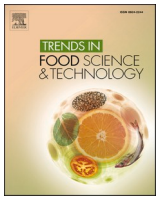
Trends in Food Science & Technology 147 (2024) 104481

Contents lists available at ScienceDirect

Trends in Food Science & Technology

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



Metabolomics and chemometrics: The next-generation analytical toolkit for the evaluation of food quality and authenticity

Pascual García-Pérez [a,b,*], Pier Paolo Becchi [a], Leilei Zhang [a], Gabriele Rocchetti [c], Luigi Lucini [a]

[a] Department for Sustainable Food Process, Università Cattolica del Sacro Cuore, Via Emilia Parmense 84, 29122, Piacenza, Italy
[b] Universidade de Vigo, Nutrition and Bromatology Group, Department of Analytical Chemistry and Food Science, Faculty of Science, E-32004, Ourense, Spain
[c] Department of Animal Science, Food and Nutrition, Università Cattolica del Sacro Cuore, Via Emilia Parmense 84, 29122, Piacenza, Italy

ARTICLE INFO

Handling Editor: Prof. F. Toldra

Keywords:
Unsupervised statistics
Supervised modeling
Machine learning
Artificial neural networks
Authenticity markers
Food integrity

ABSTRACT

Background: The advances in NMR and mass spectrometry metabolomics allows a comprehensive profiling of foods, potentially covering geographical origin, authenticity, quality and integrity issues. However, mining specific effects within the corresponding datasets is challenging due to the presence of a set of interacting factors that finally determine metabolomics signatures.

Scope and approach: This review provides an overview of the different metabolomics approaches used in food quality and authenticity, then focusing on different chemometric approaches for data interpretation. In particular, data interpretation is hierarchically presented, starting from unsupervised (PCA, hierarchical clusters) to supervised multivariate statistics like OPLS and AMOPLS multiblock ANOVA discriminant approaches. Finally, machine learning approaches like Artificial Neural Networks are discussed as the novel and emerging tool to support food integrity issues.

Key findings and conclusions: Tailored data mining approaches are advisable, rather than unique solutions, with unsupervised statistics that naively provide qualitative recognition of patterns, and supervised modeling that support markers identification. Nonetheless, machine learning approaches are emerging as a novel approach able to interpretate complex metabolomics signatures.

1. Introduction

Bearing in mind the impact and potential risks to human health caused by food adulteration, food safety and integrity issues are critical concerns in food science. Consequently, ensuring authenticity has become a shared goal among regulatory authorities, food processors, retailers, and consumers. To this aim, a series of analytical techniques emerged as promising tools, even though regulatory authorities are required to consistently update analytical methods and conditions to validate authenticity, thereby enabling effective enforcement of laws and regulations (Danezis, Tsagkaris, Brusic, & Georgiou, 2016). Determining a comprehensive chemical profile of food products provides valuable insights into their quality, geographical origin and traceability. Variations in the chemical fingerprints can occur due to various factors, including changes in metabolite levels resulting from different conditions such as production systems, raw materials' geographical origin, and adulteration practices (Dinis, Tsamba, Jamin, & Camel, 2023). In this scenario, recognizing the importance of solid and reliable analytical methods is essential to use accurate and coherent approaches to assess the quality and authenticity of food products and their derivatives, as it is the case of medicinal plants and herbs with associated nutritional and health-promoting properties.

Metabolomics has emerged in the last decades as a robust and powerful technology to ensure an easy and feasible large-scale determination of the chemical composition of food matrices with a plethora of downstream applications (García-Pérez et al., 2023). In particular, metabolomics is extremely powerful in providing information about the possible interactions between a particular metabolome and a specific environment (P. Deng et al., 2019). Omics-based approaches have been previously described to investigate the chemical fingerprint as related to the so-called "terroir" effect (Lucini, Rocchetti, & Trevisan, 2020), thus providing clear evidence that the organoleptic and nutritional quality of foods is strongly affected by the area of cultivation coupled with the other pedo-climatic conditions (Marchev et al., 2021). Metabolomics combines high-throughput analytical techniques with bioinformatics methods to provide the most information about a wide range of

* Corresponding author. Department for Sustainable Food Process, Università Cattolica del Sacro Cuore, Via Emilia Parmense 84, 29122 Piacenza, Italy.
E-mail address: pasgarcia@uvigo.es (P. García-Pérez).

https://doi.org/10.1016/j.tifs.2024.104481
Received 10 September 2023; Received in revised form 9 March 2024; Accepted 5 April 2024
Available online 7 April 2024

metabolites and their up-*vs* down-accumulation trends when considering the biological phenomenon under investigation (Chen, Li, & Xu, 2022). The high throughput outcome attributed to metabolomics leads to the generation of massive amounts of data that can be challenging to analyze to provide a defined and concise interpretation. In this sense, different chemometric and statistical methods have been widely incorporated in metabolomics workflows to enhance experimental design and to extract valuable insights from extensive and intricate datasets related to various aspects of food analysis (Medina, Perestrelo, Silva, Pereira, & Câmara, 2019; Noshad, Behbahani, & Karabagias, 2023). These chemometric techniques are crucial in handling data generated by both targeted and untargeted metabolomic platforms. In particular, multivariate statistical tools have been primarily employed to identify significant trends and patterns, facilitating the extraction of meaningful information from raw data provided by previous analytical stages.

The combination of metabolomics with multivariate statistics in food science has been massively investigated in the last years, being the topic of more than 650 scientific publications in the period 2020–2024 (source: Scopus, March 2024). On these bases, the integration of appropriate chemometric approaches is crucial for providing valid conclusions from metabolomics datasets and obtaining significant information. Due to the increasing requirements in the field of metabolomics devoted to food authenticity, many efforts are currently being developed to facilitate this task and provide a deeper insight into the information concealed beneath analytical datasets. This paradigm has prompted the development of data mining algorithms based on machine learning technologies to provide a robust and exhaustive description of highly complex data frames, simplifying the overall workflows and accelerating the knowledge in metabolomics (Qiu et al., 2023). Moreover, in the era of big data, the combination of food metabolomics with other omics technologies motivated the birth of Foodomics, based on the integration of food genomics, transcriptomics, proteomics, and metabolomics, therefore urging the design of novel deep-learning computational algorithms to merge all the generated knowledge into a single significant output (Picard, Scott-Boyer, Bodein, Périn, & Droit, 2021).

Starting from this background, the present review aims to provide an updated and comprehensive overview of recent developments in the overall workflow used in metabolomics and the description of both classical and novel chemometrics tools applied for quality and traceability purposes. The implementation of robust and meaningful metabolomics protocols, combining high throughput analytical and computational approaches, will undoubtedly assist in the current demands of the food industry and research sectors for food authenticity and quality evaluation, further enhancing the safety and reliability of the whole food supply chain.

2. Metabolomics-based approaches in food quality and authenticity

In the last years, metabolomics-based platforms have been successfully used in different fields of food science, including food safety evaluation (Shubo Li et al., 2021), quality management (Wu et al., 2022), aroma/flavor chemistry (Diez-Simon, Mumm, & Hall, 2019), and functional discovery (B. Peng, Li, & Peng, 2015). However, this review aims to present the recent trends and advances in metabolomics, mainly based on chromatography coupled to high-resolution mass spectrometry (HRMS) and nuclear magnetic resonance (NMR), specifically considering "*food quality*" and "*food authenticity*". Both terms belong to a broader and relatively recent research field known as "foodomics", including besides metabolomics, also lipidomics- and proteomics-based applications (among others) (Valdés et al., 2022). In addition, this review provides new perspectives and insights into the methodology, analytical techniques, identification levels in metabolomics and the available data processing in food analysis by LC/GC-HRMS and/or NMR. Interestingly, according to databases searching (PubMed, Scopus; September 2023), the number of scientific publications dealing with "*metabolomics*" applied to "*food authenticity*" and "*food quality*" issues has dramatically increased over a short period, moving from 2 articles in 2011 up to 38 articles in the sole 2023 (Fig. 1).

2.1. Targeted vs untargeted metabolomics and confidence levels of identification

One of the significant aims of metabolomics is to investigate a broad set of quite different metabolites (from a chemical and structural point of view) in a cell, tissue, or organism (Aderemi, Ayeleso, Oyedapo, & Mukwevho, 2021). In this regard, the very high number of mass features (metabolites) provided by a metabolomics-based approach confers a hypothesis-free relatively instantaneous overview of relevant biological processes, thanks to the collection of a series of crucial smaller-molecular-weight metabolites involved in several biochemical pathways, such as amino acids, peptides, organic acids, lipids, fatty acids, hormones, oligopeptides, and vitamins (Selamat, Rozani, & Murugesu, 2021). As far as its involvement in food authentication is concerned, this wide chemical diversity directly allows the detecting of emerging and potential food frauds (Cubero-Leon, Peñalver, & Maquet, 2014).

Under the term metabolomics, we usually consider two different analytical approaches, namely targeted and untargeted metabolomics-based workflows (Roberts, Souza, Gerszten, & Clish, 2012). The fast development of omics techniques based on untargeted analyses opened the possibility of screening many metabolites in biological samples, including foods. The untargeted nature of these approaches is of primary importance because they do not require a hypothesis *a priori* regarding the compounds impacted mainly by the phenomenon under investigation (Troisi, Richards, Scala, & Landolfi, 2022). Therefore, according to untargeted metabolomics, there is a limited knowledge of which metabolites and/or class of metabolites are detected before the gathering of data, and most importantly, the compounds identified will be strictly correlated to the comprehensive spectral database used (Bittremieux, Wang, & Dorrestein, 2022). One of the most exciting advantages of untargeted approaches is its high-throughput nature, allowing the discovery of new compounds and biomarkers in the studied samples to be further validated by *ad-hoc* targeted studies (Selamat et al., 2021). More importantly, due to its extensive nature, untargeted metabolomics must be combined with advanced data reduction techniques and multivariate statistical approaches (Chen et al., 2022; Worley & Powers, 2012). This last step is crucial and can reduce the complexity of datasets usually characterized by thousands of compounds and unknowns. Looking at the main drawbacks of this technique, the time required to process large amounts of raw data, and several difficulties in identification and classification can be encountered (Gertsman & Barshop, 2018). Other

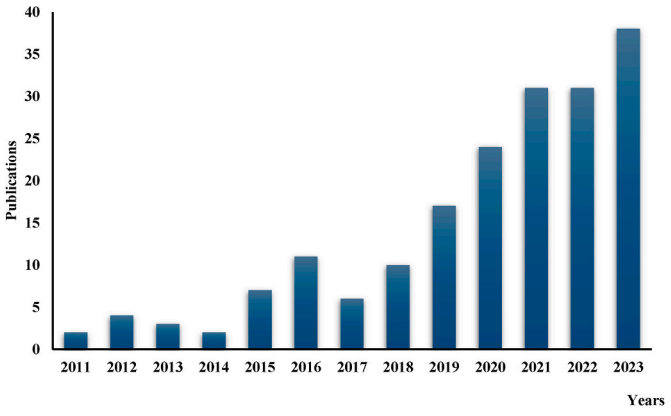


Fig. 1. Reported scientific publications by year (2011–2023) dealing with the keywords "metabolomics", "food authenticity", and "food quality" (source: PubMed and Scopus).

factors of variability depend on the specific analytical platform used, together with the impossibility of quantifying all the discriminant biomarkers detected (i.e., a semi-quantitative nature), and the need to manage false positives and false negatives, thus hampering the possibility to compare results between different studies on different food matrices (Di Minno, Gelzo, Stornaiuolo, Ruoppolo, & Castaldo, 2021). Therefore, untargeted approaches pose extreme challenges to identifying and detecting metabolites, and further research is required to validate the biomarkers identified for food quality and food authenticity applications.

Among the most helpful compound databases commonly used for compound identification as related to omics applications (Blaženović, Kind, Ji, & Fiehn, 2018), Mass Bank of North America (MoNA; auto-curating repository and metadata-centric of mass spectral records), PubChem (small molecules, metadata), ChemSpider (small molecules, curated data), KEGG (pathway database, multiple species), MetaCyc (pathway database, multiple species), BRENDA (enzymes and metabolism data), HMDB (human metabolites), FooDB (food compositional data), CHEBI (molecules of biological interest), UNPD (secondary plant metabolites), and MINE (*in silico* prediction of several metabolite species) can be mentioned. However, despite the high number of databases available to date, matching experimental reference spectra to reference databases is still a limited and complex process that covers only a fraction of the detectable metabolome. Several algorithms widely used for peak detection and mass spectral deconvolution are to date characterized by robust in-built database search algorithms, such as the NIST MS Search GUI, NIST MS PepSearch or MS-DIAL software (Blaženović et al., 2018). Commercial software from mass spectrometry vendors uses similar algorithms. An essential aspect of untargeted metabolomics is represented by the confidence levels for compound annotation. In the last years, several confidence levels have been proposed, and to date, five different levels can be listed (Blaženović et al., 2018; Zulfiqar, Gadelha, Steinbeck, Sorokina, & Peters, 2023). The "Level 0″ is characterized by an unambiguous 3D structure referring to an isolated and fully characterized pure compound (including information on its stereochemistry); the "Level 1″ annotation refers to a confident 2D structure and it is based on at least two orthogonal techniques, such as MS/MS and retention time or collision cross-section (CCS) from ion mobility; the "Level 2″ refers to a probable compound structure deriving from matching to literature data or spectral databases, and including at least two orthogonal pieces of information to exclude all other candidates; the "Level 3″ refers to possible structures of classes of compounds (including isomers and/or substructure matching); finally, "Level 4″ is the lowest confidence level and refers to the presence in the sample of unknowns. These confidence levels have been provided by the Metabolomics Standards Initiative (MSI) of the Metabolomics Society (Sumner et al., 2007), and it was further assessed by the Coordination of Standards in MetabOlomicS (COSMOS), to integrate the data access applied to different metabolomics approaches (Salek et al., 2015). According to the untargeted-based studies reviewed, it is highly recommended to integrate the level of annotation for each compound into metabolomic profiling reports.

The approach is quite different regarding targeted metabolomics, and the chemical features of the metabolites of interest must be carefully studied and considered before data acquisition (Roberts et al., 2012). In this regard, the analytical workflow (from acquisition to data elaboration and processing) is designed *ad-hoc* to provide high precision, selectivity, and reliability (Cao et al., 2020). Targeted analyses focus on a specific group of intended metabolites, with most cases requiring identification and quantification of as many metabolites within the group (Schwaiger-Haber et al., 2021). Targeted analyses are important for assessing the quantitative and absolute behavior of a specific group of metabolites in the samples under determined and/or controlled conditions (Cao et al., 2020; Roberts et al., 2012; Schwaiger-Haber et al., 2021). In targeted metabolomics applications, the extraction process must be carefully optimized according to the class of compounds of interest and the analytical approach to be performed (Roberts et al., 2012). On the other hand, the extraction step for untargeted metabolomics applications is more straightforward and usually based on the combination of polar and apolar solvents (L. Wang, Naser, Spalding, & Patti, 2019, pp. 1–15), and even solvent-free extractions for GC approaches (Zhang et al., 2016), according to analytical instruments to be used. Therefore, in targeted applications, once the chemical identity of metabolites is known, the subsequent process of deriving biological information from the acquired data can start relatively fast. To date, metabolomics-based approaches have been optimized either with MS or NMR. NMR can measure metabolite levels in intact tissue, while instruments equipped with an MS analyzer provide a more comprehensive detection range comprising metabolites of different polarities (Selamat et al., 2021). According to the literature reviewed, the future research trend will be based on using consecutive untargeted and targeted approaches to reliably and accurately classify and measure metabolites.

2.2. Hyphenated techniques and detection technologies in metabolomics

In the last years, targeted and untargeted metabolomics-based approaches allowed the detection and measurements of rather complex classes of metabolites; these latter can be measured using quite different but complementary analytical methods. New metabolite analysis and detection approaches have been developed to address and overcome some of the challenges involved with metabolomics analysis. Before mentioning the technologies available, it is important to remember that sample preparation in untargeted and targeted metabolomics should allow reproducible extractions of as many molecules as possible (Mushtaq, Choi, Verpoorte, & Wilson, 2014; Schippers et al., 2023). Overall, the detection methods or instruments used in food authentication and food quality mainly include high-performance or ultra-high-performance liquid chromatography (HPLC and UHPLC, respectively), Fourier transform infrared (FTIR), chromatography coupled with MS, and NMR (Chowdhury et al., 2023; Selamat et al., 2021). Overall, it is important to mention that no single analytical technique can isolate and detect all of the different molecules simultaneously. Therefore, different technologies should be better combined to reach this ambitious goal (the so-called multi-omics approach) (Beale, Kouremenos, & Palombo, 2016). A comprehensive overview of the most recent scientific publications dealing with the application of metabolomics (as one single approach and multi-omics approach combined with several statistics and machine learning methodologies) can be found in Table 1, considering the period 2019–2024. Overall, it is interesting to notice that this high-throughput approach has been exploited on several and quite different food matrices and product categories, such as cheese, fermented sausages, milk, meat, honey, plant foods for human nutrition, herbs, spices, nuts, among others (Table 1).

2.2.1. GC- and LC-HRMS metabolomics-based applications

In food quality and authenticity fields, mass spectrometry is often combined with chromatographic techniques involving both gas and liquid platforms. GC or LC separates the components of a complex mixture, and then the mass spectrometer detects and characterizes components. In recent years, mass spectrometry has focused primarily on instrumental changes to achieve higher mass resolution, accurate mass measurement, higher sensitivity, improved fragmentation, and improved linearity (Shuzhao Li, Dunlop, Jones, & Corwin, 2016). Usually, the most used MS detectors are based on different "in time" and "in space" platforms (Crupi, Genghi, & Antonacci, 2014), based on Orbitrap and Q-TOF technologies.

The method of GC-MS has been implemented in metabolomics for a considerable time due to its high spread capacity and robustness; it is particularly suitable for testing volatile organic compounds and compounds associated with a primary metabolism (Fiehn, 2016). Detection using GC-MS is made easier by the available *in-house* validated databases (e.g., Wiley and NIST libraries) of volatile compounds that can be used

Table 1
Overview of food metabolomics approaches and their combination with chemometrics analyses for food authentication and quality determination.

Analytical techniques	Chemometric methods	Food matrix	Main purposes	References
UHPLC-Orbitrap-HRMS	PCA, OPLS-DA	Hard cheese	Feeding strategy and Geographical origin	Becchi et al. (2023)
UHPLC-Orbitrap-HRMS	PCA, OPLS-DA	Fermented sausages	Food safety	Rocchetti et al. (2023)
UHPLC-Orbitrap-HRMS	PLS, OPLS-DA	Milk	Thermal processing	Xie et al. (2024)
UHPLC-Orbitrap-HRMS	HCA, PCA, OPLS-DA	Meat	Quality assessment	Kerner et al. (2023)
UHPLC-QTOF-HRMS	PCA, PLS, OPLS-DA	Thyme	Terroir and phenolic profiling	Rivera-Pérez et al. (2022)
UHPLC-QTOF-HRMS	HCA, PLS, OPLS-DA	Thyme	Authenticity	Rivera-Pérez, García-Pérez, Romero-González, Garrido Frenich, and Lucini (2023)
LC-QTOF-MS/MS	PCA	Milk	Feeding strategy	Tzora et al. (2022)
UHPLC-ESI-MS/MS	PCA, OPLS-DA	Tea	storage time	(S. Zhang, García-Pérez, et al., 2023)
NMR	PCA	Coffee	Aging time	Borém et al. (2023)
NMR	PCA	Wine	Soil type	Bambina et al. (2023)
NMR	ANOVA, PLS	Pummelo	Postharvest storage	(J. Liu et al., 2023)
NMR	OPLS-DA, PLS	Beef	Aging time	Bischof et al. (2023)
NMR	PCA, OPLS-DA	Pomegranate wine	Fermentation process	Girelli et al. (2023)
NMR	PCA, PLS, OPLS-DA,	EVOO	Adulteration	Beteinakis et al. (2023)
UHPLC-Orbitrap-HRMS	PCA, OPLS-DA	Wine	Traceability	Zambianchi et al. (2023)
LC-HRMS + NMR	HCA, PLS, OPLS-DA	Thyme	Geographical origin, Authenticity	Rivera-Pérez, Romero-González, and Frenich (2023)
GC–MS, UPLC-ESI-MS	PCA, OPLS-DA	Tilapia (fish)	Food deterioration	Cheng et al. (2023)
NMR	HCA, PCA	Hazelnuts	Geographical origin	Shakiba et al. (2022)
LC-QTOF-MS/MS	PCA, OPLS-DA, ANN	EVOO	Authenticity	Senizza et al. (2023)
NMR	HCA, PCA, OPLS-DA, KNN, SVM, RF	Black Tea	Geographic origin, variety, processing	(Cui, Xu, et al., 2023)
HPLC-UV/Vis-MS, UHPLC-QTOF/MS	PCA, PLS-DA	Lemon juice	Adulteration	Lyu, Yuan, Liu, Simon, and Wu (2022)
GC × GC-MS	OPLS-DA	Fish oil	Food quality	Lima et al. (2023)
GC-TOF-MS	PCA, HCA, PLS, SVM	Whole-wheat flours	Authenticity, botanical origin	Lamine et al. (2023)
GC-TOF-MS	PCA, SVM, KNN, PLS, ANNs, RF	EVOO	Authenticity	Asharindavida et al. (2023)
NMR	SVM	*Asparagus officinalis* L.	Geographic origin	Klare et al. (2020)
NMR	PCA, PLS	Maltese honey	Geographic origin	Spiteri, Lia, and Farrugia (2020)
NMR	RF	Chia, linseed, and sesame	Authenticity	Erban et al. (2019)
UHPLC-QTOF-HRMS	ANNs, OPLS-DA	*Bryophyllum* plants	Authenticity	García-Pérez et al. (2021b)
UHPLC-QTOF-HRMS	ANNs, OPLS-DA	*Bryophyllum* plants	Authenticity	García-Pérez et al. (2022)
UHPLC-QTOF-HRMS	ANNs, PLS, RF, SVM	Tomato and blueberry	Flavor identification	Colantonio et al. (2022)
GC-TOF-MS	ANNs, RF, SVM	Rougui Wuyi rock tea	Geographic origin	(Y. Peng et al., 2023)

Abbreviations: ANNs, artificial neural networks; ANOVA, analysis of variance; ESI, electrospray ionization; EVOO, extra-virgin olive oil; GC, gas chromatography; GC × GC, tandem gas chromatography; HCA, hierarchical cluster analysis; HPLC, high performance liquid chromatography; HRMS, high-resolution mass spectrometry; KNN, k-nearest neighbours; LC, liquid chromatography; MS, mass spectrometry; MS/MS, tandem mass spectrometry; NMR, nuclear magnetic resonance; OPLS-DA, orthogonal projection to latent structures discriminant analysis; PCA, principal component analysis; PLS, partial least squares; QTOF, quadrupole-time-of-flight; RF, random forest; SVM, support vector machines; TOF, time-of-flight; UHPLC, ultra-high performance liquid chromatography; UV/Vis, ultraviolet–visible spectroscopy.

to confirm the structural identity of the metabolites detected. The GC-MS or HS-SPME-GC-MS approach is usually combined with other LC-MS techniques to provide more information regarding food quality and deterioration. Cheng, Mei, and Xie (2023) evaluated through this latter the odor deterioration of tilapia (*Oreochromis mossambicus*) fillets during cold storage (Cheng et al., 2023), while Lima et al. (2023) studied the changes of spilled oil because of weathering processes through GC × GC-TOF-MS combined with ICR-FT MS (Lima et al., 2023). Recently, an exciting approach dealing with the combination of GC with ion mobility spectrometry (IMS) allowed the deep study and discrimination of botanical origin of different honey samples (Christmann, Rohn, & Weller, 2022). Also, Daygon and co-authors (2016) successfully evaluated the impact of rice cooking on several volatile compounds using untargeted GC × GC-TOF-MS combined with a sensory analysis (Daygon et al., 2016).

LC-MS is ideally designed to analyze polar compounds; however, several possibilities are available, including reverse-phase separation using *ad-hoc* chromatographic conditions (Lioupi et al., 2023; Naser et al., 2018). LC-MS is a combined analytical chemistry technique used to distinguish multiple components in complex mixtures and provide the molecular identity of individual components using high sensitivity and specificity. Nevertheless, quantifying certain chemical substances can be difficult for foodstuffs with extremely complex matrices, especially because of their potential to interact with unknown metabolites (matrix effect) (Ćeranić et al., 2020). LC equipped with high-resolution mass spectrometry provides a broader compounds coverage and fragmentation details for accurate identification. Interestingly, the ion source type (based on different soft ionization mechanisms) and the mass analyzer used may improve the spectral data obtained from the analysis. Electrospray ionization source (ESI) is a robust ion source that can detect a wide range of metabolites, from small metabolites to peptides (Ren, Zhang, Kong, & Wang, 2018). Furthermore, high-resolution mass spectrometers, such as QTOF and Orbitrap detectors, improve the limits of detection of targeted metabolites with high mass accuracy, thus allowing the determination of molecular formulas for small molecules, together with identifying some lower-abundance molecules in the food matrix that are often neglected. LC-MS platforms are widely used to determine the geographical origin and distinguish the cultivation methods of certain agricultural products, such as maize and grains, based on their metabolome (Z. Li et al., 2022; Mattoo et al., 2023). Other recent studies are available regarding applying LC-MS metabolomics in food composition, classification, adulteration and traceability. In

particular, Danesaki et al. used a combined targeted and untargeted metabolomics approach based on UHPLC-QTOF/MS to detect 1% apple juice and grape juice mixed with pomegranate juice for adulteration purposes (Dasenaki, Drakopoulou, Aalizadeh, & Thomaidis, 2019). In a previous work (Rocchetti et al., 2023), a metabolomics approach based on UHPLC-Q-Exactive Focus-MS allowed the evaluation of ripened Italian salami based on the impact of added axenic and mixed starter cultures. Becchi, Rocchetti, Vezzulli, Lambri, and Lucini (2023) recently used the same analytical approach to unravel the peculiarities of mountain grassland-based Parmigiano Reggiano PDO cheese production (Becchi et al., 2023). Overall, as carefully reviewed (Mialon, Roig, Capodanno, & Cadiere, 2023), those products most susceptible to fraud and authenticity issues are mainly olive oil, wine, honey, meat, fruits, and vegetables. They also report that the most important fraud is mislabeling, followed by adulteration and unapproved treatments or processes. Among the mislabeling practices, the most common are products wrongfully labeled as organic, from premium quality (e.g., extra virgin olive oil instead of lampant or virgin), or a wrong geographical origin or area (Mialon et al., 2023). As far as the geographical origin is concerned, the research studies reviewed suggest that a new trend is emerging in food authenticity dealing with the strong association between one or a class of metabolites with a certain geographical area, thus broadening the concept of terroir to other food products rather than the only wine (Lucini et al., 2020; Mialon et al., 2023; Roullier-Gall, Lucio, Noret, Schmitt-Kopplin, & Gougeon, 2014).

2.2.2. NMR metabolomics-based applications

NMR represents an advantageous and robust analytical technique when considering food quality and authenticity issues (Consonni & Cagliani, 2019; Selamat et al., 2021), capable of detecting different classes of metabolites in various food products, such as fermented foods, tomato paste, saffron, honey, dairy products (milk and cheese), meat and fermented meat, vinegar, and roasted coffee (among others). Some recent applications of NMR-based metabolomics as related to food authenticity and quality are reported in Table 1. Recently, the NMR approach was combined with near-infrared spectroscopy (NIR) to determine the geographical origin of hazelnuts (Shakiba et al., 2022). Also, ^{1}H NMR spectroscopy was successfully applied to identify chemical aging markers in green coffee (*Coffea arabica* L.) (Borém et al., 2023), while the same techniques allowed Cavallini and co-authors (2023) to trace the identity of Parmigiano Reggiano PDO cheese from the mountain area (Cavallini et al., 2023). ^{1}H NMR spectroscopy represents a valid alternative to traditional analytical methods, considering the possibility of simultaneously detecting several classes of chemical compounds. Moreover, quantitative determinations are also possible since the NMR signals are directly proportional to the number of resonating nuclei. As the next step, the data obtained by NMR could be processed appropriately by multivariate statistical tools, leading to disclosing sophisticated frauds, addressing the geographical origin and revealing potential biomarkers for authentication purposes (Kuballa, Brunner, Thongpanchang, Walch, & Lachenmeier, 2018; Rocchetti & O'Callaghan, 2021).

NMR offers the primary benefit of not requiring complicated sample preparation, with the possibility to determine very different chemical species using a single experiment (Wishart et al., 2022). NMR spectroscopy is also relatively easy; potential combination of compounds can be easily resolved, and almost every food metabolite can be detected from a quantitative standpoint (Selamat et al., 2021). However, the analysis is time-consuming, costly, and may have limited sensitivity (Sobolev et al., 2019). In this regard, ^{2}H (deuteron), ^{13}C, and ^{15}N are ideal metabolic tracers in detecting nuclei present at low abundance. Several strategies have been recently established to improve NMR sensitivity, such as those based on operative conditions at frequencies higher than 1.2 GHz using different cryoprobes (Schwalbe, 2017). These potential strategies increase the sensitivity of detection (mainly for low abundance analytes) while being cost-effective.

3. Chemometrics and computational learning: the next-generation toolkit for food metabolomics research

One of the major challenges of metabolomics workflows involves analysing and interpreting the vast array of acquired data, collectively known as chemometrics. This computational stage becomes a challenging task when dealing with complex and multivariate experimental designs, where a heterogeneity of factors is observed, which may interact in a complex manner to provide specific signatures at a metabolic level. In the fields of Food Science and Technology and Nutrition, the term "Foodomics" was coined to refer to the application of metabolomics and related technologies in the search for specific chemical signatures and markers associated with crucial features, such as food quality, traceability, safety and bioactivity, among others (Valdés et al., 2022). For these purposes, a wide range of bioinformatic tools are available based on either classical approaches deriving from fundamental statistics or cutting-edge technologies, mostly involving machine learning and deep learning algorithms. Interestingly, in the last years, the combined application of different analytical methodologies enabled the generation of highly significant multi-omics frameworks that provide simple and precise outcomes to shed light on the concealed knowledge that lies underneath the massive datasets obtained as a result of metabolomics workflows in combination with other high throughput technologies. This section covers an updated overview of the application of different chemometric approaches in the field of metabolomics facing the processing of the acquired data considering two learning perspectives: unsupervised learning, based on the naïve processing of data without previous assumptions, and supervised learning, based on the establishment of previous definitions and dataset labelling (X. Zhu & Goldberg, 2022). Table 2 shows an overview of the most widely applied chemometrics algorithms in metabolomics, as well as some of their major advantages and limitations.

3.1. Unsupervised chemometrics in food metabolomics: from naïvely grouping and pattern recognition to dimensionality reduction

Unsupervised learning covers a wide range of multivariate statistical approaches committed to the overall exploration of datasets for the qualitative recognition of patterns (Oliveri, Malegori, Mustorgi, & Casale, 2021), resulting in a suitable tool for the initial discovery of metabolome-wide fingerprints. Nevertheless, the absence of previous assumptions limits the applicability of unsupervised chemometrics, as no validation testing could be applied to the generated relationships between the data (Meza Ramirez, Greenop, Ashton, & Rehman, 2021). In practical terms, unsupervised chemometrics are typically employed to detect general trends, establish different groups or clusters, and identify outliers within metabolomics datasets (Pinto, 2017). In addition, unsupervised learning has proven to be efficient in reducing the complexity of data at either linear or multidimensional levels (Glielmo et al., 2021), conferring an increase in data management for downstream applications and facilitating their interpretation. This way, two groups of unsupervised learning methodologies have been mostly applied in food metabolomics, including clustering and dimensionality reduction methods. Fig. 2 includes an overview of the most relevant unsupervised algorithms.

3.1.1. Clustering

Clustering has emerged as a powerful tool to initially explore complex datasets, presenting countless applications in metabolomics, as it relies on multivariate and quantitative measurements to generate different groups according to the similarity of their metabolic profiles (Jaeger & Banks, 2023). Nevertheless, they show some limitations, as they provide limited information when dealing with multifactorial designs and they do not provide optimal outputs due to the same-level comparison between experimental groups. By far, the most extended clustering methodologies applied to metabolomics workflows are

Table 2
Advantages and limitations of the most widely applied chemometric algorithms in food metabolomics.

Learning typology	Algorithms	Examples	Advantages	Limitations
Unsupervised	Clustering	• HCA • k-means	1) Pattern recognition. 2) Initial dataset exploration. 3) Outlier identification.	1) Poor management of multifactorial designs. 2) Non-optimal outputs.
	Dimensionality reduction	• PCA • tSNE • NMF	1) Definition of linear functions. 2) Proximity-based distribution. 3) Dataset complexity reduction.	1) Require dimensionality optimization. 2) Loss of original overall variability. 3) Poor management of multifactorial designs.
Supervised	Machine learning	• PLS • OPLS • AMOPLS • SVM • KNN • RF	1) Predicted discriminating modeling. 2) Statistical variation. 3) Discriminant markers identification (VIP).	1) Random dataset splitting between training and test subsets. 2) VIP valorization in pairwise comparisons. 3) Require balanced designs. 4) Bad performance on high noise.
	Deep learning	• ANN • DNN	1) Robust predicting modeling. 2) Simplified model interpretation. 3) Response optimization.	1) Specific training for algorithm development. 2) Require *ad hoc* experimental designs. 3) Long modeling times and demanding informatic tools.

Abbreviations: HCA, hierarchical cluster analysis; PCA, principal component analysis; t-SNE, t-distributed stochastic neighbour embedding; NMF, non-negative matrix factorization; PLS, partial least squares; OPLS, orthogonal projection to latent structures; AMOPLS, ANOVA-multiblocking orthogonal projection to latent structures; SVM, support vector machines; KNN, k-nearest neighbours; VIP, variable importance in projection; RF, random forest; ANN, artificial neural networks; DNN, deep neural networks.

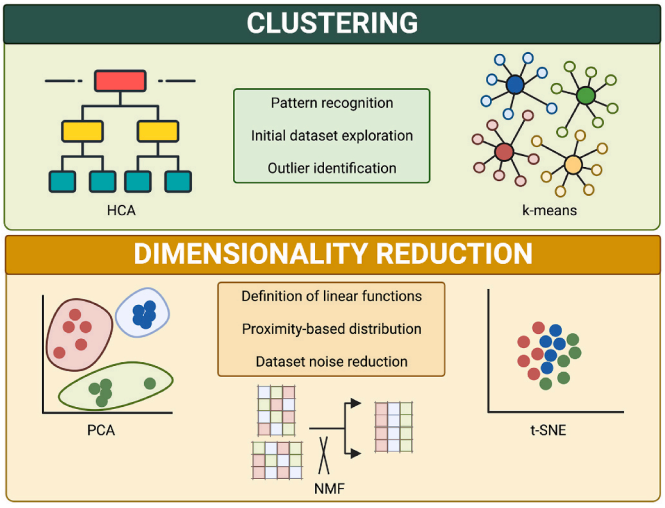


Fig. 2. Overview of the most widely applied unsupervised chemometrics algorithms in food metabolomics. HCA, hierarchical cluster analysis; PCA, principal component analysis; NMF, non-negative matrix factorization; t-SNE, t-distributed stochastic neighbour embedding.

hierarchical cluster analysis and k-means analysis.

Hierarchical cluster analysis (HCA) requires the application of an assembling algorithm to evaluate intra- and inter-group similarities and differences, usually visualized as a dendrogram that becomes a reliable resource to make decisions about the analyzed data space (Caesar, Kvalheim, & Cech, 2018). Due to its unsupervised nature, HCA is considered a multifaceted chemometric analysis for metabolomics, especially in the case of untargeted approaches, where no specific metabolites are known *a priori*, thus being useful to detect experimental artifacts or contaminants in unknown food matrices (Caesar et al., 2018). However, besides this advantageous application, HCA has been widely employed for multiple purposes in food metabolomics. Thus, Kotsanopoulos, Martsikalis, Gkafas, and Exadactylos (2022) recently reviewed the application of this clustering analysis regarding fish and seafood authenticity and adulteration, indicating that HCA was found efficient in detecting differences between fresh and frozen sea bass according to their MS spectral data, as well as differentiating three surimi products from their aromatic GC-MS profile (Kotsanopoulos et al., 2022). In the same way, HCA has been successfully applied to discern between thyme of different geographical origins and processing authentication based on their ^{1}H NMR profile (Rivera-Pérez, Romero-González, & Frenich, 2023) and LC-QTOF/MS phenolic profile (Rivera-Pérez, García-Pérez, Romero-González, Garrido Frenich, & Lucini, 2022). Following authentication, HCA was found effective in assessing the suitability of the LC-MS/MS profile to distinguish between halal and non-halal meat products (Windarsih et al., 2022). Moreover, HCA was also useful in providing added value to underexplored functional foods by their antioxidant profile, discriminating between novel wheat flour by-products (L. Zhang, García-Pérez, et al., 2023), edible flowers from different *Camellia japonica* cultivars (Pereira et al., 2023), and the supplementation of cherry laurel puree with other pectins (García-Pérez et al., 2024). Additionally, in the case of medicinal herbs, HCA assisted in the pattern recognition involving the authenticity of American ginseng based on its UHPLC-Orbitrap-MS phytochemical profile, showing a unique cluster including American and Canadian ginseng that was clearly separated from ginseng samples from different Chinese regions, this way guiding the subsequent supervised multivariate statistics approach to achieve the identification of metabolite markers responsible for that discrimination (Pang et al., 2023).

In turn, k-means is a non-hierarchical clustering algorithm that allows the unsupervised grouping of *k* non-overlapping clusters according to their similarity between variables (metabolites), displaying a significant contribution within datasets. In practical terms, k-means provides random clusters that present a center of gravity, from which associated metabolites arise with a certain distance, symbolising the weight of each metabolite related to this cluster (Toda, Goudo, Sugimoto, Hiwa, & Hiroyasu, 2023). This algorithm has been recently applied in metabolomics for food authentication and quality purposes. Thus, k-means shed light on the changes in the LC-MS/MS profile of Pu-erh tea during processing, revealing 72 significant non-volatile compounds (X. Deng, He, Han, & Chen, 2023), essentially represented by flavonoids. Moreover, k-means assisted in the authentication of white rice *via* GC-MS profiling, highlighting the influence of fatty acids and phospholipids in the geographical discrimination (Lim et al., 2018), whereas the same approach was applied to discriminate the Vietnamese and Chinese rice varieties from the Indian ones, the latter exhibiting higher amount of volatile organic compounds, such as aromatic aldehydes, terpenes, and ketones (Ch et al., 2021). In parallel, k-means was also applied to

evaluate the quality of beverages: 67 volatile compounds, predominantly esters, ketones, and alcohols, detected by GC-MS were accumulated after the lactic acid bacteria fermentation of kiwifruit juice (Lan et al., 2023); the untargeted LC-HRMS profile of cognac spirits was also subjected to k-means clustering, contributing to the discovery of new taste-active compounds (Winstel, Bahammou, Albertin, Waffo-Téguo, & Marchal, 2021); the targeted LC-MS/MS profile of red wine anthocyanins was found to discriminately accumulate during aging, featuring through k-means clustering pinotins and vitisin An as the metabolites driving this process (X.-K. Zhang, Lan, Huang, Zhao, & Duan, 2021). Concerning the metabolomics of dairy foods, k-means has been recently applied to depict that the LC-QQQ-MS/MS profile of sheep meat during aging is commanded by 407 metabolites, primarily amino acids and derivatives and fatty acids (M. Zhang et al., 2024).

3.1.2. Dimensionality reduction

Besides clustering, unsupervised algorithms may also be aimed at reducing the data dimensionality of highly complex acquired datasets, without compromising their interpretability (Eckhardt et al., 2023). The performance of these algorithms is, however, dependent on the optimization of dimesion reduction to assess the dimensionality simplification while maintaining the original dataset's variability. The most extended unsupervised methodology incorporated into metabolomics workflows is principal component analysis (PCA). However, some novel approaches with better performance for sample classification are gaining importance in recent years, dealing with high-dimensional datasets to provide a simple outcome that facilitates the downstream data analysis.

Principal component analysis (PCA) constitutes an unsupervised tool to convert high-dimensional data into linear features, explaining the total variance of the data space (Eckhardt et al., 2023). As a result, PCA modeling provides individual scatter plots reflecting the data variation observed according to the generated components, getting insight into the variables (metabolites) responsible for grouping samples through the generation of loading plots (Buvé et al., 2022). The ubiquitous application of PCA during metabolomics data processing has prompted a heterogeneous set of features that can be explored in food metabolomics, resulting in a versatile tool with countless purposes. For instance, PCA was useful in discriminating three pummelo varieties from a combined metabolomics approach based on LC-QTRAP-MS/MS and GC-MS profiles, covering 95.1% variability (C. Zhu et al., 2020), and pointing out the importance of aromatic metabolites. PCA determined that disaccharides, glucosinolates, and flavonoids ruled the discrimination between microgreens cultured under different agricultural techniques (Y. Li et al., 2023). Concerning milk quality and traceability, PCA is considered one of the most extended chemometrics approaches for a series of purposes, as reviewed earlier (Rocchetti & O'Callaghan, 2021). Moreover, PCA was proven useful in the valorization of edible flowers from different *Camellia japonica* cultivars, according to their polyphenolic profile (Pereira et al., 2023). In the same way, PCA was implicated in the authentication of ginger roots, covering 89.2% of the variance of both NMR and GC-MS profiles, pointing to a high level of fatty acids for authenticated ginseng. In contrast, the commercial products exhibited higher sugar content, suggesting an adulteration (Serag, Zayed, Mediani, & Farag, 2023). Furthermore, PCA efficiently discriminated the authenticy of natural medicines represented by turmeric and basil based on their Fourier-transformed NIR fingerprinting, showing unique metabolic signatures for Sangli and Erode turmeric led by curcumin and demethoxy-derivatives (covering 73.8% of the variance), as well as five different *Ocimum* species in the case of basil, accounting for the 67.9% of the variance (Khan et al., 2023).

In the last years, the bidimensional extent of PCA has limited its applicability when managing complex datasets, requiring novel algorithms to achieve dimensionality reduction. Among them, t-distributed stochastic neighbour embedding (t-SNE) has gained much attention as this non-linear algorithm converts high-dimensional data into low-dimensional points, rather than linear functions, to reproduce their proximity rather than distance (Spiwok & Kvrižˇ, 2020). This methodology has been earlier applied to transcriptomics facing single-cell signaling (Kobak & Berens, 2019), although some applications have been already involved in metabolomics, to date, mostly restricted to the clinical approaches devoted to intracellular dynamics imaging (Ebbels et al., 2023). In the same way, non-negative matrix factorization (NMF) emerged as a valuable tool for metabolomics, as it captures the global information of datasets through the sum of local data to preserve the global interpretation of data, reducing the overall noise (Xu et al., 2021). For instance, NMF was proved meaningful regarding peak aggregation by decreasing the co-linearity of non-discriminant metabolites implicated in classification and imputing missing values due to heterogeneous causes during the experimental stage of metabolomics workflows, thus improving the predictability of subsequent supervised algorithms (Fernandez-Albert, Llorach, Andres-Lacueva, & Perera-Lluna, 2014).

3.2. Supervised chemometrics in food metabolomics: classification algorithms to unleash the prediction power of the metabolome

Supervised chemometric approaches play a crucial role in metabolomics, assisting the initial data exploration provided by unsupervised methods. For this purpose, supervised algorithms can learn about the concealed trends and interactions found in multidimensional datasets to predict a target feature, showing a reliable outcome from data modeling supported by statistical validation (Biswas, Saran, & Wilson, 2021). Supervised modeling and validation require previous data knowledge, splitting datasets into the training set, used for model building, and the test set, employed for validating the performance of the obtained model (Pomyen et al., 2020). In recent years, supervised statistics have highlighted their learning concept thanks to the rising application of artificial intelligence based on machine learning algorithms. Besides their prediction ability, machine learning provides a robust technology contributing to data classification, becoming a convenient tool for solving problems and making specific decisions (Janiesch, Zschech, & Heinrich, 2021). These algorithms are normally combined with projection analyses, essentially known as variable importance in projection (VIP) analysis, to detect the variables playing the highest contribution on predictive modeling, facilitating the identification of markers, known as VIP markers, therefore providing a valuable outcome for food authentication and quality (Chen et al., 2022). Within machine learning, the application of neural networks has prompted the definition of deep learning, mainly applied in metabolomics for compound and structure identification and quantification rather than biological interpretations (Pomyen et al., 2020). Thus, supervised learning algorithms are classified in this section as machine learning and deep learning algorithms. Fig. 3 includes an overview of the most relevant supervised algorithms based on machine learning modeling.

3.2.1. Machine learning-based chemometrics

Machine learning (ML) applications have arisen exponentially in different research fields, including metabolomics, in recent years. The most extended objectives of ML-based chemometrics amid the available algorithms cover data regression and classification, exhibiting a valuable utility regarding food authenticity and quality: chemical markers identification (Galal, Talal, & Moustafa, 2022). Although the assistance of these algorithms have proven to be extremely useful in the field of metabolomics, they present some requirements to unleash all their performance, as they normally require balanced designs, they show a limited performance on high-noise datasets, and the value of chemical discriminant markers is optimal as a result of pairwise comparisons. In this sense, several regression analysis tools have been applied in food metabolomics, being considered an iteration of simpler algorithms, such as PCA, combined with linear and multidimensional regression to give rise to partial least squares (PLS), orthogonal projection to latent structures discriminant analysis (OPLS-DA), and the recently applied

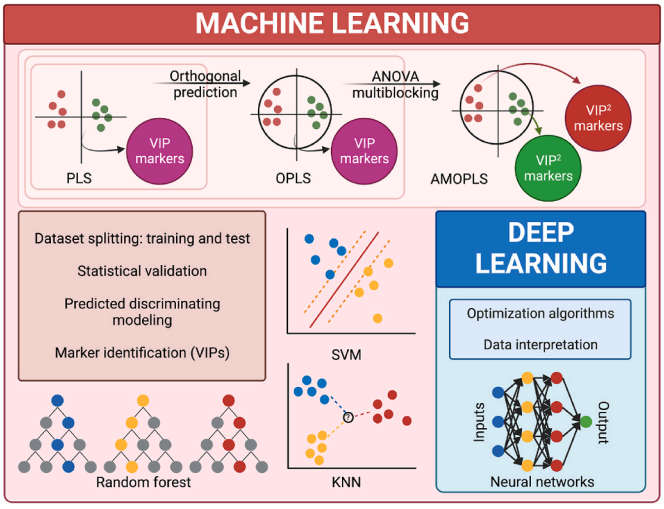


Fig. 3. Overview of the most widely applied machine learning-based supervised chemometrics algorithms in food metabolomics. PLS, partial least squares; OPLS, orthogonal projection to latent structures; AMOPLS, ANOVA-multiblocking orthogonal projection to latent structures; SVM, support vector machines; KNN, k-nearest neighbours.

ANOVA-multiblocking orthogonal partial least squares analysis (AMOPLS).

Firstly, PLS combines PCA with multiple linear regression to create latent vectors from the linear decomposition of both dependent and independent variables (for instance, categorical factors and metabolites, respectively) to predict the characterization of samples according to their metabolomic profile (Buvé et al., 2022). This algorithm was successfully applied to discriminate milk treated under different thermal processing temperatures, according to their UHPLC-Orbitrap-HRMS profile, with lysine derivatives described as major markers (Xie, Nie, Wang, Xu, & Zhang, 2024). Dinis and co-workers performed PLS analysis on the HILIC-Orbitrap-HRMS/MS profile of carrots to authenticate their geographical origin, highlighting the influence of 6-methoxymellein, arginine, and L-carnitine, as well as to discriminate their quality under conventional or organic agricultural production (Dinis et al., 2023). The application of PLS was found reliable in predictably differentiating adulterations in huajiao edible oil, mainly attributed to *Sn*-3 triacylglycerol, widely consumed in China, with 100% accuracy based on their ^{1}H NMR profile (Cui, Xia, et al., 2023). Interestingly, PLS assisted in predicting the increase in the antitumor activity of avocado seeds due to the changes in the GC-MS profile after fermentation and roasting, predicting the high contribution of amino acids and sugars to that property (Zhao et al., 2023). In the same way, the authentication of virgin coconut oil was achieved by PLS analysis *via* NMR-based metabolomics, as well as the potential adulteration with its refined counterpart, mainly affecting triglycerides and fatty acids composition (Bao, Tang, Rich, & Hatzakis, 2023). Concerning medicinal herbs, PLS assisted in the geographical origin determination of *Dioscorea nipponica* rhizome extracts, according to their LC-QTOF/MS profile, indicating that saponins and fatty acids were mainly responsible for the discrimination between different Chinese regions (Yang et al., 2023).

Based on PLS modeling, the orthogonal projection of datasets allows covering a more comprehensive range of the acquired information by metabolomics approaches, as it includes the separation of data variation into the correlated predictive variation (that is, the information provided by PLS), together with the uncorrelated variation, which represents the orthogonal component. The PLS model residuals can be used to employ the greatest proportion of the available information within the dataset without losing the predictive accuracy, resulting in an enhanced knowledge that facilitates interpretation (Pinto, 2017). Thanks to these advantages, OPLS-DA has become a milestone supervised chemometric approach devoted to food authentication and quality assessment, with countless applications available in the scientific literature. Thus, OPLS efficiently discriminated the application of different starter cultures on the UHPLC-Orbitrap-HRMS profile of ripened salami, featuring glutamyl peptides, biogenic amines, and lipid oxidized derivatives as responsible for guiding this discrimination, also impacting the sensory properties of this meat product (Rocchetti et al., 2023). Additionally, OPLS-DA was able to discriminate between the UHPLC-QTOF/MS profile of several wholegrain wheat flours from three genotypes containing differential amylose-amylopectin ratios, also impacting their nutritional outcomes, pointing at peptides, glycerophospholipids, and polyphenols as the main discriminant markers (Garcia-Perez et al., 2023). The impact of ripening on Parmigiano Reggiano cheese UHPLC-Orbitrap-HRMS profile was efficiently described by OPLS-DA, indicating the involvement of amino acids, peptides, fatty acids and conjugates, whereas polyphenols, terpenoids, glycerophospholipids and sterols were observed as responsible for authenticity by the discrimination between different cow feeding regimes (Becchi et al., 2023). The same analytical approach was combined with OPLS chemometric analysis for the quality assessment of novel functional foods, as is the case of pork meatballs enriched with blackcurrant seed extracts, which were found to possess an accumulation of terpenoids, polyphenols (especially anthocyanins), and oxo-derivatives of fatty acid hydroperoxides (Kerner et al., 2023). The UHPLC-QTRAP-MS/MS profile of black tea was subjected to OPLS modeling, revealing that aging played a positive role in its bioactive and sensory properties, motivated by the accumulation of specific discriminant markers, i.e.: ascorbic acid, salicylic acid, and benzoic acid, providing insight about the quality of this food matrix (S. Zhang, García-Pérez, et al., 2023). Recently, Rivera-Pérez and co-workers applied OPLS modeling as a meta-data mid-level fusion strategy to describe the geographical origin and processing authentication of thyme, merging the acquired datasets from different analytical approaches, i.e.: ^{1}H NMR, GC-HRMS, and UHPLC-QTOF/MS, providing a total of five highly discriminant metabolites ruling the thyme fingerprint, represented by thymol, choline, malic acid, chlorogenic acid, and α-glucose (Rivera-Pérez, Romero-González, & Frenich, 2023).

In parallel, the application of OPLS-DA was proven to be efficient in the authentication of phytochemical bioactive compounds from a wide range of extracts deriving from herbal preparations and medicinal plants. Thus, the VIP markers derived from the UHPLC-Orbitrap-HRMS profile of *Crocus sativus* flowers indicated that flavonoids and terpenoids were the most discriminant metabolites ruling the antibacterial and antioxidant properties of the derived extracts (Bellachioma, Marini, et al., 2022). Indeed, the phenolic profile of these extracts were further studied to determine the effect of infusion on the UHPLC-QTOF/MS phenolic profile of flowers, concluding that boiling affected not only phenolics but also the sensory and functional properties of *C. sativus* extracts (Bellachioma, Rocchetti, et al., 2022). OPLS modeling also contributed to the authenticity assessment of different *Lupinus* seeds used in the traditional medicine in Tunisia, reporting that the cultivar *Lupinus albus* exhibited a discriminant phenolic profile with respect to the wild accessions *Lupinus cossentinii* and *Lupinus luteus*, led by flavonoids and phenolic acids (Hassine et al., 2021). The quality assessment of *Taxus baccata* cell cultures treated with coronatine was also analyzed by OPLS, suggesting that this elicitor promoted the short-term accumulation of flavonoids and lignans that was further repressed with the cultivation time, whereas the accumulation of taxisuns and other terpenoids was enhanced over time, assisting in the productive optimization of these cultures (Perez-Matas, Garcia-Perez, Bonfill, et al., 2023; Perez-Matas, Garcia-Perez, Miras-Moreno, et al., 2023).

Overall, the effectiveness of OPLS-DA modeling has been largely assessed. Still, it presents some limitations, as the prediction ability is restricted to the previously defined experimental factors. The projection of markers is also dependent on these predefined assumptions, missing the ability to explore eventual interactions and the contribution of single factors to the predictive outcome during the interpretation of complex

and multifactorial designs. To overcome these limitations, Boccard and Rudaz (2016) explored the feasibility of OPLS supported by analysis of variance (ANOVA), providing a novel algorithm called ANOVA-multiblocking OPLS (AMOPLS). This algorithm is based on OPLS modeling, assisted by ANOVA-partitioned submatrices, providing a predictive component for each effect involved in the design of the experimental, thus indicating the particular influence of each factor and their interactions within the overall predicted discrimination with statistical support (Boccard & Rudaz, 2016). In practical terms, AMOPLS modeling provides a robust prediction associated with all the possible events observed along an experimental set, defining their discriminant contribution and enabling the extrapolation of the projected markers associated with each single effect and their interactions, known as VIP^2 markers (González-Ruiz et al., 2017). To the best of our knowledge, only a few authors have already employed this algorithm for food quality and authentication purposes, thus conferring a new highly significant chemometric platform. For instance, the terroir effect of wine was evaluated through AMOPLS predictive modeling, indicating that the single effect of the region, harvest stage and aging significantly impacted the discrimination of wine. In contrast, only the interaction region × harvest stage was found significant, providing a broad series of polyphenols and carotenoids attributed to such observed effects by developing a unique predictive model (Schmidtke et al., 2020). Equally, the effect of red wine composition during bottle aging was analyzed by AMOPLS modeling, showing an intricate significant interaction between different viticultural factors, such as vineyard location, associated with the dimethyl sulfide as the major discriminant VIP^2 marker; grape variety, related to 2-methylbutanal; and grape maturity, which was ruled mainly by the marker methanethiol (X. Zhang et al., 2020). Moreover, AMOPLS modeling was recently applied to decipher the influence of plant tissue and the exogenous application of naringenin and rosmarinic acid on the bioactive profile of the medicinal plant *Lepdidium sativum*, indicating that the exogenous treatment significantly interacted with plant tissues to modulate the phytochemical profile as well as the bioactivity of the derived herbal extracts (Salehi et al., 2023). Interestingly, it was assessed that flavonoids glycosides governed the discriminant effect attributed to tissues, as they were identified as the relative VIP^2 markers, whereas lignans were mostly associated to the effect of the treatment, additionally modulating the functional traits of extracts.

Besides regression algorithms, supervised learning techniques include a series of classification algorithms (also based on regression) that exhibit a wide range of predicting modeling approaches. Among them, support vector machines (SVM), k-nearest neighbours (KNN), and random forest (RF) algorithms have been mostly applied in the field of metabolomics. Thus, SVM relies on generating a hyperplane that separates samples from different classes, whose width depends on the number of covariables (Debik, Sangermani, Wang, Madssen, & Giskeødegård, 2022). Although it presents a high sensitivity to outliers, and therefore to model overfitting, it has been proven effective for prediction purposes (Liebal, Phan, Sudhakar, Raman, & Blank, 2020). In the same way, the KNN algorithm relies on establishing a local *k* significant feature that rules the establishment of different groups, classifying samples according to their distance from *k* (Yu et al., 2016). In contrast, RF relies on establishing decision trees to strictly classify metabolomics samples with human interpretable classification rules and the ability to extract significant features involved in the predicted classification (Erban et al., 2019). Massaro and co-workers demonstrated an SVM model to predictively discriminate the frozen and thawed European seabass and Atlantic salmon, according to their UHPLC-Orbitrap-MS/MS (Massaro et al., 2021). Concerning authenticity, SVM efficiently predicted the climate zones and geographical origin of mung beans with 100% accuracy based on their UHPLC-QTOF/MS profiles (He et al., 2023). In most cases, SVM, KNN, and RF algorithms are simultaneously applied to maximize the prediction ability of data modeling. For instance, the NMR profile of black tea from different geographical origins and industrial processing was subjected to machine learning classification, combining these three algorithms, providing the highest accuracy in the case of RF and SVM (>90% for both) with respect to KNN (86.3% accuracy) (Cui, Xia, et al., 2023). Moreover, RF could distinguish several markers associated with geographical origin, such as caffeine, malic acid, and lysine. Additionally, the UHPLC-QTOF/MS profile of two native Chinese cheery species was subjected to authenticity evaluation *via* SVM and RF, the latter exhibiting a 100% accuracy on the classification (Z. Wang et al., 2023). In parallel, in the case of food quality, SVM and RF were useful in discriminating orange juice from concentrate and not from concentrate through their UHPLC-QTOF/MS profiles, and the SVM model successfully discriminated 100% of commercial juices further subjected to the generated model.

3.2.2. *Deep learning-based chemometrics*

Deep learning-based techniques are specific types of Artificial Neural Networks (ANNs) algorithms, specifically falling under the broader category of supervised learning models (Srikanth Kumar, Lakshmana Rao, & Basaveswara Rao, 2018). To comprehend deep learning-based chemometric approaches and apply them to the food metabolomics field, a foundational understanding of ANNs is highly beneficial for this purpose.

ANNs are a machine learning technique that derives inspiration from the human nervous system and emulates the functional processes of the human brain when tackling problem-solving tasks. As illustrated in Fig. 4A, the architecture of a typical ANN comprises a collection of nodes referred to as neurons, organized into distinct layers, encompassing the "input layer", responsible for retaining input data information; the "hidden layer(s)", tasked with conducting data processing tasks; and the "output layer", responsible for delivering the final output information. The nodes of a layer are interconnected with their neighbour nodes and are associated with a specific weight value (Krogh, 2008).

The hidden layer is the location of mathematical operations that can process input data with their corresponding weights to generate output plus the bias. The mathematical operations executed by hidden neurons depend on the activation functions, which are the mathematical function that gives the linear or nonlinear regression property. Since the most challenging problems addressed by neural networks are inherently nonlinear and complex, the possibility of selecting the most suitable nonlinear function represents an added value to ANN algorithms. There are many activation functions available for a neural network to use; the most used are the sigmoid function, with output values defined in the range of (0,1); the hyperbolic tangent function, a scaled sigmoid function, defined in the range of values (−1, 1); the rectified linear unit (ReLU) function, the most used activation function, and the output is between 0 and infinity (Mukhametkhan, Krestinskaya, & James, 2018, pp. 245–249; Rasamoelina, Adjailia, & Sincak, 2020, pp. 281–286).

A deep Neural Network (DNN) is an artificial neural network characterized by a greater number of hidden layers, which develop the formation of increasingly abstract representations and extend their capacity to process complex and large amounts of labeled data, including omics data. They are typically used for complex tasks, including but not limited to image recognition, image classification, and handwriting identification (W. Liu et al., 2017). In the supervised DNN-based models, the training process consists of furnishing the DNNs with input data along with their corresponding target outputs, which is extremely important. This process facilitates the adjustment of weight values, aligning them more closely with the desired output, thus enhancing the network's ability to accurately approximate the expected results (Fig. 4B) (W. Liu et al., 2017). During each iteration cycle, known as an epoch, a series of feed-forward and backpropagation processes are executed to adjust weight values (Kanada, 2016, pp. 1472–1479). The extent to which weight adjustments occur is determined by the gradient descent method, which provides insight into how steeply the error will be reduced or increased for a change in the weight. The derived gradient descendent values at each step indicate the model's goodness of fit, predictability, and accuracy (Weng, Elsawah, & Fang, 2021). These

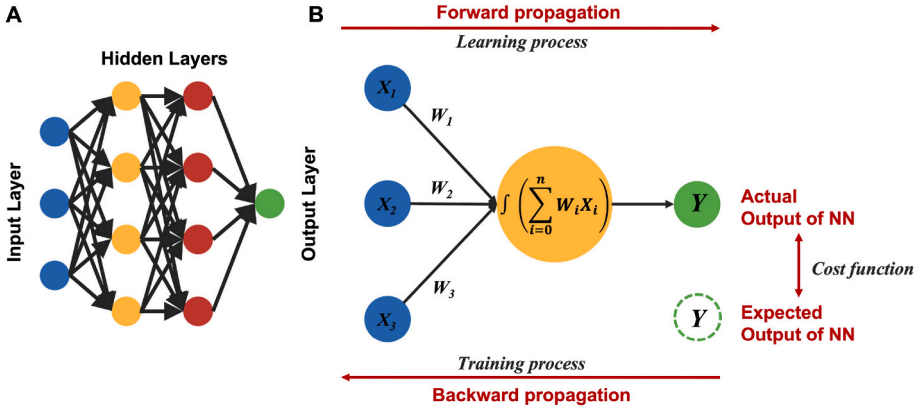


Fig. 4. (A) Architecture structure of an Artificial Neural Network (ANN). (B) Simple mathematical operations during the learning process of an ANN. In a standard neuron, the inputs x1, x2, and ×3 are connected to the hidden neuron with the corresponding synaptic weights w1, w2, and w3. The input information is multiplied with the corresponding weights to generate output.

latter are the basis of the learning process of so-called back-propagation neural network (BPNN) models.

Deep learning (DL) is increasingly important in several areas of life science, and only in recent years it has started to be employed in the agricultural and food sciences. Few examples are available in the field of food authenticity and adulteration, combining high-throughput metabolomics and deep-learning approaches. However, the application of this technology requires the trained individuals to optimize the computational workflow, whose workloads demand a significant amount of time, and are highly dependent on *ad hoc* experimental designs to unravel the maximal information available from the experimental data. To the best of our knowledge, Senizza et al. employed a combined approach of multilayer neural network and OPLS-DA to establish the authenticity of Taggiasca Ligure extra-virgin olive oil (EVOO), successfully discriminating both the geographical origin and variety of this Taggiasca Ligure EVOO (Senizza, Ganugi, Trevisan, & Lucini, 2023). The discrimination models are made using a comprehensive collection of phenolic and sterol metabolic profiles, which were determined using UHPLC-QTOF/MS as the analytical methodology. Another application of combining metabolomics and machine learning was to authenticate the geographic origin of Wuyi rock tea from two regions of China (Y. Peng et al., 2023). The authors used volatile metabolites derived by GC-TOF-MS analytical instruments to produce a series of ML algorithms, including a multilayer perceptron neural network, which resulted in the best performance with an average accuracy of 92.7% on the training data using 176 volatile features. Another example of using volatile compounds was reported by Colantonio et al. (2022), which used a targeted metabolomics approach to quantify sugars, acids, and volatiles in tomatoes and blueberries (*Vaccinium* spp.) sourced from various accessions (Colantonio et al., 2022). The authors collected these data to construct a statistical and machine-learning model capable of predicting the sensory perceptions of fruit flavor from a consumer-driven decision-making perspective.

An innovative application of combined approaches lies in characterizing medicinal plants as biofactories, particularly in the food, cosmetic, and pharmaceutical sectors, as reported by García-Pérez and co-workers (2021 and 2022) (García-Pérez et al., 2021a, 2022). In these comprehensive studies, the authors deciphered the critical factors responsible for the biosynthesis of bioactive compounds in the cell suspension cultures from medicinal plants belonging to the *Bryophyllum* subgenus. These cultures were subjected to various conditions, including elicitation with salicylic acid and/or methyl jasmonate elicitors and alterations in macro- and micro-nutrient solutions within the culture media. The authors employed a neuro-fuzzy logic machine learning approach, utilizing a comprehensive metabolic profile of *Bryophyllum* extract cultivated under specific experimental conditions. This methodology allowed for the elucidation of critical factors contributing to the production of bioactive compounds in the context of medicinal plant biofactories, as well as enabled the authenticity assessment of different species according to the UHPLC-QTOF/MS profile of the extracts derived from these medicinal plants, contributing to the optimization of the phytochemical production of health-promoting compounds (García-Pérez et al., 2021a, 2022).

4. Conclusions and future outlook

Metabolomics is a young science, compared to other omics. The advance in metabolomics equipments, in available databases, in software solutions, and the growing consensus on standards in reporting has paved the way towards the use of metabolomics in several fields, including food science. Nonetheless, the complexity of the multi-layered interacting factors that finally determine food quality and integrity requires challenging data management approaches. As a consequence, appropriate chemometric approaches become essential to harvest all the useful informations hidden in metabolomics datasets. Our review highlights as there is not a golden approach, and specific approaches tailored to specific needs are required. From one side, unsupervised approaches are important to naively provide distribution patterns across samples and treatments in a qualitative manner. On the contrary, supervised approaches are better able to model one or few parameters under study, thus highlighting specific markers of authenticity, origin, or quality. A third option is represented by machine learning approaches, a novel solution that has been recently proposed to interpretate metabolomics data in the context of food science. In this sense, the combined application of unsupervised statistics together with supervised and machine learning-based algorithms are considered as a robust analytical toolkit to unleash the whole potential of metabolomics in the field of food authentication and quality assessment. Interestingly, both machine learning and deep learning algorithms represent a valuable upgrade to cope with other critical aspects in the field of Foodomics thanks to their prediction ability, thus avoiding resource- and time-consuming experimental workflows. As a result, they provide a priceless platform for the discovery of concealed trends within highly complex studies, facilitating the current challenges of food systems, including food security, nutrition, and sustainability assessment.

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.

Acknowledgements

The work by P.G.-P. was financed by the Spanish Ministry of Universities under the application 33.50.460 A.752 and by the European Union NextGenerationEU/PRTR through a Margarita Salas contract by the Universidade de Vigo.

References

Aderemi, A. V., Ayeleso, A. O., Oyedapo, O. O., & Mukwevho, E. (2021). Metabolomics: A scoping review of its role as a tool for disease biomarker discovery in selected non-communicable diseases. *Metabolites, 11*(7). https://doi.org/10.3390/metabo11070418

Asharindavida, F., Nibouche, O., Uhomoibhi, J., Liu, J., Vincent, J., & Wang, H. (2023). Miniature spectrometer data analytics for food fraud. *Journal Fur Verbraucherschutz Und Lebensmittelsicherheit.* , Article 0123456789. https://doi.org/10.1007/s00003-023-01439-8

Bambina, P., Spinella, A., Lo Papa, G., Chillura Martino, D. F., Lo Meo, P., Corona, O., et al. (2023). 1H NMR-based metabolomics to assess the impact of soil type on the chemical composition of nero d'Avola red wines. *Journal of Agricultural and Food Chemistry, 71*(14), 5823–5835. https://doi.org/10.1021/acs.jafc.2c08654

Bao, R., Tang, F., Rich, C., & Hatzakis, E. (2023). A comparative evaluation of low-field and high-field NMR untargeted analysis: Authentication of virgin coconut oil adulterated with refined coconut oil as a case study. *Analytica Chimica Acta, 341537*.

Beale, D. J., Kouremenos, K. A., & Palombo, E. A. (2016). Microbial metabolomics: Applications in clinical, environmental, and industrial microbiology. *Microbial Metabolomics: Applications in Clinical, Environmental, and Industrial Microbiology*, 1–321. https://doi.org/10.1007/978-3-319-46326-1

Becchi, P. P., Rocchetti, G., Vezzulli, F., Lambri, M., & Lucini, L. (2023). The integrated metabolomics and sensory analyses unravel the peculiarities of mountain grassland-based cheese production: The case of Parmigiano Reggiano PDO. *Food Chemistry, 428* (June), Article 136803. https://doi.org/10.1016/j.foodchem.2023.136803

Bellachioma, L., Marini, E., Magi, G., Pugnaloni, A., Facinelli, B., Rocchetti, G., et al. (2022). Phytochemical profiling, antibacterial and antioxidant properties of Crocus sativus flower: A comparison between tepals and stigmas. *Open Chemistry, 20*(1), 431–443. https://doi.org/10.1515/chem-2022-0155

Bellachioma, L., Rocchetti, G., Morresi, C., Martinelli, E., Lucini, L., Ferretti, G., et al. (2022). Valorisation of Crocus sativus flower parts for herbal infusions: Impact of brewing conditions on phenolic profiling, antioxidant capacity and sensory traits. *International Journal of Food Science and Technology, 57*(6), 3838–3849. https://doi.org/10.1111/ijfs.15713

Beteinakis, S., Papachristodoulou, A., Kolb, P., Rösch, P., Schwarzinger, S., Mikros, E., et al. (2023). NMR-based metabolite profiling and the application of STOCSY toward the quality and authentication assessment of European EVOOs. *Molecules, 28*(4). https://doi.org/10.3390/molecules28041738

Bischof, G., Witte, F., Terjung, N., Januschewski, E., Heinz, V., Juadjur, A., et al. (2023). NMR-based comparison of the metabolome of beef from Simmental and black-and-white young bulls during wet- and dry-aging. *European Food Research and Technology, 249*(8), 2113–2124. https://doi.org/10.1007/s00217-023-04283-0

Biswas, A., Saran, I., & Wilson, F. P. (2021). Introduction to supervised machine learning. *Kidney, 2*(5), 878.

Bittremieux, W., Wang, M., & Dorrestein, P. C. (2022). The critical role that spectral libraries play in capturing the metabolomics community knowledge. *Metabolomics, 18*(12), 1–16. https://doi.org/10.1007/s11306-022-01947-y

Blaženović, I., Kind, T., Ji, J., & Fiehn, O. (2018). Software tools and approaches for compound identification of LC-MS/MS data in metabolomics. *Metabolites, 8*(2). https://doi.org/10.3390/metabo8020031

Boccard, J., & Rudaz, S. (2016). Exploring Omics data from designed experiments using analysis of variance multiblock Orthogonal Partial Least Squares. *Analytica Chimica Acta, 920*, 18–28. https://doi.org/10.1016/j.aca.2016.03.042

Borém, F. M., Abreu, G. F. de, Ferreira, A. G., Santos, M. da S., Alves, T. da C., & Alves, A. P. de C. (2023). 1H NMR spectroscopy applied to identify chemical aging markers in green coffee (Coffea arabica L.). *Food Chemistry, 405*(April 2022), 1–11. https://doi.org/10.1016/j.foodchem.2022.134667

Buvé, C., Saeys, W., Rasmussen, M. A., Neckebroeck, B., Hendrickx, M., Grauwet, T., et al. (2022). Application of multivariate data analysis for food quality investigations: An example-based review. *Food Research International, 151*, Article 110878.

Caesar, L. K., Kvalheim, O. M., & Cech, N. B. (2018). Hierarchical cluster analysis of technical replicates to identify interferents in untargeted mass spectrometry metabolomics. *Analytica Chimica Acta, 1021*, 69–77.

Cao, G., Song, Z., Hong, Y., Yang, Z., Song, Y., Chen, Z., et al. (2020). Large-scale targeted metabolomics method for metabolite profiling of human samples. *Analytica Chimica Acta, 1125*, 144–151. https://doi.org/10.1016/j.aca.2020.05.053

Cavallini, N., Strani, L., Becchi, P. P., Pizzamiglio, V., Michelini, S., & Savorani, F. (2023). Analytica Chimica Acta Tracing the identity of Parmigiano Reggiano " Prodotto di Montagna - Progetto Territorio " cheese using NMR spectroscopy and multivariate data analysis. *Analytica Chimica Acta, 1278*(April), Article 341761. https://doi.org/10.1016/j.aca.2023.341761

Ćeranić, A., Bueschl, C., Doppler, M., Parich, A., Xu, K., Lemmens, M., et al. (2020). Enhanced metabolome coverage and evaluation of matrix effects by the use of experimental-condition-matched 13C-labeled biological samples in isotope-assisted LC-HRMS metabolomics. *Metabolites, 10*(11), 434.

others Ch, R., Chevallier, O., McCarron, P., McGrath, T. F., Wu, D., Kapil, A. P., et al. (2021). Metabolomic fingerprinting of volatile organic compounds for the geographical discrimination of rice samples from China, Vietnam and India. *Food Chemistry, 334*, Article 127553.

Chen, Y., Li, E. M., & Xu, L. Y. (2022). Guide to metabolomics analysis: A bioinformatics workflow. *Metabolites, 12*(4). https://doi.org/10.3390/metabo12040357

Cheng, H., Mei, J., & Xie, J. (2023). The odor deterioration of tilapia (Oreochromis mossambicus) fillets during cold storage revealed by LC-MS based metabolomics and HS-SPME-GC-MS analysis. *Food Chemistry*, Article 136699.

Chowdhury, C. R., Kavitake, D., Jaiswal, K. K., Jaiswal, K. S., Reddy, G. B., Agarwal, V., et al. (2023). NMR-based metabolomics as a significant tool for human nutritional research and health applications. *Food Bioscience, 53*(January), Article 102538. https://doi.org/10.1016/j.fbio.2023.102538

Christmann, J., Rohn, S., & Weller, P. (2022). Finding features - variable extraction strategies for dimensionality reduction and marker compounds identification in GC-IMS data. *Food Research International, 161*(June), Article 111779. https://doi.org/10.1016/j.foodres.2022.111779

Colantonio, V., Ferrão, L. F. V., Tieman, D. M., Bliznyuk, N., Sims, C., Klee, H. J., et al. (2022). Metabolomic selection for enhanced fruit flavor. *Proceedings of the National Academy of Sciences of the United States of America, 119*(7). https://doi.org/10.1073/pnas.2115865119

Consonni, R., & Cagliani, L. R. (2019). The potentiality of NMR-based metabolomics in food science and food authentication assessment. *Magnetic Resonance in Chemistry, 57*(9), 558–578.

Crupi, P., Genghi, R., & Antonacci, D. (2014). In-time and in-space tandem mass spectrometry to determine the metabolic profiling of flavonoids in a typical sweet cherry (Prunus avium L.) cultivar from Southern Italy. *Journal of Mass Spectrometry, 49*(10), 1025–1034.

Cubero-Leon, E., Peñalver, R., & Maquet, A. (2014). Review on metabolomics for food authentication. *Food Research International, 60*, 95–107. https://doi.org/10.1016/j.foodres.2013.11.041

Cui, C., Xia, M., Chen, J., Shi, B., Peng, C., Cai, H., et al. (2023). 1H NMR-based metabolomics combined with chemometrics to detect edible oil adulteration in huajiao (Zanthoxylum bungeanum Maxim.). *Food Chemistry, 423*, Article 136305.

Cui, C., Xu, Y., Jin, G., Zong, J., Peng, C., Cai, H., et al. (2023). Machine learning applications for identify the geographical origin, variety and processing of black tea using 1H NMR chemical fingerprinting. *Food Control, 148*(November 2022), Article 109686. https://doi.org/10.1016/j.foodcont.2023.109686

Danezis, G. P., Tsagkaris, A. S., Brusic, V., & Georgiou, C. A. (2016). Food authentication: State of the art and prospects. *Current Opinion in Food Science, 10*, 22–31. https://doi.org/10.1016/j.cofs.2016.07.003

Dasenaki, M. E., Drakopoulou, S. K., Aalizadeh, R., & Thomaidis, N. S. (2019). Targeted and untargeted metabolomics as an enhanced tool for the detection of pomegranate juice adulteration. *Foods, 8*(6), 212.

Daygon, V. D., Prakash, S., Calingacion, M., Riedel, A., Ovenden, B., Snell, P., et al. (2016). Understanding the Jasmine phenotype of rice through metabolite profiling and sensory evaluation. *Metabolomics, 12*(4), 1–15. https://doi.org/10.1007/s11306-016-0989-6

Debik, J., Sangermani, M., Wang, F., Madssen, T. S., & Giskeødegård, G. F. (2022). Multivariate analysis of NMR-based metabolomic data. *NMR in Biomedicine, 35*(2), 1–21. https://doi.org/10.1002/nbm.4638

Deng, X., He, S., Han, Y., & Chen, Y. (2023). Metabolites profiling reveals the dynamic changes of non-volatiles in Pu-erh during Ganpu tea processing. *Food Chemistry, X*, Article 100774.

Deng, P., Li, X., Petriello, M. C., Wang, C., Morris, A. J., & Hennig, B. (2019). Application of metabolomics to characterize environmental pollutant toxicity and disease risks. *Reviews on Environmental Health*. https://doi.org/10.1515/reveh-2019-0030

Di Minno, A., Gelzo, M., Stornaiuolo, M., Ruoppolo, M., & Castaldo, G. (2021). The evolving landscape of untargeted metabolomics. *Nutrition, Metabolism, and Cardiovascular Diseases, 31*(6), 1645–1652. https://doi.org/10.1016/j.numecd.2021.01.008

Diez-Simon, C., Mumm, R., & Hall, R. D. (2019). Mass spectrometry-based metabolomics of volatiles as a new tool for understanding aroma and flavour chemistry in processed food products. *Metabolomics, 15*(3), 1–20. https://doi.org/10.1007/s11306-019-1493-6

Dinis, K., Tsamba, L., Jamin, E., & Camel, V. (2023). Untargeted metabolomics-based approach using UHPLC-HRMS to authenticate carrots (Daucus carota L.) based on geographical origin and production mode. *Food Chemistry, 423*(April). https://doi.org/10.1016/j.foodchem.2023.136273

Ebbels, T. M. D., van der Hooft, J. J. J., Chatelaine, H., Broeckling, C., Zamboni, N., Hassoun, S., et al. (2023). Recent advances in mass spectrometry-based computational metabolomics. *Current Opinion in Chemical Biology, 74*, Article 102288.

Eckhardt, C. M., Madjarova, S. J., Williams, R. J., Ollivier, M., Karlsson, J., Pareek, A., et al. (2023). Unsupervised machine learning methods and emerging applications in healthcare. *Knee Surgery, Sports Traumatology, Arthroscopy, 31*(2), 376–381.

Erban, A., Fehrle, I., Martinez-Seidel, F., Brigante, F., Más, A. L., Baroni, V., et al. (2019). Discovery of food identity markers by metabolomics and machine learning technology. *Scientific Reports, 9*(1), 1–19. https://doi.org/10.1038/s41598-019-46113-y

Fernandez-Albert, F., Llorach, R., Andres-Lacueva, C., & Perera-Lluna, A. (2014). Peak aggregation as an innovative strategy for improving the predictive power of LC-MS metabolomic profiles. *Analytical Chemistry, 86*(5), 2320–2325.

Fiehn, O. (2016). Metabolomics by gas chromatography–mass spectrometry: Combined targeted and untargeted profiling. *Current Protocols in Molecular Biology, 114*(1), 30–34.

Galal, A., Talal, M., & Moustafa, A. (2022). Applications of machine learning in metabolomics: Disease modeling and classification. *Frontiers in Genetics, 13*, Article 1017340.

Garcia-Perez, P., Cassani, L., Garcia-Oliveira, P., Xiao, J., Simal-Gandara, J., Prieto, M. A., et al. (2023). Algal nutraceuticals: A perspective on metabolic diversity, current food applications, and prospects in the field of metabolomics. *Food Chemistry, 409*(December 2022), Article 135295. https://doi.org/10.1016/j.foodchem.2022.135295

García-Pérez, P., Giuberti, G., Sestili, F., Lafiandra, D., Botticella, E., & Lucini, L. (2023). The functional implications of high-amylose wholegrain wheat flours: An in vitro digestion and fermentation approach combined with metabolomics. *Food Chemistry, 135959*. https://doi.org/10.1016/j.foodchem.2023.135959

García-Pérez, P., Lozano-Milo, E., Zhang, L., Miras-Moreno, B., Landin, M., Lucini, L., et al. (2022). Neurofuzzy logic predicts a fine-tuning metabolic reprogramming on elicited *Bryophyllum* PCSCs guided by salicylic acid. *Frontiers in Plant Science, 13*, Article 991557. https://doi.org/10.3389/fpls.2022.991557

García-Pérez, P., Tomas, M., Rivera-Pérez, A., Patrone, V., Giuberti, G., Cervini, M., et al. (2024). Pectin conformation influences the bioaccessibility of cherry laurel polyphenols and gut microbiota distribution following in vitro gastrointestinal digestion and fermentation. *Food Chemistry, 430*, Article 137054.

García-Pérez, P., Zhang, L., Miras-Moreno, B., Lozano-Milo, E., Landin, M., Lucini, L., et al. (2021a). The combination of untargeted metabolomics and machine learning predicts the biosynthesis of phenolic compounds in Bryophyllum medicinal plants (Genus *Kalanchoe*). *Plants, 10*(11), 1–23. https://doi.org/10.3390/plants10112430

García-Pérez, P., Zhang, L., Miras-Moreno, B., Lozano-Milo, E., Landin, M., Lucini, L., et al. (2021b). The combination of untargeted metabolomics and machine learning predicts the biosynthesis of phenolic compounds in *Bryophyllum* medicinal plants (Genus *Kalanchoe*). *Plants, 10*(11), 2430. https://doi.org/10.3390/plants10112430

Gertsman, I., & Barshop, B. A. (2018). Promises and pitfalls of untargeted metabolomics. *Journal of Inherited Metabolic Disease, 41*(3), 355–366. https://doi.org/10.1007/s10545-017-0130-7

Girelli, C. R., Papadia, P., Pagano, F., Miglietta, P. P., Cardinale, M., & Rustioni, L. (2023). Metabolomic NMR analysis and organoleptic perceptions of pomegranate wines: Influence of cultivar and yeast on the product characteristics. *Heliyon, 9*(6), Article e16774. https://doi.org/10.1016/j.heliyon.2023.e16774

Glielmo, A., Husic, B. E., Rodriguez, A., Clementi, C., Noé, F., & Laio, A. (2021). Unsupervised learning methods for molecular simulation data. *Chemical Reviews, 121*(16), 9722–9758. https://doi.org/10.1021/acs.chemrev.0c01195

González-Ruiz, V., Pezzatti, J., Roux, A., Stoppini, L., Boccard, J., & Rudaz, S. (2017). Unravelling the effects of multiple experimental factors in metabolomics, analysis of human neural cells with hydrophilic interaction liquid chromatography hyphenated to high resolution mass spectrometry. *Journal of Chromatography A, 1527*, 53–60. https://doi.org/10.1016/j.chroma.2017.10.055

Hassine, A. B., Rocchetti, G., Zhang, L., Senizza, B., Zengin, G., Mahomoodally, M. F., et al. (2021). Untargeted phytochemical profile, antioxidant capacity and enzyme inhibitory activity of cultivated and wild lupin seeds from Tunisia. *Molecules, 26*(11), 1–14. https://doi.org/10.3390/molecules26113452

He, L., Hu, Q., Yu, Y., Yu, Y., Yu, N., & Chen, Y. (2023). Discrimination of mung beans according to climate and growing region by untargeted metabolomics coupled with machine learning methods. *Food Control, 153*(June), Article 109927. https://doi.org/10.1016/j.foodcont.2023.109927

Jaeger, A., & Banks, D. (2023). Cluster analysis: A modern statistical review. *Wiley Interdisciplinary Reviews: Computational Statistics, 15*(3), Article e1597.

Janiesch, C., Zschech, P., & Heinrich, K. (2021). Machine learning and deep learning. *Electronic Markets, 31*(3), 685–695.

Kanada, Y. (2016). Optimizing neural-network learning rate by using a genetic algorithm with per-epoch mutations. *Proceedings of the international joint conference on neural networks, 2016-octob*. https://doi.org/10.1109/IJCNN.2016.7727372

Kerner, K., Kazernavičiūtė, R., Jõudu, I., Rocchetti, G., Lucini, L., Tänavots, A., et al. (2023). Evaluation of different blackcurrant seed ingredients in meatballs by using conventional quality assessment and untargeted metabolomics. *Meat Science, 200*(December 2022). https://doi.org/10.1016/j.meatsci.2023.109160

Khan, S., Rai, A. K., Singh, A., Singh, S., Dubey, B. K., Lal, R. K., et al. (2023). Rapid metabolic fingerprinting with the aid of chemometric models to identify authenticity of natural medicines: Turmeric, Ocimum, and Withania somnifera study. *Journal of Pharmaceutical Analysis, 13*(9), 1041–1057. https://doi.org/10.1016/j.jpha.2023.04.018

Klare, J., Rurik, M., Rottmann, E., Bollen, A., Kohlbacher, O., Fischer, M., et al. (2020). Determination of the geographical origin of Asparagus officinalis L. By 1H NMR spectroscopy. *Journal of Agricultural and Food Chemistry, 68*(49), 14353–14363. https://doi.org/10.1021/acs.jafc.0c05642

Kobak, D., & Berens, P. (2019). The art of using t-SNE for single-cell transcriptomics. *Nature Communications, 10*(1), 5416.

Kotsanopoulos, K., Martsikalis, P. V., Gkafas, G. A., & Exadactylos, A. (2022). The use of various statistical methods for authenticity and detection of adulteration in fish and seafood. *Critical Reviews in Food Science and Nutrition*, 1–19.

Krogh, A. (2008). What are artificial neural networks? *Nature Biotechnology, 26*(2), 195–197. https://doi.org/10.1038/nbt1386

Kuballa, T., Brunner, T. S., Thongpanchang, T., Walch, S. G., & Lachenmeier, D. W. (2018). Application of NMR for authentication of honey, beer and spices. *Current Opinion in Food Science, 19*, 57–62. https://doi.org/10.1016/j.cofs.2018.01.007

Lamine, M., Hamdi, Z., Rahali, F. Z., Mliki, A., Varzakas, T., & Gargouri, M. (2023). Non-targeting metabolite profiling and chemometric approaches for the discrimination and authentication analyses of whole-wheat flours from Tunisian durum wheat landraces (Triticum turgidum ssp. durum). *Journal of Food Measurement and Characterization, 17*(2), 1889–1898. https://doi.org/10.1007/s11694-022-01759-7

Lan, T., Lv, X., Zhao, Q., Lei, Y., Gao, C., Yuan, Q., et al. (2023). Optimization of strains for fermentation of kiwifruit juice and effects of mono-and mixed culture fermentation on its sensory and aroma profiles. *Food Chemistry X, 17*, Article 100595.

Li, S., Dunlop, A. L., Jones, D. P., & Corwin, E. J. (2016). High-resolution metabolomics: Review of the field and implications for nursing science and the study of preterm birth. *Biological Research For Nursing, 18*(1), 12–22.

Li, Z., Tan, M., Deng, H., Yang, X., Yu, Y., Zhou, D., et al. (2022). Geographical origin differentiation of rice by LC–MS-based non-targeted metabolomics. *Foods, 11*(21), 1–11. https://doi.org/10.3390/foods11213318

Li, S., Tian, Y., Jiang, P., Lin, Y., Liu, X., & Yang, H. (2021). Recent advances in the application of metabolomics for food safety control and food quality analyses. *Critical Reviews in Food Science and Nutrition, 61*(9), 1448–1469. https://doi.org/10.1080/10408398.2020.1761287

Li, Y., Zhou, B., Teng, Z., Zhang, M., Yu, L., Luo, Y., et al. (2023). Improved metabolomic approach for evaluation of phytochemicals in mustard, kale, and broccoli microgreens under different controlled environment agriculture conditions. *Journal of Agriculture and Food Research, 14*, Article 100719.

Liebal, U. W., Phan, A. N. T., Sudhakar, M., Raman, K., & Blank, L. M. (2020). Machine learning applications for mass spectrometry-based metabolomics. *Metabolites, 10*(6), 1–23. https://doi.org/10.3390/metabo10060243

Lim, D. K., Mo, C., Lee, J. H., Long, N. P., Dong, Z., Li, J., et al. (2018). The integration of multi-platform MS-based metabolomics and multivariate analysis for the geographical origin discrimination of Oryza sativa L. *Journal of Food and Drug Analysis, 26*(2), 769–777.

Lima, B. D., Martins, L. L., Pereira, V. B., Franco, D. M. M., dos Santos, I. R., Santos, J. M., et al. (2023). Weathering impacts on petroleum biomarker, aromatic, and polar compounds in the spilled oil at the northeast coast of Brazil over time. *Marine Pollution Bulletin, 189*(November 2022). https://doi.org/10.1016/j.marpolbul.2023.114744

Lioupi, A., Marinaki, M., Virgiliou, C., Begou, O., Gika, H., Wilson, I., et al. (2023). Probing the polar metabolome by UHPLC-MS. *TrAC, Trends in Analytical Chemistry, 161*, Article 117014. https://doi.org/10.1016/j.trac.2023.117014

Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y., & Alsaadi, F. E. (2017). A survey of deep neural network architectures and their applications. *Neurocomputing, 234*(October 2016), 11–26. https://doi.org/10.1016/j.neucom.2016.12.038

Liu, J., Zhou, X., Chen, D., Guo, J., Chen, K., Ye, C., et al. (2023). 1H NMR-based metabolic profiling to follow changes in pomelo cultivars during postharvest senescence. *Foods, 12*, 2001. https://doi.org/10.3390/foods12102001

Lucini, L., Rocchetti, G., & Trevisan, M. (2020). Extending the concept of terroir from grapes to other agricultural commodities: An overview. *Current Opinion in Food Science, 31*, 88–95. https://doi.org/10.1016/j.cofs.2020.03.007

Lyu, W., Yuan, B., Liu, S., Simon, J. E., & Wu, Q. (2022). Assessment of lemon juice adulteration by targeted screening using LC-UV-MS and untargeted screening using UHPLC-QTOF/MS with machine learning. *Food Chemistry, 373*(PA), Article 131424. https://doi.org/10.1016/j.foodchem.2021.131424

Marchev, A. S., Vasileva, L. V., Amirova, K. M., Savova, M. S., Balcheva-Sivenova, Z. P., & Georgiev, M. I. (2021). Metabolomics and health: From nutritional crops and plant-based pharmaceuticals to profiling of human biofluids. *Cellular and Molecular Life Sciences, 78*(19–20), 6487–6503. https://doi.org/10.1007/s00018-021-03918-3

Massaro, A., Stella, R., Negro, A., Bragolusi, M., Miano, B., Arcangeli, G., et al. (2021). New strategies for the differentiation of fresh and frozen/thawed fish: A rapid and accurate non-targeted method by ambient mass spectrometry and data fusion (part A). *Food Control, 130*(March), Article 108364. https://doi.org/10.1016/j.foodcont.2021.108364

Mattoo, A. K., Cavigelli, M. A., Mišić, D. M., Gašić, U., Maksimović, V. M., Kramer, M., et al. (2023). Maize metabolomics in relation to cropping system and growing year. *Frontiers in Sustainable Food Systems, 7*. https://doi.org/10.3389/fsufs.2023.1130089

Medina, S., Perestrelo, R., Silva, P., Pereira, J. A. M., & Câmara, J. S. (2019). Current trends and recent advances on food authenticity technologies and chemometric approaches. *Trends in Food Science and Technology, 85*(January), 163–176. https://doi.org/10.1016/j.tifs.2019.01.017

Meza Ramirez, C. A., Greenop, M., Ashton, L., & Rehman, I. ur (2021). Applications of machine learning in spectroscopy. *Applied Spectroscopy Reviews, 56*(8–10), 733–763. https://doi.org/10.1080/05704928.2020.1859525

Mialon, N., Roig, B., Capodanno, E., & Cadiere, A. (2023). Untargeted metabolomic approaches in food authenticity: A review that showcases biomarkers. *Food Chemistry, 398*(August 2022), Article 133856. https://doi.org/10.1016/j.foodchem.2022.133856

Mukhametkhan, M., Krestinskaya, O., & James, A. P. (2018). Notice of retraction: Analysis of multilayer perceptron with rectifier linear unit activation function. *Proceedings of the 2nd international conference on computing and network communications*. https://doi.org/10.1109/CoCoNet.2018.8476902. *CoCoNet 2018, May*.

Mushtaq, M. Y., Choi, Y. H., Verpoorte, R., & Wilson, E. G. (2014). Extraction for metabolomics: Access to the metabolome. *Phytochemical Analysis, 25*(4), 291–306. https://doi.org/10.1002/pca.2505

Naser, F. J., Mahieu, N. G., Wang, L., Spalding, J. L., Johnson, S. L., & Patti, G. J. (2018). Two complementary reversed-phase separations for comprehensive coverage of the

semipolar and nonpolar metabolome. *Analytical and Bioanalytical Chemistry, 410*(4), 1287–1297. https://doi.org/10.1007/s00216-017-0768-x

Noshad, M., Behbahani, B. A., & Karabagias, I. K. (2023). Volatilomic with chemometrics: A toward authentication approach for food authenticity control. *European Food Research and Technology, 249*(9), 2215–2226. https://doi.org/10.1007/s00217-023-04291-0

Oliveri, P., Malegori, C., Mustorgi, E., & Casale, M. (2021). Qualitative pattern recognition in chemistry: Theoretical background and practical guidelines. *Microchemical Journal, 162*, Article 105725. https://doi.org/10.1016/j.microc.2020.105725. October 2020.

Pang, S., Piao, X., Zhang, X., Chen, X., Zhang, H., Jin, Y., et al. (2023). Discrimination for geographical origin of Panax quinquefolius L. using UPLC Q-Orbitrap MS-based metabolomics approach. *Food Science and Nutrition, 11*(8), 4843–4852. https://doi.org/10.1002/fsn3.3461

Peng, B., Li, H., & Peng, X. X. (2015). Functional metabolomics: From biomarker discovery to metabolome reprogramming. *Protein and Cell, 6*(9), 628–637. https://doi.org/10.1007/s13238-015-0185-x

Peng, Y., Zheng, C., Guo, S., Gao, F., Wang, X., Du, Z., et al. (2023). Metabolomics integrated with machine learning to discriminate the geographic origin of Rougui Wuyi rock tea. *Npj Science of Food, 7*(1). https://doi.org/10.1038/s41538-023-00187-1

Pereira, A. G., Cassani, L., Oludemi, T., Chamorro, F., Calhelha, R. C., Prieto, M. A., et al. (2023). Untargeted metabolomics and in vitro functional analysis unravel the intraspecific bioactive potential of flowers from underexplored Camellia japonica cultivars facing their industrial application. *Industrial Crops and Products, 204*, Article 117389.

Perez-Matas, E., Garcia-Perez, P., Bonfill, M., Lucini, L., Hidalgo-Martinez, D., & Palazon, J. (2023). Impact of elicitation on plant antioxidants production in Taxus cell cultures. *Antioxidants, 12*(4). https://doi.org/10.3390/antiox12040887

Perez-Matas, E., Garcia-Perez, P., Miras-Moreno, B., Lucini, L., Bonfill, M., Palazon, J., et al. (2023). Exploring the interplay between metabolic pathways and taxane production in elicited Taxus baccata cell suspensions. *Plants, 12*(14), 1–18. https://doi.org/10.3390/plants12142696

Picard, M., Scott-Boyer, M. P., Bodein, A., Périn, O., & Droit, A. (2021). Integration strategies of multi-omics data for machine learning analysis. *Computational and Structural Biotechnology Journal, 19*, 3735–3746. https://doi.org/10.1016/j.csbj.2021.06.030

Pinto, R. C. (2017). Chemometrics methods and strategies in metabolomics. *Metabolomics: From Fundamentals to Clinical Applications*, 163–190.

Pomyen, Y., Wanichthanarak, K., Poungsombat, P., Fahrmann, J., Grapov, D., & Khoomrung, S. (2020). Deep metabolome: Applications of deep learning in metabolomics. *Computational and Structural Biotechnology Journal, 18*, 2818–2825.

Qiu, S., Guo, S., Yang, Q., Xie, Y., Tang, S., & Zhang, A. (2023). Innovation in identifying metabolites from complex metabolome—highlights of recent analytical platforms and protocols. *Frontiers in Chemistry, 11*(January), 1–3. https://doi.org/10.3389/fchem.2023.1129717

Rasamoelina, A. D., Adjailia, F., & Sincak, P. (2020). A review of activation function for artificial neural network. *Sami 2020 - IEEE 18th world symposium on applied machine intelligence and informatics, proceedings*. https://doi.org/10.1109/SAMI48414.2020.9108717

Ren, J. L., Zhang, A. H., Kong, L., & Wang, X. J. (2018). Advances in mass spectrometry-based metabolomics for investigation of metabolites. *RSC Advances, 8*(40), 22335–22350. https://doi.org/10.1039/c8ra01574k

Rivera-Pérez, A., García-Pérez, P., Romero-González, R., Garrido Frenich, A., & Lucini, L. (2022). An untargeted strategy based on UHPLC-QTOF-HRMS metabolomics to identify markers revealing the terroir and processing effect on thyme phenolic profiling. *Food Research International, 162*(July), Article 112081. https://doi.org/10.1016/j.foodres.2022.112081

Rivera-Pérez, A., García-Pérez, P., Romero-González, R., Garrido Frenich, A., & Lucini, L. (2023). UHPLC-QTOF-HRMS metabolomics insight on the origin and processing authentication of thyme by comprehensive fingerprinting and chemometrics. *Food Chemistry, 407*(August 2022). https://doi.org/10.1016/j.foodchem.2022.135123

Rivera-Pérez, A., Romero-González, R., & Frenich, A. G. (2023). Untargeted 1H NMR-based metabolomics and multi-technique data fusion: A promising combined approach for geographical and processing authentication of thyme by multivariate statistical analysis. *Food Chemistry, 420*, Article 136156.

Roberts, L. D., Souza, A. L., Gerszten, R. E., & Clish, C. B. (2012). Targeted metabolomics. *Current Protocols in Molecular Biology, 1*(SUPPL.98), 1–24. https://doi.org/10.1002/0471142727.mb3002s98

Rocchetti, G., & O'Callaghan, T. F. (2021). Application of metabolomics to assess milk quality and traceability. *Current Opinion in Food Science, 40*, 168–178. https://doi.org/10.1016/j.cofs.2021.04.005

Rocchetti, G., Rebecchi, A., Maria Lopez, C., Dallolio, M., Dallolio, G., Trevisan, M., et al. (2023). Impact of axenic and mixed starter cultures on metabolomic and sensory profiles of ripened Italian salami. *Food Chemistry, 402*(August 2022), Article 134182. https://doi.org/10.1016/j.foodchem.2022.134182

Roullier-Gall, C., Lucio, M., Noret, L., Schmitt-Kopplin, P., & Gougeon, R. D. (2014). How subtle is the "terroir" effect? Chemistry-related signatures of two "climats de Bourgogne.". *PLoS One, 9*(5), Article e97615.

Salehi, H., Zhang, L., Alp-Turgut, F. N., Arikan, B., Elbasan, F., Ozfidan-Konakci, C., et al. (2023). The exogenous application of naringenin and rosmarinic acid modulates functional traits in Lepidium sativum. *Journal of the Science of Food and Agriculture*. https://doi.org/10.1002/jsfa.13160. November.

Salek, R. M., Neumann, S., Schober, D., Hummel, J., Billiau, K., Kopka, J., et al. (2015). COordination of standards in MetabOlomicS (COSMOS): Facilitating integrated metabolomics data access. *Metabolomics, 11*(6), 1587–1597. https://doi.org/10.1007/s11306-015-0810-y

Schippers, P., Rasheed, S., Park, Y. M., Risch, T., Wagmann, L., Hemmer, S., et al. (2023). Evaluation of extraction methods for untargeted metabolomic studies for future applications in zebrafish larvae infection models. *Scientific Reports, 13*(1), 1–15. https://doi.org/10.1038/s41598-023-34593-y

Schmidtke, L. M., Antalick, G., Šuklje, K., Blackman, J. W., Boccard, J., & Deloire, A. (2020). Cultivar, site or harvest date: The gordian knot of wine terroir. *Metabolomics, 16*(5), 1–17. https://doi.org/10.1007/s11306-020-01673-3

Schwaiger-Haber, M., Stancliffe, E., Arends, V., Thyagarajan, B., Sindelar, M., & Patti, G. J. (2021). A workflow to perform targeted metabolomics at the untargeted scale on a triple quadrupole mass spectrometer. *ACS Measurement Science Au, 1*(1), 35–45. https://doi.org/10.1021/acsmeasuresciau.1c00007

Schwalbe, H. (2017). Editorial: New 1.2 GHz NMR spectrometers— new horizons? *Angewandte Chemie International Edition, 56*(35), 10252–10253. https://doi.org/10.1002/anie.201705936

Selamat, J., Rozani, N. A. A., & Murugesu, S. (2021). Application of the metabolomics approach in food authentication. *Molecules, 26*(24), 1–19. https://doi.org/10.3390/molecules26247565

Senizza, B., Ganugi, P., Trevisan, M., & Lucini, L. (2023). Combining untargeted profiling of phenolics and sterols, supervised multivariate class modelling and artificial neural networks for the origin and authenticity of extra-virgin olive oil: A case study on Taggiasca ligure. *Food Chemistry, 404*(PA), Article 134543. https://doi.org/10.1016/j.foodchem.2022.134543

Serag, A., Zayed, A., Mediani, A., & Farag, M. A. (2023). Integrated comparative metabolite profiling via NMR and GC–MS analyses for tongkat ali (Eurycoma longifolia) fingerprinting and quality control analysis. *Scientific Reports, 13*(1), 2533.

Shakiba, N., Gerdes, A., Holz, N., Wenck, S., Bachmann, R., Schneider, T., et al. (2022). Determination of the geographical origin of hazelnuts (Corylus avellana L.) by Near-Infrared spectroscopy (NIR) and a Low-Level Fusion with nuclear magnetic resonance (NMR). *Microchemical Journal, 174*(December 2021). https://doi.org/10.1016/j.microc.2021.107066

Sobolev, A. P., Thomas, F., Donarski, J., Ingallina, C., Circi, S., Cesare Marincola, F., et al. (2019). Use of NMR applications to tackle future food fraud issues. *Trends in Food Science and Technology, 91*(July), 347–353. https://doi.org/10.1016/j.tifs.2019.07.035

Spiteri, C., Lia, F., & Farrugia, C. (2020). Determination of the geographical origin of Maltese honey using 1H NMR fingerprinting. *Foods, 9*(10), 1–18. https://doi.org/10.3390/foods9101455

Spiwok, V., & Kvriž, P. (2020). Time-lagged t-distributed stochastic neighbor embedding (t-SNE) of molecular simulation trajectories. *Frontiers in Molecular Biosciences, 7*, 132.

Srikanth Kumar, K., Lakshmana Rao, A., & Basaveswara Rao, M. V. (2018). Design, synthesis, biological evaluation and molecular docking studies of novel 3-substituted-5-[(indol-3-yl)methylene]-thiazolidine-2,4-dione derivatives. *Heliyon, 4*(9), Article e00807. https://doi.org/10.1016/j.heliyon.2018.e00807

Sumner, L. W., Amberg, A., Barrett, D., Beale, M. H., Beger, R., Daykin, C. A., et al. (2007). Proposed minimum reporting standards for chemical analysis: Chemical analysis working group (CAWG) metabolomics standards initiative (MSI). *Metabolomics, 3*(3), 211–221. https://doi.org/10.1007/s11306-007-0082-2

Toda, A., Goudo, M., Sugimoto, M., Hiwa, S., & Hiroyasu, T. (2023). Comparison of classification accuracy and feature selection between sparse and non-sparse modeling of metabolomics data. *bioRxiv*, 2004–2023.

Troisi, J., Richards, S. M., Scala, G., & Landolfi, A. (2022). Approaches in untargeted metabolomics. In *Metabolomics perspectives* (pp. 237–262). Elsevier.

Tzora, A., Nelli, A., Kritikou, A. S., Katsarou, D., Giannenas, I., Lagkouvardos, I., et al. (2022). The "crosstalk" between microbiota and metabolomic profile of kefalograviera cheese after the innovative feeding strategy of dairy sheep by omega-3 fatty acids. *Foods, 11*(20). https://doi.org/10.3390/foods11203164

Valdés, A., Álvarez-Rivera, G., Socas-Rodríguez, B., Herrero, M., Ibáñez, E., & Cifuentes, A. (2022). Foodomics: Analytical opportunities and challenges. *Analytical Chemistry, 94*(1), 366–381. https://doi.org/10.1021/acs.analchem.1c04678

Wang, L., Naser, F. J., Spalding, J. L., & Patti, G. J. (2019). A protocol to compare methods for untargeted metabolomics. *Metabolic signaling: Methods and protocols*.

Wang, Z., Zhou, L., Hao, W., Liu, Y., Xiao, X., Shan, X., et al. (2023). Comparative antioxidant activity and untargeted metabolomic analyses of cherry extracts of two Chinese cherry species based on UPLC-QTOF/MS and machine learning algorithms. *Food Research International, 171*(June), Article 113059. https://doi.org/10.1016/j.foodres.2023.113059

Weng, L. C., Elsawah, A. M., & Fang, K. T. (2021). Cross-entropy loss for recommending efficient fold-over technique. *Journal of Systems Science and Complexity, 34*(1), 402–439. https://doi.org/10.1007/s11424-020-9267-9

Windarsih, A., Rohman, A., Riswanto, F. D. O., Dachriyanus, Yuliana, N. D., & Bakar, N. K. A. (2022). The metabolomics approaches based on LC-MS/MS for analysis of non-halal meats in food products: A review. *Agriculture, 12*(7), 984.

Winstel, D., Bahammou, D., Albertin, W., Waffo-Téguo, P., & Marchal, A. (2021). Untargeted LC–HRMS profiling followed by targeted fractionation to discover new taste-active compounds in spirits. *Food Chemistry, 359*, Article 129825.

Wishart, D. S., Cheng, L. L., Copié, V., Edison, A. S., Eghbalnia, H. R., Hoch, J. C., et al. (2022). *NMR and metabolomics — a roadmap for the future* (Vols. 1–20).

Worley, B., & Powers, R. (2012). Multivariate analysis in metabolomics. *Current Metabolomics, 1*(1), 92–107. https://doi.org/10.2174/2213235x130108

Wu, W., Zhang, L., Zheng, X., Huang, Q., Farag, M. A., Zhu, R., et al. (2022). Emerging applications of metabolomics in food science and future trends. *Food Chemistry X, 16*(August), Article 100500. https://doi.org/10.1016/j.fochx.2022.100500

Xie, Y., Nie, X., Wang, C., Xu, X., & Zhang, F. (2024). A metabolomics approach for monitoring thermal processing temperature of bovine milk using ultra-performance

liquid chromatography tandem Q-Exactive mass spectrometry and multivariate data analysis. *Journal of Future Foods, 4*(1), 83–90.

Xu, J., Wang, Y., Xu, X., Cheng, K.-K., Raftery, D., & Dong, J. (2021). NMF-based approach for missing values imputation of mass spectrometry metabolomics data. *Molecules, 26*(19), 5787.

Yang, X. fang, Miao, Y., Yang, D. wei, Kong, R., Yuan, B., Quan, J. ye, et al. (2023). UPLC-Q TOF-MS-Based metabolomics and anti-myocardial ischemia activity of Dioscoreae Nipponicae Rhizoma from different geographical origins. *Journal of Pharmaceutical and Biomedical Analysis, 234*(July), Article 115551. https://doi.org/10.1016/j.jpba.2023.115551

Yu, Z., Member, S., Chen, H., Liu, J., You, J., Leung, H., et al. (2016). Hybrid k -nearest neighbor classifier. *IEEE Transactions on Cybernetics, 46*(6), 1263–1275.

Zambianchi, S., Patrone, V., Becchi, P. P., Callegari, M. L., Stagnati, L., Lucini, L., et al. (2023). Metagenomic bacterial diversity and metabolomics profiling of Buttafuoco wine production. *Food Control, 148*(January), Article 109657. https://doi.org/10.1016/j.foodcont.2023.109657

Zhang, L., García-Pérez, P., Martinelli, E., Giuberti, G., Trevisan, M., & Lucini, L. (2023). Different fractions from wheat flour provide distinctive phenolic profiles and different bioaccessibility of polyphenols following in vitro digestion. *Food Chemistry, 404*(September 2022). https://doi.org/10.1016/j.foodchem.2022.134540

Zhang, X., Kontoudakis, N., Šuklje, K., Antalick, G., Blackman, J. W., Rutledge, D. N., et al. (2020). Changes in red wine composition during bottle aging: Impacts of grape variety, vineyard location, maturity, and oxygen availability during aging. *Journal of Agricultural and Food Chemistry, 68*(47), 13331–13343. https://doi.org/10.1021/acs.jafc.9b07164

Zhang, X.-K., Lan, Y.-B., Huang, Y., Zhao, X., & Duan, C.-Q. (2021). Targeted metabolomics of anthocyanin derivatives during prolonged wine aging: Evolution, color contribution and aging prediction. *Food Chemistry, 339*, Article 127795.

Zhang, S., Li, Q., Wen, S., Sun, L., Chen, R., Zhang, Z., et al. (2023). Metabolomics reveals the effects of different storage times on the acidity quality and metabolites of large-leaf black tea. *Food Chemistry, 426*(June), Article 136601. https://doi.org/10.1016/j.foodchem.2023.136601

others Zhang, M., Sun, L., Su, R., Corazzin, M., Yang, Z., Dou, L., et al. (2024). Widely targeted metabolomic analysis reveals the dynamic changes of metabolites during postmortem chilled aging in Mongolian sheep. *Food Chemistry, 431*, Article 137035.

Zhang, Q. H., Zhou, L. Di, Chen, H., Wang, C. Z., Xia, Z. N., & Yuan, C. S. (2016). Solid-phase microextraction technology for in vitro and in vivo metabolite analysis. *TrAC, Trends in Analytical Chemistry, 80*, 57–65. https://doi.org/10.1016/j.trac.2016.02.017

Zhao, Z., Hamid, N., Gutierrez-Maddox, N., Kam, R., Kantono, K., Wang, K., et al. (2023). Metabolite profiling identifies chemical markers associated with the cytotoxic properties of roasted fermented avocado seeds. *Food Research International, 113131*.

Zhu, X., & Goldberg, A. B. (2022). *Introduction to semi-supervised learning*. Springer Nature.

Zhu, C., Lu, Q., Zhou, X., Li, J., Yue, J., Wang, Z., et al. (2020). Metabolic variations of organic acids, amino acids, fatty acids and aroma compounds in the pulp of different pummelo varieties. *Lebensmittel-Wissenschaft und -Technologie, 130*, Article 109445.

Zulfiqar, M., Gadelha, L., Steinbeck, C., Sorokina, M., & Peters, K. (2023). Maw: The reproducible metabolome annotation workflow for untargeted tandem mass spectrometry. *Journal of Cheminformatics, 15*(1), 1–17. https://doi.org/10.1186/s13321-023-00695-y