UNIVERSITY OF SALERNO

DEPARTMENT OF CHEMISTRY AND BIOLOGY

and

UNIVERSITY OF BASILICATA

DEPARTMENT OF SCIENCE

PhD in Chemistry XXXIII Cycle

CHIM/01- Analytical Chemistry

MOLECULAR CARTOGRAPHY OF THE METABOLOME OF TYPICAL FOOD PRODUCTS OF THE BASILICATA REGION BY USING HIGH RESOLUTION MASS SPECTROMETRY

PhD Coordinator

Prof. Claudio Pellecchia

Tutor

Prof. Giuliana Bianco

PhD Student

Alberto Onzo

Co-Tutor

Prof. Carmine Gaeta

Dr. Rosanna Ciriello

Academic year 2019-2020

Abbreviations

PDO	**Protected Designation of Origin**
PGI	**Protected Geographical Indication**
TSG	**Traditional Specialty Garanteed**
GI	**Geographical Indication**
EU	**European Union**
WIPO	**World Intellectual Property Organization**
MS	**Mass Spectrometry**
ESI	**Electrospray Ionization**
FT	**Fourier Transform**
HRMS	**High Resolution Mass Spectrometry**
ICR	**Ion Cyclotron Resonance**
ICP	**Inductively Coupled Plasma**
OES	**Optical Emission Spectroscopy**
NMR	**Nuclear Magnetic Resonance**
GC	**Gas Chromatography**
LC	**Liquid Chromatography**
HPLC	**High Performance Liquid Chromatography**
m/z	**Mass-to-charge ratio**
IRS	**Infrared Spectroscopy**
FWHM	**Full Width Half Maximum**
EI	**Electron Impact**

EC	**Electron Capture**
IE	**Ionization Energy**
MALDI	**Matrix Assisted Laser Desorption Ionization**
DNA	**Deoxyribonucleic acid**
Q	**Quadrupole**
QIT	**Quadrupole Ion Trap**
ToF	**Time of Flight**
LIT	**Linear Ion Trap**
RF	**Radiofrequency**
FID	**Free Induction Decay**
R	**Resolution**
S/N	**Signal-to-noise ratio**
ppm	**parts-per-million**
ppb	**parts-per-billion**
MMA	**Mass Measurement Accuracy**
RFI	**Radio-frequency Interference**
DNP	**Dictionary of Natural Products**
TMS	**Trimethylsilyl**
RDBE	**Ring-plus-double bonds equivalent**
DBE	**Double bonds equivalent**
IUPAC	**International Union of Pure and Applied Chemistry**
KMD	**Kendrick Mass Defect**
KNM	**Kendrick Nominal Mass**
NM	**Nominal Mass**
KM	**Kendrick Mass**
MdiN	**Mass Difference Network**
CRAN	**Comprehensive R Archive Network**
BioC	**Bioconductor project**
ANOVA	**Analysis of Variance**
aFT	**Absorption mode FT-ICR Mass Spectrum**
mFT	**Magnitude mode FT-ICR Mass Spectrum**
OIFA	**Omics Interactive Formula Assignment**

Abstract

Protected Designation of Origin (PDO) and Protected Geographical Indication (PGI) are, together with Traditional Speciality Guaranteed (TSG), the instruments created by the European Union (EU) to protect Geographical Indications (GIs) within the European framework as indications which identify a good as originating in a specific location, where a given quality, reputation or other characteristic of the good is essentially attributable to its geographical origin. Food products with a protected geographical status distinguish from other similar products of the same category for the link with the region they originate from. Despite the improvement this quality scheme provided to the protection of unique foodstuff, the threat of food fraud is still present and sophistication of adulteration of food products is making the utilization of the most advanced technologies compulsory for labelled food product protection. Mass spectral characterization of food materials has advanced rapidly in the past few years, mostly due to the development and now routine availability of electrospray ionization (ESI). However, it is now clear that food products exist as complex mixtures and High resolution Electrospray Ionization Fourier transform—Ion Cyclotron Resonance Mass Spectrometry (ESI FT-ICR MS) at high magnetic fields is currently a techniques capable of resolving thousands of individual molecules in few minutes. In this work, a Mass Spectrometry-based phytochemical screening was performed on several traditional food products produced in the Basilicata region (Italy) labelled with geographical indication marks of quality. High Resolution ESI-FT-ICR MS data obtained from food sample analyses were used to perform a rapid evaluation of metabolome by converting accurate m/z values in putative elemental formulas. Molecular formula maps, or *molecular fingerprints*, were obtained by making 2D Van Krevelen plots, that lead to a direct identification of different classes of metabolites. The presence of important metabolite classes, i.e. fatty acid derivatives, tannins, amino acids and peptides, carbohydrates and polyphenolic derivatives, was assessed. Moreover, differences among Van Krevelen plots could be noticed from their direct comparison, thus reflecting differences in promoted biochemical pathways and suggesting the presence of biomarkers, that can eventually be identified by a target approach. Thus, molecular fingerprints prove to be an innovative tool that could be useful for food authentication and traceability.

Index

1. Introduction

Geographical indications (GIs) are a key economic asset for the EU, and Italy has 876 GI-protected products exalting the agricultural sector, promoting worldwide the Italian culture of "well-eating and well-being" [1]. EU quality schemes aim at protecting the names of specific products to promote their unique characteristics, linked to their geographical origin as well as know-how of the region and were develop to meet the increasing consumer demand for safe and high quality products, understanding quality as the sum of features, characteristics, and properties of a product, which bear on its ability to satisfy stated or implied needs. The World Intellectual Property Organization (WIPO) defines a geographical indication (GI) as "a sign used on goods that have a specific geographical origin and possess qualities, a reputation, or characteristics that are essentially attributable to that origin" [2]. At present, no "worldwide" GI right exists. As explained by the WIPO, "intellectual property rights are governed by the 'territoriality principle'. The effects of a right obtained in a particular jurisdiction are limited to the territory of that jurisdiction. At the end of the 20th century, the European Union (EU) recognized and supported the potential of differentiating quality products on a regional basis. The first regulation on geographical indications was adopted in the EU in 1992, to harmonize diverse protection instruments existing in some Member States and to create a system of registration and protection of names compatible with the single common market. Since then, EU law lays down stringent requirements guaranteeing the standards of all European products. The distinctiveness of protected productions is ensured by dedicated consortia, which state rigorous production regulations. Technical reports clearly describing the link with the territory, intended as the correlation between the geographical area and the characteristics of the product, and truthful information for guarantee rights and awareness of consumers should be provided [3]. The EU has created three labels regarding Protected Designation of Origin (PDO), Protected Geographical Indication (PGI), and Traditional Specialty Guarantee (TSG). Those schemes encourage diverse agricultural production, protect product names from misuse and imitation, and help consumers by giving them information concerning the specific character of the labelled products. PDO covers agricultural products and foodstuffs which are produced, processed, and prepared in a given geographical area (a specific place, region, or, in

exceptional cases, a country) using recognized know-how, whose quality or characteristics are essentially or exclusively due to a particular geographical environment with its inherent natural and human factors and the production steps which all take place in the defined geographical area. PGI protects agricultural products and foodstuffs closely linked to the geographical area (a specific place, region, or country), whose given quality, reputation, or other characteristics are essentially attributable to its geographical origin, and at least one of the stages of production, processing, or preparation takes place in the defined geographical area. TSG indicates the traditional character of food, either in the composition or means of production. It describes a specific product or foodstuff that results from a mode of production, processing, or composition corresponding to traditional practice for that product or foodstuff or is produced from raw materials or ingredients than are those traditionally used. GI work as product differentiators on the market by enabling consumers to distinguish between products with geographical origin-based characteristics and others without those characteristics. Usually, the price of such GI products is higher than those without GI, as it is necessary to have accurate methods for distinguishing them because financial incentives continue to drive retailers/resellers to misidentify the geographic origin of commodities and food products. The use of analytical techniques to determine the GI of food products is the best way to avoid adulterations and mislabelling after the incorporation of food to the market [4]. Strategies employed to detect adulterated or mislabelled products have relied on instrumental techniques mainly because of the sophistication of fraudulent procedures. Since the content of selected minerals and trace elements clearly reflects the soil type of a cultivation area and the environmental growing conditions for food productions [5], evaluation of trace element content has been proposed as one of the selected approaches to assure the geographical origin of food samples. In this context, techniques like Inductively Coupled Plasma Optical Emission spectroscopy (ICP–OES) and Mass Spectrometry (ICP-MS) were employed extensively since up to 60 elements can be screened per sample run in less than one minute and trace element composition can be determined in a variety of aqueous or organic matrices [6]. However, the main drawbacks of ICP-MS and ICP-OES are the expensive instrumentation and operation costs, the requirement for trained operators, and in most cases, the need of sample pre-treatment steps, which frequently includes the complete mineralization of samples. Apart from element analysis, screening and quantification of organic compounds was employed for food authentication and traceability, providing satisfying results. In this context, *metabolomics* played a crucial role in the protection of labelled food products, since through this approach multiple metabolites with features of discrimination and prediction were identified and quantified [7]. Spectroscopic techniques like UV/Vis and Near and Middle Infrared Spectroscopy were employed to examine global parameters or indexes which depend on the content of a single molecule or a specific metabolite family, such as the anisidine value in the quality of edible oils, being defined as the absorbance of a solution of a fat sample containing aldehydes which have reacted with p-anisidine, the process of fat deterioration by the peroxide value, or the general color, determined by the saturation of chlorophyll or carotenoid pigments, among others [8]. On the other hand, ^{1}H NMR is a suitable way to simultaneously determine multiple compounds belonging to specific metabolite classes. For instance, profiles generated using ^{1}H NMR, together with the utilization of multivariate statistical techniques, were employed to characterize Corsican honey [9].

Despite the advantages provided by the utilization of spectroscopic techniques, such as completeness of information in single spectra, several drawbacks, like low sensitivity and specificity, hamper the use of these methods for routine analysis. Separation techniques, like gas and liquid chromatography (GC and LC, respectively), provide higher levels of sensitivities and specificities in organic molecule analysis, leading to the separation of single components of a complex matrix. Moreover, coupling with Mass Spectrometry makes possible the obtainment of important structural information to shed light on analyte structures. These advantages allow to identify low concentration key markers unique for a certain labelled food product and to obtain important clues on their identity. For instance, separation and detection of fatty acids and triacylglycerols, sterols, and aroma have been used for the authentication of food [10–12]. LC has been extensively used in food analysis for measuring numerous compounds, e.g. carbohydrates, vitamins, additives, mycotoxins, amino acids, proteins, triglycerides, lipids, chiral compounds, and pigments [13]. On the other hand, GC is one of the most universal separation techniques used in food analysis, mainly for volatile and semivolatile composition studies, aromas, and pesticides [14]. Full scan LC and GC-MS analyses provided complex chromatograms that turned out to be very suitable as molecular *fingerprints*, i.e. unique patterns of metabolites that can be compared among each other for food authentication purposes [15]. Thus, utilization of hyphenated techniques could be thought as a milestone for metabolomic analysis and labelled food protection. However, these analyses are really time-consuming, with related analysis times that could span from 30-40 minutes to hours, and cost demanding, needing high volumes of organic solvents to be employed [13]. For these reasons, during the last years, High-resolution (HR) Mass Spectrometry is becoming the technique of choice for sensitive and selective analysis of both known and unknown compounds. Outstanding levels of accuracy and resolution led to unequivocal identification of molecules, starting from their chemical formulas. Moreover, simple direct injection full scan HRMS spectra can be obtained in few seconds and could be used as molecular fingerprints since high sensitivity and resolution allows to distinguish potential markers without any need of prior separation [15–17]. These benefits could be provided even if sample treatment steps are reduced to a minimum. HRMS data could be used to trace molecular cartography of foodstuffs, i.e. real molecular maps rich in information about molecules proposed as unique marks for a specific food product, data obtained from various techniques must be transformed into suitable descriptors. The molecular annotation network obtained is then converted by using dedicated algorithms, into molecular maps, e.g. Van Krevelen diagrams recently used to define the metabolic fingerprint of foodstuffs [15,18]. Overall, HRMS can be thought as a promising technique to characterize food matrices and to expand the range of powerful analytical techniques to be used for quality assurance and protection of labelled foodstuffs.

1.1. References

[1] Regolamento di esecuzione (UE) n. 668/2014 della Commissione, del 13 giugno 2014 , recante modalità di applicazione del regolamento (UE) n. 1151/2012 del Parlamento europeo e del Consiglio sui regimi di qualità dei prodotti agricoli e alimentari, Italy, 2014.

[2] WIPO, World Intellectual Property Organization, (2015). http://www.wipo.int/%0Ageo_indications/en/faq_geographicalindications.html.

[3] EU Council, Protection of geographical indications and designations of origin for agricultural products and foodstuffs, 1992.

[4] M. de la Guardia, A.G. Illueca, Food Protected Designation of Origin: Methodologies and Applications, Elsevier, 2013.

[5] S.A. Drivelos, C.A. Georgiou, Multi-element and multi-isotope-ratio analysis to determine the geographical origin of foods in the European Union, TrAC - Trends Anal. Chem. 40 (2012) 38–51. https://doi.org/10.1016/j.trac.2012.08.003.

[6] M. Iglesias, E. Besalú, E. Anticó, Internal standardization-atomic spectrometry and geographical pattern recognition techniques for the multielement analysis and classification of Catalonian red wines, J. Agric. Food Chem. 55 (2007) 219–225. https://doi.org/10.1021/jf0629585.

[7] Y. Picó, Chemical Analysis of Food: Techniques and Applications, Elsevier Inc., 2012. https://doi.org/10.1016/C2010-0-64808-5.

[8] A. Nawrocka, J. Lamorsk, Determination of Food Quality by Using Spectroscopic Methods, in: Adv. Agrophysical Res., InTech, 2013. https://doi.org/10.5772/52722.

[9] J.A. Donarski, S.A. Jones, A.J. Charlton, Application of cryoprobe 1H nuclear magnetic resonance spectroscopy and multivariate analysis for the verification of corsican honey, J. Agric. Food Chem. 56 (2008) 5451–5456. https://doi.org/10.1021/jf072402x.

[10] F.J. Delgado, J. González-Crespo, R. Cava, R. Ramírez, Formation of the aroma of a raw goat milk cheese during maturation analysed by SPME-GC-MS, Food Chem. 129 (2011) 1156–1163. https://doi.org/10.1016/j.foodchem.2011.05.096.

[11] S.S. Cunha, J.O. Fernandes, M.B.P.P. Oliveira, Quantification of free and esterified sterols in Portuguese olive oils by solid-phase extraction and gas chromatography-mass spectrometry, J. Chromatogr. A. 1128 (2006) 220–227. https://doi.org/10.1016/j.chroma.2006.06.039.

[12] F. Aranda, S. Gómez-Alonso, R.M. Rivera Del Álamo, M.D. Salvador, G. Fregapane, Triglyceride, total and 2-position fatty acid composition of Cornicabra virgin olive oil: Comparison with other Spanish cultivars, Food Chem. 86 (2004) 485–492. https://doi.org/10.1016/j.foodchem.2003.09.021.

[13] V. Di Stefano, G. Avellone, D. Bongiorno, V. Cunsolo, V. Muccilli, S. Sforza, A. Dossena, L. Drahos, K. Vékey, Applications of liquid chromatography-mass spectrometry for food analysis, J. Chromatogr. A. 1259 (2012) 74–85. https://doi.org/10.1016/j.chroma.2012.04.023.

[14] S.J. Lehotay, J. Hajšlová, Application of gas chromatography in food analysis, TrAC - Trends Anal. Chem. 21 (2002) 686–697. https://doi.org/10.1016/S0165-9936(02)00805-1.

[15] R. Pascale, G. Bianco, T.R.I. Cataldi, P.S. Kopplin, F. Bosco, L. Vignola, J. Uhl, M. Lucio, L. Milella, Mass spectrometry-based phytochemical screening for hypoglycemic activity of Fagioli di Sarconi beans (Phaseolus vulgaris L.), Food Chem. 242 (2018) 497–504. https://doi.org/10.1016/j.foodchem.2017.09.091.

[16] R.D. Gougeon, M. Lucio, L. Boutegrabet, D. Peyron, F. Feuillat, D. Chassagne, H. Alexandre, A. Voilley, P. Cayot, I. Gebefügi, N. Hertkorn, P. Schmitt-Kopplin, Authentication approach of the chemodiversity of grape and wine by FTICR-MS, in: ACS Symp. Ser., American Chemical Society, 2011: pp. 69–88. https://doi.org/10.1021/bk-2011-1081.ch005.

[17] A. Santarsiero, A. Onzo, R. Pascale, M.A. Acquavia, M. Coviello, P. Convertini, S. Todisco, M. Marsico, C. Pifano, P. Iannece, C. Gaeta, S. D'Angelo, M.C. Padula, G. Bianco, V. Infantino, G. Martelli, Pistacia lentiscus Hydrosol: Untargeted Metabolomic Analysis and Anti-Inflammatory Activity Mediated by NF-κB and the Citrate Pathway, Oxid. Med. Cell. Longev. 2020 (2020) 1–14. https://doi.org/10.1155/2020/4264815.

[18] A. Onzo, G. Bianco, R. Pascale, P. Iannece, C. Gaeta, Molecular Fingerprinting of traditional food products by ultra-high resolution ESI-FT-ICR Mass Spectrometry, in: XXVIII Congr. Anal. Chem. Div., Bari, Italy, 2019.

2. Methodologies

Mass spectrometry is an indispensable analytical tool in chemistry, biochemistry, pharmacy, medicine, and many related fields of science [1–9]. Mass spectrometry (MS) is employed to analyze combinatorial libraries [10,11], sequence biomolecules [12], and help explore single cells [13] or objects from outer space [14]. Structure elucidation of unknown substances, environmental and forensic analytes, quality control of drugs, foods, and polymers all rely to a great extent on mass spectrometry [2,3,8,15,16]. The information delivered by mass alone can be sufficient for the identification of elements and the determination of the molecular formula of an analyte. The relative abundance of isotopologues helps to establish which elements contribute to such a formula and to estimate the number of atoms of a contributing element. Under the conditions of certain mass spectrometric experiments, fragmentation of ions can deliver information on ionic structure. Thus, MS elucidates the connectivity of atoms within smaller molecules, identifies functional groups, determines the (average) number and eventually the sequence of constituents of macromolecules, and in some cases even yields their three-dimensional structure [17]. The basic principle of mass spectrometry (MS) is to generate ions from either inorganic or organic compounds by any suitable method, to separate these ions by their mass-to-charge ratio (*m/z*) and to detect them qualitatively and quantitatively by their respective *m/z* and abundance [17,18]. The analyte may be ionized thermally, by electric fields or by impacting energetic electrons, ions or photons. The large variety of ionization techniques and their key applications can be roughly classified by their relative hardness or softness and (molecular) mass of suitable analytes [19]. A mass spectrometer consists of an ion source, a mass analyser, and a detector which are operated under high vacuum conditions. A closer look at the front end of such a device might separate the steps of sample introduction, evaporation, and successive ionization or desorption/ionization, respectively, but it is not always trivial to identify each of these steps as clearly separated from each other [17–20]. Nowadays, the instrument is always coupled with a data system which collects and processes data from the detector. Whereas other spectroscopic methods such as nuclear magnetic resonance (NMR), infrared (IRS) or Raman spectroscopy do allow for sample recovery, mass spectrometry is destructive, i.e., it consumes the analyte [17]. This is apparent from the process of ionization and translational motion through the mass analyser to the detector during analysis. Although some sample is

consumed, it may still be regarded as practically non-destructive, because the amount of analyte needed is in the low microgram range or even by several orders of magnitude below [17]. In turn, the extremely low sample consumption of mass spectrometry makes it the method of choice when most other analytical techniques fail because they are not able to yield analytical information from nanogram amounts of sample [21]. What a Mass Spectrometry analysis returns is a so-called Mass Spectrum, a two-dimensional representation of signal intensity (ordinate) versus mass-to-charge ratios (*m/z*, abscissa). The position of a peak, as signals are usually called, reflects the *m/z* of an ion that has been created from the analyte within the ion source [17,18,20]. The intensity of this peak correlates to the abundance of that ion. Often but not necessarily, the peak at highest *m/z* results from the detection of the intact ionized molecule, the molecular ion, M+• [17]. The molecular ion peak is usually accompanied by several peaks at lower *m/z* caused by fragmentation of the molecular ion to yield fragment ions. Consequently, the respective peaks in the mass spectrum may be referred to as fragment ion peaks [22]. The most intense peak of a mass spectrum is called base peak. In most representations of mass spectral data the intensity of the base peak is normalized to 100% relative intensity. This largely helps to make mass spectra more easily comparable. The normalization can be done because the relative intensities are basically independent from the absolute ion abundances registered by the detector. Usually, spectra are represented as a bar graphs or histograms. Such data reduction is common in mass spectrometry and useful as long as peaks are well resolved. The intensities of the peaks can be obtained either from measured peak heights and the position of the signal, i.e., the *m/z* ratio, is determined from its centroid [17].

2.1. Important features of Mass Spectra

2.1.1. Isotopic Patterns

As long as we are dealing with low molecular mass ranges, it is possible to separate ions which differ by 1 Da in mass. The upper mass limit for their separation depends on the resolution of the instrument employed. Consequently, the isotopic composition of the analyte is directly reflected in the mass spectrum – it can be regarded as an elemental fingerprint [23,24]. Even if the analyte is chemically perfectly pure it represents a mixture of different isotopic compositions, provided it is not composed of monoisotopic elements only. Therefore, a mass spectrum normally superimposes the mass spectra of all isotopic species involved [23,24]. The set of peaks relative to the same species with different isotopic content, known as *isotopologues*, is called *isotopic pattern* or *distribution*. While it may seem, at the first glance, to complicate the interpretation of mass spectra, isotopic patterns are in fact an ideal source of analytical information, since relative intensity of isotopic pattern peaks reflects elemental composition of related species [25]. For example, let's consider a species whose chemical formula comprises only carbon atoms. As it can be noticed in **Figure 1**, the isotopic pattern changes by rising the carbon atom number:

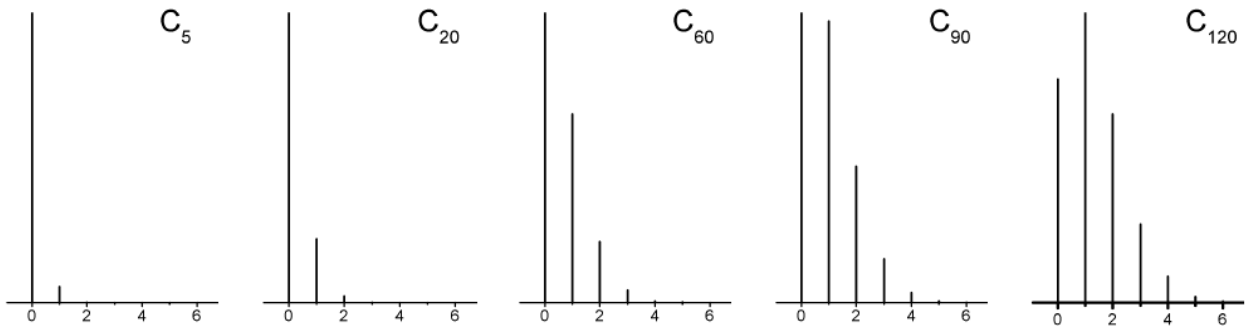

Figure 1 *Simulated isotopic pattern (calculated intensity of MS peaks versus the number of* ^{13}C *atoms) of molecules with a different number of carbon atoms (adapted from* [17]).

In a more general way, carbon consists of ^{13}C and ^{12}C in a ratio r that can be written as $r = c/(100 - c)$ where c is the abundance of ^{13}C. Then, the probability to have only ^{12}C in a molecular ion M consisting of w carbons, i.e., the probability of monoisotopic ions P_M is given by [23]:

$$P_M = \left(\frac{100 - c}{100}\right)^w$$

The probability of having exactly one ^{13}C atom in an ion with w carbon atoms is therefore:

$$P_{M+1} = w\left(\frac{c}{100 - c}\right)\left(\frac{100 - c}{100}\right)^w$$

and the ratio P_{M+1}/P_M is given as:

$$\frac{P_{M+1}}{P_M} = w\left(\frac{c}{100 - c}\right)$$

The ratio P_{M+1}/P_M reflects the one between the isotopologues M and M+1. Thus, by looking at relative intensity ratios, one can deduce the number of carbon atoms of the analysed species.

As can be noticed, the equation to calculate probability of M+2, M+3, … isotopologues turns to higher level of complexity. Moreover, with a lower number of carbon atoms, M+2, M+3, … isotopologue intensities become lower and negligible, reflecting the very low probability related to extreme combinations.

The presence of other di-or multi-isotopic elements, i.e. elements with more than one abundant isotope, together with high atom counts make the theoretical isotopic pattern calculation very complex, since one equation should be solved for every peak. Thus, two new methods were developed to simplify the process and are commonly used nowadays to obtain theoretical isotopic patterns given a specific chemical formula, i.e. the *binomial* and the *polynomial approaches* [23,26]. The former is used in case of the presence of a di-isotopic element into the chemical formula, like carbon, and consists in calculating isotopologues relative intensities as single terms of the expression $(a + b)^n$, where a and b are the isotopic abundances and n is the total atom count.

The latter, instead, is useful for the calculation of isotopic distributions of polyisotopic elements or for formulas composed of several non-monoisotopic elements. In general, the isotopic distribution of a molecule can be described by a product of polynominals:

$$(a_1 + a_2 + a_3 + \cdots + a_x)^n \left(b_1 + b_2 + b_3 + \cdots + b_y\right)^m (c_1 + c_2 + c_3 + \cdots + c_z)^l \ldots.$$

where letters *a*, *b*, *c*, … correspond to elements, underscores relate to isotopes and *n*, *m*, *l*, … to the total atom counts. Once resolved the expression, obtained terms correspond to isotopologues relative intensities, just like for the binomial approach. Of course, with higher amount of multi-isotopic element and total atom counts, simplification of related polynomial expression is impossible by hand. For this, many computer software exist that are able to calculate very complex isotopic patterns in few milliseconds [26].

Calculation of isotopic patterns is very important in *m/z* formula assignment. Indeed, matching of theoretical and observed isotopic patterns reflects the same elemental composition, thus allowing filtering of incorrect formula candidates [25].

2.1.2. Resolution and Mass Accuracy

The separation observed in a mass spectrum is termed mass resolution, R, or simply resolution. Mass resolution is given as the smallest difference in *m/z* (*Δm/z*) that can be separated for a given signal, i.e., at a given *m/z* value:

$$R = \frac{m}{\Delta m} = \frac{m/z}{\Delta m/z}$$

Accordingly, resolution is dimensionless [27,28]. The ability of an instrument to resolve neighbouring peaks is called its *mass resolving power* or simply *resolving power*. It is obtained from the peak width at a specific percentage of the peak height expressed as a function of mass, as it is shown in **Figure 2**:

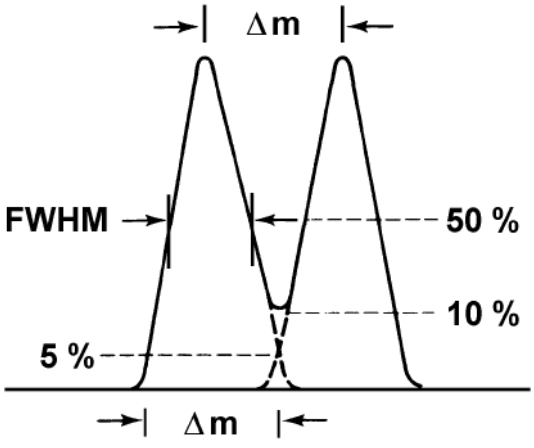

Figure 2 *The 10% valley and full width at half maximum (FWHM) definitions of resolution (adapted from* [17]).

Two neighbouring peaks are assumed to be sufficiently separated when the valley separating their maxima has decreased to 10% of their intensity [17,27]. Hence, this is known as 10% valley definition of resolution, $R_{10\%}$. The 10% valley conditions are fulfilled if the peak width at 5% relative height equals the mass difference of the corresponding ions, because then the 5% contribution of each peak to the same point of the *m/z* axis adds up to 10%. With the advent of linear quadrupole analysers, the full width at half maximum (FWHM) definition of resolution became widespread especially among instruments manufacturers. In principle, resolution is always determined from the peak width of some signal at a certain relative height and therefore, any peak can serve this purpose. Increasing resolution does not affect the relative intensities of the peaks, but increased settings of resolving power are usually obtained at the cost of transmission of the analyser, thereby reducing the absolute signal intensity [27].

Resolution is very close to another important aspect related to mass spectra, i.e. the *mass measurement accuracy*. In detail, the *absolute mass accuracy*, $\Delta m/z$, is defined as the difference between measured mass and *calculated exact mass*, i.e. the sum of monoisotopic masses of every atom of a chemical species [28]:

$$\Delta m/z = m/z_{Observed} - m/z_{Exact}$$

In general, mass accuracy is reported as *relative accuracy* dividing the absolute accuracy by the calculated exact mass. This quantity is thus expressed in parts-per-million (ppm). Accuracy describes the deviation of the experimental value from the true value, however, in practice, one never deals with exact values, but to reference ones. Accuracy is high if the values from several measurements are close to the reference value.

Higher mass measurement accuracies lead to the possibility to assign chemical formulas to observed ionic species [25,29,30]. With infinite accuracy, for example, one is able to assign a unique formula to each observed MS signal, in order to obtain important chemical information about identified analytes, like the presence of unsaturations and/or heteroatoms. In practice, it's highly probable to deal with errors in the order of several ppm, depending on the utilized instrument, thus improving the number of possible formula candidate per MS signal [25].

High resolution and accurate mass measurements are closely related and depend on each other, because mass accuracy tends to improve as peak resolution is improved. Nevertheless, they should not be confused, as performing a measurement at high resolution alone does not equally imply measuring the accurate mass [28].

2.2. Ionization: fundamentals and common techniques

Mass analyser of any mass spectrometer can only handle charged species, i.e., ions that have been created from atoms or molecules, occasionally also from radicals, zwitterions or clusters [19]. It is the task of the ion source to perform this crucial step and there is a wide range of ionization methods in use to achieve this goal for the whole variety of analytes. The classical procedure of ionization involves shooting energetic electrons on a gaseous neutral. This is called *electron ionization* (EI). Electron ionization has formerly been termed *electron*

impact ionization or simply *electron impact* (EI) [31,32]. When a neutral is hit by an energetic electron carrying several tens of electronvolts (eV) of kinetic energy, some of the energy of the electron is transferred to the neutral. If the electron, in terms of energy transfer, collides very effectively with the neutral, the energy transferred can exceed the ionization energy (IE) of the neutral. Then, from the mass spectrometric point of view, the most desirable process can occur, i.e. ionization by ejection of one electron generating a molecular ion, a positive radical ion [31]:

$$M + e^- \rightarrow M^{+\bullet} + 2e^-$$

Depending on the analyte and on the energy of the primary electrons, doubly charged and even triply charged ions may be observed:

$$M + e^- \rightarrow M^{2+} + 3e^-$$

$$M + e^- \rightarrow M^{3+\bullet} + 4e^-$$

While the doubly charged ion, M^{2+}, is an even-electron ion, the triply charged ion, $M^{3+\bullet}$, again is an odd-electron ion. In addition to the desired generation of molecular ions, several other events can result from electron-neutral interactions [17,19,31]. A less effective interaction brings the neutral into an electronically excited state without ionizing it. As the energy of the primary electrons increases, the abundance and variety of the ionized species will also increase, i.e., electron ionization may occur via different channels, each of which gives rise to characteristic ionized and neutral products. This includes the production of the following type of ions: molecular ions, fragment ions, multiply charged ions, metastable ions, rearrangement ions, and ion pairs [33,34]. The electron could also be captured by the neutral to form a negative radical ion. However, electron capture (EC) is rather unlikely to occur with electrons of 70 eV since EC is a resonance process because no electron is produced to carry away the excess energy [35]. Thus, EC only proceeds effectively with electrons of very low energy, preferably with thermal electrons. It is obvious that ionization of the neutral can only occur when the energy deposited by the electron-neutral collision is equal to or greater than the ionization energy (IE) of the corresponding neutral, defined as the minimum amount of energy that needs to be absorbed by an atom or molecule in its electronic and vibrational ground states in order to form an ion that is also in its ground states by ejection of an electron [31,33,34]. Ionization energies of most molecules are in the range of 7–15 eV. Removal of an electron from a molecule can formally be considered to occur at a σ-bond, a π-bond, or at a lone electron pair with the σ-bond being the least favored and the lone electron pair being the most favored position for charge localization within the molecule, an assumption directly reflected in the IEs. The more atoms are contained within a molecule the easier it finds a way for stabilization of the charge, e.g., by delocalization or hyperconjugation. Once the molecular ion is formed, the electron charge is never really localized in a single orbital, although assuming so is often a good working hypothesis for mass spectral interpretation [31,33]. The ionization energy represents the absolute minimum energy required for ionization of the neutral concerned. This means in turn that in order to effect ionization, the impacting electrons need to carry at least this amount of energy. If this energy were then to be quantitatively transferred during the collision, ionization would take place. Obviously, such an event is of rather low probability and therefore, the ionization

efficiency is close to zero with electrons carrying just the IE of the pertinent neutral. However, a slight increase in electron energy brings about a steady increase in ionization efficiency. Strictly speaking, every molecular species has an ionization efficiency curve of its own. Fortunately, the curves of ionization cross section vs. electron energy are all of the same type, exhibiting a maximum at electron energies around 70 eV [17,31,33].

EI is suitable for the analysis of low molecular weight molecules, providing important structural information thanks to fragmentation which occurs into the ionization source moderately. Thanks to these features, it's usually coupled with Gas Chromatography (GC), allowing the qualitative and quantitative analysis of volatile compounds after a separation step [36,37]. However, fragmentation occurs in a too high extend for higher molecular weight molecules, making this ionization technique not suitable for their analysis. Thus, softer ionization techniques were introduced, such as Chemical Ionization (CI). In chemical ionization, new ionized species are formed when gaseous molecules interact with ions, i.e., chemical ionization is based on ion–molecule reactions. Chemical ionization may involve the transfer of an electron, proton, or other ions between the reactants [19,38]. These reactants are the neutral analyte M and ions from a reagent gas. In CI, bimolecular processes are used to generate analyte ions. There are four general pathways of positive-ion formation from a neutral analyte molecule M:

$$M + [BH]^+ \rightarrow [M+H]^+ + B \text{ proton transfer}$$

$$M + X^+ \rightarrow [M+X]^+ \text{ electrophilic addition}$$

$$M + X^+ \rightarrow [M-A]^+ + AX \text{ anion abstraction}$$

$$M + X^{+\bullet} \rightarrow M^{+\bullet} + X \text{ charge exchange}$$

CI ion sources exhibit close similarity to EI ion sources. Indeed, modern EI ion sources can usually be switched to CI operation in seconds, i.e. they are constructed as EI/CI combination ion sources. In any CI plasma, ions of both polarities, positive and negative, are formed simultaneously, e.g. $[M+H]^+$ and $[M–H]^–$ ions, and it is just a matter of the polarity of the acceleration voltage which ions are extracted from the ion source [38]. Thus, negative-ion chemical ionization (NICI) mass spectra are readily obtained when one of the following processes occurs:

$$M \rightarrow [M-H]^- + H^+ \text{ dissociation of acids}$$

$$M + A^- \rightarrow [M+A]^- \text{ nucleophilic addition}$$

$$M + e^- \rightarrow [M-B]^- + B^+ + e^- \text{ ion-pair formation}$$

Similar to EI ionization, this method poses some limitations in terms of mass range (<1000) and requires specific sample characteristics with regard to thermal stability and volatility. CI is, however, better than EI with respect to the production of the molecular ion. Nevertheless, both EI and CI were not capable of ionizing the most valuable, thermally instable, polar biological compounds. Subsequently, additional soft ionization methods were developed and replaced older techniques. These include fast atom bombardment (FAB), liquid secondary ion mass spectrometry (LSIMS), matrix-assisted laser desorption ionization (MALDI), and electrospray ionization (ESI) [39–42]. Remarkably, the latter two ionization techniques have revolutionized

the usage of mass spectrometers and enabled researchers to easily study biological substances, such as glycoconjugates, proteins, and DNA [39,40,43]. Development of electrospray ionization started with the work of Dole and co-workers, who successfully introduced a polystyrene polymer (average MW = 51,000 Da) into the gas phase as a charged species [44]. Surprisingly, this ionization technique is by far one of the simplest to understand. Samples are usually dissolved in a buffer or solvent that is introduced into the mass spectrometer in the form of a spray. In ESI-MS, the sample should be soluble in a preferably polar solvent, which can be infused, under atmospheric pressure, into the ionization source via a thin needle. As the sample is being constantly sprayed, a high electrical potential is applied at the needle (3 – 4 kV), resulting in the formation of highly charged droplets (i.e., nebulization). These droplets are then driven electrically and are vaporized with the aid of a warm neutral gas (usually nitrogen). A schematic representation of the ESI source is shown in **Figure 3**:

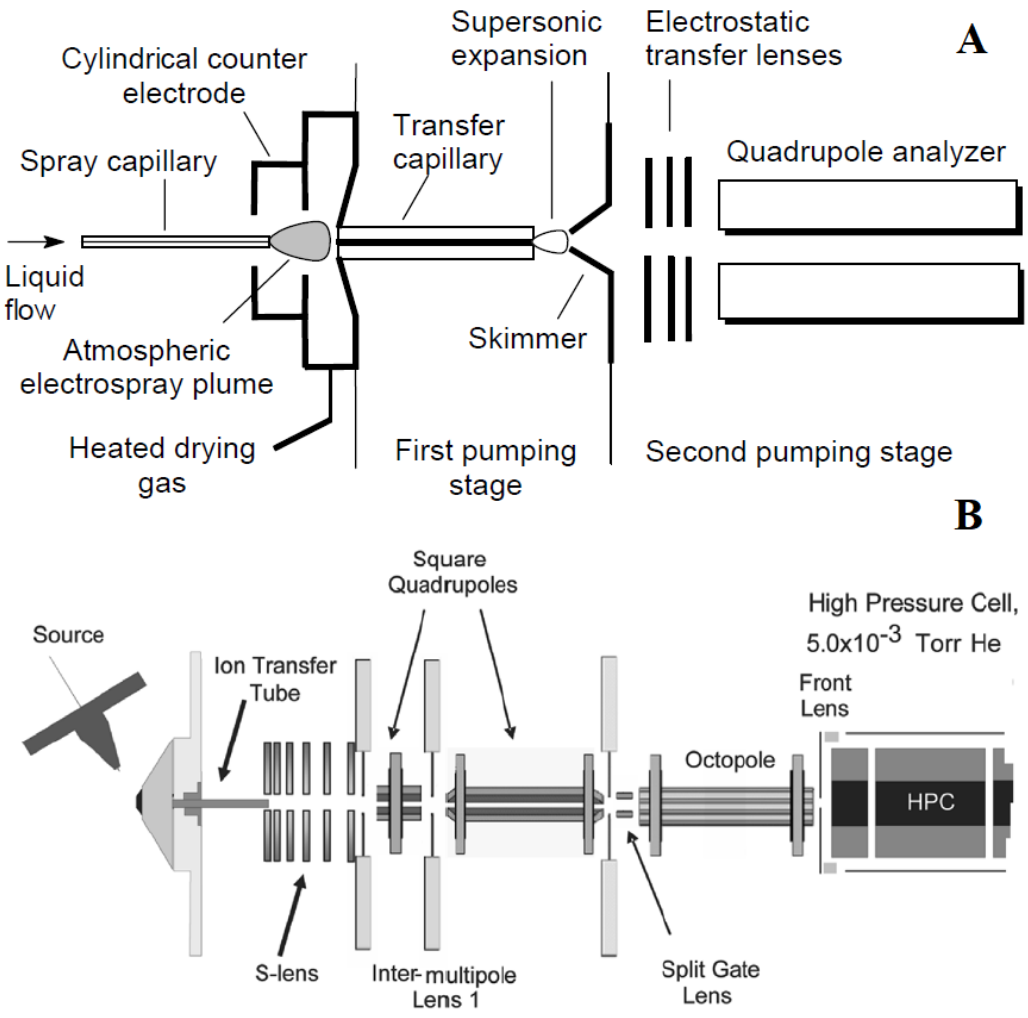

Figure 3 *Representation of an early electrospray ionization (ESI) source (A). In more recent configurations (B), the spray capillary is set at a 45° angle to the transfer capillary, preventing the clogging of capillaries and skimmers caused by the deposition of non-volatile impurities, such as buffer salts (adapted from* [17]).

Under these conditions, the droplets break down and, while shifting inside the source, their size is continuously being reduced. Eventually, the repulsive forces, also termed the coulombic forces, among the ions on the surface of the shrinking droplets become very high. These forces will ultimately exceed the surface tension of the solvent, resulting in ions that desorb into the gas phase. This theory of ESI ion formation is termed the *ion evaporation method* [45] and is believed to favour ions with relatively low *m/z* values. An alternative theory, which is supposed to be dominant in the case of ions with very high *m/z*, is the charge residue model [45], which involves continuous evaporation of the solvent accompanied by droplet fragmentation so that a single ion (probably multiply charged) is formed at the end of this process (i.e., solvent is completely evaporated). $[M+H]^+$ and $[M-H]^-$ are mainly produced during the process. However, the formation of other types of ions could take place, such as clusters, multi-charged ions (resulting from the gain or loss of more than one proton) or metal adducts [20,39,43]. In MALDI, ions are desorbed from the solid phase [40]. A sample is first dissolved in a suitable solvent and mixed with an excessive amount of an appropriate matrix. Subsequently, it is spotted on a MALDI plate and air-dried (or under a stream of nitrogen gas). Under these circumstances, the sample is co-crystallized with the matrix. The components in the mixture are brought into the gas phase via a laser beam (usually a nitrogen laser at a wavelength of 337 nm) that hits the sample-matrix crystal, leading to absorption of the laser energy by the matrix and subsequent desorption and ionization of the analytes in the sample, as it is shown in **Figure 4**:

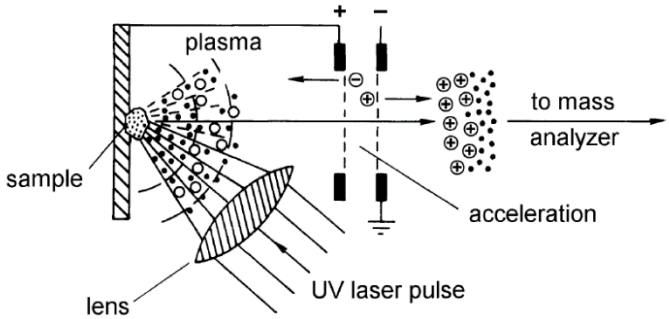

Figure 4 *Schematic representation of MALDI process (adapted from* [17]).

The mechanisms of ion formation in MALDI are a subject of continuing research [46]. The major concerns are the relationship between ion yield and laser fluence, the temporal evolution of the desorption process and its implications upon ion formation, the initial velocity of the desorbing ions, and the question whether preformed ions or ions generated in the gas phase provide the major source of the ionic species detected in MALDI [46–49]. Both ESI and MALDI are very sensitive analytical techniques utilizing analyte concentrations that are as low as picomolar. One of the main differences, however, between MALDI and ESI

is the state in which the sample is introduced to the ion source. ESI uses solvated sample that is infused into the instrument, whereas MALDI uses the solid state. Therefore, when interfaced with LC, it is possible to efficiently utilize ESI for quantitative measurements [50,51]. Even though ESI is capable of reproducing data better than MALDI, it should be noted that relative abundance of various ions in an ESI spectrum is not a real representation of the sample concentration. ESI tends to produce multiply charged species for biomolecules, such as proteins and peptides. This is the reason why ESI can, theoretically, have unlimited mass range, because very large proteins can appear at lower *m/z* values [44,45]. MALDI, however, tends to produce singly charged species and this phenomenon is of great importance for identifying the molecular ion of proteins, carbohydrates and lipids [46,48,49].

2.3. Mass Analysers

A mass analyser is the part of the instrument in which ions are separated based on their *m/z* values [17,18,20]. In a mass spectrometer, the isolation of ions is usually electrically driven, although traditional analysers, namely, magnetic sectors, employ a magnetic field that influences ion separation. From the very beginning to the present almost any physical principle ranging from time-of-flight to cyclotron motion has been employed to construct mass analysing devices. Some were extremely successful at the time of their invention, for others it took decades until their potential had fully been recognized. Currently, many analysers are widely used, namely, quadrupole (Q), quadrupole ion trap (QIT), time of flight (ToF), Fourier transform ion cyclotrone resonance (FT-ICR) and Orbitrap. These analysers vary in terms of size, price, resolution, mass range, and the ability to perform tandem mass spectrometry experiments (MS/MS). For example, QIT is capable of multiple mass spectrometric experiments (MS^n), while FT-ICR is very powerful in terms of accurate mass measurements [15,29,30,52].

2.3.1. Quadrupole analysers

Since the Nobel Prize-awarded discovery of the mass-analysing and ion-trapping properties of two- and three-dimensional electric quadrupole fields and the concomitant construction of a quadrupole (Q) mass spectrometer [53,54], this type of instrument has steadily gained importance. Modern quadrupole instruments cover up to *m/z* 2000 or even higher with good resolving power and represent a standard device in LC-MS [55]. Among the advantages of quadrupoles there are high transmission, light-weighted, compactness and comparatively low-prices, low ion acceleration voltages, and high scan speeds, since scanning is realized by solely sweeping electric potentials. A linear quadrupole mass analyser consists of four hyperbolically or cylindrically shaped rod electrodes extending in the z-direction and mounted in a square configuration (xy-plane) [56], as it is shown in **Figure 5**:

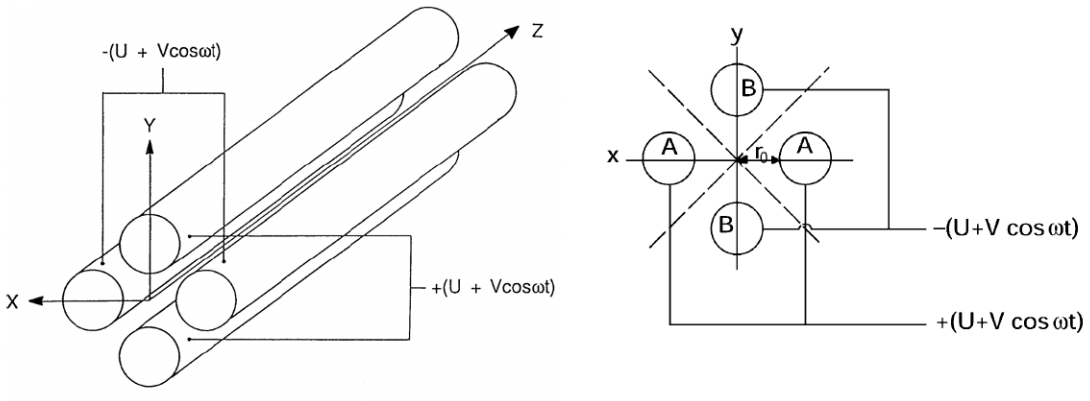

Figure 5 *Schematic representation of a linear quadrupole mass analyzer (adapted from* [17]).

The pairs of opposite rods are each held at the same potential which is composed of a DC and an AC component. As an ion enters the quadrupole assembly in the z-direction, an attractive force is exerted on it by one of the rods with its charge opposite to the ionic charge. If the voltage applied to the rods is periodic, attraction and repulsion in both the x- and y-directions will alternate in time, because the sign of the electric force also changes periodically in time [56]. If the applied voltage is composed of a DC voltage U and a radiofrequency (RF) voltage V with the frequency ω, the total potential φ_0 is given by:

$$\varphi_0 = U + V\,cos\,cos\,\omega t$$

Thus, the equations of motion are:

$$\frac{d^2x}{dt^2} + \frac{e}{mr_0^2}\varphi_0 x = 0$$

$$\frac{d^2y}{dt^2} + \frac{e}{mr_0^2}\varphi_0 y = 0$$

where r_0 is the distance between the centre of the quadrupole and the rod surface. In case of an inhomogeneous periodic field such as the above quadrupole field, there is a small average force which is always in the direction of the lower field. The electric field is zero along the asymptotes in case of hyperbolic electrodes. It is therefore possible that an ion may traverse the quadrupole without hitting the rods, provided its motion around the z-axis is stable with limited amplitudes in the xy-plane [57]. For a given set of U, V, and ω the overall ion motion can result in a stable trajectory causing ions of a certain *m/z* value or *m/z* range to pass the quadrupole. Ions oscillating within the distance $2r_0$ between the electrodes will have stable trajectories. These are transmitted through the quadrupole and detected thereafter. The path stability of a particular ion is defined by the magnitude of the RF voltage V and by the ratio U/V. From here, it is possible to obtain a stability diagram, useful to evaluate the xy-plan trajectory stability, by plotting parameters *a* and *q*, defined as the time invariant and variant fields, respectively, and obtained from the equation of motion:

$$a_x = -a_y = \frac{4eU}{m^2 r_0^2 \omega^2}$$

$$q_x = -q_y = \frac{2eV}{m^2 r_0^2 \omega^2}$$

The plot, shown in **Figure 6**, reveals the existence of regions where both x- and y-trajectories are stable, either x- or y-trajectories are stable, and no stable ion motion occurs:

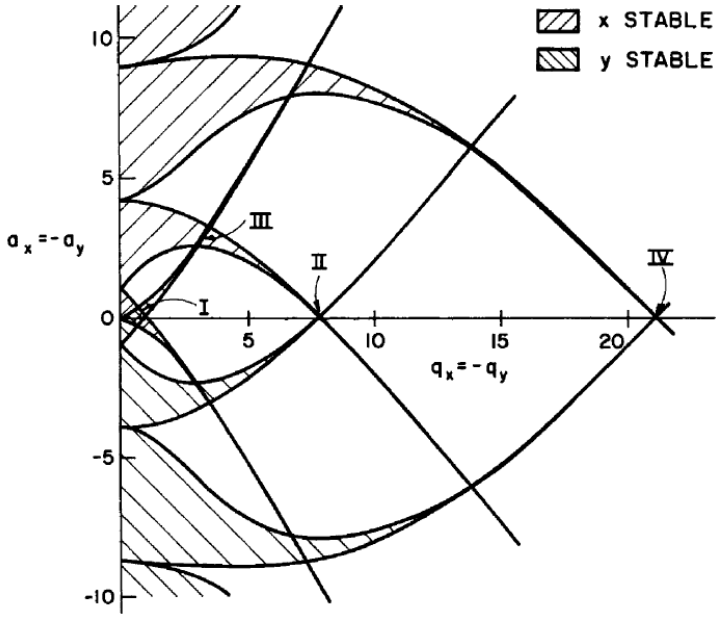

Figure 6 *Stability diagram for a linear quadrupole analyzer (adapted from* [17]).

Among the four stability regions of the first category, region I is of special interest for the normal mass-separating operation of the linear quadrupole [55–57]. If the ratio *a/q* is chosen so that 2U/V = 0.237/0.706 = 0.336, the xy-stability region shrinks to one point, the apex, of the diagram. By reducing *a* at constant *q*, i.e. reducing U relative to V, an increasingly wider *m/z* range can be transmitted simultaneously. Enough resolving power is achieved as long as only a small *m/z* range remains stable, e.g. one specific *m/z* ± 0.5 for unit resolution. Thus, the width (Δq) of the stable region determines the resolving power. By varying the magnitude of U and V at constant U/V ratio, a U/V = *constant* linked scan is obtained allowing ions of increasingly higher *m/z* to pass the quadrupole. Overall, the quadrupole analyser rather acts as a mass filter. Quadrupole analysers generally are operated at so-called unit resolution, normally restricting their use to typical low resolution (LR) applications [57]. At unit resolution adjacent peaks are just separated from each other over the entire *m/z* range, i.e. R = 20 at *m/z* 20, R = 200 at *m/z* 200, and R = 2000 at *m/z* 2000. Setting the DC voltage U to zero transforms the quadrupole into a wide band pass for ions. In the stability diagram this mode of operation is represented by an operation line equivalent to the q-axis. Such devices are commonly known as RF-only quadrupoles (q); RF-only hexapoles (h) and octopoles (o) are used analogously [58–60]. Generally, higher-order RF 2N-

multipoles differ from quadrupoles in that they do not exhibit a sharp *m/z* cut-off in transmission. Higher-order multipoles exhibit increasingly steeper potential wells, offer better ion-guiding capabilities and better wide-band pass characteristics, i.e. wider *m/z* range acceptance. This property led to the widespread application of electric quadrupoles, hexapoles, and octopoles as ion guides and collision cells [61]. From the viewpoint of the ions, they act like a hose or pipe while being fully permeable for neutrals. Thus, the RF ion guide allows residual gas to effuse through the gaps between the rods into the vacuum pumps, whereas ions are escorted into the mass analyser. RF-only quadrupole, hexapole, or octopole collision cells are part of so-called triple quadrupole mass spectrometers, which essentially represent QqQ, QhQ, or QoQ instruments, respectively, depending on the type of RF-only collision cell actually in place [62].

In general, it is also possible to prevent ions inside a multipole from escaping via either open end by creating a trapping potential well. This is possible by placing electrodes of slightly higher potential adjacent to the front and rear ends of the multipole. Such devices are known as linear (quadrupole) ion traps (LIT) [63]. While the entrance plate of the LIT is held at low potential, ions may enter the radially ion-confining RF field. The time span for ion accumulation is limited by reflection of the fast ions at the backside potential wall affording that the entrance gate must be closed before the lightest ions to be stored can exit the trap via the entrance. Storage of ions in the presence of some buffer gas, e.g., argon or nitrogen at 10^{-3}–10^{-2} mbar, then allows for their thermalization and collisional focusing towards the LIT axis [63]. The ions can be axially ejected at any convenient point in time. It's worth noting that quadrupoles are the only devices capable of mass-selective operation, whereas higher-order RF ion guides or higher-order LITs can only guide, accumulate, store, and finally release ions for subsequent *m/z* analysis. LITs are a rapidly expanding field of instrumentation. They have been established to collect ions externally before injecting them in bunches into an FT-ICR or a TOF analyser [64,65]. In those instruments the LIT serves to accumulate ions until a population suitable for the respective analyser is reached, to thermalize the ions in order to have narrow kinetic energy distributions, and to deliver this package to a mass analyser operating in batch mode [63]. For ion ejection from a LIT, there are two modes possible: one employs excitation of the ions to achieve mass-selective ejection in radial direction, the other uses mass-selective axial ejection by application of an auxiliary AC field to the rods of the LIT [63–65]. For the latter, when the ion radial secular frequency (governed by the stability parameters and the drive RF) matches that of the auxiliary AC field, ion excitation is effected in a way that also enhances its axial kinetic energy, and thus, leads to ejection from the LIT.

2.3.2. Fourier Transform Ion Cyclotron Resonance

The development that led to modern Fourier transform ion cyclotron resonance (FT-ICR) mass spectrometers began in 1932 when E. O. Lawrence applied a transverse alternating electric field orthogonally to a magnetic field to build a particle accelerator [66]. It was demonstrated that in ion cyclotron resonance (ICR) the angular frequency of the circular motion of ions is independent of the radius they are traveling on. Later, this principle was applied to construct an ICR mass spectrometer [67,68]. It was the introduction of FT-ICR in 1974 that initiated the major breakthrough [69]. Ever since, the performance of FT-ICR instruments has steadily

improved to reach unprecedented levels of resolving power and mass accuracy when superconducting magnets are employed [70]. Modern FT-ICR mass spectrometers offer higher resolving power and mass accuracy, attomol detection limits (with nanoESI or MALDI sources), high mass range and MS^n capabilities [25,29,30]. Most of the modern FT-ICR instruments represent some sort of hybrids with linear quadrupole or LIT front ends [70,71].

The ICR mass analyzer, or ICR cell, is showed in **Figure 7**:

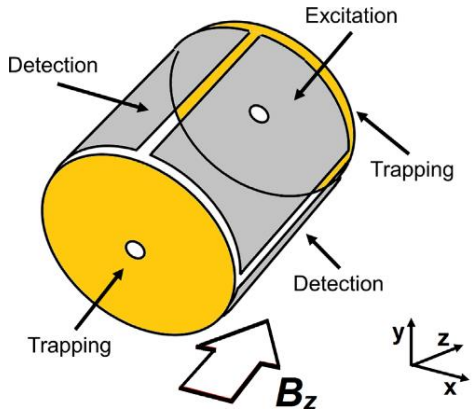

Figure 7 *A representation of the ICR cell (adapted from* [67]*).*

To understand what happens in an ICR cell, let's consider an ion entering a uniform magnetic field B, which direction is co-axial with the cell axis (*z*-direction). In this situation, it will, by action of the Lorentz force, immediately move on a circular path with a velocity *v* perpendicular to B, with negative ions circulating clockwise while positive ions moving counterclockwise [68]. The radius r_m of the ions' circular motion is determined by:

$$r_m = \frac{mv}{qB}$$

where *m* being the mass of the ionic species, *v* its velocity, *q* its charge and *B* the magnitude of the applied magnetic field. By substituting $v = \omega_c r_m$, it's possible to determine the *cyclotron angular frequency* ω_c by rearranging the previous equation:

$$\omega_c = \frac{qB}{m}$$

From here, the *cyclotron frequency* can be obtained:

$$f_c = \frac{\omega_c}{2\pi} = \frac{qB}{2\pi m}$$

One realizes that the cyclotron frequency is independent of the ions' initial velocity, but proportional to its charge and the magnetic field, and inversely proportional to its mass. Of any physical quantity, frequencies can be measured at the highest accuracy, and thus, cyclotron frequency measurements appear as ideal premises for building powerful *m/z* analysers [67,68].

Gaseous ions are not at rest but at least move arbitrarily at their thermal velocities. When such a package of thermal ions is generated within a magnetic field or is injected into it, the resulting small ion cloud contains ions that are all spinning at their respective cyclotron frequencies (circular micromotion) while the cloud as a whole remains stationary provided it has been brought to a halt within the field boundaries [68]. Therefore, the magnetic field not only acts in a *m/z*-sensitive way by imposing the cyclotron motion on the ions, but also provides ion trapping in a plane perpendicular to its field lines. In practice, the ions initially oscillate at very small "thermal" cyclotron orbits. However, their initial packet radius in the ICR cell is defined by the space that they occupied in the ion optics, which is probably ~1 mm, before being transferred into the ICR cell [67,68]. For detection, a radio frequency (RF) potential is applied to the excitation plates to increase an ions' cyclotron radius. When the RF excitation frequency equals the cyclotron frequency of the ion, it will absorb energy from the electric field and spiral up into a larger cyclotron orbit with a diameter of typically 2–5 cm, depending on the excitation amplitude, cell geometry parameters, and the excitation duration [68], as it is shown in **Figure 8**:

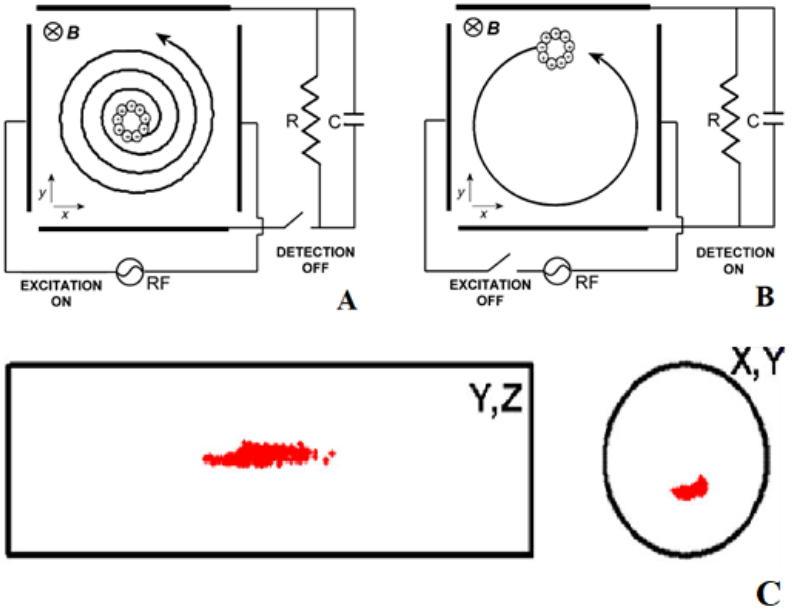

Figure 8 *Excitation (A) and image current detection (B) in FT-ICR MS (adapted from* [17]*). On the bottom (C), a simulation of the ion cloud motion during excitation in a FT-ICR cell reveals the presence of coherence on the y,z-plane too (adapted from* [68]*).*

Normally, to excite all the ions for detection, the applied RF pulse is a stepwise frequency sweep across the entire resonant frequency range of interest, and the amplitude of the pulse is usually a constant, known as an "RF chirp" [67,68,72]. During the excitation event, ions that are off-resonance do not absorb energy (mostly) and remain at the centre of the cell, while ions whose cyclotron frequency is in resonance with the excitation one rapidly spiral outwards coherently to a larger radius. Through the RF chirp, ions of the same *m/z* are excited

coherently and undergo motions as ion packets. Each ion packet has a resonance frequency (corresponding to *m/z*). For lighter ions, the spiral reaches the same radius with fewer cycles than in case of heavier ones, i.e. the spiral is steeper, because low-mass ions need less energy than high-mass ions to accelerate to a certain velocity. Cyclotron frequency measurements are thus made by *image current detection*, which relies on the fact that an ion cloud repeatedly attracts (positive ions) or repels (negative ions) the electrons of two opposite detection electrodes upon its passage [73]. The resulting minuscule image current can be amplified, transformed into a voltage signal and recorded as long as the ion motion exhibits sufficient coherence. Thus, the potential induced by the ion current is recorded simultaneously as a function of time. As the ions' motion in the cell is periodic, the signal recorded is a composite sum of N sinusoidal waves with different frequencies in the time domain, and the intensity of the signal is damped with time.

Detection of this image current allows to record a so-called *free induction decay* (FID), i.e. an oscillating signal that goes to zero after a certain time span [67,68,73]. The transient FID is recorded, and afterwards, is converted from the time to the frequency domain by means of Fourier transformation, a mathematical operation that transforms one complex-valued function of a real variable into another. This means that the complex FID caused by superimposition of many single frequencies is deconvoluted to reveal the single contributing frequencies and their respective amplitudes [67]. The frequencies are converted to *m/z* values, their amplitudes now representing the abundances of the corresponding ions. The detection efficiency is greatly improved for cylindrical cells as compared to cubic cells, because the ions pass the detection electrodes at almost constant distance to their surface, resulting in stronger image currents [68,74,75]. The image current induced in the detector plates (y-axis) is recorded as transient signal for some period of time (0.5–30 s). The excitation of the ions within the ICR cell must stop at a level low enough to avoid wall collisions of the lightest ions to be measured.

It's worth noting that in an actual experiment, the continuous image current detected cannot yield continuous time domain data. Instead, it is sampled at a certain acquisition frequency to produce a discrete transient consisting of a finite number of data points [67]. Due to the discontinuity of the data, the frequency used for sampling is crucial for reconstruction of the original signal. According to the Nyquist theorem [76], the required frequency for sampling must be at least twice the highest frequency being recorded in the transient. The instrument software determines the sampling frequency, which is usually twice the highest frequency in the spectrum (corresponding to the lowest *m/z* cut-off). Similarly, the number of data points to be acquired in the transient can be determined by the user prior to acquisition. After that, the acquisition time (T) of the transient can be calculated from the sampling frequency f_s and the data set size N (expressed in mega-point, e.g. 1 mega-point = 1 x 1024 x 1024 data points) according to the following expression:

$$T = \frac{N}{f_s}$$

During acquisition, the time domain transient with a fixed data size is then digitized at the sampling frequency rate [67].

At a first glance, the z-dimension of the cell seems to be of no importance for the function of an FT-ICR mass spectrometer. However, the z-component of thermal energy and the kinetic energy of ion injection into the ICR cell in case of an external ion source both would lead to rapid loss of the xy-trapped ions along that axis, because they would pass through the cell along the z-axis on a helical trajectory. It is therefore important to establish a trapping potential in z-direction [74,77]. Trapping of ions in a potential well implies reflection of ions between the trapping plates that induces an oscillatory motion along the z-axis with a frequency ω_z. The curved electric field close to the borders also produces an outward-bound radial force $F_r = qE(r)$ opposed to the action of the Lorentz force. The magnetic field now acts by transforming the radial force component into another circular motion of the trapped ions. The sum of these acting forces results in two other angular frequencies, i.e. the *reduced cyclotron angular frequency* ω_+ and the *magnetron frequency* ω_- [68], given by:

$$\omega_+ = \frac{\omega_c}{2} + \sqrt{\left(\frac{\omega_c}{2}\right)^2 + \frac{\omega_z^2}{2}}$$

$$\omega_- = \frac{\omega_c}{2} - \sqrt{\left(\frac{\omega_c}{2}\right)^2 + \frac{\omega_z^2}{2}}$$

Thus, the presence of the trapping potential well results in a reduction of the *unperturbed cyclotron angular frequency* ω_c and a more complex ion motion. To summarize, the three ion motional modes are represented in **Figure 9**:

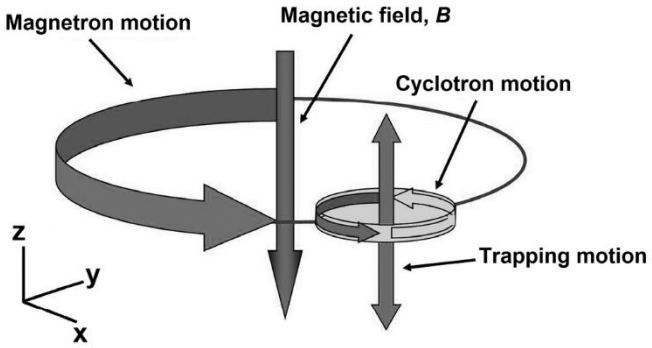

Figure 9 *The three natural ion motional modes, i.e. cyclotron rotation, magnetron rotation, and trapping oscillation (adapted from* [67]*).*

Among the three frequencies, the trapping frequency ω_z (<10 kHz) and the magnetron frequency (<100 Hz, approximately independent of *m/z*) are the consequence of the applied trapping electric field which are much lower than the cyclotron frequency (ranging from kHz to MHz), for this reason, trapping and magnetron frequencies are often ignored. However, the trapping electric fields do perturb the ion's motion and affect the measurement of the ion cyclotron frequency in case of inhomogeneity of the applied magnetic field, and these

can result in undesirable effects including radial ion diffusion, frequency shift, peak broadening, and sidebands on the peaks [74]. However, during last years, modifications of the ICR cell were apported in order to adjust trapping electric field and, thus, to reduce ion motion perturbation. Among most suitable solutions, the dynamically harmonized FT-ICR cell deserves a special attention, since a dynamic harmonization of the electric field [78] is achieved here by adding extra electrodes shaped in such a way that the averaged electric field created by these produces a counter force to the forces caused by the inhomogeneous magnetic field. This resulted in the highest resolving power achieved in peptide and protein analysis [74]. Nowadays, the dynamically harmonized FT-ICR cell is used in commercial instrumentations such as the Bruker SolariX XR, providing outstanding results [29,30,79,80].

Ion traps, ICR cells as well as quadrupole ion traps, are best operated with the number of trapped ions close to their respective optimum, because otherwise ion trajectories are distorted by coulombic repulsion [81]. Hence, external ion sources, in combination with ion transfer optics capable of controlling the number of injected ions, are ideally attached to ion traps. Currently, MALDI and even more so ESI ion sources predominate in FT-ICR [29,30,79,80]. The ion current is not solely regulated by the source but by some device to collect and store the desired amount of ions until the package is ready for injection into the ICR cell. Linear RF-multipole ion traps are normally employed for that purpose. RF-only multipoles are commonly used to transfer the ions through the boundaries of the magnetic field into the ICR cell [64]. For their injection, it is important to adjust the conditions so that the ions have low kinetic energy in z-direction in order not to overcome the shallow trapping potential. While some buffer gas is beneficial in case of LITs and QITs, ICR cells are operated at the lowest achievable pressure. The typical path from an external ion source into the ICR cell is therefore characterized by multistep differential pumping to achieve some 10^{-8}–10^{-7} Pa inside the ICR cell.

2.4. Practical aspects of Fourier Transform Ion Cyclotron Resonance Mass Spectrometry and Data Treatment

2.4.1. Signal Acquisition: Resolution, Signal-to-Noise ratio and Transient time

Mass measurement accuracy (MMA) is the key measurement parameter of a FT-ICR mass spectrometer. Indeed, FT-ICR can achieve mass accuracy in sub-ppm, or even ppb (parts per billion) level in state-of-the-art instruments [82]. With such extraordinary performance, the ion's elemental composition can be revealed without tandem MS experiments [29,30,79,80]. However, mass accuracy depends upon the mass resolving power (R) and signal-to-noise ratio (S/N), because a prerequisite for accuracy is that the peak of interest must be well resolved and distinguished from others [27]. FT-ICR is known best for its high mass resolving power:

$$R = \frac{m}{\Delta m} = \frac{\omega}{\Delta \omega}$$

where m and ω are the m/z and the cyclotron frequency values for the peak of interest, Δm and $\Delta \omega$ are the peak width at half maximum (FWHM). The resolving power is particularly important when peaks become

close neighbours (e.g., isotope peaks of highly charged ions, complex mixtures). The mass resolving power in FT-ICR can be estimated in terms of the applied magnetic field and the transient time:

$$R = \frac{m}{\Delta m} = \frac{\omega}{\Delta \omega} = \frac{qB}{m\Delta\omega} \propto \frac{qBT}{mk}$$

$$\Delta\omega \propto \frac{k}{T}$$

in which k is a peak width constant and q is the ionic charge. This peak width, $\Delta\omega$, is inversely proportional to the signal acquisition time, T, after the Fourier transform (FT). Therefore, the equations above reveal that the resolving power of any peak for a given m/z (m and q are fixed) is directly dependent on the magnetic field (B), duration of the transient (T), and the peak width constant (k). Among the three factors, building a higher magnetic field is the most straightforward way to improve resolution [70,82]. However, the expense of the magnet is the major part of the instrument cost and increases dramatically for high field. Nevertheless, efforts can be made through instrument tuning and data analysis in order to improve T and k. Thus, the users tune the instrument to acquire a time-domain signal which lasts as long as possible. Overall, R increases with the duration time of the transient, and the peak shape changes from a *sinc* function to a Lorentzian function. It could seem that R increases linearly with T, thus suggesting assuming long transient times for the analysis to reach higher resolution levels. Such an argument is only partially right. The increase of resolution is almost linear during the first 1-2 sec in the transient which agrees with theory, however, the slope starts to drop quickly after that, and the resolution reaches a maximum before facing a decrease. Such a nonlinear variation of the resolving power results from the damping of the transient. After the excitation event, the ions' coherent motion will experience frequent collisions with other ions and gas particles, which speeds up the magnetron expansion and affects ions' cyclotron motion [83–85]. In the ICR cell, the trapping potential is nearly ideal in the centre, but "squares off" quickly at larger radius. With substantial magnetron expansion or axial excitation, the ions will be exposed to increasingly inhomogeneous electric fields in both cyclotron and trapping motions, and therefore, the coherent ion packets dephase rapidly during the ion detection event, which results in an exponential decay on the signal intensity. Furthermore, in a real experiment, the Coulombic repulsion and inhomogeneous electric field in the cell will cause an ion's frequency to shift and undergo peak broadening, which further attenuate the signal-to-noise ratio S/N. The latter parameter thus seems to follow the same fashion related to resolution. But, generally, the resolution of a peak is maximized at longer acquisition times than S/N, which means the optimal data acquisition time for the two parameters are not the same. From theoretical considerations, the optimal conditions for R and S/N in terms of transient time depends on the transient shape [86]. Therefore, when acquiring a spectrum, in addition to monitoring the *m/z* spectrum, it is important to monitor the transient at the same time.

2.4.2. Absorption and Magnitude mode Mass Spectra

In FT-ICR, the original signal is a composite sum of sinusoidal waves with different frequencies recorded in the time-domain. Such a signal is Fourier transformed to a frequency domain spectrum. Fourier transformation

on a time domain signal, $F(t)$, produces a frequency spectrum with a complex output, $f(\omega)$, consisting the *absorption mode* spectrum, $A(\omega)$, and the *dispersion mode* spectrum, $iD(\omega)$ [87]:

$$f(\omega) = A(\omega) + iD(\omega)$$

The phase angle of the complex number can be calculated by the following expression:

$$\phi(\omega) = arctan\left[\frac{D(\omega)}{A(\omega)}\right]$$

Mathematically, $A(\omega)$ and $D(\omega)$ are the projection of $f(\omega)$ in the real and imaginary axis, which means they contain the same information with π/2 difference in phase. However, as projection of $f(\omega)$ on the x- or y-axis ranges from positive to negative, both $A(\omega)$ and $D(\omega)$ have positive and negative values which makes their plots difficult to interpret [88]. To overcome this problem, the vector sum of $A(\omega)$ and $D(\omega)$ is calculated to yield a phase independent *magnitude mode* spectrum, $M(\omega)$, which is plotted by all commercial and custom FT-ICR instruments [89]:

$$M(\omega) = \sqrt{((A(\omega))^2 + (D(\omega))^2)}$$

It's worth noting that, even if plotting magnitude-mode spectra allows to avoid the tedious phasing problem, using absorption mode acquisition provides a series of advantages [88]. First, the peak width at half-maximum height (Δm) is narrower than its corresponding magnitude-mode by a factor depending on the damping of the transient, from 1.7 to 2, without a concomitant loss in peak height. Second, in absorption-mode, the S/N increases by $\sqrt{2}$ compared to magnitude-mode and this is because the noise of the imaginary part of the complex number is not added to the one related to the absorption mode spectrum. Third, displaying the spectrum in absorption mode can easily distinguish the common artefacts that exist in any FT-based mass spectrometer. This is possible since artificial peaks cannot be phased. Unfortunately, in FT-ICR, the pulse program involves a large and varying phase shift before the spectral acquisition event, and such a phase shift has hindered usage of the absorption mode spectrum for almost 40 years [90,91]. The phase shift comes from two sources. First, for signal detection, ions in the ICR cell are excited to a larger orbital radius by a linear frequency sweep, and ions will undergo excitation once its cyclotron frequency matches the excitation waveform frequency. Thus, ions of different frequencies are excited at different times and accumulate a different total phase lag before the detection event. Second, once the ions are excited, a delay time (in milliseconds) is required to settle the amplifiers before the detection event. During the delay time, all ions will continue their cyclotron motion and accumulate phase [68]. In summary, before signal detection, all ions experience varying phase shifts, and any ion's phase, $\phi(\omega)$, is expected to vary with their excited frequency. The wide range of the phase shift results in the "phase wrapping problem" [91]. As any peak can be perfectly phased at a $\phi(\omega)$ between 0 and 2π, it is easy to phase a small *m/z* region in the spectrum, where the phase shift is <2π. However, according to trigonometric relationships, $\phi(\omega_i) = \phi(\omega_i) + 2n_i\pi$ for any integer n_i at any ω_i, which makes any phase angle beyond 2π return to 0–2π (known as phase wrapping). Consequently, it becomes essential to consider the additional $2n_i\pi$ into the phase angle for each peak in the spectrum. In FT-

ICR, a typical excitation bandwidth ranges from kHz to MHz, which makes the accumulation of phase substantial (e.g. >10000 π for the *m/z* range from 200 to 2000 in a 12 T system). Finding the correct n_i for each peak (thousands of peaks in a spectrum) is crucial for broadband phase correction, and this answers the question why virtually all FT-ICR instruments plot the spectrum in the magnitude-mode as calculating the phase shift accurately for each peak is difficult. Despite this, many efforts were dedicated to find a solution, some of which include the utilization of a dedicated instrumentation, while others comprise the accomplishment of too many computational steps [67]. Nevertheless, Kilgour et al. [91,92] optimized successfully a tool able to calculate absorption mode mass spectra in few milliseconds by employing a genetic algorithm regardless the type of FT-MS instrumentation, making absorption mode MS analysis accessible to common users.

2.4.3. Apodization and Zero-filling

Spectra acquired from an instrument have a definite peak shape which is associated with the instrument itself. The natural peak shape is a convolution of both *sinc* and Lorentzian functions, and the component from the *sinc* function contains undesirable sidebands at both sides of a peak, which cannot be avoided during spectral acquisition [93]. These *wiggles* contain no useful information but can interfere with the identification of adjacent peaks of low intensity. Therefore, during data processing, the time-domain transient is often multiplied by a window function prior to FT to minimize the sideband intensities and smooth the line shape. Such procedure is called *apodization* [94]. An optimal window function can smooth the sidebands but also inevitably degrades the S/N and resolving power of the spectrum. It has been found that the magnitude and absorption mode spectra do performance better with different types of apodization [93]. By multiplying the transient with a window function, the overall peak shape changes, which therefore affects the resolution, relative intensity, and S/N of the peaks after FT. After apodization, the sidebands of the peak are largely suppressed and the line shape becomes much smoother, while at the same time, the peak width is broadened. Such effect is important for complex spectra (e.g. from proteomics to petroleomics), where peak intensities vary over 1,000x throughout these ones. By using apodization, the low intensity peaks will suffer much less perturbation from adjacent intense peaks. Apodization is a standard step during FT-ICR data processing, because it smooths the peak shape and facilitates the assignment. However, the benefit varies with the spectral conditions and for different peaks in the same spectrum. Furthermore, the improved peak shape is generated at the cost of spectral resolution, and problems like frequency shifts, from image or space-charge, will often be "hidden" as the peak shape is "smoothed out". Thus, it is advisable to keep the raw spectrum and use apodization carefully.

Although most window functions produce similar results, spectra in magnitude and absorption-mode are affected differently by different types of window functions. A window with its maximum at the beginning and minimum at the end is called "half window", a symmetrical window with its minimum at two sides and its maximum in the middle is called "full window" [95,96]. In the magnitude-mode plot, the full window after FT shows a narrower peak width and suppressed sidebands compared to the half window apodization, therefore,

it is recommended in the magnitude-mode spectrum for the best line shape and better resolving power. However, a full window function will generate large negative intensities in the absorption mode after FT [95,96]. As the peaks become more closely spaced in the *m/z* domain, the negative sidebands start to interact with neighbouring peaks and can severely distort the spectrum; such a problem can be fixed by using a half window function [93]. And in addition, the half window apodization in the absorption-mode normally results in a narrower peak width compared to a full window function in the magnitude mode. An extra benefit is that the full window zeroes the signal at both beginning and end of the transient, while a half window function retains the most intense signal at the beginning. Therefore, using a half window apodization will cause less change in S/N and peak shape compared to the full window apodization. In summary, while no apodization is preferred for preservation of peak shapes and detection of space-charge or electric field inhomogeneity effects, if apodization is required to facilitate peak picking and assignment, the full window and half window functions are recommended for magnitude- and absorption-mode spectrum, respectively.

During recording, it's important to remember that a continuous image current cannot analytically be recorded; instead, a discrete transient, F(t), is sampled at 2x of the Nyquist frequency, and then recorded at N equally spaced intervals over the signal acquisition time T. Fourier transformation on the discrete time-domain signal yields N/2 complex data points in the positive frequency domain, with 1/T Hz space interval. The spectrum is plotted by connecting the individually discrete frequency points using a straight line. However, compared to a theoretical continuous spectrum, a straight-line connection inevitably distorts the spectral line shape, as in most situations the signal frequency to be determined is not exactly the one of the frequency points used by the FT. Such distortion of line shapes causes peak broadening, sideband wiggles, and position shift, which will affect the peak centroiding algorithms for determining the peak location and intensity [97,98]. Recovery of the continuous line shape can be greatly improved by *zero filling*, a method to extend the time-domain data by adding zeros at the end of the transient prior to FT. Consider if N zeros are added to the end of an N-point transient, the data size will then become 2N, therefore, the FT yields an N-points frequency spectrum, which is now equally spaced at the interval of 1/2T Hz rather than 1/T Hz. By doing this, the number of data points in the frequency spectrum is doubled, and a smoother line shape can be generated, thus improving peak shapes and resolution. The discrete spectrum approaches the true continuous spectrum with infinite zero filling. However, each zero fill doubles the data size, and the computation time for FT increases substantially with the increasing data length. Although zero filling is usually necessary to reduce the errors rising from the discrete line shape, to some extent, further zero filling is a waste of the computer memory. In general, the peak shape and resolution can be almost perfectly recovered by two zero fills.

2.4.4. Mass Calibration and Space-Charge effect

Operation of FT-ICR requires implementation of a mass calibration equation to convert experimentally measured frequencies of ions into the corresponding *m/z* value [81,99]. Extremely high mass accuracy becomes essential for unambiguously determining the elemental composition of the ions, and therefore, an accurate calibration function is required to convert the detected frequency of ions to their true *m/z*. Typically, one of

two calibration functions are used for calibrating the FT-ICR spectrum: the Ledford equation [81] and the Francl equation [99], both originating from the ion's motion in a spatially uniform magnetic field plus a three-dimensional axial quadrupolar electrostatic potential:

$$\frac{m}{z} = \frac{eB}{\omega_+} - \frac{eV_T\alpha}{a^2\omega_+^2} = \frac{A_{Ledford}}{\omega_+} - \frac{B_{Ledford}}{\omega_+^2} \text{ (Ledford Equation)}$$

$$\frac{m}{z} = \frac{eB}{\omega_+ + \left(\frac{V_T\alpha}{Ba^2}\right)} = \frac{A_{Francl}}{\omega_+ + B_{Francl}} \text{ (Francl Equation)}$$

in which α is the trapping scale factor, a is the size (i.e. distance between the trap plates) of the ICR cell, B is the magnitude of the applied magnetic field, V_T is the applied trapping DC voltage, e is the charge of an electron and ω_+ is the reduced cyclotron angular frequency. In practice, although the calibration constants are in principle known, they are usually determined by calibration. Mass calibration consists of least squares fitting for either Ledford or Francl equations to the frequencies of two or more peaks of known *m/z* values to yield either constants [100]. This equation is then used to calculate the *m/z* scale from the known frequency scale. The two calibration functions above are essentially equivalent within the usual mass range and can be interconverted [100].

The mass error from measurement caused by different experimental conditions and parameters from scan to scan (e.g. the trapping potential applied, stability of the magnetic field, variation of the ion population) is inevitable. Currently, it is possible to routinely achieve mass measurement accuracies (MMAs) better than 1 ppm level over a broad *m/z* range via internal calibration [101,102]. However, such performance is also the lower limit of the two calibration functions cited above, i.e. routine MMA <0.1 ppm can hardly be achieved by a simple implementation of the two equations. The major contributor to mass errors below 0.1 ppm is attributed to the space-charge effect rising from the columbic repulsion between ions during the detection of the time-domain signal [81]. Similar to the applied electric field, the ion space-charge causes frequency shift, and such a perturbation is highly dependent on the total ion number in the cell [103]. Although the sensitivity and dynamic range of the image current detector increases with the number of ions trapped in the cell, the accuracy of measurement tends to decrease significantly [104], and in extreme cases, the space-charge effect can cause the transient to collapse and die out in a very short time, a phenomenon called "the spontaneous loss of coherence catastrophe" [105]. The conventional calibration functions struggle to achieve sub-ppm MMA. Due to space-charge effects, particularly the "local" space-charge, the number of ions trapped in the ICR cell should be considered during the measurement to achieve better MMA. In an FT-ICR experiment, the normal principle to achieve high MMA is to acquire spectra with small ion populations (minimize space-charge), average the signal of multiple spectra, and mass calibrate internally. By contrast, external calibration can never provide accuracy better than a few ppm because the ion number for a measurement varies from experiment to experiment. Changes in both "global" and "local" space-charge conditions can severely degrade the ability for the Ledford and Francl equation to accommodate the frequency shift. Masselon et al. [106] have pointed out that in addition to the "global" space-charge for all the existing ions (which can be compensated by internal calibration), each individual ion cloud experiences different interactions with other ion clouds, called a "local"

space-charge effect, and this effect become crucial when the frequencies are measured to a precision better than 0.25 Hz (corresponding to a mass error of ~1 ppm depending on the magnetic field). Furthermore, the space-charge of local ion clouds varies with the specific ion abundance over a wide range. Internal calibration compensates for the "global" space-charge effect because the ions being detected are exposed to an identical experimental environment, and therefore, the conventional calibration laws can routinely approach MMA of ~1 ppm level [102]. Meanwhile external calibration only performs well when the calibrant and the analyte spectra are measured under the same conditions (e.g. same instrument, trapping potentials, excitation voltage, and transient duration). Neverthless, the ion population cannot always be reproduced from one experiment to another, especially in LC-MS, MALDI, and imaging experiments, whose chaotic nature produces scan-to-scan variations on the number of ions even for the same analyte. When dealing with unknown samples, an external calibration function acquired using the same experimental parameters can be applied, and external calibration offers a simpler pulse sequence and higher throughout because an internal calibrant is not always feasible. However, if the space-charge effect is ignored, the external calibration can result in mass errors of hundreds of ppm, and such error is even worse for spectra with a larger frequency range or with many components in the sample.

2.4.5. Artefacts in FT-ICR Mass Spectrometry

The peaks in a FT-ICR spectrum represent the frequency of the ions' cyclotron motion, while artefacts that result from transient distortion and radio-frequency interference (RFI) also exist as peaks in the spectrum. The artefacts are inevitable spectral features [107]. They contain no chemical information but can complicate the task of data interpretation. Artefacts induced by RFI are noise signals from the detection lines. In an FT-ICR instrument, power supplies, turbo pumps, RF oscillators, ion gauges, and even the (noisy) DC voltages on the ion optics can be the source of RFI, and thus introduce electronic peaks into the final *m/z* spectrum. Typically, electronic peaks are single peaks without isotopic patterns (although they can show some modulation by interference with other signals). Therefore, if the RFI peaks do not overlap with real peaks, they can normally, but not always, be easily recognized in the spectrum. The harmonics are the most common artefacts existing in any FT-based mass spectrometer (FT-ICR, Orbitrap, and ion trap) [108]. Different from the RFI, harmonic peaks are not real signals, they are generated inherently during the FT due to non-sinusoidal features of real signals. Odd harmonics are mostly generated by non-sinusoidal image current from detection electrodes, saturation of the amplifier or overloading the analog-to-digital converter. Even harmonics are mostly generated by non-zero magnetron radius, misalignment of the cell with the magnetic field, or imbalance of the excitation or detection amplifiers [108]. Although the intensity of the artefacts is usually minor, all these frequencies will still exist in the spectrum after FT. When the spectrum is complex, these artefacts can overlap with real peaks, deteriorate the S/N, and cause confusion for data interpretation, affecting the performance of the FT-ICR instrument. Unfortunately, they can never be completely avoided in the experiment. However, they can be easily observed in the absorption-mode spectrum by their anomalous phase variation, because neither the harmonics nor RFI experiences ion's cyclotron motion in the ICR cell, and therefore, cannot normally be phased correctly [88,91,92].

2.4.6. Smoothing and Baseline subtraction

Mass spectrometry data usually shows a varying baseline. Chemical noise in the matrix or ion overloading can cause this variation. Subtracting the baseline makes spectra easier to compare [109]. Baseline subtraction should be used whenever samples show an obvious offset, drift, or broad low-frequency peaks and before correcting the obtained spectrum with mass calibration, because the noise would affect the results of that step [109]. One strategy for removing a low-frequency baseline within the high-frequency noise and signal peaks follows three steps: estimate the most likely baseline in a small window, regress the varying baseline to the window points using a spline interpolation and smoothing, and subtract the estimated and regressed baseline from the spectrum [109,110]. Estimating the most likely background in every window is the most crucial step. Unfortunately, it's not possible to observe the true baseline using the minimum values because of the high-frequency signal noise. There are two good approaches to overcome this problem:

- Use a quantile value of the observed sample within the window. One can safely underestimate the quantile with the result that the estimated baseline is slightly conservative. On the other hand, if the quantile should be overestimated, the proportion of baseline points includes peak values [111];
- Using a probabilistic model. The second approach improves the result at the cost of computational time. The method consists of assuming that the points in every window come from a doubly stochastic model, that the source of each point can be "noise" or "peak," and that each class has its own distribution. In practice, assuming a uniform Gaussian distribution is relatively safe. Estimating the baseline implies learning the distributions and the class labels for every point, which is an unsupervised clustering problem. At the end, the mean of the "noise" class turns out to be the best baseline estimate for the window [112].

The chosen window size should be sufficiently small so that the varying trend of the baseline is not significant, and it should be sufficiently large so that a representative sample of the baseline in the window can be observed.

Nevertheless, sometimes it's necessary to filter or reduce the noise from mass spectra to improve the validity and precision of identified *m/z* values. To do so, another step is performed in Mass Spectrometry data pre-treatment, the so-called *smoothing*, that allows to smooth the spectra, making easier the work of peak detection algorithms [113]. Smoothing (also known as polynomial filtering) involves the treatment of the signal samples in order to make them fit a particular model. It consists of adjusting sample by sample the signal based on a regional polynomial fit [113]. Among the different smoothing algorithm used in Mass Spectrometry, the Savitzky and Golay's one is commonly used and it's able to smooth a mass spectrum using a least-squares digital polynomial filter [114]. The filter coefficients are derived by performing an unweighted linear least square fit using a polynomial of a given degree and a dataset of a predefined size (*m/z* window). It allows to use higher order polynomials for the fitting. As a result, the algorithm preserves signal features such as the resolution between ion peaks and the height of the peaks. One of the most important parameters in polynomial filtering is the size of the window (or spanning). It is indirectly associated with the cut-off frequency. However,

there is not a practical relation between these two so one can usually adjust the window based on experimental experience.

2.4.7. Noise level estimation: the N-sigma methodology

As can be noticed from previous statements, noise reduction or filtering is important in a FT-ICR Mass Spectrometry experiment to accurately identify analyte signals. In FT-ICR MS, the intensity of noise peaks increases with *m/z*, being a serious problem for FT-ICR MS analyses in higher *m/z* ranges. Smoothing algorithms allow to reduce the noise level, but they are not always useful for analyte identification purposes. A possible alternative could be *noise filtering*, which can be thought as a complete cut-off of noisy signals from the recorded mass spectrum. Different approaches were assumed to accomplish this task. One of them consists of a direct elimination of several MS signals whose Signal-to-Noise ratio (S/N) is lower than a preselected threshold value, defined itself as the *noise level*. S/N ranging between 2 and 20 [30,115–117] have been used which highlights the challenge of reliable noise level estimation in a wide mass range. Other studies simply removed the lowest 10% of peaks (based on intensity) [118] while a S/N of 4 effectively removed peaks with relative intensity below 0.5% in a previous study [119]. Nevertheless, the choice of S/N or a cut-off based on intensity seems arbitrary and could lead to a massive loss of information, especially in metabolomics, since S/N ratio of metabolite-related peaks is often found to be near to 1 [29,30]. Reproducibility increases with increasing S/N cut-off from 3 to 10. However, using a strict S/N threshold is not adequate for establishing peak detection reproducibility because well-defined peaks could go undetected just below the defined threshold. An efficient solution to keep reproducibility while avoiding missing analyte signal consists of the estimation of the S/N threshold and this is the key feature of the *N-sigma methodology* [120]. In practice, for each processed spectrum, the noise level is estimated based on fitting a normal distribution to a histogram of intensities. Histogram bin sizes are selected based on the Freedman-Diaconis rule [121] to ensure the histogram is representative without excessive computer processing. The noise intensity is characterised by a bi-modal normal distribution. The first mode corresponds to the lowest intensity peaks in the MS probably associated with thermal noise [122]. The second mode may correspond to a higher intensity chemical noise or artefacts, like wiggles resulting from the application of Fourier transformation [118]. Artefacts signals may have intensities similar to analyte peaks with low concentrations or ionisation efficiency which make up the second mode of the histogram. The noise level, which is subsequently used to remove peaks during the main processing stage, is thus estimated as the mean plus N standard deviations based on the fit of the first mode. In general, choosing N = 3 is conservative, while choosing common peaks from replicate experiments accounts for the second mode. However, if doing replicates is not feasible, switching to a direct cut-off approach by choosing a noise level equal to three times the mean value could be a suitable solution, but implies the exclusion of several low intensity analyte peaks.

Smoothing the mass spectrum leads to the depletion of wiggles, thus leading to the one of the second mode of the intensity distribution, lowering the probability of detection of false positives by acting following the direct cut-off approach.

2.4.8. Formula Assignment: the Seven Golden rules and the Kendrick Mass Defect analysis

FT-ICR Mass Spectrometry thus provides unprecedented levels of resolution and accuracy. The former allows to separate isobaric peaks considering their different mass defects, i.e. the decimal part of related *m/z* values, and to resolve isotopic fine structures of multicharged ions, while the latter makes possible the assignment of a reduced number of possible chemical formulas to an observed *m/z* value, thus leading to elemental composition determination [29,30,79,80]. In detail, the number of possible formulas for a specific *m/z* value depends on the experimental accuracy. Since FT-ICR Mass Spectrometers can provide accuracies in the order of sub-ppm, it's possible to make unequivocal formula assignments for low *m/z* values. However, still a huge number of possible candidates persists for high *m/z* range MS signals. To overcome this problem, several tools should be applied to filter possible results.

The Seven Golden rules [25] are a set of chemical and heuristic rules, validated on a large database consisting of 432,968 molecular formulas which covered a chemical space of more than five million compounds, that are used as constraints for finding the correct chemical formula. The first one is related to the element numbers. In detail, it's possible to assume a maximum for element numbers, in order to reduce the number of calculated formulas. The maximum element count can be chosen *a priori* considering the chemistry of the analysed sample (for example, lower nitrogen and sulphur count maxima could be assumed for formula assignment in *petroleomics* [123–125]). Otherwise, maximum element count can be deduced from the analysis of available databases. In this regard, maximum element counts deduced from the analysis of Wiley and the Dictionary of Natural Products (DNP) databases are listed in **Table 1**:

Table 1 *Maximum element counts for small molecule formula generation based on examination of the DNP and Wiley mass spectral databases.*

Mass Range (Da)	**Library**	**C max**	**H max**	**N max**	**O max**	**P max**	**S max**	**F max**	**Cl max**	**Br max**	**Si max**
< 500	DNP	29	72	10	18	4	7	15	8	5	
	Wiley	39	72	20	20	9	10	16	10	4	8
< 1000	DNP	66	126	25	27	6	8	16	11	8	
	Wiley	78	126	20	27	9	14	34	12	8	14
< 2000	DNP	115	236	32	63	6	8	16	11	8	
	Wiley	156	180	20	40	9	14	48	12	10	15

< 3000	DNP	162	208	48	78	6	9	16	11	8	

Then, the second rule, known as the *Senior rule*, relates to the chemical feasibility of a formula. In particular, it allows to understand if a certain chemical formula could correspond to a chemically existent species, or molecular graph. Senior's theorem [126] requires three essential conditions for the existence of molecular graphs [127]:

- The sum of valences or the total number of atoms having odd valences is even (that comprises the so-called *hydrogen rule* [128]);
- The sum of valences is greater than or equal to twice the maximum valence;
- The sum of valences is greater than or equal to twice the number of atoms minus 1.

The third rule deals with detected isotopic patterns. Indeed, compounds that were synthesized by natural precursors comprise monoisotopic and isotope masses according to the natural average abundance of stable isotope abundances [129]. Considering isotopic ratio abundance patterns removes most of the wrongly assigned molecular formulas from a certain mass measurement experiment. The fourth rule is related to the hydrogen-carbon ratio, while the fifth one to heteroatom-carbon ones, values that can be used as a constraint. Indeed, most typical ratios are found between 3.1 > H/C > 0.2, for example for long chain alkanes (H/C ~ 2) or polycyclic aromatic hydrocarbons (H/C ~ 0.5). As a proof, it can be noticed that more than 99.7% of all formulas in Wiley database show a H/C ratio between 0.2–3.1 [25]. Heteroatom ratios distributions are even more skewed than H/C ratios, because many formulas comprise no heteroatom at all (such as alkanes) or very few, and rare cases exist with high ratios of heteroatoms to carbon numbers. The common, the extended and the extreme ratios for small organic compounds comprised in the Wiley mass spectral database are listed in **Table 2**:

Table 2 *Common element ratios obtained from the analysis of the Wiley mass spectral database formulas for the mass range 30 Da – 1500 Da.*

Element Ratio	**Common range (99.7 % of formulas covered)**	**Extended range (99.99 % of formulas covered)**	**Extreme range (beyond 99.99 % of formulas covered)**
H/C	0.2 – 3.1	0.1 – 6	< 0.1 and 6 – 9
F/C	0 – 1.5	0 – 6	> 1.5
Cl/C	0 – 0.8	0 – 2	> 0.8
Br/C	0 – 0.8	0 – 2	> 0.8

N/C	0 – 1.3	0 – 4	> 1.3
O/C	0 – 1.2	0 – 3	> 1.2
P/C	0 – 0.3	0 – 2	> 0.3
S/C	0 – 0.8	0 – 3	> 0.8
Si/C	0 – 0.5	0 – 1	> 0.5

The latter rules only restrict unlikely high element ratios in molecular formulas, but they don't account for multiple high element counts. The sixth rule, thus, deals with too improbable combinations of high element ratios. In **Table 3**, additional constraints are listed for NOPS, NOP and NOS combinations:

Table 3 *Multiple element count restriction for compounds < 2000 Da.*

Element Counts	**Heuristic Rule**
NOPS all > 1	N< 10, O < 20, P < 4, S < 3
NOP all > 3	N < 11, O < 22, P < 6
OPS all > 1	O < 14, P < 3, S < 3
PSN all > 1	P < 3, S < 3, N < 4
NOS all > 6	N < 19, O < 14, S < 8

Finally, the seventh golden rule, or the trimethylsilyl (TMS) rule, can be applied in analyses which comprise a derivatization step which in turn involves the addiction of TMS groups to analytes. In detail, TMS groups should be subtracted from calculated formulas before assignments. Then, resulting formulas which obey the previous rules should be considered for assignment. A similar approach could be assumed by considered different types of adduct, like sodium, potassium or water ones.

The ring-plus-double bonds equivalent (RDBE), obtained by using the equation $RDBE = C + Si - 1/2(H + F + Cl + Br + I) + 1/2(N + P) + 1$, where each element symbol represents the count of atoms of an element in the chemical formula, couldn't be used as a constraint, since it doesn't account for the different valences of nitrogen, sulphur and phosphorus. However, for the 99.90% of known natural compounds, the RDBE has been found to be less than 40 [25].

Together with the Seven Golden rules, the so-called *nitrogen rule* could be used to further reduce the number of possible formulas [130]. It states that an odd nominal mass related to a neutral molecular species corresponds

to an odd number of nitrogen atoms. However, this is no more valid for nominal masses higher than 500 Da, because small non-nominal mass contributions from a large number of elements add up in higher mass regions [25]. Nevertheless, the nitrogen rule can still be useful in lower mass ranges or during assignment of elemental compositions to small fragments.

Despite the huge number of filtered formulas, the employment of the Seven Golden rules could not lead to unique assignment for every MS signal. For example, isotopologues of low intensity monoisotopic peaks could not always be distinguished from noise, thus making impossible the utilization of the third rule for candidate formula filtration.

Another useful tool able to further improve formula assignment confidence is the *Kendrick Mass Defect* analysis [131,132]. The method implies the conversion of recorded spectra from the IUPAC scale to the Kendrick one, by multiplying *m/z* values to a specific factor, which value is related to a preselected group of atoms (*building block*), thus to obtain the corresponding Kendrick masses (KM):

$$KM = M\frac{NM_{Building\ Block}}{M_{Building\ Block}}$$

Here, $NM_{Building\ Block}$ and $M_{Building\ Block}$ refer to the nominal and the exact mass of the chosen building block, respectively, and M to the accurate measured mass. The building block is defined by a chemical formula and could correspond to a known chemical functional group or a neutral molecular species. The main effect of this conversion is a global re-ordering of the peaks. In detail, peaks with the same *Kendrick Mass Defect* (KMD), defined as $KMD = KNM - KM$, where KNM is the Kendrick nominal mass, will differ each other by a certain amount of the chosen building block. As a conseguence, it will be possible to observe the formation of different families of peaks, i.e. *homologous series*, each characterized by a certain value of KMD [132]. This leads to a further filtration of candidate formulas of peaks belonging to one of them. In particular, if an unequivocal assignment has already been performed for a member of the series, it's possible to automatically select the correct chemical formula of the other ones by adding or subtracting the corresponding amount of the building block. Moreover, separation of peaks in different homologous series could be useful since different KMD values could be related to different values of RDBE and/or to the presence of heteroatoms [131,132]. Another parameter calculated in KMD analysis is the *z* score*, which insert another degree of separation among homologous series:

$$z^* = \left[mod\left(NM_{Analyte} - NM_{Building\ Block}\right)\right]/NM_{Building\ Block}$$

Thus, homologous series are organized in nominal mass series. The *z* score* employment helps distinguishing homologous series with similar KMD values, but different formula type.

This approach is particularly useful in *petroleomics*, where different hydrocarbons contained in crude oil samples often differ for their alkylation grade [124,132]. In this situation, by setting CH_2 as the chosen building block, it's possible to easily identify alkylation derivatives and to assign unique formulas. Moreover, in

metabolomics, groups of atoms corresponding to known biochemical reactions could be selected as building blocks to improve formula assignment [29].

2.4.9. Mass Spectrometry visualization tools: Kendrick plot, Van Krevelen diagram and Molecular Mass Difference Network

Analysis by ESI-FT-ICR-MS of complex samples, such as food matrices, produces large data sets with thousands of peaks [29,30]. As stated previously, molecular formulas can be determined for each individual peak because of the higher levels of accuracy reached by using FT-ICR MS technique. Once obtained chemical formulas, various information can be obtained, related to the chemistry of a certain species, such as the RDBE and the atomic ratios. Through these parameters, it would be possible to infer which kind of class of molecules our analytes belong to or which kind of derivative they are [29,30,79,80]. To simplify this kind of analysis, it's possible to employ a well-known visualization tool, i.e. the *Van Krevelen diagram* [133], which was first introduced by Van Krevelen [134] and first used for FT-MS data by Kim et al. [135], plotting the molar H/C ratios on the y-axis and the molar O/C ratios on the x-axis. Such a plot allows one not only to elucidate what compound classes are present but also to identify what reaction pathways are taking place [79,80,135]. Thus, masses in complex natural mixtures could be related by numerous chemical transformations (e.g. methylation, hydrogenation, hydration, redox, carboxylation, etc.). In **Figure 10**, an example of Van Krevelen plot is shown:

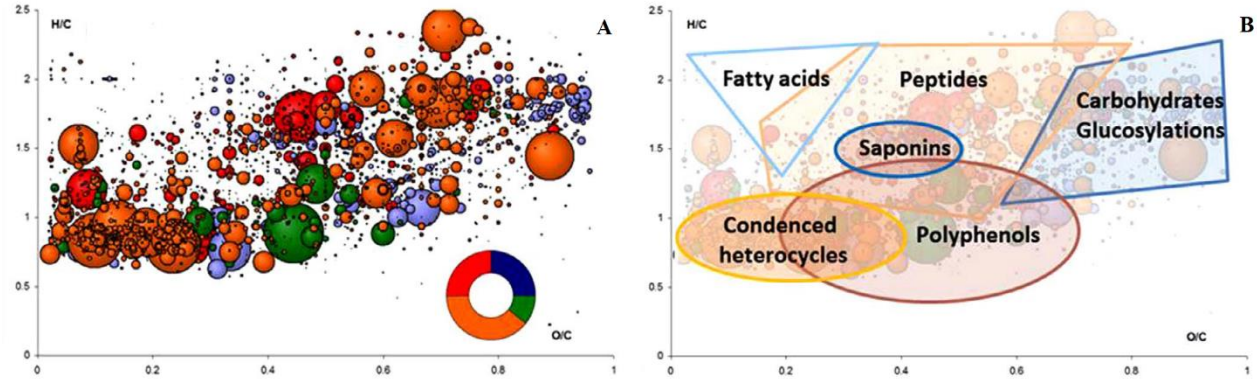

Figure 10 *In plot A, Van Krevelen plot of a sample of Fagioli di Sarconi beans. In plot B, Van Krevelen diagram with the interpretation of molecular family (CHONS (red), CHO (blue), CHON (orange) and CHOS (green) elemental compositions; adapted from* [30]*).*

In addition to the two-dimensional van Krevelen diagram, one can also display the information in three dimensions by adding ion abundance or another molar ratio (N/C, S/C, etc.) as the z-axis [136,137]. Plotting peak intensity as the third dimension provides an indication of which compound class is present with the highest abundance, but because ionization efficiency plays a large role in determining the ions' abundance, this comparison should only be used qualitatively [133]. Plotting the z-axis as either an N/C ratio or S/C ratio

disperses the elemental composition information into a third dimension where one can examine the H/C and O/C ratio of N- or S-containing molecules separated from the clusters of molecules containing only C, H, and O. If peptides are abundantly present, then the N/C ratio will highlight their presence. N/C ratios of 0.0–0.1 suggest long-chain alkyl amines, while N/C ratios of 0.1–0.4 suggest peptides and proteins [137]. Overall, two- and three-dimensional van Krevelen diagrams greatly assist in visualizing the complicated mass spectra that are acquired during the analysis of complex samples.

Another way to simplify MS data visualization and interpretation relies on KMD analysis. Indeed, as stated previously, by converting observed masses to the Kendrick scale, peaks are re-ordered in different homologous series, each of which is characterized by a KMD value. Thus, plotting the latter versus the KNM will result in a diagram in which every homologous series is characterized by points arranged on a horizontal line, each one divided by a certain amount of the chosen building block. Such a diagram is known as the *Kendrick plot* [132] and is shown in **Figure 11**:

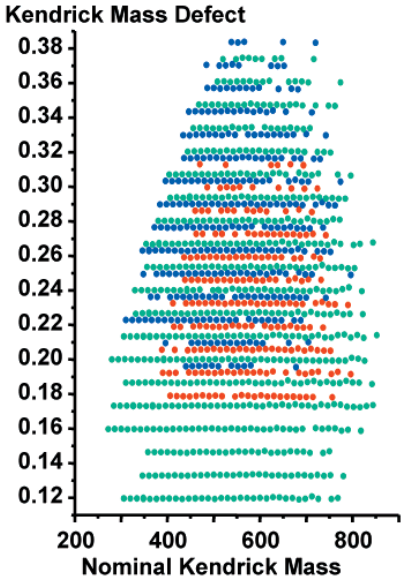

Figure 11 *Example of Kendrick plot obtained from the FT-ICR MS analysis of a sample of crude oil* [132]. *Different nominal mass or z* families are distinguished by colours (adapted from* [132]*).*

Depending on the makeup of the sample, KMD analysis using other functional groups (e.g. OCH_2, COO, CO, etc.) can be more valuable. In this way, related Kendrick plot could be used to identify species which take part to related reactions.

Despite the great potential of KMD analysis for formula assignment and data visualization, still some drawbacks persist. In detail, KMD analysis relates to one building block, making harder the identification of frequent building blocks and performing simultaneous KMD analyses related to them. Moreover, this hampers

markedly formula assignment step, making harder the identification of species that belong to no obtained homologous series. To overcome these issues, during the last years, a new approach of network reconstruction and visualisation of high-field FT-ICR mass spectra has been optimized, which offers expedient assignment of elemental formulae with improved coverage. Moreover, this approach offers new unambiguous means to depict relationships between functional group equivalents, transformations and organic molecular complexity in general. In detail, this network-based method relies on the construction of the so-called *Mass Difference Networks* (MdiN) [138–140], in which every node corresponds to an observed *m/z* value, while connecting edges correspond to pre-selected mass differences Δm, related to known atomic groups, as for the one in **Figure 12**:

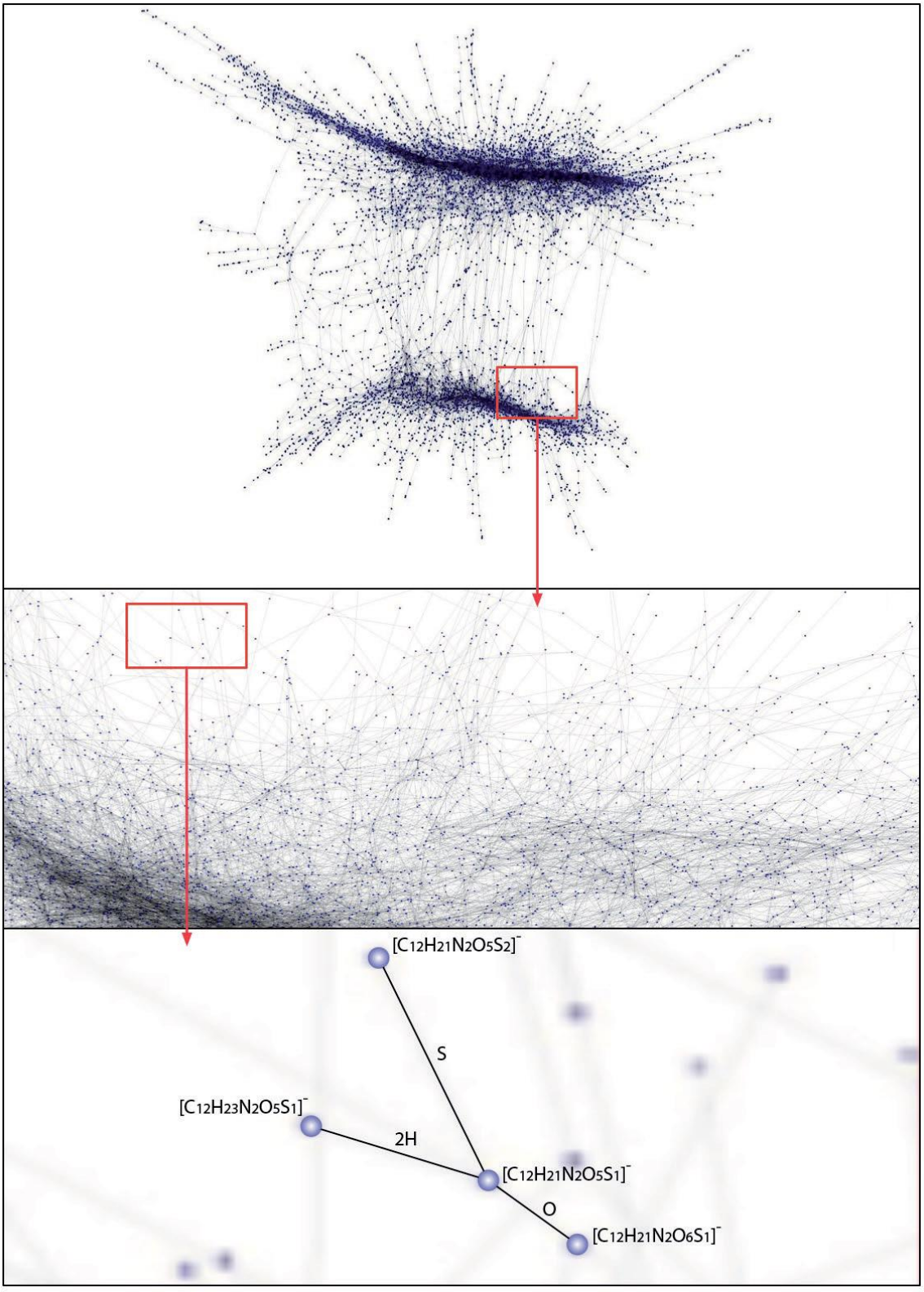

Figure 12 *Molecular network derived from a negative electrospray 12T FT-ICR mass spectrum of a secondary atmospheric aerosol (adapted from* [140]).

In this way, it's possible to highlight more than a chemical relationship among detected species by preselecting more building blocks. Furthermore, this approach could simplify isotopologue identification by adding isotopic differences, such as ^{13}C-^{12}C or ^{34}S-^{32}S, as building blocks [140]. Compositional and functional network visualisation in two or three dimensions is enabled by means of a multilevel, force-directed, layout algorithm [141]. Highly connected nodes are then arranged near the centre whereas less connected ones are assembled towards the periphery. Obtained network provide a wide range of useful information. For example, crowded clusters could form and be related to a common structural motif. Furthermore, high frequency building blocks could be identified by evaluating the occurrence of specific edges. Thus, comparison of different networks could reveal chemical pathway promotion or depletion [138–140].

Of course, formula assignment is further improved, thanks to the fact that more building blocks could be considered at once for formula calculation or filtration, leading to the accomplishment of more simultaneous KMD analyses.

2.4.10. Use of programming languages for Mass Spectrometry data treatment: the R software

As can be noticed, FT-ICR MS data face a huge processing before interpretation and results. Moreover, each step could appear very time consuming, since the dimension of obtained dataset markedly increases with the complexity of the sample. It's the case of *omics* fields, whose samples contain thousands of small molecules.

Accordingly, the use of a dedicated software is compulsory to improve laboratory efficiency nowadays, saving a considerable amount of time in performing MS routine tasks [142]. In general, mass spectrometers are commercialized together with a dedicated software system for data pre-treatment, like apodization, zero-filling, smoothing and calibration. Additional packages are offered to perform MS data analysis like formula assignment or statistics. However, related costs are too high to be widely used by research groups. Fortunately, the scientific community has developed a wide range of open source software, providing freely available advanced processing and analysis approaches. Among these, the programming and statistics environment R has emerged as one of the most popular environments to process and analyse MS datasets [143–145]. The core of the R language was started in 1997 and provided the basic functionality of a programming language, with some functions targeting statistics [146]. The real power driving the popularity of R today is the huge number of contributed packages providing algorithms and data types for a myriad of application realms. Many packages have an Open Source license. These packages are typically hosted on platforms that serve as an umbrella project and are a "home" for the developer and user communities. The Comprehensive R Archive Network (CRAN) repository contains over 14,500 packages for many application areas. The Bioconductor project (BioC for short) was started by a team around Robert Gentleman in 2004 [147], and has become a vibrant community of around 1000 contributors, working on 1741 software, 371 data and 948 annotation packages (BioC release 3.9). Among all of these, many are dedicated to MS data processing. For example, peak picking can be done using *MassSpecWavelet* [148] that applies a continuous wavelet transform-based peak detection. Moreover, many packages are dedicated to formula assignment and application of isotope pattern filtering, such as *GenFormR* and *MSbox*, while other packages, like *enviPat* and *InterpretMSspectrum*

allows to calculate theoretical isotopic patterns [148–152]. A plethora of packages exist for hyphenated MS and MS/MS techniques too, the most important of which is *xcms* [144], useful for LC and GC-MS and MS/MS data pre-treatment (peak peaking, feature extraction, data merging and data dimension reduction). Finally, R packages exist that help data interpretation through application of chemometric tools, like *MetaboAnalyst*, a package optimized for the multivariate and pathway analyses of metabolomic LC and GC-MS data [153].

R language is relatively easy to understand and allows to create personal functions able to perform multiple tasks at once, in order to save more time. Apart from MS related packages, other R software tools exist that make easier analysts' lives. For example, the R software offers a wide variety of built-in tools that allows to do huge statistics with a single function [146], like ANOVA with the function *aov()* and *prcomp()* for principal component analysis. It's possible to re-organize datasets by mean of the package *dplyr*, which comprises functions to join datasets following several criteria (like joining rows with values of a pre-selected variable in common), filter a dataset by deleting rows which don't meet a predefined prerequisite, adding or binding rows or columns, *etc*. Other packages allow you to work on character strings, like *stringr*, or to create ready publishable plots with few lines of code, like *ggplot2*. Finally, other remarkable R packages are *R Markdown* [154] and *R Shiny* [155]. The former allows users to create documents in different formats, like *pdf*, *docx* and *html*. The major advantage related to its use is that it's possible to run lines of code during document creation thus avoiding manual insertion of huge tables or graphics. Moreover, it's possible to use it to automatically generate reports at the end of a user defined pipeline. The latter, instead, makes easier the creation of interactive web apps. Indeed, with minimal syntax, it's possible to create a friendly user interface with a wide variety of widgets and to create interactive plots whose appeal changes with different user inputs.

2.5. References

[1] G. Bianco, G. Novario, G. Anzilotta, A. Palma, A. Mangone, T.R.I. Cataldi, Polybrominated diphenyl ethers (PBDEs) in Mediterranean mussels (Mytilus galloprovincialis) from selected Apulia coastal sites evaluated by GC-HRMS, J. Mass Spectrom. 45 (2010) 1046–1055. https://doi.org/10.1002/jms.1799.

[2] G. Bianco, R. Zianni, G. Anzillotta, A. Palma, V. Vitacco, L. Scrano, T.R.I. Cataldi, Dibenzo-p-dioxins and dibenzofurans in human breast milk collected in the area of Taranto (Southern Italy): First case study, Anal. Bioanal. Chem. 405 (2013) 2405–2410. https://doi.org/10.1007/s00216-013-6706-7.

[3] M. Caivano, R. Pascale, G. Mazzone, A. Buchicchio, S. Masi, G. Bianco, D. Caniani, N2O and CO2 Emissions from secondary settlers in WWTPs: Experimental results on full and pilot scale plants, in: Front. Wastewater Treat. Model. FICWTM, Springer, 2017: pp. 412–418. https://doi.org/10.1007/978-3-319-58421-8_65.

[4] G. Ventura, C.D. Calvano, I. Losito, G. Bianco, R. Pascale, F. Palmisano, T.R.I. Cataldi, Effect of pH and mobile phase additives on the chromatographic behaviour of an amide-embedded stationary phase: Cyanocobalamin and its diaminemonochloro-platinum(II) conjugate as a case study, J. Sep. Sci. 42 (2019) 1155–1162. https://doi.org/10.1002/jssc.201801060.

[5] R. Pascale, G. Bianco, D. Coviello, M. Cristina Lafiosca, S. Masi, I.M. Mancini, S.A. Bufo, L. Scrano, D. Caniani, Validation of a liquid chromatography coupled with tandem mass spectrometry method for the determination of drugs in wastewater using a three-phase solvent system, J. Sep. Sci. 43 (2020) 886–895. https://doi.org/10.1002/jssc.201900509.

[6] C.D. Calvano, G. Ventura, M. Trotta, G. Bianco, T.R.I. Cataldi, F. Palmisano, Electron-Transfer Secondary Reaction Matrices for MALDI MS Analysis of Bacteriochlorophyll a in Rhodobacter sphaeroides and Its Zinc and Copper Analogue Pigments., J. Am. Soc. Mass Spectrom. 28 (2017) 125–135. https://doi.org/10.1007/s13361-016-1514-x.

[7] T.R.I. Cataldi, G. Bianco, S. Abate, I. Losito, Identification of unsaturated N-acylhomoserine lactones in bacterial isolates of Rhodobacter sphaeroides by liquid chromatography coupled to electrospray ionization-hybrid linear ion trap-Fourier transform ion cyclotron resonance mass spectrometry., Rapid Commun. Mass Spectrom. 25 (2011) 1817–26. https://doi.org/10.1002/rcm.5054.

[8] S. Laurino, G. Grossi, P. Pucci, A. Flagiello, S.A. Bufo, G. Bianco, R. Salvia, S.B. Vinson, H. Vogel, P. Falabella, Identification of major Toxoneuron nigriceps venom proteins using an integrated transcriptomic/proteomic approach., Insect Biochem. Mol. Biol. 76 (2016) 49–61. https://doi.org/10.1016/j.ibmb.2016.07.001.

[9] G. Bianco, N. Agerbirk, I. Losito, T.R.I. Cataldi, Acylated glucosinolates with diverse acyl groups

investigated by high resolution mass spectrometry and infrared multiphoton dissociation., Phytochemistry. 100 (2014) 92–102. https://doi.org/10.1016/j.phytochem.2014.01.010.

[10] D.B. Kassel, Combinatorial chemistry and mass spectrometry in the 21st century drug discovery laboratory, Chem. Rev. 101 (2001) 255–267. https://doi.org/10.1021/cr990085q.

[11] Y.G. Shin, R.B. van Breemen, Analysis and screening of combinatorial libraries using mass spectrometry, Biopharm. Drug Dispos. 22 (2001) 353–372. https://doi.org/10.1002/bdd.278.

[12] D. Gibson, C.E. Costello, Mass spectrometry of biomolecules, Elsevier Inc., 2000. https://doi.org/10.1016/S0149-6395(00)80053-7.

[13] L. Yin, Z. Zhang, Y. Liu, Y. Gao, J. Gu, Recent advances in single-cell analysis by mass spectrometry, Analyst. 144 (2019) 824–845. https://doi.org/10.1039/c8an01190g.

[14] C. Fenselau, R. Caprioli, A.O. Nier, W.B. Hanson, A. Seiff, M.B. Mcelroy, N.W. Spencer, R.J. Duckett, T.C.D. Knight, W.S. Cook, K. Biemann, J. Oro, P. Toulmin, L.E. Orgel, A.O. Nier, D.M. Anderson, P.G. Simmonds, D. Flory, A. V. Diaz, D.R. Rushneck, J.A. Biller, T. Owen, K. Biemann, Mass spectrometry in the exploration of Mars, J. Mass Spectrom. 38 (2003) 1–10. https://doi.org/10.1002/jms.396.

[15] A. Onzo, M.A. Acquavia, T.R.I. Cataldi, M. Ligonzo, D. Coviello, R. Pascale, G. Martelli, M. Bondoni, L. Scrano, G. Bianco, Coceth Sulfate Characterization by Electrospray Ionization with Tandem Mass Spectrometry, Rapid Commun. Mass Spectrom. (2020). https://doi.org/10.1002/rcm.8884.

[16] A. Buchicchio, G. Bianco, A. Sofo, S. Masi, D. Caniani, Biodegradation of carbamazepine and clarithromycin by Trichoderma harzianum and Pleurotus ostreatus investigated by liquid chromatography - high-resolution tandem mass spectrometry (FTICR MS-IRMPD)., Sci. Total Environ. 557–558 (2016) 733–9. https://doi.org/10.1016/j.scitotenv.2016.03.119.

[17] J.H. Gross, Mass Spectrometry - A Textbook, Springer International Publishing, 2017. https://doi.org/10.1007/978-3-319-54398-7.

[18] J.R. Jocelyn Paré, V. Yaylayan, Chapter 7 Mass spectrometry: Principles and applications, Tech. Instrum. Anal. Chem. 18 (1997) 239–266. https://doi.org/10.1016/S0167-9244(97)80016-9.

[19] H. Awad, M.M. Khamis, A. El-Aneed, Mass spectrometry, review of the basics: Ionization, Appl. Spectrosc. Rev. 50 (2015) 158–175. https://doi.org/10.1080/05704928.2014.954046.

[20] A. El-Aneed, A. Cohen, J. Banoub, Mass spectrometry, review of the basics: Electrospray, MALDI, and commonly used mass analyzers, Appl. Spectrosc. Rev. 44 (2009) 210–230. https://doi.org/10.1080/05704920902717872.

[21] G.A. Valaskovic, N.L. Kelleher, D.P. Little, D.J. Aaserud, F.W. McLafferty, Attomole-Sensitivity Electrospray Source for Large-Molecule Mass Spectrometry, Anal. Chem. 67 (1995) 3802–3805. https://doi.org/10.1021/ac00116a030.

[22] D.P. Demarque, A.E.M. Crotti, R. Vessecchi, J.L.C. Lopes, N.P. Lopes, Fragmentation reactions using electrospray ionization mass spectrometry: An important tool for the structural elucidation and characterization of synthetic and natural products, Nat. Prod. Rep. 33 (2016) 432–455. https://doi.org/10.1039/c5np00073d.

[23] J.A. Yergey, A general approach to calculating isotopic distributions for mass spectrometry, Int. J. Mass Spectrom. Ion Phys. 52 (1983) 337–349. https://doi.org/10.1016/0020-7381(83)85053-0.

[24] J. Meija, Understanding isotopic distributions in mass spectrometry, J. Chem. Educ. 83 (2006) 1761. https://doi.org/10.1021/ed083p1761.2.

[25] T. Kind, O. Fiehn, Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry, BMC Bioinformatics. 8 (2007) 105. https://doi.org/10.1186/1471-2105-8-105.

[26] G. Alves, A.Y. Ogurtsov, Y.-K. Yu, Molecular Isotopic Distribution Analysis (MIDAs) with Adjustable Mass Accuracy, J. Am. Soc. Mass Spectrom. 25 (2014) 57–70. https://doi.org/10.1007/s13361-013-0733-7.

[27] A. G. Marshall, G. T. Blakney, T. Chen, N. K. Kaiser, A. M. McKenna, R. P. Rodgers, B. M. Ruddy, F. Xian, Mass Resolution and Mass Accuracy: How Much Is Enough?, Mass Spectrom. 2 (2013) S0009–S0009. https://doi.org/10.5702/massspectrometry.s0009.

[28] A. Gareth Brenton, A. Ruth Godfrey, Accurate Mass Measurement: Terminology and Treatment of Data, (2010). https://doi.org/10.1016/j.jasms.2010.06.006.

[29] A. Santarsiero, A. Onzo, R. Pascale, M.A. Acquavia, M. Coviello, P. Convertini, S. Todisco, M. Marsico, C. Pifano, P. Iannece, C. Gaeta, S. D'Angelo, M.C. Padula, G. Bianco, V. Infantino, G. Martelli, Pistacia lentiscus Hydrosol: Untargeted Metabolomic Analysis and Anti-Inflammatory Activity Mediated by NF-κB and the Citrate Pathway, Oxid. Med. Cell. Longev. 2020 (2020) 1–14. https://doi.org/10.1155/2020/4264815.

[30] R. Pascale, G. Bianco, T.R.I. Cataldi, P.S. Kopplin, F. Bosco, L. Vignola, J. Uhl, M. Lucio, L. Milella, Mass spectrometry-based phytochemical screening for hypoglycemic activity of Fagioli di Sarconi beans (Phaseolus vulgaris L.), Food Chem. 242 (2018) 497–504. https://doi.org/10.1016/j.foodchem.2017.09.091.

[31] N. Mirsaleh-Kohan, W.D. Robertson, R.N. Compton, Electron ionization time-of-flight mass spectrometry: Historical review and current applications, Mass Spectrom. Rev. 27 (2008) 237–285. https://doi.org/10.1002/mas.20162.

[32] D.R. Luffer, K.H. Schram, Electron ionization mass spectrometry of synthetic C60, Rapid Commun. Mass Spectrom. 4 (1990) 552–556. https://doi.org/10.1002/rcm.1290041218.

[33] T. Kind, O. Fiehn, Advances in structure elucidation of small molecules using mass spectrometry, Bioanal. Rev. 2 (2010) 23–60. https://doi.org/10.1007/s12566-010-0015-9.

[34] N. Ferreira, L. Sigaud, E.C. Montenegro, Molecular fragmentation by electron impact investigated using a time delayed spectroscopic technique, in: J. Phys. Conf. Ser., Institute of Physics Publishing, 2014: p. 12042. https://doi.org/10.1088/1742-6596/488/1/012042.

[35] R.W. Giese, Electron-capture mass spectrometry: Recent advances, J. Chromatogr. A. 892 (2000) 329–346. https://doi.org/10.1016/S0021-9673(00)00364-2.

[36] G.J. Feistner, N. Pascoe, K.F. Faull, K.B. Tomer, Gas chromatography/electron impact mass spectrometry, fast atom bombardment mass spectrometry, mass-analyzed ion kinetic energy spectroscopy andB/E linked scan analysis of triaryl phosphates and triethylene glycol fatty acid esters, Biol. Mass Spectrom. 19 (1990) 151–158. https://doi.org/10.1002/bms.1200190310.

[37] K.B. Scheidweiler, M.A. Huestis, A validated gas chromatographic-electron impact ionization mass spectrometric method for methylenedioxymethamphetamine (MDMA), methamphetamine and metabolites in oral fluid, J. Chromatogr. B Anal. Technol. Biomed. Life Sci. 835 (2006) 90–99. https://doi.org/10.1016/j.jchromb.2006.03.020.

[38] M.S.B. Munson, F.H. Field, Chemical Ionization Mass Spectrometry. I. General Introduction, J. Am. Chem. Soc. 88 (1966) 2621–2630. https://doi.org/10.1021/ja00964a001.

[39] J.B. Fenn, M. Mann, C.K. Meng, S.F. Wong, C.M. Whitehouse, Electrospray ionization for mass spectrometry of large biomolecules, Science (80-.). 246 (1989) 64–71. https://doi.org/10.1126/science.2675315.

[40] V. V. Laiko, M.A. Baldwin, A.L. Burlingame, Atmospheric pressure matrix-assisted laser desorption/ionization mass spectrometry, Anal. Chem. 72 (2000) 652–657. https://doi.org/10.1021/ac990998k.

[41] Y. Zhang, W. Zeng, L. Huang, W. Liu, E. Jia, Y. Zhao, F. Wang, Z. Zhu, In Situ Liquid Secondary Ion Mass Spectrometry: A Surprisingly Soft Ionization Process for Investigation of Halide Ion Hydration, Anal. Chem. 91 (2019) 7039–7046. https://doi.org/10.1021/acs.analchem.8b05804.

[42] M. Barber, R.S. Bordoli, G.J. Elliott, R.D. Sedgwick, A.N. Tyler, Fast Atom Bombardment Mass Spectrometry, Anal. Chem. 54 (1982) 645–657. https://doi.org/10.1021/ac00241a002.

[43] M.A. Baldwin, Mass spectrometers for the analysis of biomolecules, Methods Enzymol. 402 (2005) 3–48. https://doi.org/10.1016/S0076-6879(05)02001-X.

[44] J. Gieniec, L.L. Mack, K. Nakamae, C. Gupta, V. Kumar, M. Dole, Electrospray mass spectroscopy of macromolecules: Application of an ion-drift spectrometer, Biol. Mass Spectrom. 11 (1984) 259–268. https://doi.org/10.1002/bms.1200110602.

[45] M. Wilm, Principles of electrospray ionization, Mol. Cell. Proteomics. 10 (2011). https://doi.org/10.1074/mcp.M111.009407.

[46] R. Knochenmuss, Ion formation mechanisms in UV-MALDI, Analyst. 131 (2006) 966–986. https://doi.org/10.1039/b605646f.

[47] M. Niehaus, J. Soltwisch, New insights into mechanisms of material ejection in MALDI mass spectrometry for a wide range of spot sizes, Sci. Rep. 8 (2018) 1–10. https://doi.org/10.1038/s41598-018-25946-z.

[48] M. Karas, R. Krüger, Ion formation in MALDI: The cluster ionization mechanism, Chem. Rev. 103 (2003) 427–439. https://doi.org/10.1021/cr010376a.

[49] I. Fournier, C. Marinach, J.C. Tabet, G. Bolbach, Irradiation effects in MALDI, ablation, ion production, and surface modifications. Part II: 2,5-dihydroxybenzoic acid monocrystals, J. Am. Soc. Mass Spectrom. 14 (2003) 893–899. https://doi.org/10.1016/S1044-0305(03)00347-7.

[50] R. de Oliveira Silva, R.C. de Castro, M.A.L. Milhome, R.F. Do Nascimento, Liquid chromatography-electrospray ionization-tandem mass spectrometry method for determination of twenty multi-class pesticide residues in cashew, LWT - Food Sci. Technol. 59 (2014) 21–25. https://doi.org/10.1016/j.lwt.2014.05.035.

[51] P. Jiang, C.A. Lucy, Coupling normal phase liquid chromatography with electrospray ionization mass spectrometry: Strategies and applications, Anal. Methods. 8 (2016) 6478–6488. https://doi.org/10.1039/c6ay01419d.

[52] R. Pascale, M.A. Acquavia, T.R.I. Cataldi, A. Onzo, D. Coviello, S.A. Bufo, L. Scrano, R. Ciriello, A. Guerrieri, G. Bianco, Profiling of quercetin glycosides and acyl glycosides in sun-dried peperoni di Senise peppers (Capsicum annuum L.) by a combination of LC-ESI(-)-MS/MS and polarity prediction in reversed-phase separations, Anal. Bioanal. Chem. 412 (2020) 3005–3015. https://doi.org/10.1007/s00216-020-02547-2.

[53] G.L. Glish, S.A. McLuckey, T.Y. Ridley, R.G. Cooks, A new “hybrid” sector/quadrupole mass spectrometer for mass spectrometry/mass spectrometry, Int. J. Mass Spectrom. Ion Phys. 41 (1982) 157–177. https://doi.org/10.1016/0020-7381(82)85032-8.

[54] R.A. Yost, C.G. Enke, Selected Ion Fragmentation with a Tandem Quadrupole Mass Spectrometer, J. Am. Chem. Soc. 100 (1978) 2274–2275. https://doi.org/10.1021/ja00475a072.

[55] D.J. Douglas, Linear quadrupoles in mass spectrometry, Mass Spectrom. Rev. 28 (2009) 937–960. https://doi.org/10.1002/mas.20249.

[56] H. Steinwedel, W. Paul, Apparatus for separating charged particles of different specific charges, US2939952A, 1960.

[57] P.H. Dawson, ed., Quadrupole Mass Spectrometry and its Applications, Elsevier, 1976. https://doi.org/10.1016/c2013-0-04436-2.

[58] I. Szabo, New ion-optical devices utilizing oscillatory electric fields. I. Principle of operation and analytical theory of multipole devices with two-dimensional electric fields, Int. J. Mass Spectrom. Ion Process. 73 (1986) 197–235. https://doi.org/10.1016/0168-1176(86)80001-5.

[59] M.H. Friedman, A.L. Yergey, J.E. Campana, Fundamentals of ion motion in electric radio-frequency multipole fields, J. Phys. E. 15 (1982) 53–56. https://doi.org/10.1088/0022-3735/15/1/010.

[60] J.E. Campana, Elementary theory of the quadrupole mass filter, Int. J. Mass Spectrom. Ion Phys. 33 (1980) 101–117. https://doi.org/10.1016/0020-7381(80)80042-8.

[61] C.M. Whitehouse, B.A. Andrien, E.E. Gulcicek, Mass spectrometry with multipole ion guides, CA2332534C, 1999.

[62] M. Shimomura, Triple quadrupole mass spectrometer, US8803086B2, 2011.

[63] D.J. Douglas, A.J. Frank, D. Mao, Linear ion traps in mass spectrometry, Mass Spectrom. Rev. 24 (2005) 1–29. https://doi.org/10.1002/mas.20004.

[64] S.M. Peterman, C.P. Dufresne, S. Horning, The use of a hybrid linear trap/FT-ICR mass spectrometer for on-line high resolution/high mass accuracy bottom-up sequencing, J. Biomol. Tech. 16 (2005) 112–124.

[65] C.S. Ejsing, T. Moehring, U. Bahr, E. Duchoslav, M. Karas, K. Simons, A. Shevchenko, Collision-induced dissociation pathways of yeast sphingolipids and their molecular profiling in total lipid extracts: A study by quadrupole TOF and linear ion trap-orbitrap mass spectrometry, J. Mass Spectrom. 41 (2006) 372–389. https://doi.org/10.1002/jms.997.

[66] E.O. Lawrence, M.S. Livingston, The production of high speed light ions without the use of high voltages, Phys. Rev. 40 (1932) 19–35. https://doi.org/10.1103/PhysRev.40.19.

[67] Y. Qi, P.B. O'Connor, Data processing in Fourier transform ion cyclotron resonance mass spectrometry, Mass Spectrom. Rev. 33 (2014) 333–352. https://doi.org/10.1002/mas.21414.

[68] E.N. Nikolaev, Y.I. Kostyukevich, G.N. Vladimirov, Fourier transform ion cyclotron resonance (FT ICR) mass spectrometry: Theory and simulations, Mass Spectrom. Rev. 35 (2016) 219–258. https://doi.org/10.1002/mas.21422.

[69] M.B. Comisarow, A.G. Marshall, Fourier transform ion cyclotron resonance spectroscopy, Chem. Phys. Lett. 25 (1974) 282–283. https://doi.org/10.1016/0009-2614(74)89137-2.

[70] D.F. Smith, D.C. Podgorski, R.P. Rodgers, G.T. Blakney, C.L. Hendrickson, 21 Tesla FT-ICR Mass Spectrometer for Ultrahigh-Resolution Analysis of Complex Organic Mixtures, Anal. Chem. 90 (2018) 2041–2047. https://doi.org/10.1021/acs.analchem.7b04159.

[71] E. Nikolaev, Ion cyclotron resonance mass spectrometer, US7038200B2, 2004.

[72] M.B. Comisarow, A.G. Marshall, Frequency-sweep fourier transform ion cyclotron resonance spectroscopy, Chem. Phys. Lett. 26 (1974) 489–490. https://doi.org/10.1016/0009-2614(74)80397-0.

[73] A.G. Marshall, C.L. Hendrickson, Fourier transform ion cyclotron resonance detection: Principles and experimental configurations, Int. J. Mass Spectrom. 215 (2002) 59–75. https://doi.org/10.1016/S1387-3806(01)00588-7.

[74] Y.I. Kostyukevich, G.N. Vladimirov, E.N. Nikolaev, Dynamically harmonized FT-ICR cell with specially shaped electrodes for compensation of inhomogeneity of the magnetic field. Computer simulations of the electric field and ion motion dynamics, J. Am. Soc. Mass Spectrom. 23 (2012) 2198–2207. https://doi.org/10.1007/s13361-012-0480-1.

[75] E.N. Nikolaev, I.A. Boldin, R. Jertz, G. Baykut, Initial experimental characterization of a new ultra-high resolution FTICR cell with dynamic harmonization, J. Am. Soc. Mass Spectrom. 22 (2011) 1125–1133. https://doi.org/10.1007/s13361-011-0125-9.

[76] R. Oshana, Overview of Digital Signal Processing Algorithms, in: DSP Softw. Dev. Tech. Embed. Real-Time Syst., Elsevier, 2006: pp. 59–121. https://doi.org/10.1016/b978-075067759-2/50006-5.

[77] A.M. Brustkern, D.L. Rempel, M.L. Gross, Ion behavior in an electrically compensated ion cyclotron resonance trap, Int. J. Mass Spectrom. 300 (2011) 143–148. https://doi.org/10.1016/j.ijms.2010.06.027.

[78] I.A. Boldin, E.N. Nikolaev, Fourier transform ion cyclotron resonance cell with dynamic harmonization

of the electric field in the whole volume by shaping of the excitation and detection electrode assembly, Rapid Commun. Mass Spectrom. 25 (2011) 122–126. https://doi.org/10.1002/rcm.4838.

[79] R.D. Gougeon, M. Lucio, L. Boutegrabet, D. Peyron, F. Feuillat, D. Chassagne, H. Alexandre, A. Voilley, P. Cayot, I. Gebefügi, N. Hertkorn, P. Schmitt-Kopplin, Authentication approach of the chemodiversity of grape and wine by FTICR-MS, in: ACS Symp. Ser., American Chemical Society, 2011: pp. 69–88. https://doi.org/10.1021/bk-2011-1081.ch005.

[80] C. Roullier-Gall, M. Witting, R.D. Gougeon, P. Schmitt-Kopplin, High precision mass measurements for wine metabolomics, Front. Chem. 2 (2014) 102. https://doi.org/10.3389/fchem.2014.00102.

[81] E.B. Ledford, D.L. Rempel, M.L. Gross, Space Charge Effects in Fourier Transform Mass Spectrometry. Mass Calibration, Anal. Chem. 56 (1984) 2744–2748. https://doi.org/10.1021/ac00278a027.

[82] N.K. Kaiser, J.P. Quinn, G.T. Blakney, C.L. Hendrickson, A.G. Marshall, A Novel 9.4 Tesla FTICR Mass Spectrometer with Improved Sensitivity, Mass Resolution, and Mass Range, J. Am. Soc. Mass Spectrom. 22 (2011). https://doi.org/10.1007/s13361-011-0141-9.

[83] G. Vladimirov, C.L. Hendrickson, G.T. Blakney, A.G. Marshall, R.M.A. Heeren, E.N. Nikolaev, Fourier transform ion cyclotron resonance mass resolution and dynamic range limits calculated by computer modeling of ion cloud motion, J. Am. Soc. Mass Spectrom. 23 (2012) 375–384. https://doi.org/10.1007/s13361-011-0268-8.

[84] N.K. Kaiser, J.E. Bruce, Reduction of ion magnetron motion and space charge using radial electric field modulation, Int. J. Mass Spectrom. 265 (2007) 271–280. https://doi.org/10.1016/j.ijms.2007.02.040.

[85] C.L. Hendrickson, S.A. Hofstadler, S.C. Beu, D.A. Laude, Initiation of coherent magnetron motion following ion injection into a Fourier transform ion cyclotron resonance trapped ion cell, Int. J. Mass Spectrom. Ion Process. 123 (1993) 49–58. https://doi.org/10.1016/0168-1176(93)87053-U.

[86] A.G. Marshall, Theoretical Signal-to-Noise Ratio and Mass Resolution in Fourier Transform Ion Cyclotron Resonance Mass Spectrometry, Anal. Chem. 51 (1979) 1710–1714. https://doi.org/10.1021/ac50047a029.

[87] E.C. Craig, I. Santos, A.G. Marshall, N.M.M. Nibbering, Dispersion versus absorption (DISPA) method for automatic phasing of fourier transform ion cyclotron resonance mass spectra, Rapid Commun. Mass Spectrom. 1 (1987) 33–37. https://doi.org/10.1002/rcm.1290010209.

[88] D.P.A. Kilgour, S.L. Van Orden, Absorption mode Fourier transform mass spectrometry with no baseline correction using a novel asymmetric apodization function, Rapid Commun. Mass Spectrom.

29 (2015) 1009–1018. https://doi.org/10.1002/rcm.7190.

[89] Z. Liang, A.G. Marshall, Precise Relative Ion Abundances from Fourier Transform Ion Cyclotron Resonance Magnitude-Mode Mass Spectra, Anal. Chem. 62 (1990) 70–75. https://doi.org/10.1021/ac00200a013.

[90] Y. Qi, M.P. Barrow, H. Li, J.E. Meier, S.L. Van Orden, C.J. Thompson, P.B. O'Connor, Absorption-mode: The next generation of Fourier transform mass spectra, Anal. Chem. 84 (2012) 2923–2929. https://doi.org/10.1021/ac3000122.

[91] D.P.A. Kilgour, M.J. Neal, A.J. Soulby, P.B. O'Connor, Improved optimization of the Fourier transform ion cyclotron resonance mass spectrometry phase correction function using a genetic algorithm, Rapid Commun. Mass Spectrom. 27 (2013) 1977–1982. https://doi.org/10.1002/rcm.6658.

[92] D.P.A. Kilgour, R. Wills, Y. Qi, P.B. O'Connor, Autophaser: An algorithm for automated generation of absorption mode spectra for FT-ICR MS, Anal. Chem. 85 (2013) 3903–3911. https://doi.org/10.1021/ac303289c.

[93] Y. Qi, H. Li, R.H. Wills, P. Perez-Hurtado, X. Yu, D.P.A. Kilgour, M.P. Barrow, C. Lin, P.B. O'Connor, Absorption-Mode Fourier Transform Mass Spectrometry: The Effects of Apodization and Phasing on Modified Protein Spectra, J. Am. Soc. Mass Spectrom. 24 (2013) 828–834. https://doi.org/10.1007/s13361-013-0600-6.

[94] L. Zhang, Y. Zhang, S. Zhao, K.H. Chung, C. Xu, Q. Shi, Effect of apodization on FT-ICR mass spectrometry analysis of petroleum, Int. J. Mass Spectrom. 373 (2014) 27–33. https://doi.org/10.1016/j.ijms.2014.08.030.

[95] J.P. Lee, M.B. Comisarow, Advantageous Apodization Functions for Absorption-Mode Fourier Transform Spectroscopy, Appl. Spectrosc. 43 (1989) 599–604. https://doi.org/10.1366/0003702894202517.

[96] J.P. Lee, M.B. Comisarow, Advantageous Apodization Functions for Magnitude-Mode Fourier Transform Spectroscopy, Appl. Spectrosc. 41 (1987) 93–98. https://doi.org/10.1366/0003702874868016.

[97] K.L. Goodner, K.E. Milgram, K.R. Williams, C.H. Watson, J.R. Eyler, Quantitation of ion abundances in Fourier transform ion cyclotron resonance mass spectrometry, J. Am. Soc. Mass Spectrom. 9 (1998) 1204–1212. https://doi.org/10.1016/S1044-0305(98)00090-7.

[98] M.B. Comisarow, J.D. Melka, Error Estimates for Finite Zero-Filling in Fourier Transform Spectrometry, Anal. Chem. 51 (1979) 2198–2203. https://doi.org/10.1021/ac50049a032.

[99] T.J. Francl, M.G. Sherman, R.L. Hunter, M.J. Locke, W.D. Bowers, R.T. McIver, Experimental determination of the effects of space charge on ion cyclotron resonance frequencies, Int. J. Mass Spectrom. Ion Process. 54 (1983) 189–199. https://doi.org/10.1016/0168-1176(83)85017-4.

[100] S.D.H. Shi, J.J. Drader, M.A. Freitas, C.L. Hendrickson, A.G. Marshall, Comparison and interconversion of the two most common frequency-to-mass calibration functions for Fourier transform ion cyclotron resonance mass spectrometry, Int. J. Mass Spectrom. 195–196 (2000) 591–598. https://doi.org/10.1016/S1387-3806(99)00226-2.

[101] D.C. Muddiman, A.L. Oberg, Statistical evaluation of internal and external mass calibration laws utilized in Fourier transform ion cyclotron resonance mass spectrometry, Anal. Chem. 77 (2005) 2406–2414. https://doi.org/10.1021/ac048258l.

[102] P.B. O'Connor, C.E. Costello, Internal calibration on adjacent samples (InCAS) with fourier transform mass spectrometry, Anal. Chem. 72 (2000) 5881–5885. https://doi.org/10.1021/ac000770t.

[103] M.L. Easterling, T.H. Mize, I.J. Amster, Routine part-per-million mass accuracy for high-mass ions: Space-charge effects in MALDI FT-ICR, Anal. Chem. 71 (1999) 624–632. https://doi.org/10.1021/ac980690d.

[104] P.K. Taylor, I.J. Amster, Space charge effects on mass accuracy for multiply charged ions in ESI-FTICR, Int. J. Mass Spectrom. 222 (2003) 351–361. https://doi.org/10.1016/S1387-3806(02)00994-6.

[105] K. Aizikov, R. Mathur, P.B. O'Connor, The spontaneous loss of coherence catastrophe in fourier transform ion cyclotron resonance mass spectrometry, J. Am. Soc. Mass Spectrom. 20 (2009) 247–256. https://doi.org/10.1021/jasms.8b03385.

[106] C. Masselon, A. V. Tolmachev, G.A. Anderson, R. Harkewicz, R.D. Smith, Mass measurement errors caused by "local" frequency perturbations in FTICR mass spectrometry, J. Am. Soc. Mass Spectrom. 13 (2002) 99–106. https://doi.org/10.1016/S1044-0305(01)00333-6.

[107] R. Mathur, P.B. O'Connor, Artifacts in Fourier transform mass spectrometry, Rapid Commun. Mass Spectrom. 23 (2009) 523–529. https://doi.org/10.1002/rcm.3904.

[108] Z. Lin, Characterization of harmonics and multi-charged peaks obtained by Fourier transform ion cyclotron resonance mass spectrometry, Instrum. Sci. Technol. 46 (2018) 307–315. https://doi.org/10.1080/10739149.2017.1383270.

[109] K. Yang, X. Fang, R.W. Gross, X. Han, A practical approach for determination of mass spectral baselines, J. Am. Soc. Mass Spectrom. 22 (2011) 2090–2099. https://doi.org/10.1007/s13361-011-0229-2.

[110] B. Williams, S. Cornett, A. Crecelius, R. Caprioli, B. Dawant, B. Bodenheimer, An algorithm for baseline correction of MALDI mass spectra, in: Proc. Annu. Southeast Conf., ACM Press, New York, New York, USA, 2005: pp. 1137–1142. https://doi.org/10.1145/1167350.1167394.

[111] X. Liu, Z. Zhang, P.F.M. Sousa, C. Chen, M. Ouyang, Y. Wei, Y. Liang, Y. Chen, C. Zhang, Selective iteratively reweighted quantile regression for baseline correction, Anal. Bioanal. Chem. 406 (2014) 1985–1998. https://doi.org/10.1007/s00216-013-7610-x.

[112] J. Urban, J. Vaněk, D. Štys, Unsupervised adaptive filter for baseline thresholding and elimination in liquid chromatography-mass spectrometry via approximation of the standard deviation of baseline distribution in retention time domain, Acta Chromatogr. 25 (2013) 257–273. https://doi.org/10.1556/AChrom.25.2013.2.4.

[113] A.A. Trubitsyn, A.B. Tolstoguzov, V.S. Gurov, Smoothing of mass spectra by a normalizing window with automatically adjustable width, J. Anal. Chem. 69 (2014) 1259–1263. https://doi.org/10.1134/S1061934814130127.

[114] A. Savitzky, M.J.E. Golay, Smoothing and Differentiation of Data by Simplified Least Squares Procedures, Anal. Chem. 36 (1964) 1627–39. https://doi.org/10.1021/ac60214a047.

[115] P.J. Roach, J. Laskin, A. Laskin, Molecular characterization of organic aerosols using nanospray-desorption/ electrospray ionization-mass spectrometry, Anal. Chem. 82 (2010) 7979–7986. https://doi.org/10.1021/ac101449p.

[116] A.S. Wozniak, J.E. Bauer, R.L. Sleighter, R.M. Dickhut, P.G. Hatcher, Technical Note: Molecular characterization of aerosol-derived water soluble organic carbon using ultrahigh resolution electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry, Atmos. Chem. Phys. 8 (2008) 5099–5111. https://doi.org/10.5194/acp-8-5099-2008.

[117] B.P. Koch, M. Witt, R. Engbrodt, T. Dittmar, G. Kattner, Molecular formulae of marine and terrigenous dissolved organic matter detected by electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry, Geochim. Cosmochim. Acta. 69 (2005) 3299–3308. https://doi.org/10.1016/j.gca.2005.02.027.

[118] C.C.L. Wong, D. Cociorva, J.D. Venable, T. Xu, J.R. Yates, Comparison of different signal thresholds on data dependent sampling in orbitrap and LTQ mass spectrometry for the identification of peptides and proteins in complex mixtures, J. Am. Soc. Mass Spectrom. 20 (2009) 1405–1414. https://doi.org/10.1021/jasms.8b03525.

[119] T.B. Nguyen, A.P. Bateman, D.L. Bones, S.A. Nizkorodov, J. Laskin, A. Laskin, High-resolution mass spectrometry analysis of secondary organic aerosol generated by ozonolysis of isoprene, Atmos.

Environ. 44 (2010) 1032–1042. https://doi.org/10.1016/j.atmosenv.2009.12.019.

[120] K.O. Zhurov, A.N. Kozhinov, L. Fornelli, Y.O. Tsybin, Distinguishing analyte from noise components in mass spectra of complex samples: Where to cut the noise?, Anal. Chem. 86 (2014) 3308–3316. https://doi.org/10.1021/ac403278t.

[121] D. Freedman, P. Diaconis, On the histogram as a density estimator:L2 theory, Zeitschrift Für Wahrscheinlichkeitstheorie Und Verwandte Gebiete. 57 (1981) 453–476. https://doi.org/10.1007/BF01025868.

[122] A. Makarov, E. Denisov, O. Lange, S. Horning, Dynamic range of mass accuracy in LTQ orbitrap hybrid mass spectrometer, Off. J. Am. Soc. Mass Spectrom. 17 (2006) 977–982. https://doi.org/10.1021/jasms.8b02700.

[123] C.A. Hughey, R.P. Rodgers, A.G. Marshall, Resolution of 11 000 compositionally distinct components in a single electrospray ionization fourier transform ion cyclotron resonance mass spectrum of crude oil, Anal. Chem. 74 (2002) 4145–4149. https://doi.org/10.1021/ac020146b.

[124] A.G. Marshall, R.P. Rodgers, Petroleomics: The Next Grand Challenge for Chemical Analysis, Acc. Chem. Res. 37 (2004) 53–59. https://doi.org/10.1021/ar020177t.

[125] A.C. Stenson, A.G. Marshall, W.T. Cooper, Exact masses and chemical formulas of individual Suwannee River fulvic acids from ultrahigh resolution electrospray ionization Fourier transform ion cyclotron resonance mass spectra, Anal. Chem. 75 (2003) 1275–1284. https://doi.org/10.1021/ac026106p.

[126] J.K. Senior, Partitions and Their Representative Graphs, Am. J. Math. 73 (1951) 663. https://doi.org/10.2307/2372318.

[127] T. Morikawa, B.T. Newbold, Analogous Odd-Even Parities in Mathematics and Chemistry, Chemistry (Easton). 12 (2003) 445–450.

[128] J. Claesen, D. Valkenborg, T. Burzykowski, The (generalized) hydrogen rule for organic molecules, J. Mass Spectrom. 55 (2020) e4485. https://doi.org/10.1002/jms.4485.

[129] J.R. de Laeter, J.K. Böhlke, P. De Bièvre, H. Hidaka, H.S. Peiser, K.J.R. Rosman, P.D.P. Taylor, Atomic weights of the elements. Review 2000 (IUPAC Technical Report), Pure Appl. Chem. 75 (2003) 683–800. https://doi.org/10.1351/pac200375060683.

[130] V. Pellegrin, Molecular formulas of organic compounds the nitrogen rule and degree of unsaturation, J. Chem. Educ. 60 (1983) 626–633. https://doi.org/10.1021/ed060p626.

[131] C.S. Hsu, K. Qlan, Y.C. Chen, An innovative approach to data analysis in hydrocarbon characterization by on-line liquid, Anal. Chim. Acta. 264 (1992) 79–89.

[132] K. Qian, R.P. Rodgers, C.L. Hendrickson, C.A. Hughey, A.G. Marshall, Kendrick Mass Defect Spectrum: A Compact Visual Analysis for Ultrahigh-Resolution Broadband Mass Spectra, Anal. Chem. 73 (2002) 4676–4681. https://doi.org/10.1021/ac010560w.

[133] N. Kuhnert, F. Dairpoosh, G. Yassin, A. Golon, R. Jaiswal, What is under the hump? Mass spectrometry based analysis of complex mixtures in processed food-lessons from the characterisation of black tea thearubigins, coffee melanoidines and caramel, Food Funct. 4 (2013) 1130–1147. https://doi.org/10.1039/c3fo30385c.

[134] D.W. Van Krevelen, Graphical-statistical method for the study of structure and reaction processes of coal, Fuel. 29 (1950) 269–284.

[135] S. Kim, R.W. Kramer, P.G. Hatcher, Graphical Method for Analysis of Ultrahigh-Resolution Broadband Mass Spectra of Natural Organic Matter, the Van Krevelen Diagram, Anal. Chem. 75 (2003) 5336–5344. https://doi.org/10.1021/ac034415p.

[136] N. Martins, N.T. Jiménez-Morillo, F. Freitas, R. Garcia, M. Gomes da Silva, M.J. Cabrita, Revisiting 3D van Krevelen diagrams as a tool for the visualization of volatile profile of varietal olive oils from Alentejo region, Portugal, Talanta. 207 (2020) 120276. https://doi.org/10.1016/j.talanta.2019.120276.

[137] J. Heo, Y. Yoon, D.H. Kim, H. Lee, D. Lee, N. Her, A new fluorescence index with a fluorescence excitation-emission matrix for dissolved organic matter (DOM) characterization, Desalin. Water Treat. 57 (2016) 20270–20282. https://doi.org/10.1080/19443994.2015.1110719.

[138] F. Moritz, M. Kaling, J.-P. Schnitzler, P. Schmitt-Kopplin, Characterization of poplar metabotypes via mass difference enrichment analysis, Plant. Cell Environ. 40 (2017) 1057–1073. https://doi.org/10.1111/pce.12878.

[139] K. Longnecker, E.B. Kujawinski, Using network analysis to discern compositional patterns in ultrahigh-resolution mass spectrometry data of dissolved organic matter, Rapid Commun. Mass Spectrom. (2016) 2388–2394. https://doi.org/10.1002/rcm.7719.

[140] D. Tziotis, N. Hertkorn, P. Schmitt-Kopplin, Kendrick-Analogous Network Visualisation of Ion Cyclotron Resonance Fourier Transform Mass Spectra: Improved Options for the Assignment of Elemental Compositions and the Classification of Organic Molecular Complexity, Eur. J. Mass Spectrom. 17 (2011) 415–421. https://doi.org/10.1255/ejms.1135.

[141] U. Brandes, Force-Directed Graph Drawing, in: Encycl. Algorithms, Springer US, 2014: pp. 1–6.

https://doi.org/10.1007/978-3-642-27848-8_648-1.

[142] K. O'Shea, B.B. Misra, Software tools, databases and resources in metabolomics: updates from 2018 to 2019, Metabolomics. 16 (2020) 36. https://doi.org/10.1007/s11306-020-01657-3.

[143] L. Gatto, A. Christoforou, Using R and bioconductor for proteomics data analysis, Biochim. Biophys. Acta - Proteins Proteomics. 1844 (2014) 42–51. https://doi.org/10.1016/j.bbapap.2013.04.032.

[144] C.A. Smith, E.J. Want, G. O'Maille, R. Abagyan, G. Siuzdak, XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification, Anal. Chem. 78 (2006) 779–787. https://doi.org/10.1021/ac051437y.

[145] L. Gatto, S. Gibb, J. Rainer, MSnbase, Efficient and Elegant R-Based Processing and Visualization of Raw Mass Spectrometry Data, J. Proteome Res. (2020). https://doi.org/10.1021/acs.jproteome.0c00313.

[146] M.J. Crawley, The R Book, 2nd ed., Wiley, 2012.

[147] R.C. Gentleman, V.J. Carey, D.M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A.J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J.Y.H. Yang, J. Zhang, Bioconductor: open software development for computational biology and bioinformatics., Genome Biol. 5 (2004) R80. https://doi.org/10.1186/gb-2004-5-10-r80.

[148] P. Du, W.A. Kibbe, S.M. Lin, Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching, Bioinformatics. 22 (2006) 2059–2065. https://doi.org/10.1093/bioinformatics/btl355.

[149] J. Lisec, InterpretMSSpectrum: Interpreting High Resolution Mass Spectra, (2018). https://cran.r-project.org/package=InterpretMSSpectrum.

[150] M. Meringer, S. Reinker, J. Zhang, A. Muller, MS/MS Data Improves Automated Determination of Molecular Formulas by Mass Spectrometry, MATCH Commun. Math Comput. Chem. 65 (2011) 259–290.

[151] M. Loos, C. Gerber, F. Corona, J. Hollender, H. Singer, Accelerated isotope fine structure calculation using pruned transition trees, Anal. Chem. 87 (2015) 5738–5744. https://doi.org/10.1021/acs.analchem.5b00941.

[152] J. Greaves, J. Roboz, Mass Spectrometry for the Novice, CRC Press, 2013. https://doi.org/10.1201/b15436.

[153] J. Xia, N. Psychogios, N. Young, D.S. Wishart, MetaboAnalyst: A web server for metabolomic data analysis and interpretation, Nucleic Acids Res. 37 (2009) W652–W660. https://doi.org/10.1093/nar/gkp356.

[154] J. Allaire, Y. Xie, J. McPherson, J. Luraschi, K. Ushey, A. Atkins, H. Wickham, J. Cheng, C. Winston, R. Iannone, rmarkdown: Dynamic Documents for R, (2020). https://rmarkdown.rstudio.com.

[155] V. Perrier, F. Meyer, D. Granjon, shinyWidgets: Custom Inputs Widgets for Shiny, (2020). https://cran.r-project.org/package=shinyWidgets.

3. Contribute 1

Metabolic profiling of Peperoni di Senise PGI peppers by using High Resolution Mass Spectrometry and data elaboration with AutoVectis Pro

A. Onzo[1], M. Acquavia[1], R. Pascale[1], P. Iannece[2], C. Gaeta[2], Y. Tsybin[3], G. Bianco[1]*

[1]Università degli Studi della Basilicata, Dipartimento di Scienze, Via dell'Ateneo Lucano 10, Potenza, Italy;

[2]Università degli Studi di Salerno, Dipartimento di Chimica e Biologia, Via Giovanni Paolo II 132, Fisciano, Italy

[3]Spectroswiss Sarl, EPFL Innovation Park, Building I, 1015 Lausanne, Switzerland.

3.1. Abstract

Pepper fruits (genus Capsicum) are an excellent source of health-related compounds, such as ascorbic acid (vitamin C), carotenoids (provitamin A), tocopherols (vitamin E), flavonoids and capsaicinoids, each of them known for biological activities such as antioxidant, anti-inflammatory and anticarcinogenic ones. They have been used for fresh and cooked consumption, as well as for medicinal purposes, such as treatment of asthma, coughs, sore throats, and toothache. During recent years, many efforts were dedicated to the metabolic profiling of pepper fruits, in order to obtain a complete overview of the diversity of biocomponents present in analyzed samples and, thus, to be able to make some speculation on macroscopic health-promoting properties. However, as far as we know, none of the assumed approach was able to simultaneously identify all the metabolite classes and possible derivatives of model compounds present in pepper fruits with a simple and fast direct analysis. This was accomplished in this work on *Peperoni di Senise* peppers (*Capsicum Annuum L.*), a typical food product cultivated in Basilicata (Southern Italy), protected with a PGI quality mark, known for their unique taste, by performing a direct-injection Fourier Transform Ion Cyclotron Mass Spectrometry technique. Moreover, several data pre-treatment steps were followed starting from recorded free induction decay (FID) to finally obtain the absorption mode FT-ICR mass spectrum, by using the new commercial software AutoVectis Pro. This tool allowed to increase the information on metabolic profile of *Peperoni di Senise PGI* through the identification of a higher number of compounds.

3.2. Introduction

A huge number of wild and cultivated species belong to the genus *Capsicum*. *Capsicum* plants are grown all over the world, principally in tropical and subtropical countries. The genus *Capsicum*, pepper, belongs to the family *Solanaceae* and consists of up to 30 species [1,2]. Among these, the species *Capsicum Annuum L.* (pepper) is one of the five major cultivated and marketed species. The market for *Capsicum* saw a great expansion during the last years [3]. World pepper production in 2018 reached ~0.8 million tons [4]. Peppers are used as a fresh or cooked vegetable, a condiment, or a spice. The industry uses peppers as a spice or colouring agent in many food products. Capsicum fruits are a rich source of capsaicinoids, carotenoids (some of them with provitamin A activity), flavonoids, and vitamins, such as ascorbic acid (vitamin C), and tocopherols (vitamin E) [5,6]. The amount and composition of these metabolites vary among genotypes and are affected by many conditions such as fruit maturity, cultivation systems, geographical origin and processing methods [7,8]. The presence of these particular compounds provides *Capsicum* fruits some of their very well-known macroscopic properties. In detail, capsaicinoids are responsible for the hot taste of chili peppers, also known as pungency and any variation in their chemical structures, including the structure of the acyl moiety, affects the level of the pungency [9,10]. Furthermore, the colours of pepper fruits, green, red, yellow, brown, and orange, derive from carotenoids, except for the purple-fruited pepper, in which anthocyanins (flavonoid derivatives) contribute to the purple colour [11]. The presence of some of these metabolites may be employed in some defence mechanisms against various biotic and abiotic stresses [12,13]. Polyphenols such as feruloyl O-glucosides, kaempferol O-pentosyldihexosides, and dihydroxyflavone O-hexoses can act as phytoanticipins. Other polyphenols are known phytoalexins, such as N-caffeoyl putrescine and caffeoyl O-hexoside, which are induced in *C. Annuum* fruits upon infection with the fungus *Colletotrichum gloeosporioides*. In addition, capsaicin is suggested to be responsible of the defence of pepper plants against fruit-eating animals and *Fusarium* fungi. Ascorbic acid in bell pepper fruits may protect the plant against physiological disorders caused by environmental stresses, such as the calcium deficiency known as blossom-end rot [12,13]. The biochemical content of *Capsicum Annuum L.* species is not only valuable for the plant itself but may also be advantageous for human health. Different techniques were used to shed some light on metabolomic profile of peppers, such as UV/Vis, Infrared (IR) and NMR spectroscopy [14–17]. However, satisfying levels of specificity and sensitivity were reached only after the breakthrough of hyphenated chromatographic techniques, like gas and liquid chromatography coupled to Mass Spectrometry (GC and LC-MS), which allowed the separation of matrix components and, thus, a more accurate identification and quantification of specific metabolites [18–27]. With them, an improvement of the knowledge of present metabolite derivatives was possible, thus making possible to understand to which kind of biochemical pathways specific types of biocompounds are subjected and under which condition this takes place. Furthermore, extraction techniques could be optimized to isolate one or more specific metabolites, a common task to accomplish in order to prepare efficient nutraceutical products [18,24,28–31]. However, utilization of these techniques in metabolomic analysis is hampered by a series of drawbacks, such as long time of analysis and its high costs (related to the utilization of high amounts of eluents). High Resolution Fourier Transform Ion Cyclotron Resonance Mass Spectrometry (FT-ICR MS) is

able to provide a complete information of classes of metabolites present in analysed samples, being capable to identify an enormous number of ionic species simultaneously in a simple and fast direct analysis [32–36]. The promises of this approach are remarkable, simplifying and accelerating metabolomic experimental designs for biomechanistic and nutraceutical formulation purposes [32–36]. By the way, a particular attention should be paid to FT-ICR data pre-treatment. Indeed, obtained raw data cannot be readily used for observation discussion, as mass spectra could contain artefacts, such as wiggles and harmonics, which could lead to wrong formula assignments of observed accurate mass-to-charge (*m/z*) ratios [37,38]. Moreover, probability of loss of information is remarkable since FT-ICR mass spectra of complex matrices is usually noisy [39]. This latter feature could make difficult or impossible to distinguish low intensity ionic species from noise peaks. To overcome these problems, absorption mode mass spectra could be employed instead of more common magnitude mode ones, since related peaks are way narrower, thus leading to a marked improvement of the overall resolution [37]. Moreover, absorption mode mass spectra are characterized by lower levels of noise, thus allowing to identify more ionic species [37,38]. Finally, in absorption mode, artefacts could be easily identified and deleted [37]. However, obtaining a readable absorption mode mass spectrum is not straightforward, since a proper correction of the phase shifting of the ions should be made prior to result formulation, and this is not easy for complex matrix analysis [37]. For this reason, utilization of dedicated tools is compulsory to efficiently perform a proper data pre-treatment process and to extrapolate a useful absorption mode mass spectrum from raw data, which is virtually not possible with current commercial FT-ICR instrumentation [37,38]. In this work, a complete metabolic profile of a methanolic extract of *Peperoni di Senise PGI* peppers was obtained by using FT-ICR MS and elaborating raw data with the new software AutoVectis Pro, with which was possible to perform a phase correction step in few milliseconds and, thus, to obtain the absorption mode mass spectrum, thanks to which a higher number of ionic species could be identified. Moreover, utilization of a well-known visualization tool, i.e. the Van Krevelen plot, allowed to better interpret our results, leading to the identification of the classes of metabolites present in our sample and to the evaluation of the diversity of related derivatives.

3.3. Materials and Methods

Chemicals

Sodium trifluoroacetate (NaTFA, 98%) and methanol were purchased from Sigma-Aldrich (Milano, Italy). Methanol LC-MS grade was used for the analysis. Pure nitrogen (99.996%) was delivered to the MS system as the sheath gas.

Sample Preparation

Extracts of *Peperoni di Senise* peppers PGI sun-dried peppers (*Capsicum Annuum L.*) were obtained by following a modified procedure based on a previously reported method [24]. Peppers were grounded to a fine powder using a home miller and residual water was eliminated by lyophilization (24 h). 500 mg of each sample were extracted by using 1.5 mL of MeOH as solvent. Metabolites were extracted by means of the Ultrasound

Assisted Extraction (UAE) technique for 15 min at room temperature (Sonorex Super RK 100/H sonicator; Bandelin electronic, Berlin, Germany) with a 35 kHz automatic frequency control and a high-frequency power of 80 W. Extracts were passed through a PTFE 0.22 µm filter and were injected into the MS system without any further pre-treatment. A blank sample was prepared by applying every step on 1.5 mL of MeOH.

Mass spectrometry analysis

ESI (-) Fourier Transform Ion Cyclotron Resonance Mass Spectrometry (ESI-FT-ICR) technique was used to untargeted analysis of the sample. High-resolution Mass Spectra were acquired on a Bruker (Bruker Daltonik GmbH, Bremen, Germany) solariX XR Fourier transform ion cyclotron resonance mass spectrometer (FT-ICR-MS) equipped with a 7T superconducting magnet and an ESI source. The capillary voltage was set to 3.9 kV, with a nebulizer gas pressure of 1.2 bar and dry gas flow rate of 4 L/min at 200 °C. Spectra were acquired with a Time Domain size of 16 mega-word, an accumulation time of 0.1 s and a mass range of 100-2000 m/z. Moreover, the average number of scans was set to 50. Before the analysis, the mass spectrometer was externally calibrated with NaTFA. High accuracies were reached, with a root mean square (RMS) error lower than 0.1 ppm. Once recorded, FT-ICR mass spectra were submitted to several data pre-treatment steps. More specifically, recorded free induction decays (FIDs) were subjected to apodization and related absorption and magnitude mode mass spectra (aFT and mFT, respectively) were obtained. On the former, phase correction, mass recalibration and baseline correction have been performed, while the latter was processed by means of smoothing, choosing the Savitzki-Golay (SK) algorithm with a 0.001 Da range and performing 10 cycles. Finally, noise filtering has been performed on mass spectra by following the N-Sigma methodology approach [39]. More specifically, noise level has been estimated and peaks showing a signal-to-noise ratio (S/N) higher than 2 were retained. Thus, obtained FT-ICR mass spectra were exported to peak lists. From these, possible elemental formulas were calculated for each MS signal. To obtain unequivocal formulas, several constraints were applied, such as atoms number limitations, i.e. $C \leq 100$, $H \leq 200$, $O \leq 80$, $N \leq 5$ and $S \leq 1$ [36], restrictions on atoms to carbon number ratios, i.e. $0.2 \leq H/C \leq 3.1$, $O/C \leq 2$, $N/C \leq 1.3$ and $S/C \leq 0.8$, RDBE > 0, nitrogen rule (for *m/z* ratio values lower or equal to 500) and isotopic pattern filtering [40]. Moreover, Kendrick mass defect (KMD) was performed to help formulas assignment for both magnitude and absorption mode mass spectra. For the KMD analysis, building blocks with a higher number of occurrences were identified and chosen for the analysis. For this step, experimental mass differences values were examined and only those comprised in the range ± 1 mDa of the building block exact mass were considered [41–44]. To further improve the reliability of results, building blocks with occurrences lower than a threshold value (properly chosen to remove all the noisy data) were excluded, being higher the probability for these to have occurred randomly [45]. The threshold was set to 30 for the mFT, while it was set to 200 for the aFT. HRMS data were processed by using AutoVectis Pro (v.8.9, Spectroswiss, Lausanne, Switzerland) and R software (v3.6.3, www.r-project.org).

3.4. Results and Discussion

Direct-injection High Resolution ESI(-)-FT-ICR MS data were used to obtain a general description of metabolome of *Peperoni di Senise* peppers PGI. Obtained mass spectra showed an enormous number of peaks, thus revealing the high complexity of the sample. However, most of them could be related to noise and artifacts, thus hampering the straight identification of observed metabolites by formula assignment [37,38]. It is, thus, important to face several pre-treatment steps to obtain more reliable MS data, from which an accurate metabolic profile could be deduced. It's in this context that the choice of working with absorption or magnitude mode mass spectra could make the difference. Utilization of aFT could lead to the improvement of peak resolutions and signal-to-noise ratios, lowering the probability of losing information and boosting the number of species identified with a single direct analysis. Moreover, artifact identification is possible with aFT. Despite all these advantages, a series of disadvantages made impractical the employment of aFT for scientific purposes for at least 40years [37]. Indeed, ion phase shifting makes these unreadable, because of the presence of intense negative peaks, which could show a marked asymmetry too. Moreover, phase shifting depends from the cyclotron frequency, so every identified ion needs a different phase correction [37,38]. These features make the obtainment of complex matrices aFT almost impossible without the utilization of a dedicated tool, with which a full control of the FID elaboration is possible. In this sense, the new software AutoVectis Pro [46] was crucial for data pre-treatment, allowing us to work directly on the FID, starting from apodization to aFT phase correction [47–49]. Firstly, a full control of the peak peaking process was possible, allowing tuning of related parameters such as the magnitude of points interval to consider for centroid calculation. Moreover, aFT and mFT could be readily calculated from the apodized FID (**Figure 1A**). AutoVectis Pro software, thus, allowed to make a direct comparison between them. In this way, it was possible to appreciate peak shrinkage which characterizes aFT, leading to an improvement of related resolution, and artefact depletion (**Figure 1B**). Thus, by considering the latter, it's possible to perform a more reliable peak quality evaluation, that is not always possible by looking at the mFT alone. Obtainment of the aFT not only led to a marked reduction of artifacts, like wiggles and harmonics, but to the improvement of peak resolutions and S/N ratios too, thus increasing the number of resolved MS signals and observed ion species. However, before this, a proper phase correction should be carried out to obtain a readable aFT. During the last years, several approaches were optimized to solve the phase correction problem [37]. However, some of them needs the utilization of a dedicated MS apparatus, while others shown to be really time consuming because of the higher number of computational steps. AutoVectis Pro employs a genetic algorithm able to deduce calibration coefficients in few milliseconds [47,49]. In detail, it generates several phase correction functions starting from a predefined one (which order is defined by the user) and applying random mutation on related frequency values. After the best phase correction function from the initial population has been identified, the step is iterated till a full optimization is obtained [49]. Thanks to the aFT, more ionic species could be distinguished and identified and a complete information on metabolic profile of the analyzed sample could be retrieved. This last aspect was evident during the formula assignment step. Indeed, a *m/z* peak list of 4906 members could be extrapolated from the aFT, a marked improvement compared to the 901 peaks obtained from the mFT. In each case, however, the number of MS signals is huge, making necessary to find another way to visualize results to simplify their analysis and

to obtain desired information from mass spectra. In light of this, accurate m/z values were subjected to the KMD analysis [44] to identify present homologous series and to simplify molecular formula assignment. It's worth noting how aFT analysis provided a more complete information of the metabolic profile of our sample by leading to the discovery of new homologous series and related members, as can be noticed by the analysis of related Kendrick plots (**Figure 2**). Moreover, this last feature helped the identification of other high frequency building blocks (**Table 1**), that could be chosen to perform other KMD analyses to extend the range of assigned peaks. For aFT, 1175 unequivocal formulas were obtained, a marked improvement compared to the 351 formulas obtained for the mFT. To best interpret our results, a well-known visualization method was assumed, i.e. Van Krevelen diagrams, in which elemental compositions are plotted depending on their O/C and H/C ratios [32–36]. Thanks to these plots, simply by looking at the positions of the spots, it's possible to assign every observed metabolite to a specific metabolic class [32–36]. From the analysis of aFT and mFT related Van Krevelen plots (**Figure 3**), the presence of the same important classes of metabolites could be deduced, i.e. fatty acids and related derivatives, carotenoids, amino acids and peptides, carbohydrates and polyphenols. As a matter of fact, peppers are a rich source of this kind of metabolites, some of them important for the improvement of human health. In detail, peppers show high levels of carotenoids and polyphenols, well-known compounds for their antioxidant activity toward free radicals and reactive oxygen and their potential anticancer activity [24,50]. Moreover, peppers contain anthocyanins, metabolites belonging to the classes of polyphenols, that give color to the fruit [51]. The presence of a high concentration of these compounds gives peppers beneficial properties and this makes this kind of fruit suitable for medical purposes, such as treatment of asthma, sore throats and toothache [24,50,52]. As can be deduced from the analysis of obtained Van Krevelen plots, sun-dried *Peperoni di Senise PGI* peppers still show a wide diversity of these kind of metabolites, thus supporting the hypothesis of retainment of macroscopic health promoting properties. Moreover, comparison of aFT and mFT related Van Krevelen plots allowed to appreciate the advantages provided by the former, supporting what was already argued from the analysis of related Kendrick plots. Indeed, point density is higher for the former, providing a better idea of the range of different metabolites belonging to a specific class.

3.5. Conclusions

The new commercial AutoVectis Pro software demonstrated to be crucial in maximizing the reliable information provided by a single High Resolution MS spectrum obtained from the analysis of a very complex matrix. The possibility to work on recorded FID directly and to calculate absorption mode FT-ICR MS spectra in few seconds allowed us to obtain a complete overview of the metabolic profile of a methanolic extract of a sample of *Peperoni di Senise PGI* pepper. More specifically, the utilization of the software AutoVectis Pro led to the identification of an enormous number of species, as could be deduced from related Kendrick and Van Krevelen plot, minimizing loss of information and allowing the complete characterization of metabolic profile of the sample. This work underlines how the implementation of absorption mode MS calculation is necessary

to gain more clues on analyzed matrix and how AutoVectis Pro could be the perfect solution to make this step easy, fast and straightforward.

3.6. References

[1] S.K. Basu, A.K. De, Capsicum: The genus Capsicum, Taylor & Francis Group, London, 2003.

[2] F. Di Dato, M. Parisi, T. Cardi, P. Tripodi, Genetic diversity and assessment of markers linked to resistance and pungency genes in Capsicum germplasm, Euphytica. 204 (2015) 103–119. https://doi.org/10.1007/s10681-014-1345-4.

[3] Product and Market Development: World Markets in the Spice Trade 2000−2004, Geneva, 2006.

[4] FAOSTAT, (n.d.). http://www.fao.org/faostat/en/#data/QC (accessed September 16, 2020).

[5] L.R. Howard, R.E.C. Wildman, Handbook of Nutraceuticals and Functional Foods, CRC Press, Boca Raton, FL, 2007.

[6] Y. Wahyuni, A.R. Ballester, E. Sudarmonowati, R.J. Bino, A.G. Bovy, Metabolite biodiversity in pepper (Capsicum) fruits of thirty-two diverse accessions: Variation in health-related compounds and implications for breeding, Phytochemistry. 72 (2011) 1358–1370. https://doi.org/10.1016/j.phytochem.2011.03.016.

[7] L.R. Howard, S.T. Talcott, C.H. Brenes, B. Villalon, Changes in phytochemical and antioxidant activity of selected pepper cultivars (Capsicum species) as influenced by maturity, J. Agric. Food Chem. 48 (2000) 1713–1720. https://doi.org/10.1021/jf990916t.

[8] F. Márkus, H.G. Daood, J. Kapitány, P.A. Biacs, Change in the carotenoid and antioxidant content of spice red pepper (Paprika) as a function of ripening and some technological factors, J. Agric. Food Chem. 47 (1999) 100–107. https://doi.org/10.1021/jf980485z.

[9] E.R. Naves, L. de Ávila Silva, R. Sulpice, W.L. Araújo, A. Nunes-Nesi, L.E.P. Peres, A. Zsögön, Capsaicinoids: Pungency beyond Capsicum, Trends Plant Sci. 24 (2019) 109–120. https://doi.org/10.1016/j.tplants.2018.11.001.

[10] X.J. Luo, J. Peng, Y.J. Li, Recent advances in the study on capsaicinoids and capsinoids, Eur. J. Pharmacol. 650 (2011) 1–7. https://doi.org/10.1016/j.ejphar.2010.09.074.

[11] G.J. Lightbourn, R.J. Griesbach, J.A. Novotny, B.A. Clevidence, D.D. Rao, J.R. Stommel, Effects of anthocyanin and carotenoid combinations on foliage and immature fruit color of Capsicum annuum L., J. Hered. 99 (2008) 105–111. https://doi.org/10.1093/jhered/esm108.

[12] S. Park, W.Y. Jeong, J.H. Lee, Y.H. Kim, S.W. Jeong, G.S. Kim, D.W. Bae, C.S. Lim, J.S. Jin, S.J. Lee, S.C. Shin, Determination of polyphenol levels variation in Capsicum annuum L. cv. Chelsea (yellow bell pepper) infected by anthracnose (Colletotrichum gloeosporioides) using liquid chromatography-tandem mass spectrometry, Food Chem. 130 (2012) 981–985. https://doi.org/10.1016/j.foodchem.2011.08.026.

[13] B. Schulze, D. Spiteller, Capsaicin: Tailored Chemical Defence Against Unwanted “Frugivores,” ChemBioChem. 10 (2009) 428–429. https://doi.org/10.1002/cbic.200800755.

[14] J. Lim, G. Kim, C. Mo, M. Kim, Design and Fabrication of a Real-Time Measurement System for the Capsaicinoid Content of Korean Red Pepper (Capsicum annuum L.) Powder by Visible and Near-Infrared Spectroscopy, Sensors. 15 (2015) 27420–27435. https://doi.org/10.3390/s151127420.

[15] I. Domínguez-Martínez, O.G. Meza-Márquez, G. Osorio-Revilla, J. Proal-Nájera, T. Gallardo-Velázquez, Determination of capsaicin, ascorbic acid, total phenolic compounds and antioxidant activity of Capsicum annuum L. var. serrano by mid infrared spectroscopy (Mid-FTIR) and chemometric analysis, J. Korean Soc. Appl. Biol. Chem. 57 (2014) 133–142. https://doi.org/10.1007/s13765-013-4295-y.

[16] D. Lee, M. Kim, B.H. Kim, S. Ahn, Identification of the Geographical Origin of Asian Red Pepper (<scp> *Capsicum annuum* </scp> L.) Powders Using ^{1}H NMR Spectroscopy, Bull. Korean Chem. Soc. 41 (2020) 317–322. https://doi.org/10.1002/bkcs.11974.

[17] M. Ritota, F. Marini, P. Sequi, M. Valentini, Metabolomic characterization of italian sweet pepper (Capsicum annum L.) by means of HRMAS-NMR spectroscopy and multivariate analysis, J. Agric. Food Chem. 58 (2010) 9675–9684. https://doi.org/10.1021/jf1015957.

[18] G. Bianco, R. Pascale, C.F. Carbone, M.A. Acquavia, T.R.I. Cataldi, P. Schmitt-Kopplin, A. Buchicchio, D. Russo, L. Milella, Determination of soyasaponins in Fagioli di Sarconi beans (Phaseolus vulgaris L.) by LC-ESI-FTICR-MS and evaluation of their hypoglycemic activity., Anal. Bioanal. Chem. 410 (2018) 1561–1569. https://doi.org/10.1007/s00216-017-0806-8.

[19] G. Bianco, A. Buchicchio, T.R.I. Cataldi, Structural characterization of major soyasaponins in traditional cultivars of Fagioli di Sarconi beans investigated by high-resolution tandem mass spectrometry., Anal. Bioanal. Chem. 407 (2015) 6381–9. https://doi.org/10.1007/s00216-015-8810-3.

[20] A. Buchicchio, G. Bianco, A. Sofo, S. Masi, D. Caniani, Biodegradation of carbamazepine and clarithromycin by Trichoderma harzianum and Pleurotus ostreatus investigated by liquid chromatography - high-resolution tandem mass spectrometry (FTICR MS-IRMPD)., Sci. Total Environ. 557–558 (2016) 733–9. https://doi.org/10.1016/j.scitotenv.2016.03.119.

[21] G. Bianco, N. Agerbirk, I. Losito, T.R.I. Cataldi, Acylated glucosinolates with diverse acyl groups investigated by high resolution mass spectrometry and infrared multiphoton dissociation., Phytochemistry. 100 (2014) 92–102. https://doi.org/10.1016/j.phytochem.2014.01.010.

[22] G. Bianco, R. Zianni, G. Anzillotta, A. Palma, V. Vitacco, L. Scrano, T.R.I. Cataldi, Dibenzo-p-dioxins and dibenzofurans in human breast milk collected in the area of Taranto (Southern Italy): First case study, Anal. Bioanal. Chem. 405 (2013) 2405–2410. https://doi.org/10.1007/s00216-013-

6706-7.

[23] G. Bianco, G. Novario, G. Anzilotta, A. Palma, A. Mangone, T.R.I. Cataldi, Polybrominated diphenyl ethers (PBDEs) in Mediterranean mussels (Mytilus galloprovincialis) from selected Apulia coastal sites evaluated by GC-HRMS, J. Mass Spectrom. 45 (2010) 1046–1055. https://doi.org/10.1002/jms.1799.

[24] R. Pascale, M.A. Acquavia, T.R.I. Cataldi, A. Onzo, D. Coviello, S.A. Bufo, L. Scrano, R. Ciriello, A. Guerrieri, G. Bianco, Profiling of quercetin glycosides and acyl glycosides in sun-dried peperoni di Senise peppers (Capsicum annuum L.) by a combination of LC-ESI(-)-MS/MS and polarity prediction in reversed-phase separations, Anal. Bioanal. Chem. 412 (2020) 3005–3015. https://doi.org/10.1007/s00216-020-02547-2.

[25] G. Ventura, C.D. Calvano, I. Losito, G. Bianco, R. Pascale, F. Palmisano, T.R.I. Cataldi, Effect of pH and mobile phase additives on the chromatographic behaviour of an amide-embedded stationary phase: Cyanocobalamin and its diaminemonochloro-platinum(II) conjugate as a case study, J. Sep. Sci. 42 (2019) 1155–1162. https://doi.org/10.1002/jssc.201801060.

[26] F. Lelario, G. Bianco, S.A. Bufo, T.R.I. Cataldi, Establishing the occurrence of major and minor glucosinolates in Brassicaceae by LC-ESI-hybrid linear ion-trap and Fourier-transform ion cyclotron resonance mass spectrometry., Phytochemistry. 73 (2012) 74–83. https://doi.org/10.1016/j.phytochem.2011.09.010.

[27] G. Bianco, F. Lelario, F.G. Battista, S.A. Bufo, T.R.I. Cataldi, Identification of glucosinolates in capers by LC-ESI-hybrid linear ion trap with Fourier transform ion cyclotron resonance mass spectrometry (LC-ESI-LTQ-FTICR MS) and infrared multiphoton dissociation., J. Mass Spectrom. 47 (2012) 1160–9. https://doi.org/10.1002/jms.2996.

[28] M.H. Sarafian, M. Gaudin, M.R. Lewis, F.P. Martin, E. Holmes, J.K. Nicholson, M.E. Dumas, Objective set of criteria for optimization of sample preparation procedures for ultra-high throughput untargeted blood plasma lipid profiling by ultra performance liquid chromatography-mass spectrometry, Anal. Chem. 86 (2014) 5766–5774. https://doi.org/10.1021/ac500317c.

[29] S. Malovaná, F.J. García Montelongo, J.P. Pérez, M.A. Rodríguez-Delgado, Optimisation of sample preparation for the determination of trans-resveratrol and other polyphenolic compounds in wines by high performance liquid chromatography, Anal. Chim. Acta. 428 (2001) 245–253. https://doi.org/10.1016/S0003-2670(00)01231-9.

[30] S. Kim, D.Y. Lee, G. Wohlgemuth, H.S. Park, O. Fiehn, K.H. Kim, Evaluation and optimization of metabolome sample preparation methods for Saccharomyces cerevisiae, Anal. Chem. 85 (2013) 2169–2176. https://doi.org/10.1021/ac302881e.

[31] S. Ferrari, Biological elicitors of plant secondary metabolites: Mode of action and use in the production of nutraceutics, Adv. Exp. Med. Biol. 698 (2010) 152–166. https://doi.org/10.1007/978-1-4419-7347-4_12.

[32] R.D. Gougeon, M. Lucio, L. Boutegrabet, D. Peyron, F. Feuillat, D. Chassagne, H. Alexandre, A. Voilley, P. Cayot, I. Gebefügi, N. Hertkorn, P. Schmitt-Kopplin, Authentication approach of the chemodiversity of grape and wine by FTICR-MS, in: ACS Symp. Ser., American Chemical Society, 2011: pp. 69–88. https://doi.org/10.1021/bk-2011-1081.ch005.

[33] C. Roullier-Gall, M. Witting, R.D. Gougeon, P. Schmitt-Kopplin, High precision mass measurements for wine metabolomics, Front. Chem. 2 (2014) 102. https://doi.org/10.3389/fchem.2014.00102.

[34] C. Roullier-Gall, D. Hemmler, M. Gonsior, Y. Li, M. Nikolantonaki, A. Aron, C. Coelho, R.D. Gougeon, P. Schmitt-Kopplin, Sulfites and the wine metabolome, Food Chem. 237 (2017) 106–113. https://doi.org/10.1016/j.foodchem.2017.05.039.

[35] A. Santarsiero, A. Onzo, R. Pascale, M.A. Acquavia, M. Coviello, P. Convertini, S. Todisco, M. Marsico, C. Pifano, P. Iannece, C. Gaeta, S. D'Angelo, M.C. Padula, G. Bianco, V. Infantino, G. Martelli, Pistacia lentiscus Hydrosol: Untargeted Metabolomic Analysis and Anti-Inflammatory Activity Mediated by NF-κB and the Citrate Pathway, Oxid. Med. Cell. Longev. 2020 (2020) 1–14. https://doi.org/10.1155/2020/4264815.

[36] R. Pascale, G. Bianco, T.R.I. Cataldi, P.S. Kopplin, F. Bosco, L. Vignola, J. Uhl, M. Lucio, L. Milella, Mass spectrometry-based phytochemical screening for hypoglycemic activity of Fagioli di Sarconi beans (Phaseolus vulgaris L.), Food Chem. 242 (2018) 497–504. https://doi.org/10.1016/j.foodchem.2017.09.091.

[37] Y. Qi, P.B. O'Connor, Data processing in Fourier transform ion cyclotron resonance mass spectrometry, Mass Spectrom. Rev. 33 (2014) 333–352. https://doi.org/10.1002/mas.21414.

[38] E.N. Nikolaev, Y.I. Kostyukevich, G.N. Vladimirov, Fourier transform ion cyclotron resonance (FT ICR) mass spectrometry: Theory and simulations, Mass Spectrom. Rev. 35 (2016) 219–258. https://doi.org/10.1002/mas.21422.

[39] A.T. Zielinski, I. Kourtchev, C. Bortolini, S.J. Fuller, C. Giorio, O.A.M. Popoola, S. Bogialli, A. Tapparo, R.L. Jones, M. Kalberer, A new processing scheme for ultra-high resolution direct infusion mass spectrometry data, Atmos. Environ. 178 (2018) 129–139. https://doi.org/10.1016/j.atmosenv.2018.01.034.

[40] T. Kind, O. Fiehn, Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry, BMC Bioinformatics. 8 (2007) 105. https://doi.org/10.1186/1471-2105-8-105.

[41] K. Longnecker, E.B. Kujawinski, Using network analysis to discern compositional patterns in ultrahigh-resolution mass spectrometry data of dissolved organic matter, Rapid Commun. Mass Spectrom. (2016) 2388–2394. https://doi.org/10.1002/rcm.7719.

[42] A.G. Marshall, R.P. Rodgers, Petroleomics: The Next Grand Challenge for Chemical Analysis, Acc. Chem. Res. 37 (2004) 53–59. https://doi.org/10.1021/ar020177t.

[43] F. Moritz, M. Kaling, J.-P. Schnitzler, P. Schmitt-Kopplin, Characterization of poplar metabotypes via mass difference enrichment analysis, Plant. Cell Environ. 40 (2017) 1057–1073. https://doi.org/10.1111/pce.12878.

[44] K. Qian, R.P. Rodgers, C.L. Hendrickson, C.A. Hughey, A.G. Marshall, Kendrick Mass Defect Spectrum: A Compact Visual Analysis for Ultrahigh-Resolution Broadband Mass Spectra, Anal. Chem. 73 (2002) 4676–4681. https://doi.org/10.1021/ac010560w.

[45] E. V. Kunenkov, A.S. Kononikhin, I. V. Perminova, N. Hertkorn, A. Gaspar, P. Schmitt-Kopplin, I.A. Popov, A. V. Garmash, E.N. Nikolaev, Total mass difference statistics algorithm: A new approach to identification of high-mass building blocks in electrospray ionization fourier transform ion cyclotron mass spectrometry data of natural organic matter, Anal. Chem. 81 (2009) 10106–10115. https://doi.org/10.1021/ac901476u.

[46] AutoVectis - Kilgour Laboratory, (n.d.). http://www.kilgourlab.com/autovectis/ (accessed January 30, 2021).

[47] D.P.A. Kilgour, R. Wills, Y. Qi, P.B. O'Connor, Autophaser: An algorithm for automated generation of absorption mode spectra for FT-ICR MS, Anal. Chem. 85 (2013) 3903–3911. https://doi.org/10.1021/ac303289c.

[48] D.P.A. Kilgour, S.L. Van Orden, Absorption mode Fourier transform mass spectrometry with no baseline correction using a novel asymmetric apodization function, Rapid Commun. Mass Spectrom. 29 (2015) 1009–1018. https://doi.org/10.1002/rcm.7190.

[49] D.P.A. Kilgour, M.J. Neal, A.J. Soulby, P.B. O'Connor, Improved optimization of the Fourier transform ion cyclotron resonance mass spectrometry phase correction function using a genetic algorithm, Rapid Commun. Mass Spectrom. 27 (2013) 1977–1982. https://doi.org/10.1002/rcm.6658.

[50] R. Arimboor, R.B. Natarajan, K.R. Menon, L.P. Chandrasekhar, V. Moorkoth, Red pepper (Capsicum annuum) carotenoids as a source of natural food colors: analysis and stability—a review, J. Food Sci. Technol. 52 (2015) 1258–1271. https://doi.org/10.1007/s13197-014-1260-7.

[51] G.J. Lightbourn, R.J. Griesbach, J.A. Novotny, B.A. Clevidence, D.D. Rao, J.R. Stommel, Effects of Anthocyanin and Carotenoid Combinations on Foliage and Immature Fruit Color of Capsicum annuum L., J. Hered. 99 (2008) 105–111. https://doi.org/10.1093/jhered/esm108.

[52] M. Imran, M.S. Butt, H.A.R. Suleria, Capsicum annuum Bioactive Compounds: Health Promotion Perspectives, in: Springer, Cham, 2018: pp. 1–22. https://doi.org/10.1007/978-3-319-54528-8_47-1.

3.7. Figures

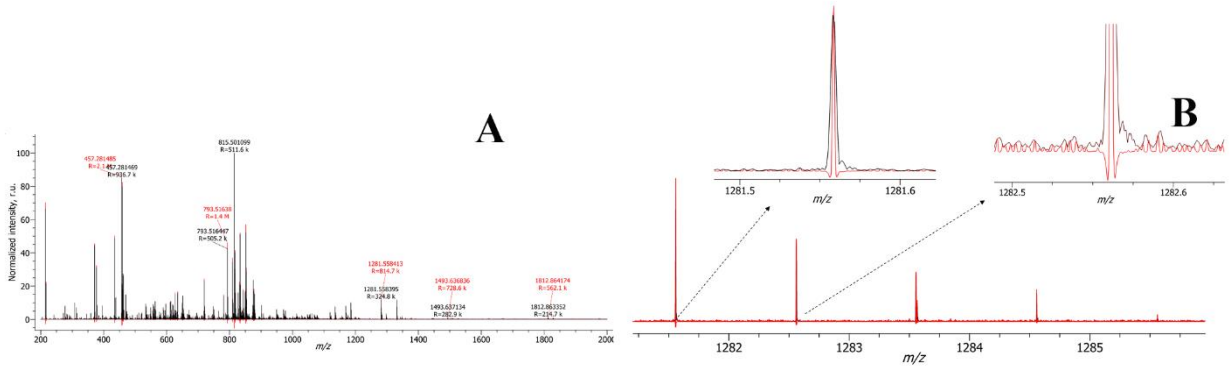

Figure 1 *In plot A, the ESI(-)-FT-ICR magnitude (black line) and absorption mode (red line) mass spectra of a sample of sun-dried Peperoni di Senise (PGI) peppers. In plot B, a detail of the spectra, in which an improvement of peak resolution and a reduction of related wiggles (artefacts, see text) can be appreciated.*

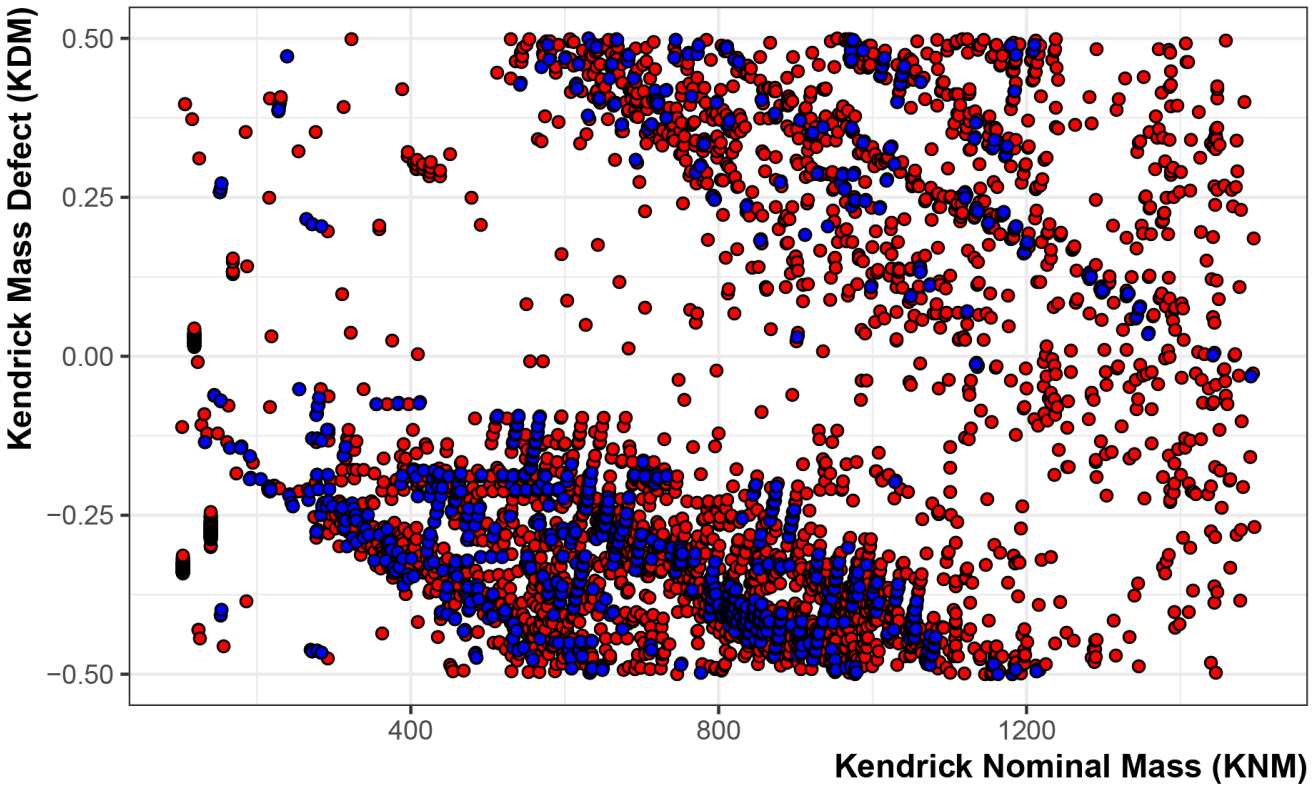

Figure 2 *Kendrick plot obtained from absorption (red dots) and magnitude (blue dots) mode MS, by choosing CH_2 as building block.*

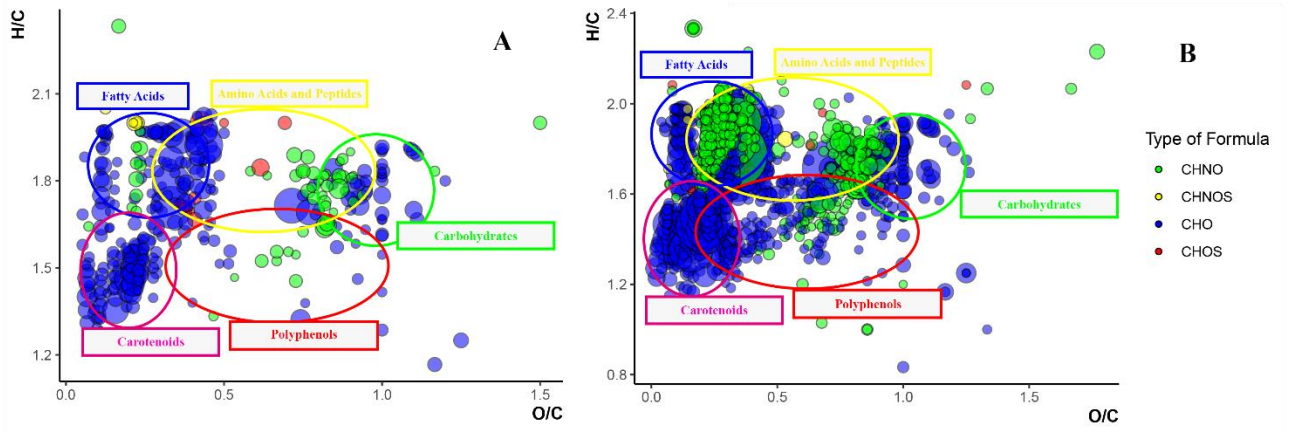

Figure 3 *Van Krevelen plots of an extract of sun-dried Peperoni di Senise PGI peppers (Capsicum Annuum L.), obtained from the elaboration of the magnitude (plot A) and absorption (plot B) mode MS, respectively. Types of formula are distinguished by colors, i.e. blue for CHO, green for CHON, red for CHONS and yellow for CHOS type.*

3.8. Tables

Table 1 *Building block occurrences in both absorption and magnitude mode mass spectra. Mass differences which value was in the range ±1 mDa of the building block exact mass were considered.*

Building Block	Exact mass	Reaction	Magnitude Mode MS	Absorption Mode MS
CH_2	14.016	Methylation	54	441
H_2	2.016	Hydrogenation	104	1033
C_1	12.000	C-insertion	31	340
O_1	15.995	O-insertion	42	455
CO_2	43.990	Carboxylation	37	253
C_2H_2	26.016	Decarboxylative Condensation	73	537
H_2O	18.011	Hydrolysis/Condensation	43	355
C_2H_4	28.032	Alkylation	48	645
$C_6H_{10}O_5$	162.053	Glucose addiction	62	291
CH_2O	30.011	Hydroxymethyl transfer	39	380
C_2H_2O	42.011	Hydroxypyruvic acid addiction	0	310
CO	27.995	Formyl transfer	0	288
C_2H_4O	44.026	Pyruvic Acid addiction	0	246
$C_2H_4O_2$	60.021	Hydroxypyruvic acid addiction/ hydrogenation	0	206
$C_3H_2O_2$	70.006	2-Ketosuccinate addiction	0	206
$C_3H_6O_2$	74.037	3-Hydroxy-2-oxobutanoic acid addiction	0	275
$C_5H_8O_4$	132.043	Pentose addiction (condensation)	0	212

4. Contribute 2

Untargeted metabolomic analysis by High Resolution Mass Spectrometry for the characterization of new Italian wine varieties

A. Onzo[1], M. Acquavia[1], R. Pascale[1], P. Iannece[2], C. Gaeta[2], Y. Tsybin[3], G. Bianco[1]*

[1]*Università degli Studi della Basilicata, Dipartimento di Scienze, Via dell'Ateneo Lucano 10, Potenza, Italy;*

[2]*Università degli Studi di Salerno, Dipartimento di Chimica e Biologia, Via Giovanni Paolo II 132, Fisciano, Italy*

[3]*Spectroswiss Sarl, EPFL Innovation Park, Building I, 1015 Lausanne, Switzerland.*

4.1. Abstract

The chemical composition of wine samples comprises numerous bioactive compounds that are responsible for unique flavor and health-promoting properties. Moreover, most of them might account for all wine specific features such as cultivar, vintage, origin, and quality. Thus, it's important to have a complete overview of metabolic profile of new wine products in order to obtain peculiar information in terms of their bioactivity, quality and traceability. To achieve this aim, in this work a Mass Spectrometry-based phytochemical screening was performed on seven new wine products from Villa D'Agri in the Basilicata region (Italy), i.e. *Aglianico bianco*, *Plavina*, *Guisana*, *Giosana*, *Malvasia ad acino piccolo*, *Colata Murro* and *Santa Sofia*. High Resolution Mass Spectrometry data obtained from sample analysis were employed to perform a rapid analysis of metabolome by converting accurate mass-to-charge ratio (*m/z*) values in putative elemental formulas in order to better understand the chemical composition of the samples. Molecular formula maps were obtained by making 2D Van Krevelen plots, that led to a direct identification of different classes of metabolites. The presence of important metabolite classes, i.e. fatty acid derivatives, tannins, amino acids and peptides, carbohydrates and phenolic derivatives, was assessed. Moreover, the comparison of obtained metabolomic maps revealed some differences among profiles, thus suggesting their employment as *metabolic fingerprints*. This study shed some light on metabolic composition of seven new Italian wine varieties, improving their value in terms of related bioactive compound content. Moreover, unique metabolomic fingerprints were obtained for each of them, which can be used as innovative tools for their authentication.

4.2. Introduction

Wine has been part of human culture for 6000 years, being employed for dietary and socioreligious purposes [1,2]. Its first production goes back to antiquity, as does the discovery of its healthful benefits, now largely attributed to the antimicrobial activity of ethanol [2]. Throughout ancient times, the conversion of grapes into wine was considered a gift from the gods and the best wines were thus reserved for the elite of society [1]. Today, wine is an integral component of the culture of many countries, a form of entertainment and a beverage of choice for supporters of its health benefits. Unlike many foodstuffs, wine's attractions rely not on bold consistent flavours, but upon a wide array of sensations that make its charm difficult to define [3]. Indeed, wine producers are considered sellers of sensory experiences. Regardless of the region in which the wine is produced or the economic status of the consumer, all wines are expected to be pleasant experiences [3]. In most of the world's wine regions, at least until around the middle of the 1980s, wine has been obtained from grapes, following a complex process known as "*winemaking*", that comprises the fermentation of grape must with the indigenous yeasts present on the grapes when harvested, or introduced from the equipment and cellar during the vinification process [4–7]. Nowadays, the practice of adding selected yeasts to slightly sulphited musts has become widespread to ensure that must fermentation is rapid and complete and can produce wines of reproducible character and quality [5,6,8]. The biological process of winemaking is the result of a series of biochemical transformations carried on by the action of several enzymes from different microorganisms, especially yeasts, which are responsible for the principal part of the process, alcoholic fermentation [4,6]. Result of the winemaking process is, thus, the production of a series of biocompounds that are considered to be responsible for the most appreciated properties of this beloved product, such as its taste and related biological activity [7,9]. It's for this reason that a lot of effort has been spent through the last years to improve the knowledge of metabolic profile of different wine varieties. Knowledge of the chemical composition of grape and wine provided a wide series of advantages, such as the possibility to shed some light on how winemaker process could influence the metabolic profile of the final product, thus allowing its optimization for the maximization of certain wine properties [8–12]. The number of compounds identified in wine increased dramatically since the development of gas chromatography (GC), high pressure liquid chromatography (HPLC), thin-layer chromatography (TLC), infrared spectroscopy (IRS), and nuclear magnetic resonance (NMR) [13–21]. In detail, coupling of Mass Spectrometry (MS) technique to GC and to HPLC has been especially valuable in identifying unknown compounds [22–25]. More than 500 compounds have been recognized in wine thus far, of which ~160 are esters. Concentrations range between 10^{-1} and 10^{-6} mg/L generally. At these levels, the individual compounds play very little or no role in the human organoleptic (taste) perception, but collectively they may be very significant [26–28]. The number of aromatic and sapid substances derived from grapes are relatively few compared to that of by-products of yeast activity during fermentation. Wines generally contain 0.8 – 1.2 g/L of aromatic compounds, of which the most common are alcohols, volatile acids, and fatty acid esters [29,30]. Alcohols often constitute 50% of all volatile substances in wines [24,25]. Carbonyls, phenols, lactones, terpenes, acetals, hydrocarbons, sulphur, and nitrogen compounds are present in much lower concentrations, but they are more important qualitatively and contribute specific sensory

characteristics relevant to the fragrance of a wine [31–33]. The taste and mouth-feel sensations are due primarily to the few compounds that occur individually at concentrations > 100 mg/L, like water, ethanol, organic acids, sugars, and glycerol [34–36]. Tannins occur in red wine and rarely in significant amounts in white wines [37,38]. The principal grape sugars are glucose and fructose, and they occur in roughly equal proportions at maturity, whereas overripe grapes often have a higher proportion of fructose [39,40]. Polysaccharides are generally low in content. They are partially water soluble and are extracted into the juice during crushing and pressing [41]. The most important and abundant alcohol in wine is ethanol [42,43]. Under standard fermentation conditions, ethanol can accumulate to ~14–15%, but generally ethanol concentrations in wine range between 10–13%. Different factors control ethanol production, like sugars, temperature, and yeast strain. Ethanol is crucial to the stability, aging, and sensory properties of wine [42,43]. Other potentially significant higher alcohols in wine are the straight-chain alcohols: 1-propanol, 2-methyl-1-propanol, 2-methyl-1-butanol, and 3-methyl-1-butanol [44]. The formation of higher alcohols occurs as a by-product of yeast fermentation and is markedly influenced by vinification parameters, such as temperature, presence of oxygen, suspended solids, and yeast strain [5,7,8]. Carboxylic acids like tartaric, malic, lactic, succinic, oxalic, fumaric, and citric acids control the pH of wine [39,45]. Phenols are important for the characteristics and quality of red wines. Their concentration in white wine is much lower. Phenols and related compounds can affect the appearance, taste, mouth-feel, fragrance, and antimicrobial properties of wine [44]. The two primary phenol groups that occur in grapes and wine are the flavonoids and the nonflavonoids. The most common flavonoids in wine are flavonols, catechins (flavan-3-ols), and, in red wines, anthocyanins [46–48]. Flavonoids come primarily from the skins, seeds, and stems of the fruit [49]. In red wines, they commonly constitute > 85% of the phenol content, while, in white wines, flavonoids typically comprise < 20% of the total phenolic content. The amount of flavonoids extracted during vinification is influenced by many factors, including temperature, length of skin contact, mixing, type of fermentation vessel, ethanol concentrations, SO_2, yeast strain and pH [50–54]. Hydroxycinnamic acid derivatives commonly occur esterified to sugars, organic acids, or various alcohols. The concentration of phenolics in wine increases during skin fermentation and subsequently begins to fall as phenols aggregate and precipitate with proteins and yeast cell remnants [32,33]. During fining and ripening, phenols continues to decrease, and aging has a further dramatic effect on their reduction. Seen the huge diversity of organic compounds present in wine varieties and their synergic activity resulting in pleased macroscopic properties, together with the deep correlation between compound levels and winemaking process, it's of great interest to have an idea of the metabolomic profile of new wine varieties, in order to use obtained information for related winemaking process optimization, quality control, authentication and traceability [55–58]. This could be accomplished using hyphenated techniques by following an untargeted analysis approach [58]. However, high costs and long analysis times hamper their routine employment for this task. Nevertheless, High Resolution Mass Spectrometry (HRMS) technique showed to be the technique of choice for untargeted metabolomic analysis, even if used by following a shotgun, or direct-injection, approach [59–62]. Indeed, HRMS has been already used for wine characterization, allowing the identification of thousands of metabolites with a single direct analysis, with labour times of few minutes [28,59,60]. Moreover, related results provided

useful biochemical mechanistic information, highlighting metabolic differences among samples subjected to different winemaking conditions [28,59,60]. In this study, Fourier transform ion cyclotron resonance mass spectrometry (FT-ICR MS) was employed to profile the molecular profile of of seven new Italian wine varieties, i.e. *Aglianico bianco*, *Plavina*, *Guisana*, *Giosana*, *Malvasia ad acino piccolo*, *Colata Murro* and *Santa Sofia*, produced in Villa D'Agri in the Basilicata region (Italy). Maps of main metabolites were proposed and discussed.

4.3. Materials and Methods

Wine Samples

Wine samples were obtained from new germoplasms cultivated in the Pollino region, a natural area located in Basilicata (Italy, and were provided by the the Agency for Development and Innovation in Agriculture (ALSIA, Agenzia Lucana di Sviluppo e di Innovazione in Agricoltura). Key features of these new wine varieties, together with a detailed organoleptic description, are describer elsewhere [63].

Chemicals

Sodium trifluoroacetate (NaTFA, 98%) and Methanol (MeOH, LC-MS grade) were purchased from Sigma-Aldrich (Milano, Italy). Pure nitrogen (99.996%) was delivered to the MS system as the sheath gas. Wine samples were provided by ALSIA (Agenzia Lucana di Sviluppo e di Innovazione in Agricoltura).

Sample Preparation

Wine samples were prepared by following a previous method [60]. In detail, 20 µL of sample were diluted by adding 1 mL of MeOH. The solution was vortexed, passed through a PTFE 0.22 µm syringe filter and directly injected into the HRMS instrument. For every sample, 3 replicates have been prepared, together with a blank sample obtained by subjecting 20 µL of MeOH to the whole sample preparation step.

Mass spectrometry analysis

ESI (±) Fourier Transform Ion Cyclotron Resonance Mass Spectrometry (ESI-FT-ICR MS) technique was used for the untargeted analysis of the sample. High-resolution Mass Spectra were acquired on a Bruker (Bruker Daltonik GmbH, Bremen, Germany) solariX XR Fourier transform ion cyclotron resonance mass spectrometer (FT-ICR-MS) equipped with a 7T superconducting magnet and an ESI source. The capillary voltage was set to 3.9 and -4.5 kV for negative and positive ionization modes, respectively, with a nebulizer gas pressure of 1.2 bar and dry gas flow rate of 4 L/min at 200 °C. Spectra were acquired with a Time Domain size of 16 mega-word, an accumulation time of 0.1 s and a mass range of 100-2000 m/z. Moreover, the average number of scans was set to 50. Before the analysis, the mass spectrometer was externally calibrated with NaTFA. High accuracies were reached, with a root mean square (RMS) error lower than 0.1 ppm. FT-ICR mass spectra were subjected to several data pre-treatment steps. In detail, recorded free induction decays (FIDs) were subjected to apodization and related absorption mode mass spectra was obtained. Phase correction, mass recalibration and baseline correction have been performed, together with blank subtraction [64–66] and noise

filtering by following the N-Sigma methodology approach [67]. More specifically, noise level has been estimated and peaks showing a signal-to-noise ratio (S/N) higher than 2 were retained. Thus, obtained FT-ICR mass spectra were exported to peak lists. From these, possible elemental formulas were calculated for each MS signal. To obtain unequivocal formulas, several constraints were applied, such as atoms number limitations, i.e. $C \leq 100$, $H \leq 200$, $O \leq 80$, $N \leq 5$ and $S \leq 1$ [59–62], restrictions on atoms to carbon number ratios, i.e. $0.2 \leq H/C \leq 3.1$, $O/C \leq 2$, $N/C \leq 1.3$ and $S/C \leq 0.8$, RDBE > 0, nitrogen rule (for *m/z* ratio values lower or equal to 500) and isotopic pattern filtering. Moreover, Kendrick mass defect (KMD) was performed to help formulas assignment [59,60]. In detail, building blocks with a higher number of occurrences were identified and chosen for the analysis. For this step, experimental mass differences values were examined and only those comprised in the range ± 1 mDa of the building block exact mass were considered [68,69]. To further improve the reliability of results, building blocks with occurrences lower than a threshold value (properly chosen to remove all the noisy data) were excluded, being higher the probability for these to have occurred randomly [70]. HRMS data were processed by using AutoVectis Pro (v.8.9, Spectroswiss, Lausanne, Switzerland) and R software (v3.6.3, www.r-project.org).

4.4. Results and Discussion

Direct-injection High Resolution ESI(±)-FT-ICR MS data were used to obtain a general description of metabolome of new Italian wine varieties samples. In detail, five types of white wines, i.e. *Aglianico bianco*, *Guisana*, *Giosana*, *Malvasia ad acino piccolo* and *Santa Sofia*, and two types of red ones, i.e. *Colata Murro* and *Plavina*, were analyzed. Obtained mass spectra showed a huge number of peaks (**Figure 1**), thus revealing the wide diversity of metabolites present in our samples. However, it should be pointed out that artifacts occurrence cannot be considered negligible in FT-ICR [71,72]. Moreover, noisy data hamper the identification of low intensity ionic species, thus making overall data elaboration not easy. Working with absorption mode mass spectra turned out to be the best solution to overcome these issues. Indeed, peak resolution and signal-to-noise ratios (S/N) were markedly improved, leading to the identification of a higher number of ionic species. Of course, utilization of a dedicated tool to accomplish this task, i.e. the AutoVectis Pro software, was crucial to perform a quick phase correction step efficiently and, thus, to obtain a readable absorption mode mass spectrum, task that couldn't be achieved for almost 40 years [64,65,71]. Despite the advantages provided by the absorption mode mass spectrum, it's still difficult to deduce something by simply looking at MS spectra. To best interpret our results, MS signals were assigned to unique elemental formulas (see *Materials and Methods*) and results were employed with a well-known visualization tool, i.e. the Van Krevelen plot, made by plotting elemental formula on a 2D diagram, setting the H/C and the O/C ratios as the y- and the x-axis, respectively [59–61]. Identified ionic species are, thus, spread all over the plot and their position is crucial to classify them in one of the major metabolite classes. Thanks to the analysis of Van Krevelen plots, the presence of specific types of metabolites in wine samples was proposed, i.e. carbohydrates, polyphenols, amino acids and peptides and unsaturated fatty acids (**Figure 1**). In detail, from the analysis of Van Krevelen plots, differences among metabolic profiles can be noticed, some of them reflecting what was already found in the literature. More specifically, by looking at the negative ionization mode results, it can be seen how every type

of wine showed a higher density of points on the upper right part of the plot, indicating a wider diversity of carbohydrates and glycoconjugates. Furthermore, red wines seem to contain a higher amount of unsaturated glycoconjugate compounds, supporting what was already found regarding the occurrence of phenolic compounds in wine as glycoconjugates mainly [7,44,47,48,73]. A higher density of points in the middle part of Van Krevelen plot of red wines, moreover, indicates the presence of more phenolic derivatives, most probably flavonoids [59,60]. However, white wines show a little cluster of points in the middle left part of the plot (**Figure 1B**), absent for red wine varieties, which could be related to the presence of low oxygen content phenolic acids, such as hydroxycinnamic acid derivatives, which are known to be responsible for the typical yellowish colour of white wines [74,75]. Overall, among the same types of wine, difference could be spot related to the absence and the presence of points in specific Van Krevelen plots, suggesting the fact that some of the identified derivatives are present only in some wine samples, just like the unsaturated CHNO formula type compounds present for the *Giosana* and *Santa Sofia* white wines (see *Supplementary Material*) or the aliphatic amides present in the *Colata Murro* red wine sample only, suggesting the utilization of Van Krevelen plots to rapidly detect potential markers for quality control purposes. For what concerns positive ionization mode results, the analysis of related Van Krevelen plots shows how profiles look similar among the different types of wine (**Figure 2**). In this case, amino acids and peptides, together with their aliphatic derivatives, and aliphatic amides seems to be predominant, together with other CHO formula type compounds most probably related to unsaturated fatty acid derivatives. Interestingly, no remarkable differences among red and white wine metabolic profiles could be uncovered by looking at positive ionization mode Van Krevelen plots. Moreover, the *Malvasia ad acino piccolo* white wine Van Krevelen plot shows the highest density of points, indicating how wider is the range of the diversity of identified classes of metabolites.

4.5. Conclusions

High Resolution Mass Spectrometry technique was successfully used in this work for the characterization of metabolome of new Italian type wine varieties, thus confirming its suitability for quick and efficient untargeted metabolic analysis of natural samples. Useful information about types of metabolite present in wine samples were obtained, since a classification of identified species has been made possible by the utilization of the Van Krevelen plot, a well-known visualization tool useful for HRMS data interpretation. Results helped to identify principal classes of metabolites and to spot principal differences among related metabolic profiles and, thus, are very promising for the employment of Van Krevelen plots as metabolomic *fingerprints* useful for potential marker identification, for quality control, authentication and traceability.

4.6. References

[1] G.J. Soleas, E.P. Diamandis, D.M. Goldberg, Wine as a biological fluid: History, production, and role in disease prevention, J. Clin. Lab. Anal. 11 (1997) 287–313. https://doi.org/10.1002/(SICI)1098-2825(1997)11:5<287::AID-JCLA6>3.0.CO;2-4.

[2] J.B. German, R.L. Walzem, The Health Benefits of Wine, Annu. Rev. Nutr. 20 (2000) 561–593. https://doi.org/10.1146/annurev.nutr.20.1.561.

[3] Zoecklein, ed., Wine Analysis and Production, Springer US, 1995. https://doi.org/10.1007/978-1-4757-6978-4.

[4] R.B. Boulton, V.L. Singleton, L.F. Bisson, R.E. Kunkee, Principles and Practices of Winemaking, Springer US, 1999. https://doi.org/10.1007/978-1-4757-6255-6.

[5] P. Romano, M. Ciani, G.H. Fleet, Yeasts in the Production of Wine, Springer-Verlag, New York, 2019. https://doi.org/10.1007/978-1-4939-9782-4.

[6] P.J. Chambers, I.S. Pretorius, Fermenting knowledge: the history of winemaking, science and yeast research, EMBO Rep. 11 (2010) 914–920. https://doi.org/10.1038/embor.2010.179.

[7] M.V. Moreno-Arribas, M.C. Polo, Winemaking Biochemistry and Microbiology: Current Knowledge and Future Trends, Crit. Rev. Food Sci. Nutr. 45 (2005) 265–286. https://doi.org/10.1080/10408690490478118.

[8] J.A. Suárez-Lepe, A. Morata, New trends in yeast selection for winemaking, Trends Food Sci. Technol. 23 (2012) 39–50. https://doi.org/10.1016/j.tifs.2011.08.005.

[9] E.C. Kritzinger, F.F. Bauer, W.J. Du Toit, Role of glutathione in winemaking: A review, J. Agric. Food Chem. 61 (2013) 269–277. https://doi.org/10.1021/jf303665z.

[10] K.L. Sacchi, L.F. Bisson, D.O. Adams, A Review of the Effect of Winemaking Techniques on Phenolic Extraction in Red Wines, Am. J. Enol. Vitic. 56 (2005) 197–206.

[11] A. Di Lorenzo, N. Bloise, S. Meneghini, A. Sureda, G. Tenore, L. Visai, C. Arciola, M. Daglia, Effect of Winemaking on the Composition of Red Wine as a Source of Polyphenols for Anti-Infective Biomaterials, Materials (Basel). 9 (2016) 316. https://doi.org/10.3390/ma9050316.

[12] B. Ayestarán, L. Martínez-Lapuente, Z. Guadalupe, C. Canals, E. Adell, M. Vilanova, Effect of the winemaking process on the volatile composition and aromatic profile of Tempranillo Blanco wines, Food Chem. 276 (2019) 187–194. https://doi.org/10.1016/j.foodchem.2018.10.013.

[13] S. Agatonovic-Kustrin, C.G. Hettiarachchi, D.W. Morton, S. Razic, Analysis of phenolics in wine by high performance thin-layer chromatography with gradient elution and high resolution plate imaging,

J. Pharm. Biomed. Anal. 102 (2015) 93–99. https://doi.org/10.1016/j.jpba.2014.08.031.

[14] A. Romano, H. Klebanowski, S. La Guerche, L. Beneduce, G. Spano, M.L. Murat, P. Lucas, Determination of biogenic amines in wine by thin-layer chromatography/ densitometry, Food Chem. 135 (2012) 1392–1396. https://doi.org/10.1016/j.foodchem.2012.06.022.

[15] E. Revilla, J.M. Ryan, Analysis of several phenolic compounds with potential antioxidant properties in grape extracts and wines by high-performance liquid chromatography-photodiode array detection without sample preparation, J. Chromatogr. A. 881 (2000) 461–469. https://doi.org/10.1016/S0021-9673(00)00269-7.

[16] A. Visconti, M. Pascale, G. Centonze, Determination of ochratoxin A in wine by means of immunoaffinity column clean-up and high-performance liquid chromatography, J. Chromatogr. A. 864 (1999) 89–101. https://doi.org/10.1016/S0021-9673(99)00996-6.

[17] L. Culleré, A. Escudero, J. Cacho, V. Ferreira, Gas Chromatography-Olfactometry and Chemical Quantitative Study of the Aroma of Six Premium Quality Spanish Aged Red Wines, J. Agric. Food Chem. 52 (2004) 1653–1660. https://doi.org/10.1021/jf0350820.

[18] R. López, M. Aznar, J. Cacho, V. Ferreira, Determination of minor and trace volatile compounds in wine by solid-phase extraction and gas chromatography with mass spectrometric detection, J. Chromatogr. A. 966 (2002) 167–177. https://doi.org/10.1016/S0021-9673(02)00696-9.

[19] R. Godelmann, F. Fang, E. Humpfer, B. Schütz, M. Bansbach, H. Schäfer, M. Spraul, Targeted and nontargeted wine analysis by 1H NMR spectroscopy combined with multivariate statistical analysis. differentiation of important parameters: Grape variety, geographical origin, year of vintage, J. Agric. Food Chem. 61 (2013) 5610–5619. https://doi.org/10.1021/jf400800d.

[20] H.S. Son, G.S. Hwang, K.M. Kim, H.J. Ahn, W.M. Park, F. Van Den Berg, Y.S. Hong, C.H. Lee, Metabolomic studies on geographical grapes and their wines using 1H NMR analysis coupled with multivariate statistics, J. Agric. Food Chem. 57 (2009) 1481–1490. https://doi.org/10.1021/jf803388w.

[21] D. Cozzolino, M.J. Kwiatkowski, M. Parker, W.U. Cynkar, R.G. Dambergs, M. Gishen, M.J. Herderich, Prediction of phenolic compounds in red wine fermentations by visible and near infrared spectroscopy, in: Anal. Chim. Acta, Elsevier, 2004: pp. 73–80. https://doi.org/10.1016/j.aca.2003.08.066.

[22] Y. Wang, F. Catana, Y. Yang, R. Roderick, R.B. Van Breemen, An LC-MS method for analyzing total resveratrol in grape juice, cranberry juice, and in wine, J. Agric. Food Chem. 50 (2002) 431–435. https://doi.org/10.1021/jf010812u.

[23] L. Jaitz, K. Siegl, R. Eder, G. Rak, L. Abranko, G. Koellensperger, S. Hann, LC-MS/MS analysis of phenols for classification of red wine according to geographic origin, grape variety and vintage, Food

Chem. 122 (2010) 366–372. https://doi.org/10.1016/j.foodchem.2010.02.053.

[24] M. Aznar, R. López, J.F. Cacho, V. Ferreira, Identification and quantification of impact odorants of aged red wines from Rioja, GC-olfactometry, quantitative GC-MS, and odor evaluation of HPLC fractions, J. Agric. Food Chem. 49 (2001) 2924–2929. https://doi.org/10.1021/jf001372u.

[25] J. Bosch-Fusté, M. Riu-Aumatell, J.M. Guadayol, J. Caixach, E. López-Tamames, S. Buxaderas, Volatile profiles of sparkling wines obtained by three extraction methods and gas chromatography-mass spectrometry (GC-MS) analysis, Food Chem. 105 (2007) 428–435. https://doi.org/10.1016/j.foodchem.2006.12.053.

[26] N.A. Bokulich, T.S. Collins, C. Masarweh, G. Allen, H. Heymann, S.E. Ebeler, D.A. Millsa, Associations among wine grape microbiome, metabolome, and fermentation behavior suggest microbial contribution to regional wine characteristics, MBio. 7 (2016). https://doi.org/10.1128/mBio.00631-16.

[27] A. Cuadros-Inostroza, P. Giavalisco, J. Hummel, A. Eckardt, L. Willmitzer, H. Peña-Cortés, Discrimination of wine attributes by metabolome analysis, Anal. Chem. 82 (2010) 3573–3580. https://doi.org/10.1021/ac902678t.

[28] C. Roullier-Gall, D. Hemmler, M. Gonsior, Y. Li, M. Nikolantonaki, A. Aron, C. Coelho, R.D. Gougeon, P. Schmitt-Kopplin, Sulfites and the wine metabolome, Food Chem. 237 (2017) 106–113. https://doi.org/10.1016/j.foodchem.2017.05.039.

[29] R.A. Peinado, J. Moreno, J.E. Bueno, J.A. Moreno, J.C. Mauricio, Comparative study of aromatic compounds in two young white wines subjected to pre-fermentative cryomaceration, Food Chem. 84 (2004) 585–590. https://doi.org/10.1016/S0308-8146(03)00282-6.

[30] M. Vilanova, C. Martínez, First study of determination of aromatic compounds of red wine from Vitis vinifera cv. Castañal grown in Galicia (NW Spain), Eur. Food Res. Technol. 224 (2007) 431–436. https://doi.org/10.1007/s00217-006-0322-0.

[31] R. Longo, J.W. Blackman, P.J. Torley, S.Y. Rogiers, L.M. Schmidtke, Changes in volatile composition and sensory attributes of wines during alcohol content reduction, J. Sci. Food Agric. 97 (2017) 8–16. https://doi.org/10.1002/jsfa.7757.

[32] R. Harrison, Practical interventions that influence the sensory attributes of red wines related to the phenolic composition of grapes: a review, Int. J. Food Sci. Technol. 53 (2018) 3–18. https://doi.org/10.1111/ijfs.13480.

[33] C. Coetzee, W.J. Du Toit, Sauvignon blanc wine: Contribution of Ageing and Oxygen on Aromatic and Non-aromatic Compounds and Sensory Composition: A Review, South African J. Enol. Vitic. 36 (2015) 347–365. https://doi.org/10.21548/36-3-968.

[34] S. Vidal, L. Francis, S. Guyot, N. Marnet, M. Kwiatkowski, R. Gawel, V. Cheynier, E.J. Waters, The mouth-feel properties of grape and apple proanthocyanidins in a wine-like medium, J. Sci. Food Agric. 83 (2003) 564–573. https://doi.org/10.1002/jsfa.1394.

[35] S. Vidal, L. Francis, P. Williams, M. Kwiatkowski, R. Gawel, V. Cheynier, E. Waters, The mouth-feel properties of polysaccharides and anthocyanins in a wine like medium, Food Chem. 85 (2004) 519–525. https://doi.org/10.1016/S0308-8146(03)00084-0.

[36] S. Vidal, P. Courcoux, L. Francis, M. Kwiatkowski, R. Gawel, P. Williams, E. Waters, V. Cheynier, Use of an experimental design approach for evaluation of key wine components on mouth-feel perception, Food Qual. Prefer. 15 (2004) 209–217. https://doi.org/10.1016/S0950-3293(03)00059-4.

[37] A.P. Nel, Tannins and anthocyanins: From their origin to wine analysis - A review, South African J. Enol. Vitic. 39 (2018) 1–20. https://doi.org/10.21548/39-1-1503.

[38] M.J. Herderich, P.A. Smith, Analysis of grape and wine tannins: Methods, applications and challenges, Aust. J. Grape Wine Res. 11 (2005) 205–214. https://doi.org/10.1111/j.1755-0238.2005.tb00288.x.

[39] M. Calull, R.M. Marcé, F. Borrull, Determination of carboxylic acids, sugars, glycerol and ethanol in wine and grape must by ion-exchange high-performance liquid chromatography with refractive index detection, J. Chromatogr. A. 590 (1992) 215–222. https://doi.org/10.1016/0021-9673(92)85384-6.

[40] E.F. Lopez, E.F. Gomez, Simultaneous Determination of the Major Organic Acids, Sugars, Glycerol, and Ethanol by HPLC in Grape Musts and White Wines, J. Chromatogr. Sci. 34 (1996) 254–257. https://doi.org/10.1093/chromsci/34.5.254.

[41] S. Escot, M. Feuillat, L. Dulau, C. Charpentier, Release of polysaccharides by yeasts and the influence of released polysaccharides on colour stability and wine astringency, Aust. J. Grape Wine Res. 7 (2001) 153–159. https://doi.org/10.1111/j.1755-0238.2001.tb00204.x.

[42] A. Rakotovao, C. Berthonneche, A. Guiraud, M. de Lorgeril, P. Salen, J. de Leiris, F. Boucher, Ethanol, Wine, and Experimental Cardioprotection in Ischemia/Reperfusion: Role of the Prooxidant/Antioxidant Balance, Antioxid. Redox Signal. 6 (2004) 431–438. https://doi.org/10.1089/152308604322899503.

[43] M. Iriti, E.M. Varoni, Moderate Red Wine Consumption in Cardiovascular Disease: Ethanol Versus Polyphenols, in: Mediterr. Diet An Evidence-Based Approach, Elsevier Inc., 2015: pp. 143–151. https://doi.org/10.1016/B978-0-12-407849-9.00014-2.

[44] K.H. Čiča, M. Pezer, J. Mrvčić, D. Stanzer, J. Čačić, V. Jurak, M. Krajnović, J.G. Kljusurić, Identification of phenolic and alcoholic compounds in wine spirits and their classification by use of multivariate analysis, J. Serbian Chem. Soc. 84 (2019) 663–677. https://doi.org/10.2298/JSC190115020H.

[45] D. Tusseau, C. Benoit, Routine high-performance liquid chromatographic determination of carboxylic acids in wines and champagne, J. Chromatogr. A. 395 (1987) 323–333. https://doi.org/10.1016/S0021-9673(01)94121-4.

[46] R. Boulton, The Copigmentation of Anthocyanins and Its Role in the Color of Red Wine: A Critical Review, Am. J. Enol. Vitic. 52 (2001).

[47] N. Castillo-Muñoz, S. Gómez-Alonso, E. García-Romero, I. Hermosín-Gutiérrez, Flavonol profiles of Vitis vinifera red grapes and their single-cultivar wines, J. Agric. Food Chem. 55 (2007) 992–1002. https://doi.org/10.1021/jf062800k.

[48] N. Castillo-Muñoz, S. Gómez-Alonso, E. García-Romero, M.V. Gómez, A.H. Velders, I. Hermosín-Gutiérrez, Flavonol 3-O-glycosides series of Vitis vinifera Cv. Petit Verdot red wine grapes, J. Agric. Food Chem. 57 (2009) 209–219. https://doi.org/10.1021/jf802863g.

[49] I. Fernandes, R. Pérez-Gregorio, S. Soares, N. Mateus, V. de Freitas, Wine Flavonoids in Health and Disease Prevention, Molecules. 22 (2017) 292. https://doi.org/10.3390/molecules22020292.

[50] S.D. Cohen, J.M. Tarara, G.A. Gambetta, M.A. Matthews, J.A. Kennedy, Impact of diurnal temperature variation on grape berry development, proanthocyanidin accumulation, and the expression of flavonoid pathway genes, J. Exp. Bot. 63 (2012) 2655–2665. https://doi.org/10.1093/jxb/err449.

[51] J.A. Kennedy, M.A. Matthews, A.L. Waterhouse, Effect of Maturity and Vine Water Status on Grape Skin and Wine Flavonoids, Am. J. Enol. Vitic. 53 (2002).

[52] F. Fang, J.M. Li, Q.H. Pan, W.D. Huang, Determination of red wine flavonoids by HPLC and effect of aging, Food Chem. 101 (2007) 428–433. https://doi.org/10.1016/j.foodchem.2005.12.036.

[53] F. Fang, J.M. Li, P. Zhang, K. Tang, W. Wang, Q.H. Pan, W.D. Huang, Effects of grape variety, harvest date, fermentation vessel and wine ageing on flavonoid concentration in red wines, Food Res. Int. 41 (2008) 53–60. https://doi.org/10.1016/j.foodres.2007.09.004.

[54] M. Nardini, I. Garaguso, Effect of Sulfites on Antioxidant Activity, Total Polyphenols, and Flavonoid Measurements in White Wine, Foods. 7 (2018) 35. https://doi.org/10.3390/foods7030035.

[55] I.S. Arvanitoyannis, M.N. Katsota, E.P. Psarra, E.H. Soufleros, S. Kallithraka, Application of quality control methods for assessing wine authenticity: Use of multivariate analysis (chemometrics), Trends Food Sci. Technol. 10 (1999) 321–336. https://doi.org/10.1016/S0924-2244(99)00053-9.

[56] K. Chira, M. Jourdes, P.L. Teissedre, Cabernet sauvignon red wine astringency quality control by tannin characterization and polymerization during storage, Eur. Food Res. Technol. 234 (2012) 253–261. https://doi.org/10.1007/s00217-011-1627-1.

[57] M. Amargianitaki, A. Spyros, NMR-based metabolomics in wine quality control and authentication,

Chem. Biol. Technol. Agric. 4 (2017) 1–12. https://doi.org/10.1186/s40538-017-0092-x.

[58] J.D. Nunes-Miranda, G. Igrejas, E. Araujo, M. Reboiro-Jato, J.L. Capelo, Mass Spectrometry-Based Fingerprinting of Proteins and Peptides in Wine Quality Control: A Critical Overview, Crit. Rev. Food Sci. Nutr. 53 (2013) 751–759. https://doi.org/10.1080/10408398.2011.557514.

[59] C. Roullier-Gall, M. Witting, D. Tziotis, A. Ruf, R.D. Gougeon, P. Schmitt-Kopplin, Integrating analytical resolutions in non-targeted wine metabolomics, Tetrahedron. 71 (2015) 2983–2990. https://doi.org/10.1016/j.tet.2015.02.054.

[60] C. Roullier-Gall, M. Witting, R.D. Gougeon, P. Schmitt-Kopplin, High precision mass measurements for wine metabolomics, Front. Chem. 2 (2014) 102. https://doi.org/10.3389/fchem.2014.00102.

[61] R. Pascale, G. Bianco, T.R.I. Cataldi, P.S. Kopplin, F. Bosco, L. Vignola, J. Uhl, M. Lucio, L. Milella, Mass spectrometry-based phytochemical screening for hypoglycemic activity of Fagioli di Sarconi beans (Phaseolus vulgaris L.), Food Chem. 242 (2018) 497–504. https://doi.org/10.1016/j.foodchem.2017.09.091.

[62] A. Santarsiero, A. Onzo, R. Pascale, M.A. Acquavia, M. Coviello, P. Convertini, S. Todisco, M. Marsico, C. Pifano, P. Iannece, C. Gaeta, S. D'Angelo, M.C. Padula, G. Bianco, V. Infantino, G. Martelli, Pistacia lentiscus Hydrosol: Untargeted Metabolomic Analysis and Anti-Inflammatory Activity Mediated by NF-κB and the Citrate Pathway, Oxid. Med. Cell. Longev. 2020 (2020) 1–14. https://doi.org/10.1155/2020/4264815.

[63] V. Alba, C. Bergamini, M. Gasparro, F. Mazzone, S. Roccotelli, D. Antonacci, A.R. Caputo, Basivin_SUD: la ricerca del germoplasma viticolo in Basilicata, 2nd ed., n.d.

[64] D.P.A. Kilgour, R. Wills, Y. Qi, P.B. O'Connor, Autophaser: An algorithm for automated generation of absorption mode spectra for FT-ICR MS, Anal. Chem. 85 (2013) 3903–3911. https://doi.org/10.1021/ac303289c.

[65] D.P.A. Kilgour, S.L. Van Orden, Absorption mode Fourier transform mass spectrometry with no baseline correction using a novel asymmetric apodization function, Rapid Commun. Mass Spectrom. 29 (2015) 1009–1018. https://doi.org/10.1002/rcm.7190.

[66] D.P.A. Kilgour, M.J. Neal, A.J. Soulby, P.B. O'Connor, Improved optimization of the Fourier transform ion cyclotron resonance mass spectrometry phase correction function using a genetic algorithm, Rapid Commun. Mass Spectrom. 27 (2013) 1977–1982. https://doi.org/10.1002/rcm.6658.

[67] A.T. Zielinski, I. Kourtchev, C. Bortolini, S.J. Fuller, C. Giorio, O.A.M. Popoola, S. Bogialli, A. Tapparo, R.L. Jones, M. Kalberer, A new processing scheme for ultra-high resolution direct infusion mass spectrometry data, Atmos. Environ. 178 (2018) 129–139. https://doi.org/10.1016/j.atmosenv.2018.01.034.

[68] K. Longnecker, E.B. Kujawinski, Using network analysis to discern compositional patterns in ultrahigh-resolution mass spectrometry data of dissolved organic matter, Rapid Commun. Mass Spectrom. (2016) 2388–2394. https://doi.org/10.1002/rcm.7719.

[69] F. Moritz, M. Kaling, J.-P. Schnitzler, P. Schmitt-Kopplin, Characterization of poplar metabotypes via mass difference enrichment analysis, Plant. Cell Environ. 40 (2017) 1057–1073. https://doi.org/10.1111/pce.12878.

[70] E. V. Kunenkov, A.S. Kononikhin, I. V. Perminova, N. Hertkorn, A. Gaspar, P. Schmitt-Kopplin, I.A. Popov, A. V. Garmash, E.N. Nikolaev, Total mass difference statistics algorithm: A new approach to identification of high-mass building blocks in electrospray ionization fourier transform ion cyclotron mass spectrometry data of natural organic matter, Anal. Chem. 81 (2009) 10106–10115. https://doi.org/10.1021/ac901476u.

[71] Y. Qi, P.B. O'Connor, Data processing in Fourier transform ion cyclotron resonance mass spectrometry, Mass Spectrom. Rev. 33 (2014) 333–352. https://doi.org/10.1002/mas.21414.

[72] E.N. Nikolaev, Y.I. Kostyukevich, G.N. Vladimirov, Fourier transform ion cyclotron resonance (FT ICR) mass spectrometry: Theory and simulations, Mass Spectrom. Rev. 35 (2016) 219–258. https://doi.org/10.1002/mas.21422.

[73] M.V. Moreno-Arribas, M.C. Polo, Wine chemistry and biochemistry, Springer-Verlag, New York, 2009. https://doi.org/10.1007/978-0-387-74118-5.

[74] A.F. Ortega, A. Lopez-Toledano, M. Mayen, J. Merida, M. Medina, Changes in Color and Phenolic Compounds During Oxidative Aging of Sherry White Wine, J. Food Sci. 68 (2003) 2461–2468. https://doi.org/10.1111/j.1365-2621.2003.tb07046.x.

[75] Á.F. Recamales, A. Sayago, M.L. González-Miret, D. Hernanz, The effect of time and storage conditions on the phenolic composition and colour of white wine, Food Res. Int. 39 (2006) 220–229. https://doi.org/10.1016/j.foodres.2005.07.009.

4.7. Figures

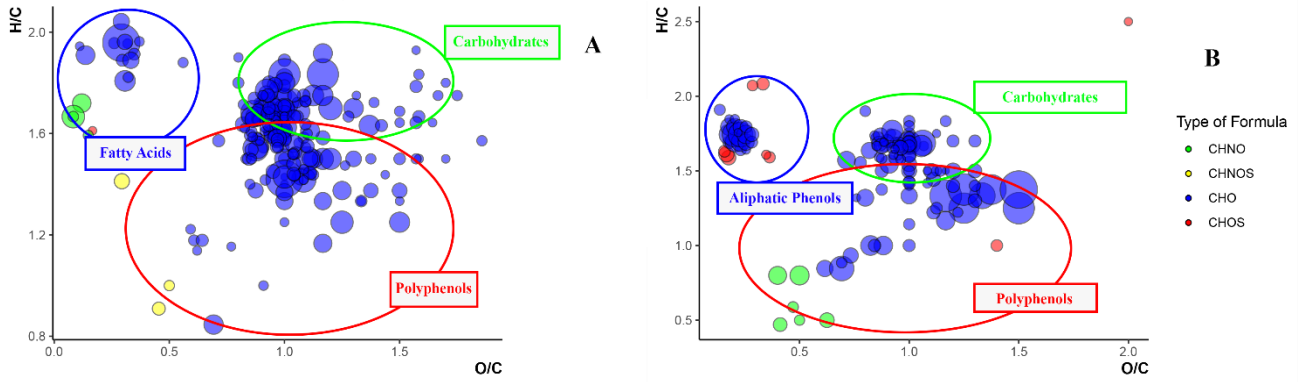

Figure 2 *Van Krevelen plots of Colata Murro (A) and Aglianico Bianco (B) red and white wine samples, respectively, obtained from related ESI(-)-FT-ICR MS data. Types of formula are distinguished by colors (green for CHNO, yellow for CHNOS, blue for CHO and red for CHOS).*

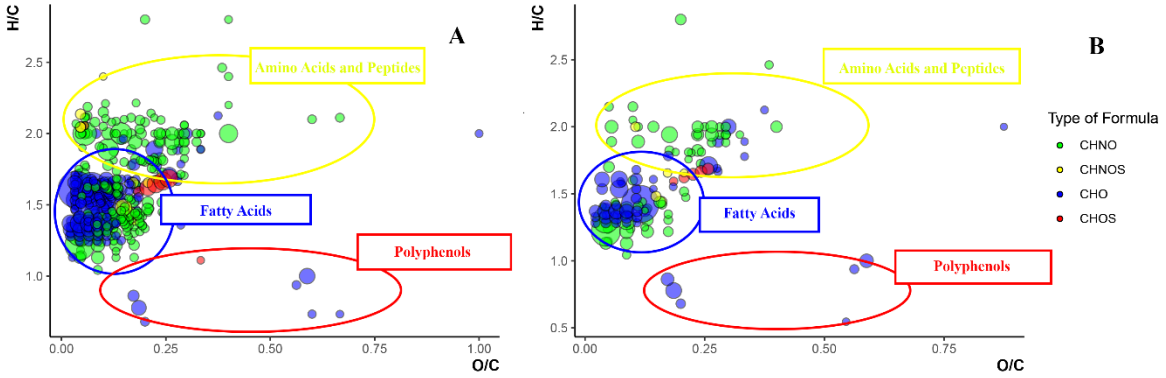

Figure 2 *Van Krevelen plots of Malvasia ad acino piccolo (A) and Colata Murro (B) white and red wine samples, respectively, obtained from related ESI(+)-FT-ICR MS data. Types of formula are distinguished by colors (green for CHNO, yellow for CHNOS, blue for CHO and red for CHOS).*

4.8. Supplementary Material

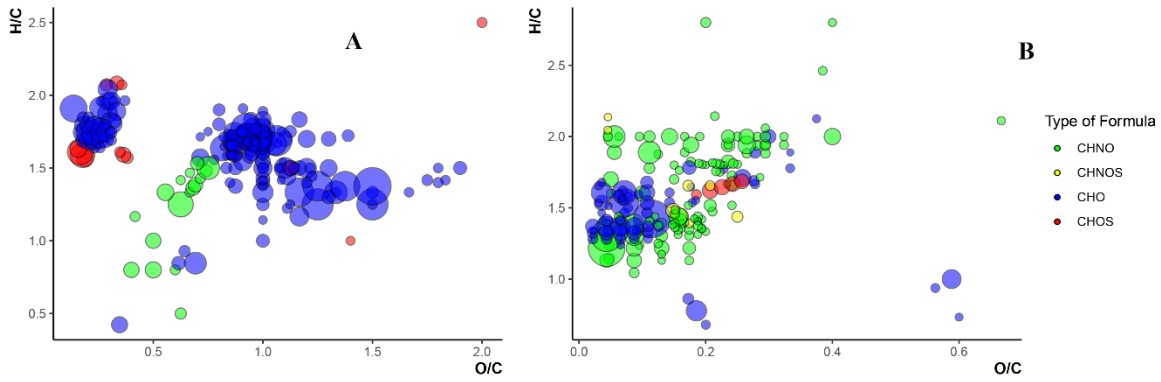

Figure S1 *Van Krevelen plot of Giosana white wine sample, obtained from related ESI(-) (A) and ESI(+)-FT-ICR MS (B) data, respectively. Types of formula are distinguished by colors (green for CHNO, yellow for CHNOS, blue for CHO and red for CHOS).*

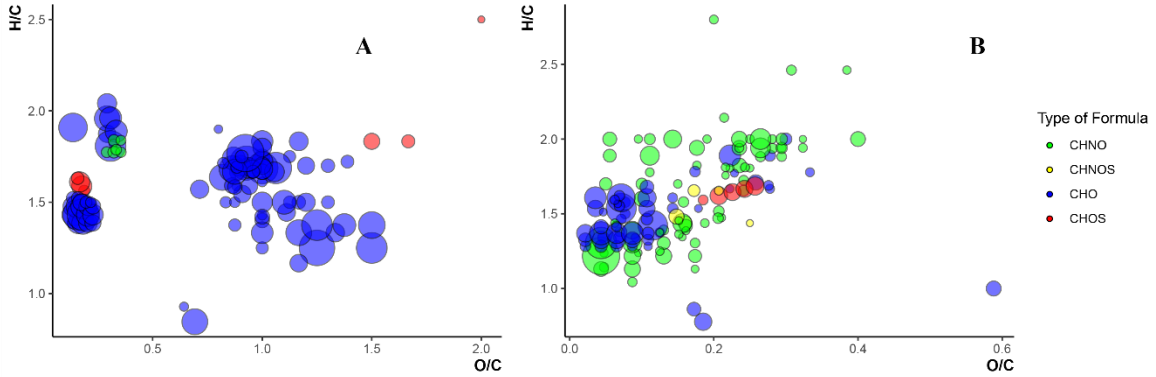

Figure S2 *Van Krevelen plot of Guisana white wine sample, obtained from related ESI(-) (A) and ESI(+)-FT-ICR MS (B) data, respectively. Types of formula are distinguished by colors (green for CHNO, yellow for CHNOS, blue for CHO and red for CHOS).*

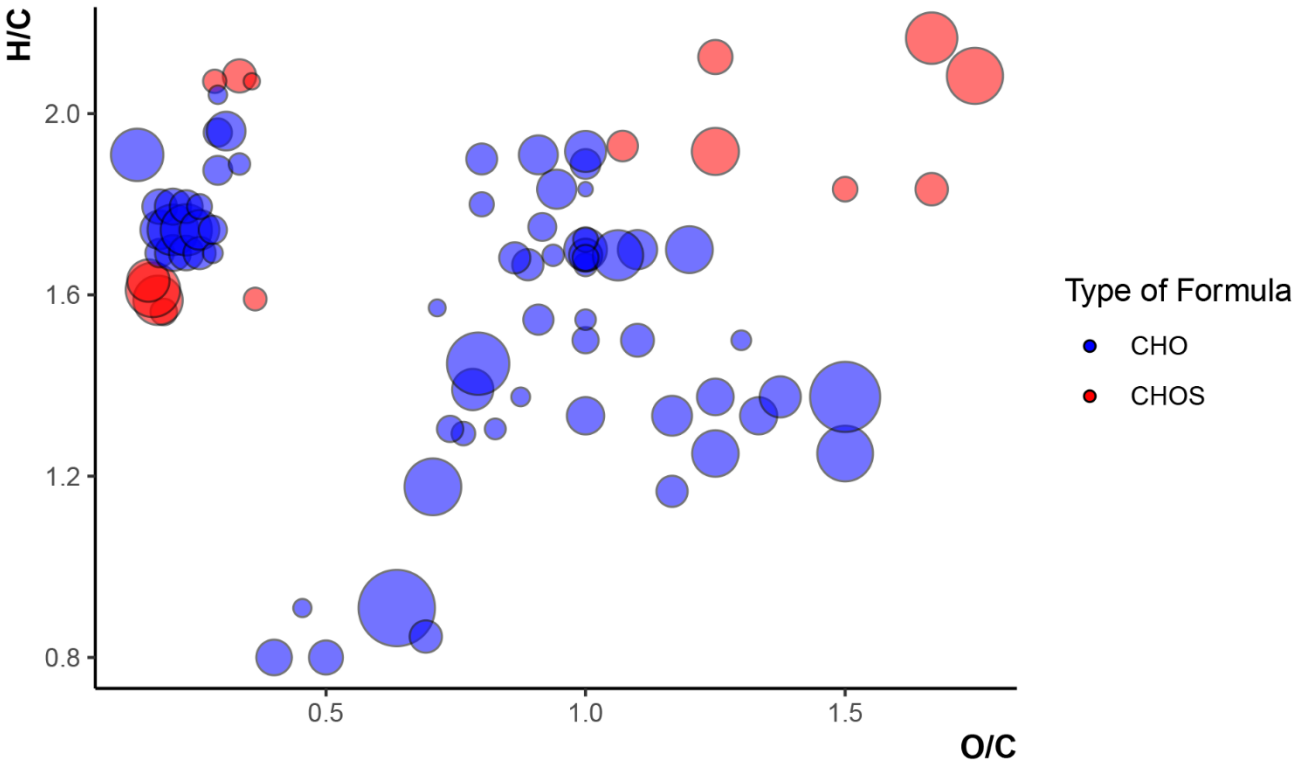

Figure S3 *Van Krevelen plot of Malvasia ad acino piccolo white wine sample, obtained from related ESI(-)-FT-ICR MS data. Types of formula are distinguished by colors (green for CHNO, yellow for CHNOS, blue for CHO and red for CHOS).*

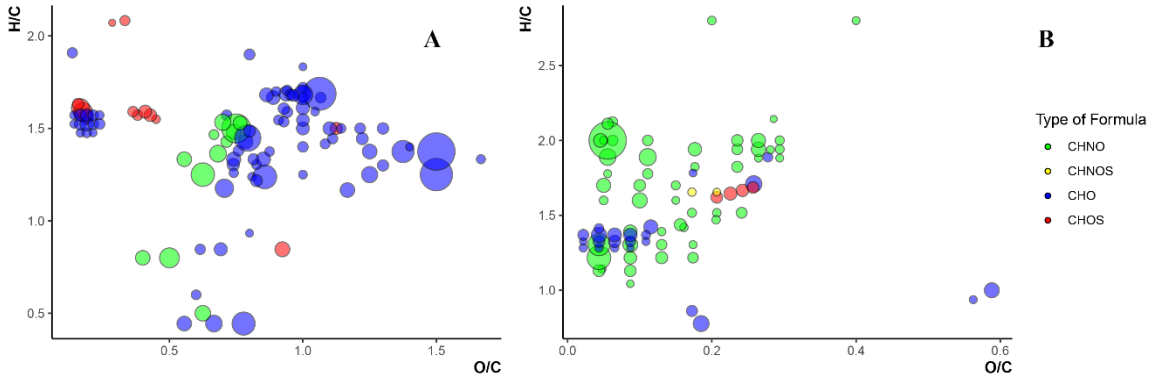

Figure S4 *Van Krevelen plot of Santa Sofia white wine sample, obtained from related ESI(-) (A) and ESI(+)-FT-ICR MS (B) data, respectively. Types of formula are distinguished by colors (green for CHNO, yellow for CHNOS, blue for CHO and red for CHOS).*

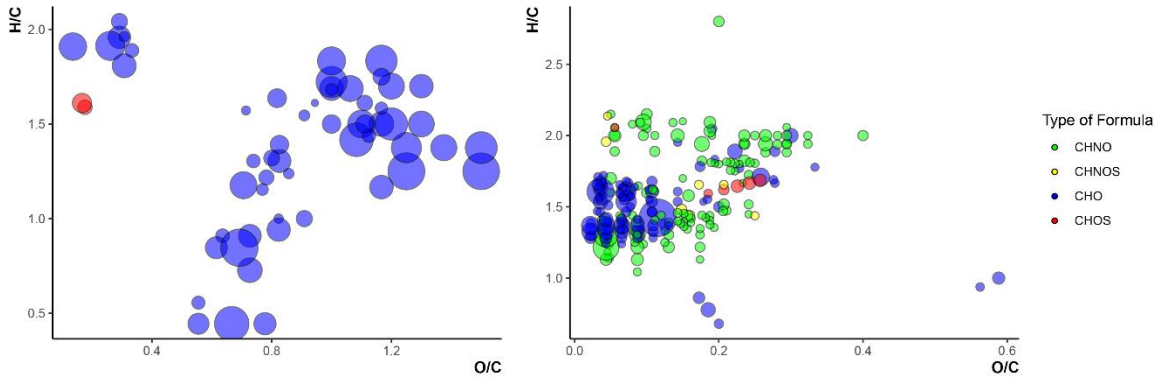

Figure S5 *Van Krevelen plot of Plavina white wine sample, obtained from related ESI(-) (A) and ESI(+)-FT-ICR MS (B) data, respectively. Types of formula are distinguished by colors (green for CHNO, yellow for CHNOS, blue for CHO and red for CHOS).*

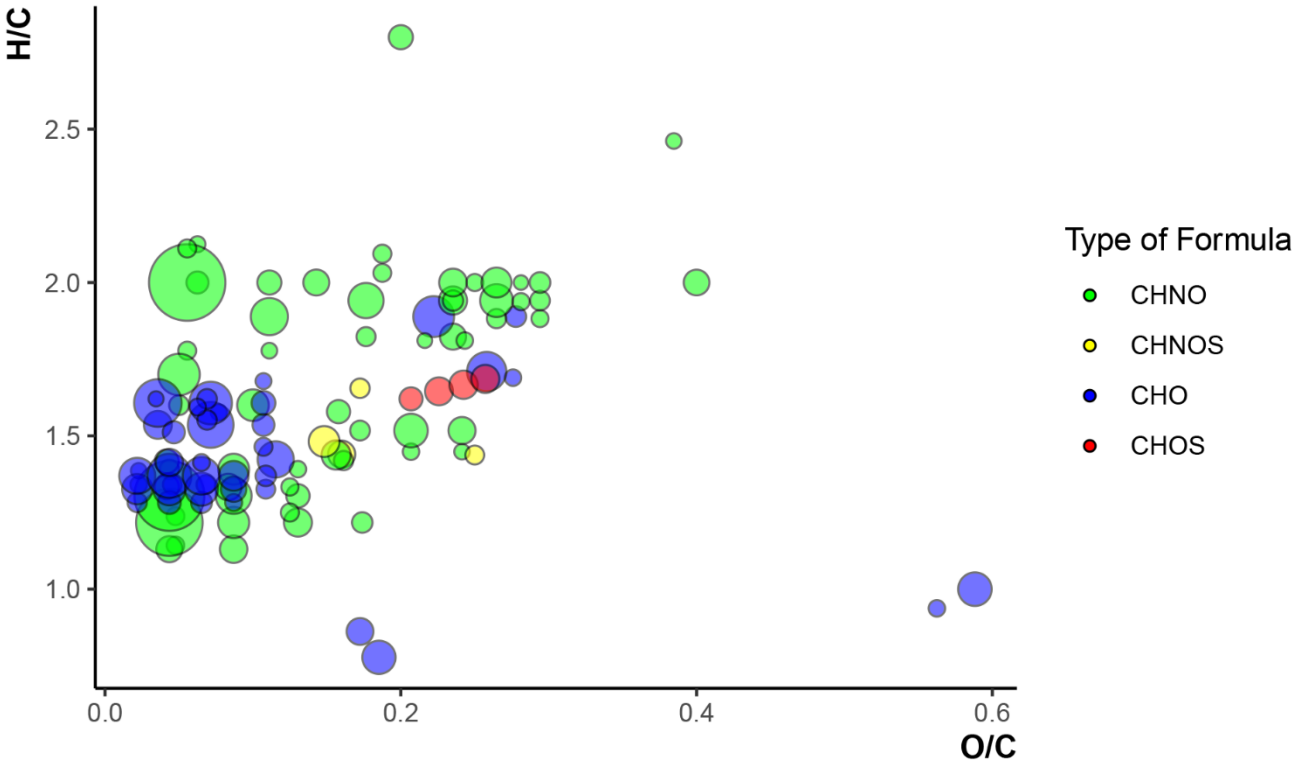

Figure S6 *Van Krevelen plot of Aglianico Bianco white wine sample, obtained from related ESI(+)-FT-ICR MS data. Types of formula are distinguished by colors (green for CHNO, yellow for CHNOS, blue for CHO and red for CHOS).*

5. Contribute 3

OIFA: a R Shiny app for the interactive elaboration of Metabolomic High Resolution Mass Spectrometry data

A. Onzo[1], M. Acquavia[1], R. Pascale[1], P. Iannece[2], C. Gaeta[2], Y. Tsybin[3], G. Bianco[1]*

[1]*Università degli Studi della Basilicata, Dipartimento di Scienze, Via dell'Ateneo Lucano 10, Potenza, Italy;*

[2]*Università degli Studi di Salerno, Dipartimento di Chimica e Biologia, Via Giovanni Paolo II 132, Fisciano, Italy*

[3]*Spectroswiss Sarl, EPFL Innovation Park, Building I, 1015 Lausanne, Switzerland.*

5.1. Abstract

High Resolution Mass Spectrometry is becoming the technique of election for the identification of thousands of metabolites in very complex matrices, such as petroleum, human tissue or food products. Among the advantages provided by its employment, the most important one consists of providing higher levels of accuracies, thanks to which elemental formulas can be calculated and assigned to Mass Spectrometry signals. However, high accuracy is not enough to obtain unique formula assignments for higher mass-to-charge ratios (*m/*z). To reduce the number of possible candidates, several tools and filters are employed. Among these, Kendrick plot and Molecular Network use is promising, since these approaches exploit the chemistry of analyzed samples to accomplish this task, providing important mechanistic information and clues on the variety of derivatives present in it too. In this work, a new R Shiny app, i.e. the Omics Interactive Formula Assignment (OIFA) software, is presented, which allowed us to assume an interactive approach for the elemental formula assignment process by directly working on Kendrick plot and Molecular Networks by means of a point-and-click approach. In this way, low resolved homologous series can be identified easily, elemental formulas for single members can be calculated, most frequent building blocks can be identified and immediately used for formula assignment purposes and systematic effects acting on data could be uncovered. OIFA was crucial for the elaboration of metabolomic High Resolution Mass Spectrometry data of two Italian typical food products, i.e. the *Fagioli Bianchi di Rotonda PDO* beans and the *Melanzane Rosse di Rotonda PDO* eggplants, leading to the identification of present metabolites and their classification by means of Van Krevelen plots.

5.2. Introduction

Mass spectrometry (MS) technology saw a remarkable evolution during last years, becoming an indispensable tool for compound characterization [1–3]. In detail, development of softer ionization techniques and specific types of analyzers made this technique able to provide useful information about analytes [4,5]. As an example, Matrix-Assisted Laser Desorption and Electrospray Ionizations (MALDI and ESI, respectively) can be used as a way to ionize polar higher molecular weight and thermal labile compounds with a reduced or depleted degree of fragmentation, thus assuring the observation of relative pseudomolecular ions, through which the monoisotopic mass of the analytes of interest can be deduced, which can be more accurate if the soft ionization source is coupled to accurate mass measurement analyzers such as Orbitrap and Fourier-Transform Ion Cyclotron resonance (FT-ICR) cell [6–8]. This leads to the calculation of compound chemical formulas, crucial for the identification of unknown species and to shed some light about their chemistry. Indeed, knowledge of chemical formulas could lead to the differentiation of metabolites present in natural samples or to the evaluation of the range of derivatives of an analyte of interest, like pollutants in an air sample [9,10]. Orbitrap and FT-ICR analyzers are able to identify thousands of ions simultaneously, with accuracies in the range of parts-per-million (ppm) and sub-ppm, assuring a huge filtering of possible formulas for a single accurate mass value [10–17]. However, higher accuracies alone cannot provide unique formula assignments for accurate masses > 200 Da [18], thus making compulsory the utilization of other tools to further filter formulas of higher accurate mass species. To this aim, different strategies are assumed. One of these relies on the matching of experimental and theoretical isotopic patterns, the calculation of which is possible from a chemical formula by using natural isotopic abundances and employing a polynomial expression [1,18]. A mass spectrum of sufficient resolution can report the ionic species containing the monoisotopic elements [M] and an isotopic element [M + 1] or elements [M + n]. Both the presence of isotopic peaks and their relative abundance to the monoisotopic peak can provide evidence for constraining the number of elemental formulas. In particular, this is useful when analyzers such as FT-ICR and Orbitrap are employed, able to provide a dramatical improvement of peak resolution, thus leading to the observation of isotopic fine structures [19]. To apply chemical and heuristic rules for elemental formula filtering consists in another strategy commonly adopted. These include relatively well-known chemical rules such as the nitrogen, the hydrogen and ring-plus-double-bond equivalence rules [18], and established principles relating to valence electron theory (i.e. Lewis ‘octet rule’) [18]. Finally, an effective method for analyte identification uses accurate mass data that is searched in a chemical database such as Chemical Abstracts Service (CAS) registry or ChemSpider [20]. Searches can be carried out using a range of properties, but typically using elemental formula or molecular weight as the active search parameter where results are ranked according to the highest number of references within the database. Despite the impressive reduction of candidate formulas supplied by these strategies, still it’s not possible to obtain unique assignments for accurate mass values >500 Da. Moreover, chemical database search not always provides reliable results, since some derivatives observed in mass spectra could have not ever been reported before [21]. To overcome this problem, different visualization tools where introduced to guide formula assignment of high mass-to-charge (*m/z*) species, i.e. the Kendrick plot and the Molecular Network Analysis

[22,23]. The former can be obtained by converting observed accurate mass values to the so-called Kendrick scale by means of a linear expression, which depends on the choice of a specific *building block*, i.e. a group of atoms which usually corresponds to a specific chemical reaction [24]. Direct consequence of this conversion is the organization of MS signals in homologous series, whose members show the same mass defect (or Kendrick Mass Defect, KMD, since accurate masses have been converted to the Kendrick scale) [24]. For this reason, by plotting MS signals as points in a 2D Kendrick Nominal Mass/KMD plot, it's possible to observe different homologous series as points ordered on a line parallel to the x-axis [22]. The key feature of homologous series is the fact that members differ from each other by a certain amount of the chosen building block, thus, by assuming that one member has been successfully assigned to one elemental formula, it would be possible to calculate formulas of other members by adding or subtracting the right amount of building block unit [22]. A step forward to this approach is represented by the Molecular Network Analysis, thanks to which more than one building block can be considered simultaneously by applying network analysis on experimental MS data [21,23]. In detail, networks are built by assuming that nodes correspond to experimental *m/z* ratios and edges to mass differences, which in turn are related to a specific building block [25]. Ionic species are, then, organized in clusters, which can differ from each other by the amount of several heteroatoms or by their RDBE value, depending from which kind of building blocks are considered, and a single unique formula assignment allows the calculation of one elemental formula for every node of the cluster [25]. Utilization of these tools revolutionized the way High Resolution MS is used in metabolomics, making it the technique of choice for untargeted *omics* analysis, such as *petroleomics* and *metabolomics*, being able here to identify thousands of compounds with a single direct-injection analysis and providing valuable information on the chemistry of samples [10,26–29]. However, full automation of these processes could lead to errors, especially when systematic effects influence raw data and/or a huge number of artefacts, such as wiggles, are present in recorded mass spectrum [30]. These aspects could get in wrong assignments, which can propagate all over the MS data by using cited tools. Thus, in this work, a new interactive app, i.e. the Omics Interactive Formula Assignment (OIFA) software, developed by means of the R Shiny package [31], is presented and used to perform a metabolomic profile of two Italian typical food products, i.e. the *Fagioli Bianchi di Rotonda PDO* (Protected designation of Origin) beans and the *Melanzane Rosse di Rotonda PDO* eggplants. Direct-injection High Resolution Mass Spectrometry data were elaborated with OIFA, which allowed to carry on a more reliable formula assignment step thanks to the possibility to directly interact with related Kendrick plot and Molecular Network. This approach allowed to make a distinction among low resolved homologous series and to select starting points for calculation of other members elemental formula. Moreover, a closer look to longer homologous series allowed us to identify the presence of systematic effects and to measure their entity.

5.3. Materials and Methods

Chemicals

Sodium trifluoroacetate (NaTFA, 98%) and ethanol (EtOH, 96%) were purchased from Sigma-Aldrich (Milano, Italy). Pure nitrogen (99.996%) was delivered to the MS system as the sheath gas. Ultra-pure water was employed and was produced using a Milli-Q RG system from Millipore (Bedford, MA, USA). *Fagioli*

Bianchi di Rotonda PDO beans (*Phaseolus Vulgaris L.*) and *Melanzane Rosse di Rotonda PDO* eggplants (*Solanum melongena L.*) were provided by the Agency for Development and Innovation in Agriculture (ALSIA, Agenzia Lucana di Sviluppo e di Innovazione in Agricoltura, Rotonda, Italy).

Sample Preparation

Extracts of *Fagioli Bianchi di Rotonda PDO* beans and *Melanzane Rosse di Rotonda PDO* eggplants were obtained by following modified procedures based on two previously reported methods, with slight modifications [10,32]. Beans were grounded to a fine powder using a home miller, while eggplants were subjected to lyophilization prior to the grinding step. 10 mL of a solution of EtOH and ultrapure water in a 7:3 ratio were added to 1 g of the bean sample, while 10 mL of MeOH was added to 200 mg of the eggplant one. Metabolites were extracted by means of the Ultrasound Assisted Extraction (UAE) technique at room temperature (Sonorex Super RK 100/H sonicator; Bandelin electronic, Berlin, Germany) with a 35 kHz automatic frequency control and a high-frequency power of 80 W. The UAE was applied for 6 h for the bean sample and for 15 min for the eggplant one. Extracts were passed through a PTFE 0.22 µm filter and were injected into the MS system without any further pre-treatment. Blank samples were prepared by applying every preparation step to 10 mL of EtOH/H_2O 7:3 and to 10 mL of MeOH, respectively. Extracts were stored under -20°C prior to the analysis.

Mass Spectrometry

ESI (±) Fourier Transform Ion Cyclotron Resonance Mass Spectrometry (FT-ICR MS) technique was used for the untargeted metabolomic analysis of the samples. High Resolution Mass Spectra were acquired on a Bruker (Bruker Daltonik GmbH, Bremen, Germany) solariX XR Fourier transform ion cyclotron resonance mass spectrometer (FT-ICR-MS) equipped with a 7T superconducting magnet and an ESI source. The capillary voltage was set to 3.9 and -4.5 kV for negative and positive ionization modes, respectively, with a nebulizer gas pressure of 1.2 bar and dry gas flow rate of 4 L/min at 200 °C. Spectra were acquired with a Time Domain size of 16 mega-word, an accumulation time of 0.1 s and a mass range of 100-2000 *m/z*. Moreover, the average number of scans was set to 50. Before the analysis, the mass spectrometer was externally calibrated with NaTFA. Once recorded, FT-ICR mass spectra were submitted to several data pre-treatment steps. In detail, recorded free induction decays (FIDs) were subjected to apodization and related absorption mode mass spectra were obtained. Phase correction, mass recalibration and baseline correction have been performed [33,34]. Finally, noise filtering has been performed on mass spectra by following the N-Sigma methodology approach [9]. More specifically, noise level has been estimated and peaks showing a signal-to-noise ratio (S/N) higher than 2 were retained [10]. Thus, obtained FT-ICR mass spectra were exported to peak lists. From these, possible elemental formulas were assigned to each MS signal by direct comparison of *m/z* ratio values to an in-house formula database. The latter was obtained by considering several constraints too, such as atoms number limitations, i.e. $C \leq 100$, $H \leq 200$, $O \leq 80$, $N \leq 5$ and $S \leq 1$ [10,29], restrictions on atoms to carbon number ratios, i.e. $0.2 \leq H/C \leq 3.1$, $O/C \leq 2$, $N/C \leq 1.3$ and $S/C \leq 0.8$, RDBE > 0, hydrogen and nitrogen rule (the latter applied for $m/z \leq 500$ only) [18]. So obtained data were subjected to an interactive Kendrick mass

defect (KMD) analysis to help formula filtering and calculation by means of the OIFA R Shiny app. A list of building blocks was previously prepared, each of them corresponding to common biochemical reactions [21], and chosen subsequently for the analysis. To further improve the reliability of results, an interactive Molecular Network analysis was carried on. For this step, experimental mass differences values were examined and only those comprised in the range ± 1 mDa of the building block exact mass were considered for the construction of molecular networks [21,25]. Building blocks with occurrences lower than a threshold value (properly chosen to remove all the noisy data [35]) were excluded, being higher the probability for these to have occurred randomly [35]. FT-ICR MS raw data pre-processing has been done by using AutoVectis Pro (v.8.9, Spectroswiss, Lausanne, Switzerland), while interactive KMD and Molecular Network analyses were done by means of the OIFA tool, developed by using the R software (v3.6.3, www.r-project.org) and employing different R packages [31,36–38].

5.4. Results and Discussion

High Resolution Mass Spectrometry is a powerful technique for the untargeted analysis of complex natural matrices, being able to identify thousands of metabolites with a single direct analysis [10,26,29]. However, obtained data are just too complex to be interpreted at first sight, a feature that makes the utilization of dedicated visualization tools compulsory for compound identification and classification. OIFA R Shiny app is able to build interactive Kendrick plots and Molecular Networks from data, allowing a critical evaluation of identified homologous series and to monitor the formula assignment process step-by-step. OIFA presents a friendly GUI, through which one is able to upload personal data for elaboration. Firstly, a database of building blocks should be uploaded, without which the software would not be able to make the Kendrick plot and Molecular Network. Further, it's possible to upload personal data, which would consist in a dataset in which *m/z* data are reported. It's worth noting, here, that few elemental formulas should already be assigned, in order to calculate other formulas through KMD and Molecular Network analyses [22,23]. Once done, the software calculates related Kendrick plot and Molecular Network in two different tabs immediately. For the former (**Figure 1**), it's possible to perform a series of operations by assuming a point-and-click approach, together with the utilization of action buttons. Principally, it's possible to choose the building block considered for the KMD analysis by selecting it on the left panel of the GUI. Hovering on a point of the Kendrick plot allow to display information of that ionic species, such as the *m/z* value, the intensity and KMD related properties, i.e. the KMD, the Kendrick Mass (KM), the Kendrick Nominal Mass (KNM) and the *z**, another useful parameter in KMD analysis which adds another level of differentiation of ionic species in different nominal mass families [24]. Then, clicking on a point of the Kendrick plot allows to highlight it, together with the other members belonging to the same homologous series, showing the same *z** and a KMD value lying in a range defined by the KMD mean value of the series ± an error (ΔKMD). For what concerns this aspect, since obtained data consists of accurate mass values, single members show different KMD values, randomly distributed around a mean. Obviously, the higher is the mass measurement accuracy, the narrower is this range [22,24]. So, points of the Kendrick plot belonging to the same homologous series should appear randomly distributed around a horizontal line ideally parallel to the x-axis. Deviations from this situation would be caused by the presence of

systematic effects, derived from, for example, space-charge interactions among the ions inside the ICR cell [30]. The ΔKMD value considered for the identification of members of the homologous series can be set by the user on the left panel. Once the homologous series is selected, it can be isolated by pressing on the unselected point legenda (**Figure 2**). This allows to focus the attention of the user on selected homologous series completely, making possible the observation of systematic effects affecting the data, suggested by a drift of the ideal horizontal KMD line. Further, once clicked on a point, a table is displayed under the Kendrick plot, showing assigned formulas for members of the highlighted homologous series (NA would occur if no formula has been assigned to the corresponding MS signal). It's possible, then, to calculate formulas for all the member or, if there are elemental formulas assigned to members already, to filter them by clicking on the starting point and using related action buttons. For the calculation, it's possible to decide if to use or not the Nitrogen rule by using related radio buttons, since it's not respected by >500 m/z ratios [18]. Another useful function provided by the software allows to identify all the homologous series related to the other building blocks present in the uploaded database to which a selected point belongs to, together with related total number of members and KMD standard deviation. Thanks to this feature, the user can know in advance which of the building blocks should be selected to calculate a formula for the selected point (**Table 1**). Despite complete automation of the process could lead to errors, especially when artefacts are present, it's still possible to switch to it by dedicated action buttons. In particular, two action buttons are present to perform automatic formula filtering or calculation, depending on if elemental formula candidates are present in the *m/z* dataset already or they should be calculated from unique formula assignments. Other two tabs are related to Molecular Network analysis, one of which containing the neural network in which all the ionic species are connected by building blocks forming clusters. In detail, the app doesn't consider all the building blocks provided by the user for the construction of the network but focalizes on the ones with a number of occurrences higher than a threshold value, set by the user on the left panel. Setting this parameter is not straightforward and needs a careful observation of mass differences among experimental peaks [35]. To this aim, the software identifies and collects all the experimental mass differences, comparing them to the building blocks present in uploaded database. In this way, mass difference values within an error (set by the user) are correlated to specific building blocks and their occurrence is calculated. Thus, a plot of mass differences and related occurrences is provided to the user into the second tab (**Figure 3**). Mass differences that were successfully assigned to a building block can be highlighted by using dedicated radio buttons under the plot and x- and y-axis range can be set by the user. The analysis of the plot allows a rough estimation of the noise [35], which can be cut off by setting the proper threshold value on the left panel. Building blocks related to the noise are, then, not considered for the making of the neural network, since the probability of having occurred randomly is not negligible for them [35]. As for the Kendrick plot, the Molecular Network can be used with a point-and-click approach (**Figure 4**). More specifically, clicking on the nodes allows to display its information as a data table, such as *m/z*, peak intensity, assigned formula and Kendrick properties. Moreover, it's possible to interact with the network through dedicated action and radio buttons, such as displaying labels of the edges (thus revealing corresponding building blocks) or calculating or filtering elemental formulas of the other nodes of the cluster

starting from the selected one, depending on which kind of action button is used. Here too, the process can be automated, starting calculation from assigned nodes. OIFA software has been used in this work for the elaboration of High Resolution FT-ICR MS data of two Italian typical food products, i.e. the *Fagioli Bianchi di Rotonda PDO* beans and the *Melanzane Rosse di Rotonda PDO* eggplants, cultivated in the Basilicata region. Obtained raw data were pre-treated to calculate de-noised, phase corrected and full informative absorption mode mass spectra, the peaks of which were preliminary assigned to elemental formulas contained in a database subjected to a series of constraints (see *Materials and Methods*). Then, redundance resolution and further elemental formula calculation were performed by the OIFA software. Kendrick plots were immediately calculated (**Figure 1**) and used for formula assignment as described previously. Furthermore, selection of long homologous series made possible the assessment of the presence of a systematic error by the observation of a change in the angular coefficient of the ideal horizontal KMD line (**Figure 2**). The presence of the same effect for other homologous series noticed by assuming this approach assured that the same error affects all the dataset. For this reason, recalibration with a reference list of ubiquitous fatty acids [39,40] was performed before further performing formula assignment. The analysis of mass difference occurrences allowed the rough evaluation of a threshold value for the noise cut-off (**Figure 3**). Obtained networks show the presence of different clusters (**Figure 4**), defining formula families which differ from each other by the presence or the absence of nitrogen, sodium and potassium atoms. Unique formula assignments are distinguished here by colour (red for unique assignments, blue for redundances and grey for not assigned MS signals) and could be selected to assign elemental formulas to nodes of the cluster they belong. In this way, 52 and 192 reliable unique formulas were obtained from the elaboration of the *Fagioli Bianchi di Rotonda PDO* bean and the *Melanzane Rosse di Rotonda PDO* eggplant datasets, respectively. Furthermore, so obtained formulas were used to make a Van Krevelen plot (**Figure 5**), through which the classification of identified metabolites in major classes and an evaluation of related derivatives could be accomplished [10,29,41]. In particular, for what concerns *Fagioli Bianchi di Rotonda PDO* beans, nitrogen and sulphur-bearing amino acids and peptides, together with related alkyl derivatives, were identified. Moreover, the presence of few unsaturated fatty acid derivatives could be deduced by CHO formula type points lying on the left middle part of the plot. Observation seems to agree to what is reported in literature for *Phaseolus vulgaris L.* cultivars, since beans are rich sources of amino acids and peptides and the presence of high RDBE fatty acids, such as linolenic and linoleic acids, was accessed too [42]. Despite this, *Fagioli Bianchi di Rotonda PDO* bean extract seems to be pour for what concerns metabolite diversity, unlike most of the other bean cultivars [10,42]. For *Melanzane Rosse di Rotonda PDO* eggplants, two major metabolite classes could be identified, i.e. fatty acids and carbohydrates. Few points suggest the presence of carotenoids and peptide derivatives too. Moreover, the presence of high RDBE CHNO-type compounds (points in the middle part of the Van Krevelen plot) could indicate the presence of nitrogen-containing steroidal glycosides, commonly found in members of the genus *Solanum* [43]. However, unlike what was found for different cultivars of the species *Solanum melongena L.* [43], few CHO formula type points are present in the middle part of the Van Krevelen plot, suggesting a low diversity of phenolic derivatives.

5.5. Conclusions

A new interactive tool for the reliable assignment of elemental formulas to accurate *m/z* ratios, i.e. the Omics Interactive Formula Assignment (OIFA) R Shiny app, has been shown in this work and employed to elaborate High Resolution MS data obtained from the direct analysis of two Italian typical food products, i.e. the *Fagioli Bianchi di Rotonda PDO* beans and the *Melanzane Rosse di Rotonda PDO* eggplants. Thanks to this new tool, the identification and isolation of low resolved homologous series and elemental formula calculation of related members was possible simply by working on related Kendrick plots. Moreover, by means of this tool, identification of higher occurrence building blocks and making of Molecular Network was possible, remarkably improving the elemental formula assignment step. This was crucial for the identification of metabolites present in our sample and their characterization through the making of Van Krevelen plots, thanks to which principal classes of metabolites into the analysed extracts were identified. The source code of the OIFA R Shiny app can be found in the *Supplementary Material*, together with custom functions used by the app.

5.6. References

[1] J.H. Gross, Mass Spectrometry - A Textbook, Springer International Publishing, 2017. https://doi.org/10.1007/978-3-319-54398-7.

[2] M. Stobiecki, Application of mass spectrometry for identification and structural studies of flavonoid glycosides, Phytochemistry. 54 (2000) 237–256. https://doi.org/10.1016/S0031-9422(00)00091-1.

[3] R. Pascale, M.A. Acquavia, T.R.I. Cataldi, A. Onzo, D. Coviello, S.A. Bufo, L. Scrano, R. Ciriello, A. Guerrieri, G. Bianco, Profiling of quercetin glycosides and acyl glycosides in sun-dried peperoni di Senise peppers (Capsicum annuum L.) by a combination of LC-ESI(-)-MS/MS and polarity prediction in reversed-phase separations, Anal. Bioanal. Chem. 412 (2020) 3005–3015. https://doi.org/10.1007/s00216-020-02547-2.

[4] J.M. Daniel, S.D. Friess, S. Rajagopalan, S. Wendt, R. Zenobi, Quantitative determination of noncovalent binding interactions using soft ionization mass spectrometry, Int. J. Mass Spectrom. 216 (2002) 1–27. https://doi.org/10.1016/S1387-3806(02)00585-7.

[5] R. Jirásko, M. Holčapek, Structural analysis of organometallic compounds with soft ionization mass spectrometry, Mass Spectrom. Rev. 30 (2011) 1013–1036. https://doi.org/10.1002/mas.20309.

[6] D.J. Harvey, Matrix-assisted laser desorption/ionization mass spectrometry of carbohydrates and glycoconjugates, Int. J. Mass Spectrom. 226 (2003) 1–35. https://doi.org/10.1016/S1387-3806(02)00968-5.

[7] J. Schiller, R. Süß, J. Arnhold, B. Fuchs, J. Leßig, M. Müller, M. Petković, H. Spalteholz, O. Zschörnig, K. Arnold, Matrix-assisted laser desorption and ionization time-of-flight (MALDI-TOF) mass spectrometry in lipid and phospholipid research, Prog. Lipid Res. 43 (2004) 449–488. https://doi.org/10.1016/j.plipres.2004.08.001.

[8] J.B. Fenn, M. Mann, C.K. Meng, S.F. Wong, C.M. Whitehouse, Electrospray ionization for mass spectrometry of large biomolecules, Science (80-.). 246 (1989) 64–71. https://doi.org/10.1126/science.2675315.

[9] A.T. Zielinski, I. Kourtchev, C. Bortolini, S.J. Fuller, C. Giorio, O.A.M. Popoola, S. Bogialli, A. Tapparo, R.L. Jones, M. Kalberer, A new processing scheme for ultra-high resolution direct infusion mass spectrometry data, Atmos. Environ. 178 (2018) 129–139. https://doi.org/10.1016/j.atmosenv.2018.01.034.

[10] R. Pascale, G. Bianco, T.R.I. Cataldi, P.S. Kopplin, F. Bosco, L. Vignola, J. Uhl, M. Lucio, L. Milella, Mass spectrometry-based phytochemical screening for hypoglycemic activity of Fagioli di Sarconi beans (Phaseolus vulgaris L.), Food Chem. 242 (2018) 497–504.

https://doi.org/10.1016/j.foodchem.2017.09.091.

[11] G. Bianco, R. Pascale, C.F. Carbone, M.A. Acquavia, T.R.I. Cataldi, P. Schmitt-Kopplin, A. Buchicchio, D. Russo, L. Milella, Determination of soyasaponins in Fagioli di Sarconi beans (Phaseolus vulgaris L.) by LC-ESI-FTICR-MS and evaluation of their hypoglycemic activity., Anal. Bioanal. Chem. 410 (2018) 1561–1569. https://doi.org/10.1007/s00216-017-0806-8.

[12] G. Bianco, A. Buchicchio, T.R.I. Cataldi, Structural characterization of major soyasaponins in traditional cultivars of Fagioli di Sarconi beans investigated by high-resolution tandem mass spectrometry., Anal. Bioanal. Chem. 407 (2015) 6381–9. https://doi.org/10.1007/s00216-015-8810-3.

[13] G. Bianco, F.G. Battista, A. Buchicchio, C.G. Amarena, P. Schmitt-Kopplin, A. Guerrieri, Structural characterization of arginine vasopressin and lysine vasopressin by fourier-transform ion cyclotron resonance mass spectrometry and infrared multiphoton dissociation, Eur. J. Mass Spectrom. 21 (2015) 211–219. https://doi.org/10.1255/ejms.1339.

[14] G. Bianco, F. Lelario, F.G. Battista, S.A. Bufo, T.R.I. Cataldi, Identification of glucosinolates in capers by LC-ESI-hybrid linear ion trap with Fourier transform ion cyclotron resonance mass spectrometry (LC-ESI-LTQ-FTICR MS) and infrared multiphoton dissociation., J. Mass Spectrom. 47 (2012) 1160–9. https://doi.org/10.1002/jms.2996.

[15] M. De Bonis, G. Bianco, M. Amati, S. Belviso, T.R.I. Cataldi, F. Lelj, An interplay between infrared multiphoton dissociation Fourier-transform ion cyclotron resonance mass spectrometry and density functional theory computations in the characterization of a tripodal quinolin-8-olate Gd(III) complex., J. Am. Soc. Mass Spectrom. 24 (2013) 589–601. https://doi.org/10.1007/s13361-012-0570-0.

[16] F. Lelario, G. Bianco, S.A. Bufo, T.R.I. Cataldi, Establishing the occurrence of major and minor glucosinolates in Brassicaceae by LC-ESI-hybrid linear ion-trap and Fourier-transform ion cyclotron resonance mass spectrometry., Phytochemistry. 73 (2012) 74–83. https://doi.org/10.1016/j.phytochem.2011.09.010.

[17] J. Gubitosa, V. Rizzi, P. Fini, A. Laurenzana, G. Fibbi, C. Veiga-Villauriz, F. Fanelli, F. Fracassi, A. Onzo, G. Bianco, C. Gaeta, A. Guerrieri, P. Cosma, Biomolecules from Snail mucus (Helix aspersa) conjugate Gold nanoparticles, exhibiting potential wound healing and anti-inflammatory activity, Soft Matter. (2020). https://doi.org/10.1039/D0SM01638A.

[18] T. Kind, O. Fiehn, Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry, BMC Bioinformatics. 8 (2007) 105. https://doi.org/10.1186/1471-2105-8-105.

[19] N. Stoll, E. Schmidt, K. Thurow, Isotope Pattern Evaluation for the Reduction of Elemental Compositions Assigned to High-Resolution Mass Spectral Data from Electrospray Ionization Fourier

Transform Ion Cyclotron Resonance Mass Spectrometry, J. Am. Soc. Mass Spectrom. 17 (2006) 1692–1699. https://doi.org/10.1016/j.jasms.2006.07.022.

[20] J.L. Little, A.J. Williams, A. Pshenichnov, V. Tkachenko, Identification of "known unknowns" utilizing accurate mass data and chemspider, J. Am. Soc. Mass Spectrom. 23 (2012) 179–185. https://doi.org/10.1007/s13361-011-0265-y.

[21] F. Moritz, M. Kaling, J.-P. Schnitzler, P. Schmitt-Kopplin, Characterization of poplar metabotypes via mass difference enrichment analysis, Plant. Cell Environ. 40 (2017) 1057–1073. https://doi.org/10.1111/pce.12878.

[22] K. Qian, R.P. Rodgers, C.L. Hendrickson, C.A. Hughey, A.G. Marshall, Kendrick Mass Defect Spectrum: A Compact Visual Analysis for Ultrahigh-Resolution Broadband Mass Spectra, Anal. Chem. 73 (2002) 4676–4681. https://doi.org/10.1021/ac010560w.

[23] D. Tziotis, N. Hertkorn, P. Schmitt-Kopplin, Kendrick-Analogous Network Visualisation of Ion Cyclotron Resonance Fourier Transform Mass Spectra: Improved Options for the Assignment of Elemental Compositions and the Classification of Organic Molecular Complexity, Eur. J. Mass Spectrom. 17 (2011) 415–421. https://doi.org/10.1255/ejms.1135.

[24] C.S. Hsu, K. Qlan, Y.C. Chen, An innovative approach to data analysis in hydrocarbon characterization by on-line liquid, Anal. Chim. Acta. 264 (1992) 79–89.

[25] K. Longnecker, E.B. Kujawinski, Using network analysis to discern compositional patterns in ultrahigh-resolution mass spectrometry data of dissolved organic matter, Rapid Commun. Mass Spectrom. (2016) 2388–2394. https://doi.org/10.1002/rcm.7719.

[26] A.G. Marshall, R.P. Rodgers, Petroleomics: The Next Grand Challenge for Chemical Analysis, Acc. Chem. Res. 37 (2004) 53–59. https://doi.org/10.1021/ar020177t.

[27] A.C. Stenson, A.G. Marshall, W.T. Cooper, Exact masses and chemical formulas of individual Suwannee River fulvic acids from ultrahigh resolution electrospray ionization Fourier transform ion cyclotron resonance mass spectra, Anal. Chem. 75 (2003) 1275–1284. https://doi.org/10.1021/ac026106p.

[28] E.N. Nikolaev, Y.I. Kostyukevich, G.N. Vladimirov, Fourier transform ion cyclotron resonance (FT ICR) mass spectrometry: Theory and simulations, Mass Spectrom. Rev. 35 (2016) 219–258. https://doi.org/10.1002/mas.21422.

[29] A. Santarsiero, A. Onzo, R. Pascale, M.A. Acquavia, M. Coviello, P. Convertini, S. Todisco, M. Marsico, C. Pifano, P. Iannece, C. Gaeta, S. D'Angelo, M.C. Padula, G. Bianco, V. Infantino, G. Martelli, Pistacia lentiscus Hydrosol: Untargeted Metabolomic Analysis and Anti-Inflammatory Activity Mediated by NF-κB and the Citrate Pathway, Oxid. Med. Cell. Longev. 2020 (2020) 1–14.

https://doi.org/10.1155/2020/4264815.

[30] Y. Qi, P.B. O'Connor, Data processing in Fourier transform ion cyclotron resonance mass spectrometry, Mass Spectrom. Rev. 33 (2014) 333–352. https://doi.org/10.1002/mas.21414.

[31] W. Chang, J. Cheng, J. Allaire, Y. Xie, J. McPherson, shiny: Web Application Framework for R, (2020). https://cran.r-project.org/package=shiny (accessed December 2, 2020).

[32] S.B. Wu, R.S. Meyer, B.D. Whitaker, A. Litt, E.J. Kennelly, A new liquid chromatography-mass spectrometry-based strategy to integrate chemistry, morphology, and evolution of eggplant (Solanum) species, J. Chromatogr. A. 1314 (2013) 154–172. https://doi.org/10.1016/j.chroma.2013.09.017.

[33] D.P.A. Kilgour, R. Wills, Y. Qi, P.B. O'Connor, Autophaser: An algorithm for automated generation of absorption mode spectra for FT-ICR MS, Anal. Chem. 85 (2013) 3903–3911. https://doi.org/10.1021/ac303289c.

[34] D.P.A. Kilgour, M.J. Neal, A.J. Soulby, P.B. O'Connor, Improved optimization of the Fourier transform ion cyclotron resonance mass spectrometry phase correction function using a genetic algorithm, Rapid Commun. Mass Spectrom. 27 (2013) 1977–1982. https://doi.org/10.1002/rcm.6658.

[35] E. V. Kunenkov, A.S. Kononikhin, I. V. Perminova, N. Hertkorn, A. Gaspar, P. Schmitt-Kopplin, I.A. Popov, A. V. Garmash, E.N. Nikolaev, Total mass difference statistics algorithm: A new approach to identification of high-mass building blocks in electrospray ionization fourier transform ion cyclotron mass spectrometry data of natural organic matter, Anal. Chem. 81 (2009) 10106–10115. https://doi.org/10.1021/ac901476u.

[36] V. Perrier, F. Meyer, D. Granjon, shinyWidgets: Custom Inputs Widgets for Shiny, (2020). https://cran.r-project.org/package=shinyWidgets.

[37] B.V. Almende, T. Benoit, R. Titouan, visNetwork: Network Visualization using "vis.js" Library, (2019). https://cran.r-project.org/package=visNetwork.

[38] C. Sievert, Interactive Web-Based Data Visualization with R, plotly, and shiny, (2020). https://plot.ly.

[39] C. Roullier-Gall, M. Witting, R.D. Gougeon, P. Schmitt-Kopplin, High precision mass measurements for wine metabolomics, Front. Chem. 2 (2014) 102. https://doi.org/10.3389/fchem.2014.00102.

[40] R.D. Gougeon, M. Lucio, L. Boutegrabet, D. Peyron, F. Feuillat, D. Chassagne, H. Alexandre, A. Voilley, P. Cayot, I. Gebefügi, N. Hertkorn, P. Schmitt-Kopplin, Authentication approach of the chemodiversity of grape and wine by FTICR-MS, in: ACS Symp. Ser., American Chemical Society, 2011: pp. 69–88. https://doi.org/10.1021/bk-2011-1081.ch005.

[41] N. Kuhnert, F. Dairpoosh, G. Yassin, A. Golon, R. Jaiswal, What is under the hump? Mass spectrometry based analysis of complex mixtures in processed food-lessons from the characterisation of black tea

thearubigins, coffee melanoidines and caramel, Food Funct. 4 (2013) 1130–1147. https://doi.org/10.1039/c3fo30385c.

[42] F.G.B. Los, A.A.F. Zielinski, J.P. Wojeicchowski, A. Nogueira, I.M. Demiate, Beans (Phaseolus vulgaris L.): whole seeds with complex chemical composition, Curr. Opin. Food Sci. 19 (2018) 63–71. https://doi.org/10.1016/j.cofs.2018.01.010.

[43] N. Gürbüz, S. Uluişik, A. Frary, A. Frary, S. Doğanlar, Health benefits and bioactive compounds of eggplant, Food Chem. 268 (2018) 602–610. https://doi.org/10.1016/j.foodchem.2018.06.093.

5.7. Figures

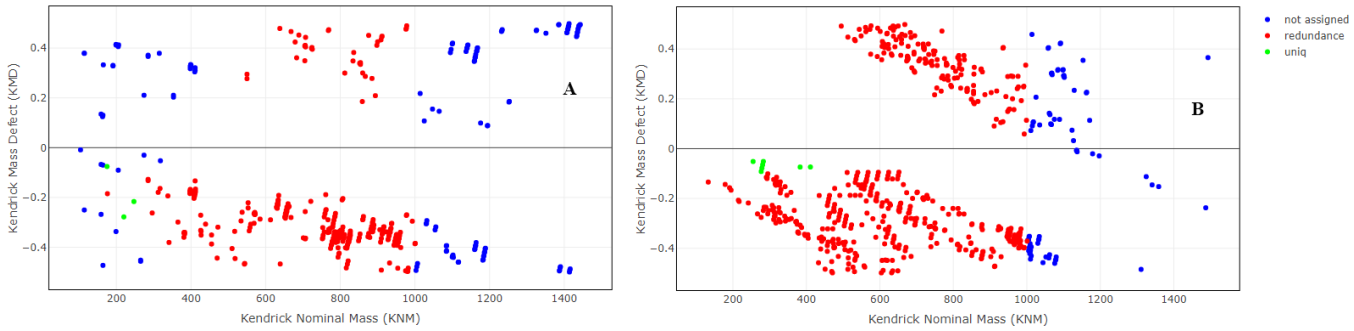

Figure 3 *Kendrick plots obtained from ESI-FT-ICR MS data of Fagioli Bianchi di Rotonda PDO beans (A) and Melanzane Rosse di Rotonda PDO eggplants (B). Unique assignments, redundances and not assigned peaks are distinguished by colour.*

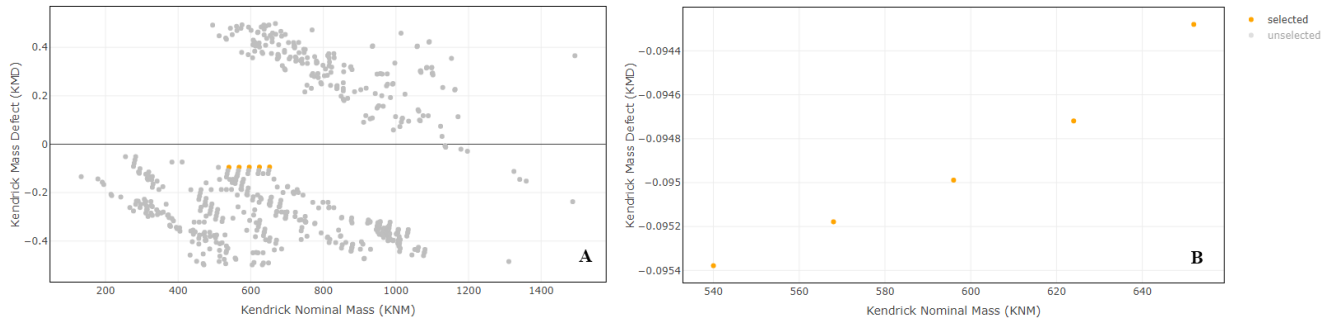

Figure 2 *Homologous series highlighting example in plot A. In plot B, selected homologous series is isolated by clicking on the unselected label in the legenda, making related points to disappear.*

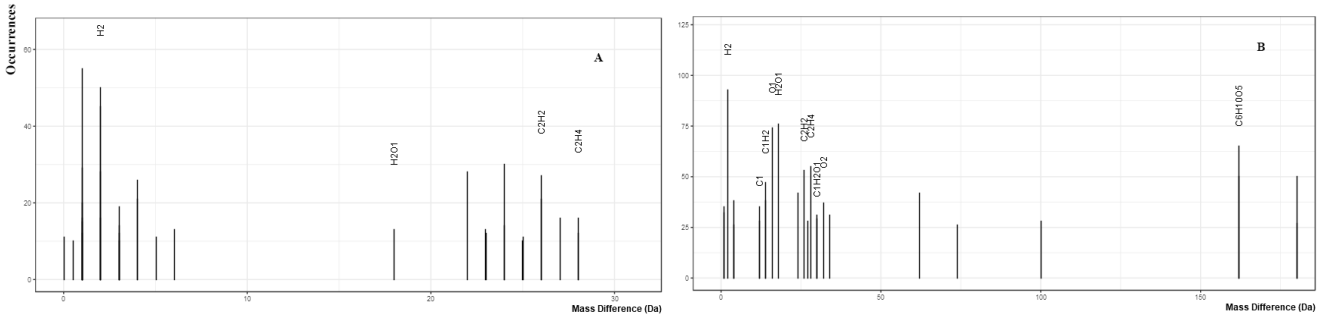

Figure 3 *Mass difference occurrence plots related to Fagioli Bianchi di Rotonda PDO beans (A) and Melanzane Rosse di Rotonda PDO eggplants (B), respectively. Mass differences successfully associated to building blocks present into the uploaded database are indicated by labels.*

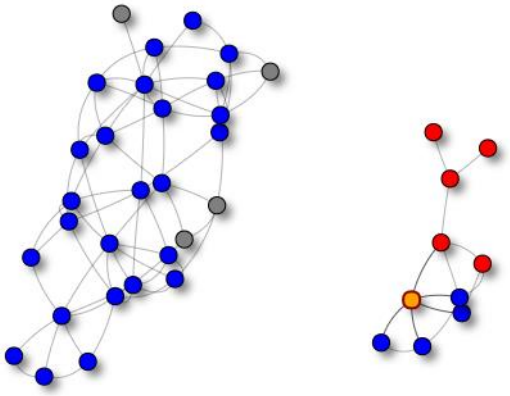

Figure 4 *Example of clusters belonging to Melanzane Rosse di Rotonda PDO eggplant Molecular Network. Nodes are distinguished by color according to the number of related formula candidates (blue for redundancies, red for unique assignments and grey for unassigned peaks).*

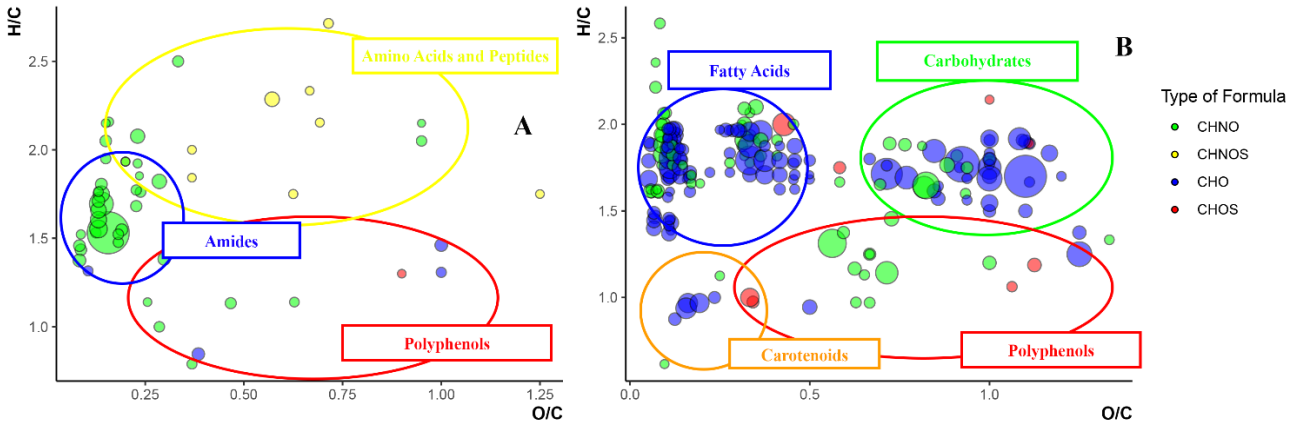

Figure 5 *Van Krevelen plots of Fagioli Bianchi di Rotonda PDO beans (A) and Melanzane Rosse di Rotonda PDO eggplants (B), respectively. Formula types are distinguished by color (blue for CHO, green for CHNO, red for CHOS and yellow for CHNOS).*

5.8. Tables

Table 1. *Homologous series data for m/z 539.50704, obtained by identification of homologous series related to different building blocks the selected peak belongs to. Since the initial number of formula candidates for selected peak is 4, it's possible to deduce that in every identified homologous series, the selected peak is the lowest redundance level member.*

Building Block	Members	KMD standard deviation	Minimum redundance level
CH_2	4	0.00028080	4
C_2H_4	4	0.00028080	4
C_4H_6	3	0.00049930	4
C_2H_2	3	0.00017078	4
$H_{-1}N_{-1}O$	3	0.00029740	4
H_2	2	0.00050516	4
C_1	2	0.00005657	4
C_3H_5NO	2	0.00033499	4
$C_2H_5NO_{-1}$	2	0.00010923	4
C_9H_9NO	2	0.00021062	4

5.9. Supplementary Material

Omics Interactive Formula Assignment R Shiny app source code

Here the OIFA source code is reported. It comprises a series of custom functions that the app employs to accomplish its tasks. Custom functions, together with the *ui* and *server* objects, should be loaded into the R software Global Environment before running the R Shiny app with the command > `shinyApp(ui = ui, server = server)`.

```
#Loading needed packages

require(shiny)
require(plotly)
require(visNetwork)
require(shinyWidgets)
require(dplyr)
require(ggplot2)
require(foreach)
require(doParallel)
require(visNetwork)
require(stringr)
require(purrr)
require(rcdk)
require(InterpretMSSpectrum)
require(numbers)
#They should be installed to use the app

############Custom Functions############

#Functions to load into the Global Environment since they are used by the
#app.

#The function calculates Kendrick properties of the input peak list,
#which is supposed to be comprised by two columns, i.e. mz and Intensity.
#It's possible to change the unit base fraction by specifying it
#into the baseunit_fraction argument.

kendrick_properties_calculator<-function(building_block = "C1H2", df,
                                         bb_database,
                                         baseunit_fraction = 1){
  db<-bb_database

  building_block_choosen<-db[(
    db$Formula == building_block
  ),]
```

```
  building_block_choosen<-building_block_choosen[,1]

  building_block_choosen<-c((round(building_block_choosen,
                                   digits = 0)/baseunit_fraction),
                            (building_block_choosen/baseunit_fraction))

  current_data<-df
  current_data$KM<-NA
  current_data$NM<-NA
  current_data$KNM<-NA
  current_data$KMD<-NA
  current_data$zstar<-NA

  bbratio<-as.numeric(building_block_choosen[1])/as.numeric(
    building_block_choosen[2])

  current_data<-within(current_data, KM <- current_data[,1]*(bbratio))

  current_data$NM<-round(current_data[,1])

  current_data$KNM<-round(current_data$KM)

  current_data<-within(current_data,
                       zstar <- numbers::mod(NM,
                                             as.numeric(
                                               building_block_choosen[1]))
                       - as.numeric(building_block_choosen[1]))

  current_data<-within(current_data, KMD <- (KM - KNM))

  #KMD is defined as the difference between the mass  in Kendrick scale
  #minus the related nominal mass (KNM)

  return(current_data)
}

#The function allows to separate peaks into different nominal mass families
#according to peak Kendrick properties.
#The function returns a list, in which every object is related to a nominal
#mass family

zstar_separation<-function(starting_data){
  cl<-starting_data
  cl$zstar<-cl$zstar %>% as.factor()

  output<-lapply(c(1:length(levels(cl$zstar))), function(i){
    current_zstar<-levels(cl$zstar)[i]
    subset(cl, zstar == current_zstar)
  })
```

```
  #For every value of the vector levels(cl$zstar), a data.frame in which
  #zstar == current_zstar is returned

  return(output)

}


#The function works on every zstar family and divides related peaks into
#different homologous series

kmd_separation<-function(zstar_set = NULL, masslist = NULL,
                         deltaKMD, bb_mass_integer = 14){
  zlist<-zstar_set
  df<-masslist

  if(is.null(zlist)){
    zlist<-zstar_set(df)
  }

  bb_int<-bb_mass_integer

  totalkmdlist<-map(zlist, .f = function(x){ #map applica la funzione .f

    cur_mz<-x
    y<-list(NA)


    index<-0
    while(nrow(cur_mz)>0){
      #The first KMD value is taken into account. Then, other KMD values
      #within a set range +-deltaKMD are selected to isolate
      #an homologous series. Once collected, the homologous series
      #is substracted from the starting zstar dataset.
      #The loop ends when there are no other values to collect.

      cur_kmd_value<-cur_mz$KMD[1]

      cur_homoseries<-cur_mz[(
        cur_mz$KMD<(cur_kmd_value+deltaKMD) &
          cur_mz$KMD>(cur_kmd_value-deltaKMD)
      ),]

      cur_mz<-cur_mz[!(
        cur_mz[,1] %in% cur_homoseries[,1]
      ),]

      if(nrow(cur_homoseries)>1){
        cur_homoseries<-q_calculator(chem_data = cur_homoseries,
                                     bbinteger = bb_mass_integer)
```

```
      } else{
        cur_homoseries$q<-NA
      }

      index<-index+1

      y[[index]]<-cur_homoseries
    }

    y

  })

  return(totalkmdlist)

}

#The function allows to delete or keep homologous series in which there's at
#least one member showing a q value equal to an integer, specified into the
#q_value argument. If discard_logical == T, than homologous series that
#satisfy this requisite are discarded.

kmd_q_filtering<-function(kmd_set, q_value = 1,
                          discard_logical = F){

  kmd<-kmd_set

  if(discard_logical==F){

    output<-keep(kmd, function(w){
      sum(w$q == q_value, na.rm = T) > 0
    })

  } else{

    output<-discard(kmd, function(w){
      sum(w$q == q_value, na.rm = T) > 0
    })

  }

  return(output)

}

#The function allows to discard peaks which don't belong to any homologous
#series. The function is used for automatic formula calculation
#through Kendrick Mass Defect analysis or to simplify the Kendrick plot.
```

```
kendrick_noise_filter<-function(masslist = NULL, kmdset = NULL,
                                bb_database = NULL, deltaKMD = NULL,
                                chosen_bb = "C1H2"){

  kmdlist<-kmdset
  db<-bb_database
  df<-masslist


  if(is.null(kmdlist)){

    df<-kendrick_properties_calculator(building_block = chosen_bb,
                                       bb_database = db, df = df)
    zset<-zstar_separation(df)
    kmdlist<-kmd_separation(zstar_set = zset, deltaKMD = deltaKMD,
                            bb_mass_integer = round(db[(
                              db$Formula==chosen_bb
                            ),1]))

  }

  output<-unlist(kmdlist, recursive = F)

  output<-keep(output, function(i){
    (nrow(i)>1 & sum(i$q>0, na.rm = T)>0)
  })

  return(output)

}

#The funcion calculates the q value for every member of the homologous series,
#i.e. the difference between the NM of one member and the NM of the previous
#one, divided by the nominal mass of the chosen building block.
#Practically, it's the number of building block units which separates
#the two members

q_calculator<-function(chem_data, bbinteger=14){
  df<-chem_data
  df$q<-NA

  if(nrow(df)>1){
    for (o in seq(2, nrow(df), 1)){

      df$q[o]<-((df$NM[o]-df$NM[o-1])/bbinteger)

    }
  }

  return(df)
```

```
}

#The function filtrates candidate formulas of homologous series members
#starting from the lowest redundance level member.
#The function needs a building block database to work, which in turn
#comprises building block exact mass, atom counts and formula.
#The function applies only on homologous series in which members show a
different
#number of candidate formulas.


kendrickfilter<-function(homoseriesdf, bb_database,
                         bb_formula = "C1H2",
                         max_q_value){

  db<-bb_database
  bb<-bb_formula
  selected_bb<-db[(db$Formula == bb), purrr::map_lgl(db, is.numeric)]
  #Only numeric variables of building block database are retained.

  df<-homoseriesdf

  df_smallest_redun<-df[,1] %>% table() %>% as.data.frame()
  #It's supposed that the first column of df is related to m/z ratios

  df_smallest_redun[,1]<-df_smallest_redun[,1] %>% as.character() %>%
    as.numeric()

  if(nrow(df_smallest_redun)>1){

    z<-names(selected_bb)

    selected_bb<-lapply(c((max_q_value*(-1)):max_q_value),
                        function(w){
                          selected_bb*w
                        })

    selected_bb<-do.call(rbind, selected_bb)
    #selected_bb contains data of the same building block, multiplied by a
    #different w factor. selected_bb is used for other members
    #formula calculation.

    if(min(df_smallest_redun$Freq)!=max(df_smallest_redun$Freq)){
      df_smallest_redun<-df[(
        df[,1] %in% (subset(df_smallest_redun, Freq == min(Freq))[,1])
      ),]
    }else{
      df_smallest_redun<-df[(
        df[,1] == df_smallest_redun[1,1]
      ),]
```

```
}



df2<-df_smallest_redun[, z]#Assure that the order of variables
#is the same of the one of the building block database.
#It's important that the name of the variable related to the exact mass
#is the same for building block database and peak list.

assignments<-lapply(c(1:ncol(df2)), function(i){
  y<-lapply(c(1:nrow(selected_bb)), function(j){
    df2[,i]+selected_bb[j,i]
  })
  unlist(y)
})

assignments<-do.call(cbind, assignments)


assignments<-assignments %>% as.data.frame()

names(assignments)<-names(df2)

assignments<-create_Formula(assignments)

output<-df[(
  df$Formula %in% assignments$Formula
),]

#The following chunk is necessary to control if no m/z value
#was deleted after filtration process. This can happen if
#no other matching has been observed.

k<-df[,1] %>% table() %>% as.data.frame()
y<-output[,1] %>% table() %>% as.data.frame()

if(nrow(k) == nrow(y)){

  return(output)

} else{

  #If an information loss is observed,
  #the starting dataset is returned

  return(df)

}
```

```
  }else{
    return(df)
  }

}

#The function performs an automatic formula filtering through KMD analysis
#by working on every homologous series.

kmd_analysis<-function(chemlist, bb_database, deltaKMD = 0.001,
                       chosen_bb = "C1H2"){

  df<-chemlist
  db<-bb_database
  mz_df<-chemlist[,1:2]%>%table()%>%as.data.frame()
  mz_df<-mz_df[(
    mz_df$Freq>0
  ),1:2]
  mz_df[,1]<-mz_df[,1]%>%as.character()%>%as.numeric()
  mz_df[,2]<-mz_df[,2]%>%as.character()%>%as.numeric()
  mz_df<-mz_df[order(mz_df[,1]),]
  #assuming that m/z ratios are into the first column and the intensity values
  #into the second one

  mz_df<-kendrick_properties_calculator(building_block = chosen_bb,
                                        df = mz_df, bb_database = db
  )
  zset<-zstar_separation(mz_df)
  kmdset<-kmd_separation(zstar_set = zset, masslist = mz_df,
                         deltaKMD = deltaKMD, bb_mass_integer = round(
                           db[(db$Formula==chosen_bb), 1]
                         ))
  kmdset2<-kendrick_noise_filter(kmdset = kmdset)

  kmdset2<-kmd_q_filtering(kmd_set = kmdset2, q_value = 0, discard_logical = T)
  #Discard overlapping homologous series.

  finaldata<-map(kmdset2, function(x){

    cur_q_value<-max(x$q, na.rm = T)*10
    chem<-df[(
      df[,1]%in%x[,1]
    ),]
    assigned<-chem[(is.na(chem$Formula)==F),]
    not_assigned<-chem[(is.na(chem$Formula)==T),]

    if(nrow(assigned)>0){
      output<-kendrickfilter(homoseriesdf = assigned,
```

```
                              bb_database = db,
                              bb_formula = chosen_bb,
                              max_q_value = cur_q_value)
      if(nrow(not_assigned)>0){
        output<-rbind(output, not_assigned)

        output

      } else{
        output
      }
    } else{
      not_assigned
    }

  })

  finaldata<-do.call(rbind, finaldata)

  finaldata<-rbind(finaldata, df[!(df[,1]%in%finaldata[,1]),])

  finaldata<-finaldata[order(finaldata[,1]),]

  return(finaldata)

}

#The function performs an automatic formula calculation process, acting in the
#same way of the kmd_analysis() function, but starting calculation from
#lowest redundance level member of homologous series

kmd_form_calc<-function(chemlist, bb_database, deltaKMD = 0.001,
                        chosen_bb = "C1H2", ppm = 5){

  df<-chemlist
  db<-bb_database
  selected_bb<-db[(db$Formula==chosen_bb),map_lgl(db, is.numeric)]

  mz_df<-chemlist[,1:2]%>%table()%>%as.data.frame()
  mz_df<-mz_df[(
    mz_df$Freq>0
  ),1:2]
  mz_df[,1]<-mz_df[,1]%>%as.character()%>%as.numeric()
  mz_df[,2]<-mz_df[,2]%>%as.character()%>%as.numeric()
  mz_df<-mz_df[order(mz_df[,1]),]

  mz_df<-kendrick_properties_calculator(building_block = chosen_bb,
                                        df = mz_df, bb_database = db
  )
```

```
zset<-zstar_separation(mz_df)

kmdset<-kmd_separation(zstar_set = zset, masslist = mz_df,
                       deltaKMD = deltaKMD, bb_mass_integer = round(
                         db[(db$Formula==chosen_bb), 1]
                       ))

kmdset2<-kendrick_noise_filter(kmdset = kmdset)

kmdset2<-kmd_q_filtering(kmd_set = kmdset2, q_value = 0, discard_logical = T)


finaldata<-map(kmdset2, function(x){

  cur_q_value<-max(x$q, na.rm = T)*10
  chem<-df[(
    df[,1]%in%x[,1]
  ),]
  assigned<-chem[(is.na(chem$Formula)==F),]
  not_assigned<-chem[(is.na(chem$Formula)==T),]
  cur_bb<-lapply(seq((cur_q_value*(-1)),cur_q_value, 1), function(i){
    selected_bb*i
  })
  cur_bb<-do.call(rbind, cur_bb)

  if(nrow(assigned)>0 & nrow(not_assigned)>0){

    #Calculation takes place when assigned and not assigned peaks
    #are present into the same homologous series

    redundances<-assigned[,1]%>%table()%>%as.data.frame()
    redundances[,1]<-redundances[,1]%>%as.character()%>%as.numeric()

    if(sum(redundances$Freq==min(redundances$Freq, na.rm = T))==1){

      starting_formulae<-assigned[(
        assigned[,1]%in%redundances[(redundances$Freq==min(
          redundances$Freq, na.rm = T)),1]
      ), ]
      #If there's only one lowest redundance level member...
    } else{

      #...otherwise, the lowest m/z ratio related one is selected

      smallest_redundances<-redundances[(redundances$Freq==min(
        redundances$Freq, na.rm = T)),1]
      starting_formulae<-assigned[(
        assigned[,1]%in%smallest_redundances[1,1]
      ), ]
    }
```

```
    assignments<-candidate_formulae_calculator(exp_form = starting_formulae,
                                               building_blocks = cur_bb)
    #Candidate formula calculation

    new_assignments<-assignment_script(masslist = not_assigned[,1:2],
                                       err_ppm = ppm,
                                       formula_db = assignments)

    if(nrow(new_assignments[(is.na(new_assignments$Formula)==F),])>0){

      #to continue, it's important that at least one new assignment has
      #been obtained successfully

      assigned<-rbind(assigned[, names(new_assignments)], new_assignments)

      not_assigned<-not_assigned[!(
        not_assigned[,1]%in%assigned[,1]
      ),]

      if(nrow(not_assigned)==0){
        #If every member has been assigned to a candidate elemental formula
        assigned
      } else{
        #If not, binds assigned members data to not assigned members one
        rbind(assigned, not_assigned[, names(assigned)])
      }


    } else{

      #If no new assignment has been obtained, starting homologous series
      #is returned

      chem
    }

  } else{
    chem
  }

})

finaldata<-do.call(rbind, finaldata)

finaldata<-rbind(finaldata, df[!(df[,1]%in%finaldata[,1]),])

return(finaldata)
```

```
}

#Given a preselected m/z ratio and a building block dataset, together
#with a predefined error for KMD (deltaKMD), the following function allows
#to calculate the number of members of every homologous series
#the selected m/z ratio belongs to.

homo_series_numbers<-function(df, mz_ratio, bb_db, deltaKMD = 0.001,
                              chemlist = NULL){

  mz<-df
  choosen_mz<-mz_ratio
  db<-bb_db
  chem<-chemlist #The argument is needed to know the minimum redundance
  #multiplicity

  output<-lapply(c(1:nrow(db)), function(i){

    k<-kendrick_properties_calculator(building_block = db$Formula[i],
                                      df = mz, bb_database = db)
    k$building_block<-db$Formula[i]

    k

    #Calculate Kendrick parameters for every experimental m/z ratio

  })

  output<-map(output, function(x){

    choosen_mz_row<-x[(
      as.character(x[,1]) %in% as.character(choosen_mz)
    ),]

    homologous_series<-x[(
      x$zstar == choosen_mz_row$zstar &
        x$KMD>(choosen_mz_row$KMD-deltaKMD) &
        x$KMD<(choosen_mz_row$KMD+deltaKMD)
    ),]
    homologous_series<-homologous_series[order(homologous_series[,1]),]

    ifelse(nrow(homologous_series)>1,
           yes = homologous_series$sdKMD<-sd(homologous_series$KMD, na.rm = T)
,
           no = homologous_series$sdKMD<-homologous_series$KMD
    )
    #This chunk allows the calculation of the KMD standard deviation in a
    #series, just to have an idea of the "quality" of found homologous
    #series
```

```
    if(is.null(chem)==F){

      homologous_chem<-chem[(
        as.character(chem[,1]) %in% as.character(homologous_series[,1])
      ),] #Extrapolate calculate formulas for homologous series members

      redundance<-homologous_chem[(
        is.na(homologous_chem$Formula)==F
      ),1] %>% table() %>% as.data.frame()

      #Only assigned peaks will be considered to calculate minimum redundance
      #value

      redundance[,1]<-redundance[,1] %>% as.character() %>% as.numeric()
      redundance[,2]<-redundance[,2] %>% as.character() %>% as.numeric()
      #From factor to numeric

      data.frame(Formula = x$building_block[1],
                 Members = nrow(homologous_series),
                 sdKMD = homologous_series$sdKMD[1],
                 Minimum_Redun = min(redundance[,2], na.rm = T))

    } else{

      data.frame(Formula = x$building_block[1],
                 Members = nrow(homologous_series),
                 sdKMD = homologous_series$sdKMD[1])

    }


  })

  output<-do.call(rbind, output)

  output<-output[(output$Members<50 &
                    output$Members>1),] #Longer series are not feasible
  #Be aware! At higher deltaKMD, longer series are obtained, but it's obvious
  #members don't really belong to these (too high error).
  #Furthermore, only 2 or higher member series are of interest.

  output<-output[order(output$Members, decreasing = T),]

  return(output)


}

#The function allows to assign candidate formulas (formula_db argument)
```

```
#to m/z ratios given as an input data.frame (masslist argument).
#The error, the ionization mode and the application of the nitrogen rule
#can be set by the err_ppm, mode and nitrogen_rule_applier arguments.
#The function allows to perform isotopologue collection and isotopic
#pattern matching score too, if isot_pattern_filtering == T.

assignment_script<-function(masslist, err_ppm = 5,
                            formula_db, isot_pattern_filtering = F,
                            isodiff = list(C13 = c(1.003, 50),
                                           N15 = c(0.997,1),
                                           S34 = c(1.995,10)),
                            int_precision = 0.02,
                            digits = 3,
                            err_da = 0.001,
                            mode = "positive",
                            nitrogen_rule_applier = T){
  df<-masslist
  names(df)<-c("mz", "Intensity")
  z<-names(df)
  cur_db<-formula_db

  if(isot_pattern_filtering==T){
    isolist<-isot_collector(masslist = df, dmatrix = NULL,
                            daerr = err_da, dmatrix_digits = digits,
                            isot_diff_list = isodiff)
    isoset<-do.call(rbind, isolist)
    df<-df[!(
      df[,1] %in% isoset$`M+1`
    ),]
  }


  output<-lapply(c(1:nrow(df)), function(i){

    cur_mz<-df[i,1]
    cur_i<-df[i,2]
    nm<-round(cur_mz)

    candidate_formulae<-cur_db[(
      cur_db$exact.mass>(cur_mz-0.1)&
        cur_db$exact.mass<(cur_mz+0.1)
    ),]

    candidate_formulae<-within(candidate_formulae,
                               ppm<-((cur_mz-exact.mass)/exact.mass)*1e+6)
    candidate_formulae<-candidate_formulae[(
      abs(candidate_formulae$ppm)<err_ppm
    ),]

    y<-data.frame(mz = rep(cur_mz, nrow(candidate_formulae)),
```

```
                Intensity = rep(cur_i, nrow(candidate_formulae)),
                candidate_formulae)


  y


})

output<-do.call(rbind, output)


if(nrow(output)>0){
  output<-create_Formula(output)
  #Nitrogen rule applier

  if(nitrogen_rule_applier == T){

    output2<-output[(
      output[,1]>500
    ),]

    output<-output[(
      output[,1]<=500
    ),]

    #Nitrogen rule is applied only to less or equal than 500 m/z ratios
    #to increase accuracy of assignments.

    output<-anti_join(output, output[(
      (round(output[,1])%%2==0) &
        output$N%%2==0
    ),])

    output<-anti_join(output, output[(
      (round(output[,1])%%2!=0) &
        output$N%%2!=0
    ),])

    output<-rbind(output, output2)


  }



  output<-full_join(x = output, y = df[!(df[,1] %in% output[,1]),],
                    by = z)
  output<-output[order(output[,1]),]
```

```
    if(isot_pattern_filtering==T){

      output<- isotopic_pattern_filtering(chemlist = output,
                                          isolist = isolist,
                                          mode = mode)

    } else{
      output$mSigma<-NA
    }

    return(output)
  } else{

    df$Formula<-NA
    return(df)

  }



}

#The function identifies and collects isotopologues.
#isot_diff, err_da and rel_int_thresh arguments are related to
#isotopologue difference, difference error and relative intensity threshold,
#respectively, considered to distinguish monoisotopic peaks from isotopologues
#and depend from the element and related isotope considered.

isotopologue_identifier<-function(masslist, dmatrix = NULL,
                                  isot_diff = 1.003,
                                  err_da = 0.001,
                                  rel_int_thresh = 50,
                                  dmatrix_digits = 3){
  mass<-masslist
  ddf<-dmatrix

  if(is.null(ddf)){
    ddf<-dmatrixcalculator(masslist1 = mass,
                           number_of_digits = dmatrix_digits)
  }

  isotopologues<-which((ddf <(isot_diff+err_da) &
                          ddf >(isot_diff-err_da)),
                       arr.ind = T) %>% as.data.frame()

  if(nrow(isotopologues)>0){

    isotopologues<-isotopologues[,c(2,1)]
```

```
    isotopologues$row<-mass[isotopologues$row, 1]
    isotopologues$col<-mass[isotopologues$col, 1]

    isotopologues<-lapply(c(1:nrow(isotopologues)), function(i){
      currow<-isotopologues[i,]
      currow[3]<-mass[(mass[,1] %in% currow[1]),2]
      currow[4]<-mass[(mass[,1] %in% currow[2]),2]
      currow
    })
    isotopologues<-do.call(rbind, isotopologues)

    names(isotopologues)<-c("M", "M+1", "Intensity", "I+1")

    isotopologues<-mutate(isotopologues,
                          isotopologues$`I+1`<-(isotopologues$`I+1`/
                                                  isotopologues$Intensity)*100)

    isotopologues<-isotopologues[(
      isotopologues[,5]<=rel_int_thresh
    ),c(1:4)]

    return(isotopologues)
  }

}

#The function returns a list in which different families of isotopologues
#(one per considered element) are collected in different objects.
#It's possible to specify elements, mass differences and intensity thresholds
#through the isot_diff_list argument.

isot_collector<-function(masslist, dmatrix = NULL,
                         isot_diff_list = list(C13 = c(1.003, 50),
                                               N15 = c(0.997,1),
                                               S34 = c(1.995,10)),
                         daerr = 0.001,
                         dmatrix_digits = 3){
  mass<-masslist
  ddf<-dmatrix

  if(is.null(ddf)){
    ddf<-dmatrixcalculator(masslist1 = mass,
                           number_of_digits = dmatrix_digits)
  }

  output<-lapply(c(1:length(isot_diff_list)), function(i){

    isotopologue_identifier(masslist = mass, dmatrix = ddf,
                            isot_diff = isot_diff_list[[i]][1],
```

```
                              err_da = daerr,
                              rel_int_thresh = isot_diff_list[[i]][2],
                              dmatrix_digits = dmatrix_digits)

  })

  names(output)<-names(isot_diff_list)

  return(output)


}

#The function works on a chemlist and compare theoretical isotopic patterns to
#the observed ones. It's assumed that isotopologues are not present into the
#chemlist and are collected into a separated list (isot_collector() output)

isotopic_pattern_filtering<-function(chemlist, isolist, mode = "positive",
                                     err_ppm = 5, rel_int_precision = 10){

  output<-chemlist

  output$mScore<-NA

  isoset<-do.call(rbind, isolist)

  assign_iso<-output[(output[,1] %in% isoset$M),]
  assign_iso<-assign_iso[(is.na(assign_iso$Formula)==F),]
  mass_e<-0

  if(mode == "positive"){
    mass_e<-chemdict[c("e"),1]*(-1)
  }
  if(mode == "negative"){
    mass_e<-chemdict[c("e"),1]*(+1)
  }

  assign_iso2<-lapply(c(1:nrow(assign_iso)), function(x){

    cur_row<-assign_iso[x,]


    obs_iso<-isofinder(exp_mz = assign_iso[x,1], isolist = isolist) %>%
      as.data.frame()

    int_precision<-(min(obs_iso[,2], na.rm = T)*rel_int_precision)/100

    min_int_value<-(min(obs_iso[,2], na.rm = T)-int_precision)
```

```
    the_iso<-get.isotopes.pattern(get.formula(
      mf = as.character(assign_iso$Formula[x])),
      minAbund = min_int_value) %>% as.data.frame()
    the_iso[,1]<-the_iso[,1]+mass_e

    ddf<-dmatrixcalculator(masslist1 = the_iso,
                           masslist2 = obs_iso,
                           number_of_digits = 5,
                           zero_and_neg_to_na = F)
    ddf<-(ddf/t(obs_iso)[1,])*1e6 #Convert abs errors in ppm

    #Create a difference matrix in ppm to select
    #isotopologues nearest to the observed pattern
    #in terms of ppm


    k<-which(abs(ddf)<err_ppm, arr.ind = T)

    the_iso<-the_iso[k[,1],] %>% as.data.frame()

    if(nrow(obs_iso)==nrow(the_iso)){

      cur_row$mScore<-mScore(obs = t(obs_iso), the = t(the_iso),
                             dabs = 0, dppm = err_ppm,
                             int_prec = int_precision)
      cur_row

    } else{

      cur_row$mScore<-0
      cur_row

    }

  })

  assign_iso2<-do.call(rbind, assign_iso2)

  output<-rbind(assign_iso2, output[!(output[,1] %in% assign_iso2[,1]),])
  output<-output[order(output[,1]),]

  return(output)


}

#The function allows filter the input peak list deleting all the isotopologues.
#The function works as isot_collector(), which is effectively used
#to obtain a data.frame of isotopologues, useful for their elimination
```

```r
#from input peak list.

isot_filter<-function(masslist, isot_list = NULL,
                      dmatrix = NULL,
                      isot_diff_list = list(C13 = c(1.003, 50),
                                            N15 = c(0.997,1),
                                            S34 = c(1.995,10)),
                      daerr = 0.001,
                      dmatrix_digits = 3){

  isotlist<-isot_list
  df<-masslist
  ddf<-dmatrix

  if(is.null(ddf)){
    ddf<-dmatrixcalculator(masslist1 = df, number_of_digits = dmatrix_digits)
  }

  isodifflist<-isot_diff_list


  if(is.null(isotlist)){

    isotlist<-isot_collector(masslist = df, dmatrix = ddf,
                             isot_diff_list = isodifflist,
                             daerr = daerr,
                             dmatrix_digits = dmatrix_digits)

  }

  isoset<-do.call(rbind, isotlist)

  output<-df[!(
    df[,1] %in% isoset$'M+1'
  ),]

  return(output)

}

#The function allows to collect isotopologues of a specific experimental
#m/z ratio from an isotopologues set, created by using the function
#isot_collector()

isofinder<-function(exp_mz, isolist){


  final<-lapply(c(1:length(isolist)), function(x){
```

```
  isoset<-isolist[[x]]

  if(!is.null(isoset)){

    if(nrow(isoset)>0){

      output<-isoset[(isoset$M==exp_mz), c("M", "Intensity")]

      if(nrow(output)>0){

        repeat{

          output2<-isoset[(isoset$'M'==output[nrow(output), 1]),
                          c("M+1", "I+1")]

          if(nrow(output2)>0){

            names(output2)<-names(output)
            output<-rbind(output, output2)

          } else{
            break
          }

        }

        if(x==1){

          output

        } else{

          output[2:nrow(output),]

        }

      }

    }

  }



})

final<-do.call(rbind, final)

final<-mutate(final, Intensity=(Intensity/max(Intensity)))
```

```
  return(final)

}

#The function calculates a mass difference matrix (dmatrix or ddf), in which
#every column and
#row correspond to a m/z ratio. Thus, every matrix cell contains a difference
#value among two m/z ratios. Through the number_of_digits argument, it's
#possible to round mass difference values to a specified number of digits.
#If zero_and_neg_to_na == T, zero and negative values are converted to NA.


dmatrixcalculator<-function(masslist1, masslist2 = NULL,
                            number_of_digits = 3,
                            zero_and_neg_to_na = T){

  if(is.null(masslist2)){
    masslist2<-masslist1
  }

  dmatrix<-as.data.frame(matrix(data = NA, nrow = nrow(masslist1),
                                ncol = nrow(masslist2)))

  rownames(dmatrix)<-as.character(masslist1[,1])
  colnames(dmatrix)<-as.character(masslist2[,1])

  for (k in seq(1, nrow(masslist2), 1)){
    dmatrix[,k]<-masslist1[,1] - as.numeric(masslist2[,1][k])
  }
  #Columns of the dmatrix are equal to the difference among the entire peak
list
  #and a single peak value

  neg_values_positions<-which(dmatrix < 0, arr.ind = T)

  zero_positions<-which(dmatrix == 0, arr.ind = T)
  dmatrix<-round(dmatrix, number_of_digits)

  if(zero_and_neg_to_na == T){
    dmatrix[neg_values_positions]<-NA
    dmatrix[zero_positions]<-NA
  }#zero and negative value are converted in NA. This assures that these
  #values are not considered by functions which take the dmatrix
  #as an input file


  return(dmatrix)
}
```

```
#The function returns a mass difference occurrence matrix (pmatrix or ppf),
#in which every mass difference value of the dmatrix is collected,
#together with its occurrence. p_threshold is an occurrence threshold value
#which is useful to delete mass differences which occurrence is lower
#than it. If abs_frequency==F, a probability value is returned in place of
#the absolute occurrence (see Anal. Chem. 2009, 81, 10106-10115).

pmatrixcalculator<-function(dmatrix, abs_frequency = F, p_threshold = 0.1){
  z<-as.matrix(dmatrix)
  k<-as.data.frame(table(z, useNA = "no"))
  names(k)<-c("d", "p")
  k$d<-as.character(k$d)
  k$d<-as.numeric(k$d)

  if(abs_frequency == F){
    k$p<-(k$p)/((nrow(dmatrix)^2)/2)     #(nrow(dmatrix)-1)

    k$p<-((k$p)/max(k$p, na.rm = T))*100
    #Probability values are calculated as the number of occurrences
    #on the total number of mass difference.

    k<-subset(k, p >= p_threshold)
  } else{
    k<-subset(k, p >= p_threshold)
  }

  return(k)
}

#The function returns a list in which edges data are collected for every m/z
#ratio. The output list is necessary for Molecular Network calculations.
#Indeed, every list object comprises building block data, such as
#exact mass and atomic counts, which are then added to calculate candidate
#formulas of adjacent nodes.
#The function scans a mass difference matrix, filter matrix objects
#retaining all the successfully assigned mass differences and creates
#the output list.

diff_identifier<-function(database, dmatrix = NULL, error_arg = 0.001,
                          pmatrix = NULL, p_threshold = 50){

  db<-database
  err<-error_arg
  ddf<-dmatrix
  df<-data.frame(Mass = (colnames(ddf)%>%as.numeric()))
  ppf<-pmatrix

  if(is.null(ppf)){
    ppf<-pmatrixcalculator(dmatrix = ddf, abs_frequency = T,
                           p_threshold = p_threshold)
```

```
}

ppf<-ppf[(
  ppf$d != 0
),]

reaction_bb<-lapply(c(1:nrow(ppf)), function(i){
  cur_d<-ppf$d[i]
  cur_p<-ppf$p[i]
  cur_sign<-NA
  ifelse(cur_d>0, yes = cur_sign<-1,
         no = cur_sign<-(-1))
  cur_bb<-db[(
    (db[,1]>(abs(cur_d)-err)) &
      (db[,1]<(abs(cur_d)+err))
  ),]

  if(nrow(cur_bb)>0){
    data.frame(d = rep(cur_d, nrow(cur_bb)),
               p = rep(cur_p, nrow(cur_bb)),
               Formula = cur_bb$Formula,
               Sign = rep(cur_sign, nrow(cur_bb)))
  }

})

reaction_bb<-do.call(rbind, reaction_bb)

edges<-lapply(c(1:nrow(reaction_bb)), function(u){

  k<-which(ddf == reaction_bb$d[u], arr.ind = T)%>%
    as.data.frame()
  k$Formula<-rep(reaction_bb$Formula[u], nrow(k))
  k$Sign<-rep(reaction_bb$Sign[u], nrow(k))
  k[,2]<-df[k[,2],1]
  k[,2:ncol(k)]

})

edges<-do.call(rbind, edges)
edges<-edges[order(edges[,1]),]
y<-mass_collector(edges)

bblist<-lapply(c(1:nrow(y)), function(j){

  sub_edges<-edges[(
    edges[,1]==y[j,1]
  ),]

  output<-lapply(c(1:nrow(sub_edges)), function(u){
```

```
      db[(
        db$Formula == sub_edges$Formula[u]
      ), map_lgl(db, is.numeric)]*sub_edges$Sign[u]
    })

    do.call(rbind, output)

  })

  names(bblist)<-(y[,1]%>%as.character())

  return(bblist)

}

#The function calculates formulas adding building block data to a starting
node.
#The m/z ratio related to this is supposed to be into the first column.
#Moreover, it's supposed that exact mass column is named in the same way for
#both input datasets (starting node and building block data).
#seven_golden_rules_applier = T applies the seven golden rules to filter
#calculated formulas


candidate_formulae_calculator<-function(starting_formulae, building_blocks,
                                        seven_golden_rules_applier = T){

  assigned<-starting_formulae

  cur_bblist<-building_blocks
  cur_bblist<-cur_bblist[, map_lgl(cur_bblist, is.numeric)]
  #Only numeric vectors of cur_bblist are retained

  assigned<-assigned[, names(cur_bblist)]
  #The order of columns is the same for assigned and cur_bblist.
  #Be aware! Name of exact mass column and atom count ones must be the same
  #to avoid errors.

  y<-lapply(c(1:nrow(cur_bblist)), function(i){
    j<-lapply(c(1:ncol(assigned)), function(w){
      (assigned[,w]+cur_bblist[i,w])%>%as.data.frame()
    })
    j<-do.call(cbind, j) #Formulae are obtained by doing a columnwise sum
  })
  y<-do.call(rbind, y)

  names(y)<-names(cur_bblist)

  negative_atoms_index<-which(y<0,arr.ind = T)%>%
```

```
    as.data.frame()

  if(nrow(negative_atoms_index)>0){

    negative_atoms_index<-negative_atoms_index[,1]%>%table()%>%
      as.data.frame() #Only rows needed

    y<-y[-(negative_atoms_index[,1]%>%as.character()%>%
             as.numeric()),]
    #Delete all the formulas that are not chemically feasible,
    #i.e. negative atom count ones

  }

  if(seven_golden_rules_applier == T){

    y<-y[(
      ((y$H+y$Na+y$K)/y$C)>=0.2 &
        ((y$H+y$Na+y$K)/y$C)<=3.1 & #Atomic ratios filter, modify here or
comment to
        (y$O/y$C)<=2 & #avoid the application of this type of filter
        (y$N/y$C)<=1.3 &
        (y$P/y$C)<=0.3 &
        (y$S/y$C)<=0.8
    ),]

    y<-element_counter_rule_applier(y)

    z<-names(y)

    y<-HC_OC_DBE_DBE_o_calculator(y)

    y<-y[(
      y$DBE>=0 & y$DBE<40
    ),]


    y<-y[,z]

  }

  return(y)

}


#The function filter formulas of a Molecular Network cluster starting from a
#node/s, specified here through the argument selected_nodes. The function
#accepts a character vector containing m/z ratios related to starting nodes.
```

```
#The function, firstly, uses diff_identifier() (if the output list is not
#given in input through the argument bb_list) and uses building block
#data to calculate candidate formulas and filter redundances.
#If bb_list != NULL, dmatrix, pmatrix, p_threshold, exp_bb_err and digits
#are useless, since they are needed to calculate bb_list (diff_identifier()
#output).
#User formula data should be given to the function through chemlist argument.
#If masslist == NULL, the peak list is deduced from chemlist.
#The function is supposed to work on formula data which are organized
#just like the assignment_script() function output.

net_filtering_analysis<-function(chemlist,
                                 masslist = NULL,
                                 bb_database,
                                 selected_nodes,
                                 bb_list = NULL,
                                 dmatrix = NULL,
                                 digits = 3,
                                 pmatrix = NULL,
                                 p_threshold = 10,
                                 exp_bb_err = 0.001){

  db<-bb_database
  df<-chemlist

  if(is.null(masslist)){
    mz_df<-df[,1:2]%>%table()%>%as.data.frame()
    mz_df<-mz_df[(mz_df$Freq>0),1:2]
    mz_df[,1]<-mz_df[,1]%>%as.character()%>%as.numeric()
    mz_df[,2]<-mz_df[,2]%>%as.character()%>%as.numeric()
    mz_df<-mz_df[order(mz_df[,1]),]
  } else{
    mz_df<-masslist
  }

  not_assigned<-df[(is.na(df$Formula)==T),]

  if(nrow(not_assigned)>0){
    df<-df[!(df[,1] %in% not_assigned[,1]),]
  }

  #Difference matrix calculation. Negative values and zero are included.

  if(is.null(dmatrix) & is.null(bb_list)){
    ddf<-dmatrixcalculator(masslist1 = mz_df,
                           zero_and_neg_to_na = F,
                           number_of_digits = digits)
  } else{
    if(!is.null(dmatrix) & is.null(bb_list)){
      ddf<-dmatrix
```

```
  }
}

#Difference-occurrence matrix (Pmatrix) calculation. Zero is discarded.

if(is.null(pmatrix) & is.null(bb_list)){
  ppf<-pmatrixcalculator(dmatrix = ddf, abs_frequency = T,
                         p_threshold = p_threshold)
  ppf<-ppf[(ppf$d!=0),]


} else{
  ppf<-pmatrix
}

#Making a list in which building blocks are grouped by m/z values

if(is.null(bb_list)){

  collected_bb<-diff_identifier(database = db, dmatrix = ddf,
                                pmatrix = ppf, error_arg = exp_bb_err,
                                p_threshold = p_threshold)
} else{

  collected_bb<-bb_list

}

#Here, a loop starts, during which correct candidate formulae are
#collected. Starting from the selected node, formulae are calculated
#and compared with candidate formula starting data.frame (chemlist).
#Those which would match are collected and related nodes are considered
#to repeat the step. When no other node is collected, the loop stops
#and incorrect candidate formulae are discarded.


add_to_uniq<-df[(df[,1] %in% as.numeric(selected_nodes)),]


repeat{

  cur_bblist<-collected_bb[(names(collected_bb) %in% as.character(
    add_to_uniq[,1]))]
  #Only building blocks related to assigned m/z ratios will be considered
  #for formula calculation

  if(length(cur_bblist)==0){
    print("There are no edges for this node.")
    break
```

```
  }


  assignm<-lapply(c(1:length(cur_bblist)), function(x){

    cur_formulae<-add_to_uniq[(
      as.character(add_to_uniq[,1]) %in% (names(cur_bblist)[x])
    ), ]

    #For every building block group, formulae related to the node
    #are selected for calculation

    candidate_formulae_calculator(starting_formulae = cur_formulae,
                                  building_blocks = cur_bblist[[x]])

  })


  assignm<-do.call(rbind, assignm)

  assignm<-create_Formula(assignm)

  cur_df<-df[!(df$Formula%in%add_to_uniq$Formula),]

  selected_formulae<-cur_df[(
    cur_df$Formula %in% assignm$Formula
  ),]

  if(nrow(selected_formulae)>0){

    add_to_uniq<-rbind(add_to_uniq, selected_formulae)

  } else{

    break
  }

}

if(length(cur_bblist)==0){

  return(df)

} else{

  output<-rbind(df[!(
    df[,1] %in% add_to_uniq[,1]
  ),], add_to_uniq)
```

```
    if(nrow(not_assigned)>0){
      output<-rbind(output, not_assigned)
    }

    output<-output[order(output[,1]),]

    return(output)
  }

}

#The function returns a data.frame comprising all the edges data. The output
#is supposed to be suitable for the making of Molecular Network through the
#visNetwork() function. dmatrix, database, value_to_edges_df,
#value_frequency_thresh and exp_bb_err are arguments for the
#mass difference matrix, the building block database, the occurrence matrix
#(if NULL, it's made by the function itself, deleting all the mass differences
#which occurrence is lower than value_frequency_thresh), the occurrence
threshold
#value and the error considered for mass difference building block assignment.

make_edges_list<-function(dmatrix, database, value_to_edges_df = NULL,
                          value_frequency_thresh = 0,
                          exp_bb_err = 0.001){

  ddf<-dmatrix
  db<-database
  err<-exp_bb_err

  #Occurrence matrix calculation

  if(is.null(value_to_edges_df)){
    ppf<-pmatrixcalculator(dmatrix = ddf,
                           abs_frequency = T,
                           p_threshold = value_frequency_thresh)
  } else{
    ppf<-value_to_edges_df
  }


  reaction_bb<-lapply(c(1:nrow(ppf)), function(i){
    cur_d<-ppf$d[i]
    cur_p<-ppf$p[i]
    cur_sign<-NA
    ifelse(cur_d>0, yes = cur_sign<-1,
           no = cur_sign<-(-1))
    cur_bb<-db[(
      (db[,1]>(abs(cur_d)-err)) &
        (db[,1]<(abs(cur_d)+err))
    ),]
```

```
  if(nrow(cur_bb)>0){
    data.frame(d = rep(cur_d, nrow(cur_bb)),
               p = rep(cur_p, nrow(cur_bb)),
               Formula = cur_bb$Formula)
  }

})

reaction_bb<-do.call(rbind, reaction_bb)
#Collects successfully assigned mass difference occurrence data

reaction_bb$Formula<-reaction_bb$Formula %>% as.character()

z<-names(reaction_bb)

z[1]<-c("from")

z[2]<-c("to")

#Edge data.frame making. It looks for m/z ratios into the mass difference
#matrix: every mass difference corresponds to a series of peaks
#differing for that amount of Da (within a exp_bb_err error).

edges_list<-foreach(p = c(1:nrow(reaction_bb)), .combine = rbind) %do% {
  k<-which(matrix(ddf == reaction_bb[p,1], ncol = ncol(ddf),
                  nrow = nrow(ddf)), arr.ind = T)
  x<-lapply(c(1:nrow(k)),function(i){

    reaction_bb[p,(map_lgl(reaction_bb, is.character)+
                     map_lgl(reaction_bb, is.factor) == 1)]
  })

  x<-do.call(rbind, x)

  data.frame(from = as.numeric(names(ddf[k[,2]])),
             to = as.numeric(names(ddf[k[,1]])),
             x)
} %>% as.data.frame()

names(edges_list)<-z

edges_list$from<-edges_list$from %>% as.character()
edges_list$to<-edges_list$to %>% as.character()

edges_list$id<-c(1:nrow(edges_list)) %>% as.character()


return(edges_list)
```

```
}

#The function automizes the formula calculation process through Molecular
#Network Analysis. The function consider all the assigned peaks
#as starting nodes. The function works as net_filtering_analysis() for
#what concerns the building block data list making ad employing.

net_form_calc<-function(starting_assignments,
                        masslist = NULL,
                        bb_db = NULL,
                        dmatrix = NULL, digits = 3, pmatrix = NULL,
                        bblist = NULL,
                        p_threshold = 30, exp_bb_err = 0.001,
                        ppm = 5){


  df<-starting_assignments      #Objects initialization
  ddf<-dmatrix
  ppf<-pmatrix
  db<-bb_db
  bb_list<-bblist
  mz_df<-masslist

  assigned<-df[(is.na(df$Formula)==F),]

  if(nrow(assigned)>0){

    k<-assigned[,1] %>% table() %>% as.data.frame()

    output<-net_form_local_calc(chemlist = df, masslist = mz_df,
                                dmatrix = ddf, pmatrix = ppf, digits = digits,
                                bblist = bb_list, p_threshold = p_threshold,
                                err_da = exp_bb_err, ppm = ppm, bb_db = db,
                                selected_nodes = as.character(k[,1]))

  }



  return(output)

}

#This function allows to calculate and to assign new chemical formulas
#through molecular network analysis starting from selected nodes.
#The m/z ratios of starting node should be provided as a character vector.
#The function takes a starting chemlist as an input, a list of m/z values,
#a dmatrix (if related argument is NULL, the function will calculate
#it), a pmatrix(if related argument is NULL, the function will calculate
```

```
#it), with related frequency threshold value (p_threshold)
#a building block database (shouldn't be NULL if bblist must be
#calculated), a mass difference error in Da (err_da)
#and a list object in which building blocks are grouped by m/z
#ratio (core object of the script).
#If this last argument is provided, no one of the formers is necessary for
#the function to work.
#Last but not least, an experimental error in ppm must be provided to the
#function for the assignment (ppm argument).

net_form_local_calc<-function(chemlist, masslist = NULL,
                              selected_nodes,
                              bb_db = NULL,dmatrix = NULL,
                              digits = 3, pmatrix = NULL, bblist = NULL,
                              p_threshold = 30, err_da = 0.001,
                              ppm = 5){

  df<-chemlist      #Objects initialization
  ddf<-dmatrix
  ppf<-pmatrix
  db<-bb_db

  if(is.null(masslist)){
    mz_df<-df[,1:2]%>%table()%>%as.data.frame()
    mz_df<-mz_df[(mz_df$Freq>0),1:2]
    mz_df[,1]<-mz_df[,1]%>%as.character()%>%as.numeric()
    mz_df[,2]<-mz_df[,2]%>%as.character()%>%as.numeric()
    mz_df<-mz_df[order(mz_df[,1]),]
  } else{
    mz_df<-masslist
  }


  if(is.null(ddf) & is.null(bblist)){
    ddf<-dmatrixcalculator(masslist1 = mz_df, zero_and_neg_to_na = F,
                           number_of_digits = digits)
  }#Dmatrix calculation (with negative value)

  if(is.null(ppf) & is.null(bblist)){
    ppf<-pmatrixcalculator(dmatrix = ddf, abs_frequency = T,
                           p_threshold = p_threshold)
    ppf<-ppf[(ppf$d!=0),]

  }#Pmatrix calculation

  if(is.null(bblist)){
    bblist<-diff_identifier(database = db, dmatrix = ddf,
                            pmatrix = ppf, error_arg = err_da,
                            p_threshold = p_threshold)
  }
```

```
assigned<-lapply(c(1:length(selected_nodes)), function(x){
  df[(
    as.character(df[,1])%in%selected_nodes[x]
  ),]
})

assigned<-do.call(rbind, assigned)
#Data.frame containing candidate formulae of selected nodes


not_assigned<-mz_df[!(
  as.character(mz_df[,1]) %in% as.character(assigned[,1])
),]

#Initialization of core data.frames: selected nodes are in "assigned", from
# which formulae are calculated by adding and substracting building blocks
#(bblist). These formulae will be used for assignment.


repeat{


  cur_bblist<-bblist[(names(bblist) %in% as.character(assigned[,1]))]
  #Only building blocks related to assigned m/z ratios will be considered
  #for formula calculation


  assignm<-lapply(c(1:length(cur_bblist)), function(x){

    cur_formulae<-assigned[(
      as.character(assigned[,1]) %in% (names(cur_bblist)[x])
    ), ]

    #For every building block group, formulas related to the node
    #are selected for calculation

    candidate_formulae_calculator(starting_formulae = cur_formulae,
                                  building_blocks = cur_bblist[[x]])

  })


  assignm<-do.call(rbind, assignm)


  #########Delete duplicates########

  y<-assignm[,1] %>% table() %>% as.data.frame()
```

```
  assignm<-lapply(c(1:nrow(y)), function(x){

    newdata<-assignm[(
      as.character(assignm[,1]) %in% as.character(y[x,1])
    ),]

    #For every exact mass value, the function extrapolate related duplicate
    #data frame and retain its first row

    newdata[1,]

  })

  assignm<-do.call(rbind, assignm)


  #######

  new_assignm<-assignment_script(masslist = not_assigned,
                                 err_ppm = ppm,
                                 formula_db = assignm)
  #Calculated formulas are used for assignment of unassigned m/z ratios.
  #If no assignment was possible, the function returns the starting masslist
  #with a column 'Formula' with NAs.

  new_assignm<-new_assignm[(
    is.na(new_assignm$Formula)==F
  ),]

  if(nrow(new_assignm)>0){


    assigned<-rbind(assigned[,names(new_assignm)], new_assignm)

    not_assigned<-not_assigned[!(
      not_assigned[,1] %in% assigned[,1]
    ),]

  } else{
    break
  }
}

to_be_assigned<-assigned[!(assigned$Formula %in% df$Formula),]
#Only new calculated formula will be added to the initial data frame


output<-rbind(df[!(
```

```
    as.character(df[,1]) %in% as.character(to_be_assigned[,1])
  ), names(to_be_assigned)], to_be_assigned)

  #As it is, the function deletes already calculated formulas for
  #assigned m/z ratios.


  output<-output[order(output[,1]),]

  return(output)

}

#The function creates a character variable into the input data frame containing
#elemental formulas.

create_Formula<-function(df){
  x<-df

  atom_vector<-c( "C","H" , "N"   ,"O",  'S','P','Br','Cl','I'   ,'Na'  ,'K',
                  "He" , "Li",  "Be",  "B"   ,
                  'F'   ,'Ne',  'Mg',  'Al',  'Si'  ,
                  'Ar',     'Ca',
                  'Sc' , 'Ti'  ,'V'  , 'Cr' , 'Mn'  ,'Fe',  'Co'  ,'Ni',
                  'Cu'  ,'Zn' , 'Ga'  ,'Ge' , 'As' , 'Se'  ,  'Kr'  ,
                  'Rb' , 'Sr' , 'Y' ,  'Zr' ,
                  'Nb'  ,'Mo'  ,'Tc'  ,'Ru'  ,'Rh'  ,'Pd' , 'Ag'  ,'Cd' ,
                  'In'  ,'Sn'  ,'Sb'  ,'Te'  ,'Xe'  ,'Cs' , 'Ba'  ,
                  'La'  ,'Ce'  ,'Pr' , 'Nd' ,
                  'Pm'  ,'Sm'  ,'Eu'  ,'Gd'  ,'Tb'  ,'Dy'  ,'Ho'  ,'Er'  ,
                  'Tm'  ,'Yb'  ,'Lu'  ,'Hf'  ,'Ta'  ,'W'   ,'Re'  ,'Os'  ,
                  'Ir'  ,'Pt'  ,'Au'  ,'Hg' ,
                  'Tl'  ,'Pb'  ,'Bi'  ,'Po'  ,'At'  ,'Rn'  ,'Fr' , 'Ra'  ,
                  'Ac' , 'Th'  ,'Pa' , 'U'   ,'Np'  ,'Pu' , 'Am'  ,'Cm' ,
                  'Bk'  ,'Cf'  ,'Es'  ,'Fm' ,
                  'Md'  ,'No'  ,'Lr'  ,'Rf'  ,'Db'  ,'Sg'  ,'Bh',  'Hs'  ,
                  'Mt',  'Ds'  ,'Rg',  'Cn'  ,'Nh'  ,'Fl',  'Mc'  ,'Lv',
                  'Ts'  ,'Og' )

  #Elements present into the user data frame are identified
  #Be aware! Only variables related to atomic counts have to be named
  #using the appropriate elemental symbol, in order to avoid errors in
  #formula elaboration.
  #Example: peak intensity name "I" should be avoided, since
  #the same symbol is used for iodine.

  cur_elements<-atom_vector[atom_vector %in% names(x)]

  x$Formula<-""
```

```
  for (i in c(1:(length(cur_elements)))){

    x<-mutate(x, Formula = paste(x$Formula, cur_elements[i],
                                 x[,cur_elements[i]], sep = ""))

  }


  for(i in c(1:length(cur_elements))){
    x$Formula<-str_replace(x$Formula,
                           capture.output(
                             cat(cur_elements[i], "0", sep = "")
                           ), "")
  }

  return(x)
}



##########Omics Interactive Formula Assignment App#############


#User interface


ui<-fluidPage(
  sidebarLayout(
    sidebarPanel( #Making the left panel

      fileInput("ddb", label = "Upload your building block
                                .csv file:", accept = c("text/csv")),

      #Action button for building block database upload

      conditionalPanel(condition = "input.bb != null",
                       fileInput("upload",
                                 label = "Upload your
                                       chemlist.csv file :"
                       ),

                       #Once uploaded, two other action buttons will
                       #appear through which personal data can be uploaded.

                       fileInput("isotope",
                                 label = "Upload related
                                       isotopologues .R file:")

                       #Action button for isotopologue list upload
```

```
                #(output of the isot_collector() function)
),


uiOutput("bb_list"), #Select box for building block selection

uiOutput("mscore"), #Select box for m/z selection before isotopic
#pattern matching score calculation

radioButtons("mode", label = "Ionization mode:",
             choices = list(positive = "positive",
                            negative = "negative")),
#Selection of the source ionization mode through radio buttons,
#useful for isotopic pattern matching score calculation only

conditionalPanel(condition = "input.mz != null",
                 actionButton(inputId = "ipf",
                              label = "Calculate mScore for
                                    selected m/z ratio")),
#Perform isotopic pattern matching score calculation

numericInput(
  "delta",
  label = "Insert deltaKMD value :",
  min = 1e-6,
  max = 1e-1,
  step = 0.0001,
  value = 0.001
),

#Selection of KMD error for homologous series identification


numericInput("Da_err", "Insert the value of differences error
                            in Da:",
             min = 1e-8,
             max = 1000,
             value = 0.001,
             step = 0.0001),

#Mass difference error for building block identification

numericInput("ppm", "Insert the mass error
                            in ppm for formula calculation:",
             min = 0.01,
             max = 100,
             value = 5,
             step = 0.01),
#ppm error for formula calculation
```

```
      numericInput("threshold", label = "Insert the building block
                                  frequency threshold",
                   min = 0,
                   value = 30,
                   step = 1),
      #Threshold value for low occurrence mass differences elimination

      textInput("delete", label = "Type the molecular formula to be
                      deleted:"),
      actionButton("godelete", label = "Delete Formula"),
      #Delete single elemental formula

      textInput("keep", label = "Type the molecular formula to be
                      kept:"),
      actionButton("gokeep", label = "Keep Formula"),
      #Keep typed elemental formula and delete all the other formula
      #candidates for related m/z ratio

      helpText("All the other molecular formulae related to
                    the same m/z ratio will be deleted."),


      helpText(
        "Note: Local KMD and Network analysis will be performed
                on points related to
                     a clicked one, i.e. with the same Z* value and a KMD
                     comprised into the range KMD +/- deltaKMD, while
                     Network analysis can be performed after clicked on
                     the starting node. Once performed,
                     these analyses will change the input file, with no chance
to
                     undo."),


      helpText(
        "Note: To start the formula calculation, a starting
                point has to be
                selected on relative plot. The total KMD filtering allows
                to filter candidate formulae by considering every homologous
                series and starting from related lower redundance level
members."),


      downloadButton('download', 'Download Assignments as .csv file')
      #Download output dataset as a .csv file


    ),
```

```
mainPanel(
  tabsetPanel(#inserting and modifying different tabs
    tabPanel(title = "KMD Analysis",
             plotlyOutput("kmd_plot"),
             #Kendrick plot

             fluidRow(
               column(4,   #Action buttons to perform KMD
                      #analysis formula filtering

                      h3("Kendrick Mass Defect analysis"),
                      h4("Total formula filtering"),
                      actionButton("gototalkmd",
                                   "Perform Total KMD Filtering"),
                      h4("Local Formula filtering"),
                      actionButton("gokmd",
                                   "Perform Local KMD Filtering")),

               column(4, #Action buttons for formula calculation

                      h3("KMD Formula calculation"),
                      h4("Local formula calculation"),
                      actionButton("kmd_form_calc",
                                   label = "Perform
                                KMD local formula calculation"),
                      h4("Total Formula calculation"),
                      actionButton("kmdcalc", label = "Perform
                                KMD total formula calculation"),

                      radioButtons(inputId = "nitrorule",
                                   label = "Apply Nitrogen Rule for
                                              calculated formulae",
                                   choices = list(yes = T,
                                                  no = F),
                                   selected = T)),
               #Radio Buttons for nitrogen rule application
               #during elemental formula calculation

               column(4,
                      radioButtons("filter",
                                   label = "Leave homologous series only:",
                                   choices = list(yes = "yes",
                                                  no = "no"),
                                   selected = "no"),
                      #Automatically deletes from Kendrick plot
                      #all the points which don't belong to any
                      #homologous series
```

```
                    h4("Number of members per m/z ratio"),
                    helpText("Identifies homologous series
                                        selected m/z ratio belongs to and
                                        calculates the number of members
                                        together with related KMD standard
                                        deviation."),
                    actionButton("homonumber",
                                 label = "Perform
                                number of members calculation"))
          ),
          tableOutput("subkmd"),
          #Every time a point of the Kendrick plot is selected,
          #homologous series data appear under the plot.


          tableOutput("homonumbertable")),


  tabPanel(title = "Mass Difference analysis",
          #Mass Difference data related tab.

          plotOutput("tmds_plot"),
          #Mass Difference occurrences plot

          fluidRow(
            column(3,
                    radioButtons("tmds",
                                 label = "Show related building blocks:",
                                 choices =
                                   list(yes = "yes", no = "no",
                                        nothing = "Show no label"),
                                 selected = "no")),
            #Radio buttons for Mass Difference value and associated
            #building blocks displaying into the plot

            column(3,
                    numericRangeInput("xrange", "X-axis:",
                                      value = c(0,500)),
                    numericRangeInput("yrange", "Y-axis:",
                                      value = c(0,500)),
                    downloadButton('downloadtmds',
                                   'Download TMDS plot as .pdf file'),
                    downloadButton('downloadtmdsdata',
                                   'Download TMDS data as .csv file')
            )
          )
  ),

  tabPanel(title = "Network Analysis",
```

```
                #Molecular Network analysis related tab.

                visNetworkOutput("network_proxy"), #Molecular Network
                fluidRow(
                  column(4,
                         h3("Network Analysis"),
                         h4("NA local filtering"),
                         actionButton("go_network",
                                      label = "Perform Local
                                Network Analysis") #Perform formula filtering
from
                         #a clicked starting node
                  ),
                  column(4,
                         h3("NA Formula calculation"),
                         h4("Local Formula calculation"),
                         actionButton("net_local_calc",
                                      label = "Perform
                                NA local formula calculation"),
                         #Perform formula calculation from a clicked
                         #starting node

                         h4("Automatic Formula calculation"),
                         actionButton("net_form_calc",
                                      label = "Perform
                                NA automatic formula calculation")
                         #Perform formula calculation from assigned nodes
                  ),
                  column(4,
                         radioButtons("edgeslab", label =
                                        "Show edge labels:",
                                      choices = list(yes = "yes",
                                                     no = "no"),
                                      selected = "no"))
                  #Display edge labels
                ),
                tableOutput("subNA")),
        #Under the network, a table with clicked node data will appear

        tabPanel(title = "Experimental m/z data.frame",
                 tableOutput("masslist")) #Uploaded m/z dataset will be shown
        #in this tab
      )
    )

  )

)
```

```
server <- function(input, output) {

  options(shiny.maxRequestSize=100*1024^2) #Increase maximum upload file
  #size to 100MB

  #Definition of dynamic objects, such as the m/z dataset and the isotopologue
  #list

  chem<-reactiveVal()
  isolist<-reactiveVal()

  observeEvent({input$upload
    input$ddb},{

      #When personal data are uploaded, the following code
      #assure that the dataset is assigned to the reactive value chem().
      #Of course, uploading of a building block dataset should be performed
      #before this step, since Kendrick plot and Molecular Network
      #would be made immediately after personal data upload

      if(!is.null(input$ddb) & !is.null(input$upload)){
        newchem<-read.csv(input$upload$datapath)
        newchem<-kendrick_properties_calculator(df = newchem,
                                                bb_database = db(),
                                                building_block = input$bb)
        chem(newchem)
      }
    }, ignoreNULL = T)

  observeEvent(input$godelete, {

    #The following chunck assure the elimination of the elemental formula
    #to delete from the dataset

    req(input$upload)
    current_data<-chem()
    assigned<-current_data[(is.na(current_data$Formula)==F),]

    new_data<-assigned[!(assigned$Formula %in% input$delete),]

    new_data<-rbind(current_data[!(current_data[,1] %in% new_data[,1]),],
                    new_data)
    new_data<-new_data[order(new_data[,1]),]

    chem(new_data)

  }, ignoreNULL = T)

  observeEvent(input$gokeep, {
```

```
  #The following chunck assures the input elemental formula to be kept
  #into the dataset, deleting all the other candidates for related m/z ratio
  #This code works when the Keep formula action button is pressed.

  req(input$upload)
  current_data<-chem()

  assigned<-current_data[(is.na(current_data$Formula)==F),]

  to_keep_formula<-assigned[(assigned$Formula %in% input$keep),]

  new_data<-rbind(assigned[!(assigned[,1] %in% to_keep_formula[,1]),],
                  to_keep_formula)

  new_data<-rbind(current_data[!(current_data[,1] %in% new_data[,1]),],
                  new_data)
  new_data<-new_data[order(new_data[,1]),]

  chem(new_data)

}, ignoreNULL = T)

observeEvent({input$bb},{
  req(input$upload)

  #When another building block is selected from the left panel,
  #the following chunck allows to recalculate Kendrick properties
  #of the uploaded peak list

  current_data<-chem()
  newchem2<-kendrick_properties_calculator(df = current_data,
                                           bb_database = db(),
                                           building_block = input$bb)
  chem(newchem2)
}, ignoreNULL = F)

observeEvent({input$isotope},{

  load(input$isotope$datapath)

  isolist(cur_list)

}, ignoreNULL = T)

homo_numbers<-reactiveVal()
#Table collecting data of different homologous series the selected
#peak belongs to
```

```
observeEvent({input$homonumber}, {

  #The following chunck allows the identification of different homologous
  #series the selected m/z ratio belongs to by recalculating Kendrick
  #properties for every building block present into the uploaded database

  req(instant_data_chem()) #Clicking on a point of the Kendrick plot
  #is compulsory to let this chunck work

  a <- event_data(event = "plotly_click",
                  source = "sorgente")
  cur_mz<-a$key[1]

  withProgress(message = "Calculation in progress...",
               value = 1, {
                 newdata<-homo_series_numbers(df = masslist(),
                                              mz_ratio = cur_mz,
                                              bb_db = db(),
                                              deltaKMD = input$delta,
                                              chemlist = chem())
                 newdata$Choosen_mz_ratio<-cur_mz
               })

  homo_numbers(newdata)

}, ignoreNULL = T)

output$homonumbertable<-renderTable({

  #Table of different homologous series of the selected point

  req(input$upload)

  if(!is.null(event_data(event = "plotly_click",
                         source = "sorgente"))){
    homo_numbers()
  }

}, digits = 8)

output$mscore<-renderUI({

  #Select box related to the isotopic pattern matching score calculation
  #with which you can select the m/z ratio for which the score
  #can be calculated

  req(input$isotope)

  isoset<-do.call(rbind, isolist())
  isoset<-isoset[order(isoset$'M'),]
```

```
  mz_values<-as.character(isoset$'M')

  mz_list<-lapply(c(1:nrow(isoset)), function(x){
    mz_values[x]
  })
  names(mz_list)<-mz_values

  selectInput("mz", label = "Select the m/z ratio:",
              choices = mz_list)

})

observeEvent(input$ipf,{

  #The following chunck allows the calculation of the isotopic
  #pattern matching score when dedicated action button is clicked

  req(input$mz)
  req(input$upload)

  current_data<-chem()
  cur_iso<-isolist()

  sub_data<-current_data[(as.character(current_data[,1]) %in%
                            input$mz),]

  sub_data<-isotopic_pattern_filtering(chemlist = sub_data,
                                       isolist = cur_iso,
                                       mode = input$mode,
                                       err_ppm = input$ppm
  )

  new_data<-rbind(current_data[!(as.character(current_data[,1]) %in%
                                   input$mz),],
                  sub_data)

  new_data<-new_data[order(new_data[,1]),]

  chem(new_data)


}, ignoreNULL = T)

observeEvent(input$go_network,{

  #Allows filtering of formulas from selected node

  req(input$current_node_id)
```

```
    req(db())

    withProgress(message = "Network Analysis in progress...",
                 value = 1,{
                   current_data<-chem()

                   output<-net_filtering_analysis(chemlist = current_data,
                                                  selected_nodes =
input$current_node_id$nodes[[1]],
                                                  bb_database = db(),
                                                  exp_bb_err = input$Da_err,
                                                  bb_list = bblist())
                   incProgress(amount = 1)

                   chem(output)
                 })

    visNetworkProxy("network_proxy") %>%
      visUpdateNodes(nodes = nodes())

  }, ignoreNULL = T)

  observeEvent(input$edgeslab, {

    #Allows to display edges labels and to hide them when dedicated radio
    #buttons are used

    req(edges())

    switch(input$edgeslab,
           yes = {
             cur_edges<-edges()
             cur_edges$label<-cur_edges$Formula
             edges_to_remove<-cur_edges$id

             visNetworkProxy("network_proxy") %>%
               visRemoveEdges(id = edges_to_remove) %>%
               visUpdateEdges(edges = cur_edges)},
           no = {
             cur_edges<-edges()
             cur_edges$label<-c("")
             edges_to_remove<-cur_edges$id

             visNetworkProxy("network_proxy") %>%
               visRemoveEdges(id = edges_to_remove) %>%
               visUpdateEdges(edges = cur_edges)})

  }, ignoreNULL = T)

  observeEvent(input$gokmd,{
```

```
  #Filtering formula locally using KMD analysis

  req(instant_data_chem())

  current_data<-chem()
  cur_homo_chem<-instant_data_chem()

  homoseries<-masslist()[(as.character(masslist()[,1])%in%
                            as.character(cur_homo_chem[,1])),]
  homoseries<-homoseries[order(homoseries[,1]),]
  homoseries<-q_calculator(homoseries,
                           bbinteger = round(db()[(db()$Formula
                                                   %in% input$bb),
                                                  1])) #q_calculator function
  #needs the building block nominal mass to be specified.

  filtered_data<-kendrickfilter(homoseriesdf = cur_homo_chem,
                                bb_database = db(),
                                bb_formula = input$bb,
                                max_q_value = max(homoseries$q,
                                                  na.rm = T))
  new_data<-current_data[!(
    current_data[,1] %in% filtered_data[,1]
  ),]

  new_data<-bind_rows(new_data, filtered_data)

  chem(new_data)

})

observeEvent(input$gototalkmd, {

  #Automatic filtering of elemental formula through KMD analysis
  #by considering every identified homologous series

  req(chem())

  withProgress(message = "KMD automatic filtering in progress...",
               value = 1,{
                 current_data<-chem()
                 new_data<-kmd_analysis(chemlist = current_data,
                                        bb_database = db(),
                                        deltaKMD = input$delta,
                                        chosen_bb = input$bb)
                 incProgress(amount = 1)
                 chem(new_data)
               })
```

```
    visNetworkProxy("network_proxy") %>%
      visUpdateNodes(nodes = nodes())

  }, ignoreNULL = T)

  observeEvent(input$net_form_calc, {
    req(network())

    #Formula calculation related to Network Analysis from every
    #assigned node

    withProgress(message = "NA automatic formula calculation in progress...",
                 value = 1,{
                   current_data<-chem()
                   new_data<-net_form_calc(starting_assignments = current_data,
                                           bblist = bblist(),
                                           exp_bb_err = input$Da_err,
                                           ppm = input$ppm)
                   incProgress(amount = 1)
                   chem(new_data)
                 })

    visNetworkProxy("network_proxy") %>%
      visUpdateNodes(nodes = nodes())
  })

  observeEvent(input$net_local_calc, {
    req(network())

    #Formula calculation related to Network Analysis from
    #selected node

    withProgress(message = "NA local formula calculation in progress...",
                 value = 1,{
                   current_data<-chem()
                   new_data<-net_form_local_calc(chemlist = current_data,
                                                 selected_nodes =
input$current_node_id$nodes[[1]],
                                                 bblist = bblist(),
                                                 err_da = input$Da_err,
                                                 ppm = input$ppm)
                   incProgress(amount = 1)

                 })

    chem(new_data)

    visNetworkProxy("network_proxy") %>%
      visUpdateNodes(nodes = nodes())
  }, ignoreNULL = T)
```

```
  observeEvent(input$kmd_form_calc, {

    #Formula calculation through KMD analysis from selected homologous series
    #member

    req(instant_data_chem()) #Clicking on a point is compulsory to let
    #the following chunck running

    current_data<-chem()

    a <- event_data(event = "plotly_click",
                    source = "sorgente")

    withProgress(message = "Calculation in progress...",
                 {
                   homoseries<-
masslist()[(masslist()[,1]%in%instant_data_chem()[,1]),]
                   homoseries<-homoseries[order(homoseries[,1]),]
                   homoseries<-q_calculator(homoseries,
                                            bbinteger =
round(db()[(db()$Formula
                                                                    %in%
input$bb),
                                                                   1]))
                   q_value<-max(homoseries$q, na.rm = T)
                   q_value<-q_value*10

                   starting<-instant_data_chem()[(
                     as.character(instant_data_chem()[,1]) %in%
                       as.character(a$key[1])
                   ),]
                   not_assigned<-homoseries[!(
                     as.character(homoseries[,1]) %in%
                       as.character(starting[,1])
                   ),]

                   bb<-db()[(db()$Formula==input$bb),map_lgl(db(), is.numeric)]

                   bb<-lapply(c((q_value*(-1)):q_value), function(i){
                     bb*i
                   })
                   bb<-do.call(rbind, bb)

                   if(nrow(starting)>0 & nrow(not_assigned)>0){
                     form_db<-candidate_formulae_calculator(starting_formulae =
starting,
                                                            building_blocks =
bb,
```

```
seven_golden_rules_applier = F)

                      new_assigned<-assignment_script(masslist =
not_assigned[,1:2],
                                                      err_ppm = input$ppm,
                                                      formula_db = form_db,
                                                      nitrogen_rule_applier =
input$nitrorule)



                      new_assigned<-kendrick_properties_calculator(df =
new_assigned,

building_block = input$bb,
                                                                   bb_database =
db())

                      z<-names(starting)



if(nrow(new_assigned[(is.na(new_assigned$Formula)==F),])>0){

                        new_assigned<-new_assigned[,z]

                        new_data<-rbind(current_data[!(
                          as.character(current_data[,1]) %in% as.character(
                            new_assigned[,1])
                        ), names(new_assigned)], new_assigned)

                        chem(new_data)
                      } else{
                        chem(current_data)
                      }
                    } else{
                      chem(current_data)
                    }
                  }, value = 1)

  })

  observeEvent(input$kmdcalc, {
    req(network())

    #Automatic formula calculation through KMD analysis considering every
    #homologous series and starting from related lowest redundance
    #level members

    withProgress(message = "KMD formula calculation in progress...",
```

```
                value = 1,{
                  current_data<-chem()
                  new_data<-kmd_form_calc(chemlist = current_data,
                                          bb_database = db(),
                                          deltaKMD = input$delta,
                                          chosen_bb = input$bb,
                                          ppm = input$ppm)
                  incProgress(amount = 1)

                })

  chem(new_data)

  visNetworkProxy("network_proxy") %>%
    visUpdateNodes(nodes = nodes())
}, ignoreNULL = T)


masslist <- reactive({
  req(chem())

  #Another peak list is created here apart from uploaded one
  #without elemental formula data to speed up KMD and NA calculations

  masslist <- chem()[, 1:2] %>%
    table() %>%
    as.data.frame() %>%
    subset(Freq > 0)

  masslist[,1] <- masslist[,1] %>%
    as.character() %>%
    as.numeric()

  masslist[,2] <- masslist[,2] %>%
    as.character() %>%
    as.numeric()

  masslist<-masslist[order(masslist[,1]),]

  masslist <- kendrick_properties_calculator(df = masslist,
                                             bb_database = db(),
                                             building_block = input$bb)
  masslist$Freq[(masslist[,1] %in%
                   chem()[(is.na(chem()$Formula)==T),1])]<-0

  masslist$Redun <- c("uniq", "redundance")[((masslist$Freq != 1) +
                                               1)]
  masslist$Redun[(masslist$Freq==0)]<-"not assigned"

  return(masslist)
```

```
})

ddf<-reactive({

  #Mass difference matrix calculation. Row and column names are m/z ratio
  #and every matrix object contains the difference between m/z values
  #related to the row and the column the mass difference belongs to

  req(input$upload)
  dmatrixcalculator(masslist1 = masslist(),
                    number_of_digits = (floor(log10(input$Da_err))*(-1)),
                    zero_and_neg_to_na = T)
  #Negative and zero values of ddf() are converted to NA
})

complete_ddf<-reactive({
  req(input$upload)
  dmatrixcalculator(masslist1 = masslist(),
                    number_of_digits = (floor(log10(input$Da_err))*(-1)),
                    zero_and_neg_to_na = F)
  #Here, ddf negative and zero values are retained. This kind of ddf
  #is useful for network calculation
})

p_df<-reactive({
  req(ddf())

  #Collecting every different mass difference value in a matrix object
  #with related occurrences. Mass differences with occurrences lower
  #than a certain value input$threshold are automatically deleted

  pmatrixcalculator(dmatrix = ddf(),
                    abs_frequency = T,
                    p_threshold = input$threshold)
})

db<-reactiveVal()
#Building block database reactive value

observeEvent(input$ddb, {

  new_db<-read.csv(input$ddb$datapath)

  #Uploaded building block database is assigned to related reactive value
  #db()

  db(new_db)

})
```

```
output$bb_list<-renderUI({

  req(db())

  #Select box through which the KMD building block can be choosen among
  #those uploaded

  current_db<-db()
  bb<-current_db$Formula
  bbl<-lapply(c(1:length(bb)), function(i){
    bb
  })
  names(bbl)<-bb

  selectInput(
    "bb",
    label = "Choice current building block :",
    choices = bbl,
    selected = 1
  )


})

text_pdf<-reactive({

  #Matrix object containing assigned mass difference values
  #together with relative occurrences.
  #Useful for edge matrix calculation

  req(p_df())
  req(input$ddb)

  value_df<-p_df()
  err<-input$Da_err
  db<-db()

  reaction_bb<-lapply(c(1:nrow(value_df)), function(i){
    cur_d<-value_df$d[i]
    cur_p<-value_df$p[i]
    cur_sign<-NA
    ifelse(cur_d>0, yes = cur_sign<-1,
           no = cur_sign<-(-1))
    cur_bb<-db[(
      (db[,1]>(abs(cur_d)-err)) &
        (db[,1]<(abs(cur_d)+err))
    ),]

    if(nrow(cur_bb)>0){
      data.frame(d = rep(cur_d, nrow(cur_bb)),
```

```
                  p = rep(cur_p, nrow(cur_bb)),
                  Formula = cur_bb$Formula,
                  Sign = rep(cur_sign, nrow(cur_bb)))
    }

  })

  reaction_bb<-do.call(rbind, reaction_bb)

  reaction_bb$Formula<-reaction_bb$Formula %>% as.character()

  return(reaction_bb)

})


bblist<-reactive({

  #bblist collects edges data for every peak of the uploaded data.
  #It's the core of the network and is used for formula filtering
  #and calculation through Network Analysis

  output_list<-diff_identifier(database = db(),
                               dmatrix = complete_ddf(),
                               pmatrix = NULL,
                               p_threshold = input$threshold,
                               error_arg = input$Da_err
  )

  return(output_list)

})


color_var <- reactive({
  req(chem())

  #Reactive value useful for homologous series highlighting

  if (!is.null(event_data(event = "plotly_click",
                          source = "sorgente"))) {
    color_var <- c("unselected", "selected")[((masslist()[,1] %in%
                                                 instant_data_chem()[,1]) +
                                                1)]

    return(color_var)
  } else{
    color_var <- masslist()$Redun
    return(color_var)
```

```
  }
})

color_vector <- reactive({
  req(chem())

  #Colors related to the Kendrick plot points.
  #The color vector must be a reactive value because it's supposed to
  #change whenever a point of the plot is selected or not

  if (!is.null(event_data(event = "plotly_click",
                          source = "sorgente"))) {
    return(c("orange", "grey"))
  } else{
    return(c("blue", "red", "green"))
  }
})


instant_data_chem <- reactive({

  #Dataset containing selected point information

  req(chem())

  a <- event_data(event = "plotly_click",
                  source = "sorgente")

  current_data<-chem()

  insta_data <- current_data[(current_data[,1] == a$key), ]

  row_to_add <-current_data[(
    (current_data$KMD < (insta_data$KMD[1] + input$delta)) &
      (current_data$KMD > (insta_data$KMD[1] - input$delta)) &
      (current_data$zstar == insta_data$zstar[1])
  ),]

  row_to_add<-row_to_add[order(row_to_add[,1]),]


  return(row_to_add)
})

filtered_kmdset<-reactive({

  #List object containing every identified homologous series
  #Useful for Kendrick plot making
```

```
  req(masslist())

  output<-kendrick_noise_filter(masslist = masslist()[,1:2],
                                bb_database = db(),
                                chosen_bb = input$bb,
                                deltaKMD = input$delta)
  return(output)

})

filtered_masslist<-reactive({

  req(filtered_kmdset())

  #Peak list containing m/z ratios which belong to at least 1 homologous
  #series. Useful for Kendrick plot making

  filtered<-do.call(rbind, filtered_kmdset())

  if(length(nrow(filtered)>0)!=0){
    filtered<-filtered[order(filtered[,1]),]
    output<-masslist()[(masslist()[,1]%in%filtered[,1]),]

    return(output)

  } else{

    return(masslist())

  }




})

kmdplot<-reactive({

  req(masslist())

  #Complete Kendrick plot
  #It's a plot_ly object, assuring interactivity

  plot_ly(
    data = masslist(),
    x = masslist()$NM,
    y = masslist()$KMD,
    type = "scatter",
```

```
      mode = "markers",
      hoverinfo = "text",
      text = paste(
        "</br>m/z = ",
        masslist()[, 1],
        "</br>Z* = ",
        masslist()$zstar,
        "</br>KMD = ",
        masslist()$KMD,
        "</br>NM = ",
        masslist()$NM,
        "</br>N. of Formulae = ",
        masslist()$Freq
      ),
      color = color_var(),
      colors = color_vector(),
      key = ~ mz,
      source = "sorgente"
    ) %>%
      config(scrollZoom = T) %>%
      layout(xaxis = list(title = "Kendrick Nominal Mass (KNM)",
                          font = list(family = "Times New Roman"),
                          linecolor = "black",
                          linewidth = 0.5,
                          mirror = T),
             yaxis = list(title = "Kendrick Mass Defect (KMD)",
                          font = list(family = "Times New Roman"),
                          linecolor = "black",
                          linewidth = 0.5,
                          mirror = T))

})

color_var_filtered <- reactive({
  req(chem())
  req(filtered_masslist())

  #If only homologous series have to be displayed (radio button)
  #The reactive value is a vector of colors for the filtered
  #Kendrick plot

  if (!is.null(event_data(event = "plotly_click",
                          source = "sorgente"))) {
    color_var <- c("unselected",
                   "selected")[((filtered_masslist()[,1] %in%
                                   instant_data_chem()[,1]) +
                                  1)]

    return(color_var)
  } else{
    color_var <- filtered_masslist()$Redun
```

```
    return(color_var)
  }
})

filtered_kmdplot<-reactive({

  req(filtered_masslist())

  #Kendrick plot with homologous series only


  plot_ly(
    data = filtered_masslist(),
    x = filtered_masslist()$NM,
    y = filtered_masslist()$KMD,
    type = "scatter",
    mode = "markers",
    hoverinfo = "text",
    text = paste(
      "</br>m/z = ",
      filtered_masslist()[, 1],
      "</br>Z* = ",
      filtered_masslist()$zstar,
      "</br>KMD = ",
      filtered_masslist()$KMD,
      "</br>NM = ",
      filtered_masslist()$NM,
      "</br>N. of Formulae = ",
      filtered_masslist()$Freq
    ),
    color = color_var_filtered(),
    colors = color_vector(),
    key = ~ mz,
    source = "sorgente"
  ) %>%
    config(scrollZoom = T) %>%
    layout(xaxis = list(title = "Kendrick Nominal Mass (KNM)",
                        font = list(family = "Times New Roman"),
                        linecolor = "black",
                        linewidth = 0.5,
                        mirror = T),
           yaxis = list(title = "Kendrick Mass Defect (KMD)",
                        font = list(family = "Times New Roman"),
                        linecolor = "black",
                        linewidth = 0.5,
                        mirror = T))

})

output$subkmd<-renderTable({
```

```
  #A table collecting all the selected homologous series data

  req(input$upload)

  if(!is.null(event_data(event = "plotly_click",
                         source = "sorgente"))){
    instant_data_chem()
  }
})


output$kmd_plot <- renderPlotly({
  req(filtered_kmdplot())
  req(kmdplot())

  #Depending on the selected radio button, the Kendrick plot
  #or the filtered one is returned to the user

  switch(input$filter,
         yes = filtered_kmdplot(),
         no = kmdplot())

})



tmdsplot<-reactive({

  req(p_df())

  #Mass Difference occurrence plot
  #This is made by employing ggplot2 package, since
  #plot_ly leads to errors
  #This plot shows mass difference values as labels

  pl<-ggplot(data = p_df(), aes(x = d, y = p))

  output<-pl + geom_bar(stat = "identity", fill = "gray",
                        color = "black") +
    coord_cartesian(xlim = input$xrange, ylim = input$yrange) +
    theme_bw() +
    theme(axis.title.x = element_text(hjust = 1, face = "bold"),
          axis.title.y = element_text(hjust = 1, face = "bold")) +
    xlab(label = "Mass Difference (Da)") +
    ylab(label = "Occurrence")

  return(output + geom_text(data = p_df(), aes(x = p_df()$d,
                                               y = (p_df()$p +20),
```

```
                                                label = p_df()$d)))

})

labeled_tmdsplot<-reactive({

  req(text_pdf())

  #Mass difference occurrence plot with labels indicating successfully
  #assigned mass differences

  if(nrow(text_pdf())>0){

    pl<-ggplot(data = p_df(), aes(x = d, y = p))

    output<-pl + geom_bar(stat = "identity", fill = "gray",
                          color = "black") +
      coord_cartesian(xlim = input$xrange, ylim = input$yrange) +
      theme_bw() +
      theme(axis.title.x = element_text(hjust = 1, face = "bold"),
            axis.title.y = element_text(hjust = 1, face = "bold")) +
      xlab(label = "Mass Difference (Da)") +
      ylab(label = "Occurrence")

    output + geom_text(data = text_pdf(), aes(x = text_pdf()$d,
                                              y = (text_pdf()$p +20),
                                              label = text_pdf()$Formula,
                                              angle = 90))

  } else{

    return(tmdsplot())
  }

})

nolabel_tmds_plot<-reactive({

  req(p_df())

  #Mass difference occurrence plot with no label at all

  pl<-ggplot(data = p_df(), aes(x = d, y = p))

  output<-pl + geom_bar(stat = "identity", fill = "gray",
                        color = "black") +
    coord_cartesian(xlim = input$xrange, ylim = input$yrange) +
    theme_bw() +
    theme(axis.title.x = element_text(hjust = 1, face = "bold"),
```

```
          axis.title.y = element_text(hjust = 1, face = "bold")) +
    xlab(label = "Mass Difference (Da)") +
    ylab(label = "Occurrence")

  return(output)

})

output$tmds_plot<-renderPlot({

  req(tmdsplot())
  req(labeled_tmdsplot())

  #Depending on the selected radio button, a type of mass difference
  #occurrence plot is returned to the user

  switch (input$tmds,
          "yes" = labeled_tmdsplot(),
          "no" = tmdsplot(),
          "Show no label" = nolabel_tmds_plot()
  )
})

id_nodes_to_remove<-reactiveVal()

nodes<-reactive({

  #Nodes data frame

  req(input$ddb)
  req(edges())

  nodes_data<-masslist()[,1:3]

  names(nodes_data)<-c("id", "I%", "Freq")

  nodes_data$id<-nodes_data$id %>%
    as.character()

  #nodes_data$size<-nodes_data$size %>%
  #    as.character() %>%
  #    as.numeric()

  #nodes_data<-mutate(nodes_data, size = (size/max(size))*100)

  nodes_data$shape<-"dot"

  nodes_data$shadow<-TRUE
```

```
  nodes_data$color.highlight.background<-"orange"

  nodes_data$color.highlight.border<-"darkred"

  nodes_data$color.border<-"black"

  nodes_data$color.background<-c("red", "blue")[(
    (nodes_data$Freq != 1)+1
  )]

  nodes_data$color.background[(nodes_data$Freq==0)]<-c("gray")

  nodes_data$title<-paste("m/z = ", nodes_data$id,
                          "</br> N. of Formulae = ", nodes_data$Freq,
                          "</br> I % =", nodes_data$'I%')
  nodes_data$label<-paste("m/z = ", nodes_data$id,
                          "\nN. of Formulae = ", nodes_data$Freq,
                          "\nI % =", nodes_data$'I%')


  #nodes_data<-nodes_data[(nodes_data$id %in% edges()$from |
  #                          nodes_data$id %in% edges()$to),]

  return(nodes_data)


})

edges<-reactive({

  #Edges data frame

  req(input$ddb)
  req(db())
  req(ddf())
  req(p_df())

  data_edges<-make_edges_list(dmatrix = ddf(),
                              database = db(),
                              value_to_edges_df = p_df(),
                              exp_bb_err = input$Da_err)


  return(data_edges)

})

id_edges_to_remove<-reactiveVal()
#To update edges in returned Molecular Network every time the mass
```

```
#difference error or the occurrence threshold value are changed,
#the creation of a reactive value that collect current edges IDs is
#compulsory

observeEvent({input$ddb
  input$upload}, {
    req(edges())

    id_edges_to_remove(edges()$id)

  })

network<-reactiveVal()
#The Molecular Network is assigned to the reactive value network()

observeEvent({input$upload
  input$ddb},{

    #This chunck allows the making of the Molecular Network from
    #uploaded data.
    #It runs after uploaded personal peak list

    if(!is.null(input$ddb) & !is.null(input$upload)){
      nodes_data<-nodes()
      edges_data<-edges()

      new_network<-visNetwork(nodes = nodes_data, edges = edges_data) %>%
        visOptions(nodesIdSelection = T, highlightNearest = T)  %>%
        visEvents(select = 'function(nodes_data){
                    Shiny.onInputChange("current_node_id",nodes_data);
                    ;}')

      network(new_network)
    }
  })

output$network_proxy<-renderVisNetwork({

  network()

})

output$subNA<-renderTable({

  #Table collecting all the selected node data, returned to the
  #user under the Molecular Network after having selected a node

  req(network())
```

```
  if(!is.null(input$current_node_id$nodes[[1]])){
    chem()[(as.character(chem()[,1]) %in%
              input$current_node_id$nodes[[1]]),]
  } else{
    paste("")
  }

})


observeEvent({input$threshold
  input$Da_err}, {

    #The chunck allows to update the network whenever the mass difference
    #error and the occurrence threshold are changed

    req(input$ddb)

    new_edges_data<-edges()

    visNetworkProxy("network_proxy") %>%
      visRemoveEdges(id = id_edges_to_remove()) %>%
      visUpdateEdges(edges = new_edges_data)

    id_edges_to_remove(new_edges_data$id)

  }, ignoreNULL = T)

output$masslist<-renderTable({
  masslist()
}, digits = 8)


output$download <- downloadHandler(

  # This function returns a string which tells the client
  # browser what name to use when saving the file.
  filename = function() {
    paste(input$upload, "csv", sep = ".")
  },

  # This function should write data to a file given to it by
  # the argument 'file'.
  content = function(file) {

    # Write to a file specified by the 'file' argument
    write.csv(chem(), file, row.names = FALSE)
  }
)
```

```
  tmds_to_download<-reactive({
    req(input$upload)

    switch (input$tmds,
            "yes" = labeled_tmdsplot(),
            "no" = tmdsplot(),
            "Show no label" = nolabel_tmds_plot())

  })

  output$downloadtmds <- downloadHandler(

    #This chunck allows to download the mass difference occurrence plot
    #as a .pdf high quality file

    filename = function() {
      paste(input$upload, "pdf", sep = ".")
    },


    content = function(file) {

      ggsave(filename = file, plot = tmds_to_download(), dpi = 300,
             width = 6.38, height = 3)

    }
  )

  output$downloadtmdsdata <- downloadHandler(

    #Download mass difference occurrence data as a data table (.csv file)

    filename = function() {
      paste(input$upload, "csv", sep = ".")
    },


    content = function(file) {

      # Write to a file specified by the 'file' argument
      write.csv(text_pdf(), file, row.names = FALSE)
    }
  )

}

#Run the application after having defined ui and server objects
```

```r
#into the Global Environment
shinyApp(ui = ui, server = server)
```

6. Conclusions

In this Ph. D. thesis work, High Resolution FT-ICR Mass Spectrometry was used to obtain useful information on metabolic profile of traditional Italian food products produced in the Basilicata region (Italy) and to evaluate its potentials to be used routinely for food authentication and traceability. More specifically, HRMS analyses were conducted on *Peperoni di Senise PGI* peppers, *Fagioli Bianchi di Rotonda PDO* beans, *Melanzane Rosse di Rotonda PDO* eggplants and seven types of red and white wines obtained from new germplasms growth in the Pollino natural area, i.e. *Aglianico Bianco*, *Santa Sofia*, *Malvasia ad Acino Piccolo*, *Guisana*, *Giosana*, *Colata Murro* and *Plavina* wines, by assuming a direct-injection approach. Related MS spectra were obtained after few seconds and led to the identification of thousands of different compounds with a single direct analysis. The high levels of resolution and accuracy reached led to the construction of *metabolomic fingerprints* or 2D Van Krevelen plots which made easy the identification of classes of biocompounds present in analysed matrices and the evaluation of related diversity. Moreover, the comparison of Van Krevelen diagrams of analysed wines supported their utilization to make a distinction among them on a molecular basis, since differences between obtained fingerprints were present. Furthermore, the usefulness of dedicated software to maximize the reliable information gettable from HRMS was demonstrated. In this sense, the new commercial software AutoVectis Pro was employed successfully to delete artifacts, i.e. *wiggles* and RFI peaks, which could led to errors during the execution of formula assignment step, and to identify a higher number of ionic species by increasing MS signal resolution and S/N ratio. Moreover, a new R Shiny app, i.e. the OIFA software, has been developed in this work to make the formula assignment step interactive and, thus, more accurate, letting the user to interact with built Kendrick plot and Molecular Network in order to guide formula filtering and calculation reducing the risk of misassignments. To summarize, High Resolution FT-ICR MS, together with the utilization of dedicated software, was able to provide a huge amount of reliable and useful information on the metabolome of analysed traditional Italian food products under the form of *molecular fingerprints* or *cartographies*, which could be used for their protection against adulteration and food fraud. However, despite

the promising starting point provided by this work, still a lot should be done to fully demonstrate the validity of this approach for routine food product certification and traceability by performing a validation study on selected matrices.

Publications

Articles

- A. Onzo, R. Pascale, M. A. Acquavia, P. Cosma, J. Gubitosa, C. Gaeta, P. Iannece, Y. Tsybin, V. Rizzi, A. Guerrieri, R. Ciriello, G. Bianco, *Untargeted analysis of pure snail slime and snail slime-induced Au nanoparticles metabolome with MALDI FT-ICR MS*, J. of Mass Spec. 2021, 56(5), e4722, doi: 10.1002/jms.4722;
- A. Santarsiero, A. Onzo, R. Pascale, M. A. Acquavia, M. Coviello, P. Convertini, S. Todisco, M. Marsico, C. Pifano, P. Iannece, C. Gaeta, S. D'Angelo, M. C. Padula, G. Bianco, V. Infantino, G. Martelli, *Pistacia lentiscus Hydrosol: Untargeted Metabolomic Analysis and Anti-Inflammatory Activity Mediated by NF-κB and the Citrate Pathway*, Oxidative Medicine and Cellular Longevity 2020, doi: 10.1155/2020/4264815;
- J. Gubitosa, V. Rizzi, P. Fini, A. Laurenzana, G. Fibbi, C. Veiga-Villauriz, F. Fanelli, F. Fracassi, A. Onzo, G. Bianco, C. Gaeta, A. Guerrieri, P. Cosma, *Biomolecules from snail mucus (Helix aspersa) conjugated gold nanoparticles, exhibiting potential wound healing and anti-inflammatory activity*, Soft Matter 2020, doi: 10.1039/D0SM01638A;
- R. Pascale, M. A. Acquavia, T. R. I. Cataldi, A. Onzo, D. Coviello, S. A. Bufo, L. Scrano, R. Ciriello, Guerrieri, G. Bianco, *Profiling of quercetin glycosides and acyl glycosides in sun-dried peperoni di Senise peppers (Capsicum annuum L.) by a combination of LC-ESI(-)-MS/MS and polarity prediction in reversed-phase separations*, Anal. Bioanal. Chem. 2020, 412, 3005-3015, doi: 10.1007/s00216-020-02547-2;
- A. Onzo, M.A. Acquavia, T.R.I. Cataldi, M. Ligonzo, D. Coviello, R. Pascale, G. Martelli, M. Bondoni, L. Scrano, G. Bianco, *Coceth Sulfate Characterization by Electrospray Ionization with Tandem Mass Spectrometry*, Rapid Comm. Mass Spectrom. 2020, doi: 10.1002/rcm.8884;
- R. Pascale, A. Onzo, R. Ciriello, L. Scrano, S. A. Bufo, G. Bianco, *LC/MS Based Food Metabolomics*, Reference Module in Food Science 2020, doi: 10.1016/B978-0-08-100596-5.22774-1.

Poster

- A. Onzo, R. Pascale, D. Coviello, P. Iannece, C. Gaeta, G. Bianco, *Nontargeted food fingerprinting of typical food products of the Basilicata region (Italy)*, 6th MS Food Day, 25-27 September 2019, Camerino, Italy;
- A. Onzo, G. Bianco, R. Ciriello, C. Gaeta, *Molecular cartography of Metabolome of typical food products of Basilicata region (Italy)*, EU FTICR Mass Spectrometry 1st Advanced School, 14-18 April 2019, Lisbon, Portugal;

- A. Onzo, G. Bianco, J. Gubitosa, V. Rizzi, P. Fini, A. Laurenzana, G. Fibbi, C. Veiga-Villauriz, F. Fanelli, F. Fracassi, C. Gaeta, A. Agostiano, P. Cosma, A. Guerrieri, *Untargeted metabolomic analysis of snail slime by ultra-high resolution MALDI-FT-ICR Mass Spectrometry*, XXVIII Congress of the Analytical Chemistry Division, 22-26 September 2019, Bari, Italy;
- R. Pascale, A. Onzo, G. Bianco, P. Schmitt Kopplin, T. R. I. Cataldi, L. Milella, *Target and Non-Target Analysis of Fagioli di Sarconi beans extracts (Phaseolus Vulgaris L.) by using Fourier-Transform Ion Cyclotron Resonance Mass Spectrometry*, 13th European Fourier-Transform Mass Spectrometry Workshop, 24-27 April 2018, Freising, Germany;
- A.Onzo, D. Coviello, G. Bianco, R. Pascale, Acquavia M. A., P. Schmitt Kopplin, P. Iannece, C. Gaeta, *Non-Target Metabolomic Analysis of Peperoni di Senise peppers PGI (Capsicum Annuum L.) and Evaluation of their Nutraceutical Properties*, XXII International Mass Spectrometry Conference, 26-31 August 2018, Florence, Italy.

Oral Communications

- M. A. Acquavia, R. Pascale, T. R. I. Cataldi, A. Onzo, D. Coviello, S. A. Bufo, L. Scrano, R. Ciriello, A. Guerrieri, G. Bianco, *QUEdb: a new comprehensive database of quercetin glycosides and acyl-glycosides derivatives*, Banche Dati e tools informatici in Spettrometria di Massa, 11 December 2019, Rome;
- A. Onzo, G. Bianco, R. Pascale, P. Iannece, C. Gaeta, *Molecular Fingerprinting of traditional food products by ultra-high resolution ESI-FT-ICR Mass Spectrometry*, XXVIII Congress of the Analytical Chemistry Division, 22-26 September 2019, Bari, Italy;
- A. Onzo, G. Bianco, G. Martelli, V. Infantino, R. Pascale, P. Iannece, C. Pifano, C. Gaeta, *Molecular fingerprinting of Pistacia Lentiscus hydrosol by ultra-high resolution MALDI-FT-ICR Mass Spectrometry*, 3rd MS-NatMed Day, 19-21 June 2019, Aboca, Italy;
- A. Onzo, G. Bianco, P. Iannece, P. Schmitt-Kopplin, C. Gaeta, *Phytochemical screening of new italian wine varieties by using High Resolution Mass Spectrometry*, 3rd Mass Spectrometry Wine Day, 16-17 May 2019, San Michele dell'Adige, Italy.

Attended Workshops and Schools

Workshops

- ***6th Mass Spectrometry Food Day***, Università degli Studi di Camerino, Camerino, Italy, 25-27 September 2019;
- ***XXVIII Congress of the Analytical Chemistry Division***, Università degli Studi di Bari Aldo Moro, Bari, Italy, 22-26 September 2019;
- ***3rd Mass Spectrometry NatMedDay / MASSA 2019***, Aboca, Italy, 19-21 June 2019;
- ***3rd Mass Spectrometry Wine Day***, Fondazione Edmund Mach, San Michele dell'Adige, Italy, 16-17 May 2019;
- ***XXII International Mass Spectrometry Conference***, Fortezza da Basso, Florence, Italy, 26-31 August 2018;
- ***13th European Fourier-Transform Mass Spectrometry Workshop***, Kardinal-Doepfner-Haus, Freising, Germany, 24-27 April 2018.

Schools

- ***School of Multivariate Analysis***, Università degli Studi di Genova, 21/09/2020 – 25/09/2020, Genova, Italy;
- ***Detection, Diagnosis, and Health Concerns of Toxic Chemical and Biological Agents***, NATO Science for Peace and Security Programme Advanced Study Institute G5535, 29/09/2019 – 05/10/2019, Cetraro, Italy;
- ***Validazione dei Metodi LC-MS ed LC-MS/MS nella ricerca scientifica***, Consiglio Nazionale delle Ricerche – Istituto di Chimica Biomolecolare, 12/09/2019 – 13/09/2019, Pozzuoli, Italy;
- ***3rd short course of the EU FT-ICR Mass Spectrometry network***, Dipartimento di Chimica e Tecnologie del Farmaco, La Sapienza Università di Roma, 25/06/2019 – 27/06/2019, Rome, Italy;
- ***School of Experimental Design***, Università degli Studi di Genova, 27/05/2019 – 31/05/2019, Genova, Italy;
- ***EU FT-ICR Mass Spectrometry 1st Advanced School***, Faculdade de Ciências, Universidade de Lisboa, 14/04/2019 – 18/04/2019, Lisbone, Portugal;
- ***2nd Short Course of the EU-FT-ICR network***, Université Paris Sud, 05/11/2018 – 07/11/2018, Orsay, France;
- ***Applicazione della Spettrometria di Massa in campo ambientale e alimentare***, Università degli Studi della Basilicata, 02/07/2018 – 20/07/2018, Potenza, Italy;

- ***2nd European Fourier-Transform Mass Spectrometry School***, Helmholtz Zentrum Muenchen, 23/04/2018 – 24/04/2018, Munich, Germany;
- ***22th Mass Spectrometry School***, Società Chimica Italiana, Divisione di Spettrometria di Massa, 12/03/2018 – 16/03/2018, Siena, Italy;
- ***Corso di Formazione sulla valutazione del rischio in ambienti di lavoro***, Università degli Studi della Basilicata, 06/02/2018 – 06/02/2018, Potenza, Italy.