Journal of Chromatography A 1711 (2023) 464467

Contents lists available at ScienceDirect

Journal of Chromatography A

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

Theoretical modeling and machine learning-based data processing workflows in comprehensive two-dimensional gas chromatography—A review

Meriem Gaida *, Pierre-Hugues Stefanuto, Jean-François Focant

Organic and Biological Analytical Chemistry Group (OBiAChem), MolSys Research Unit, Liège University, Belgium

ARTICLE INFO

Keywords:
Comprehensive two-dimensional gas chromatography
Method development
Modeling
Data processing
Machine Learning

ABSTRACT

In recent years, comprehensive two-dimensional gas chromatography (GC × GC) has been gradually gaining prominence as a preferred method for the analysis of complex samples due to its higher peak capacity and resolution power compared to conventional gas chromatography (GC). Nonetheless, to fully benefit from the capabilities of GC × GC, a holistic approach to method development and data processing is essential for a successful and informative analysis. Method development enables the fine-tuning of the chromatographic separation, resulting in high-quality data. While generating such data is pivotal, it does not necessarily guarantee that meaningful information will be extracted from it. To this end, the first part of this manuscript reviews the importance of theoretical modeling in achieving good optimization of the separation conditions, ultimately improving the quality of the chromatographic separation. Multiple theoretical modeling approaches are discussed, with a special focus on thermodynamic-based modeling. The second part of this review highlights the importance of establishing robust data processing workflows, with a special emphasis on the use of advanced data processing tools such as, Machine Learning (ML) algorithms. Three widely used ML algorithms are discussed: Random Forest (RF), Support Vector Machine (SVM), and Partial Least Square–Discriminate Analysis (PLS-DA), highlighting their role in discovery-based analysis.

1. Introduction

The origins of separation science can be traced back to the early beginnings of the 20th century. Walking down memory lane, one cannot ignore the remarkable evolution that the field has undergone over the course of many decades. Even though the early-developed methods are drastically different from those used in the field today, the basics remain conceptually the same: separating multiple components of a mixture based on their preferential interactions with a specific material [1]. It was not until 1957 that modern gas chromatography (GC) made its debut in the field of separation science thanks to Michael Golay [1,2]. GC promptly attracted a lot of attention due to its versatility and ability to provide fast and accurate separations [3]. It was soon recognized as a valuable analytical technique and a method of choice in various fields, including food, pharmacology, environmental science, and forensics, to name a few [4]. The fundamental concepts of GC are easy to grasp. GC is a technique that aims at separating volatile and semi-volatile organic compounds (VOCs) present within a given sample based on their interactions with a chromatographic column. This separation is achieved by using a chromatographic column containing a special coating, known as the stationary phase, and a mobile phase, typically an inert gas, to ensure analyte migration through the column. In GC, the choice of stationary phase is critical because it determines the selectivity of the separation process. This selectivity relies on the type of interactions between the analyte and the stationary phase, including various forces such as dipole interactions, hydrogen bonding, and van der Waals forces. To optimize the separation, it is essential that the selected stationary phase closely matches the properties of the compounds being separated. A wide variety of columns are commercially available, each coated with stationary phases of different polarities. Hence, by selecting a column with a specific polarity, it is possible to fine-tune the separation of analytes and to improve the chromatographic performance. Practically, upon injection of a VOC-containing sample, the sample progresses through the column, and the VOCs interact with the stationary phase. Commonly, compounds with lower boiling points, i.e. more volatile compounds, migrate faster through the column and result in shorter retention times. Conversely, compounds with higher boiling points, i.e. less volatile compounds, interact more with the column coating and

* Corresponding author.

https://doi.org/10.1016/j.chroma.2023.464467
Received 24 June 2023; Received in revised form 15 October 2023; Accepted 17 October 2023
Available online 19 October 2023
0021-9673/© 2023 Elsevier B.V. All rights reserved.

hence elute later. Note that, while the volatility of the compounds can greatly influence the order of elution, it is however not the sole determinant. As mentioned earlier, the selectivity of the stationary phase also influences the separation process. This is why GC can effectively separate compounds with similar volatility, i.e. boiling points, but different chemical structures based on their interactions with the stationary phase [5].

Refusal to settle for the status quo and acknowledgment of the shortcomings of GC, including limited peak capacity, reduced resolution power, and lower sensitivity, led to the introduction of comprehensive two-dimensional gas chromatography (GC × GC) by Liu and Philips in 1991, nearly three decades after GC was first introduced [6]. Despite both methods relying on the same fundamental separation protocol: injection, analyte separation with a chromatographic column, and detection, GC × GC remains a more sophisticated technique from a hardware perspective. GC × GC uses two columns with different selectivities to separate compounds based on different chemical properties. Commonly, the first-dimension (^{1}D) column is of low to mid-polarity, while the second-dimension (^{2}D) column is of mid to high polarity. Occasionally, reversed-phase column sets, i.e. the polarity of both dimensions is inverted, may be used to accommodate specific sample requirements [7]. Both columns are connected in series using a special interface called a modulator, the heart of the GC × GC separation. The modulator ensures that the ^{1}D effluent is repeatedly trapped and released into narrower bands, which are then reinjected into the ^{2}D column during a process called modulation. The duration of each modulation cycle of trapping and releasing is defined by a modulation period (P_M) [8,9]. Due to the mass conservation principle, the modulator, with its focusing effects, helps improve the detection limits of the technique by increasing the signal-to-noise ratio (S/N), thereby enhancing sensitivity. Moreover, one of the key advantages of GC × GC resides in its increased peak capacity. This makes it a highly suitable method to separate, with good resolution, complex samples that for instance would not be accurately resolved by conventional one-dimensional GC (1D-GC) due to overlaps of co-eluting peaks. Coupling GC × GC with other techniques, specifically mass spectrometry (MS) significantly enhances the analytical capabilities of the technique and widens its scope of applications. For instance, in the academic setting, GC × GC–MS has become a go-to method in various omics fields due to its efficiency in identifying potential biomarkers, making it a potentially useful non-invasive medical diagnostic tool [10–12]. Owing to their high acquisition rates, i.e. 100 to 500 spectra per second, time-of-flight-MS (ToF-MS) detectors enable accurate reconstruction of chromatographic peaks, including narrow and rapidly eluting ^{2}D peaks, thereby enhancing the sensitivity of the separation. These detectors are frequently used for both targeted and untargeted analyses due to their ability to acquire full mass range spectra with good sensitivity making them indispensable for analyzing complex matrices. Moreover, the introduction of fast scanning quadrupole MS (QMS) has further expanded the GC × GC scope of applications thanks to their high scanning rates (up to 10,000 u/s) [13].

The increased GC × GC resolving power is often perceived as a double-edged sword. On the one hand, the addition of a second separation dimension can significantly increase the complexity of the optimization process, as it introduces additional experimental parameters that need to be acknowledged. The interactions between all these parameters can swiftly become difficult to assess, hence requiring attentive consideration during the method development stage [7,14]. On the other hand, the large amount of data generated during the chromatographic separation can at times be overwhelming for researchers who are interested in the method but lack expertise in GC × GC, leading to concerns about how to interpret the data. Method development is an essential step in any analytical workflow as it enables the generation of high-quality data (Fig. 1). Nevertheless, method development alone cannot guarantee the efficiency of the data processing step. In this regard, to fully leverage the potential of GC × GC, a holistic approach to both method development and data processing is important. In other words, the two steps need to be optimized in tandem to ensure an

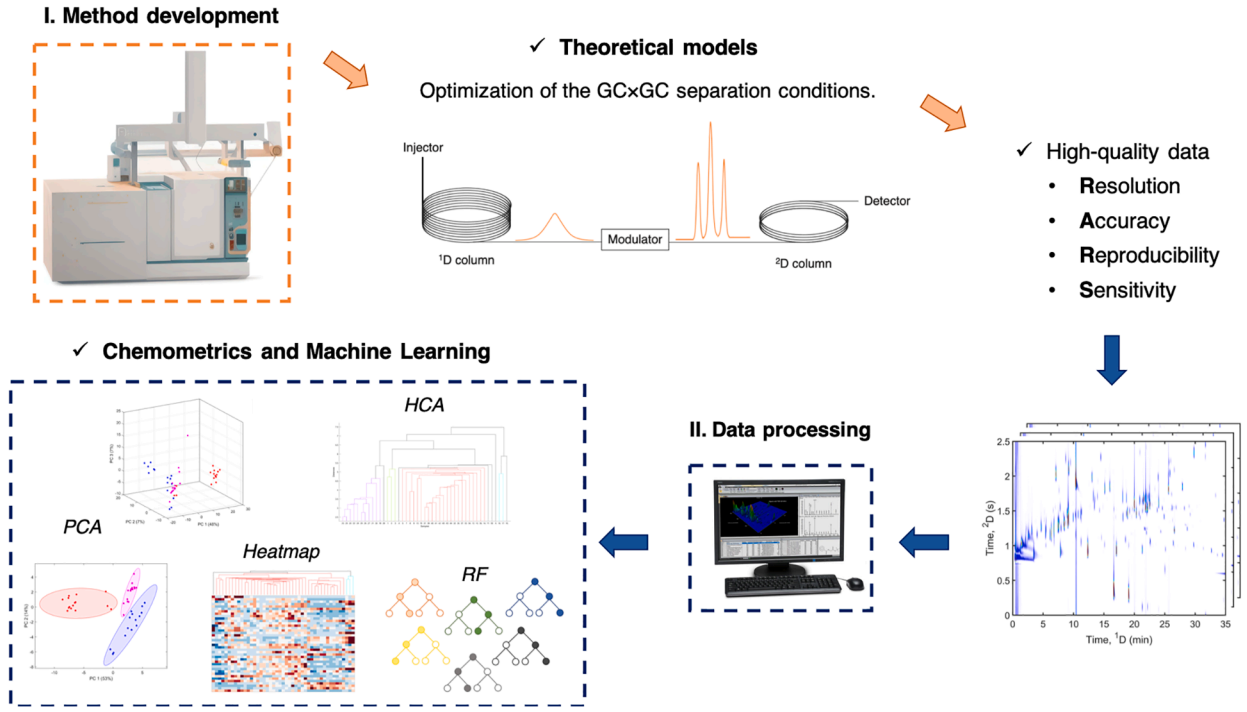

Fig. 1. Flow chart highlighting the synergistic relationship between method development and data processing. This figure illustrates the iterative process of developing a robust analytical workflow for GC × GC separations. As part of method development, theoretical modeling enables the optimization of the GC × GC experimental conditions and enhances the quality of the chromatographic separation, leading to the generation of high-quality data. In the data processing step, the combined use of chemometrics and machine learning algorithms helps alleviate the complexity of the produced data and enables the depiction of meaningful chemical information from complex datasets.

efficient and robust analysis.

Theoretical modeling plays an important role in method development in GC in general and in GC × GC in particular. One of its key benefits is its ability to help understand how different experimental parameters, such as the temperature program and the carrier gas flow rate, impact the chromatographic separation quality [15]. Additionally, it can guide the selection of appropriate column sets that help to achieve a good orthogonal separation mechanism [16]. Furthermore, theoretical modeling can aid in the prediction of retention times and peak shapes of analytes in complex mixtures, facilitating the potential identification of unknown compounds [17,18]. Ultimately, these advantages aim to optimize the separation conditions to achieve the best possible separation space occupation. Developing new theoretical models is a highly beneficial step in any analytical chemistry workflow. Nevertheless, as previously mentioned, generating high-quality data does not guarantee that the data will be easily processed and interpreted. Therefore, it is important to develop robust data processing workflows suitable for the analysis of the datasets at hand. In recent years, machine learning (ML) algorithms have become increasingly popular in GC × GC as they have demonstrated high capabilities in dealing with large sets of data and are progressively proving to be a great asset in extracting meaningful chemical information [19–23]. Their versatility makes them suitable to perform multiple tasks such as feature selection, pattern recognition, prediction, and classification. Additionally, they can greatly reduce the time and effort allotted to data analysis compared to mainstream methods that tend to be time-consuming and more prone to human error [24]. Furthermore, it is worth noting that their use can also be extended beyond data processing as they can play an important role in method development by predicting analyte retention, as highlighted by previous studies [25–27]. However, for the scope of this review, we focus exclusively on the application of ML algorithms in the data processing step.

The present review is structured into two main sections. The first section provides a critical evaluation of the current state of theoretical modeling in GC × GC by providing an overview of the conducted research and highlighting the challenges that lie ahead. Particular attention is paid to thermodynamic modeling. In the second section, we highlight the importance of robust data processing workflows in extracting meaningful chemical information. Specifically, we focus on the rise of ML algorithms as efficient and powerful tools for handling large datasets. We discuss their advantages over traditional statistical methods and provide a comprehensive understanding of how they can be incorporated into the data processing workflow thoughtfully and informedly.

2. Theoretical modeling

Theoretical modeling in GC × GC dates back to the early days of the technique and the first prediction model was pioneered by Beens et al. in 1998 [28]. The perpetual concern of every theoretical work conducted in the frame of GC × GC was and still is to harness the wealth of knowledge acquired over the years from GC modeling and effectively transpose it to a more complex separation system. Theoretical modeling in GC × GC reached its peak between the early 2000s and the early 2010s, with few studies published outside of this time frame. The present review does not intend to provide a detailed explanation of the work conducted during this period, nor to provide the explicit mathematical formulation of every prediction model. For more in-depth explanations, the readers are directed to the appropriate references. However, the focus of this review is to outline what was accomplished in the field of theoretical modeling in GC × GC and identify the research topics that still require further attention. A special focus is attributed to thermodynamic-based modeling, as it is one of the most commonly used approaches in the field. Note that despite the growing popularity of GC × GC, there is a marked scarcity of literature reviews focused on the subject of theoretical modeling of GC × GC separations and consequently retention time predictions. This particular subject is often briefly and superficially mentioned in the context of 1D-GC advancements [29].

Theoretical modeling has made significant advancements over the years. Multiple efforts were dedicated to developing approaches capable of accurately describing all aspects of the chromatographic separations. This yielded a wide variety of methods, from models describing the kinetics and thermodynamics of the separation to computer-based models [30–34]. Despite these achievements, there is still room for improvement as some aspects may require further refinement to achieve greater accuracy. In recent years, it is apparent that the main focus of research in GC × GC has shifted towards the practical applications of the technique, as well as data processing. This is portrayed by the number of excellent and highly-informative review papers that were recently published in this regard [35–39]. This shift in focus is most likely caused by the introduction of advanced hardware, which encompasses cutting-edge technological improvements in the modulator structure and functionalities, as well as the use of high-resolution MS detectors. Consequently, it appears that theoretical modeling has taken a backseat to technological progress. While modeling is currently no longer a major focus of research, it remains an important aspect of method development, as it ensures a thorough understanding of the separation mechanism and helps assess the full potential of the method. Thus, it is important to find a balance between the urgent need for applications and the importance of fundamental research to avoid fast and short-term fixes.

2.1. Early-developed models

Over the years, numerous theoretical models were introduced to provide a thorough understanding of various aspects of GC × GC. In this section, we provide a brief compilation of some of the early-developed prediction models and briefly discuss their underlying concepts. Note that, herein, to provide a comprehensive overview of the topic, we include a few additional recent references that either build upon early-developed models or use less common approaches.

Most of the first prediction models were dedicated to calculating chromatographic parameters, with a special focus on the retention index (RI). Briefly, the RI of an analyte is calculated based on its retention time relative to the retention of a series of reference compounds, typically *n*-alkanes. In other words, the analyte's RI corresponds to an interpolation of its retention time between two bracketing reference compounds. As opposed to the absolute retention times, RIs are system-independent constants that provide a standardized measure for the identification and quantification of compounds making them a quite popular tool in GC. Numerous works proposed RI-based retention time prediction models. Some of the earliest works were published by Beens et al. [28] and Vendeuvre et al. [40]. Both studies proposed similar methodologies for retention time calculations including the use of linear retention indices (LRI) for the prediction of the ^{1}D retention times ($^{1}t_r$) and Kovats indices for the prediction of the ^{2}D retention times ($^{2}t_r$). Good agreements between the experimental and the predicted chromatograms were achieved, even though higher secondary retention time deviations were noticed. Despite their wide applicability and usefulness, RIs still present numerous limitations in that they depend on both the temperature program and the type of stationary phase coating. In this context, Seeley et al. closely examined the repercussions of ignoring these dependencies on the calculation accuracy of the RIs. Although a less accurate model was produced, this approach could still be used as an *a priori* model that provides a rough idea of the space occupation of a specific analyte [41]. Commonly, RIs are calculated through single-column measurements. However, multiple efforts were dedicated to adapting these calculations to the GC × GC framework [42–48]. RI-calculations for GC × GC separations present numerous challenges in comparison to the more straightforward calculations in 1D-GC. This topic was extensively discussed in the literature [30,46,47,49–52]. Besides the need to report two distinct RI values, one for the ^{1}D separation and another for the ^{2}D

separation, adjusted retention time calculations are needed for ^{2}D RI generation. These calculations necessitate ^{2}D column dead time measurements, which are rather challenging to achieve experimentally. Moreover, while a series of *n*-alkanes are often used to calculate RIs, they might not be the most suitable reference compounds for ^{2}D-RI calculations. Rationally, more polar compounds tend to be more retained by the ^{2}D column compared to the *n*-alkanes. Therefore, the compound of interest will most likely not be bracketed by the alkanes making it difficult to report an RI value. This limitation was acknowledged by multiple works that suggested the use of more polar mixtures such as fatty acid methyl esters (FAMEs) [30,43], ketones [42], or the Phillips mix compounds [53]. The starting point for ^{2}D-RI calculations consists in generating the so-called isovolatility curves for the reference standards. Several approaches were used to generate these curves using serial or continuous injections of reference compounds [28,30,42–44,54]. Most of these methods require collecting a significant amount of retention data and could potentially lead to a decrease in the ^{1}D peak resolution due to the use of long modulation periods [31]. Although early developed models heavily relied on the use of RIs, multiple studies showed that other numerical approaches involving retention factor (*k*) [32] and flow calculations [55,56] could be considered viable options for retention time predictions. Regardless of the used approach, greater $^{2}t_{r}$ modeling errors were overall registered. For instance, relative $^{2}t_{r}$ errors of 10 % [40,41], 5 % [28], and 2 % [32] were reported.

Most of the aforementioned approaches limit their investigation to small sets of compounds usually belonging to the same chemical families, such as hydrocarbons [28,40], alkanes, and pyridines [32]. Furthermore, they rely on the use of system-dependent single-column GC data. In other words, they require prior measurements of every analyte's retention data on a specific stationary phase using a specific temperature program. This leads to a subsequent amount of measurements before establishing the actual prediction model. To circumvent this issue, the solvation parameter model was introduced for retention data predictions. This model aims at describing the intermolecular interactions between the solute (analyte) and the stationary phase through the use of a set of solvation descriptors. Each descriptor corresponds to a physical property, i.e. solute size, solute polarizability, hydrogen bond acidity of the solute, hydrogen bond basicity of the solute, and the excess polarizability of the solute [57]. Initially, the solvation parameter model is intended for the prediction of retention factors. Nevertheless, it can be mathematically transformed to calculate RIs instead and subsequently to predict retention times on both dimensions. In this context, standard errors of 1 % and 5 % were reported for $^{1}t_{r}$ and $^{2}t_{r}$ predictions, respectively [58]. Computer-based models involving molecular simulations [59] and/or correlations of the chemical structure to physicochemical properties and biological activities such as quantitative structure-retention relationship models (QSRR) were also deployed for retention time predictions in GC × GC [60,61]. These models correlate the chemical structure of compounds to their retention behavior.

Most of these early works sought to investigate the same research topics. One of their focuses was to accurately assess the interactions between the analytes and the column's stationary phase and to optimize the column combinations accordingly. Additionally, these models were often deployed to assist in determining the ideal temperature programming conditions necessary to achieve maximum separation efficiency and good chromatographic peak shape.

2.2. Thermodynamic-based modeling

The GC × GC separation can be regarded as a thermodynamic process because it involves analyte partitioning between two phases: the stationary phase and the mobile phase. This distribution process is ruled by thermodynamic principles. Therefore, since the early days of theoretical modeling in GC × GC, researchers sought to develop models that describe the thermodynamics of the separation. For instance, Zhu et al. harnessed thermodynamic properties to predict RIs across different column temperature conditions [30]. Jaramillo et al. introduced a new model that enables the thermodynamic modeling of the isovolatility curves of a set of alkanes [31]. Lu et al. used thermodynamics to predict the retention times of multiple pyridines [32]. Given the prediction accuracy yielded by thermodynamic-based models, these models have become a significant tool in retention time prediction in GC × GC. An overview of the available literature reveals a particular emphasis on relating the GC equilibrium constant (*K*) to the well-known thermodynamic indices, enthalpy (ΔH), entropy (ΔS), and molar heat capacity (ΔC_p) [62]. This thermodynamic treatment of *K* enables a better understanding of the solute-stationary phase intermolecular interactions.

2.2.1. Thermodynamic expressions of the GC equilibrium constant (K)

The GC equilibrium constant (*K*) also known as the chromatographic partition coefficient is at the heart of thermodynamic-based modeling. Several methods for calculating *K* were reported in the literature. One of the first thermodynamic expressions of *K* consisted in the use of the two-parameter *van't Hoff* model, in which the equilibrium constant for a specific analyte *i*, K_i is described using Eq. (1), where *R* is the molar gas constant.

$$\ln(K_i) = -\frac{\Delta H_i}{RT} + \frac{\Delta S_i}{R} \tag{1}$$

Even though it was convenient to express *K* as a means of a two-parameter model, this simplified expression retains one substantial shortcoming. It disregards the temperature-dependency of the thermodynamic indices since the model treats them as constants. As a result of this approximation, the two-parameter model was shown to lead to poor model performance regarding retention time prediction in GC [63–65] and could introduce systematic errors, as reported by Karolat and Harynuk [66].

As an extension to the two-parameter model, Clarke and Glew introduced a six-parameter model (Eq. (2)) that acknowledges the effect of temperature on the thermodynamic indices [67].

$$\ln(K_i) = A + \frac{B}{T} + C\ ln(T) + DT + ET^{2} + FT^{3} \tag{2}$$

The *A*, *B*, *C*, *D*, *E*, and *F* terms are constants that are temperature-independent but analyte and stationary phase-dependent. The contribution of the last three terms was deemed negligible and the expression was reduced to a three-parameter model (Eq. (3)) [67], which has since become the most popular thermodynamic expression of *K*.

$$\ln(K_i) = A + \frac{B}{T} + C\ ln(T) \tag{3}$$

The *A*, *B*, and *C* terms are related to the thermodynamic indices ΔH, ΔS, and ΔC_p evaluated at a reference temperature T_{ref} [68]. In this respect, Blumberg introduced a new approach termed *K-centric modeling* that aims to provide more refined calculations for the thermodynamic indices. This approach no longer estimates these indices at a reference temperature T_{ref} but rather considers a predetermined reference distribution coefficient (K_{ref}) for the calculations. This model was shown to be particularly interesting in cases where the analytes exhibit either too small or too large retention factors at the selected T_{ref}. Furthermore, it offers valuable insights for cases where a compound is evenly distributed between the mobile and the stationary phase ($k = 1$) [62]. This new calculation procedure was exploited by Stevenson et al. to generate retention maps that describe analyte distribution in the GC × GC separation space [69].

In this review, the *A*, *B*, and *C* terms (Eq. (3)) will be referred to as thermodynamic parameters. They are commonly determined through single-column isothermal retention time measurements [70,66]. Since *K* is related to the chromatographic retention factor through the $K = \beta k$ expression (β is the phase ratio), isothermal measurements of every analyte's retention factor (k_i) on a specific stationary phase enable the calculations of K_i. The *A*, *B*, and *C* terms are then calculated through

curve-fitting of the K_i vs. T data points. Despite its proven efficiency and accuracy, this calculation method is often considered time-consuming. To address this issue, multiple efforts were dedicated to reducing the operator time by no longer collecting the data using isothermal runs but by using a series of temperature-programmed runs [71–74]. One of the first works to adopt this approach was the computer-based model introduced by Dorman et al. [33]. McGinitie et al. also used the same methodology for the estimation of the thermodynamic parameters for the three-parameter thermodynamic model (Eq. (3)). In this context, the total data collection time for a set of ten compounds was reduced from 41.6 h to 2.0 h [75]. Furthermore, they investigated the impact of manual versus automatic injection conditions on the accuracy of the measurements [76]. In addition, they developed a calibration method to account for variations in the column geometry, which enhanced the calculation of the thermodynamic parameters. With this method, thermodynamic data obtained for a given stationary phase can be used for predictions across columns with the same stationary phase but different geometries [77]. Karolat and Harynuk offered a new take on the calculation of the thermodynamic parameters through their additive thermodynamic model. This model breaks down the molecule into its fundamental building blocks and estimates the thermodynamic contribution of every block. These blocks can then be rearranged to form other molecules for which the thermodynamic parameters are calculated by simply adding the contributions of the individual blocks. This method proved to be efficient for the retention time predictions of a series of alcohols and ketones [66].

2.2.2. *Performance of thermodynamic-based models*

Often during thermodynamic modeling of GC separations, the analytes' movements down the GC column are monitored through an iterative procedure, also known as the modified version of the time summation model [17,31,68,78–81]. During this process, the chromatographic separation is divided into small isothermal time intervals, typically as short as 0.1 s [33,68,80]. While the overall strategy remains the same, each thermodynamic model has its unique specificities, intending to reduce the errors associated with the calculations. Thermodynamic-based models proved to be efficient in accurately predicting retention times and their performance was assessed by using a wide variety of compounds (Fig. 2). Although the model performance varied from one compound to the other most likely due to the use of the simplified thermodynamic model (Eq. (1)), Dorman et al. reported a good agreement between the calculated and measured retention times of the Grob mix compounds with a variance of less than 1 % between the theory and the experiment [33]. McGinitie et al. worked with a single standard mixture of alkanes, alcohols, and ketones for which average errors of 0.64 % for ^1t_r and 2.22 % for ^2t_r predictions were obtained [17].

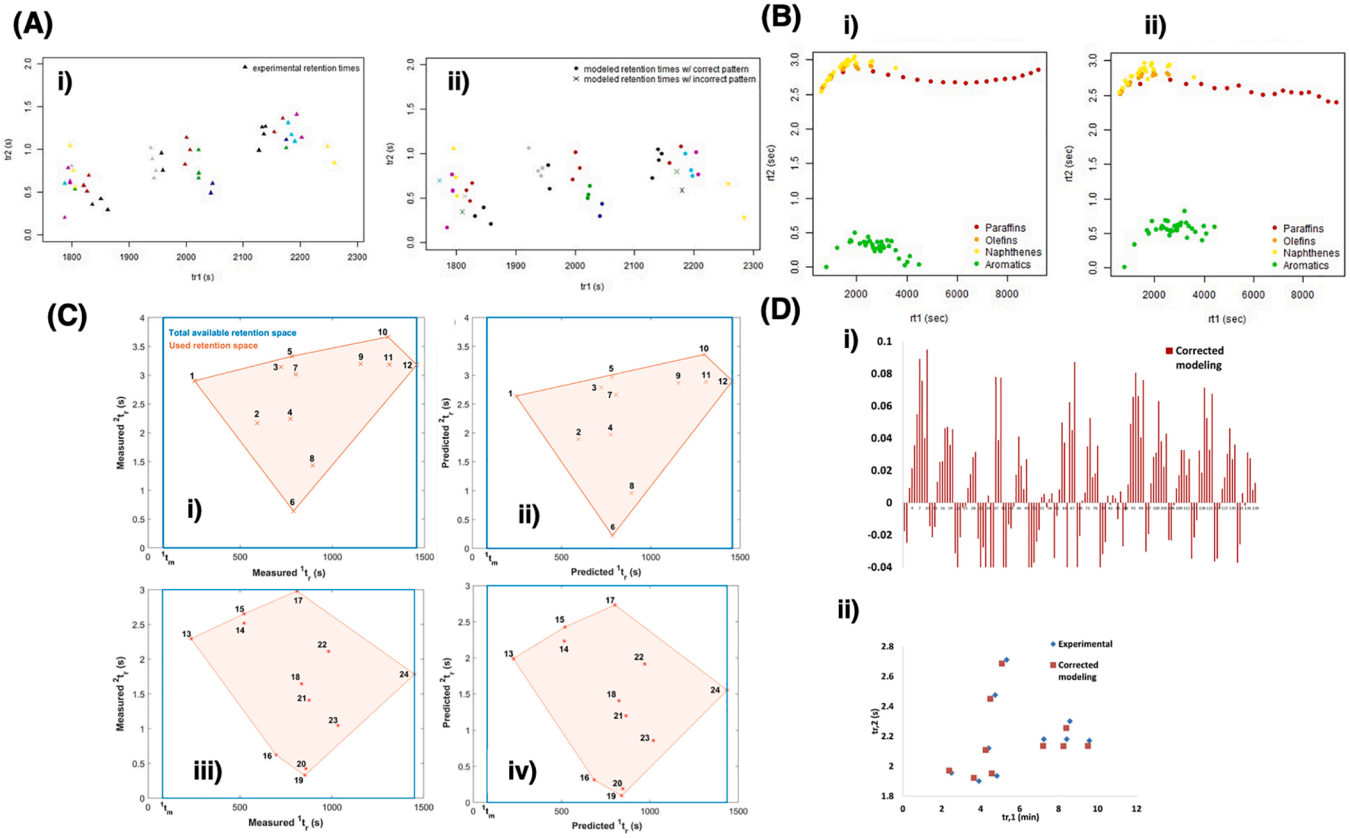

Fig. 2. Examples of GC × GC chromatograms predicted using thermodynamic-based modeling. A) (i) Experimental and (ii) predicted dioxin congeners separations. Reproduced from C. Stultz, R. Jaramillo, P. Teehan, F. Dorman, Comprehensive two-dimensional gas chromatography thermodynamic modeling and selectivity evaluation for the separation of polychlorinated dibenzo-p-dioxins and dibenzofurans in fish tissue matrix, J. Chromatogr. A 1626 (2020). https://doi.org/10.1016/J.CHROMA.2020.461311. B) (i) Experimental and (ii) modeled GC × GC chromatograms of hydrocarbons. Reproduced from R. Jaramillo, F.L. Dorman, Retention time prediction of hydrocarbons in cryogenically modulated comprehensive two-dimensional gas chromatography: A method development and translation application, J. Chromatogr. A (2019). C) Comparison between the experimental and predicted separation space occupation of two sets of standard mixtures: the Grob mix compounds (i and ii) and the fragrance mix compounds (iii and iv). Adapted with permission from M. Gaida, F.A. Franchina, P.-H. Stefanuto, J.-F. Focant, Top-Down Approach to Retention Time Prediction in Comprehensive Two-Dimensional Gas Chromatography–Mass Spectrometry, Anal. Chem. 94 (2022) 17,081–17,089. https://doi.org/10.1021/acs.analchem.2c03107. Copyright 2023 American Chemical Society. D) (i) Second-dimension separation modeling errors for multiple separations of Grob mix compounds. (ii) An experimental and corrected modeled GC × GC separation of the Grob mix analytes. Reproduced from R. Jaramillo, F.L. Dorman, Retention time prediction in thermally modulated comprehensive two-dimensional gas chromatography: Correcting second dimension retention time modeling error, J. Chromatogr. A 1581–1582 (2018) 116–124. https://doi.org/10.1016/j.chroma.2018.10.054.

Silva et al. on the other hand reported average retention time errors of 3.6 s for ^{1}D and 0.2 s for ^{2}D for endogenous steroids [82]. Regardless of the chemical composition of the used samples, most of the published works reported higher modeling errors for the $^{2}t_{r}$ predictions in comparison with those of the $^{1}t_{r}$. Also, higher $^{2}t_{r}$ modeling errors are often reported for cryogenically-modulated systems in comparison to flow-modulated ones. For example, a relative deviation of 7 % for $^{2}t_{r}$ predictions for a series of *n*-alkanes and polyaromatic hydrocarbons (PAH) was reported for a cryogenically-modulated system [70], whereas only a mean error of 2.2 % was reported for the $^{2}t_{r}$ prediction of a series of *n*-alkanes in a flow-modulated run [83]. The underlying reasons behind this trend will be further discussed in the following section. Numerous recent publications focused on reducing the discrepancy between the experimental and predicted $^{2}t_{r}$ either by empirical corrections of the modeling errors or by implementations of newer modeling strategies. By acknowledging the impact of the carrier gas velocity and the elution temperature on the $^{2}t_{r}$ modeling error, Jaramillo et al. introduced an empirical correction model that reduced the average error of a mixture of alkanes from 0.197 to 0.017 s [80]. Moreover, Burel et al. introduced a new expression of the equilibrium constant that endorses the effect of temperature and pressure and thus managed to reduce the ^{2}D modeling error from 7.3 to 2.2 % [83]. Newer modeling strategies featured the use of thermodynamically-modeled isovolatility curves for retention time predictions. This approach gave satisfactory results and yielded average modeling errors of 11 s and 0.09 s on both dimensions, respectively [31]. Recently, we introduced a new modeling approach that breaks down the GC × GC run into two separate 1D-GC and treats the modulator as a second injection device. In this context, separate 1D-GC retention time predictions are conducted and then combined to account for the GC × GC separation space. This system-independent approach gave favorable results in terms of separation space description and occupancy [18,68].

2.2.3. Advantages, challenges, and limitations

Thermodynamic-based models can be distinguished based on their inherent ability to account for the experimental variations that may occur during chromatographic separation. Specifically, the thermodynamic parameters, i.e. *A*, *B*, and *C*, are calculated based on experimental measurements. Therefore, they accurately describe the effects of temperature on the analyte's retention factor. Furthermore, from an experimental point of view, thermodynamic-based models are often considered a slightly superior approach to RI-based models. Even though ^{1}D-RI measurements are conducted straightforwardly, ^{2}D-RI estimations were proven to be instrumentally challenging and time-consuming [54].

While modeling the GC × GC ^{1}D separation is moderately simple since the separation primarily depends on the compounds' volatility, i.e. boiling point, the same cannot be inferred for the ^{2}D modeling. The fast pace of the ^{2}D separation, the presence of the modulator interface, the limited user control over the experimental parameters especially in cryogenically-modulated systems, and the intricacy between all the involved factors all combined significantly increase the complexity of the ^{2}D separation modeling.

When describing ^{2}D retention, researchers often resort to the use of multiple assumptions to overcome the complexity of the system. The most commonly used assumption consists in perceiving the ^{2}D separation as isothermal and occurring at the ^{1}D elution temperature. Furthermore, the modulator interface has been subject to numerous simplifying assumptions, especially in thermally-modulated systems. The modeling of these systems is significantly challenging compared to flow-modulated systems. Specifically, because it is difficult to assess the temperature variations that occur within the modulator as a result of the constant cooling and heating processes. This dilemma, in part, explains why fewer modeling works are conducted using thermal modulation [18,33,70,79–82]. The thermal modulation of GC × GC systems is often viewed as a dynamic process. When passing through the modulator, the analyte goes through very fast sorption and desorption phenomena due to the periodic cold and hot jets. So far, none of the conducted research was able to assess how the analyte's passage through the modulator affects its retention behavior instead some assumptions were made. Furthermore, it is still difficult to estimate the moment in time when each analyte is trapped by the cold jet and to calculate the equilibration time needed by the analyte to reach the preset temperature after going through the cold jets. In this regard, Jaramillo et al. modeled the analyte retention within the modulator using the modulator's offset temperature and estimated the time the analytes are trapped by the cold jet to the sum of the retention time and the modulation period [80]. Recently, we offered a new take on GC × GC modeling by omitting altogether the modulator from the modeling process. Instead, it was simply considered as a second injection device, thus offering a system-independent approach [18]. Another relevant reason why thermally modulated systems are less commonly used in theoretical modeling compared to flow-modulated systems is due in part to the limited user control over pressure. With thermal modulators, there exists only one pressure control point, at the inlet of the ^{1}D column, while flow modulators allow pressure control at the head of both dimensions' columns. Therefore, most of the experimental parameters that govern the separation, such as the carrier gas velocity and the pressure profile must be calculated theoretically. This can result in increased modeling errors. Note that, along with the preferential use of flow modulators, the majority of the published works model systems with flame ionization detectors (FID) instead of MS detectors to ease the calculation process since the outlet pressure is easily assessed in systems using FIDs. However, in reality, most commercially and industrially available GC × GC systems use thermal modulators and MS detectors. Hence, it is important to allot more time to understand the intricacies of these systems and to develop models that could be readily used in practical settings. Overall, despite the necessary use of approximations, good modeling performances were achieved by most of the aforementioned methods. Nevertheless, more work and further refinements are still needed to avoid oversimplifying the GC × GC system, particularly at the modulator stage.

Despite the wide variety of available prediction models, GC × GC still lacks a fully automated system-independent method. Moreover, most of the available software packages only allow for theoretical simulations of 1D-GC. Additionally, it is important to create a comprehensive database that encompasses all the retention data that was collected across different instruments, configurations, and stationary phases. In this regard, we acknowledge the recent publication of a retention database by Brehmer et al. [84]. This database can tremendously help in the data collection step by significantly reducing the amount of time dedicated to the initial measurements. It not only streamlines method development but also allocates more time for the challenging data processing step, a topic we will cover in the following section. Furthermore, this database can also facilitate the sharing of knowledge among researchers, ultimately advancing and reviving the field of theoretical modeling in GC × GC.

3. Machine learning algorithms for GC × GC–MS data processing

Undoubtedly, the high dimensionality of the data generated by GC × GC–MS often calls for a robust and effective data processing step to extract meaningful information [85]. Traditional chemometric tools may, at times, not be enough to fully leverage the potential of the technique as they often rely on the quality and format of the collected data. Furthermore, some of these tools, may not be designed to operate directly on complex three-dimensional data structures such as the ones generated by GC × GC–MS. As a result, there has been a growing interest in using advanced ML algorithms for GC × GC–MS data processing. Briefly, ML is a sub-field of artificial intelligence (AI) that involves training algorithms on large datasets, allowing them to recognize patterns and relationships within the data, and make decisions about other new sets based on the learned information. One of the key features of ML

algorithms is their ability to iteratively adjust their parameters over time to improve the overall performance and accuracy of the model [86]. ML algorithms can be used in both supervised and unsupervised fashion. Supervised learning considers sample class affiliations during the training of the model, and it mainly focuses on building regression and/or classification models. For instance, random forest (RF) is a supervised ML algorithm that simultaneously trains multiple decision trees to enhance the accuracy of the predictions. On the other hand, unsupervised learning does not consider sample class memberships and helps depict clustering patterns among the data and/or reduces its dimensionality (Fig. 3). Principal component analysis (PCA) may be one of the most popular unsupervised ML algorithms. It is commonly used as a dimensionality reduction technique. In other words, it helps transform high-dimensional datasets into lower-dimensional datasets while making sure to express as much of the original variation as possible. It has come to our attention that frequently PCA is discussed separately from other ML algorithms. One possible explanation for this differentiation could be that PCA only serves a specific purpose in data analysis. While other ML algorithms are versatile and can serve a wider range of tasks, such as classification, regression, pattern recognition, and prediction, PCA is primarily used for exploratory data analysis. Its initial function is to screen high-dimensional data by reducing its dimensionality.

Lately, the processing of GC × GC–MS data has been the subject of several informative review papers. Stefanuto et al. provided a thorough and detailed overview of the entire data processing scheme, from preprocessing to validation [36], while Pollo et al. discussed the use of fundamental chemometric tools in the analysis of -omics-related datasets [35]. On the other hand, Stilo et al. provided a comprehensive overview of the various approaches employed in chromatographic fingerprinting [37], while Jimenez-Carvelo and Cuadros-Rodriguez critically discussed the use of untargeted data analysis approaches in foodomics [87]. More recently, Trinklein et al. discussed the latest developments in chemometric tools and their application in nontargeted analysis [38]. While some of these reviews have mentioned the incorporation of ML in GC × GC–MS data processing, the depth, and extent of these discussions have so often been limited. ML algorithms are frequently introduced as "the next big thing" in data processing. However, it is important to acknowledge that they have already received wide acceptance in the GC × GC field and are yielding highly promising results. Their feasibility has been demonstrated in diverse domains, and they have been successful in analyzing GC × GC–MS data collected from different matrices.

3.1. Algorithm selection

Similar to other analytical techniques, the processing of GC × GC–MS data is a tedious procedure that involves numerous steps. Although the details of some of these steps are beyond the scope of this review, it is however worth stressing their importance for accurate and robust data analysis. Before the use of ML algorithms, GC × GC data undergoe a preprocessing step [36]. This initial phase is of the utmost importance since it significantly affects the interpretation of the outcome of the data processing step. It ensures the removal of outliers and helps standardize the data to make it ready for use in model building [88]. Briefly, pre-processing includes multiple steps. Before engaging in data processing, one needs to make sure that the data is suitable for analysis including checking its format and identifying any missing entries in the datasets. If so, data imputation techniques can be performed [89]. Once data verification is completed, normalization and scaling are conducted through a variety of techniques. Transformation techniques may also be applied to correct the noise in the data [88]. Once the data is in good shape, an exploratory data processing step is often conducted using unsupervised multivariate statistical tools, such as PCA and hierarchical cluster analysis (HCA) to reveal any underlying patterns in the data or to highlight clustering trends, if any, among the samples based on batch-related effects [36,90,91]. PCA aims to provide a visual

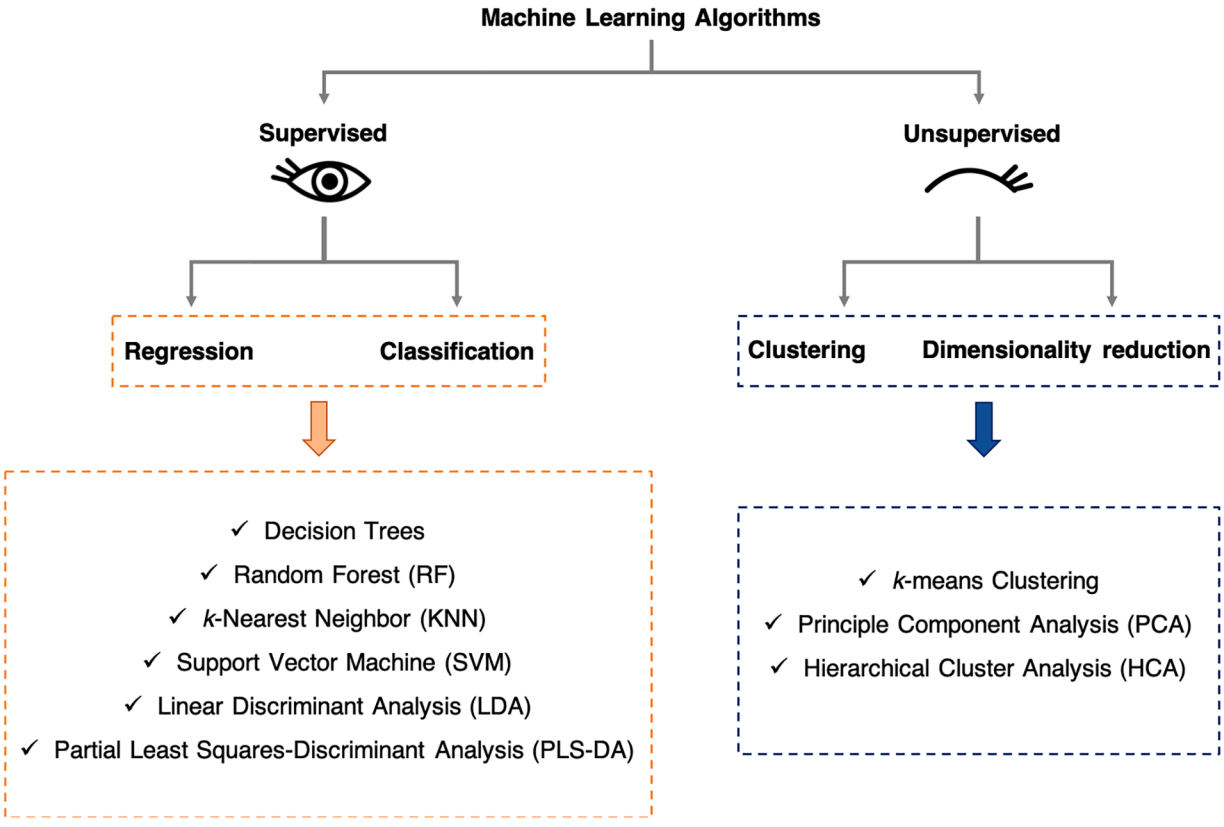

Fig. 3. Overview of the most commonly used machine learning algorithms in GC × GC data processing.

interpretation of the data using as few components as possible. These components are called principal components (PC) and are ordered based on the amount of variation they explain in the original dataset. PCA is used to depict trends and relationships among the measured variables. HCA, on the other hand, is a clustering technique that groups the measured variables into clusters based on their distance or similarity. After completion of the data preparation and gaining a global understanding of the datasets, ML algorithms can be employed to cater to the specific requirements of the scientific problem being investigated. In this review, we do not aim to provide an exhaustive analysis of how these algorithms operate, as numerous previous reviews have already extensively covered the topic [92–95]. We also do not aim to list all the available ML algorithms. Instead, our focus is on three of the commonly used ML algorithms in the field of GC × GC: random forest (RF), support vector machine (SVM), and partial least squares-discriminant analysis (PLS-DA) and their practical applications for the analysis of GC × GC–MS data. Nevertheless, to ensure that the reader can grasp the concepts behind these algorithms, we provide a brief fundamental understanding of the functioning of the discussed ML techniques. Moreover, it is worth noting the emergence of deep learning algorithms as a promising tool for large dataset analysis [96–98]; however, this topic falls outside the scope of this review.

The selection of the appropriate ML algorithm is of paramount importance to achieve the desired performance in data analysis. With a multitude of ML algorithms available, each serving different purposes and leaning on different statistical and mathematical concepts, it is important to choose the most suitable one for the specific analytical task. Generally, cross-validation (CV) is used to assess an ML model's performance. In other words, it serves as an indicator of how well the model is likely to perform when applied to a new set of data. It requires partitioning the data into two different sets: the first set is called a training set and is used to build the model. The second set is referred to as the validation set and is used to evaluate the accuracy of the model.

RF and SVM have emerged as the most frequently used ML algorithms in GC × GC–MS data analysis [19,99–104], closely followed by PLS-DA [102,105–107]. Therefore, this section will primarily focus on exploring these algorithms (Fig. 4). However, it is important to acknowledge the potential applicability of other ML-based algorithms, including linear discriminant analysis (LDA) [20,105], Monte-Carlo neural network (MCNN) [108], cluster resolution (CR) [109], and even customized ML software [110].

3.1.1. Random forest

RF is one of the most popular ML algorithms widely used in classification and regression tasks. Being an ensemble algorithm, RF operates by building a forest of decision trees, with each tree operating independently on a randomly selected subset of the original dataset. At the outset, the original dataset is divided into numerous subsets, referred to as bootstrapped datasets. Generally, these datasets are generated by random selection of two-thirds of the samples in the initial dataset, with a possibility for a sample to be represented more than once. The remaining one-third of the samples are assigned to a subset known as the out-of-bag (OOB) dataset. This set acts as an external validation dataset that enables the calculation of the misclassification error of the model trained using the bootstrapped dataset. This error is called the OOB error and is an important metric for evaluating the model's performance [111].

In GC × GC–MS data analysis, RF is commonly employed in a supervised fashion for classification purposes [19,100,101,105]. In this context, each decision tree in the forest generates a class prediction and the overall prediction of the model is determined by the majority vote across all trees. The accuracy of an RF model largely depends on its hyperparameters, therefore careful tuning of these parameters is necessary for achieving optimal results [112]. Among these parameters, we can cite the number of decision trees, the number of predictor variables to sample, and the minimum leaf size. Selecting the optimal number of decision trees is a rather challenging task, in the sense that there is no rule of thumb regarding what the optimal value should be. As a matter of fact, not only it depends on the quality, complexity, and size of the dataset, the optimal number of decision trees also depends on the desired level of accuracy and the purpose of the study. Nevertheless, as a general guideline, for an accurate and robust RF model, it is recommended to use a sufficiently large number of trees, but not too large to avoid overfitting [112]. Moreover, it is wise to explore a range of values at the beginning of the analysis and perform cross-validation (CV) or monitor the OOB error to evaluate the model's performance under different settings. Then, the user could settle for the value that entails the best answer to the research question at hand. For instance, Strozier et al. employed an RF model with 5000 trees to classify chemical threat agents analyzed by GC × GC-ToF-MS according to their origin [113]. Andersen et al. also used 5000 trees to identify putative blood-based multiple sclerosis biomarkers [103]. In contrast, Beccaria et al. in their study to distinguish between patients with a specific pulmonary infection and those with other pathologies, opted for 1000 decision trees [101]. These studies serve as examples of how the number of trees can be adapted across different analyses. Unlike the decision tree number, most of the other important hyperparameters often have numerical recommendations that are typically set by default in software packages, such as R [114] or MATLAB [115]. As an example, the number of predictor variables to sample (n), is set to the $\sqrt{N}$ in classification problems, where N is the total number of variables in the original dataset, and $N/3$ in regression problems. This is based on the original work conducted by Breiman, where it was shown that using relatively small n improves the performance of the RF model in that it reduces the correlation between decision trees [111]. Another example is the minimum leaf size. It is set to 1 for classification tasks and 5 for regression tasks. This means that each terminal leaf node will have at least one sample assigned to it for classification and 5 for regression. These default values can be used as a starting point for model building. Nevertheless, for better model performance, these default values can be further optimized by using various optimization techniques, including but not limited to Bayesian Optimization (BO), grid search, random search, and genetic algorithms [116].

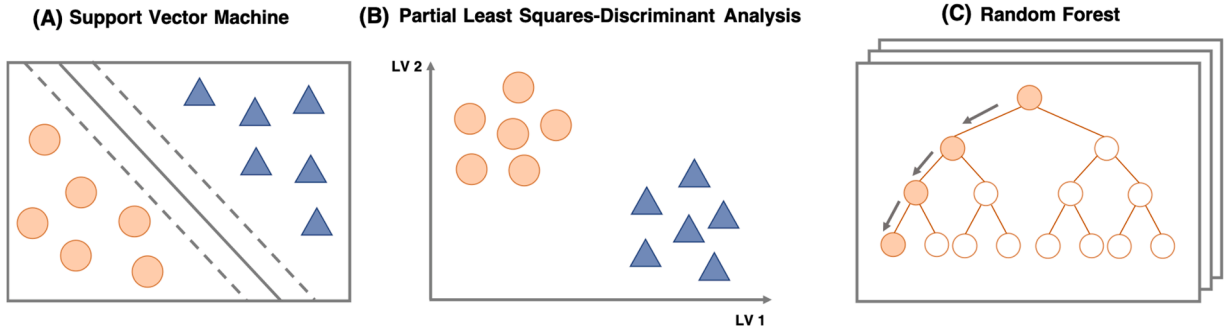

Fig. 4. Graphical representation of the three most commonly used machine learning algorithms in GC × GC data processing.

3.1.2. Support vector machine

SVM is a supervised ML algorithm that can solve both classification and regression problems. Lately, it has been repeatedly used for classification purposes of data acquired through GC × GC–MS analyses [20,102,105,117,118] The central idea in SVM consists in segregating input data into different classes using a so-called decision boundary. This boundary could be a line, a hyperplane, a curve, or a surface depending on the relationship between the input and output data. The data points that are the closest to the decision boundary are referred to as support vectors and the distances between the support vectors and the decision boundary are called margins. Intuitively, the further the data points are from the decision boundary, the better the separation is. Therefore, SVM seeks to maximize the margins to find the most optimal decision boundary that makes the separation between classes as wide as possible. The algorithm is referred to as linear SVM if the decision boundary is linear, i.e. straight line or a hyperplane. If the decision boundary is non-linear, i.e. a curved line or a complex surface, the algorithm is called non-linear SVM. In this context, kernel functions are used to transform the data from its lower-dimensional space to a higher-dimensional space where a linear decision boundary can be applied. The kernel functions used in SVM include linear, polynomial, radius, and sigmoid kernels. Their choice depends on the nature of the data and the problem to solve [94,119]. Therefore, similarly to RF models, for greater model accuracy, SVM hyperparameters need to be tuned [119]. Often, the kernel type is the first parameter to be adjusted as it is vital for the classification of the data. Depending on the chosen kernel type, other hyperparameters may also need to be optimized. As an example, for the analysis of GC × GC-QMS measurements of crude oils, Guilherme et al. opted for a radial basis function (RBF) kernel and used the grid search optimization technique to determine the most optimal C and γ values [105]. These hyperparameters are specific to the RBF kernel. The C parameter serves as a trade-off between model accuracy and margin maximization. Small values of C result in wider margins, which reduces overfitting but may increase the likelihood of misclassification errors. Conversely, larger values of C result in narrower margins, decreasing the probability of misclassification but potentially increasing overfitting. The γ hyperparameter affects the shape of the decision boundary, where larger values lead to a more flexible and complex boundary and smaller values to a more restricted boundary. Therefore, the authors investigated a search space formed by all possible combinations between $0.001 \leq C \leq 100$, and $10^{-6} \leq \gamma \leq 10$ [105]. Similarly, Rist et al. fine-tuned the C parameter when using SVM with a linear kernel [117]. In addition to hyperparameter tuning, Reichenbach et al. assessed the performance of multiple SVM models by benchmarking different types of kernels. Specifically, they explored the effects of varying the degree of the polynomial (cubic and quadratic), for poly-type kernels, on the accuracy of the model. Additionally, they investigated the use of a Gaussian kernel, which is an RBF kernel, with a moderate γ value, and kernels with small to moderate scales of hyperparameters. By benchmarking these different kernels, the authors aimed to identify the optimal kernel configuration for their wine classification task [20].

3.1.3. Partial least squares-discriminant analysis

PLS-DA is a variation of the PLS regression algorithm that simultaneously operates as a dimensionality reduction technique and a discriminant analysis technique serving both regression and classification tasks [95]. Its goal is to find a linear relationship between the input data (X), i.e. the measured variables, and the output data (Y), i.e. the samples class membership through the use of components labeled as latent variables (LV). These LVs are constructed to capture the maximum covariance between X and Y. While the output variables in classification tasks are categorical, PLS-DA only handles continuous variables. Therefore, the first step in PLS-DA model building is to transform categorical variables into continuous variables by using dummy codes such as −1, 0, and +1. To summarize the essence of the algorithm, PLS-DA calculates weight and loading vectors for each LV. The first LV is selected to maximize the covariance between X and Y, while subsequent components are selected to maximize the covariance between X and Y variables that were not explained by the previous components [95]. Typically, the selected number of LVs in a PLS-DA model is defined through the CV process. Venetian blind CV along with *k*-fold CV are popular methods for PLS-DA hyperparameter tuning [102,105,107]. Briefly, Venetian blind CV divides the dataset into multiple non-overlaying subsets of equal size. Each subset acts exactly once as a validation set, while the other remaining sets are used as training sets. This process is repeated multiple times based on the frequency chosen by the user. Then, the results are averaged to obtain an estimate of the model's accuracy. On the other hand, *k*-fold CV randomly divides the original dataset into *k* subsets, often referred to as folds, of equal size. Typically, one of the folds is held apart as a validation set, while the other remaining folds are used for model training. This process is repeated *k* times, each time with a different fold held out for validation. The model's performance is reported as the average performance across the *k* validation sets. For example, through Venetian blind CV, some authors were able to optimize the PLS-DA model by reducing the number of LVs down to 5 [105] and 3 [107], while others performed 5-fold CV to select the optimal number of LVs [102]. PLS-DA models can be used for prediction purposes, meaning that they can predict the class membership of a new sample based on its measured variables. In this context, the new set of data is projected into the LV space using the weight vectors, and their corresponding Y variables are calculated using the loading vectors. PLS-DA can also be used for feature selection tasks by ranking the X variables based on their influence on the model.

3.2. Performance assessment of ml algorithms in GC × GC–MS data analysis

Given the complexity of GC × GC–MS data, it is common practice to benchmark multiple ML algorithms. This process enables users to assess different algorithms and identify the ones that exhibit superior performance and align best with the specific analytical objectives of the study. Typically, several factors are considered when determining the most suitable technique. These factors include the model's training time, which is particularly important when investigating large datasets, as well as model complexity. Simpler models are generally preferred as they are easier to interpret. Scalability is also a pivotal factor in that it ensures that the algorithm can effectively handle high-dimensional datasets, such as those produced by GC × GC–MS, where the number of variables exceeds the number of samples. Furthermore, resource requirements also need to be considered. However, the most significant parameter to acknowledge is the accuracy of the model. Commonly, model performance is evaluated using CV, and specific performance evaluation metrics can be calculated using the predictions made by the trained model on the validation set (Table. 1) [22,107,117,120,121]. Note that the formula displayed in Table 1 are applicable for binary classification tasks, i.e. classification tasks involving only two classes. However, these calculations can be easily altered to accommodate multi-class classifications by either calculating the performance measures for each class individually or by averaging across all classes.

The performance of an ML model can also be assessed through graphical representations such as the receiver operating characteristic (ROC) curve (Fig. 5). This curve plots the true positive rate, i.e. sensitivity, against the false positive rate, i.e. 1-specificity, using different threshold values. Based on this curve, a scalar value called the area under the ROC curve (AUROC) can be calculated. The AUROC ranges between 0 and 1, with 0.5 indicating a poor model performance (no prediction ability) and 1 representing a perfect classification [122]. ROC curves and AUROC calculations are commonly used when benchmarking different algorithms since they allow for easier and more straightforward comparison between different models [102]. Confusion matrices (CM) also enable a visual interpretation of the performance of an ML model (Fig. 5). The CM allows for a clear understanding of the

Table 1
Performance metrics used to evaluate machine learning models. TP: True positive, TN: True negative, FP: False positive, and FN: False negative.

Metric	Formula	Description
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	Ability of the model to predict the classes of the input data.
Misclassification error	$\frac{FP+FN}{TP+TN+FP+FN}$ *or* $1-accuracy$	Proportion of instances that are incorrectly classified by the model.
Precision	$\frac{TP}{TP+FP}$	Proportion of true positive predictions among all positive predictions made by the model.
Recall	$\frac{TP}{TP+FN}$	Proportion of positive instances that are correctly predicted by the model out of all the actual positive instances in the input dataset.
Specificity	$\frac{TN}{TN+FP}$	Proportions of true negative instances that are correctly classified by the model.
F1-score	$2\times\frac{Precision\times Recall}{Precision+Recall}$	Indicator of the overall accuracy of the model by balancing between precision and recall.

model's performance by providing a detailed visual breakdown of the model's predictions through numeric representations of true positive, true negative, false positive, and false negative predictions. It is often represented as a matrix plot or a heatmap [113,123].

In the spirit of comparing and assessing different ML algorithms, Reichenbach et al. estimated the accuracy of 17 different ML techniques. These techniques, including decision trees, discriminant analysis, SVM, *k*-nearest neighbor (KNN), and ensemble methods, were applied to classify wines based on their varieties, geographic regions, vintages, and wineries. The performance evaluation of all these algorithms was conducted using the leave-one-out CV (LOOCV) [20]. Similar to other CV methods, LOOCV divides the input data into a training set and a validation set. Notably, in LOOCV, the validation consists of only one sample, while the remaining samples form the training set. This procedure is repeated for each sample in the input dataset. Among the 17 compared methods, SVM-based models achieved the highest average accuracies, reaching approximately 90 %. Ensemble techniques followed closely with average accuracies reaching around 87 %. Conversely, decision trees were among the least successful methods as they exhibited relatively lower performance. Similarly, in a study conducted by Guilherme et al., SVM outperformed other algorithms, such as PLS-DA, quadratic discriminant analysis (QDA), LDA, and KNN [105]. When classifying individuals according to their sex and age using metabolite profiles from urine and plasma, SVM along with PLS-DA, and generalized linear model net (glmnet) demonstrated comparable levels of accuracy [117]. Furthermore, when comparing SVM to PLS-DA and RF, RF exhibited superior overall performance [102]. These findings emphasize the absence of one-algorithm-suits-all since the choice of methods highly depends on the requirements and characteristics of the investigated data.

3.3. Feature selection

Feature selection (FS) is among the most appealing aspects of ML algorithms and is extensively used in the context of GC × GC–MS data analysis. FS aims to reduce the initial pool of variables (features) to a subset of the most informative and discriminative features (Fig. 6). Not only it significantly reduces the computational time since only a limited number of features is used during model training, FS also enhances model performance and mitigates overfitting issues. In light of all these advantages, FS is usually carried out before constructing a prediction model that can reliably classify external samples. Particularly in GC × GC–MS data processing, where hundreds to thousands of compounds can be present in a single chromatogram, FS plays a crucial role in eliminating irrelevant, redundant, and noisy data, which significantly increases the quality and the interpretability of the results. Typically, in ML-based GC × GC–MS data processing workflows, a common sequence of steps is often followed. First, a classification model is built using the entire dataset, followed by a validation step to access the accuracy of the model. Subsequently, an FS step is applied to identify the most class-distinguishing features. Following FS, a reduced model is trained solely using the subset of features identified by FS, which is then applied for the classification of new sets of data. FS is performed differently from one ML algorithm to the other [124]. In this section, however, we refrain from diving into the specifics of FS in ML algorithms since it is beyond the scope of this review. Nevertheless, we will discuss some examples illustrating the efficacy of combining GC × GC–MS analysis with FS using ML algorithms and emphasizing its applicability across various fields. For in-depth insights into the various methods of performing FS with ML algorithms, we recommend referring to the following references [125–127].

The majority of the studies primarily focused on using FS for biomarker discovery. In a recent investigation by Cen et al., the effect of COVID-19 vaccines on the exhaled VOC profile of individuals was compared to that of non-vaccinated individuals. Their approach involved two steps. They initially selected a subset of 12 candidate biomarkers using variable importance in projection (VIP) scores calculated within a PLS-DA model. Then, these biomarkers were validated using an RF model [120]. In another COVID-19 study, VIP scores were calculated to distinguish between healthy individuals and those infected by the disease. This study identified putative biomarkers that could

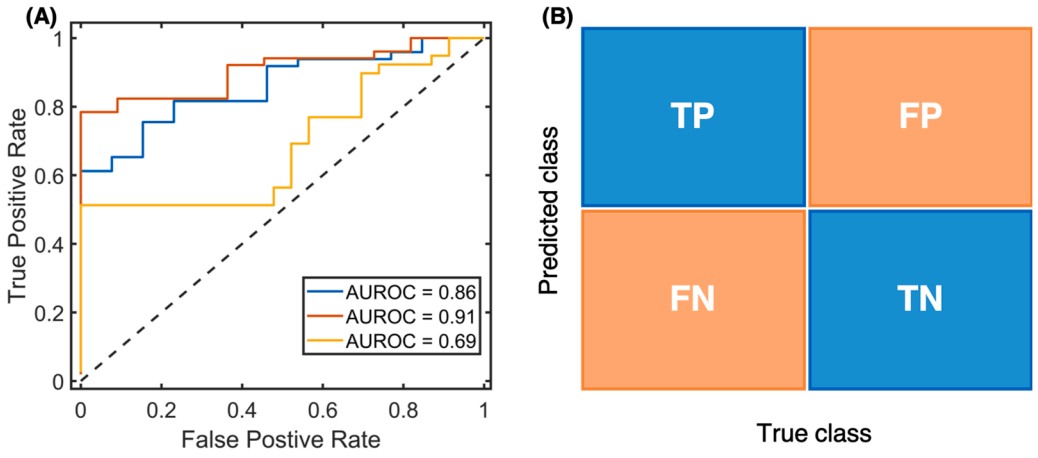

Fig. 5. Graphical representation of the A) Receiver Operating Characteristic (ROC) curve and the Area Under the ROC curve (AUROC) and the B) Confusion matrix (CM). TP: True Positive, TN: True Negative, FP: False Positive, and FN: False Negative.

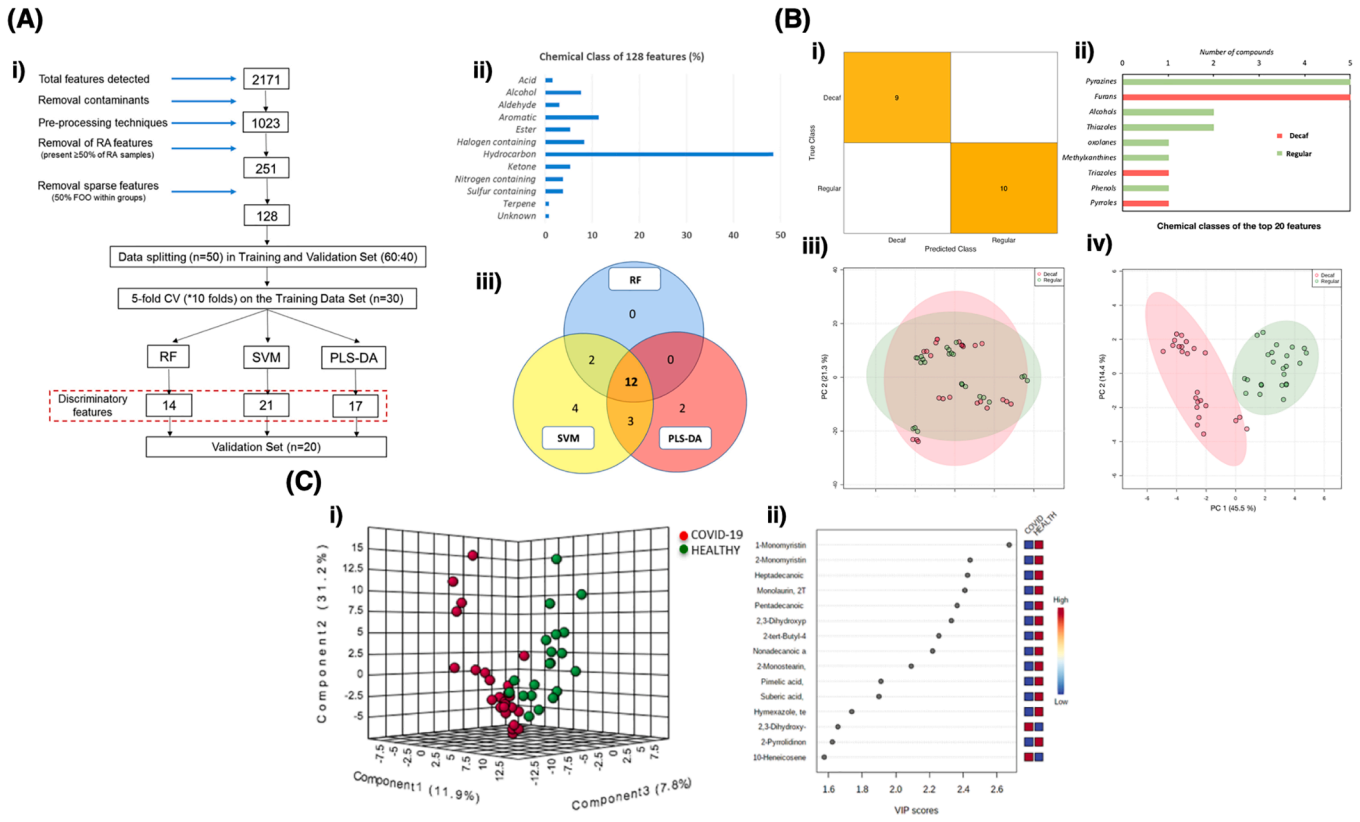

Fig. 6. Examples of identification of class-distinguishing analytes using machine learning (ML) algorithms feature selection techniques. A) Identification of potential Tuberculosis markers using 3 ML algorithms: Random Forest (RF), Support Vector Machine (SVM), and Partial Least Squares-Discriminant Analysis (PLS-DA). i) Feature selection workflow. ii) Chemical classes of the identified class-distinguishing analytes. iii) Venn Diagram representing the features selected from the 3 ML algorithms. Reproduced with permission from M. Beccaria, C. Bobak, B. Maitshotlo, T.R. Mellors, G. Purcaro, F.A. Franchina, C.A. Rees, M. Nasir, A. Black, J.E. Hill. Exhaled Human Breath Analysis in Active Pulmonary Tuberculosis Diagnostics by Comprehensive Gas Chromatography-Mass Spectrometry and Chemometric Techniques, J. Breath Res. 13 (2019) 016,005. Date of publication: 05 November 2018. https://doi.org/10.1088/1752-7163/aae80e. © IOP Publishing. B) Top 20 features identified using RF to distinguish between regular and decaffeinated coffee. i) Confusion matrix. ii) Chemical classes of the Top 20 features. iii) Unsupervised PCA scores plot. iv) Supervised PCA scores plot. Reproduced from Y. Zou, M. Gaida, F.A. Franchina, P.H. Stefanuto, J. Focant. Distinguishing between Decaffeinated and Regular Coffee by HS-SPME-GC × GC-TOFMS, Chemometrics, and Machine Learning, Molecules. 27 (2022) 16. C) identification of features that effectively discriminate between COVID-19 infected patients and healthy patients using a PLS-DA model. i) PCA performed using the reduced dataset. ii) VIP scores plot of the most class-distinguishing analytes in each group. Reproduced from E. Barberis, E. Amede, S. Khoso, L. Castello, P.P. Sainaghi, M. Bellan, P.E. Balbo, G. Patti, D. Brustia, M. Giordano, R. Rolla, A. Chiocchetti, G. Romani, M. Manfredi, R. Vaschetto. Metabolomics Diagnosis of Covid-19 from Exhaled Breath Condensate, Metabolites. 11 (2021).

allow for a noninvasive diagnosis and monitoring of the disease. These biomarkers were further validated using a genetic algorithm-based ML model [106]. Additionally, in the same context of the COVID-19 pandemic, Favela et al. employed FS to investigate the chemical composition of facemasks [110]. Furthermore, Beccaria et al. identified potential biomarkers for tuberculosis diagnosis in exhaled breath [102], while Andersen et al. identified blood-based biomarkers for multiple sclerosis using FS in conjunction with RF [103].

FS was also applied to the study of various food matrices. Zou et al. combined FS with RF to distinguish between regular and decaffeinated coffee. The newly discovered markers were then used to construct a prediction model capable of predicting the type of new coffee samples [19]. Lima et al. used a PLS-DA model to assess the authenticity of fish oil supplements [107], while Paiva et al. used the same ML algorithm to classify beers and to identify 31 analytes that are correlated with consumer preferences [128]. Li et al. used multiple ML models to assess the authenticity of orange juice and identified key compounds that could differentiate between biological origins, geographical origins, harvesting years, and processing methods [121]. FS was also employed in other studies involving the differentiation of virgin and recycled polyethylene terephthalate [123], the study of the algal metabolome [109], and the characterization of jet fuel properties [108].

3.4. ML algorithms vs. traditional statistical methods

Comparing ML algorithms to traditional statistics is not as easy as one might think. Some can even argue that it is like comparing apples to pears since the two fields operate in a notably different manner. Instead of thinking about using them interchangeably or favoring one over the other, it may be more beneficial to consider their use based on the specific objectives and requirements of the analysis. For instance, if the goal is to build a classification model and then use it to make predictions, ML algorithms are more suitable for the task. On the other hand, if the objective is to explore relationships between variables and to test hypothesis, a statistical model may then be more appropriate. Both approaches have their strengths and limitations, hence the decision should be based on the problem at hand.

One of the benefits of ML algorithms is their high flexibility to handle various and complex types of data. As opposed to traditional statistical methods, that often require assumptions like normal distribution and equal variances across classes, ML algorithms do not rely on such a priori assumptions. This feature becomes particularly handy when dealing with datasets that have unbalanced class sizes and/or inherent biological variations within the same class.

ML algorithms also excel in handling high-dimensional datasets, where the number of variables exceeds the number of samples. This

scenario is often, if not always, encountered in GC × GC studies. Oppositely, traditional statistical methods face challenges in analyzing such datasets since they are primarily designed for low-dimensional data, where the number of samples is larger than the number of variables [129]. The high-dimensionality of the data poses numerous challenges when attempting to plot a specific response against all variables in a linear fashion. This difficulty arises not only due to the high number of variables but also because the data itself may exhibit multiple responses, making it more complex to visualize and interpret the relationships between variables and responses. This is where ML algorithms come in handy. By leveraging advanced techniques such as dimensionality reduction and feature selection, they tremendously and effectively decrease the complexity of the data.

ML algorithms are known for their high adaptability as they are capable of learning and adjusting their predictions based on the underlying data distribution. They help automate the process of model-building and decision-making, resulting in considerable time and effort savings in comparison to manual data curation. Nevertheless, it is important to acknowledge that despite their advantages, ML algorithms are not immune to certain challenges and limitations. Some algorithms do not enable an easy and straightforward interpretation of the results. It can be, at times, difficult to understand the rationale behind certain predictions and outcomes. Furthermore, ML algorithms are not exempt from overfitting, which occurs when the model becomes overly tailored to the training set and hence fails to perform for external datasets. An accurate and efficient use of ML algorithms often require solid expertise in algorithm selection, hyperparameter tuning, and data interpretation. Moreover, in light of the exponential increase in the use of ML-based methods in particular and AI in general, ethical considerations have nowadays become increasingly significant. The apprehension about AI surpassing human intelligence is now more relevant than ever.

4. Final remarks

It is undeniable that GC × GC has now entered its golden age and has become an indispensable tool in the analysis of volatile and semi-volatile molecules across different fields. Nevertheless, in order to unlock the full potential of the technique, a thorough method development step needs to be combined with a robust and accurate data processing workflow. In this review, through an extensive examination of the literature, we provided a comprehensive overview of the current state-of-the-art in theoretical modeling and ML-based data processing workflows in GC × GC.

As part of method development, theoretical modeling in GC × GC plays an important role in optimizing separation conditions, thereby increasing efficiency and selectivity. Extracting meaningful information from GC × GC data is challenging due to the high-dimensionality of the data. Therefore, robust data processing workflows are of the utmost importance to effectively and efficiently analyze the data. To this end, the integration of ML algorithms into the data processing workflow can be highly beneficial. Algorithms such as RF, SVM, and PLS-DA have demonstrated great promise in GC × GC data analysis, particularly in biomarker discovery through feature selection.

Despite the remarkable progress in both method development and data processing in GC × GC, several challenges remain. There is still a need for the development of more accurate, automated, and system-independent theoretical models. Furthermore, collaborations between chemists, statisticians, and computer scientists can be highly beneficial to fully explore the potential of ML algorithms in order to use them in an informed manner.

CRediT authorship contribution statement

Meriem Gaida: Conceptualization, Methodology, Writing – original draft. **Pierre-Hugues Stefanuto:** Writing – review & editing. **Jean-François Focant:** Writing – review & editing, Supervision, Project administration, Funding acquisition.

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.

Data availability

No data was used for the research described in the article.

Acknowledgments

This research was funded by the FWO/FNRS Belgium EOS Grant 30897864 “Chemical Information Mining in a Complex World”, F.R.S.-F.N.R.S, and Léon Fredericq Foundation scientific grants.

References

[1] K.D. Bartle, P. Myers, History of gas chromatography, TrAC Trends Anal. Chem. 21 (2002) 547–557, https://doi.org/10.1016/S0165-9936(02)00806-3.
[2] M.J.E. Golay, J. Chromatogr. Libr. 17 (1979) 109–114, https://doi.org/10.1016/S0301-4770(08)60640-5.
[3] G.A. Eiceman, J. Gardea-Torresdey, E. Overton, K. Carney, F. Dorman, Gas chromatography, Anal. Chem. 74 (2002) 2771–2780, https://doi.org/10.1021/AC020210P.
[4] I. Špánik, A. Machyňáková, Recent applications of gas chromatography with high-resolution mass spectrometry, J. Sep. Sci. 41 (2018) 163–179, https://doi.org/10.1002/JSSC.201701016.
[5] C.F. Poole, Gas Chromatography, second ed., Elsevier, 2021.
[6] Z. Liu, J.B. Phillips, Comprehensive two-dimensional gas chromatography using an on-column thermal modulator interface, J. Chromatogr. Sci. 29 (1991) 227–231, https://doi.org/10.1093/CHROMSCI/29.6.227.
[7] J. Mommers, S. van der Wal, Column selection and optimization for comprehensive two-dimensional gas chromatography: a review, Crit. Rev. Anal. Chem. 51 (2021) 183–202, https://doi.org/10.1080/10408347.2019.1707643.
[8] H.J. Cortes, B. Winniford, J. Luong, M. Pursch, Comprehensive two dimensional gas chromatography review, J. Sep. Sci. 32 (2009) 883–904, https://doi.org/10.1002/JSSC.200800654.
[9] C. Meinert, U.J. Meierhenrich, A new dimension in separation science: comprehensive two-dimensional gas chromatography, Angew. Chem. - Int. Ed. 51 (2012) 10460–10470, https://doi.org/10.1002/ANIE.201200842.
[10] 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.
[11] 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 1114–1115 (2019) 146–153, https://doi.org/10.1016/j.jchromb.2019.01.029.
[12] 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. 200 (2019) 444–453, https://doi.org/10.1164/rccm.201811-2210OC.
[13] M. Libarondi, Comparing the capabilities of time-of-flight and quadrupole mass spectrometers, LCGC Suppl. 8 (2010) 28–33. https://www.chromatographyonline.com/view/comparing-capabilities-time-flight-and-quadrupole-mass-spectrometers-0. accessed May 23, 2023.
[14] J. Harynuk, T. Górecki, Experimental variables in GC × GC: a complex interplay, Am. Lab. 39 (2007) 36–39.
[15] X. Lu, H. Kong, H. Li, C. Ma, J. Tian, G. Xu, Resolution prediction and optimization of temperature programme in comprehensive two-dimensional gas chromatography, J. Chromatogr. A 1086 (2005) 175–184, https://doi.org/10.1016/j.chroma.2005.05.105.
[16] F.L. Dorman, P.D. Schettler, C.M. English, D.V. Patwardhant, Predicting gas chromatographic separation and stationary-phase selectivity using computer modeling, Anal. Chem. 74 (2002) 2133–2138, https://doi.org/10.1021/ac0110496.
[17] 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.
[18] M. Gaida, F.A. Franchina, P.-H. Stefanuto, J.-F. Focant, Top-down approach to retention time prediction in comprehensive two-dimensional gas chromatography–mass spectrometry, Anal. Chem. 94 (2022) 17081–17089, https://doi.org/10.1021/acs.analchem.2c03107.

[19] Y. Zou, M. Gaida, F.A. Franchina, P.H. Stefanuto, J. Focant, Distinguishing between decaffeinated and regular coffee by HS-SPME-GC × GC-TOFMS, chemometrics, and machine learning, Molecules 27 (2022) 16.
[20] 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 1595 (2019) 158–167, https://doi.org/10.1016/J.CHROMA.2019.02.027.
[21] M. Beccaria, T.R. Mellors, J.S. Petion, C.A. Rees, M. Nasir, H.K. Systrom, J. W. Sairistil, M.A. Jean-Juste, V. Rivera, K. Lavoile, P. Severe, J.W. Pape, P. F. Wright, J.E. Hill, Preliminary investigation of human exhaled breath for tuberculosis diagnosis by multidimensional gas chromatography—Time of flight mass spectrometry and machine learning, J. Chromatogr. B. 1074–1075 (2018) 46–50, https://doi.org/10.1016/J.JCHROMB.2018.01.004.
[22] G. Purcaro, C.A. Rees, J.A. Melvin, J.M. Bomberger, J.E. Hill, Volatile fingerprinting of pseudomonas aeruginosa and respiratory syncytial virus infection in an in vitro cystic fibrosis co-infection model, J. Breath Res. 12 (2018), 046001, https://doi.org/10.1088/1752-7163/AAC2F1.
[23] M. Beccaria, F.A. Franchina, M. Nasir, T. Mellors, J.E. Hill, G. Purcaro, Investigating bacterial volatilome for the classification and identification of mycobacterial species by HS-SPME-GC-MS and machine learning, Molecules 26 (2021), https://doi.org/10.3390/MOLECULES26154600.
[24] J.G. Carbonell, R.S. Michalski, T.M. Mitchell, An overview of machine learning. Mach. Learn., Morgan Kaufmann, 1983, pp. 3–23, https://doi.org/10.1016/B978-0-08-051054-5.50005-4.
[25] D.D. Matyushin, A.Y. Sholokhova, A.K. Buryak, Deep learning based prediction of gas chromatographic retention indices for a wide variety of polar and mid-polar liquid stationary phases, Int. J. Mol. Sci. 22 (2021) 9194, https://doi.org/10.3390/IJMS22179194/S1.
[26] Y. Kobayashi, K. Yoshida, Automated retention time prediction of new psychoactive substances in gas chromatography, Procedia Comput. Sci. 207 (2022) 654–663, https://doi.org/10.1016/J.PROCS.2022.09.120.
[27] G.M. Randazzo, A. Bileck, A. Danani, B. Vogt, M. Groessl, Steroid identification via deep learning retention time predictions and two-dimensional gas chromatography-high resolution mass spectrometry, J. Chromatogr. A 1612 (2020), 460661, https://doi.org/10.1016/j.chroma.2019.460661.
[28] J. Beens, R. Tijssen, J. Blomberg, Prediction of comprehensive two-dimensional gas chromatographic separations. A theoretical and practical exercise, J. Chromatogr. A 822 (1998) 233–251, https://doi.org/10.1016/S0021-9673(98)00649-9.
[29] G. Castello, P. Moretti, S. Vezzani, Retention models for programmed gas chromatography, J. Chromatogr. A 1216 (2009) 1607–1623, https://doi.org/10.1016/j.chroma.2008.11.049.
[30] S. Zhu, X. Lu, Y. Qiu, T. Pang, H. Kong, C. Wu, G. Xu, Determination of retention indices in constant inlet pressure mode and conversion among different column temperature conditions in comprehensive two-dimensional gas chromatography, J. Chromatogr. A 1150 (2007) 28–36, https://doi.org/10.1016/J.CHROMA.2006.09.026.
[31] R. Jaramillo, F.L. Dorman, Thermodynamic modeling of comprehensive two dimensional gas chromatography isovolatility curves for second dimension retention indices based analyte identification, J. Chromatogr. A 1622 (2020), https://doi.org/10.1016/J.CHROMA.2020.461111.
[32] X. Lu, H. Kong, H. Li, C. Ma, J. Tian, G. Xu, Resolution prediction and optimization of temperature programme in comprehensive two-dimensional gas chromatography, J. Chromatogr. A 1086 (2005) 175–184, https://doi.org/10.1016/J.CHROMA.2005.05.105.
[33] 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.
[34] C. Stultz, R. Jaramillo, P. Teehan, F. Dorman, Comprehensive two-dimensional gas chromatography thermodynamic modeling and selectivity evaluation for the separation of polychlorinated dibenzo-*p*-dioxins and dibenzofurans in fish tissue matrix, J. Chromatogr. A 1626 (2020), https://doi.org/10.1016/J.CHROMA.2020.461311.
[35] B.J. Pollo, C.A. Teixeira, J.R. Belinato, M.F. Furlan, I.C. de M. Cunha, C.R. Vaz, G. V. Volpato, F. Augusto, Chemometrics, comprehensive two-dimensional gas chromatography and "omics" sciences: basic tools and recent applications, TrAC Trends Anal. Chem. 134 (2021), 116111, https://doi.org/10.1016/J.TRAC.2020.116111.
[36] P.H. Stefanuto, A. Smolinska, J.F. Focant, Advanced chemometric and data handling tools for GC × GC-TOF-MS: application of chemometrics and related advanced data handling in chemical separations, TrAC - Trends Anal. Chem. 139 (2021), https://doi.org/10.1016/j.trac.2021.116251.
[37] 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. 134 (2021), 116133, https://doi.org/10.1016/J.TRAC.2020.116133.
[38] T.J. Trinklein, C.N. Cain, G.S. Ochoa, S. Schöneich, L. Mikaliunaite, R.E. Synovec, Recent advances in GC × GC and chemometrics to address emerging challenges in nontargeted analysis, Anal. Chem. 95 (2023) 264–286, https://doi.org/10.1021/acs.analchem.2c04235.
[39] 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.
[40] C. Vendeuvre, F. Bertoncini, D. Thiébaut, M. Martin, M.C. Hennion, Evaluation of a retention model in comprehensive two-dimensional gas chromatography, J. Sep. Sci. 28 (2005) 1129–1136, https://doi.org/10.1002/JSSC.200401933.
[41] J.V. Seeley, S.K. Seeley, Model for predicting comprehensive two-dimensional gas chromatography retention times, J. Chromatogr. A 1172 (2007) 72–83, https://doi.org/10.1016/J.CHROMA.2007.09.058.
[42] R.J. Western, P.J. Marriott, Retention correlation maps in comprehensive two-dimensional gas chromatography, J. Sep. Sci. 25 (2002) 832–838, https://doi.org/10.1002/1615-9314.
[43] R.J. Western, P.J. Marriott, Methods for generating second dimension retention index data in comprehensive two-dimensional gas chromatography, J. Chromatogr. A 1019 (2003) 3–14, https://doi.org/10.1016/J.CHROMA.2003.09.006.
[44] S. Bieri, P.J. Marriott, Generating multiple independent retention index data in dual-secondary column comprehensive two-dimensional gas chromatography, Anal. Chem. 78 (2006) 8089–8097, https://doi.org/10.1021/ac060869l.
[45] J.S. Arey, R.K. Nelson, L. Xu, C.M. Reddy, Using comprehensive two-dimensional gas chromatography retention indices to estimate environmental partitioning properties for a complete set of diesel fuel hydrocarbons, Anal. Chem. 77 (2005) 7172–7182, https://doi.org/10.1021/ac051051n.
[46] Y. Nolvachai, C. Kulsing, P.J. Marriott, Multidimensional gas chromatography in food analysis, TrAC - Trends Anal. Chem. 96 (2017) 124–137, https://doi.org/10.1016/J.TRAC.2017.05.001.
[47] C. Von Mühlen, P.J. Marriott, Retention indices in comprehensive two-dimensional gas chromatography, Anal. Bioanal. Chem. 401 (2011) 2351–2360, https://doi.org/10.1007/S00216-011-5247-1/TABLES/1.
[48] B. d'Acampora Zellner, C. Bicchi, P. Dugo, P. Rubiolo, G. Dugo, L. Mondello, Linear retention indices in gas chromatographic analysis: a review, Flavour Fragr. J. 23 (2008) 297–314, https://doi.org/10.1002/FFJ.1887.
[49] M. Jiang, Facile approach for calculation of second dimensional retention indices in comprehensive two dimensional gas chromatography with single injection, Anal. Chem. 91 (2019) 4085–4091, https://doi.org/10.1021/acs.analchem.8b05717.
[50] C. Veenaas, P. Haglund, A retention index system for comprehensive two-dimensional gas chromatography using polyethylene glycols, J. Chromatogr. A 1536 (2018) 67–74, https://doi.org/10.1016/J.CHROMA.2017.08.062.
[51] M. Jiang, C. Kulsing, Y. Nolvachai, P.J. Marriott, Two-dimensional retention indices improve component identification in comprehensive two-dimensional gas chromatography of saffron, Anal. Chem. 87 (2015) 5753–5761, https://doi.org/10.1021/acs.analchem.5b00953.
[52] D.M. Mazur, I.G. Zenkevich, V.B. Artaev, O.V. Polyakova, A.T. Lebedev, Regression algorithm for calculating second-dimension retention indices in comprehensive two-dimensional gas chromatography, J. Chromatogr. A 1569 (2018) 178–185, https://doi.org/10.1016/J.CHROMA.2018.07.038.
[53] J.M.D. Dimandja, G.C. Clouden, I. Colón, J.F. Focant, W.V. Cabey, R.C. Parry, Standardized test mixture for the characterization of comprehensive two-dimensional gas chromatography columns: the Phillips mix, J. Chromatogr. A 1019 (2003) 261–272, https://doi.org/10.1016/J.CHROMA.2003.09.027.
[54] S. Bieri, P.J. Marriott, Dual-injection system with multiple injections for determining bidimensional retention indexes in comprehensive two-dimensional gas chromatography, (2008). https://doi.org/10.1021/ac071367q.
[55] J. Beens, H.G. Janssen, M. Adahchour, U.A.T. Brinkman, Flow regime at ambient outlet pressure and its influence in comprehensive two-dimensional gas chromatography, J. Chromatogr. A 1086 (2005) 141–150, https://doi.org/10.1016/j.chroma.2005.05.086.
[56] J. Harynuk, T. Górecki, Flow model for coupled-column gas chromatography systems, J. Chromatogr. A 1086 (2005) 135–140, https://doi.org/10.1016/J.CHROMA.2005.06.008.
[57] C.F. Poole, S.K. Poole, Column selectivity from the perspective of the solvation parameter model, J. Chromatogr. A 965 (2002) 263–299, https://doi.org/10.1016/S0021-9673(01)01361-9.
[58] J.V. Seeley, E.M. Libby, K.A.H. Edwards, S.K. Seeley, Solvation parameter model of comprehensive two-dimensional gas chromatography separations, J. Chromatogr. A 1216 (2009) 1650–1657, https://doi.org/10.1016/J.CHROMA.2008.07.060.
[59] C. Kulsing, Y. Nolvachai, A.X. Zeng, S.T. Chin, B. Mitrevski, P.J. Marriott, From molecular structures of ionic liquids to predicted retention of fatty acid methyl esters in comprehensive two-dimensional gas chromatography, Chempluschem 79 (2014) 790–797, https://doi.org/10.1002/CPLU.201300410.
[60] A.A. D'Archivio, A. Incani, F. Ruggieri, Retention modelling of polychlorinated biphenyls in comprehensive two-dimensional gas chromatography, Anal. Bioanal. Chem. 399 (2011) 903–913, https://doi.org/10.1007/s00216-010-4326-z.
[61] C. Veenaas, A. Linusson, P. Haglund, Retention-time prediction in comprehensive two-dimensional gas chromatography to aid identification of unknown contaminants, Anal. Bioanal. Chem. 410 (2018) 7931–7941, https://doi.org/10.1007/s00216-018-1415-x.
[62] L.M. Blumberg, Distribution-centric 3-parameter thermodynamic models of partition gas chromatography, J. Chromatogr. A 1491 (2017) 159–170, https://doi.org/10.1016/j.chroma.2017.02.047.
[63] S. Vezzani, P. Moretti, G. Castello, Fast and accurate method for the automatic prediction of programmed-temperature retention times, J. Chromatogr. A 677 (1994) 331–343, https://doi.org/10.1016/0021-9673(94)80161-4.
[64] F. Aldaeus, Y. Thewalim, A. Colmsjö, Prediction of retention times of polycyclic aromatic hydrocarbons and n-alkanes in temperature-programmed gas chromatography, Anal. Bioanal. Chem. 389 (2007) 941–950, https://doi.org/10.1007/s00216-007-1528-0.

[65] F.R. Gonzalez, A.M. Nardillo, Retention index in temperature-programmed gas chromatography, J. Chromatogr. A 842 (1999) 29–49, https://doi.org/10.1016/S0021-9673(99)00158-2.
[66] B. Karolat, J. Harynuk, Prediction of gas chromatographic retention time via an additive thermodynamic model, J. Chromatogr. A 1217 (2010) 4862–4867, https://doi.org/10.1016/j.chroma.2010.05.037.
[67] E.C.W. Clarke, D.N. Glew, Evaluation of thermodynamic functions from equilibrium constants, Trans. Faraday Soc. 62 (1966) 539, https://doi.org/10.1039/tf9666200539.
[68] M. Gaida, F.A. Franchina, P.-H. Stefanuto, J.-F. Focant, Modeling approaches for temperature-programmed gas chromatographic retention times under vacuum outlet conditions, J. Chromatogr. A 1651 (2021), 462300, https://doi.org/10.1016/j.chroma.2021.462300.
[69] K.A.J.M. Stevenson, L.M. Blumberg, J.J. Harynuk, Thermodynamics-based retention maps to guide column choices for comprehensive multi-dimensional gas chromatography, Anal. Chim. Acta 1086 (2019) 133–141, https://doi.org/10.1016/j.aca.2019.08.011.
[70] S. Zhu, S. He, D.R. Worton, A.H. Goldstein, Predictions of comprehensive two-dimensional gas chromatography separations from isothermal data, J. Chromatogr. A 1233 (2012) 147–151.
[71] 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.
[72] S. Hou, K.A.J.M. Stevenson, J.J. Harynuk, A simple, fast, and accurate thermodynamic-based approach for transfer and prediction of GC retention times between columns and instruments Part II: estimation of target column geometry, J. Sep. Sci. 41 (2018) 2553–2558, https://doi.org/10.1002/JSSC.201701344.
[73] 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 III: retention time prediction on target column, J. Sep. Sci. 41 (2018) 2559–2564, https://doi.org/10.1002/jssc.201701345.
[74] J. Leppert, T. Brehmer, M. Wüst, P. Boeker, Estimation of retention parameters from temperature programmed gas chromatography, J. Chromatogr. A 1699 (2023), 464008, https://doi.org/10.1016/J.CHROMA.2023.464008.
[75] 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.
[76] T.M. McGinitie, J.J. Harynuk, Considerations for the automated collection of thermodynamic data in gas chromatography, J. Sep. Sci. 35 (2012) 2228–2232, https://doi.org/10.1002/JSSC.201200192.
[77] T.M. McGinitie, H. Ebrahimi-Najafabadi, J.J. Harynuk, A standardized method for the calibration of thermodynamic data for the prediction of gas chromatographic retention times, J. Chromatogr. A (2014), https://doi.org/10.1016/j.chroma.2014.01.019.
[78] H. Snijders, H.G. Janssen, C. Cramers, Optimization of temperature-programmed gas chromatographic separations I. Prediction of retention times and peak widths from retention indices, J. Chromatogr. A 718 (1995) 339–355, https://doi.org/10.1016/0021-9673(95)00692-3.
[79] A. Barcaru, A. Anroedh-Sampat, H.G. Janssen, Retention time prediction in temperature-programmed, comprehensive two-dimensional gas chromatography: modeling and error assessment, J. Chromatogr. A 1368 (2014) 190–198.
[80] R. Jaramillo, F.L. Dorman, Retention time prediction in thermally modulated comprehensive two-dimensional gas chromatography: correcting second dimension retention time modeling error, J. Chromatogr. A 1581–1582 (2018) 116–124, https://doi.org/10.1016/j.chroma.2018.10.054.
[81] R. Jaramillo, F.L. Dorman, Retention time prediction of hydrocarbons in cryogenically modulated comprehensive two-dimensional gas chromatography: a method development and translation application, J. Chromatogr. A 1612 (2019) 460696.
[82] 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.
[83] A. Burel, M. Vaccaro, Y. Cartigny, S. Tisse, G. Coquerel, P. Cardinael, Retention modeling and retention time prediction in gas chromatography and flow-modulation comprehensive two-dimensional gas chromatography: the contribution of pressure on solute partition, J. Chromatogr. A 1485 (2017) 101–119, https://doi.org/10.1016/j.chroma.2017.01.011.
[84] T. Brehmer, B. Duong, M. Marquart, L. Friedemann, P.J. Faust, P. Boeker, M. Wüst, J. Leppert, Retention database for prediction, simulation, and optimization of GC separations, ACS Omega 8 (2023) 19708–19718, https://doi.org/10.1021/acsomega.3c01348.
[85] C. Quiroz-Moreno, M.F. Furlan, J.R. Belinato, F. Augusto, G.L. Alexandrino, N.G. S. Mogollón, RGCxGC toolbox: an R-package for data processing in comprehensive two-dimensional gas chromatography-mass spectrometry, Microchem. J. 156 (2020), 104830, https://doi.org/10.1016/J.MICROC.2020.104830.
[86] D. Dhall, R. Kaur, M. Juneja, Machine learning: a review of the algorithms and its applications, Lect. Notes Electr. Eng. 597 (2020) 47–63, https://doi.org/10.1007/978-3-030-29407-6_5.
[87] A.M. Jimenez-Carvelo, L. Cuadros-Rodríguez, Data mining/machine learning methods in foodomics, Curr. Opin. Food Sci. 37 (2021) 76–82, https://doi.org/10.1016/J.COFS.2020.09.008.
[88] 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.
[89] P.S. Gromski, Y. Xu, H.L. Kotze, E. Correa, D.I. Ellis, E.G. Armitage, M.L. 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.
[90] 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.
[91] 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, https://doi.org/10.1016/J.CHROMA.2005.04.078.
[92] K. Fawagreh, M.M. Gaber, E. Elyan, Random forests: from early developments to recent advancements, Syst. Sci. Control Eng. 2 (2014) 602–609, https://doi.org/10.1080/21642583.2014.956265.
[93] G. Biau, Analysis of a random forests model, J. Mach. Learn. Res. 13 (2012) 1063–1095.
[94] S. Salcedo-Sanz, J.L. Rojo-Álvarez, M. Martínez-Ramón, G. Camps-Valls, Support vector machines in engineering: an overview, Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 4 (2014) 234–267, https://doi.org/10.1002/WIDM.1125.
[95] L.C. Lee, C.Y. Liong, A.A. Jemain, Partial least squares-discriminant analysis (PLS-DA) for classification of high-dimensional (HD) data: a review of contemporary practice strategies and knowledge gaps, Analyst 143 (2018) 3526–3539, https://doi.org/10.1039/C8AN00599K.
[96] V.B. Mathema, K. Duangkumpha, K. Wanichthanarak, N. Jariyasopit, E. Dhakal, N. Sathirapongsasuti, C. Kitiyakara, Y. Sirivatanauksorn, S. Khoomrung, CRISP: a deep learning architecture for GC × GC–TOFMS contour ROI identification, simulation and analysis in imaging metabolomics, Brief. Bioinform. 23 (2022), https://doi.org/10.1093/BIB/BBAB550.
[97] J. Oh, A. Oldani, T. Lee, L. Shafer, Deep learning algorithms for assessing sustainable jet fuels from two-dimensional gas chromatography, in: AIAA Sci. Technol. Forum Expo. AIAA SciTech Forum 2022, 2022, https://doi.org/10.2514/6.2022-0228.
[98] T. Cajka, J. Hajslova, F. Pudil, K. Riddellova, Traceability of honey origin based on volatiles pattern processing by artificial neural networks, J. Chromatogr. A 1216 (2009) 1458–1462, https://doi.org/10.1016/J.CHROMA.2008.12.066.
[99] 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.
[100] G. Purcaro, C.A. Rees, J.A. Melvin, J.M. Bomberger, J.E. Hill, Volatile fingerprinting of Pseudomonas aeruginosa and respiratory syncytial virus infection in an in vitro cystic fibrosis co-infection model, J. Breath Res. 12 (2018), https://doi.org/10.1088/1752-7163/aac2f1.
[101] M. Beccaria, T.R. Mellors, J.S. Petion, C.A. Rees, M. Nasir, H.K. Systrom, J. W. Sairistil, M.A. Jean-Juste, V. Rivera, K. Lavoile, P. Severe, J.W. Pape, P. F. Wright, J.E. Hill, Preliminary investigation of human exhaled breath for tuberculosis diagnosis by multidimensional gas chromatography—Time of flight mass spectrometry and machine learning, J. Chromatogr. B 1074–1075 (2018) 46–50, https://doi.org/10.1016/j.jchromb.2018.01.004.
[102] M. Beccaria, C. Bobak, B. Maitshotlo, T.R. Mellors, G. Purcaro, F.A. Franchina, C. A. Rees, M. Nasir, A. Black, J.E. Hill, Exhaled human breath analysis in active pulmonary tuberculosis diagnostics by comprehensive gas chromatography-mass spectrometry and chemometric techniques, J. Breath Res. 13 (2019), https://doi.org/10.1088/1752-7163/aae80e.
[103] S.L. Andersen, F.B.S. Briggs, J.H. Winnike, Y. Natanzon, S. Maichle, K.J. Knagge, L.K. Newby, S.G. Gregory, Metabolome-based signature of disease pathology in MS, Mult. Scler. Relat. Disord. 31 (2019) 12–21, https://doi.org/10.1016/J.MSARD.2019.03.006.
[104] E.B. Franklin, L.D. Yee, B. Aumont, R.J. Weber, P. Grigas, A.H. Goldstein, Ch3MS-RF: a random forest model for chemical characterization and improved quantification of unidentified atmospheric organics detected by chromatography-mass spectrometry techniques, Atmos. Meas. Tech. 15 (2022) 3779–3803, https://doi.org/10.5194/AMT-15-3779-2022.
[105] G.L. Alexandrino, P.S. Prata, F. Augusto, Discriminating lacustrine and marine organic matter depositional paleoenvironments of Brazilian crude oils using comprehensive two-dimensional gas chromatography-quadrupole mass spectrometry and supervised classification chemometric approaches, Energy Fuels 31 (2017) 170–178, https://doi.org/10.1021/ACS.ENERGYFUELS.6B01925.
[106] E. Barberis, E. Amede, S. Khoso, L. Castello, P.P. Sainaghi, M. Bellan, P.E. Balbo, G. Patti, D. Brustia, M. Giordano, R. Rolla, A. Chiocchetti, G. Romani, M. Manfredi, R. Vaschetto, Metabolomics diagnosis of covid-19 from exhaled breath condensate, Metabolites 11 (2021), https://doi.org/10.3390/METABO11120847.
[107] R.A.M. Lima, S.M.M. Ferraz, V.G.K. Cardoso, C.A. Teixeira, L.W. Hantao, Authentication of fish oil (omega-3) supplements using class-oriented chemometrics and comprehensive two-dimensional gas chromatography coupled to mass spectrometry, Anal. Bioanal. Chem. (2022), https://doi.org/10.1007/S00216-022-04428-2.

[108] C. Hall, B. Creton, B. Rauch, U. Bauder, M. Aigner, Probabilistic mean quantitative structure-property relationship modeling of jet fuel properties, Energy Fuels 36 (2022) 463–479, https://doi.org/10.1021/ACS.ENERGYFUELS.1C03334.
[109] M.D. Sorochan Armstrong, O.R. Arredondo Campos, C.C. Bannon, A.P. de la Mata, R.J. Case, J.J. Harynuk, Global metabolome analysis of *Dunaliella tertiolecta, Phaeobacter italicus* R11 Co-cultures using thermal desorption—Comprehensive two-dimensional gas chromatography—Time-of-flight mass spectrometry (TD-GC × GC-TOFMS), Phytochemistry 195 (2022), https://doi.org/10.1016/J.PHYTOCHEM.2021.113052.
[110] K.A. Favela, M.J. Hartnett, J.A. Janssen, D.W. Vickers, A.J. Schaub, H.A. Spidle, K.S. Pickens, Nontargeted analysis of face masks: comparison of manual curation to automated GCxGC processing tools, J. Am. Soc. Mass Spectrom. 32 (2021) 860–871, https://doi.org/10.1021/JASMS.0C00318.
[111] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32, https://doi.org/10.1023/A:1010933404324.
[112] P. Probst, M.N. Wright, A. Boulesteix, Hyperparameters and tuning strategies for random forest, WIREs Data Min. Knowl. Discov. 9 (2019), https://doi.org/10.1002/widm.1301.
[113] E.D. Strozier, D.D. Mooney, D.A. Friedenberg, T.P. Klupinski, C.A. Triplett, Use of comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometric detection and random forest pattern recognition techniques for classifying chemical threat agents and detecting chemical attribution signatures, Anal. Chem. 88 (2016) 7068–7075, https://doi.org/10.1021/ACS.ANALCHEM.6B00725.
[114] R Core Team, R: A Language and Environment for Statistical Computing., (n.d.). https://www.r-project.org/.
[115] I. The MathWorks, MATLAB and Statistics Toolbox, (n.d.).
[116] L. Yang, A. Shami, On hyperparameter optimization of machine learning algorithms: theory and practice, Neurocomputing 415 (2020) 295–316, https://doi.org/10.1016/J.NEUCOM.2020.07.061.
[117] M.J. Rist, A. Roth, L. Frommherz, C.H. Weinert, R. Krüger, B. Merz, D. Bunzel, C. Mack, B. Egert, A. Bub, B. Görling, P. Tzvetkova, B. Luy, I. Hoffmann, S. E. Kulling, B. Watzl, Metabolite patterns predicting sex and age in participants of the Karlsruhe metabolomics and nutrition (KarMeN) study, PLoS One 12 (2017), https://doi.org/10.1371/JOURNAL.PONE.0183228.
[118] P.H. Stefanuto, R. Romano, C.A. Rees, M. Nasir, L. Thakuria, A. Simon, A.K. Reed, N. Marczin, J.E. Hill, Volatile organic compound profiling to explore primary graft dysfunction after lung transplantation, Sci. Rep. 12 (1) (2022) 1–10, https://doi.org/10.1038/s41598-022-05994-2.
[119] J. Shawe-Taylor, S. Sun, A review of optimization methodologies in support vector machines, Neurocomputing 74 (2011) 3609–3618, https://doi.org/10.1016/J.NEUCOM.2011.06.026.
[120] Z. Cen, B. Lu, Y. Ji, J. Chen, Y. Liu, J. Jiang, X. Li, X. Li, Virus-induced breath biomarkers: a new perspective to study the metabolic responses of COVID-19 vaccinees, Talanta 260 (2023), https://doi.org/10.1016/J.TALANTA.2023.124577.
[121] S. Li, Y. Hu, W. Liu, Y. Chen, F. Wang, X. Lu, W. Zheng, Untargeted volatile metabolomics using comprehensive two-dimensional gas chromatography-mass spectrometry—A solution for orange juice authentication, Talanta 217 (2020), 121038, https://doi.org/10.1016/J.TALANTA.2020.121038.
[122] A.P. Bradley, The use of the area under the ROC curve in the evaluation of machine learning algorithms, Pattern Recognit. 30 (1997) 1145–1159, https://doi.org/10.1016/S0031-3203(96)00142-2.
[123] H. Li, X. Wu, S. Wu, L. Chen, X. Kou, Y. Zeng, D. Li, Q. Lin, H. Zhong, T. Hao, B. Dong, S. Chen, J. Zheng, Machine learning directed discrimination of virgin and recycled poly(ethylene terephthalate) based on non-targeted analysis of volatile organic compounds, J. Hazard. Mater. 436 (2022), https://doi.org/10.1016/J.JHAZMAT.2022.129116.
[124] J. Cai, J. Luo, S. Wang, S. Yang, Feature selection in machine learning: a new perspective, Neurocomputing 300 (2018) 70–79, https://doi.org/10.1016/J.NEUCOM.2017.11.077.
[125] Y. Saeys, I. Inza, P. Larrañaga, A review of feature selection techniques in bioinformatics, Bioinformatics 23 (2007) 2507–2517, https://doi.org/10.1093/BIOINFORMATICS/BTM344.
[126] P. Cunningham, B. Kathirgamanathan, S.J. Delany, Feature selection tutorial with python examples, (2021). https://doi.org/10.48550/arXiv.2106.06437.
[127] P. Dhal, C. Azad, A comprehensive survey on feature selection in the various fields of machine learning, Appl. Intell. 52 (4) (2021) 4543–4581, https://doi.org/10.1007/S10489-021-02550-9.
[128] A.C. Paiva, L.W. Hantao, Exploring a public database to evaluate consumer preference and aroma profile of lager beers by comprehensive two-dimensional gas chromatography and partial least squares regression discriminant analysis, J. Chromatogr. A 1630 (2020), 461529, https://doi.org/10.1016/J.CHROMA.2020.461529.
[129] I.M. Johnstone, D.M. Titterington, Statistical challenges of high-dimensional data, Philos. Trans. R. Soc. A Math. Phys. Eng. Sci. 367 (2009) 4237–4253, https://doi.org/10.1098/RSTA.2009.0159.