Trends in Analytical Chemistry 139 (2021) 116251

Contents lists available at ScienceDirect

Trends in Analytical Chemistry

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

Advanced chemometric and data handling tools for GC×GC-TOF-MS

Application of chemometrics and related advanced data handling in chemical separations

Pierre-Hugues Stefanuto [a, *], Agnieszka Smolinska [b], Jean-François Focant [a]

[a] Organic and Biological Analytical Chemistry Group (OBIACHEM), MolSys Research Unit, Liège University, Belgium
[b] NUTRIM School of Nutrition and Translational Research in Metabolism, Department Pharmacology & Toxicology, Maastricht University, the Netherlands

ARTICLE INFO

Article history:
Available online 8 March 2021

Keywords:
Comprehensive two-dimensional gas chromatography
GCxGC-MS
Chemometrics
Data handling
Processing
Multidimensional data

ABSTRACT

Comprehensive two-dimensional gas chromatography coupled to mass spectrometry (GC×GC-MS) has become a mature technique. GC×GC-MS can now be used to conduct large scale studies, giving full access to its high-resolution power for targeted and mostly untargeted screening. The current challenges are now localized on the data management side, where powerful chemometric tools are required to unlock GC×GC-MS full potential.

This manuscript reviews and discusses recent advances in the development of specific chemometrics for GC×GC-MS. It is designed as a guide to users who desire to establish robust and reproducible GC×GC-MS data processing workflows. Each of the critical steps of data preprocessing, feature selection, model building, and validation are described and considered in detail. Finally, some future perspectives on the development of the next generation of chemometric tools based on Artificial Intelligence, especially machine learning, are critically discussed.

1. Introduction

Chromatography is an indispensable tool for analytical chemists. It allows the separation and the characterization of individual compounds in complex samples. However, in the quest for the better elucidation of complex mixture composition, there is a constant need for development towards multidimensional chromatography set-ups. These ideally being combined to accurate detection techniques, creating efficient hyphenated separation science solutions. Furthermore, the constant development of instrumental technology allows users to increase the number of analytical dimensions in their set-up, adding more chromatographic or detection levels [1].

In that context, gas chromatography (GC), next to liquid chromatography (LC), is one of the most commonly found chromatographic system in analytical laboratories. It is the "go for" instrument for the characterization of volatile and semi-volatile molecules. With the rising need for untargeted measurements, comprehensive two-dimensional gas chromatography (GC×GC) has been developed to provide a full overview of complex sample composition [1,2]. In brief, GC×GC relies on the combination of two chromatographic dimensions through a specific device, the modulator. This combination significantly increases the separation capacity of the analytical system. In addition, the modulator allows a comprehensive transfer from the first (^{1}D) to the second (^{2}D) GC dimension [3,4]. Three main types of modulators exist: phase-ratio, cryogenic, and flow-based systems [1,5]. Through the constant development of modulator hardware, most of them are now highly reliable [1,5]. On the detector side, next to the versatile but non-specific flame ionization detectors (FIDs), mass analyzers such as fast acquisition time-of-flight mass spectrometers (TOF-MS) are the most commonly used in GC×GC, followed by the last generations of fast scanning quadrupoles [1,6]. This multiplication of chromatographic dimension and detection resolution is generating more and more data for the analysis of a single sample (Fig. 1). When considering a system such as GC×GC-TOF-MS, sets of multichannel information are generated after completion of the chromatographic event during which MS data are acquired at high frequency. Such sets of information are typically reconstructed and presented under the form of a chromatography plane containing 2D peaks made of

* Corresponding author. University of Liège, Molecular System, Organic Biological Analytical Chemistry Group, Quartier Agora, Allée du Six Août, 11, 4000 Liège, Sart-Tilman, Belgium.
E-mail address: phstefanuto@uliege.be (P.-H. Stefanuto).

https://doi.org/10.1016/j.trac.2021.116251
0165-9936/

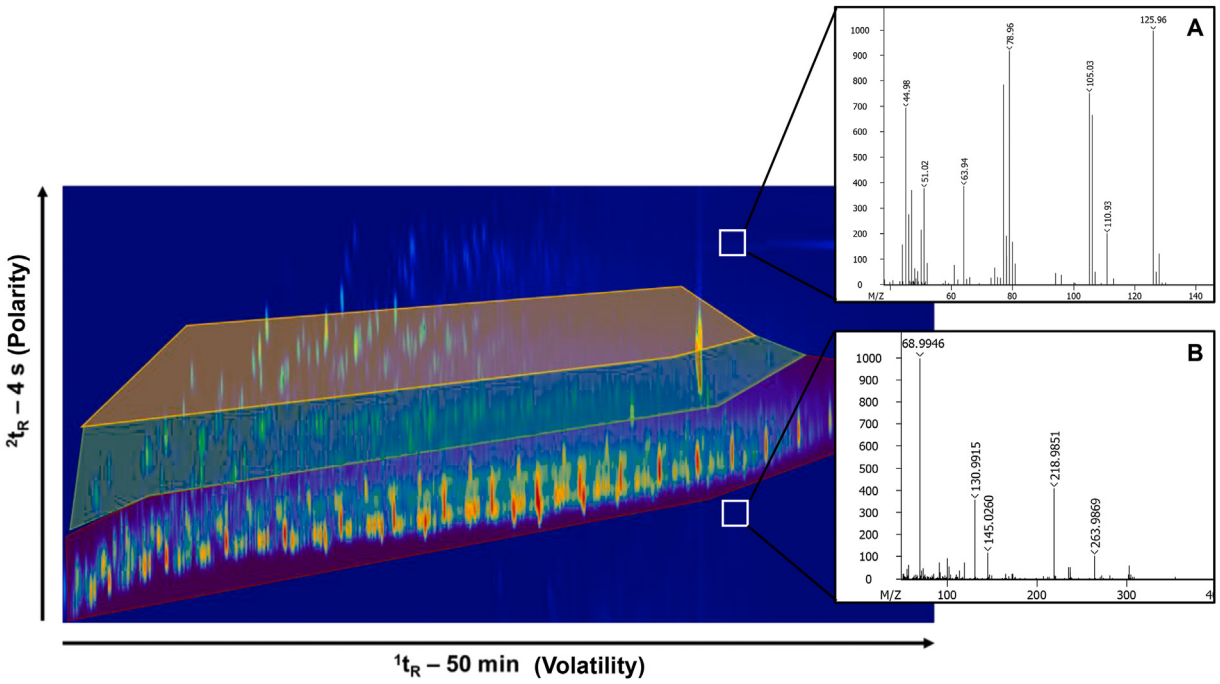

Fig. 1. Visualization of GC×GC-TOF-MS data using a contour plot depicting the chromatographic separation and relative intensities of sample constituents. Beyond the already complex image containing hundreds of chromatographic signal, MS information are available (e.g. compounds A and B) at typical frequencies of 20–200 Hz over mass ranges in the order of 30–400 amu. This results in data matrices of more than 1.10^8 points.

recombined modulated GC signals that are themselves made of ion currents recorded by the MS analyzer (Fig. 1).

In recent years, high-resolution/high accuracy MS analyzers, such as HRTOF-MS, have been more commonly used in combination with GC×GC. Commercial systems are able to reach 50,000 in resolution, with an acquisition frequency of 200 Hz and a mass accuracy at the low ppm level. As a direct result, sizes of already very large data files have further grown to a stage where their manipulation has become a potential issue. Moreover, with the development of robust and user-friendly GC×GC systems, the technique has transitioned from "single sample analysis" to large scale "omics" screening [7,8]. The combination of high dimensional data files with large numbers of samples further raises some challenges for GC×GC users. First, the data acquisition and storage capacity need to be adapted to the large data file. Then, it requires the introduction of robust chemometric tools and processing workflows in order to access and extract the reliable information that will practically answer the original research question.

Chemometric tools are useful for GC×GC applications in targeted and untargeted screening [9]. However, they are most commonly used to handle untargeted investigations. In targeted studies, the analyst knows which compounds to look for in the sample. In those situations, MS signal deconvolution tools are generally sufficient to extract the required information [2,9–11]. Indeed, deconvolution aims to resolve overlapping signals from different analytes, so that each analyte can be associated with its own cleaned signal and used for further analysis and reporting. Thus, the process of targeted studies will not be discussed in this review, information on the topic can be found elsewhere [2,9–11]. For untargeted research, the situation is significantly more complex. The analyst does not have *a priori* knowledge about the sample composition or the relevant compounds to monitor. In that case, advanced data handling tools are necessary to narrow down the field of investigation. Advanced chemometrics generally relies on a multi-step workflow of mathematical and statistical operations, which allows to robustly and reliably extract the relevant information. This workflow relies on preprocessing methods (signal and data), exploratory data analysis, feature selection, classification methods, and validation tools (Fig. 2). All these steps need to be carefully considered, properly implemented, and optimized for the specific needs of a given study in order to ensure a minimal robustness and reliability of the study outcomes (Fig. 2).

This review critically discusses the different steps of this workflow in the specific context of GC×GC-(HR)MS research. Moreover, the impact of raw data quality on the output of the data processing and the necessity to optimize properly the method (from sampling to data processing) in regard to the analytical question will also be discussed. Finally, perspectives of future developments, such as the implementation of machine learning tools, computing power requirement, and data visualization, will be examined.

Note: some aspects of GC×GC-MS developments in hardware (e.g., modulators and MS detectors) and preprocessing (e.g., deconvolution) have been covered in previous publications. For those aspects, only a brief summary and relevant references are provided.

2. Alignment, preprocessing, and exploratory analysis

2.1. Type of data

Before discussing the best way to apprehend GC×GC-MS data, it is important to understand what type of information is generated by such a multidimensional technique. GC×GC-MS generates quantitative data. Quantitative variables are measured on continuous scale (interval or ratio scale). They are not discrete and can take any value [12]. In most of the studies, GC×GC-MS data are investigated using the integrated peak area for every feature detected and a mass spectral information for the identification. Those quantitative data are processed with qualitative goals, which

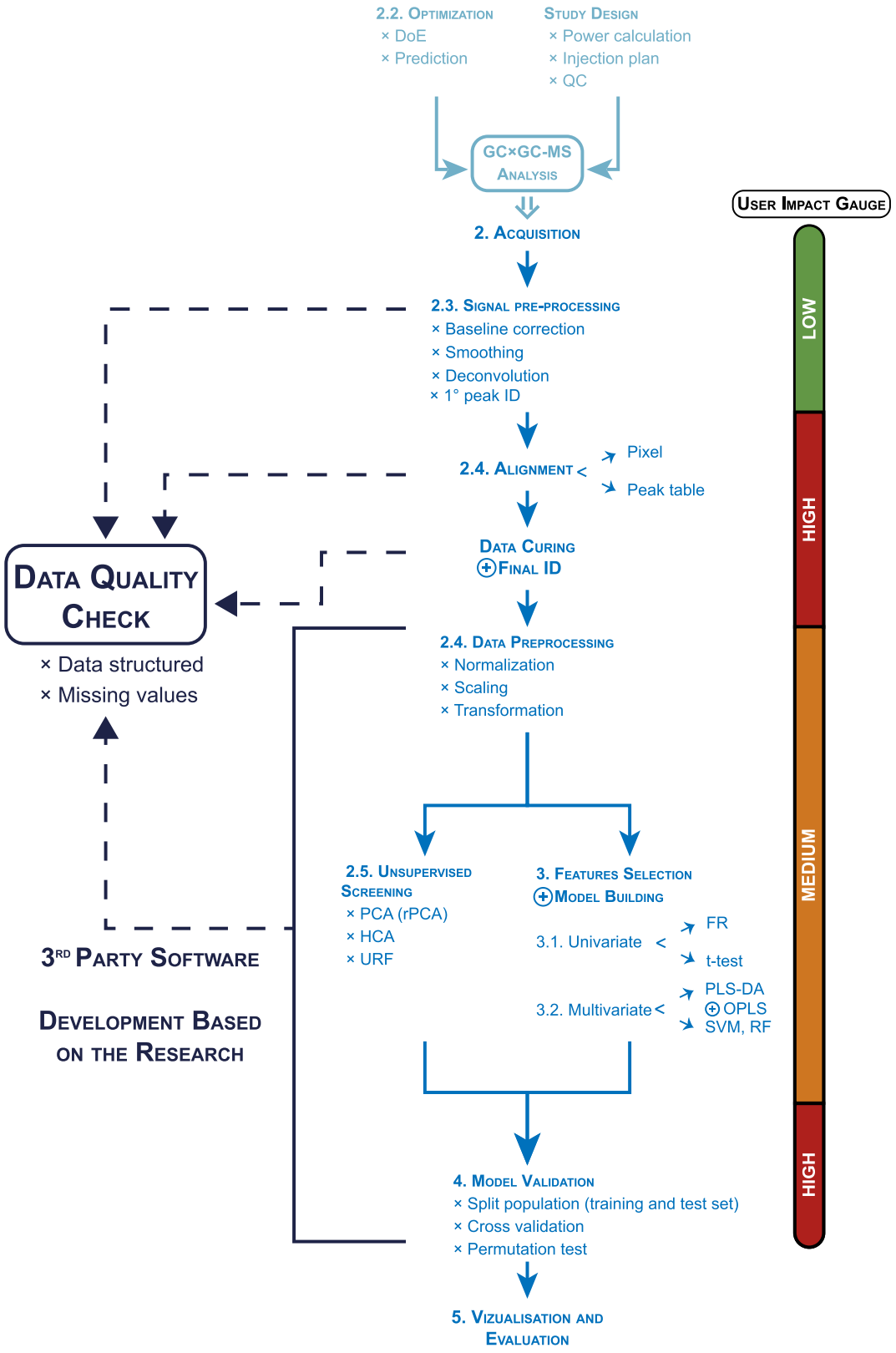

Fig. 2. General data processing workflow for GC×GC-(HR)MS data. The objective of this figure is to provide a guide to user during data analysis. It emphasized the processing steps where user decisions have a large impact on the output.

include compound identification, exploratory data analysis, feature selection, and clustering analysis.

A typical GC×GC-MS data file is acquired at high acquisition frequency (i.e., from 50 to 500 Hz, or mass spectra per second), each acquired data point contains two retention times and a complete mass spectrum [1,13]. The resulting file can easily climb to hundreds of gigabytes, especially when HRTOF-MS profile data is used. Such file sizes require dedicated attention for storage but also robust data reduction tools to ensure usability. The introduction of high-resolution mass spectrometers compatible with GC×GC system has opened a new door for chemometric tools, providing access to increased compounds identification capacities (especially combined with soft ionization) and MS-Based processing (e.g., Kendricks Mass Defect plot) [13–15]. Moreover, tandem ionization mode is also available for GC×GC-MS, increasing the analytical power of the technique but also the complexity of the data [16]. A comprehensive review of the utilization of HRMS data from GC×GC system is available elsewhere [13].

2.2. Method development, study design, and quality control

For robust data processing, an important aspect to take into account is the inherent quality of the raw data since the output of model building can only be as good as the quality of the raw data [10,12,17]. Indeed, most of the models have been designed for quantitative data. Users have to make sure that the data correspond to the chemometric tools applied, in term of data type, quality, and dimensionality. When a semi-quantitative approach is used, one has to remember that the detector response must be linear over the working range. Moreover, statistical methods can be impacted by the dynamic range and the sample absolute concentration. To facilitate the building of the model, a robustly optimized analytical method is required. In order to guarantee data quality, chemometric tools, such as design of experiment [18–21] or in silico optimization [22–27], in combination with a properly defined quality assurance (QA) system are required (Fig. 2).

When running large scale untargeted chromatographic research, a challenging point is to deal with batch effects, i.e., all systematic technical variations that can be observed between different batches of measurements. These batch-related variations are unrelated to the research question and can easily overpass valuable sample-related variations. For chromatographic data, these variations are typically retention time shifting and signal intensity variations. The best way to detect and correct for batch effects is by using an adapted study design.

Study design can simply be defined as the framework established to answer a specific research question. It guarantees that the collected information will be robust enough to help to address a specific scientific question by controlling a maximum of the variability coming from confounding effects. In omics studies, the study design is made of multi-step analytical experiments to answer a biological research question. A proper study design relies on an optimized analytical workflow (used to define the standard operating procedure (SOP)) and a proper QA system built on a reliable population (e.g., using power calculation) and representative quality control (QC) samples [19,20,28]. Power calculation helps to establish the number of independent samples required to answer the research question. Different approaches exist, they generally rely on preliminary data to assess the variability of the population.

A key factor of successful analytical design is a proper QC method selection. The QC allows the detection and minimization of batch effects [20,29]. Moreover, chemometrics tools can be employed to reduce batch effects [28]. A proper QC sample should represent the qualitative and quantitative composition of the entire sample set. Different types of QC can be used and combined: internal standard spiking, external standard mixture, and pooled QC. The standard-based approaches are generally applied for targeted or semi-targeted studies. In all types of studies, they can also be used as reliable peaks and used as anchors during data alignment. In the context of untargeted research, the pooled QC approach is usually the way to go [20,29]. A polled QC can be described as a representative average of the chemical composition as it is made of a fraction of all the samples involved in the study. The main limitation of this approach is that some matrices are still difficult to transfer into a QC format. For example, exhaled breath research is relying more and more on GC-MS and GC×GC-HRTOF-MS [30]. However, there is currently no robust way to pool breath samples to create such a QC nor any type of breath reference material available. In this type of situation, analytical scientists need to rely on internal and external standard-based strategies in combination with the implementation of proper injection strategies (e.g., randomization). This approach provides the ability to perform signal correction and evaluate the analytical precision even in the context of an untargeted study [20,31].

2.3. Acquisition and signal preprocessing

GC×GC-MS data acquisition can be performed with any classical GC software. However, in order to handle and to generate 2D data, signal reconstruction requires a dedicated software. Several commercial solutions are available, the most commonly used are: ChromaTOF™ (LECO Corp., USA), GC Image™ (Zoex Corp., USA), Chromspace® (SepSolve Anaytical Ltd, UK), AnalyzerPro® XD (SpectralWorks Ltd, UK), Canvas-2DGC® (J&X Technologies, China), ChromSquare® (Shimadzu Corp., Japan), GasPedal® (Decodon Software UG, Germany), but also open-source options, such as OpenChrom® (LabLicate, Germany) and Guineu [32,33]. Next to the acquisition comes the preprocessing. Preprocessing aims to transform the raw data into cleaned data that can be used for further processing. Cleaned data have been corrected for any unwanted experimental variations and contaminations [12,17]. For GC×GC-MS, we can differentiate signal preprocessing and data preprocessing. Signal preprocessing represents the preprocessing steps usually conducted in separation sciences. These steps are generally automatized and included in most of the GC×GC software. Data preprocessing is more “omics” oriented. It is not specific to chromatography but it is mandatory for large scale study. Data preprocessing generally has a larger user impact than signal preprocessing (Fig. 2). From this distinction, one can differentiate “chromatographic or MS signal” from “feature”. Therefore, in the present manuscript, the raw data prior to be treated during the signal preprocessing will be called “chromatographic or MS signal”. The preprocessed signal will be referred to as a “feature” from the data preprocessing step to the end of the processing.

All GC×GC-MS software have different levels of functionalities but they all allow basic signal preprocessing, such as 2D reconstruction, baseline correction, smoothing, deconvolution, and peak identification [33] (Fig. 2). Baseline fluctuations are generally present in GC×GC-MS. They are mostly due to low frequency detector noises and system fluctuations (e.g., flow, temperature). Baseline correction can be performed on unfolded GC×GC-MS data using the same methods than in 1D GC-MS. Next, smoothing is also conducted to correct for random detector noises at higher frequency. The application of smoothing algorithms generally allows to increase signal to noise ratio (S/N) and overall data quality. As for baseline correction, 1D GC-MS approaches are generally applied. Finally, even if GC×GC-MS offers an increased peak capacity, chromatographic coelutions can still occur. A proper MS deconvolution is needed to resolve such coeluting signals. GC×GC-MS are third order data. The most commonly used approaches for signal

treatment are parallel factor analysis (PARAFAC and PARAFAC2), and multivariate curve resolution with alternating least squares (MCR-ALS). All these preprocessing steps are briefly described here but they have been extensively described previously in the literature [2,9,10,17]. GC×GC-MS signal preprocessing can be classified as a minimal user impact step (Fig. 2). Indeed, most users do not usually modify basic set-ups and generally apply the default factors included in their GC×GC-MS software package.

The peak detection and reconstruction steps have a 'low to medium user impact', i.e., users' decisions do not or only slightly influence the process. The reconstruction of a chromatographic peak from different slices is based on S/N, retention times, and MS library matches. The S/N and match thresholds are user-dependent but, even if no consensus values exist, most of the studies that are found in the literature tend to use similar settings (e.g., S/N of 100 and match factor of 700) [34]. Peak identification is also conducted automatically by GC×GC-MS software. The identification generally relies on electron impact ionization mass spectra libraries (NIST, Wiley, etc.). Some software also includes linear retention indices (LRI) and mass accuracy calculation (when an HRMS detector is used). The primary peak identification is thus to be seen as a 'low user impact' step where users do not influence the identification, unless unproperly low or high matching threshold values are set (Fig. 2).

After signal preprocessing, final compound identification can be performed using a multifactor approach (Fig. 2). GC×GC-MS provides several identification factors: two retention times, a structured separation image, an electron impact-based fragmentogram, and in some cases high resolution or soft ionization MS. However, to properly exploit the multimodal identification capacities of GC×GC-MS, combining the chromatographic and MS information, manual user interventions are requested. Thus, this multi-criteria decision process is of 'high user impact'. The potential of GC×GC-MS for multimodal compounds evaluation has been described elsewhere [27,35].

2.4. Data preprocessing: alignment, normalization and scaling

Once signal preprocessing has properly taken place, data preprocessing is to be considered. As expected, GC×GC chromatograms generated from replicated injections of a same sample or a series of samples can suffer from some retention time (t_R) shifts. These fluctuations usually originate from column degradation, carrier gas pressure and temperature variations, maintenance on the instrument… The action of correction of these retention time fluctuations is called data alignment (Fig. 2).

Data alignment is a crucial part of the GC×GC-MS data processing. Interestingly, it is generally described by users as the determining step to unlock a GC×GC method's potential [36] as users usually have high expectations for alignment tools, far above the simple fact of correction of t_R fluctuations. It is in fact also often expected from data alignment that the consistent peak integration is guaranteed through the entire set of data. Indeed, slight signal fluctuations at the MS level can result in the automated selection of different parent ions used for peak integration. This interfere with the collection of accurate quantitative information across the data as a compound found across different samples needs to be integrated on the same quantitative ion or group of ions. Moreover, and especially in the context Omics, users also expect the integration of feature selection tools, the data visualization options, and the feature identification to be supported directly in the alignment tool package.

The retention time alignment itself can be addressed using manual alignment from the peak tables [37]. However, for large scale untargeted research the task can quickly become overwhelming. The development of robust alignment algorithms is primordial for the development of GC×GC-MS. With these modern alignment tools, extra-options such as the ones listed above can also be implemented, which can lead to fully and powerful integrated solutions [38].

For automated data alignment, correlation time warping and parametric time warping algorithms have been extensively used for GC data alignment [39]. These warping approaches use stretching and shifting of a 1D chromatogram in order to match a target/reference chromatogram [10]. For GC×GC-MS data alignment, multiple commercial software solutions exist. These software have recently been benchmarked using a same data set made of standards and real samples in order to estimate their relative performances over a standard data processing approach [33]. The data set is open source and available online for any user who would desire to manipulate a reference data set to estimate the robustness of a selected method [40,41]. Some of these software offers an alignment option, but it is usually limited in terms of the maximum number of samples that can be considered. Nevertheless, the continuous development of alignment solutions is a good indicator of the growing utilization of GC×GC-MS for large-scale studies. It is important to note if the main objective of these software is data alignment, the way it is practically carried out can vary. Indeed, the features to be aligned are not defined in the same way. Some software relies on data points, or tiles, or peak-regions analysis. An exhaustive definition of these different features is available in a previous review [42].

One of the first commercial solution, i.e., LECO Corp. ChromTOF™ Statistical Compare feature, relies on peak table alignment, so that preprocessing parameters such as peak matching factor and retention time shifting tolerance have an important impact on the final results [34]. Such a solution requires large computer power and might be difficult to implement for larger data sets. Nevertheless, Statistical Compare has been widely and successfully used for the last few years. One way to reduce these challenges is to use a tile-based approach, which aims to match tiles across chromatograms. A recent prototype software application has been developed based on Synovec's group research on that topic [16,43,44]. The tile approach relies on the definition of tiles which cut chromatograms into sub-pieces that are used for sample processing. The chromatographic signal is integrated for each tile, so that it is possible to compare the area from one tile to the corresponding tile in other chromatograms. By fine tuning of the tile size, it is possible to account for retention time variation, but even more importantly, detect tiles of interest by assessing the variation between chromatograms. If the tile contains a single compound, it is thus possible to extrapolate the variation to chemical features [16,42–44].

Other alignment software is based on pixel or peak region alignment, or even a combination of both). As for the tile method, this computational approach allows users to speed up the processing time and reduce computer power needs. Moreover, it exploits the image dimension of the multidimensional data (Fig. 1). The most used pixel-based and peak region based software is GC Image™ from Zoex Corp. Such software approach relies on the definition of a template of peaks, combining peak area and reliable alignment markers. Based on this template, every chromatogram is scanned in order to locate the alignment markers and to compute the geometrical transformation to align the data [8,45].

Independently to software type, manual curing of the data is required prior to pursuing any complementary processing. The removal of system artifacts is essential to ensure the quality of the data processing workflow. The data curing has a 'high user impact' as a peak removed by one user could be selected by another one. More automated solutions such as mass spectra filtration [29],

chromatographic area inclusion exclusion, blank subtraction, etc. exist to support the curing, but they are not yet widely used.

Following alignment, data quality checks need to be performed. For GC×GC-MS data, this step mostly implies missing value management [46]. The classical approach relies on the exclusion of features absent in more than 75 or 80% of the sample from one sample class. The remaining empty value can be managed by small value imputation using half of the smallest detected value.

When the data have been acquired and aligned properly, next comes the correction of the signal intensity. This can be performed through normalization, transformation, and scaling (Fig. 2). Normalization accounts for the unwanted variations in the signal of a compound [17]. For chromatographic data, these unwanted variations correspond to peak area fluctuations due to the sample itself or the analytical protocol. This can be corrected through normalization. Stefanuto et al. have demonstrated the influence of normalization on GC×GC-MS data generated from beer aroma. The correct normalization of peak intensity provides an increase in classification accuracy [18]. A proper normalization can have a large impact on overall model quality. There are different ways to conduct data normalization.

For chromatographic data, the most common approach to normalize the data is the use of internal standard [10]. This method relies on spiking samples with a known quantity of a non-native compound (isotopically labelled or not found in the sample). This method is highly powerful for targeted analysis [21,47–49]. However, this method is more complex for untargeted studies. It assumes that the concentrations and the response factor of each analyte is similar, which is usually not the case for biological samples. Nevertheless, internal standard spiking is a powerful monitoring tool for QC purposes, allowing monitoring of the instrument response to one or multiple reliable compounds through all analysis [21,30]. When no internal standard is used, sum-normalization (or total area normalization) is often applied in untargeted studies. This method divides each peak area by the total area measures in the chromatogram. This method assumes that each sample contains the same number of analytes, in the same concentration range, and with the same response factors. When these assumptions are not met, which is generally the case, total area normalization can introduce unwanted variations and trends in the data set [10]. Fig. 3 compares the impact of different normalization techniques on variable correlation. The study of variable correlation indicates if the normalization is affecting the data structure and the relationship between variables.

In the last years, probabilistic quotient normalization (PQN) has received a growing interest [10,29,30,35,50]. This method aims to determine the most probable normalization factor. It assesses the distribution of quotients using a reference sample calculated as the median chromatogram compared to the data from a test (sample) chromatogram [51]. As shown on Fig. 3, the PQN approach is not affecting the variable correlation nor the data structure.

After removing unwanted variations through the normalization process. It is important to ensure that all variables can be compared, especially for multivariate analysis. Scaling is used to equalize the contribution of each variable to the model. If the model is sensitive to scaling (e.g., principal component analysis (PCA)), the variables with the highest variance will dominate the projection and small variations may disappear [17]. In practice, variable value scaling is performed by subtracting the mean calculated for the variable, this step is called mean centering, and dividing by a dispersion metric. The most commonly used dispersion metrics for scaling are the standard deviation (autoscaling or z-score calculation), the difference between the maximum and the minimum values (range scaling), or the square root of the standard deviation (pareto

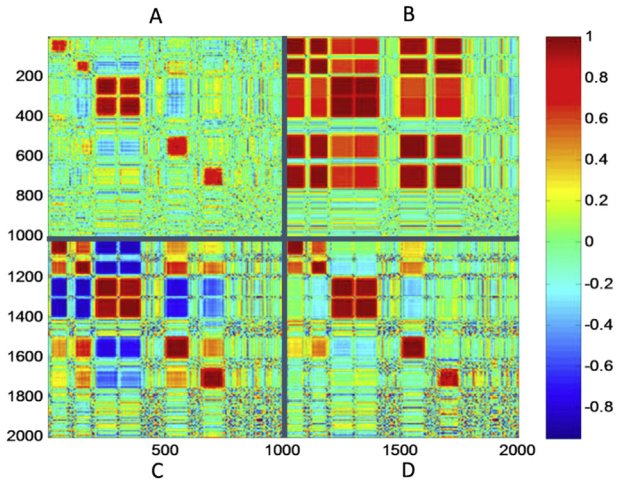

Fig. 3. The effect of two normalization procedures on the correlation structure (i.e., between features) in data. A) Correlation structure between features in the original data; B) Correlation structure between features in the data after adding effect size (i.e., instrumental/experimental error); C) Correlation structure after normalizing to total area; D) Correlation structure after normalizing using PQN. As it can be seen, the total area normalization produces spurious correlation between features, while the correlation structure after PQN normalization resembles the correlation in the original data (block A).

scaling) [17,52]. Table 1 is summarizing the different scaling approaches and their advantages.

Data transformation (e.g., log transformed) represents an alternative to scaling. Transformation aims to correct for heteroscedastic noise. Data transformation is different from scaling since it modifies individual elements rather than the complete variables [17]. Purcaro et al. described the effect of log transformation on GC×GC-MS data produced from cell cultures. In the study, they showed that the classification resolution could be improved using log data [29].

The importance of preprocessing is well accepted in the GC×GC-MS field. However, the combination of tools and the ways of implementation still need to be improved [7,53]. An important point to keep in mind is the need to properly optimize the preprocessing strategy. Corrections should only be made to correct artifacts that are present in the data and avoid generating new variations. So, the preprocessing strategy needs to be tailored based on the research question and be adapted if the question is modified. To perform this optimization, evaluation tools are required. There are different types of tools to evaluate preprocessing effects. Three are generally used in analytical chemistry. They generally rely on a classification-based research question [17].

- "trial and error" approach aims to test different preprocessing strategies and to select the one providing the best performance regarding the research question [17].
- "visualization" requires looking inside the data to see if the artifacts are corrected. For example, Pesesse et al. used PCA to evaluate batch effect optimal correction [28]. Stefanuto et al. used hierarchical cluster analysis (HCA) to optimize the normalization and feature selection parameters [18].
- "quality metrics" aims to quantify the effect of the preprocessing. A good example was developed by Purcaro et al. who used classification resolution calculated as the Euclidean distances between groups [29].

An interesting development is the implementation of unsupervised visualization tools, such as principal component analysis (PCA), in some commercially available GC×GC-MS processing

Table 1
Different scaling approaches and their advantages and disadvantages.

Method	Formula	Goal	Advantages	Disadvantages
Mean centering	$\tilde{x}_{ij} = x_{ij} - \bar{x}_i$	Substract the variable mean so the new mean is zero	Introduce a zero in the data and reduce a certain level of multicolinearity	–
Autoscaling	$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{S_i}$	Ideal for metabolite correlations	All metabolites have the same importance	Inflation of measurement errors
Range scaling	$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{x_{i\,max} - x_{i\,min}}$	Ideal for metabolite response range comparisons	All metabolites have the same importance - keep relation to the biology	Inflation of measurement errors and sensitive to outliers
Pareto scaling	$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{\sqrt{S_i}}$	Ideal to maintain data structure and to smooth outliers	Maintain the data closed to the orignal	sensitive to large fold changes

software. This will help users to monitor the effect of signal and data preprocessing directly. Other unsupervised and exploratory data analysis tools are described in the following section.

2.5. Unsupervised and exploratory data analysis

These optimization metrics rely on exploratory data analysis tools. The most commonly used are principal component analysis (PCA) and hierarchical cluster analysis (HCA) [9,10,18,54] (Fig. 4).

PCA is probably the most applied technique for the unsupervised screening of high dimensional data, such as GC×GC-MS. It aims to display the covariance structure of the data on a small number of components [56]. These components are linear combinations of the variables. The first component corresponds to the direction in which the projected observations display the larger amount of variance. The following components are orthogonal to the previous ones and also try to maximize the display variance. This way of projecting the data is, however, highly sensitive to outliers, which makes PCA a great tool for data quality check but can also limit its application. To overpass this sensibility to outliers, robust PCA techniques or machine learning tools (e.g., unsupervised random forest) have been developed [56,57]. Due to the complexity of GC×GC-MS, the implementation of such emerging techniques would be highly beneficial.

HCA is another way to visualize sample clustering using a dendrogram. In HCA, distance between samples in the variable space is measured. The distance calculation method can vary according to the data structure (e.g. Euclidean, Mahala Nobis). HCA are usually displayed as a combination of dendrogram (showing the distance between samples) and a heatmap (showing the variable intensity).

In addition to the unsupervised visualization capacity, these exploratory tools are also really useful to visualize the data structure after feature selection, regardless of the feature selection approach [2,18,28,58].

3. Features selection and model building

3.1. Univariate approaches

Biomarker selection, i.e., finding disease or condition-specific markers, is a crucial aspect of various fields. This part of the analysis usually focuses on selecting those features that are relevant to the studied problem. One way of selecting the significant features is applying univariate statistics, such as Fisher ratio (FR), parametric or nonparametric testing. Since this approach is applied into many different features, one should always use multi-testing correction, such as Bonferroni, Benjamin or Benjamini & Hochberg, for

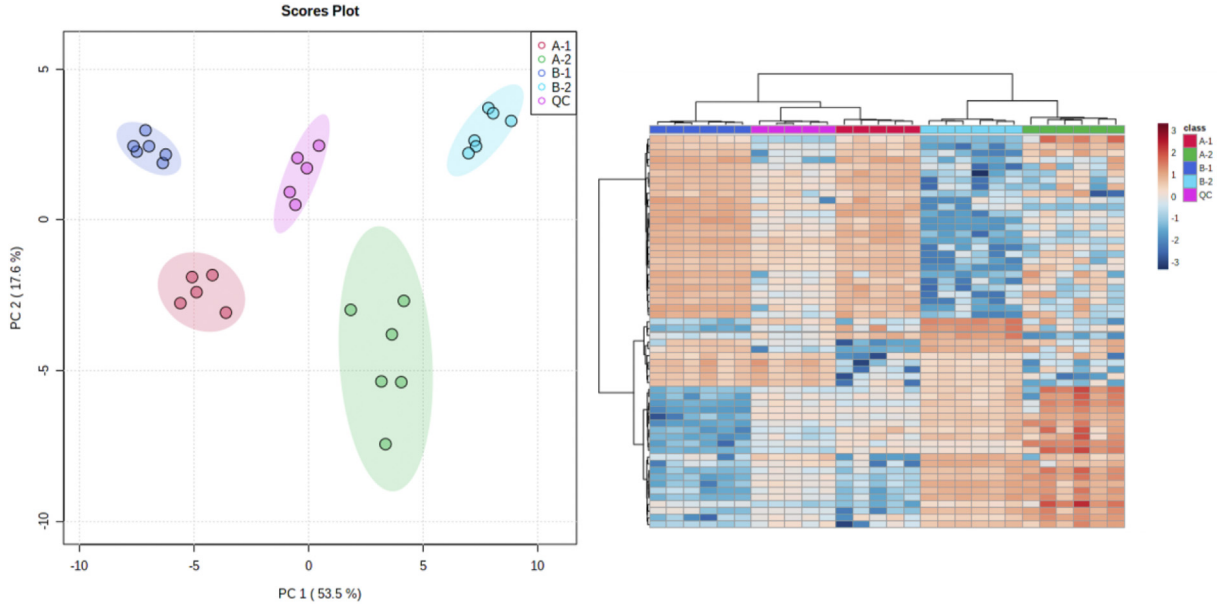

Fig. 4. Unsupervised visualization using PCA and HCA. It also shows how pool QC represent the sample variability. The data were generated by growing two bacteria (1 and 2) in two different media (A and B) using suspension cultures. The samples were filter sterilized and sampled using SPME. More information on the method can be found elsewhere [50,55].

adjusting the probabilities obtained from the statistical test. In the field of GC×GC-MS, the Fisher ratio (FR) approach is popular and has been widely used. The approach was introduced by Synovec's group in 2006, more information can be found in previous reviews [9,10,59,60]. In brief, FR is calculated as the between-class variance divided by the within-class variance. The resulting metric can be compared to F critical value to identify significant compounds [18,53]. Since its introduction, FR has been implemented in the most used alignment and chromatogram comparison tools. It has evolved from compounds FR comparison, to tile-based or peak-zone comparison. Some variants of FR have been also implemented, such as control-normalized FR analysis. In this approach the between class variance is normalized to the variance of the control class only [58]. One should keep in mind that this approach, although very simple, may discard many relevant information. Since the univariate prefiltering is based on stochastic quantities, it does not add stability to the selection done. Moreover, it can lead to overlooking important aspects of the data. The univariate prefiltering of the variables is very easy to understand and trace, which can result in quick identification of single or uncorrelated features variation. Nevertheless, one should realize that FR is only for supervised analysis, not unsupervised analysis. FR strategy should be complemented for fully untargeted investigation (e.g., for metabolomics), as it is not delivering any benefits for multivariate models [28,29]. The common way of representing significance of features is volcano plot as shown in Fig. 5. This simple visualization allows selecting features which are the most significant and with the highest fold change (top left and right corners in Fig. 5).

3.2. Multivariate approaches

Various multivariate techniques can be used to select the discriminatory/relevant features in the data. They are usually divided into linear and nonlinear techniques. As the name indicates, linear methods look for linear tendency, relation between features, while nonlinear are able to find more complex connections among features.

The widely used technique that allows selecting important features is partial least squares (PLS) analysis [61,62]. This approach

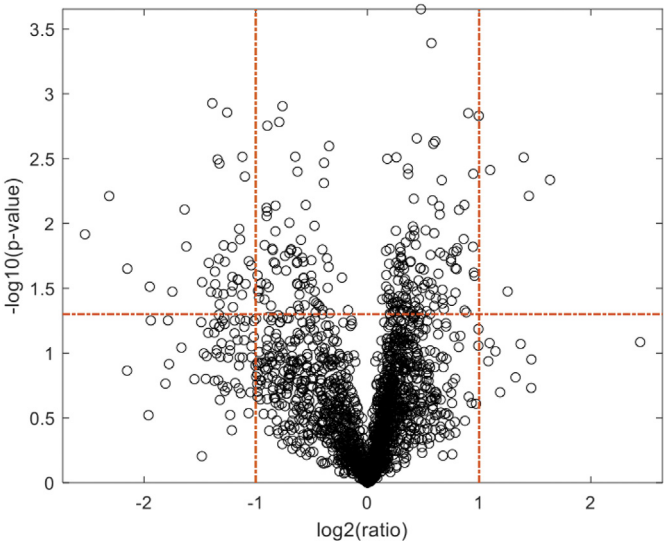

Fig. 5. Volcano plot representing relation between significance of the feature and its change fold. The red dashed lines correspond to different threshold for p-values and fold change. The significant features can be selected using either change fold, p-values or the combination of both.

is able to find relevant features for regression as well as for classification problems. It is similar to PCA, however instead of maximizing variance in the data it focuses on capturing most of the information in the data with respect to a response/class vector Y in a linear way. Consequently, PLS creates new variables, the *latent variables* (LVs) that are linear combinations of the original features, which have incorporated the response/class information of every sample, too. A very important advantage of PLS is its possibility to be applied into regression problems as well as binary but also into multiple class problems. The aid of PLS is its ability to cope with highly collinear data, thus it is a suitable technique for GC×GC-MS data. Since PLS is a supervised technique, it requires an optimization step, which aims at finding the optimal number of LVs to be used then validated before any final predictions/classifications are made. Feature importance in the PLS model can be obtained from the regression coefficient (the higher the most relevant). However, optimal threshold selection is not always straightforward. Therefore, several feature extraction procedures, which allow simple selection of a threshold, are directly available for PLS. The most common ones are variable importance in projection (VIP), selectivity ratio and significance multivariate correlation [63,64].

A modification of PLS is orthogonal PLS (OPLS) [65,66]. It is based on dividing the overall variation in the data into response predictive (i.e., linearly related to the class/response vector) and orthogonal (i.e., uncorrelated to the response). The main benefit of OPLS is the interpretation of the model since irrelevant information is directly filtered out and the relevant information is captured by first LV. The remaining variation is called structure noise and is present in the data due to differences in e.g., diet, age and gender. It is important to point out that OPLS and PLS have comparable prediction power.

GC×GC-MS is often applied to reveal complex properties of the studied data. Thus, it is expected that nonlinear relationships among features are present. Therefore, it may be preferable to implement supervised approaches that consider both linear and nonlinear relationships in the data; tree-based techniques, which are ensemble techniques, are such approaches. The most renowned and broadly used ensemble techniques are random forest (RF), adaptive boosting (AdaBoost) and gradient boosting. RF is the most commonly applied technique, which aims at building predictive models by constructing many fully-grown and decorrelated trees in the process called bagging or aggregative bootstrapping algorithm. RF enables not only building a powerful predictive model, but it provides the information on compounds that have the largest contribution into the model. The advantage of RF is the internal validation of the model by means of out-of-bag cases, which are the observations that are not used to build a tree. The overall performance of the forest is assessed by the out-of-bag error (i.e., wrongly predicted/classified out-of-bag observations) of all the trees present in the forest, and thus giving unbiased evaluation of RF performance.

RF gained its popularity thanks to its applicability to different types of data [30,35,67]. Moreover, it does not require any data scaling and is robust to any outliers present in the data. The large advantage of RF is its simplicity in usage since it does not require many parameters optimization as the default parameters suggested in various programs deliver models with relatively good performance. Another advantage of RF is the possibility of data visualization by means of proximity matrix, which is a square, similarity matrix which shows how similar each pair of samples is. Proximity in RF is obtained by calculating the number of the times two samples ended up in the same terminal node. For instance, if the RF consists of trees and the pair of samples ended up in the same terminal node in 100 of the 1000 trees, then the proximity for this pair of samples is 0.1. The proximity matrix contains numbers

between 0 and 1, while 0 indicates highest degree dissimilarity, while 1 corresponds to the opposite. Note that proximity can be further used to visualize any groups and trends in the data by applying PCA into it.

Other machine learning models have been applied for GC×GC-MS data investigation. In 2019, Reichenbach et al. have compared 22 machine learning methods to investigate wine aroma analysis using pattern-based approaches, using the full power of the GC×GC image generation dimension [68]. Using the same basis, deep learning is also an emerging direction to investigate (*see* section 5 *on deep learning).*

4. Model (cross-)validation

The usage of supervised modeling, such as PLS, OPLS or RF, requires proper validation of the technique (Fig. 2). This step is particularly relevant to ensure the certainty of the findings. There are various ways of assessing the predictive ability of the constructed supervised model [69–71], as outlined below. The most optimal approach is dividing data into a training set to train the supervised model, a validation set to optimize the supervised model, and a test set to assess the prediction power of the model. Different ways of electing those sets exist and they can be divided into two main categories, random selection and selection taking into account data distribution. The first approach, although the simplest one, leads at each run to different samples being selected into training and validation/test sets, therefore it requires multiple runs to ensure different scenarios. In order to overcome this, it is possible to define random seed during sets selection. This guarantees selecting the same samples into training and validation/test sets. The second approach looks at the distribution of the samples and aims at selecting representative training set and/or test set [72].

The division into training and validation/test sets is only possible when sufficient number of samples are available. If that is not the case the alternative approach is to use cross-validation (CV) strategy or bootstrapping. The CV can be based on simple procedure, where each time only one sample is excluded and used for testing the model while the remaining samples are used for building the model, or more complex where k-number of samples can be excluded. The important characteristics of CV is its overestimation of the predictive power of the supervised model. Hence, it is recommended to apply double CV (called nested CV) or bootstrapping.

As a final evaluation of the predictive ability of the supervised model, permutation test can be used. The permutation test is based on random rearrangement of the dependent and independent variables of the training set and the construction of a supervised model using these randomly permuted data. In the subsequent step, the prediction for test/validation set samples is obtained and the number of misclassifications/prediction error is obtained. The assumption of the permutation test is that the test set should be wrongly predicted for a randomly permuted training set.

Another crucial part of supervised techniques is the scaling of the data after division into the training, validation, and test sets. It is essential that samples used for validating the performance of the machine learning model should always be scaled (e.g., by autoscaling or pareto scaling) using parameters obtained from the training samples.

5. Model visualization and figures of merits

When properly built and (cross-)validated, models can be evaluated through their figures of merit, especially for classification ones. As for any analytical figure of merits, the strength of the metric is completely dependent on the robustness of the method. The most commonly used are accuracy, false positive rate (1-specificity), false negative rate, true positive rate (also call sensitivity), true negative rate.

Different visualization tools exist to efficiently communicate these figures of merits. The most common ones are the confusion matrix and the receiver operating characteristic curve (ROC) (Fig. 6). Area under the receiver operating characteristic curve (AUC – ROC curve) is a performance measurement for classification problem at various thresholds settings. ROC is a probability curve, and AUC represents degree or measure of separability. When AUC is high, the model performance are high too.

Interesting to point, unsupervised tools, such as PCA and HCA, can also be applied to visualize the data structure after feature selection. It allows quick comparison between before and after the selection of the features.

6. Deep learning as future perspective

GC×GC-MS itself is well known for its ability to efficiently communicate chemical composition through the 2D chromatogram image. Indeed, the structured separation gives a powerful visualization and represent a real fingerprint of the sample (Fig. 1). Based on this fingerprint it is usually possible to visually compare sample between them. However, this approach is limited by human ability to process the information, making it applicable for small population and obvious sample differences. Nevertheless, this image-based approach can be extended to large scale study using proper data handling tools.

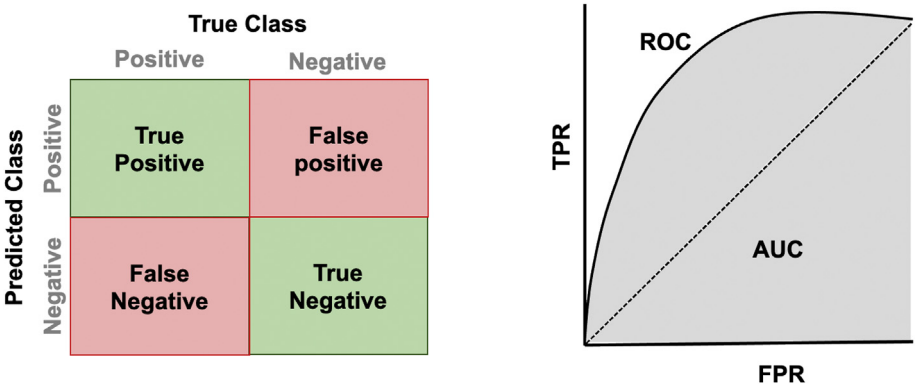

Fig. 6. Visualization tools for classification models: on the left, confusion matrix and on the right, the receiver operating characteristic (ROC) curve. TPR: True Positive Rate; FPR: False Positive Rate.

Artificial intelligence (AI) has been gaining in popularity over the last few years. In AI, computing systems carry out decisions that are usually performed by humans. Deep learning is a type of AI that concerns the ability of computers to be taught to 'think' by their own. It has often been presented as the largest development within the machine learning field. It has been successfully applied in the field of image and speech recognition. Deep learning enables to analyze the data at a more abstract level since it does not dictate any features in the data, but it aims at selecting, or crafting the relevant information. The approach taken by deep learning deviates completely from the standard techniques explained above. The deep learning algorithm is applied to data without any preprocessing, scaling or feature extraction. Deep learning teaches itself where to find relevant information in the data and ignore irrelevant parts. This very attractive characteristic of the deep learning algorithm makes it suitable for GC×GC-MS data. Yet, up till now it has not been applied to GC×GC-MS data.

Looking toward the feature application of deep learning, GC×GC-MS should definitely be the next step. This type of data can be considered as image and deep learning can omit several problems connected to data preprocessing and analysis of complex GC×GC-MS data. Therefore, if one is interested in model performance and not biological understanding of the model, deep learning, which works as a black box, would be the preferable approach.

7. Conclusion and perspectives

The continuous increase of large-scale GC×GC-MS studies has brought new challenges to the field. Nowadays, separation scientists need to develop new study designs and also manage larger data sets in a robust way. Indeed, successful, long term research requires good practice in terms of QA/QC, which is challenging for untargeted analysis.

In addition, complete control of the data workflow from acquisition to the model validation is required. This manuscript critically reviewed and discussed the literature in the intent to provide a starting guide to establish such a processing workflow. The critical steps, namely the preprocessing, including alignment, feature selection, and model validation, are identified and discussed. The critical impact of processing steps on the data structure, such as normalization, scaling, and transformation are described. Fig. 2 provides a practical workflow to guide the elaboration of robust GC×GC-MS processing workflow and inform users on the critical steps that have high impact on the model output.

In the future, the implementation of machine learning and deep learning will most probably contribute to change the way one conducts multi-dimensional chromatography, from the very early steps of optimization of analytical conditions to final model building and validation to answer the research question. Although not really yet touched, some other areas of chemometrics, such as Bayesian statistic, are worth to be investigated. Our capacity to unlock GC×GC-MS full potential is at least partly relying on the correct implementation and understanding of these chemometric tools.

Authors contribution

PH. S., A. S., and JF. F. have designed the structure of the manuscript. PH. S. and A. S. have developed the content of the manuscript. PH. S., A. S., and JF. F. have revised the manuscript.

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

This work was developed in the context of research supported by La Fondation Leon Fredericq and the FWO/FNRS Belgium EOS grant 30897864 "Chemical Information Mining in a Complex World".

References

[1] S. Nicholas, Basic Multidimensional Gas Chromatography, Else, Elsevier, 2020. http://library1.nida.ac.th/termpaper6/sd/2554/19755.pdf.
[2] S.E. Prebihalo, K.L. Berrier, C.E. Freye, H.D. Bahaghighat, N.R. Moore, D.K. Pinkerton, R.E. Synovec, Multidimensional gas chromatography: advances in instrumentation, chemometrics, and applications, Anal. Chem. 90 (2018) 505–532. https://doi.org/10.1021/acs.analchem.7b04226.
[3] J.C. Giddings, Sample dimensionality: a predictor of order-disorder in component peak distribution in multidimensional separation, J. Chromatogr. A (1995). https://doi.org/10.1016/0021-9673(95)00249-M.
[4] J.C. Giddings, Concepts and comparisons in multidimensional separation, J. High Resolut. Chromatogr. 10 (1987) 319–323. https://doi.org/10.1002/jhrc.1240100517.
[5] H.D. Bahaghighat, C.E. Freye, R.E. Synovec, Recent advances in modulator technology for comprehensive two dimensional gas chromatography, Trends Anal. Chem. 113 (2019) 379–391. https://doi.org/10.1016/j.trac.2018.04.016.
[6] P.Q. Tranchida, F.A. Franchina, P. Dugo, L. Mondello, Comprehensive two-dimensional gas chromatography-mass spectrometry: recent evolution and current trends, Mass Spectrom. Rev. (2014) 1–11. https://doi.org/10.1002/mas.21443.
[7] E.A. Higgins Keppler, C.L. Jenkins, T.J. Davis, H.D. Bean, Advances in the application of comprehensive two-dimensional gas chromatography in metabolomics, TrAC - Trends Anal. Chem. 109 (2018) 275–286. https://doi.org/10.1016/j.trac.2018.10.015.
[8] C. Cordero, J. Kiefl, S.E. Reichenbach, C. Bicchi, Characterization of odorant patterns by comprehensive two-dimensional gas chromatography: a challenge in omic studies, TrAC - Trends Anal. Chem. 113 (2019) 364–378. https://doi.org/10.1016/j.trac.2018.06.005.
[9] K.L. Berrier, S.E. Prebihalo, R.E. Synovec, Advanced data handling in comprehensive two-dimensional gas chromatography, in: Sep. Sci. Technol. (New York), 2020. https://doi.org/10.1016/B978-0-12-813745-1.00007-6.
[10] P.E. Sudol, K.M. Pierce, S.E. Prebihalo, K.J. Skogerboe, B.W. Wright, R.E. Synovec, Development of gas chromatographic pattern recognition and classification tools for compliance and forensic analyses of fuels: a review, Anal. Chim. Acta 1132 (2020) 157–186. https://doi.org/10.1016/j.aca.2020.07.027.
[11] Z. Zeng, J. Li, H.M. Hugel, G. Xu, P.J. Marriott, Interpretation of comprehensive two-dimensional gas chromatography data using advanced chemometrics, Trends Anal. Chem. 53 (2014) 150–166. https://doi.org/10.1016/j.trac.2013.08.009.
[12] E. Szymańska, J. Gerretzen, J. Engel, B. Geurts, L. Blanchet, L.M.C. Buydens, Chemometrics and qualitative analysis have a vibrant relationship, TrAC - Trends Anal. Chem. (2015). https://doi.org/10.1016/j.trac.2015.02.015.
[13] T.M. Gröger, U. Käfer, R. Zimmermann, Gas chromatography in combination with fast high-resolution time-of-flight mass spectrometry: technical overview and perspectives for data visualization, TrAC - Trends Anal. Chem. (2020). https://doi.org/10.1016/j.trac.2019.115677.
[14] A. Giri, M. Coutriade, A. Racaud, K. Okuda, J. Dane, R.B. Cody, J.F. Focant, Molecular characterization of volatiles and petrochemical base oils by photoionization GC×GC-TOF-MS, Anal. Chem. 89 (2017) 5395–5403. https://doi.org/10.1021/acs.analchem.7b00124.
[15] A. Giri, M. Coutriade, A. Racaud, P.-H. Stefanuto, K. Okuda, J. Dane, R.B. Cody, J.-F. Focant, Compositional elucidation of heavy petroleum base oil by GC × GC-EI/PI/CI/FI-TOFMS, J. Mass Spectrom. 54 (2019). https://doi.org/10.1002/jms.4319.
[16] C.E. Freye, N.R. Moore, R.E. Synovec, Enhancing the chemical selectivity in discovery-based analysis with tandem ionization time-of-flight mass spectrometry detection for comprehensive two-dimensional gas chromatography, J. Chromatogr. A (2018). https://doi.org/10.1016/j.chroma.2018.01.008.
[17] 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. https://doi.org/10.1016/j.trac.2013.04.015.
[18] P.H. Stefanuto, K.A. Perrault, L.M. Dubois, B. L'Homme, C. Allen, C. Loughnane, N. Ochiai, J.F. Focant, Advanced method optimization for volatile aroma profiling of beer using two-dimensional gas chromatography time-of-flight mass spectrometry, J. Chromatogr. A 1507 (2017) 45–52. https://doi.org/10.1016/j.chroma.2017.05.064.
[19] G. Purcaro, P.-H. Stefanuto, F.A. Franchina, M. Beccaria, W.F. Wieland-Alter, P.F. Wright, J.E. Hill, SPME-GC×GC-TOF MS fingerprint of virally-infected cell culture: sample preparation optimization and data processing evaluation, Anal. Chim. Acta (2018). https://doi.org/10.1016/j.aca.2018.03.037.
[20] W.B. Dunn, I.D. Wilson, A.W. Nicholls, D. Broadhurst, The importance of experimental design and QC samples in large-scale and MS-driven untargeted metabolomic studies of humans, Bioanalysis 4 (2012) 2249–2264. https://doi.org/10.4155/bio.12.204.

[21] F.A. Franchina, L.M. Dubois, J.F. Focant, In-depth Cannabis multiclass metabolite profiling using sorptive extraction and multidimensional gas chromatography with low- A nd high-resolution mass spectrometry, Anal. Chem. (2020). https://doi.org/10.1021/acs.analchem.0c01301.
[22] F.L. Dorman, P.D. Schettler, L.A. Vogt, J.W. Cochran, Using computer modeling to predict and optimize separations for comprehensive two-dimensional gas chromatography, J. Chromatogr. A 1186 (2008) 196–201. https://doi.org/10.1016/j.chroma.2007.12.039.
[23] A.C.A. Silva, H. Ebrahimi-Najafadabi, T.M. McGinitie, A. Casilli, H.M.G. Pereira, F.R. Aquino Neto, J.J. Harynuk, Thermodynamic-based retention time predictions of endogenous steroids in comprehensive two-dimensional gas chromatography, Anal. Bioanal. Chem. 407 (2015) 4091–4099. https://doi.org/10.1007/s00216-015-8627-0.
[24] T.M. McGinitie, H. Ebrahimi-Najafabadi, J.J. Harynuk, Rapid determination of thermodynamic parameters from one-dimensional programmed-temperature gas chromatography for use in retention time prediction in comprehensive multidimensional chromatography, J. Chromatogr. A 1325 (2014) 204–212. https://doi.org/10.1016/j.chroma.2013.12.008.
[25] S. Hou, K.A.J.M. Stevenson, J.J. Harynuk, A simple, fast, and accurate thermodynamic-based approach for transfer and prediction of gas chromatography retention times between columns and instruments Part I: estimation of reference column geometry and thermodynamic parameters, J. Sep. Sci. 41 (2018) 2544–2552. https://doi.org/10.1002/jssc.201701343.
[26] T.M. McGinitie, J.J. Harynuk, Prediction of retention times in comprehensive two-dimensional gas chromatography using thermodynamic models, J. Chromatogr. A 1255 (2012) 184–189. https://doi.org/10.1016/j.chroma.2012.02.023.
[27] P.H. Stefanuto, J.F. Focant, Columns and column configurations, in: Sep. Sci. Technol. (New York), 2020. https://doi.org/10.1016/B978-0-12-813745-1.00003-9.
[28] R. Pesesse, P.-H. Stefanuto, F. Schleich, R. Louis, J.-F. Focant, Multimodal chemometric approach for the analysis of human exhaled breath in lung cancer patients by TD-GC×GC-TOFMS, J. Chromatogr. B Anal. Technol. Biomed. Life Sci. (2019) 1114–1115. https://doi.org/10.1016/j.jchromb.2019.01.029.
[29] G. Purcaro, P.-H. Stefanuto, F.A. Franchina, M. Beccaria, W.F. Wieland-Alter, P.F. Wright, J.E. Hill, SPME-GC×GC-TOF MS fingerprint of virally-infected cell culture: sample preparation optimization and data processing evaluation, Anal. Chim. Acta (2018) 1–10. https://doi.org/10.1016/j.aca.2018.03.037.
[30] F.N. Schleich, D. Zanella, P.-H. Stefanuto, K. Bessonov, A. Smolinska, J.W. Dallinga, M. Henket, V. Paulus, F. Guissard, S. Graff, C. Moermans, E.F.M. Wouters, K. Van Steen, F.-J. van Schooten, J.-F. Focant, R. Louis, Exhaled volatile organic compounds are able to discriminate between neutrophilic and eosinophilic asthma, Am. J. Respir. Crit. Care Med. (2019). https://doi.org/10.1164/rccm.201811-2210OC. rccm.201811-2210OC.
[31] G. Stavropoulos, D.M.A.E. Jonkers, Z. Mujagic, G.H. Koek, A.A.M. Masclee, M.J. Pierik, J.W. Dallinga, F.J. Van Schooten, A. Smolinska, Implementation of quality controls is essential to prevent batch effects in breathomics data and allow for cross-study comparisons, J. Breath Res. (2020). https://doi.org/10.1088/1752-7163/ab7b8d.
[32] S. Castillo, I. Mattila, J. Miettinen, M. Orešič, T. Hyötyläinen, Data analysis tool for comprehensive two-dimensional gas chromatography/time-of-flight mass spectrometry, Anal. Chem. 83 (2011) 3058–3067. https://doi.org/10.1021/ac103308x.
[33] B.A. Weggler, L.M. Dubois, N. Gawlitta, T. Gröger, J. Moncur, L. Mondello, S. Reichenbach, P. Tranchida, Z. Zhao, R. Zimmermann, M. Zoccali, J.-F. Focant, A unique data analysis framework and open source benchmark data set for the analysis of comprehensive two-dimensional gas chromatography software, J. Chromatogr. A (2020). https://doi.org/10.1016/j.chroma.2020.461721.
[34] H.D. Bean, J.E. Hill, J.-M.D. Dimandja, Improving the quality of biomarker candidates in untargeted metabolomics via peak table-based alignment of comprehensive two-dimensional gas chromatography–mass spectrometry data, J. Chromatogr. A 1394 (2015) 111–117. https://doi.org/10.1016/j.chroma.2015.03.001.
[35] P.H. Stefanuto, D. Zanella, J. Vercammen, M. Henket, F. Schleich, R. Louis, J.F. Focant, Multimodal combination of GC × GC-HRTOFMS and SIFT-MS for asthma phenotyping using exhaled breath, Sci. Rep. 10 (2020) 1–11. https://doi.org/10.1038/s41598-020-73408-2.
[36] P.-H. Stefanuto, K.A. Perrault, D. Stoll, Multidimensional Chromatography Workshop, 2019. www.multidimensionalchromatography.com.
[37] S. Stadler, P.-H. Stefanuto, M. Brokl, S.L. Forbes, J.-F. Focant, Characterization of volatile organic compounds from human analogue decomposition using thermal desorption coupled to comprehensive two-dimensional gas chromatography-time-of-flight mass spectrometry, Anal. Chem. 85 (2013) 998–1005. https://doi.org/10.1021/ac302614y.
[38] M.J. Wilde, B. Zhao, R.L. Cordell, W. Ibrahim, A. Singapuri, N.J. Greening, C.E. Brightling, S. Siddiqui, P.S. Monks, R.C. Free, Automating and extending comprehensive two-dimensional gas chromatography data processing by interfacing open-source and commercial software, Anal. Chem. (2020). https://doi.org/10.1021/acs.analchem.0c02844.
[39] Z. Bankó, J. Abonyi, Correlation based dynamic time warping of multivariate time series, Expert Syst. Appl. (2012). https://doi.org/10.1016/j.eswa.2012.05.012.
[40] L. Dubois, D. Zanella, F. Franchina, J.-F. Focant, P.-H. Stefanuto, Dataset_GCxGC_FruityBeer, 2021. https://doi.org/10.7910/DVN/RJTYEO.
[41] L. Dubois, D. Zanella, F. Franchina, J.-F. Focant, P.-H. Stefanuto, Methods_GCxGC_FruityBeer, 2021. https://doi.org/10.7910/DVN/U6HYH0.
[42] F. Stilo, C. Bicchi, A.M. Jimenez-Carvelo, L. Cuadros-Rodriguez, S.E. Reichenbach, C. Cordero, Chromatographic fingerprinting by comprehensive two-dimensional chromatography: fundamentals and tools, TrAC Trends Anal. Chem. (2020). https://doi.org/10.1016/j.trac.2020.116133.
[43] B.A. Parsons, L.C. Marney, W.C. Siegler, J.C. Hoggard, B.W. Wright, R.E. Synovec, Tile-based Fisher ratio analysis of comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry (GC × GC–TOFMS) data using a null distribution approach, Anal. Chem. 87 (2015) 3812–3819. http://pubs.acs.org/doi/abs/10.1021/ac504472s.
[44] B.A. Parsons, D.K. Pinkerton, B.W. Wright, R.E. Synovec, Chemical characterization of the acid alteration of diesel fuel: non-targeted analysis by two-dimensional gas chromatography coupled with time-of-flight mass spectrometry with tile-based Fisher ratio and combinatorial threshold determination, J. Chromatogr. A 1440 (2016) 179–190. https://doi.org/10.1016/j.chroma.2016.02.067.
[45] S.E. Reichenbach, X. Tian, Q. Tao, E.B. Ledford Jr., Z. Wu, O. Fiehn, Informatics for cross-sample analysis with comprehensive two-dimensional gas chromatography and high-resolution mass spectrometry (GCxGCâ€"HRMS), Talanta 83 (2011) 1279–1288. https://doi.org/10.1016/j.talanta.2010.09.057.
[46] P. Gromski, Y. Xu, H. Kotze, E. Correa, D. Ellis, E. Armitage, M. Turner, R. Goodacre, Influence of missing values substitutes on multivariate analysis of metabolomics data, Metabolites 4 (2014) 433–452. https://doi.org/10.3390/metabo4020433.
[47] J.F. Focant, G. Eppe, M.L. Scippo, A.C. Massart, C. Pirard, G. Maghuin-Rogister, E. De Pauw, Comprehensive two-dimensional gas chromatography with isotope dilution time-of-flight mass spectrometry for the measurement of dioxins and polychlorinated biphenyls in foodstuffs: comparison with other methods, J. Chromatogr. A 1086 (2005) 45–60. https://doi.org/10.1016/j.chroma.2005.05.090.
[48] K.A. Perrault, P.-H. Stefanuto, B.H. Stuart, T. Rai, J.-F. Focant, S.L. Forbes, Detection of decomposition volatile organic compounds in soil following removal of remains from a surface deposition site, Forensic Sci. Med. Pathol. 11 (2015) 376–387. https://doi.org/10.1007/s12024-015-9693-5.
[49] P.-H. Stefanuto, K.A. Perrault, R.M. Lloyd, B.H. Stuart, T. Rai, S.L. Forbes, J.-F. Focant, Exploring new dimensions in cadaveric decomposition odour analysis, Anal. Methods. 7 (2015) 2287–2294. https://doi.org/10.1039/C5AY00371G.
[50] C.A. Rees, A. Burklund, P.-H. Stefanuto, J.D. Schwartzman, J.E. Hill, Comprehensive volatile metabolic fingerprinting of bacterial and fungal pathogen groups, J. Breath Res. 12 (2018). https://doi.org/10.1088/1752-7163/aa8f7f.
[51] F. Dieterle, A. Ross, G. Schlotterbeck, H. Senn, Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in1H NMR metabonomics, Anal. Chem. 78 (2006) 4281–4290. https://doi.org/10.1021/ac051632c.
[52] R.A. van den Berg, H.C.J. Hoefsloot, J.A. Westerhuis, A.K. Smilde, M.J. van der Werf, Centering, scaling, and transformations: improving the biological information content of metabolomics data, BMC Genom. (2006). https://doi.org/10.1186/1471-2164-7-142.
[53] P.H. Stefanuto, K.A. Perrault, S. Stadler, R. Pesesse, H.N. LeBlanc, S.L. Forbes, J.F. Focant, GCxGC-TOFMS and supervised multivariate approaches to study human cadaveric decomposition olfactive signatures, Anal. Bioanal. Chem. 407 (2015) 4767–4778.
[54] K.M. Pierce, J.L. Hope, K.J. Johnson, B.W. Wright, R.E. Synovec, Classification of gasoline data obtained by gas chromatography using a piecewise alignment algorithm combined with feature selection and principal component analysis, J. Chromatogr. A 1096 (2005) 101–110. http://linkinghub.elsevier.com/retrieve/pii/S0021967305008484.
[55] C.A. Rees, P.-H. Stefanuto, S.R. Beattie, K.M. Bultman, R.A. Cramer, J.E. Hill, Sniffing out the hypoxia volatile metabolic signature of Aspergillus fumigatus, J. Breath Res. 11 (2017). https://doi.org/10.1088/1752-7163/aa7b3e.
[56] M. Hubert, P.J. Rousseeuw, K. Vanden Branden, ROBPCA: a new approach to robust principal component analysis, Technometrics (2005). https://doi.org/10.1198/004017004000000563.
[57] N.L. Afanador, A. Smolinska, T.N. Tran, L. Blanchet, Unsupervised random forest: a tutorial with case studies, J. Chemom. 30 (2016) 232–241. https://doi.org/10.1002/cem.2790.
[58] S.E. Prebihalo, G.S. Ochoa, K.L. Berrier, K.J. Skogerboe, K.L. Cameron, J.R. Trump, S.J. Svoboda, J.K. Wickiser, R.E. Synovec, Control-normalized Fisher ratio analysis of comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry data for enhanced biomarker discovery in a metabolomic study of orthopedic Knee-Ligament injury, Anal. Chem. (2020). https://doi.org/10.1021/acs.analchem.0c03456.
[59] K.M. Pierce, J.C. Hoggard, J.L. Hope, P.M. Rainey, A.N. Hoofnagle, R.M. Jack, B.W. Wright, R.E. Synovec, Fisher ratio method applied to third-order separation data to identify significant chemical components of metabolite extracts, Anal. Chem. 78 (2006) 5068–5075. http://pubs.acs.org/doi/abs/10.1021/ac0602625.
[60] K.M. Pierce, B. Kehimkar, L.C. Marney, J.C. Hoggard, R.E. Synovec, Review of chemometric analysis techniques for comprehensive two dimensional separations data, J. Chromatogr. A 1255 (2012) 3–11. https://doi.org/10.1016/j.chroma.2012.05.050.
[61] S. Wold, M. Sjöström, L. Eriksson, PLS-regression: a basic tool of chemometrics, Chemom. Intell. Lab. Syst. (2001). https://doi.org/10.1016/S0169-7439(01)00155-1.
[62] P.S. Gromski, H. Muhamadali, D.I. Ellis, Y. Xu, E. Correa, M.L. Turner, R. Goodacre, A tutorial review: metabolomics and partial least squares-

discriminant analysis - a marriage of convenience or a shotgun wedding, Anal. Chim. Acta (2015). https://doi.org/10.1016/j.aca.2015.02.012.
[63] M. Farrés, S. Platikanov, S. Tsakovski, R. Tauler, Comparison of the variable importance in projection (VIP) and of the selectivity ratio (SR) methods for variable selection and interpretation, J. Chemom. (2015). https://doi.org/10.1002/cem.2736.
[64] T.N. Tran, N.L. Afanador, L.M.C. Buydens, L. Blanchet, Interpretation of variable importance in partial least squares with significance multivariate correlation (sMC), Chemom. Intell. Lab. Syst. (2014). https://doi.org/10.1016/j.chemolab.2014.08.005.
[65] J. Trygg, S. Wold, Orthogonal projections to latent structures (O-PLS), J. Chemom. (2002). https://doi.org/10.1002/cem.695.
[66] N. Di Giovanni, M.A. Meuwis, E. Louis, J.F. Focant, Untargeted serum metabolic profiling by comprehensive two-dimensional gas chromatography-high-resolution time-of-flight mass spectrometry, J. Proteome Res. 19 (2020) 1013−1028. https://doi.org/10.1021/acs.jproteome.9b00535.
[67] M. Nasir, H.D. Bean, A. Smolinska, C.A. Rees, E.T. Zemanick, J.E. Hill, Volatile molecules from bronchoalveolar lavage fluid can “rule-in” Pseudomonas aeruginosa and “rule-out” Staphylococcus aureus infections in cystic fibrosis patients, Sci. Rep. (2018). https://doi.org/10.1038/s41598-017-18491-8.
[68] S.E. Reichenbach, C.A. Zini, K.P. Nicolli, J.E. Welke, C. Cordero, Q. Tao, Benchmarking machine learning methods for comprehensive chemical fingerprinting and pattern recognition, J. Chromatogr. A (2019). https://doi.org/10.1016/j.chroma.2019.02.027.
[69] S. Patil, A. Patil, V.M. Phalle, Life prediction of bearing by using adaboost regressor, SSRN Electron. J. (2019). https://doi.org/10.2139/ssrn.3398399.
[70] S. González, F. Herrera, S. García, Managing monotonicity in classification by a pruned adaboost, in: Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), 2016. https://doi.org/10.1007/978-3-319-32034-2_43.
[71] V.F. Rodriguez-Galiano, B. Ghimire, J. Rogan, M. Chica-Olmo, J.P. Rigol-Sanchez, An assessment of the effectiveness of a random forest classifier for land-cover classification, ISPRS J. Photogramm. Remote Sens. (2012). https://doi.org/10.1016/j.isprsjprs.2011.11.002.
[72] F. Westad, F. Marini, Validation of chemometric models - a tutorial, Anal. Chim. Acta (2015). https://doi.org/10.1016/j.aca.2015.06.056.