Lü, D.X. He, Chemical variations of the essential oils in flower heads of Chrysanthemum indicum L. from China, Chem. Biodivers. 7 (2010) 2951–2962, doi:10.1002/cbdv.201000034.
[116] Y. Zaouali, C. Messaoud, A.B. Salah, M. Boussaïd, Oil composition variability among populations in relationship with their ecological areas in Tunisian Rosmarinus officinalis L, Flavour Fragr. J., 20 (2005) 512–520, doi:10.1002/ffj.1428.
[117] R.S. Verma, R.K. Verma, R.C. Padalia, A. Chauhan, A. Singh, H.P. Singh, Chemical diversity in the essential oil of Indian Valerian (Valeriana jatamansi Jones), Chem. Biodivers. 8 (2011) 1921–1929, doi:10.1002/cbdv.201100059.
[118] S. Sutour, P. Bradesi, J. Casanova, F. Tomi, Composition and chemical variability of Mentha suaveolens ssp. suaveolens and M. suaveolens ssp. insularis from Corsica, Chem. Biodivers. 7 (2010) 1002–1008, doi:10.1002/cbdv.200900365.
[119] A. Carrubba, R. la Torre, A.D. Prima, F. Saiano, G. Alonzo, Statistical analyses on the essential oil of italian coriander (coriandrum sativum L.) fruits of different ages and origins, J. Essent. Oil Res. 14 (2002) 389–396, doi:10.1080/10412905.2002.9699899.
[120] P. Rubiolo, F. Belliardo, C. Cordero, E. Liberto, B. Sgorbini, C. Bicchi, Headspace - solid-phase microextraction fast GC in combination with principal component analysis as a tool to classify different chemotypes of chamomile flower-heads (Matricaria recutita L.), Phytochem. Anal. 17 (2006) 217–225, doi:10.1002/pca.919.
[121] S. Maietti, D. Rossi, A. Guerrini, C. Useli, C. Romagnoli, F. Poli, R. Bruni, G. Sacchetti, A multivariate analysis approach to the study of chemical and functional properties of chemo-diverse plant derivatives: Lavender essential oils, Flavour Fragr. J. 28 (2013) 144–154, doi:10.1002/ffj.3145.
[122] F. Maggi, F. Papa, C. Giuliani, L. Maleci Bini, A. Venditti, A. Bianco, M. Nicoletti, R. Iannarelli, G. Caprioli, G. Sagratini, M. Cortese, M. Ricciutelli, S. Vittori, Essential oil chemotypification and secretory structures of the neglected vegetable Smyrnium olusatrum L. (Apiaceae) growing in central Italy, Flavour Fragr. J. 30 (2015) 139–159, doi:10.1002/ffj.3221.
[123] M. Leonardi, K.E. Ambryszewska, B. Melai, G. Flamini, P.L. Cioni, F. Parri, L. Pistelli, Essential-oil composition of Helichrysum italicum (Roth) G. Don ssp. italicum from Elba Island (Tuscany, Italy), Chem. Biodivers. 10 (2013) 343–355, doi:10.1002/cbdv.201200222.
[124] D.V. Lakušić, M.S. Ristić, V.N. Slavkovska, J.B. Šinžar-Sekulić, B.S. Lakušić, Environment-related variations of the composition of the essential oils of rosemary (Rosmarinus officinalis L.) in the Balkan penninsula, Chem. Biodivers. 9 (2012) 1286–1302, doi:10.1002/cbdv.201100427.
[125] C. Kapetanos, A. Karioti, S. Bojović, P. Marin, M. Veljić, H. Skaltsa, Chemical and principal-component analyses of the essential oils of Apioideae taxa (Apiaceae) from Central Balkan, Chem. Biodivers. 5 (2008) 101–119, doi:10.1002/cbdv.200890000.
[126] M. Jug-Dujaković, M. Ristić, D. Pljevljakušić, Z. Dajić-Stevanović, Z. Liber, K. Hančević, T. Radić, Z. Šatović, High diversity of indigenous populations of dalmatian sage (Salvia officinalis L.) in essential-oil composition, Chem. Biodivers. 9 (2012) 2309–2323, doi:10.1002/cbdv.201200131.
[127] I. Ben El Hadj Ali, A. Guetat, M. Boussaid, Chemical and genetic variability of Thymus algeriensis Boiss. et Reut. (Lamiaceae), a North African endemic species, Ind. Crops Prod. 40 (2012) 277–284, doi:10.1016/j.indcrop.2012.03.021.
[128] V. Dunkić, D. Kremer, I. Dragojević Müller, E. Stabentheiner, S. Kuzmić, R. Jurišić Grubešić, L. Vujić, I. Kosalec, M. Randić, S. Srečec, N. Bezić, Chemotaxonomic and micromorphological traits of Satureja montana L. and S. subspicata Vis. (Lamiaceae), Chem. Biodivers. 9 (2012) 2825–2842, doi:10.1002/cbdv.201100376.
[129] L. De Cooman, E. Everaert, D. De Keukeleire, Quantitative analysis of hop acids, essential oils and flavonoids as a clue to the identification of hop varieties, Phytochem. Anal. 9 (1998) 145–150, doi:10.1002/(SICI)1099-1565(199805/06)9:3⟨145::AID-PCA393⟩3.0.CO;2-K.
[130] H. Chograni, Y. Zaouali, C. Rajeb, M. Boussaid, Essential oil variation among natural populations of Lavandula multifida L. (Lamiaceae), Chem. Biodivers. 7 (2010) 933–942, doi:10.1002/cbdv.200900201.
[131] J.B. Boti, A. Bighelli, C. Cavaleiro, L. Salgueiro, J. Casanova, Chemical variability of Juniperus oxycedrus ssp. oxycedrus berry and leaf oils from Corsica, analysed by combination of GC, GC-MS and 13C-NMR, Flavour Fragr. J., 21 (2006) 268–273, doi:10.1002/ffj.1579.
[132] P.C.S. Barbosa, R.S. Medeiros, P.T.B. Sampaio, B.G. Vieira, L.S.M. Wiedemann, V.F. Veiga, Influence of abiotic factors on the chemical composition of copaiba oil (Copaifera multijuga hayne): Soil composition, seasonality and diameter at breast height, J. Braz. Chem. Soc. 23 (2012) 1823–1833, doi:10.1590/S0103-50532012005000049.
[133] P. Champagnat, G. Figueredo, J.C. Chalchat, A.P. Carnat, J.M. Bessière, A study on the composition of commercial Vetiveria zizanioides oils from different geographical origins, J. Essent. Oil Res. 18 (2006) 416–422, doi:10.1080/10412905.2006.9699129.
[134] A.J.D. Silvestre, J.A.S. Cavaleiro, B. Delmond, C. Filliatre, G. Bourgeois, Analysis of the variation of the essential oil composition of Eucalyptus globulus Labill. from Portugal using multivariate statistical analysis, Ind. Crops Prod. 6 (1997) 27–33, doi:10.1016/S0926-6690(96)00200-2.
[135] J. Tian, X. Zeng, S. Zhang, Y. Wang, P. Zhang, A. Lü, X. Peng, Regional variation in components and antioxidant and antifungal activities of Perilla frutescens essential oils in China, Ind. Crops Prod. 59 (2014) 69–79, doi:10.1016/j.indcrop.2014.04.048.
[136] M. Kovačevič, M. Kač, Determination and verification of hop varieties by analysis of essential oils, Food Chem 77 (2002) 489–494, doi:10.1016/S0308-8146(02)00114-0.
[137] R. Ravi, M. Prakash, K.K. Bhat, Characterization of aroma active compounds of cumin (Cuminum cyminum L.) by GC-MS, E-Nose, and sensory techniques, Int. J. Food Prop. 16 (2013) 1048–1058, doi:10.1080/10942912.2011.576356.

[138] Z. Xiang, X.Q. Wang, X.J. Caib, S. Zenga, Metabolomics study on quality control and discrimination of three curcuma species based on gas chromatograph-mass spectrometry, Phytochem. Anal. 22 (2011) 411–418, doi:10.1002/pca.1296.
[139] S. Cui, J. Wang, L. Yang, J. Wu, X. Wang, Qualitative and quantitative analysis on aroma characteristics of ginseng at different ages using E-nose and GC-MS combined with chemometrics, J. Pharm. Biomed. Anal. 102 (2015) 64–77, doi:10.1016/j.jpba.2014.08.030.
[140] Y. Akakabe, M. Sakamoto, Y. Ikeda, M. Tanaka, Identification and characterization of volatile components of the Japanese sour citrus fruit Citrus nagato-yuzukichi Tanaka, Biosci. Biotechnol. Biochem. 72 (2008) 1965–1968, doi:10.1271/bbb.80144.
[141] J. Bernáth, I. Novák, K. Szabó, Z. Seregély, Evaluation of selected oregano (Origanum vulgare L. subsp. hirtum Ietswaart) lines with traditional methods and sensory analysis, J. Herbs Spices Med. Plants 11 (2006) 19–26, doi:10.1300/J044v11n04_03.
[142] R. Ravi, M. Prakash, K.K. Bhat, Aroma characterization of coriander (Coriandrum sativum L.) oil samples, Eur. Food Res. Technol. 225 (2007) 367–374, doi:10.1007/s00217-006-0425-7.
[143] A. Gorji-Chakespari, A.M. Nikbakht, F. Sefidkon, M. Ghasemi-Varnamkhasti, E.L. Valero, Classification of essential oil composition in Rosa damascena Mill. genotypes using an electronic nose, J. Appl. Res. Med. Aromat. Plants 4 (2017) 27–34, doi:10.1016/j.jarmap.2016.07.004.
[144] S.M. Amin, H.M. Hassan, A.E.N.G. El Gendy, A.A. El-Beih, T.A. Mohamed, A.I. Elshamy, A. Bader, K.A. Shams, R. Mohammed, M.E.F. Hegazy, Comparative chemical study and antimicrobial activity of essential oils of three Artemisia species from Egypt and Saudi Arabia, Flavour Fragr. J. 34 (2019) 450–459, doi:10.1002/ffj.3525.
[145] Z. Liu, S. Kuang, M. Qing, D. Wang, D. Li, Metabolite profiles of essential oils and SSR molecular markers in Juniperus rigida Sieb. et Zucc. from different regions: A potential source of raw materials for the perfume and healthy products, Ind. Crops Prod. 133 (2019) 424–434, doi:10.1016/j.indcrop.2019.03.034.
[146] S. Škobić, M.D. Marčetić, T. Kundaković-Vasović, J. Crnobarac, Nitrogen fertilization and the essential oils profile of the rhizomes of different sweet flag populations (Acorus calamus L.), Ind. Crops Prod. 142 (2019), doi:10.1016/j.indcrop.2019.111871.
[147] I. Zribi, J. Bleton, F. Moussa, M. Abderrabba, GC-MS analysis of the volatile profile and the essential oil compositions of Tunisian Borago officinalis L.: Regional locality and organ dependency, Ind. Crops Prod. 129 (2019) 290–298, doi:10.1016/j.indcrop.2018.12.021.
[148] K.C. Camargo, L.P. Duarte, D.M. Vidal, H.V. Pereira, R.C.G. Pereira, M.G. de Aguilar, G.F. de Sousa, S.A.V. Filho, M.O. Mercadante-Simoes, M. Messias, D.M. de Oliveira, Chemodiversity of essential oils from nine species of Celastraceae, Chem. Biodivers. 17 (2020) e2000107, doi:10.1002/cbdv.202000107.
[149] X. Yan, W. Wang, Z. Chen, Y. Xie, Q. Li, Z. Yu, H. Hu, Z. Wang, Quality assessment and differentiation of Aucklandiae Radix and Vladimiriae Radix based on GC-MS fingerprint and chemometrics analysis: Basis for clinical application, Anal. Bioanal. Chem. 412 (2020) 1535–1549, doi:10.1007/s00216-019-02380-2.
[150] I. Yener, M.A. Yilmaz, O.T. Olmez, M. Akdeniz, F. Tekin, N. Hasimi, M.H. Alkan, M. Ozturk, A. Ertas, A detailed biological and chemical investigation of sixteen Achillea species' essential oils via chemometric approach, Chem. Biodivers. 17 (2020) e1900484, doi:10.1002/cbdv.201900484.
[151] A.E. Altyar, M.L. Ashour, F.S. Youssef, Premna odorata: Seasonal metabolic variation in the essential oil composition of its leaf and verification of its anti-ageing potential via in vitro assays and molecular modelling, Biomolecules 10 (2020), doi:10.3390/biom10060879.
[152] M. Allenspach, C. Valder, D. Flamm, F. Grisoni, C. Steuer, Verification of chromatographic profile of primary essential oil of Pinus sylvestris L. combined with chemometric analysis, Molecules 25 (2020), doi:10.3390/molecules25132973.
[153] H. Ao, J. Wang, L. Chen, S. Li, C. Dai, Comparison of volatile oil between the fruits of Amomum villosum Lour. and Amomum villosum Lour. var. xanthioides T. L. Wu et Senjen based on GC-MS and chemometric techniques, Molecules 24 (2019), doi:10.3390/molecules24091663.
[154] R.N. Alolga, M. Chavez Leon, G. Osei-Adjei, V. Onoja, GC-MS-based metabolomics, antibacterial and anti-inflammatory investigations to characterize the quality of essential oil obtained from dried Xylopia aethiopica fruits from Ghana and Nigeria, J. Pharm. Pharmacol. 71 (2019) 1544–1552, doi:10.1111/jphp.13150.
[155] M. Sabatino, M. Fabiani, M. Bozovic, S. Garzoli, L. Antonini, M.E. Marcocci, A.T. Palamara, G. De Chiara, R. Ragno, Experimental data based machine learning classification models with predictive ability to select in vitro active antiviral and non-toxic essential oils, Molecules 25 (2020), doi:10.3390/molecules25102452.
[156] D.J. Beale, P.D. Morrison, A.V. Karpe, M.S. Dunn, Chemometric analysis of lavender essential oils using targeted and untargeted GC-MS acquired data for the rapid identification and characterization of oil quality, Molecules 22 (2017), doi:10.3390/molecules22081339.
[157] A. Elaissi, S. Moumni, K. Roeleveld, M.Larbi Khouja, Chemical characterization of five Tunisian eucalyptus essential oils species, Chem. Biodivers. 17 (2020) e1900378, doi:10.1002/cbdv.201900378.
[158] J.E. Jackson, A User's Guide To Principal Components, John Wiley & Sons, Inc., New York, 1991.
[159] R. Bro, A.K. Smilde, Principal component analysis, Anal. Methods 6 (2014) 2812–2831, doi:10.1039/c3ay41907j.
[160] Y. Roggo, P. Chalus, L. Maurer, C. Lema-Martinez, A. Edmond, N. Jent, A review of near infrared spectroscopy and chemometrics in pharmaceutical technologies, J. Pharm. Biomed. Anal. 44 (2007) 683–700, doi:10.1016/j.jpba.2007.03.023.
[161] M. Barker, W. Rayens, Partial least squares for discrimination, J. Chemom. 17 (2003) 166–173, doi:10.1002/cem.785.
[162] J.A. Westerhuis, H.C.J. Hoefsloot, S. Smit, D.J. Vis, A.K. Smilde, E.J.J. van Velzen, J.P.M. van Duijnhoven, F.A. van Dorsten, Assessment of PLSDA cross validation, Metabolomics 4 (2008) 81–89, doi:10.1007/s11306-007-0099-6.
[163] P. Geladi, B.R. Kowalski, Partial least-squares regression: A tutorial, Anal. Chim. Acta 185 (1986) 1–17, doi:10.1016/0003-2670(86)80028-9.
[164] L. Breiman, Random Forests, Machine Learning 45 (2001) 5–32, doi:10.1023/a:1010933404324.
[165] P.S. Gromski, Y. Xu, E. Correa, D.I. Ellis, M.L. Turner, R. Goodacre, A comparative investigation of modern feature selection and classification approaches for the analysis of mass spectrometry data, Anal. Chim. Acta 829 (2014) 1–8, doi:10.1016/j.aca.2014.03.039.
[166] S. Wold, M. Sjöström, L. Eriksson, PLS-regression: A basic tool of chemometrics, Chemom. Intell. Lab. Syst 58 (2001) 109–130, doi:10.1016/S0169-7439(01)00155-1.
[167] Y.J. Park, S.A. Baek, Y. Choi, J.K. Kim, S.U. Park, Metabolic profiling of nine Mentha species and prediction of their antioxidant properties using chemometrics, Molecules 24 (2019), doi:10.3390/molecules24020258.
[168] L. Lebanov, L. Tedone, M. Kaykhaii, M.R. Linford, B. Paull, Multidimensional gas chromatography in essential oil analysis. Part 1: Technical developments, Chromatographia 82 (2019) 377–398, doi:10.1007/s10337-018-3649-3.
[169] L. Lebanov, L. Tedone, M. Kaykhaii, M.R. Linford, B. Paull, Multidimensional gas chromatography in essential oil analysis. Part 2: Application to characterisation and identification, Chromatographia 82 (2019) 399–414, doi:10.1007/s10337-018-3651-9.
[170] L.W. Hantao, H.G. Aleme, M.P. Pedroso, G.P. Sabin, R.J. Poppi, F. Augusto, Multivariate curve resolution combined with gas chromatography to enhance analytical separation in complex samples: A review, Anal. Chim. Acta 731 (2012) 11–23, doi:10.1016/j.aca.2012.04.003.
[171] M. Jalali-Heravi, R.S. Moazeni-Pourasil, H. Sereshti, Elimination of chromatographic and mass spectrometric problems in GC-MS analysis of Lavender essential oil by multivariate curve resolution techniques: Improving the peak purity assessment by variable size moving window-evolving factor analysis, J. Chromatogr. B Analyt. Technol. Biomed. Life Sci. 983-984 (2015) 83–89, doi:10.1016/j.jchromb.2015.01.005.
[172] M. Jalali-Heravi, R.S. Moazeni-Pourasil, H. Sereshti, Thorough analysis of Iranian spearmint essential oil: combination of chemometrics and gas chromatography-mass spectrometry, Anal. Methods 6 (2014), doi:10.1039/c3ay42215a.
[173] M. Jalali-Heravi, H. Parastar, H. Sereshti, Development of a method for analysis of Iranian damask rose oil: Combination of gas chromatography-mass spectrometry with chemometric techniques, Anal. Chim. Acta 623 (2008) 11–21, doi:10.1016/j.aca.2008.05.078.
[174] C. Ruckebusch, L. Blanchet, Multivariate curve resolution: A review of advanced and tailored applications and challenges, Anal. Chim. Acta 765 (2013) 28–36, doi:10.1016/j.aca.2012.12.028.
[175] M. Jalali-Heravi, B. Zekavat, H. Sereshti, Characterization of essential oil components of Iranian geranium oil using gas chromatography-mass spectrometry combined with chemometric resolution techniques, J. Chromatogr. A 1114 (2006) 154–163, doi:10.1016/j.chroma.2006.02.034.
[176] M. Jalali-Heravi, H. Parastar, H. Ebrahimi-Najafabadi, Self-modeling curve resolution techniques applied to comparative analysis of volatile components of Iranian saffron from different regions, Anal. Chim. Acta 662 (2010) 143–154, doi:10.1016/j.aca.2010.01.013.
[177] M. Jalali-Heravi, B. Zekavat, H. Sereshti, Use of gas chromatography-mass spectrometry combined with resolution methods to characterize the essential oil components of Iranian cumin and caraway, J. Chromatogr. A 1143 (2007) 215–226, doi:10.1016/j.chroma.2007.01.042.
[178] 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, doi:10.1007/s10337-017-3322-2.
[179] 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), doi:10.1016/j.talanta.2020.121208.
[180] Q.-Q. Zhang, S.-H. Lei, X.-L. Wang, L. Wang, C.-J. Zhu, Discrimination of phytoplankton classes using characteristic spectra of 3D fluorescence spectra, Spectrochim. Acta A Mol. Biomol. Spectrosc. 63 (2006) 361–369, doi:10.1016/j.saa.2005.05.041.
[181] C. Wittrup, Comparison of chemometric methods for classification of fungal extracts based on rapid fluorescence spectroscopy, J. Chemom. 14 (2000) 765–776, doi:10.1002/1099-128X(200009/12)14:5/6⟨765::AID-CEM625⟩3.0.CO;2-S.
[182] C.A. Stedmon, R. Bro, Characterizing dissolved organic matter fluorescence with parallel factor analysis: A tutorial, Limnol. Oceanogr-Meth. 6 (2008) 572–579, doi:10.4319/lom.2008.6.572.

[183] H. Nikpour, M. Mousavi, H. Asadollahzadeh, Qualitative and quantitative analysis of Teucrium polium essential oil components by GC-MS coupled with MCR and PARAFAC methods, Phytochem. Anal. 29 (2018) 590–600, doi:10.1002/pca.2772.
[184] R. Bro, Review on multiway analysis in chemistry - 2000-2005, Crit. Rev. Anal. Chem. 36 (2006) 279–293, doi:10.1080/10408340600969965.
[185] A. de Juan, R. Tauler, Comparison of three-way resolution methods for non-trilinear chemical data sets, J. Chemom. 15 (2001) 749–771, doi:10.1002/cem.662.
[186] S.A. Bortolato, A.C. Olivieri, Chemometric processing of second-order liquid chromatographic data with UV-Vis and fluorescence detection. A comparison of multivariate curve resolution and parallel factor analysis 2, Anal. Chim. Acta 842 (2014) 11–19, doi:10.1016/j.aca.2014.07.007.
[187] R. Kohavi, A study of cross-validation and bootstrap for accuracy estimation and model selection, Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (1995) 1137–1143.
[188] D.I. Broadhurst, D.B. Kell, Statistical strategies for avoiding false discoveries in metabolomics and related experiments, Metabolomics 2 (2006) 171–196, doi:10.1007/s11306-006-0037-z.
[189] J. Shao, Linear model selection by cross-validation, J. Amer. Statistical Assoc. 88 (1993) 486–494, doi:10.1080/01621459.1993.10476299.
[190] R.D. Snee, Validation of regression models: Methods and examples, Technometrics 19 (1977) 415–428, doi:10.1080/00401706.1977.10489581.
[191] R.W. Kennard, L.A. Stone, Computer aided design of experiments, Technometrics 11 (1969) 137–148, doi:10.1080/00401706.1969.10490666.
[192] R.K.H. Galvão, M.C.U. Araujo, G.E. José, M.J.C. Pontes, E.C. Silva, T.C.B. Saldanha, A method for calibration and validation subset partitioning, Talanta 67 (2005) 736–740, doi:10.1016/j.talanta.2005.03.025.
[193] E. Szymanska, E. Saccenti, A.K. Smilde, J.A. Westerhuis, Double-check: validation of diagnostic statistics for PLS-DA models in metabolomics studies, Metabolomics 8 (2012) 3–16, doi:10.1007/s11306-011-0330-3.
[194] J. Riedl, S. Esslinger, C. Fauhl-Hassek, Review of validation and reporting of non-targeted fingerprinting approaches for food authentication, Anal. Chim. Acta 885 (2015) 17–32, doi:10.1016/j.aca.2015.06.003.
[195] C.A. Smith, G. O'Maille, E.J. Want, C. Qin, S.A. Trauger, T.R. Brandon, D.E. Custodio, R. Abagyan, G. Siuzdak, METLIN: A metabolite mass spectral database, Ther. Drug Monit. 27 (2005) 747–751, doi:10.1097/01.ftd.0000179845.53213.39.
[196] J. Chong, O. Soufan, C. Li, I. Caraus, S. Li, G. Bourque, D.S. Wishart, J. Xia, MetaboAnalyst 4.0: towards more transparent and integrative metabolomics analysis, Nucleic Acids Res 46 (2018) 486–494, doi:10.1093/nar/gky310.
[197] T. Skov, A.H. Honoré, H.M. Jensen, T. Næs, S.B. Engelsen, Chemometrics in foodomics: Handling data structures from multiple analytical platforms, TrAC, Trends Anal. Chem. 60 (2014) 71–79, doi:10.1016/j.trac.2014.05.004.
[198] A.K. Smilde, J.A. Westerhuis, S. de Jong, A framework for sequential multiblock component methods, J. Chemom. 17 (2003) 323–337, doi:10.1002/cem.811.
[199] A.K. Smilde, M.J. van der Werf, S. Bijlsma, B.J.C. van der Werff-van der Vat, R.H. Jellema, Fusion of mass spectrometry-based metabolomics data, Anal. Chem. 77 (2005) 6729–6736, doi:10.1021/ac051080y.