Natural Products Analysis
https://www.sciltp.com/journals/npa

Review

Chemoinformatic Tools Used in Untargeted Metabolomic Approaches by Mass Spectrometry Applied for Natural Product Analyses

Augustin Bildstein [1], Sylvie Chollet [2] and Céline Rivière [1,*]

[1] Joint Research Unit 1158 BioEcoAgro, University of Lille, JUNIA, UPJV, University of Liège, INRAE, University of Artois, University Littoral Côte d'Opale, F-59650 Villeneuve d'Ascq, France
[2] Adrianor, F-62217 Tilloy-lès-Mofflaines, France
* Correspondence: celine.riviere@univ-lille.fr

How To Cite: Bildstein, A.; Chollet, S.; Rivière, C. Chemoinformatic Tools Used in Untargeted Metabolomic Approaches by Mass Spectrometry Applied for Natural Product Analyses. *Natural Products Analysis* **2025**, *1*(1), 100010. https://doi.org/10.53941/npa.2025.100010

Received: 1 November 2025
Revised: 14 December 2025
Accepted: 14 December 2025
Published: 22 December 2025

Abstract: The chemistry of natural products has undergone a major transformation in the last twenty years, largely due to the development of powerful coupling techniques such as LC-HRMS/MS. These techniques, combined with supervised and unsupervised multivariate statistical analyses, are used for untargeted metabolomic studies for a wide range of applications. They have also enabled the development of dereplication approaches, thus accelerating often lengthy purification processes by focusing on biologically active metabolites of unknown structure. These dereplication approaches have been further strengthened in recent years by the development of molecular networks, based on the principle of grouping compounds according to their fragmentation profile in mass spectrometry. One of the current challenges remains the annotation of a large number of variables with a high degree of confidence. This will require enriching existing databases, and more recently, leveraging artificial intelligence. The latter, integrating in-silico virtual screening and chemoinformatic approaches, is now emerging as a powerful tool for predicting biological activity.

Keywords: mass spectrometry; untargeted metabolomics; chemometrics, molecular networking; dereplication

1. Introduction

Untargeted metabolomics is an approach aimed at analyzing all the metabolites of an organism, while targeted metabolomics analyzes some specific compounds of interest that have been previously characterized. Untargeted metabolomics, using high resolution mass spectrometry applied to the analysis of natural products, follow a standardized workflow. After proper sample preparation, this workflow begins with equipment selection (analyzer type, source type, ionization mode) and data acquisition, including choosing the appropriate acquisition mode. Once the data are acquired, metabolomic analysis follows a conventional process involving various software programs and chemoinformatic tools used at key stages. The choice of analytical parameters and statistical methods presents a major challenge for interpreting the results. Thus, through the example of an untargeted analysis of a hop collection including wild, heirloom, and commercial genotypes, we present some common limitations and difficulties related to the choice of analytical parameters. Emphasis will also be placed on dereplication, encompassing all the methods used to rapidly identify known compounds, on the contribution of molecular networks to dereplication strategies and on the importance of annotating identified variables with different confidence levels.

2. Mass Spectrometry-Based Untargeted Metabolomics

2.1. Introduction to Omics Approaches

The suffix "-omics" refers to a field of recent disciplines aimed at characterizing and quantifying, at different scales, a wide range of problematic biochemical variables in organisms [1,2]. Genomics, which focuses on the genome, literally "complete set of genes" [3] and which today refers to the entirety of DNA, was the first discipline to adopt this approach. Similar terms were subsequently used to designate the complete set of gene transcripts—transcriptome (total RNA), the complete set of translation products—proteome (total proteins), and today the complete set of metabolites—metabolome (including endometabolome and exometabolome). Thus, the term 'omics' refers to the study of cellular metabolism (Figure 1). Metabolomics has rapidly established itself as a technique of choice for phytochemical analyses and the study of specialized metabolites [4–6].

"Omics" approaches rely on high-throughput analysis technologies that allow for the characterization of a large number of individuals. The goal is to obtain an overview of the organisms studied—in our case, the hop metabolome—and to establish correlations with other qualitative and/or quantitative variables.

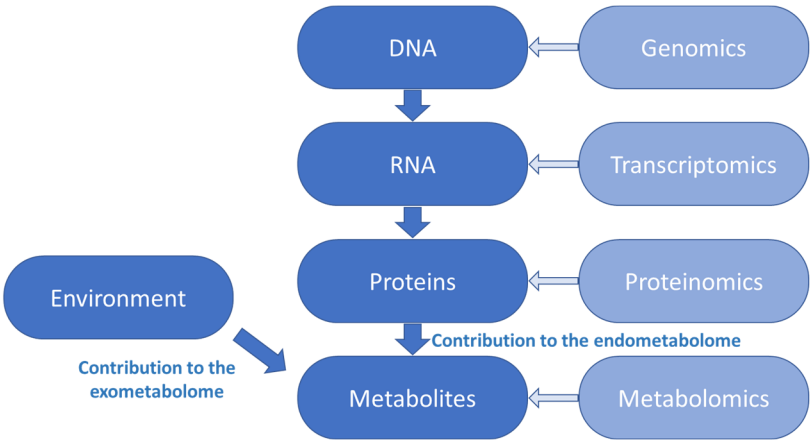

Figure 1. "-omics" approaches at different scales.

2.2. Types of Analyses in Metabolomics

Two main approaches are distinguished in metabolomics [7].

- the targeted approach, which analyzes certain compounds of interest that have been previously characterized;
- the untargeted approach, which tends to analyze a panel of compounds, as broad as possible, without any preconceived notions.

The untargeted approach allows for a more complete description of the metabolome description of the organism under study, including unknown or low-abundance compounds, and thus limits biases related to variable selection. However, it exhibits some sensitivity to data detection and processing parameters, which can lead to variability in the results [8].

Untargeted analysis can be performed by mass spectrometry (MS) or nuclear magnetic resonance (NMR) [9]. In MS, detection is generally coupled to a chromatographic method, either ultra-high-performance liquid chromatography (UHPLC) or gas chromatography (GC) [10]. The analyzers used are preferably high-resolution (HRMS) analyzers such as time-of-flight, orbitrap, or Fourier transform analyzers, providing high accuracy in measuring masses. The use of tandem mass spectrometry (MS/MS), which effectively couples these analyzers to a quadrupole and a collision cell, also provides structural information useful for identifying certain variables [11].

MS data acquisition is mainly done via three modes [10,12]:

- *Full scan acquisition*: This mode allows for obtaining the mass-to-charge ratios (m/z) and the relative abundance of metabolites for each sample. The MS1 data are then processed by specialized software (such as MZmine) to extract reliable variables, which are then statistically analyzed to identify those that vary significantly among the samples. Subsequently, an MS2 analysis on a pooled sample provides the fragmentation spectra necessary for the chemical identification of the metabolites.
- *Data-dependent acquisition (DDA)*: In this mode, the analyzer first acquires MS1 spectra via a full scan. Then, the quadrupole selects, based on user-defined parameters such as an inclusion list or intensity-dependent thresholds (hence the term "data-dependent"), precursor ions for MS2 analysis to obtain MS/MS fragmentation spectra. One limitation of this mode is that low-abundance metabolites cannot be fragmented.

- *Data-independent acquisition (DIA)*: In this mode, after a full scan acquisition, the analyzer systematically fragments all precursor ions within a defined, generally quite wide, *m*/*z* range. Fragment matching to the parent ion is performed post-acquisition by deconvolution using algorithms. This mode was introduced to overcome the limitations of DDA, particularly its inability to fragment low-abundance metabolites. However, the generated MS2 spectra can contain fragments from multiple precursors, making the correct assignment of fragments to each metabolite complex. The MSE acquisition, a specific technology developed by Waters®, is classified as a DIA. It improves assignments. It consists of alternating two MS/MS functions: One with low collision energy (inducing little or no fragmentation) and one with high collision energy [13]. An ion mobility cell upstream of the collision cell facilitates the allocation of parent ions by similarity of drift times in the mobility cell (drift time) [14].

Once the data are acquired, metabolomic analysis follows a classic process involving various chemoinformatic tools, often in the form of open-source software developed by and for researchers (Figure 2) [6,15,16]. These software programs and platforms are fully aligned with the principles of open science through data sharing and encourage collaboration via discussion forums.

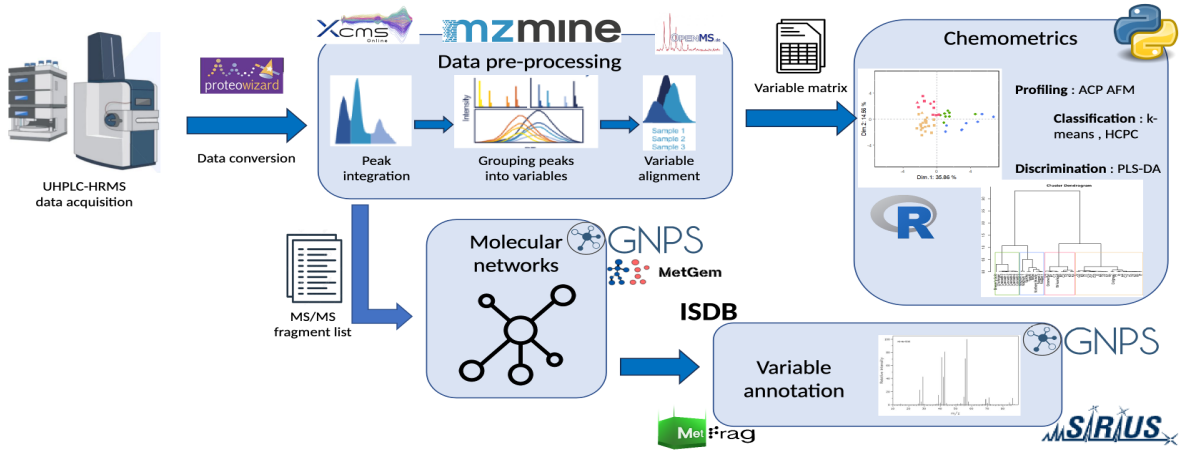

Figure 2. Standard process for metabolomic analyses (inspired by Ramos et al. [15]). Logos of the main open-source tools associated with these steps are shown.

3. Preprocessing of Mass Spectrometry Data

Before any bioinformatic analysis, the acquired data must be cleaned, grouped, and aligned [17–19]. Several free software programs such as MzMine4 [19], OpenMS [20,21] or XCMS [22], as well as software from manufacturers like Progenesis QI (Waters, Milford, MA, USA), allow this preprocessing, according to algorithms that sometimes differ: For example, in Progenesis QI, the alignment of chromatograms precedes the detection of peaks.

3.1. Importing Data

Free and open-source software, or software licensed for academic purposes, uses so-called “open” formats, such as .mzXML or more recently .mzML, which is currently the most widespread. A preliminary file conversion step is necessary, using tools such as ProteoWizard (open source) [23] or proprietary software like MassHunter Qualitative Analysis (Agilent, Santa Clara, CA, USA).

3.2. Peak Detection

The acquired mass signals are processed and integrated to generate peaks representative of each detected ion. Detection follows several steps, the parameters of which must be optimized: Minimum intensity, number of scans, mass tolerance, etc. These parameters depend on both the instrumentation used and the matrix being analyzed. Filters that are too permissive can incorporate redundant, uninformative or background noise variables, while filters that are too strict risk losing low-intensity signals [24].

3.3. Grouping Peaks into Variables

In MS acquisition, a single compound can generate multiple signals corresponding to its different isotopes, adducts, in-source fragments… Several steps allow these signals to be grouped into a single variable. These steps

include deisotopization (grouping of isotopic clusters), adduct grouping, or, in the case of electron impact (EI) analysis, fragment deconvolution. Indeed, EI is a hard ionization that induces numerous fragments in the source [25]. Grouping algorithms rely on similarities in retention time, peak shapes, and mass differences. Excessively restrictive tolerances can lead to duplicates; conversely, excessively broad tolerances can merge signals from different compounds, resulting in information loss [24].

3.4. Variable Alignment

The variables for all analyzed individuals are aligned according to retention time and the shape of the chromatographic peaks [26]. This alignment generates an intensity matrix, in which each case corresponds to the area under the curve of a variable for a given individual. This data matrix can then be filtered to eliminate redundant, uninformative, or from background noise variables [27]. Typically, removing variables present in analytical blanks eliminates column impurities, while comparison with quality control (QC) samples—that is, an equivolumic mixture of the different analyzed samples—ensures good selection of variables representative of the dataset [18]. Furthermore, it is not uncommon to identify variables present only in a single sample, often attributable to analytical or preprocessing artifacts.

3.5. Exporting Spectral Data

In addition to the data matrix, preprocessing software can generate a list of MS/MS spectra associated with the aligned variables. Several formats exist for representing this spectral data, each with its own specific syntax. Among them are the .msp format produced by Progenesis, the .mgf format generated by MzMine, among others, and the .ms file, specific to the SIRIUS software, which allows the aggregation of several lists of fragments obtained under different fragmentation conditions into a single variable. These files list the fragmentation spectra (*m*/*z*—intensity) associated with each variable and also include a set of metadata (identification, retention time, collision energy, adduct type, etc.) (Figure 3). These formats have the particularity of being readable and editable like simple text files, either manually or automatically. The .mgf format is a widely adopted format within the metabolomic community. It is recognized by most of the chemoinformatic tools and allows for the translation of a significant amount of spectral information with minimal storage cost. This enables the sharing and reuse of data within the scientific community, fully aligning with current open science principles [15].

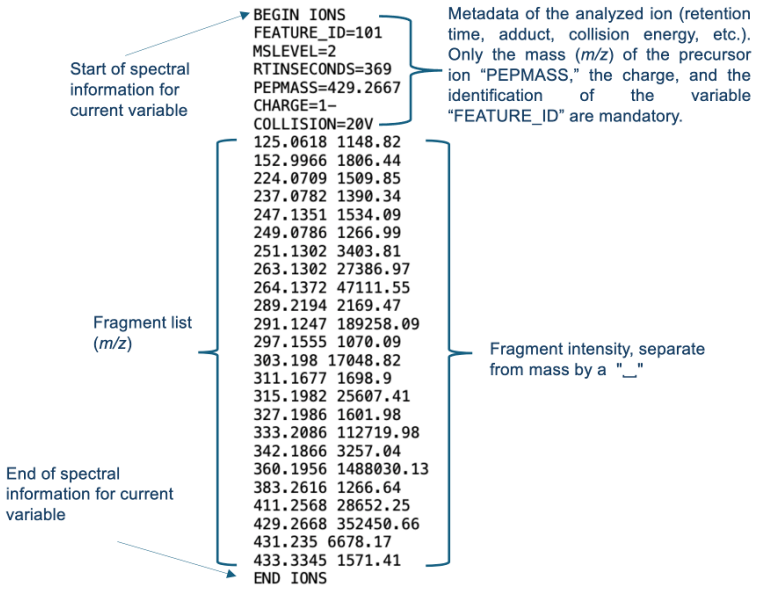

Figure 3. Syntax of a .mgf file.

4. Chemometrics

Preprocessed mass spectrometry data matrices contain a significant amount of information that is uninterpretable in its raw form. Chemometrics encompasses all the statistical analyses used to interpret these complex data through dimensionality reduction, classification, and integration with other datasets.

Chemometric studies can be performed using numerous statistical software programs, both licensed and open-source. Open-source software using programming languages such as R or Python offers greater freedom in

customizing parameters and creating graphical representations thanks to code enhanced by numerous packages. Many multivariate statistical analyses have been developed for data processing, including both unsupervised and supervised analyses [28]. Each method has its advantages and limitations, and the choice of methods will depend primarily on the objectives of the statistical interpretation. The commonly used statistical methods are presented below.

4.1. Dimensional Reduction by Principal Component Analysis

Dimensional reduction methods allow complex matrices to be projected into a lower dimension space (two or three axes), thus facilitating the visualization of underlying structures.

Principal component analysis (PCA) transforms initially correlated variables into new variables, called principal components. These components are orthogonal linear correlations, i.e., uncorrelated to each other. It allows for a reduction in data dimensionality while retaining maximum information, particularly to facilitate visualization and interpretation [29]. In the example shown below, observations are represented as score plot (Figure 4A,C), while variables appear loading plot (Figure 4B,D).

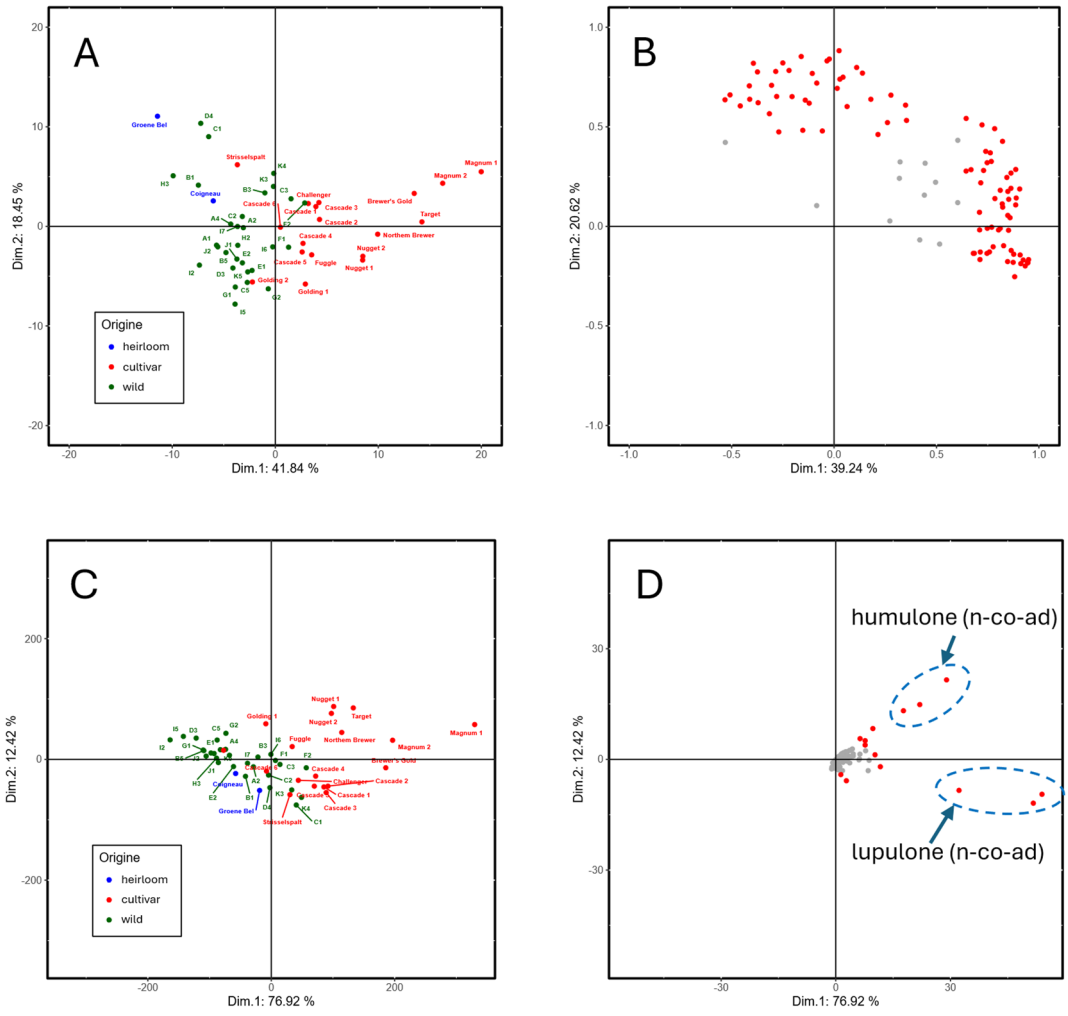

Figure 4. Principal component analyses: Score plots (**A**,**C**) and loading plots (**B**,**D**) at different scales: Standardized scale (**A**,**B**) and Pareto scale (**C**,**D**). Variables in red represent variables contributing to the construction of components 1 and 2 ($\cos^2 > 0.5$).

To ensure all variables contribute equally, the data are generally centered (subtracting the mean) and scaled (dividing by the standard deviation) (Figure 4A,B). Another scaling method used in metabolomics is Pareto scaling, which involves dividing by the square root of the standard deviation after centering (Figure 4C,D). This approach partially preserves the quantity effect of variables [30]. In an example of untargeted metabolomic analysis applied to hop resin (*Humulus lupulus* L.), the application of the Pareto scaling highlighted the major hop compounds (major alpha and beta acids) in hop samples (Figure 4D) that primarily contributed to axis construction,

while centered scaled scaling revealed more low-abundance metabolites contributing to the chemical diversity of the hop collection (Figure 4B).

4.2. Dimensional Reduction by Regression

Another approach is partial least squares discriminant analysis (PLS-DA). This is a supervised method (with prior knowledge of the classes), unlike PCA, which is unsupervised. PLS-DA groups variables into latent components that maximize the separation between defined groups [31,32]. This method allows for the identification of variables specific to a category. A variant widely used in metabolomics is orthogonal PLS-DA (OPLS-DA), which separates variables into a predictive component (correlated with the groups) and an orthogonal component (not statistically correlated with the groups) [33]. This distinction improves interpretation, particularly in metabolomics where a large number of variables can complicate model construction.

4.3. Classification

The classification aims to group individuals according to the distribution of variables, that is, in metabolomics, according to their chemical similarity. Unsupervised methods, such as *k-means* or hierarchical clustering analysis (HCA) allow for the identification of groupings [34,35]. Generally, grouping is based on the coordinates of the first principal components to focus the classification on the variables most representative of the group's variance; this is referred to as HCPC (Hierarchical Clustering on Principal Component) [36].

4.4. Towards Multi-Table Approaches

One of the major advantages of metabolomics data lies in its integration with other sources of information, whether they are of a "-omics" nature, such as genomics [6] (these are referred to as "multi-omics" approaches), or related to biological or sensory characteristics. Multi-table approaches allow for the integration of this data and the observation of correlations between different types of variables.

Unsupervised multiple factor analysis (MFA) allows for the description of individuals based on several datasets, similar to PCA [37]. It makes it possible to analyze several tables of variables simultaneously, and to obtain results, in particular, charts, that allow studying the relationship between the individuals, variables, and tables [38].

Finally, in a supervised manner, PLS regression (PLS-R), applied to two blocks of data, allows us to explore the covariance between datasets. Regression models can be constructed to predict a response Y (e.g., biological activities) from a metabolomic table X. This is referred to as PLS-R1 when the response Y is univariate (a vector) and PLS-R2 when the response Y is multivariate (a matrix). The model constructs latent variables by maximizing the covariance between X and Y. Graphical representations resulting from this model include score plots (projection of individuals) and loading plots (projection of variables from X and Y). The variables contributing to the construction of the model, i.e., those that allow for a better prediction of the Y response, are evaluated using the VIP (Variable Importance in Projection) score [39].

5. Molecular Networking

5.1. Principles of Molecular Networking

Molecular networks are a schematic representation of a set of metabolites grouped according to the similarity of their mass fragmentation profiles, whether obtained by MS/MS [40] or EI-MS [41]. The similarity of fragmentation profiles can reflect a structural similarity between two compounds.

Each spectrum is modeled as a normalized, multidimensional vector, where each dimension represents an m/z value and its relative intensity. The vectors are compared pairwise, and a cosine value ranging from 0 to 1 is calculated. The higher the cosine score between two vectors, the greater the similarity these compounds share based on their fragmentation profile (common fragments and/or losses), and therefore their structural proximity. Networks are constructed from all the cosine scores of the analyzed variables. The compounds, also called "nodes" in their graphical form, are connected by "bridges" if the cosine score exceeds a user-defined threshold. Groups of nodes then form and are arranged into a network (Figure 5) [40].

Initially developed for the study of metabolites of microbiological origin [42], this approach has become widely adopted in the study of natural products, including specialized plant metabolites [15]. Indeed, networks are well-suited to grouping naturally occurring chemical classes. Furthermore, molecular networks provide valuable and relevant structural information for unidentified compounds. Finally, the addition of metadata, made possible by the Featured Based Molecular Networking (FBMN) function offered by the Global Natural Products Social

Molecular Networking (GNPS) platform, allows for the integration of chromatographic information from preprocessing steps [40].

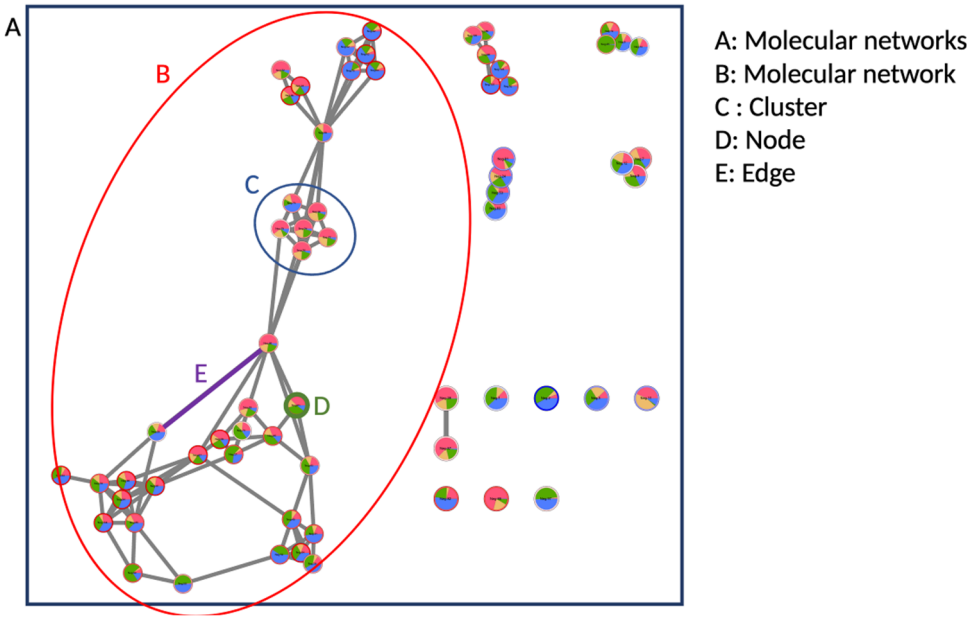

Figure 5. Example of molecular networks with associated terminology (adapted from Nothias-Scaglia et al. [40]). The segments of the pie chart represent the proportions of each variable in the identified hop chemotypes.

5.2. Software for Constructing, Visualizing, and Editing Molecular Networks

The first networks were built using scripts [42,43] and visualized using the open-source software Cytoscape [44]. Since then, the development of the online GNPS platform has enabled widespread adoption of the practice within the scientific community. This widespread adoption is due to the implementation of user-friendly online tools for network construction and dereplication, as well as a strong community aspect fostered by an active forum, educational documentation, and the sharing of experimental data on the platform [45].

More recently, the open-source software MetGem offers the advantage of building networks locally [46]. Local analysis eliminates the data import step and significantly reduces processing time. Optimizing construction parameters, where several trials are sometimes necessary to find the best representation, is facilitated. Furthermore, the software provides a network visualization and editing tool while also allowing export to Cytoscape.

The graphical editing of molecular networks is of particular interest, as it reveals metadata, whether qualitative or quantitative. Similar to chemometrics, graphical representation provides visual support for data interpretation. In the example given, representing the nodes as pie charts reflected the distribution of variables among the different hop chemotypes (Figure 5).

5.3. Construction Parameters and Limitations of Molecular Networks

The construction of molecular networks will depend largely on user-defined parameters and requires optimization based on the acquired data. As with the preprocessing of untargeted metabolomics data, mass tolerance will be primarily determined by the sensitivity of the MS analyzer. The key parameters for network construction will be the cosine score threshold value, the minimum number of fragments, and/or common neutral losses. The choice of these parameters will depend heavily on the quality of the MS/MS spectra obtained during acquisition, the number of integrated compounds, and the molecular families being studied. Generally, filters are used to limit the weight of minor fragments in the calculation of cosine scores.

When a dataset contains a large number of compounds, some groups can become too large to interpret. The “top-k” parameter limits the number of connections per node to the k closest compounds based on their cosine score. This results in a segmentation of the group into interconnected subgroups within a network.

One of the main limitations of molecular networks lies in their sensitivity to construction parameters, which can lead to very different architectures [47,48] Furthermore, the binary interpretation of networks (similar or dissimilar) can lead to erroneous assignments of chemical classes. In addition, compounds fragmented in a way that provides little information will not be well integrated into molecular networks.

5.4. t-SNE Networks

To overcome the limitations of molecular networks, a new representation, based on the t-SNE algorithm, has been developed as a complement to the classical approach. This functionality was recently developed for the MetGem software [46].

t-SNE (t-distributed Stochastic Neighbor Embedding) is a dimensionality reduction method that allows multidimensional data (such as MS/MS spectra) to be projected into a two-dimensional space while preserving local proximities.

The advantage of t-SNE representations lies in their ability to better reflect proximity relationships between compounds by including the notion of inter- and intra-group proximity. They also allow us to circumvent certain biases induced by the network architecture [46]. Our example presented in Figure 6 clearly illustrates this bias in the architecture of classical molecular networks. The non-oxygenated sesquiterpenes of hops, annotated in red and blue, were not grouped together in the classical representation (Figure 6A). Although they exhibited high cosine scores, the top-k parameter caused the blue subgroup to separate from the main cluster. This structural proximity was more faithfully reproduced in the t-SNE representation (Figure 6B), where the blue and red sesquiterpenes are grouped together.

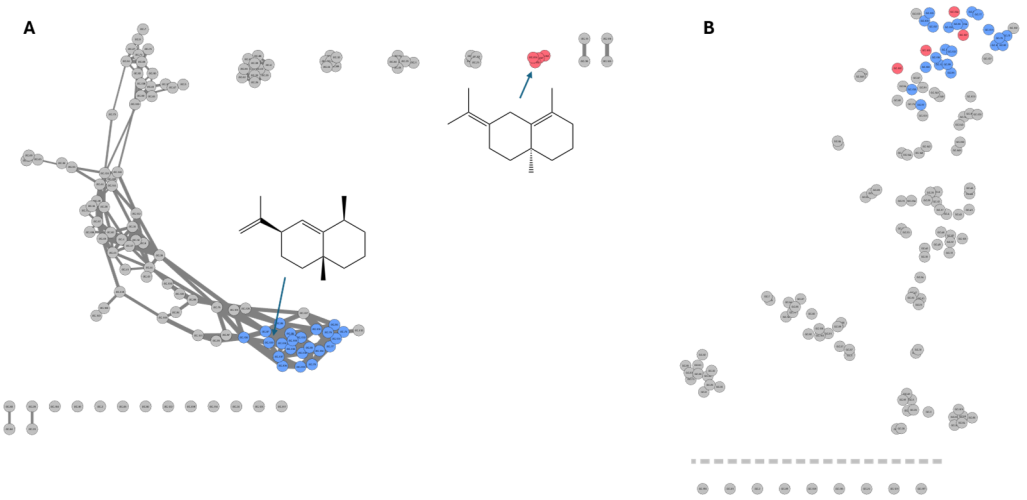

Figure 6. Comparison of visualization of a classical molecular network (**A**) and t-SNE (**B**) based on mass fragmentation profiles obtained by EI-MS. Red cluster is represented by selin-4,7(11)-diene derivatives whereas blue cluster is characterized by selin-5,11-diene derivatives.

However, unlike classical molecular networks, the t-SNE network does not offer absolute grouping in the sense that no bridges are built between the nodes. The grouping is relative, making interpretation more difficult. Classical and t-SNE representations are thus complementary [46].

6. Identification of Compounds by Dereplication

6.1. General Principles of Dereplication

In natural product chemistry, dereplication refers to methods for identifying known compounds in a complex mixture using computer tools, allowing purification efforts to then focus on compounds of unknown nature [49]. It relies primarily on search algorithms that compare fragmentation profiles obtained by HRMS/MS or electron impact (EI-MS) with those available in reference databases [49]. Several tools allow for the dereplication of variables, such as MzMine 4 or the GNPS platform, which automatically integrates dereplication during network construction [47]. Open-source databases include reference spectra or experimental data, such as the user-contributed GNPS database [45]. Dereplication can also rely on other spectroscopic methods, such as carbon-13 NMR (^{13}C NMR), or even the combination of HRMS/MS and ^{13}C NMR to improve identification, particularly in the presence of isomers [50,51].

6.2. In-Silico Dereplication

In the absence of a direct match with a database, or as a complement, identification can rely on in-silico methods [9,52,53]. First, the molecular formula of the unknown compound can be determined thanks to the

precision of the masses acquired by high-resolution mass spectrometry, as well as the isotopic distribution. The open-source software SIRIUS reconstructs the fragmentation tree of unknown compounds by analyzing fragments and neutral losses and proposes the molecular formula most consistent with the tree (Figure 7) [54]. Tools such MetFrag [55] compare experimental MS/MS spectra to simulated spectra from chemical structures from generic databases such as PubChem [56] or specialized in the field of natural products such as LOTUS [57]. Further, CSI:FingerID, integrated into SIRIUS, establishes a molecular fingerprint in experimental MS/MS spectra, then searches for candidate metabolites in structural databases [58].

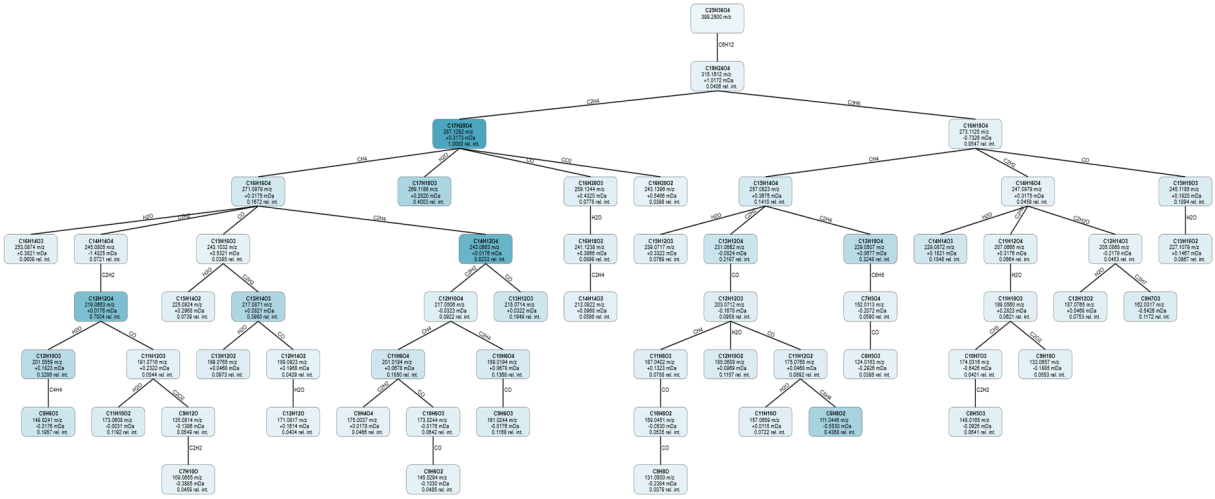

Figure 7. Fragmentation tree of co-lupulone treated by SIRIUS software.

6.3. Dereplication Limits

Dereplication methods allow for the rapid identification of a large number of variables. However, the confidence level of this identification should be considered with caution. Indeed, there is significant variability in fragmentation profiles acquired by HRMS/MS, making identification by spectral comparison to experimental or in-silico databases impractical. This variability is due, in part, to different experimental analysis conditions, such as the intensity of the analyzed compound, the collision energy, the matrix, and the ionization conditions and types, such as electrospray ionization (ESI) or atmospheric pressure chemical ionization (APCI), which can modify fragmentation [17].

On the other hand, even under similar experimental conditions, variations in fragmentation are observed depending on the equipment used, particularly the type of analyzer [59]. This inherent lack of robustness in the MS/MS fragmentation tandem must be taken into account when assessing the confidence level of proposed identifications. It is therefore essential for the chemist to confirm or refute these proposals by ensuring consistency with the scientific literature: Consistency between the proposed identification and existing chemotaxonomic data for the plant or species of the same genus; consistency between the retention time obtained by chromatography and the polarity of the candidate (logP); and consistency with the groupings of molecular networks [52].

The identification of compounds analyzed by GC-MS is facilitated by the inherent robustness of the source fragmentation by electron impact, as well as by the uniformity of the ionization parameters (70 V) in the databases [60]. Furthermore, comparing experimental retention indices with those from databases confirms the identification. A mixture of n-alkanes injected under the same experimental chromatographic conditions allows the retention indices of the compounds to be determined according to the formula of Kovátz [61]:

$$RI = 100\ (n + \frac{t_{r(x)} - t_{r(n)}}{t_{r(N)} - t_{r(n)}})$$

With:

RI: Retention Index

t_r: Retention Time

x: Unknown compound

n: The number of carbon atoms in the smallest n-alkane

N: The number of carbon atoms in the largest n-alkane

Finally, the main limitation of dereplication is that it only allows for the putative identification of previously described compounds [62]. However, the characterization of new compounds of natural origin remains a major

challenge, both for understanding plant metabolic pathways and for expanding chemical libraries of potentially valuable molecules, for example, in health. Furthermore, putative identification can erroneously lead to assigning a compound to a known molecule when it is actually a novel metabolite, which can result in the loss of an original discovery. To address these problems, the recently developed MS2DECIDE tool identifies "new" natural compounds, i.e., those not yet listed in databases [63]. This method allows for the prioritization of future research on these unknown compounds and enhance natural product database.

6.4. Characterization of Unannotated Compounds

In the presence of unannotated compounds, the putative annotation of chemical families nevertheless provides valuable information. It can initially be based on extrapolating the groupings observed in molecular networks, illustrating the complementarity between network representation and dereplication. ConCISE (Consensus Classifications of In Silico Elucidations) is a tool developed for this purpose, enabling the fusion of molecular networking, spectral libraries matching, and in-silico class predictions to establish accurate presumptive classifications for entire subnetworks [64]. A more detailed analysis of MS/MS fragmentation profiles, compared to reference profiles from the same hypothetical chemical family, allows for a more precise assignment [52]. Furthermore, recent chemoinformatics tools facilitate this approach. The CANOPUS algorithm, integrated into SIRIUS, identifies fragmentation motifs characteristic of certain chemical classes and thus proposes categorization by family [58]. Similarly, the MSNovelist de novo identification tool, also implemented in SIRIUS, generates candidate structures from observed fragments [65]. Finally, taking taxonomic criteria into account can significantly improve the effectiveness of proposed annotations for variables from databases [66]. These approaches, still emerging, remain to be developed with caution but constitute promising prospects.

6.5. Identification Levels

Since identification by dereplication methods is putative, chemists should interpret the results with caution, taking into account the level of confidence in the identification [18]. Currently, for confirmed identification, analysis of standards, whether commercially available or purified in the laboratory, is necessary. Structure resolution is achievable only by NMR or crystallography. In metabolomics analysis, there are five levels of compound identification, including levels 1 to 4 proposed by the Metabolomics Standards Initiative [67] and, more recently, by the Metabolomics Society working group, with the creation of an additional level 0 [53]:

Level 0: Unambiguous identification of the compound's 3D structure, including complete stereochemistry.

Level 1: Identification of the compound's 2D structure, confirmed by comparison with a standard requiring at least two independent and orthogonal methods, analyzed under identical experimental conditions (e.g., retention time and mass spectrum, retention time and NMR spectrum, exact mass and tandem mass, isotope diagram and exact mass, complete ^{1}H and/or $1^{3}C$ NMR spectra, 2D NMR).

Level 2: Putative identification of the compound by dereplication (without a reference molecule, based on physicochemical properties and/or spectral similarity with public/commercial spectral libraries).

Level 3: Putative identification of the chemical class (based on the physicochemical properties characteristic of the chemical class, or by spectral similarity with known compounds of a chemical class, which may rely on molecular networks) with identification of the molecular formula.

Level 4: Unknown compounds (although unidentified or unclassified, these metabolites can nevertheless be differentiated and quantified using spectral data).

7. Conclusions

Non-targeted approaches in metabolomics are a preferred strategy for describing living organisms. This approach has numerous applications, particularly in the field of natural products, where it can accelerate the discovery of new natural compounds. Combined with powerful chemometric tools, variables of interest can be identified, for example, according to their taxonomic affiliation or by their correlation with biological activities. Furthermore, the many dereplication tools allow for the rapid identification of already described metabolites and direct research toward potentially valuable unknown compounds. These tools have experienced exponential growth over the last 20 years, primarily due to the development of increasingly innovative algorithms that facilitate identification. These tools are of major interest for the automated interpretation of MS/MS spectra, the annotation of variables with different confidence levels, and, beyond that, the prediction of biological properties. The advent of artificial intelligence, if used judiciously by researchers, may also contribute to this transformation.

Author Contributions

Authors contributed equally to retrieving information from the published literature, designing the layout, manuscript preparation and editing. All authors have read and agreed to the published version of the manuscript.

Funding

The authors thank the Haut-de-France Region, the French Foundation of Brewery and Maltery, Junia-ISA, the French inter-branch organization of hop “Interhoublon”, the French brewery union “Syndicat des Brasseurs des Hauts-de-France” for the funding of the Ph.D. of Augustin Bildstein, as well as the CPER project “BiHauts Eco de France” (Netwhop project) for the financial support.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

Not applicable.

Acknowledgments

The authors sincerely thank Mehdi Beniddir and Samuel Bertrand for their expert insight during Augustin Bildstein's thesis defense, which enriched this article.

Conflicts of Interest

The authors declare no conflicts of interest. Given the role as Editorial Board Member, Céline Rivière had no involvement in the peer review of this paper and had no access to information regarding its peer-review process. Full responsibility for the editorial process of this paper was delegated to another editor of the journal.

Use of AI and AI-Assisted Technologies

No AI tools were used in the preparation of this paper.

References

1. Joyce, A.R.; Palsson, B.Ø. The Model Organism as a System: Integrating “omics” Data Sets. *Nat. Rev. Mol. Cell Biol.* **2006**, *7*, 198–210. https://doi.org/10.1038/nrm1857.
2. Dai, X.; Shen, L. Advances and Trends in Omics Technology Development. *Front. Med.* **2022**, *9*, 911861. https://doi.org/10.3389/fmed.2022.911861.
3. Winkler, H. *Verbreitung Und Ursache Der Parthenogenesis Im Pflanzen—Und Tierreiche*; G. FIscher: Jena, Germany, 1920; pp. 1–248.
4. Wolfender, J.-L.; Litaudon, M.; Touboul, D.; et al. Innovative Omics-Based Approaches for Prioritisation and Targeted Isolation of Natural Products—New Strategies for Drug Discovery. *Nat. Prod. Rep.* **2019**, *36*, 855–868. https://doi.org/10.1039/C9NP00004F.
5. Beniddir, A.M.; Bin Kang, K.; Genta-Jouve, G.; et al. Advances in Decomposing Complex Metabolite Mixtures Using Substructure- and Network-Based Computational Metabolomics Approaches. *Nat. Prod. Rep.* **2021**, *38*, 1967–1993. https://doi.org/10.1039/D1NP00023C.
6. Tsugawa, H.; Rai, A.; Saito, K.; et al. Metabolomics and Complementary Techniques to Investigate the Plant Phytochemical Cosmos. *Nat. Prod. Rep.* **2021**, *38*, 1729–1759. https://doi.org/10.1039/D1NP00014D.
7. Schrimpe-Rutledge, A.C.; Codreanu, S.G.; Sherrod, S.D.; et al. Untargeted Metabolomics Strategies—Challenges and Emerging Directions. *J. Am. Soc. Mass Spectrom.* **2016**, *27*, 1897–1905. https://doi.org/10.1007/s13361-016-1469-y.
8. Patti, G.J.; Yanes, O.; Siuzdak, G. Metabolomics: The Apogee of the Omics Trilogy. *Nat. Rev. Mol. Cell Biol.* **2012**, *13*, 263–269. https://doi.org/10.1038/nrm3314.

9. Wolfender, J.-L.; Nuzillard, J.-M.; van der Hooft, J.J.J.; et al. Accelerating Metabolite Identification in Natural Product Research: Toward an Ideal Combination of Liquid Chromatography-High-Resolution Tandem Mass Spectrometry and NMR Profiling, in Silico Databases, and Chemometrics. *Anal. Chem.* **2019**, *91*, 704–742. https://doi.org/10.1021/acs.analchem.8b05112.
10. Wolfender, J.-L.; Marti, G.; Thomas, A.; et al. Current Approaches and Challenges for the Metabolite Profiling of Complex Natural Extracts. *J. Chromatogr. A* **2015**, *1382*, 136–164. https://doi.org/10.1016/j.chroma.2014.10.091.
11. Sasse, M.; Rainer, M. Mass Spectrometric Methods for Non-Targeted Screening of Metabolites: A Future Perspective for the Identification of Unknown Compounds in Plant Extracts. *Separations* **2022**, *9*, 415. https://doi.org/10.3390/separations9120415.
12. Guo, J.; Huan, T. Comparison of Full-Scan, Data-Dependent, and Data-Independent Acquisition Modes in Liquid Chromatography-Mass Spectrometry Based Untargeted Metabolomics. *Anal. Chem.* **2020**, *92*, 8072–8080. https://doi.org/10.1021/acs.analchem.9b05135.
13. An Overview of the Principles of MSE, the Engine that Drives MS Performance. *Waters White Paper* (*P/N: 720004036EN)*, October 2011.
14. Waters Corporation. An Added Dimension for Metabolite ID Studies Using Ion Mobility Combined with MS^E. Available online: https://www.waters.com/nextgen/xg/fr/library/application-notes/2011/an-added-dimension-for-metabolite-id-studies-using-ion-mobility-combined-with-mse.html?srsltid=AfmBOoqB1sOe5QVdeiAX8PqOvcwfP4O6GLsVSOeckdDm9ZQsLKaUDWjj (accessed on 1 December 2025).
15. Ramos, A.E.F.; Evanno, L.; Poupon, E.; et al. Natural Products Targeting Strategies Involving Molecular Networking: Different Manners, One Goal. *Nat. Prod. Rep.* **2019**, *36*, 960–980. https://doi.org/10.1039/C9NP00006B.
16. Medina-Franco, J.L.; Sánchez-Cruz, N.; López-López, E.; et al. Progress on Open Chemoinformatic Tools for Expanding and Exploring the Chemical Space. *J. Comput. Aided. Mol. Des.* **2022**, *36*, 341–354. https://doi.org/10.1007/s10822-021-00399-1.
17. Wolfender, J.-L.; Glauser, G.; Boccard, J.; et al. MS-Based Plant Metabolomic Approaches for Biomarker Discovery. *Nat. Prod. Commun.* **2009**, *4*, 1934578X0900401019. https://doi.org/10.1177/1934578X0900401019.
18. Dunn, W.B.; Broadhurst, D.; Begley, P.; et al. Procedures for Large-Scale Metabolic Profiling of Serum and Plasma Using Gas Chromatography and Liquid Chromatography Coupled to Mass Spectrometry. *Nat. Protoc.* **2011**, *6*, 1060–1083. https://doi.org/10.1038/nprot.2011.335.
19. Schmid, R.; Heuckeroth, S.; Korf, A.; et al. Integrative Analysis of Multimodal Mass Spectrometry Data in MZmine 3. *Nat. Biotechnol.* **2023**, *41*, 447–449. https://doi.org/10.1038/s41587-023-01690-2.
20. Sturm, M.; Bertsch, A.; Gröpl, C.; et al. OpenMS—An Open-Source Software Framework for Mass Spectrometry. *BMC Bioinform.* **2008**, *9*, 163. https://doi.org/10.1186/1471-2105-9-163.
21. Röst, H.L.; Sachsenberg, T.; Aiche, S.; et al. OpenMS: A Flexible Open-Source Software Platform for Mass Spectrometry Data Analysis. *Nat. Methods* **2016**, *13*, 741–748. https://doi.org/10.1038/nmeth.3959.
22. Tautenhahn, R.; Patti, G.J.; Rinehart, D.; et al. XCMS Online: A Web-Based Platform to Process Untargeted Metabolomic Data. *Anal. Chem.* **2012**, *84*, 5035–5039. https://doi.org/10.1021/ac300698c.
23. Chambers, M.C.; Maclean, B.; Burke, R.; et al. A Cross-Platform Toolkit for Mass Spectrometry and Proteomics. *Nat. Biotechnol.* **2012**, *30*, 918–920. https://doi.org/10.1038/nbt.2377.
24. Guo, J.; Huan, T. Mechanistic Understanding of the Discrepancies between Common Peak Picking Algorithms in Liquid Chromatography—Mass Spectrometry-Based Metabolomics. *Anal. Chem.* **2023**, *95*, 5894–5902. https://doi.org/10.1021/acs.analchem.2c04887.
25. Allen, F.; Pon, A.; Greiner, R.; et al. Computational Prediction of Electron Ionization Mass Spectra to Assist in GC/MS Compound Identification. *Anal. Chem.* **2016**, *88*, 7689–7697. https://doi.org/10.1021/acs.analchem.6b01622.
26. Lange, E.; Tautenhahn, R.; Neumann, S.; et al. Critical Assessment of Alignment Procedures for LC-MS Proteomics and Metabolomics Measurements. *BMC Bioinform.* **2008**, *9*, 375. https://doi.org/10.1186/1471-2105-9-375.
27. Broadhurst, D.I.; Kell, D.B. Statistical Strategies for Avoiding False Discoveries in Metabolomics and Related Experiments. *Metabolomics* **2006**, *2*, 171–196. https://doi.org/10.1007/s11306-006-0037-z.
28. Tugizimana, F.; Piater, L.; Dubery, I. Plant Metabolomics: A New Frontier in Phytochemical Analysis. *S. Afr. J. Sci.* **2013**, *109*, 11. https://doi.org/10.1590/sajs.2013/20120005.
29. Meglen, R.R. Examining Large Databases: A Chemometric Approach Using Principal Component Analysis. *Mar. Chem.* **1992**, *39*, 217–237. https://doi.org/10.1016/0304-4203(92)90103-H.
30. van den Berg, R.A.; Hoefsloot, H.C.; Westerhuis, J.A.; et al. Centering, Scaling, and Transformations: Improving the Biological Information Content of Metabolomics Data. *BMC Genom.* **2006**, *7*, 142. https://doi.org/10.1186/1471-2164-7-142.
31. Fonville, J.M.; Richards, S.E.; Barton, R.H.; et al. The Evolution of Partial Least Squares Models and Related Chemometric Approaches in Metabonomics and Metabolic Phenotyping. *J. Chemom.* **2010**, *24*, 636–649. https://doi.org/10.1002/cem.1359.
32. Marini, F. Classification Methods in Chemometrics. *Curr. Anal. Chem.* **2010**, *6*, 72–79.
33. Bylesjö, M.; Rantalainen, M.; Cloarec, O.; et al. OPLS Discriminant Analysis: Combining the Strengths of PLS-DA and SIMCA Classification. *J. Chemom.* **2006**, *20*, 341–351. https://doi.org/10.1002/cem.1006.

34. Saraçli, S.; Doğan, N.; Doğan, İ. Comparison of Hierarchical Cluster Analysis Methods by Cophenetic Correlation. *J. Inequal. Appl.* **2013**, *2013*, 203. https://doi.org/10.1186/1029-242X-2013-203.
35. Granato, D.; Santos, J.S.; Escher, G.B.; et al. Use of Principal Component Analysis (PCA) and Hierarchical Cluster Analysis (HCA) for Multivariate Association between Bioactive Compounds and Functional Properties in Foods: A Critical Perspective. *Trends Food Sci. Technol.* **2018**, *72*, 83–90. https://doi.org/10.1016/j.tifs.2017.12.006.
36. Argüelles, M.; Benavides, C.; Fernández, I. A New Approach to the Identification of Regional Clusters: Hierarchical Clustering on Principal Components. *Appl. Econ.* **2014**, *46*, 2511–2519. https://doi.org/10.1080/00036846.2014.904491.
37. Boccard, J.; Rudaz, S. Harnessing the Complexity of Metabolomic Data with Chemometrics. *J. Chemom.* **2014**, *28*, 1–9. https://doi.org/10.1002/cem.2567.
38. Escofier, B.; Pagès, J. *Analyses Factorielles Simples et Multiples: Cours et Études de Cas*, *Sciences Sup*, 5th ed.; Dunod: Paris, France, 2023; ISBN 978-2-10-085957-3.
39. Chong, I.-G.; Jun, C.-H. Performance of Some Variable Selection Methods When Multicollinearity Is Present. *Chemom. Intell. Lab. Syst.* **2005**, *78*, 103–112. https://doi.org/10.1016/j.chemolab.2004.12.011.
40. Nothias-Scaglia, L.-F.; Esposito, M.; Costa, J.; et al. Les réseaux moléculaires, une approche bio-informatique globale pour interpréter les données de spectrométrie de masse tandem. *Spectra Anal.* **2015**, *307*, 73–78,
41. Elie, N.; Santerre, C.; Touboul, D. Generation of a Molecular Network from Electron Ionization Mass Spectrometry Data by Combining MZmine2 and MetGem Software. *Anal. Chem.* **2019**, *91*, 11489–11492. https://doi.org/10.1021/acs.analchem.9b02802.
42. Watrous, J.; Roach, P.; Alexandrov, T.; et al. Mass Spectral Molecular Networking of Living Microbial Colonies. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, E1743–E1752. https://doi.org/10.1073/pnas.1203689109.
43. Frank, A.M.; Bandeira, N.; Shen, Z.; et al. Clustering Millions of Tandem Mass Spectra. *J. Proteome Res.* **2008**, *7*, 113–122. https://doi.org/10.1021/pr070361e.
44. Shannon, P.; Markiel, A.; Ozier, O.; et al. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* **2003**, *13*, 2498–2504. https://doi.org/10.1101/gr.1239303.
45. Wang, M.; Carver, J.J.; Phelan, V.V.; et al. Sharing and Community Curation of Mass Spectrometry Data with Global Natural Products Social Molecular Networking. *Nat. Biotechnol.* **2016**, *34*, 828–837. https://doi.org/10.1038/nbt.3597.
46. Olivon, F.; Elie, N.; Grelier, G.; et al. MetGem Software for the Generation of Molecular Networks Based on the T-SNE Algorithm. *Anal. Chem.* **2018**, *90*, 13900–13908. https://doi.org/10.1021/acs.analchem.8b03099.
47. Nothias, L.-F.; Petras, D.; Schmid, R.; et al. Feature-Based Molecular Networking in the GNPS Analysis Environment. *Nat. Methods* **2020**, *17*, 905–908. https://doi.org/10.1038/s41592-020-0933-6.
48. Schmid, R.; Petras, D.; Nothias, L.-F.; et al. Ion Identity Molecular Networking for Mass Spectrometry-Based Metabolomics in the GNPS Environment. *Nat. Commun.* **2021**, *12*, 3832. https://doi.org/10.1038/s41467-021-23953-9.
49. Hubert, J.; Nuzillard, J.-M.; Renault, J.-H. Dereplication Strategies in Natural Product Research: How Many Tools and Methodologies behind the Same Concept? *Phytochem. Rev.* **2017**, *16*, 55–95. https://doi.org/10.1007/s11101-015-9448-7.
50. Bruguière, A.; Derbré, S.; Dietsch, J.; et al. MixONat, a Software for the Dereplication of Mixtures Based on ^{13}C NMR Spectroscopy. *Anal. Chem.* **2020**, *92*, 8793–8801. https://doi.org/10.1021/acs.analchem.0c00193.
51. Hubert, J.; Kotland, A.; Henes, B.; et al. Deciphering the Phytochemical Profile of an Alpine Rose (*Rhododendron ferrugineum* L.) Leaf Extract for a Better Understanding of Its Senolytic and Skin-Rejuvenation Effects. *Cosmetics* **2022**, *9*, 37. https://doi.org/10.3390/cosmetics9020037.
52. Allard, P.-M.; Péresse, T.; Bisson, J.; et al. Integration of Molecular Networking and In-Silico MS/MS Fragmentation for Natural Products Dereplication. *Anal. Chem.* **2016**, *88*, 3317–3323. https://doi.org/10.1021/acs.analchem.5b04804.
53. Blaženović, I.; Kind, T.; Ji, J.; et al. Software Tools and Approaches for Compound Identification of LC-MS/MS Data in Metabolomics. *Metabolites* **2018**, *8*, 31. https://doi.org/10.3390/metabo8020031.
54. Xing, S.; Shen, S.; Xu, B.; et al. BUDDY: Molecular Formula Discovery via Bottom-up MS/MS Interrogation. *Nat. Methods* **2023**, *20*, 881–890. https://doi.org/10.1038/s41592-023-01850-x.
55. Ruttkies, C.; Schymanski, E.L.; Wolf, S.; et al. MetFrag Relaunched: Incorporating Strategies beyond In Silico Fragmentation. *J. Cheminform.* **2016**, *8*, 3. https://doi.org/10.1186/s13321-016-0115-9.
56. White, J. PubMed 2.0. *Med. Ref. Serv. Q.* **2020**, *39*, 382–387. https://doi.org/10.1080/02763869.2020.1826228.
57. Rutz, A.; Sorokina, M.; Galgonek, J.; et al. The LOTUS Initiative for Open Knowledge Management in Natural Products Research. *eLife* **2022**, *11*, e70780. https://doi.org/10.7554/eLife.70780.
58. Dührkop, K.; Shen, H.; Meusel, M.; et al. Searching Molecular Structure Databases with Tandem Mass Spectra Using CSI:FingerID. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 12580–12585. https://doi.org/10.1073/pnas.1509788112.
59. Hoang, C.; Uritboonthai, W.; Hoang, L.; et al. Tandem Mass Spectrometry across Platforms. *Anal. Chem.* **2024**, *96*, 5478–5488. https://doi.org/10.1021/acs.analchem.3c05576.
60. Ausloos, P.; Clifton, C.L.; Lias, S.G.; et al. The Critical Evaluation of a Comprehensive Mass Spectral Library. *J. Am. Soc. Mass Spectrom.* **1999**, *10*, 287–299. https://doi.org/10.1016/S1044-0305(98)00159-7.

61. Kováts, E. Gas-Chromatographische Charakterisierung Organischer Verbindungen. Teil 1: Retentionsindices Aliphatischer Halogenide, Alkohole, Aldehyde Und Ketone. *Helv. Chim. Acta* **1958**, *41*, 1915–1932. https://doi.org/10.1002/hlca.19580410703.
62. Salem, M.A.; Perez de Souza, L.; Serag, A.; et al. Metabolomics in the Context of Plant Natural Products Research: From Sample Preparation to Metabolite Analysis. *Metabolites* **2020**, *10*, 37. https://doi.org/10.3390/metabo10010037.
63. Mejri, Y.; Cailloux, O.; Otogo N'Nang, E.; et al. MS2DECIDE: Aggregating Multiannotated Tandem Mass Spectrometry Data with Decision Theory Enhances Natural Products Prioritization. *Chem. Methods* ***2025***, *5,* e202400088. https://doi.org/10.1002/cmtd.202400088.
64. Quinlan, Z.A.; Koester, I.; Aron, A.T.; et al. ConCISE: Consensus Annotation Propagation of Ion Features in Untargeted Tandem Mass Spectrometry Combining Molecular Networking and In Silico Metabolite Structure Prediction. *Metabolites* **2022**, *12*, 1275. https://doi.org/10.3390/metabo12121275.
65. Stravs, M.A.; Dührkop, K.; Böcker, S.; et al. MSNovelist: De Novo Structure Generation from Mass Spectra. *Nat. Methods* **2022**, *19*, 865–870. https://doi.org/10.1038/s41592-022-01486-3.
66. Rutz, A.; Dounoue-Kubo, M.; Ollivier, S.; et al. Taxonomically Informed Scoring Enhances Confidence in Natural Products Annotation. *Front. Plant Sci.* **2019**, *10*, 1329. https://doi.org/10.3389/fpls.2019.01329.
67. Sumner, L.W.; Amberg, A.; Barrett, D.; et al. Proposed Minimum Reporting Standards for Chemical Analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). *Metabolomics* **2007**, *3*, 211–221. https://doi.org/10.1007/s11306-007-0082-2.