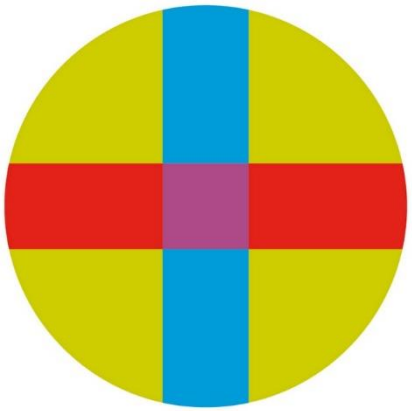
UNIVERSITY CEU - SAN PABLO

POLYTECHNIC SCHOOL

BIOMEDICAL ENGINEERING DEGREE



BACHELOR THESIS

Design, validation and implementation of a MS^2 based search and an in-source fragments detector in MS^1 based searches

Author: María Postigo Fliquete
Director: Alberto Gil De La Fuente,
Abraham Otero Quintana

July 2018

UNIVERSIDAD SAN PABLO-CEU

ESCUELA POLITÉCNICA SUPERIOR

División de Ingeniería

Datos del alumno

NOMBRE:

Datos del Trabajo

TÍTULO DEL PROYECTO:

Tribunal calificador

PRESIDENTE:	FDO.:

SECRETARIO:	FDO.:

VOCAL:	FDO.:

Reunido este tribunal el _____/_____/_____, acuerda otorgar al Trabajo Fin de Grado presentado por Don_____la calificación de _____.

ACKNOWLEDGMENTS

Quisiera agradecer a varias entidades y personas la ayuda que me ha sido prestada durante el desarrollo de este trabajo de fin de grado en ingeniería biomédica. Entre ellas y, en primer lugar, a mis directores, tanto del proyecto como de las prácticas previas, Alberto Gil de La Fuente, Abraham Otero Quintana y Sergio Saugar García, por todo lo que me han enseñado y transmitido, antes y durante el desarrollo de este proyecto.

Gracias a todos los profesores del Grado de Ingeniería Biomédica de la Universidad San Pablo CEU, por su exigencia, cercanía y empatía. Os voy a echar mucho de menos. Gracias especialmente a Cristina Sánchez, por escucharme y aconsejarme sabiamente en todo momento.

Gracias a la Universidad San Pablo CEU por la oportunidad de poder realizar este proyecto junto con un laboratorio tan actualizado e innovador como es el CEMBIO. Entre los miembros del mismo gracias a Joanna Godzień por estar siempre dispuesta a resolverme dudas y con una sonrisa de oreja a oreja.

Gracias a Charito, mi compi de laboratorio (y de penas, a veces) por aguantar todos mis berrinches y hacerme reír en momentos que realmente lo he necesitado.

A Miguel por haber soportado mis numerosos dramas y crisis, no sólo durante este proyecto, sino en toda la carrera, y por esforzase siempre en ayudarme a encontrar una solución.

Por último, les quiero agradecer a mis queridísimos padres el inmenso esfuerzo que han hecho para que mi educación sea lo mejor posible. Gracias por cuidarnos a todos tan bien, siempre anteponiendo nuestras necesidades a las vuestras. Os quiero mucho.

ABSTRACT

Metabolomics is a sub-field of molecular biology that involves the analysis of small molecules (> 1000 Da), known as metabolites. The metabolites are the substrates, intermediates and products of metabolism. This field has been used for studying several diseases such as cystic fibrosis, cancer, central nervous system diseases, diabetes and cardiac disease. Metabolomics may lead to more accurate biomarkers discovery, which will help in diagnosis, prevention and monitorisation of the risk of disease, among other purposes. Within the metabolomics work-flow, the identification of metabolites is the main bottleneck in metabolomic studies, which consists in the recognition of the metabolites analysed in the metabolomics instrumentation.

Mass spectrometry (MS) is the principal spectrometric method employed for metabolite detection due to its high sensitivity. Tandem mass spectrometry (MS^n) is the application of two or more stages of MS analysis. The identification of metabolites from their MS^2 (MS/MS) data is a key task in metabolomics. This is not a trivial task since the metabolites have different chemical and physical properties that makes them difficult to identify using only MS^1 data. These diversities cause the formation of different alterations over the metabolites, like adducts, fragments, multimers and/or multiple charged ions. Some alterations can be produced in the ion source of the MS, where the molecules can be affected before entering the mass spectrometer. These alterations difficult the identification of the compounds, and can lead to the misidentifications of them.

In this project we want to create an automatic search for in-source fragments based on the complete spectrum data of the experiment, as well as an algorithm to perform MS/MS based searches.

RESUMEN

La metabolómica es una sub-área de la biología molecular que tiene como objetivo el análisis de moléculas de pequeño tamaño (<1000 Da) denominadas metabolitos. Los metabolitos son los sustratos, moléculas intermediarias y productos del metabolismo. Esta ciencia ha sido utilizada para el estudio de distintas enfermedades, como la fibrosis quística, el cáncer, enfermedades del sistema nervioso central, la diabetes y las enfermedades cardíacas. El uso de la metabolómica puede conducir al descubrimiento de biomarcadores más precisos y así ayudar a diagnosticar, prevenir y controlar el riesgo de enfermedades, entre otros propósitos. El principal cuello de botella en los estudios metabolómicos es la identificación de los metabolitos, que consiste en el reconocimiento de los metabolitos analizados mediante la instrumentación metabolómica.

La espectrometría de masas (MS) es la principal técnica espectrométrica utilizada para la detección de metabolitos debido a su alta sensibilidad. La espectrometría de masas en tándem (MS^n) es la aplicación de dos o más etapas de análisis de MS. La identificación de metabolitos a partir de sus datos obtenidos por MS^2 es una tarea clave en metabolómica. Esta identificación no es una tarea trivial, pues los metabolitos tienen diferentes propiedades químicas y físicas que hacen que puedan sufrir alteraciones, y que se formen aductos, fragmentos, multímeros e iones con múltiples cargas. Estas alteraciones se pueden producir en la fuente de ionización del espectrómetro de masas, dónde las moléculas de la muestra pueden verse afectadas antes de entrar en el analizador de masas. Estas alteraciones dificultan la identificación de metabolitos y pueden dar lugar a la incorrecta identificación de algunos de los mismos.

En este proyecto, el objetivo es la creación de una búsqueda automática para fragmentos formados en la fuente de ionización a partir de búsqueda de masas en bloque y la creación de un algoritmo para la identificación de metabolitos a partir de la información de MS/MS.

INDEX

FIGURE INDEX

TABLE INDEX

EQUATION INDEX

INTRODUCTION

1.1 Omics' sciences

The ability to study biological systems at a cellular and molecular level has been transformed in the past two decades due to the development of omics' sciences [1]. The word omics refers to a biological science field ending with -omics. This suffix means pertaining to a comprehensive field of study while the ending -ome means the complete set of the molecules belonging to the corresponding field. The most relevant omics' sciences are genomics, transcriptomics, proteomics and metabolomics, which study the genome, transcriptome, proteome and metabolome, respectively. They are discussed individually hereunder [2], [3]:

Genomics: science that studies the DNA molecules. It studies the genome, which can be defined as the complete set of genes inside a cell [4]. Its goal is to quantify and characterise those genes, which direct the production of proteins. This science determines what an organism can potentially do.

Transcriptomics: science that studies the set of all messenger RNA molecules in a biological sample, looking into gene expression patterns [2]. Since mRNA is used as a pattern to synthetise proteins, transcriptomics studies what a cell is planning to do.

Proteomics: science that studies the structural and functional role of every protein within a cell, dynamic protein products and their interactions. The proteome is highly dynamic and it changes in response to different stimuli [3]. Therefore, proteomics represents what makes things to happen within the cell.

Metabolomics: science that studies small sized molecules known as metabolites. These molecules are the compounds that participate in cell's metabolism. It represents what has happened and what is happening in the cell [5].

The interrelations and interactions between these four sciences is known as the omics' cascade (see Figure 1). Overall, the objective of omics' sciences is to identify, characterise, and quantify all biological molecules involved in the structure, function, and dynamics of a cell, tissue, or organism [2].

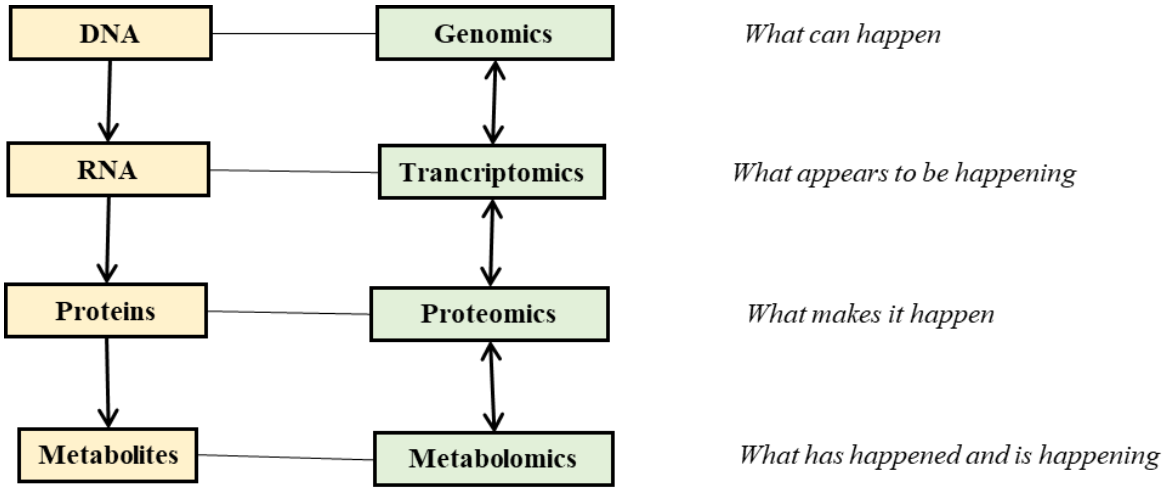


Figure 1 Omics' cascade.

1.1.1 Metabolomics

Metabolomics is a sub-field of molecular biology that involves the analysis of small molecules with molecular weights less than 1000 Da, present in biological samples (urine, blood, tissues etc.). These small molecules, known as metabolites, are the substrates, intermediates and products involved in metabolism [5], [6]. The set of metabolites is known as metabolome, initially defined in the late 1990s, which is often divided into the exometabolome (represents metabolites outside the cell) and the endometabolome (represents intracellular metabolites) [7].

Metabolomics is the most recent of the omics sciences (genomics, transcriptomics and proteomics) and it provides a greater amount of information than these other techniques since their study is focused on a specific chemical compound (deoxyribonucleic acid, ribonucleic acid or proteins, respectively) whilst metabolomic studies the abundance of a wide variety of chemical compounds. The study of the metabolome has some advantages over these other -omics sciences. The metabolome is more tractable (generally, a ten-fold difference in the number of metabolites *vs* genes), for example, S. cerevisiae has 584 metabolites and more than 6600 genes [9]. In addition, the number of approved tests arising from metabolomics nowadays are 334 whilst approved tests using genomics are 108, followed by 5 from transcriptomics and 0 tests have been approved from proteomics at the moment [8]. Moreover, since the metabolome is the final downstream product of the genome, it is closest to the function

or the phenotype of the cell. Furthermore, the metabolome is a high-throughput strategy and its costs per analysis are lower than the transcriptomics and proteomics experiments. Nevertheless, high specification analytical instruments are expensive for all of them [5], [12]. Although the number of metabolites is often less than the number of genes or proteins, their chemical composition is much more varied, which poses a challenge in their identification.

This science is a rapidly growing field. The number of papers written about metabolomics has been rising in recent years. Since technologies and analysis mechanisms continue improving, it is expected that the number of metabolomic studies will keep increasing. The metabolic mechanism behind biological processes and human diseases is more understandable thanks to metabolomics. Its main goal is to comprehend the overall metabolism changes under different conditions. Metabolomics' field has been used for studying several diseases such as cystic fibrosis [13], cancer [14], central nervous system diseases [15], diabetes [16] and cardiac disease [17]. Metabolomics usage may lead to more accurate biomarkers discovery, which will help in diagnosis, prevention and monitorisation of the risk of disease [11].

The metabolomic investigations can be classified in two approaches that are complementary: untargeted and targeted. The first one, also known as "global metabolomics", can be defined as: "The dynamic, qualitative and quantitative analysis of all small molecules (< 1000 Da) in a cell-type, tissue, body fluid or organism" [12]. Untargeted metabolomics refers to a global analysis of the metabolic changes in response to an event (for example, a disease or environmental factors). Otherwise, when the study is focused on the analysis of individual groups of metabolites related to a specific metabolic pathway or a specific class of compounds, we talk about targeted metabolomics. Both approaches differ in many aspects, including the level of quantitation and experimental accuracy, but the major difference among them is that, in targeted analysis, the target metabolites are known before data acquisition starts. Therefore, the specific metabolites are quantified whilst untargeted experiments are carried out for hypothesis generation (there is no previous knowledge) and include relative quantification of all the observed metabolites. Consequently, untargeted

metabolomics provides a huge amount of information since it studies the abundance of a considerable number of chemical compounds within a sample [5], [19].

Whether an untargeted or targeted approach to metabolomics is taken, the fundamental work-flow is essentially similar (see Figure 2). In general, there are seven steps in the metabolomics work-flow. These include experimental design and sample collection, sample preparation, metabolomics analysis, feature generation, statistical analysis (in clinical trials), metabolite identification and biological pathway interpretation. The experimental design must be carefully planned, and the subsequent step is to prepare the sample for metabolomics analysis. The metabolomics analysis allows obtaining metabolite fingerprints by different analytical platforms such as LC-MS (Liquid Chromatography – Mass Spectrometry), GC-MS (Gas Chromatography – Mass Spectrometry), CE-MS (Capillary Electrophoresis – Mass Spectrometry) and/or NMR (Nuclear Magnetic resonance). The fourth step is the use of metabolomic software tools for automatic peak detection, alignment and features (retention time and *m/z* pairs) generation. Several software tools are available, either public or commercial. The fifth step is the statistical analysis, which narrows down the features to study since it allows the user to focus on the statistically different features between the experimental and the control group. It only takes place in clinical trials, where the metabolism of two or more different groups is compared. The sixth is the metabolite identification, and the last one is the biological interpretation [19], [20].

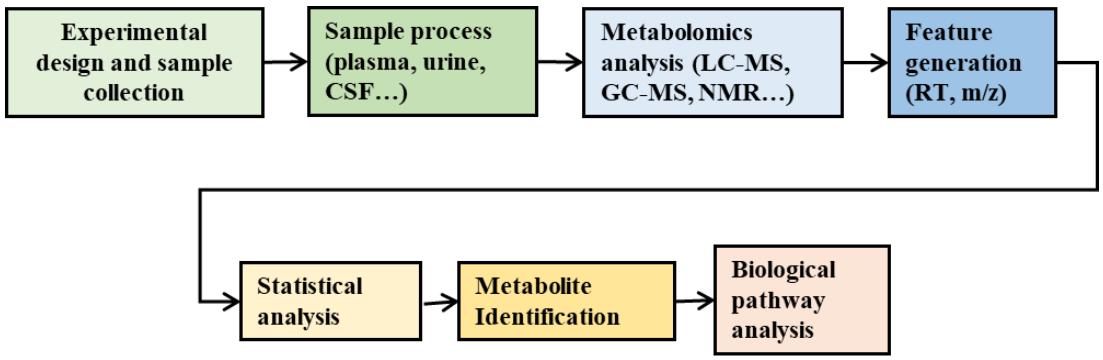


Figure 2 Untargeted metabolomics work-flow [20].

In this project, only the untargeted metabolomics will be considered. To facilitate the work-flow's understanding, Figure 3 illustrates a more computationally

oriented pipeline. The methodological pipeline starts with the spectral data processing to generate the metabolic features. Once the features are generated, some analysis methods can be applied to the complete set of features to build models that aims to describe the observed data. Moreover, after features' generation, metabolite identification can be performed by querying the experimental masses of the features against metabolomic databases. Finally, with the obtained models and the putative identifications, a biological interpretation can be performed [52].

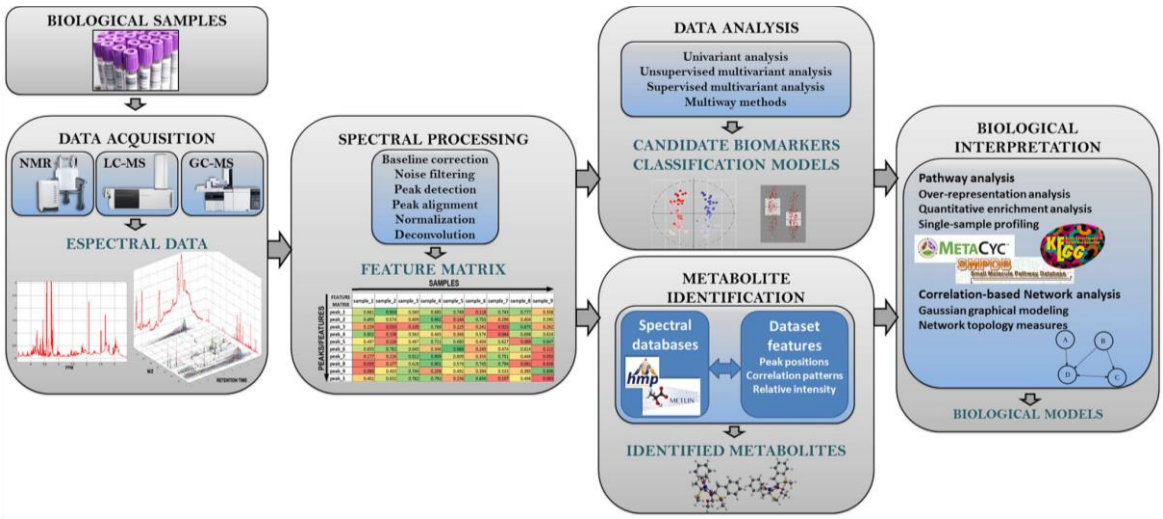


Figure 3 Untargeted metabolomic pipeline [52].

The metabolite identification represents the major bottleneck for untargeted metabolomic studies (see Figure 4) [21]. Metabolites' experimental masses (EM) obtained by mass spectrometry (MS) are searched against multiple metabolomic databases. During this identification is necessary the study of the *m/z* detected by MS, the isotopic profile, the charge state, the ion adduct formed and the retention time. Nowadays there are different software tools that facilitate and automate this process, some of them combining different metabolomic databases. Due to the development of these tools and the database sharing, in a near future it is expected that the bottleneck of metabolite identification in metabolomics will be gradually solved [5], [20].

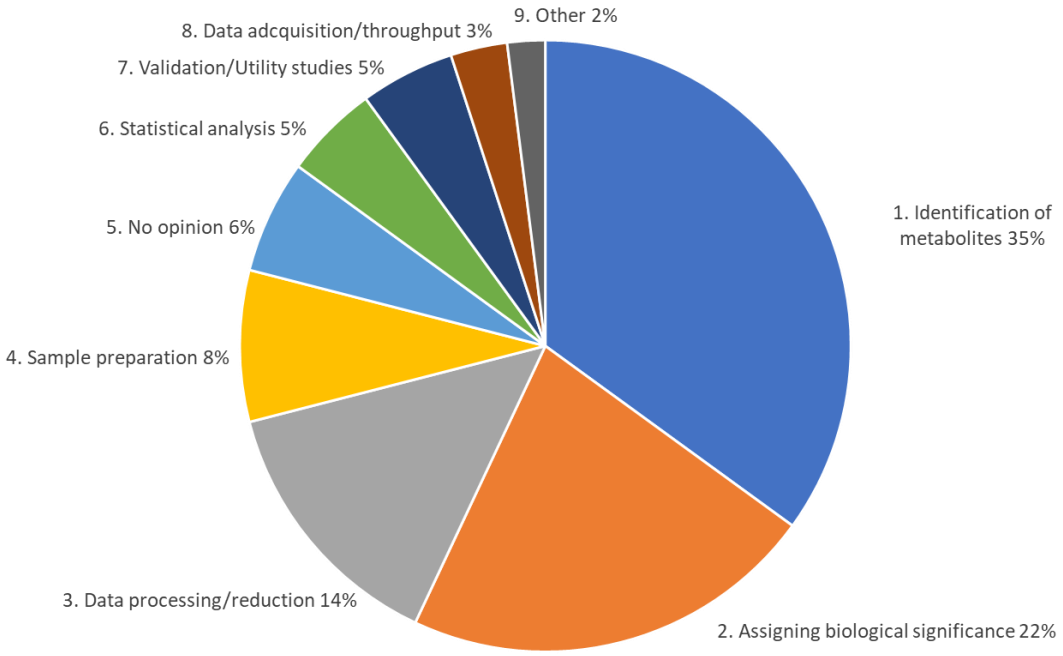


Figure 4 Time percentages for each task in a metabolomic study [21].

1.2 Mass spectrometry

Due to the complexity and the dynamic nature of the metabolome, multiple analytical platforms are needed to cover the full spectra of metabolites. The most common spectrometric systems used in metabolomics are mass spectrometry, nuclear magnetic resonance spectroscopy, and Fourier transform infrared (FT-IR) spectroscopy [19],[22].Within these three, NMR spectroscopy and MS have evolved as the most popular techniques in metabolomic studies [26].

Nuclear Magnetic resonance is a powerful and very useful technique for structure characterisation in untargeted studies. It has been applied for metabolites' analysis in biological fluids and cell extracts. NMR can provide a rapid analysis time and is more reproducible than MS. Nevertheless, in some cases the information provided by NMR spectroscopy is not enough to fully characterise a metabolite, and it only detects metabolites with a mass between 200 μg and 500 mg. On the other hand, MS advantages over NMR are its higher sensitivity, the fact that it can detect metabolites with a molecular weight from 1 till 100pg (10^{-12} g), its higher selectivity, and its applicability to either targeted and untargeted approaches. This technique's

sensitivity difference (10^{-6} units) makes MS applicable over a broader range of metabolites than NMR [5], [7], [26], [27]. Because of this, CEMBIO's laboratory, the principal client from this project, uses MS. CEMBIO is the Metabolomics and Bioanalysis Excellence Center (*Centro de Excelencia Metabolómica y Bioanálisis*, http://www.metabolomica.uspceu.es/) from San Pablo CEU University. Therefore, this project is based on the metabolite identification from MS data.

Mass spectrometer was invented over 100 years ago and it has provided new capabilities to determine atomic and molecular masses since then [12]. Today, the mass spectrometry is a key technology for the metabolomics research. Over the last years, the enormous technological advances in MS have led to a fast increase in the amount and complexity of the data produced [28]. Compared to proteomics, where peptides have a different length of possible combinations of a set of 20 amino acids arranged linearly, the metabolites are combinations of a diverse set of elements (e.g., C, H, O, S, N, and P). They have chemical and physical diversities that makes them difficult to identify based on MS data [22], including the following:

- Isotopes: atoms with the same number of protons but a different number of neutrons. Therefore, they have different atomic weights. Isotopes are different forms of the same element [23].

- Adducts: products formed by the direct union of two molecules. In metabolomics, molecules can bind metabolites forming adducts. For LC-MS, the metabolites need to be charged in order to be detected by the device. For this reason, an ionisation source is used and, consequently, the adducts are formed. The most common adducts are generated by addition ($[M+H]^+$) or subtraction ($[M-H]^-$) of a proton. Nevertheless, other adducts such as $[M+Na]^+$, $[M+K]^+$ and $[M+Cl]^-$ are common in biological samples [5].

- Fragments: molecules that arises from a precursor molecule breaking. During the ionisation, it is possible that some molecule's weak bonds break. Therefore, the resulting spectrum will be composed by the metabolite's *m/z* and the corresponding fragments' *m/z* [24].

- Multimers: metabolites conformed by several molecules of the original metabolite. Multimers may be formed when the sample's concentration is very high. Sometimes they present neutral losses. Neutral losses occur when a molecule's bond is broken and it loses a determined molecule (e.g. water – H_2O). In this case, their masses do not exactly correspond to 2x or 3x the original metabolite's mass [25].

- Multiple ion charges may be produced during ionisation process (n= 2, 3, …). The corresponding mass spectrum will have peaks in m/2, m/3, … With m as the original metabolite's mass [5].

1.2.1 The mass spectrometer

All the mass spectrometers have an ion source, a mass analyser and an ion detector (see Figure 5). The ion source (see the different options Table 1) converts the analyte molecules into ions. This ionisation is necessary since mass spectrometers use electric and magnetic fields and neutral molecules do not respond to them [29]. Next, the mass analyser (see the different options in Table 2) separates the ionised analytes based on their mass and charge (*m/z*) ratio. Then, the detector records the number of ions emerging from the analyser at each *m/z* value [30]. Finally, the information is transmitted to a computer and presented through a mass spectrum, which provides a plot of ion abundance *vs* mass to charge ratio (*m/z*).

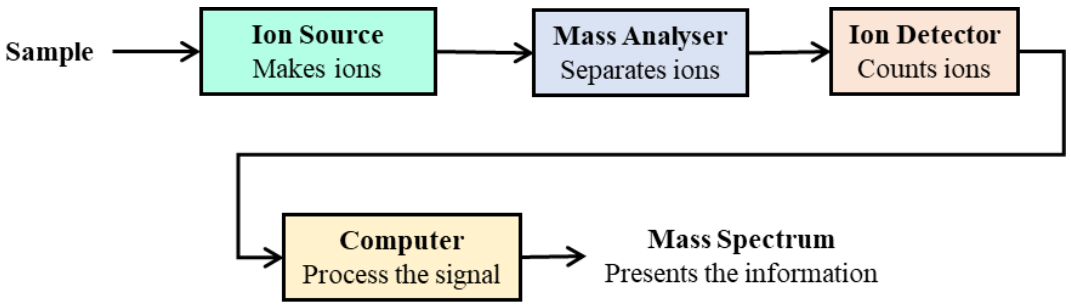


Figure 5 Mass spectrometer basic structure.

Table 1 Range of ion sources currently available [12].

Ion Sources	
Atmospheric pressure chemical ionisation (APCI)	Employed in LC–MS applications for ionisation of polar and semi-polar metabolites at atmospheric pressures. Operates by passing the LC eluent through a heated glass tube (up to 500 ◦ C) which produces explosive vaporisation to generate a gaseous collection of solvent and metabolite molecules. A corona discharge needle provides ionisation of the solvent molecules (which are in a large excess) followed by ion or charge transfer to metabolite molecules. Degradation of molecules can occur at the high temperatures used, though minimal fragmentation of molecular ions is observed. The technique is described as soft-ionisation
Atmospheric pressure photo ionisation (APPI)	Employed in LC–MS applications for ionisation of non-polar metabolites at atmospheric pressures and acts as a complementary tool to ESI and APCI. Not commonly available from many instrument manufacturers. Electron ejection from molecules is produced by photons emitted from discharge UV lamps.
Chemical ionisation (CI)	Employed in GC–MS applications for ionisation of polar and non-polar metabolites at vacuum pressures. Chemical reagent gases (methane, ammonia and others) are introduced into an EI source, in large excess to metabolite molecules, and are ionised by electron bombardment. Ion or charge transfer creates ionisation of metabolites. This is a soft-ionisation technique producing minimal fragmentation of the molecular ion, tandem mass spectrometry can be used to dissociate the molecular ion.
Electron impact (EI)	Employed in GC–MS applications for ionisation of polar and non-polar metabolites at vacuum pressures. Bombardment of the metabolite molecules by electrons creates positively charged ions. The process imparts significant energy to the molecular ion which is released by covalent bond fission to produce a mass spectrum which is characteristic of the metabolite's chemical structure. Fragmentation patterns are highly reproducible and can be used to deduce metabolite structure. Libraries of these mass spectra are also available for the identification of metabolites.
Electrospray ionisation (ESI)	Operates by passing the LC eluent through a capillary held at high voltages (2–5 kV). Ion formation occurs in the capillary followed by nebulisation and desolvation to provide transfer of ions from liquid to gas phase and introduction to the mass spectrometer vacuum. The source acts as an electrochemical cell. Minimal fragmentation of the molecular ion is observed, fragmentation can be produced in-source or by tandem mass spectrometry. The technique is described as soft-ionisation.
Matrix-assisted laser desorption/ionisation (MALDI)	In MALDI, a laser pulse is fired at a solid analyte that has been co-crystallised with a chemical matrix. This creates a vapor plume of analytes, matrix, and their ions. The matrix role is to protect the biomolecule from being destroyed by the laser beam and to facilitate vaporisation and ionisation. The advantages of this technique are its high speed, robustness, and relative immunity to contaminants and biochemical buffers. [31]

Table 2 Range of mass analysers currently available [12].

Mass Analysers	
Fourier transform ion cyclotron resonance (FTICR)	Expensive mass analyser that attains the highest mass resolution (R > 100 000) and mass accuracy (less than 1 ppm) currently available. Ions orbit in a cell operating at ultra-high vacuums (10^{-10} atmospheres) and in a high magnetic field strength (>7 Tesla). The orbital frequency, dependent on m/z, is detected as an image current and is converted from time to frequency domains with a Fourier transform.
Linear quadrupole (Q)	A low mass resolution and cheap mass analyser. DC and RF potentials are applied to four parallel rods, each opposite pair being electrically connected. These potentials are varied so as to provide a mass filter where ions of a chosen mass have trajectories of amplitude less than half the radius of the quadrupoles and traverse the mass analyser whereas ions of lower and higher mass are lost by

	collisions with the rods because their amplitudes are too great. The DC and RF potentials are varied (while keeping their ratio constant) so as to provide a mass scan where ions of increasing mass traverse the mass analyser.
Orbitrap	A new introduction which operates in a similar manner to FTICR, though without the requirement of high magnetic strength fields and at a lower purchase cost. Another high mass resolution (up to 100 000) and mass accuracy (<1 ppm) instrument. A coaxial inner spindle-like electrode is surrounded by two outer barrel-like electrodes with an electric field applied between electrodes. Ions orbit the central electrode in axial and radial directions, through a balance of electrostatic and centrifugal forces and the orbital frequency is detected as an image current by the outer electrodes. A Fourier transform is employed to convert it from the time to frequency domain. The instrument is marketed as a hybrid instrument, a linear ion trap is coupled to the Orbitrap so as to collect ions before periodic introduction into the Orbitrap mass analyser.
Quadrupole ion trap (QIT) and linear ion trap (LIT)	A low mass resolution analyser that provides ion storage which allows MS^n experiments to be performed to provide structural information. The QIT is constructed of a ring and two cap electrodes and operates with the application of DC and RF potentials to constrict ions in stable oscillatory orbits. A helium bath gas is used to stabilise ion populations. Detection occurs by destabilisation of orbits and ion ejection to a detector. Linear ion traps apply a 2D quadrupole field rather than 3D field as applied in QIT. These are physically larger and allow trapping of larger ion populations providing greater sensitivity.
Quadrupole-time of flight (Q-TOF)	A hybrid instrument allowing tandem mass spectrometry experiments with detection of product ions performed at high mass resolutions and accuracies. Constructed of quadrupole and TOF mass analysers separated by a higher-pressure collision cell used to provide collisionally induced dissociation (CID) of selected ions.
Time of flight (TOF)	A simple system which measures the flight time from source to detector in a vacuum, which is dependent on *m/z*, lower masses are detected first. The use of ion mirrors (reflectrons) focuses ions of the same *m/z* but different kinetic energies which increases mass resolution (4000–15 000 FWHM) and allows accurate mass measurements.
Triple quadrupole (QQQ)	Mass analyser commonly applied for MS/MS experiments for structural characterisation or selective and sensitive targeted analysis of limited number of metabolites. Two quadrupoles (Q1 and Q3) are separated by a quadrupole (collision cell) operating at a higher pressure. Ions accelerated from Q1 to collision cell undergo collision-induced dissociation (CID) followed by mass analysis in Q3. A range of MS/MS experiments are possible including product ion scanning, neutral loss scanning and single/multiple reaction monitoring (SRM/MRM).

1.2.2 Tandem mass spectrometry

Tandem mass spectrometry, abbreviated MS^n, performs more than one stage of MS analysis, applying a voltage to the molecules of the sample and fragmenting them. This fragmentation aids to the identification of the metabolites in untargeted approaches. The identification of metabolites from their MS^2 (also named MS/MS) data is a key task in metabolomics. Most commonly in MS/MS, a first analyser is used to isolate a precursor ion, which then undergoes a fragmentation to yield product ions and neutral fragments (see Figure 6). The product ions are analysed by the second spectrometer [32].The result is a mass spectrum of the molecule and it provides information about how the molecule is fragmented.

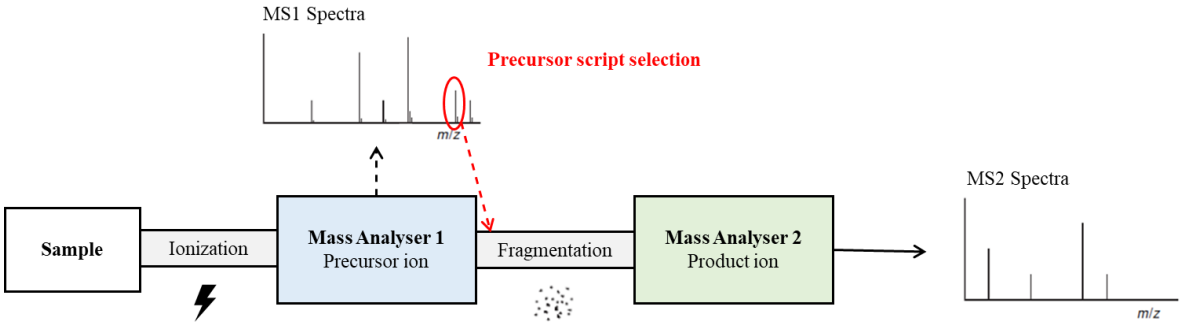
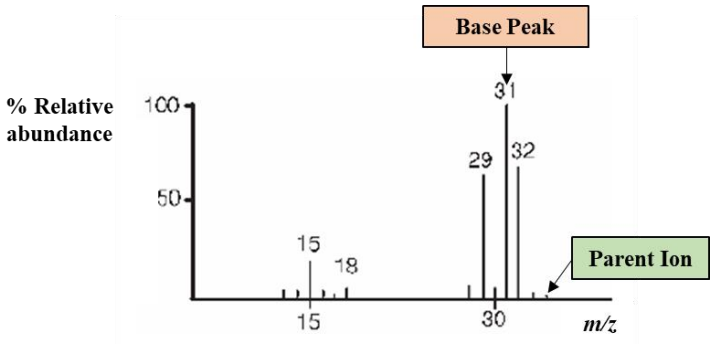


Figure 6 MS/MS general scheme.

The equipment provides as a result a set of ion abundances, also referred as intensities, *vs* mass to charge (*m/z*) ratio for each of the produced ions (mass spectrum). This data can be plotted having in the x-axis the *m/z* and in the y-axis the intensities, which are expressed as relative (from 0 to 100%) or represented in a table (see Figure 7). Each pair of *m/z* and intensity is known as a peak. The most intense peak is called the base peak. The precursor ion is the peak representing the metabolite of interest after it has been ionised. This peak is the highest *m/z* in the spectrum and it does not always appear due to the fragmentation during the procedure.



m/z	Relative abundance (%)	*m/z*	Relative abundance (%)
12	0.33	28	6.3
13	0.72	29	64
14	2.4	30	3.8
15	13	31	100
16	0.21	32	66
17	1.0	33	0.73
18	0.9	34	~ 0.1

Figure 7 Mass spectrum as a plot and as a table [32].

1.2.3 Separation techniques

MS coupled to a range of diverse chromatographic platforms is commonly used for untargeted metabolomic studies. As previously mentioned in the section 1.2, MS use is broader than NMR (see Figure 8). The chromatographic platforms aim to separate the different components of the sample and can be gas chromatography (GC–MS), liquid chromatography (LC–MS), ultra-performance liquid chromatography (UPLC–MS, also referred to as ultra-high-performance liquid chromatography—UHPLC–MS), and capillary electrophoresis (CE-MS) [33]. Separation techniques improve MS analytical performance since they spread the complexity of the sample over the time. Figure 8 shows that liquid chromatography is the most applied chromatography-MS tool. The main reason is that it provides a wide range of metabolome information requiring a minimum sample pre-treatment [34].

In Gas Chromatography (GC) the sample to be analysed may be a liquid solution or a collection of molecules adsorbed on a surface. A basic prerequisite for GC analysis is the transfer of the analytes into the gas phase without a thermal decomposition or rearrangement. During the transfer into the GC, the sample is volatilised by rapid exposure to a zone of relatively high temperature (200-300°C) and mixed with a stream of carrier gas (Ar, He, N2, or H2). GC-MS is an indispensable tool in metabolomics. It is limited to volatile, thermally stable, and energetically stable compounds [38].

Liquid Chromatography (LC) is an analytical chromatographic technique useful for separating ions or molecules that are dissolved in a solvent. It requires taking a sample and injecting it into the instrument, the solvent and analytes must be in liquid state. Usually the LC used in LC-MS is HPLC, an analytical technique that couples high-resolution chromatographic separation with sensitive and specific mass spectrum detection. This is the most employed technique due to its high versatility without losing too much reproducibility, high resolution and high mass accuracy (see Table 3). It allows the characterisation of almost every compound type according to the sample preparation, the ion source configuration and the separation column. GC-MS and LC-MS can be complementary techniques and detect complementary sets of metabolites [39], [40].

Capillary Electrophoresis (CE) is a family of related techniques that employs narrow (20-200 µm) capillaries to separate large and small molecules. These separations are facilitated using high voltages that generate electroosmotic and electrophoretic flow of buffer solutions and ionic species, respectively, within the capillary [41].

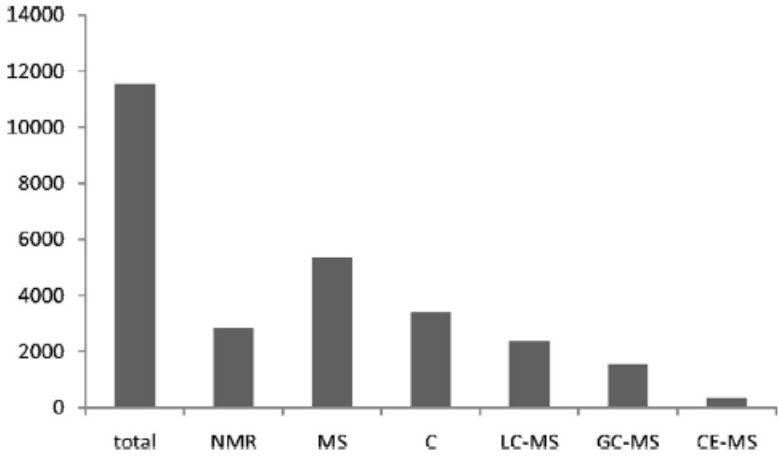


Figure 8 Metabolomics papers searched in the Web of Science on May 27, 2014 from [34].

A chromatogram is a plot containing 3-dimensional data consisting on a distribution of mass-to-charge ratio (*m/z*) and intensities over the time (see Figure 9), starting at the time of sample's injection [32], [35]. Therefore, a chromatogram describes an analyte by the retention time (RT) (time from injection to elution), the peak width and the area under the peak or the peak height [32].

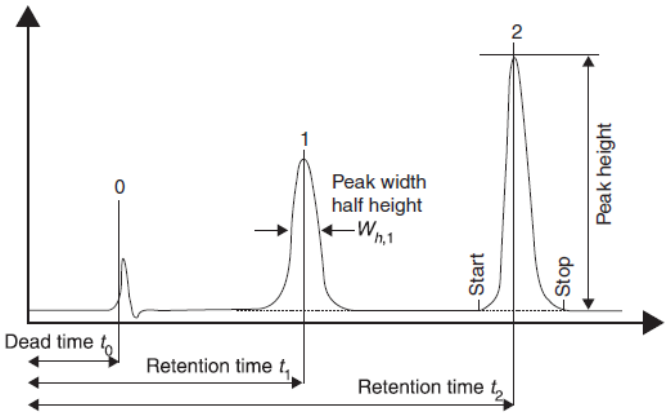


Figure 9 Simple Chromatogram. Time vs intensity [7].

Every separation technique platform has its advantages and limitations. There have been recent technological advances in all of them as well as in the mass spectrometers, but no platform is the perfect tool for untargeted metabolomics. All have their strengths and weakness (see Table 3) [35].

Table 3 Comparison among different separation techniques coupled with MS. Three stars means that the technique is adequate. A single star means that it is not adequate [7].

		LC-MS	GC-MS	CE-MS
Metabolite chemistry	High LogP	**	*	*
	Low LogP	**	***	***
	Negative charge	***	***	**
	Positive charge	**	**	***
	m/z < 80	**	***	*
	m/z > 80	***	***	**
Sample type	Tissue	**	**	**
	Bio-fluids	***	***	***
	Cell culture	**	**	**
Metabolomic approach	Targeted	**	***	*
	Untargeted	***	**	**
Analytical specifications	Sample preparation	***	*	***
	Throughput	**	**	**
	MS mass accuracy/ resolution	**	*	*
	Inter-day reproducibility	**	***	*
	Nº metabolites features	***	**	**
Data	Databases	**	***	**
	Data analysis	**	**	**
	Information	**	**	**

1.2.4 Metabolite identification

Metabolite identification remains as one of the major bottlenecks in metabolomic studies. In untargeted metabolomics, metabolite identification is achieved through multiple mass-based search in metabolomic databases followed by manual verification. The chemical analyst starts searching the *m/z* value of a molecule of interest against one or several metabolomic databases within an established tolerance.

In tandem spectrometry, the identification can be performed using the mass of the precursor ion (the mass of the molecule obtained through MS) and/or the fragmentation pattern (MS/MS peaks). The use of the precursor ion mass restricts the searches to the metabolites with a mass that matches it. However, sometimes the precursor ion mass has suffered any type of alteration and the researcher may be

interested in comparing only the fragmentation pattern, ignoring the mass of the precursor ion.

In CEMBIO, the typical mass spectrometry identification work-flow is as follows: by applying Mass Spectrometry (MS^1) over a sample, for example a blood sample, a set of precursor masses is obtained. If the mass spectrometer is coupled to a separation technique instrument (for example LC-MS), a set of precursor masses, their corresponding intensities and RT pairs are obtained. These 3-dimensional data are known as features in metabolomics. The precursor masses allow obtaining a set of putative annotations. Imagine that, for example, in the metabolomic study, they are interested in detecting Asparagine in the blood and within the masses provided by the MS^1 there is 132.0576, which can be from Asparagine, but it can also belong to other metabolites (Atropaldehyde, D-Asparagine, Ornithine, Glycyl-glycine...). The *m/z* of interest from the MS^1 (132.0576 in our example) is subjected to MS^2 obtaining the corresponding MS/MS spectra. From the MS/MS of the *m/z* of interest two paths can be followed: obtaining tentative identifications or reaching a true identification. Obtaining tentative identifications consists on performing spectral matching of the input MS/MS against databases spectra (see Figure 10). A true identification implies comparing the MS/MS from a reference standard obtained under the same analytical conditions as the MS/MS of the molecule of interest in the sample [36]. In other words, the metabolites of the putative identifications are subjected to MS/MS fragmentation and compared against the MS^2 of interest. Sometimes, before performing tandem mass spectrometry, the features are filtered according to the biological knowledge (e.g. the ontology) reducing the complexity of the data. For example, if one of the putative identifications is a plant's metabolite, we will discard that theoretical identification since the sample analysed belongs to a human being. Regardless of the path followed, the MS/MS spectrum from the molecule of interest is compared manually by visual inspection against the spectra of the putative annotations to confirm the identity of the molecules. Nevertheless, the putative identifications are not usually unique and several identifications for the same unknown metabolite are often obtained [5], [7], [19], [27], [29].

This process can be automated, saving a great amount of time to the chemical analyst. The true identification supposes a considerable amount of time and money for

comparing a list of reference standards. Instead of performing the MS/MS analysis of all the reference standards of the putative identifications, obtaining a list of tentative identifications against spectra databases can improve the efficiency of the study. There exist several databases with experimental and predicted MS/MS data of different compounds. Many of these spectra libraries are publicly available [19]. The spectrum from the molecule of interest can be compared against the spectra from the databases instead of performing MS/MS over the putative annotations. The spectra having a great similarity will be subjected to MS/MS to obtain the true identification if the user is interested in providing a higher confidence. Different tools exist with the capability of performing spectral matching (HMDB, MetFrag, Metlin…), but they are not easy to integrate in CEMBIO's analytic pipeline. Due to the integration difficulty, before this project, an automatic putative identifications' search and the subsequent manual spectral matching by visual inspection or the use of an external tool was performed. This project aims to integrate and automate both steps. The new functionality was improved and integrated in CEU Mass Mediator (CMM) tool during this project. The tool will be described in Chapter 2, and the functionality implementation details will be explained in Chapter 3.

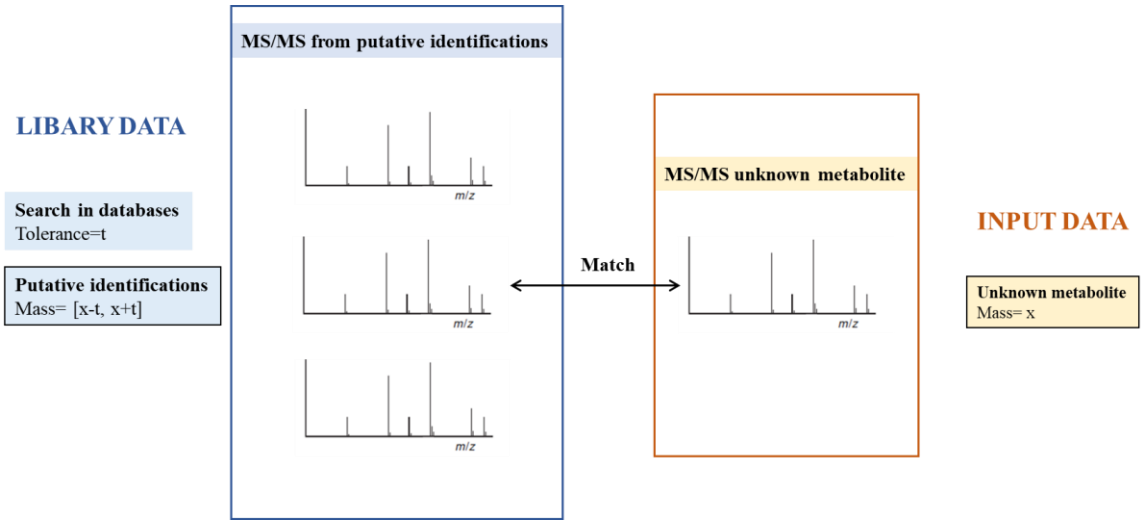


Figure 10 Spectral matching process.

1.2.5 Levels of confidence

Metabolite identification has distinct levels of confidence depending on standards availability and the information contained in databases. A confidence classification has been established by Chemical Analysis Working Group (CAWG) of the Metabolomics Standards Initiative (MSI) due to these differences. The level 0 was recently added and corresponds to compounds unequivocally identified by isolation and complete stereo-chemical characterisation. The level 1 consists on compound identification by comparing a minimum of two independent orthogonal data (e.g. RT and mass spectrum) directly to an authentic reference standard. The level 2 corresponds to putatively annotated compounds obtained by analysis of spectral data and/or similarity to data in a public database. The level 3 involves putatively characterised compounds, the data allows to place the metabolite in a class (for example, a phosphocholine). The last level, the level 4, corresponds to unknown signals [36], [37].

According to this classification, the compounds that are putatively annotated from MS and from MS/MS information correspond to the level 2. Since these two methods are not equally confident an additional level was proposed by Schrimpe-Rutledge and colleagues where the level 2 corresponds to compounds putatively identified by MS/MS data whilst the level 3 involves putative annotations from accurate mass. Both classifications are illustrated in Figure 11.

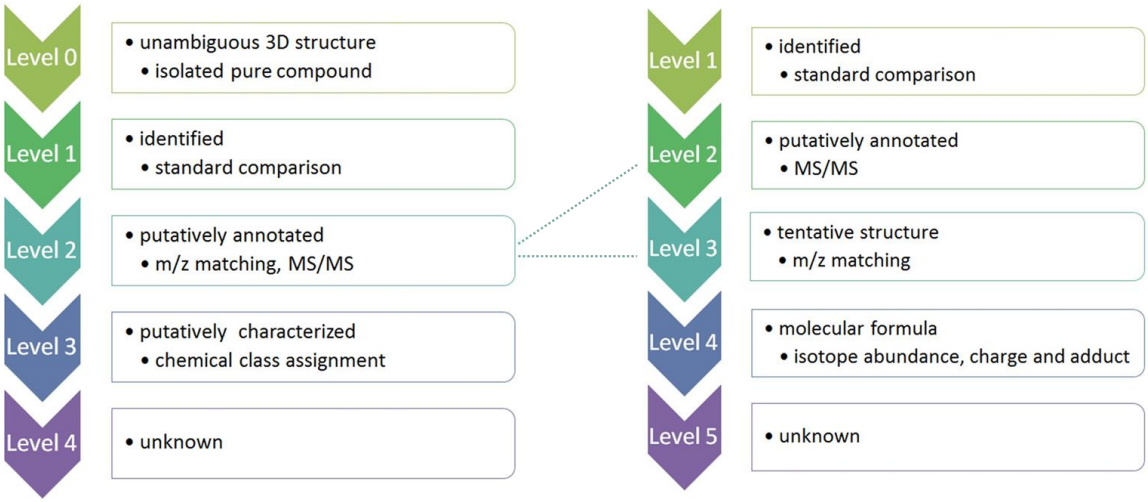


Figure 11 Confidence levels in metabolite identification [36].

1.3 Objectives

This thesis will focus on the integration of a MS/MS based search functionality previously developed, and on the development and the integration of an in-source fragment detector feature in CMM. In order to achieve these two tasks, the project has the following objectives:

1. HMDB MS/MS spectra database integration: The integration of the MS-MS spectra files from HMDB.

2. Algorithms development: Spectral matching and scoring algorithms design.

   Tasks 1 and 2 were performed before this bachelor thesis during an internship period in the biomedical engineering department of the university CEU-San Pablo.

3. MS/MS search scoring algorithm improvement.

4. MS/MS search integration in CMM: The MS/MS based search will be integrated in the CMM online web tool.

5. A new data model will be designed and implemented to group features by their RT. The current data model for the different searches (simple, advanced, batch simple, batch advanced) must be modified to achieve our goals.

6. The new data model should integrate the LC-MS searches, so a refactoring of this code is necessary.

7. Fragment search from MS/MS data: design the algorithms required to assign possible fragments within the features grouped by their RT.

1.4 Motivation

The present Biomedical Engineering bachelor thesis has arisen from the CEMBIO's interest in having a software tool capable of providing MS/MS identification support. This project aims to integrate the required functionalities into

CMM. According to the data provided by the CEMBIO, the results filtering for a single study can take one or two months of a researcher's work, depending on the analyst's work experience. The corresponding pipeline, which consists on several database querying, is a mechanical process that can be partially automated [5].

1.5 Memory structure

The following chapters compose the memory:

I. Chapter 0: a general view of the project with an introduction to metabolomics, and mass spectrometry. This chapter contains the objectives and motivation of the project.

II. Chapter 2: presents a brief state of the art regarding some of the existing metabolomic databases and software tools relevant for this project, and a description of CMM web tool.

III. Chapter 3: describes the design and integration of the functionalities in CMM; a MS/MS based search feature and an in-source fragment detector from LC-MS data. A code refactoring regarding simple and advanced CMM searches is also presented in this section.

IV. Chapter 4: presents the developed features performances and the output of the integrated functionalities. An analysis of the execution time for the LC-MS search is accomplished.

V. Chapter 5: the results of the project are discussed in this section.

VI. Chapter 6: it recapitulates the problem to be solved, analyses and discusses the results and presents some ideas about the future work.

The project described in this bachelor thesis has concluded in the implementation of a couple of required functionalities at http://ceumass.eps.uspceu.es. The generated code is accessible from https://github.com/albertogilf/ceuMassMediator.

2 STATE OF THE ART

2.1 Metabolomic databases

The metabolomic databases purpose is to organise the metabolites in a way that helps the researchers to identify and analyse the metabolites easier. The metabolite identification is performed principally through mass-based searches: researchers query the *m/z* value of a molecular ion of interest against the database/s within a tolerance range. For MS/MS data it is similar, but several peaks (*m/z* and intensity pairs) from the same precursor ion are compared instead of a single *m/z* value. When the MS data is obtained, it is compared with several metabolomic databases. This process is known as spectral matching. Various databases have been developed to help in spectral matching. Regarding to the spectral databases, the content can be experimental, which is obtained through MS/MS analysis and/or predicted, obtained by *in-silico* fragmentation [42], [43]. Some of the better-known metabolomic databases include:

- **Human Metabolome Database (HMDB):** First described in 2007. Nowadays it is considered the standard resource for human metabolomic studies. HMDB is an exhaustive, freely available web resource that contains detailed information about the human metabolome [44], [45].

- **METLIN:** is a repository for metabolites' mass spectral data. It contains over 240,000 metabolites and 72,000 high-resolution MS/MS spectra. Metabolites can be searched by compound name, mass, formula and structure. The METLIN Metabolite Database contains information from different mass analysers' results: MS/MS, LC-MS and Fourier Transform mass spectrometry (FT-MS). These data can be searched by peak lists and mass range [45].

- **MetaboLights:** Before MetaboLights was conceived, a few metabolomic databases existed. Nevertheless, they were specialised in specific domains and techniques. MetaboLights arises from the EMBL-EBI goal of obtaining a completely open-access, cross-species and cross-

technique metabolomics repository. It is a state of the art database, a repository, designed to provide a unified platform [46]. It has not MS/MS search functionality but it contains several MS/MS spectra in its database.

- **LIPIDMAPS:** created in 2003 to identify and quantify lipid species in mammalian cells and quantitate their changes in response to perturbation. This tool supports tasks such as drawing lipid structures and predicting possible structures from MS data [47], [48]. The database contains MS/MS spectra but lacks a corresponding search tool.

- **Kyoto Encyclopaedia of Genes and Genomes (KEGG):** created in 1995, it is an integrated database resource consisting of 18 databases. These databases are categorised into systems information, genomic information, chemical information and health information. Within systems information it contains metabolic pathways from a wide variety of organisms. These pathways are hyperlinked to metabolite and protein/enzyme information [45], [49]. The database does not have MS/MS data since its content is more oriented to biological interpretation rather than to metabolite identification.

- **MassBank:** is the first public repository of high-resolution mass spectra of metabolites obtained under unstandardised experimental conditions. It is a distributed database where each research group provides data from its own MassBank data servers on the Internet. Moreover, it provides free tools to prepare and manage data [45], [50]. Within these tools, there is a quick search that allows to search by peaks, both MS and MS^n searches.

Before this project, the spectra data from HMDB were imported into CMM database. The strong points of this database are its content of human metabolites (since CEMBIO works mainly with samples collected from mammals) and its open access. CMM tool will be explained in section 2.3.

The spectra information from HMDB is in XML format. Nowadays, the Human Metabolome Database is at version 4.1. Many brand-new data have been added since HMDB 3.0, large quantities of predicted metabolites, predicted NMR, GC-MS, and MS/MS as well as predicted RTs data.

Over the last 10 years, HMDB has grown from a small specific database to the largest, most exhaustive human-specific metabolomic "knowledge base" in the world. This development has been accelerated by the severe bottleneck in metabolite identification crisis. Table 4 illustrates the development experienced by HMDB from its version 1.0 until version 4.0 [44].

Table 4 Comparison between the coverage in HMDB 1.0, 2.0, 3.0 and 4.0 [44].

Category	HMDB 1.0	HMDB 2.0	HMDB 3.0	HMDB 4.0
Total number of metabolites	2180	6408	40 153	114 100
Number of detected & quantified metabolites	883	4413	16 714	18 557
Number of detected, not quantified metabolites	1297	1995	2798	3271
Number of expected metabolites	0	0	20 641	82 274
Number of predicted metabolites*	0	0	0	9548
Number of unique synonyms	27 700	43 882	199 668	1 231 398
Number of cmpds with expt. MS/MS spectra	390	799	1249	2265
Number of cmpds with expt. GC/MS spectra	0	279	1220	2544
Number of cmpds with expt. NMR spectra	385	792	1054	1494
Number of cmpds with pred. MS/MS spectra*	0	0	0	98 601
Number of cmpds with pred. GC/MS spectra*	0	0	0	26 880
Number of experimental NMR spectra	765	1580	2032	3840
Number of experimental MS/MS spectra	1180	2397	5776	22 198
Number of experimental GC/MS spectra	0	279	1763	7418
Number of predicted MS/MS spectra*	0	0	0	279 972
Number of predicted GC/MS spectra*	0	0	0	38 277
Number of metabolic pathway maps	26	58	442	25 570
Number of compounds with disease links	862	1002	3105	5498
Number of compounds with concentration data	883	4413	5027	7552
Number of predicted molecular properties	2	2	10	24
Number of metabolite-SNP interactions*	0	0	0	6777
Number of metabolite-drug interactions*	0	0	0	2497
No. of metabolites w. sex/diurnal/age variation*	0	0	0	2901
Number of metabolic reactions*	0	0	0	18,192
Number of defined ontology terms*	0	0	0	3150
Number of HMDB data fields	91	102	114	130

* New for HMDB 4.0

2.2 Metabolomic software tools

Metabolomics' field had a great expansion over the last two decades, both as experimental sciences and bioinformatics tools [51]. The metabolomics Society was established in 2004 to promote the increase, use and comprehension of metabolomics. METLIN, the first metabolomic database, was also established in 2004. The Human Metabolome Project was started on 2005 and its metabolite information is kept in the Human Metabolome Database, which produced its first sketch on 2007. There are several metabolomic databases and tools, each having different information and functionalities [11]. NMR and MS recent technological advances have led to an increase in the amount of the metabolomic data obtained from each sample. Therefore, it is necessary to have bioinformatics tools capable of dealing with such complex and voluminous data generated [52].

The most relevant bioinformatics tools for this project are those capable of identifying metabolites through their MS/MS data and those able to group features that come from the same analyte. A feature is a two-dimensional signal composed by a chromatographic peak (RT) and a mass spectrometry peak (*m/z*) [11].

Regarding the software tools capable of clustering features, we highlight CAMERA. Feature clustering consists on grouping the features that may come from the same analyte, so all of them elutes the separation column together (RT). It is a R package that integrates several algorithms for metabolites characterisation and identification. It is designed to post-process lists of features obtained experimentally (for example by LC-MS), and it performs a RT-based grouping and a graph-based grouping (peak shape analysis) over them [56]. This software aims to detect adducts from experimental data and does not perform any search against databases.

Once the features' relationships are established, grouped by their RT and with the adducts detected, the MS and MS/MS search can be performed with a new dataset that has a lower complexity.

Within those tools capable to identify metabolites through MS/MS data we outline HMDB (Human Metabolome Database). Within its features the most relevant

for us is the LC-MS/MS Search. This search allows users to submit MS/MS data and search its correspondent metabolite against a library of LC-MS/MS spectra [45].

Other tools relevant for MS/MS are MetFrag and MyCompoundID. MetFrag is a freely available web tool for metabolites annotation through MS/MS data. It matches the input spectrum against spectra obtained by *in-silico* fragmentation of the putative annotations from different databases (PubChem, KEGG, ChemSpider, MetaCyc, FOR-IDENT, LipidMaps, ChEBI, HMDB) [54]. On The other hand, MyCompoundID is a web-based resource for metabolite identification that performs spectral matching against a spectral library which contains experimental and predicted MS/MS. Within its features, we outline the MS Search, MS/MS Search and M-RT-MS/MS Search [55].

Nevertheless, even if their functionalities are useful, these tools were not easy to integrate in CMM. CAMERA's principal drawback is that its data processing is done before MS based search. We aim to establish the fragment relationship among features after the peak detection and the features clusterisation to check if some features can be fragments of some others with a greater mass within the same RT. Moreover, CAMERA's input is the mzXML files from the source whilst in CMM the input data are the detected and previously clusterised features. Therefore, the use of CAMERA is previous to the metabolite annotation, and its feature processing is useful for the annotation performed by CMM.

Regarding HMDB, its main disadvantage is the lack of an API (Application Programming Interface) in order to be able to use from CMM, such as MyCompoundID. Moreover, this last one is out of date, its last update was on 2013. Finally, MetFrag's principal inconvenience is its identification functionality from fragmentation algorithms while we aim to perform metabolite's annotations against experimental or *in-silico* data, depending on the user's choice.

2.3 CEU Mass Mediator

CMM is an online tool that provides a unified search in the metabolomic databases HMDB, Metlin, KEGG, LipidMaps, and MINE. This mediator system has been designed to help researchers when performing metabolite annotation and arises

from the need of the CEMBIO to have a tool to perform the identification from MS data [57], [58].

The unified access to metabolomic databases allows the chemical analyst to abstract from how the information is structured within each database. This time-saving access supposes a considerable optimisation in the time of the compounds identification [57].

The available features in CMM are visible on its web interface (see Figure 12). The different functionalities are explained hereunder:

- **Simple search:** allows the user to search a metabolite from a single EM in Daltons (Da) that can be neutral or *m/z*. A search tolerance must be given in ppm or mDa. The user must specify the databases, metabolite's type, ionisation mode and adducts over which the search is performed.

- **Advanced search:** the search is performed over the *m/z* or the neutral mass including some extra information that allows filtering results, such as:

    - Retention time (RT): the amount of time spent by a compound on the separation column after it has been injected on it.

    - Composite spectrum (CS): a spectrum created from the sum of all co-eluting *m/z* ions that are related (including isotopes, adducts and dimmers). It is used to calculate the original mass from the alterations within the CS.

    - Chemical alphabet: the set of possible elements of the putative annotations (CHNOPS, CHNOPS+Cl, All). Deuterium can either be included or not.

    - Modifiers: mobile phase modifier used. The adduct formation may change depending on this modifier.

- **Batch search:** simple search performed over a set of EMs obtained by MS.

- **Batch advanced search:** advanced search over a set of EMs obtained from MS.

- **Browse search:** allows to perform a search over the name and/or formula of the compound, specifying in which databases and over which type of metabolites perform the search.

- **Pathway displayer:** this feature extracts the information from a list of already identified compounds. It determines which pathway that set of compounds belongs to. A rank of pathways is given as output.

- **Long chain oxidation:** finds the non-oxidised fatty acids from long chain oxidised glycerophosphocholines. The input data includes oxidised and non-oxidised fatty acids and the precursor ion.

- **Short chain oxidation:** provides the name, molecular formula and ppm error for each oxidation type in separated bookmarks. The input data includes the non-oxidised fatty acid and precursor ion.

  Long and short chain oxidation functionalities are designed for annotation of oxidised phosphatidylcholines. Information about fatty acids chain(s) and the precursor ion is used to predict the type of oxidation and to compute the putative candidates' identifiers. In case of multiple candidates, a list of the expected fragments and neutral losses is provided to help in the annotation process.

- **Spectral quality controller:** is a tool for semi-automated assessment of the quality of MS/MS spectra. Based on the information about signal intensity, noise, number of scans, co-elution and cross-talk, the overall score is computed together with a quality tag assigning spectrum as excellent, acceptable or inadequate.

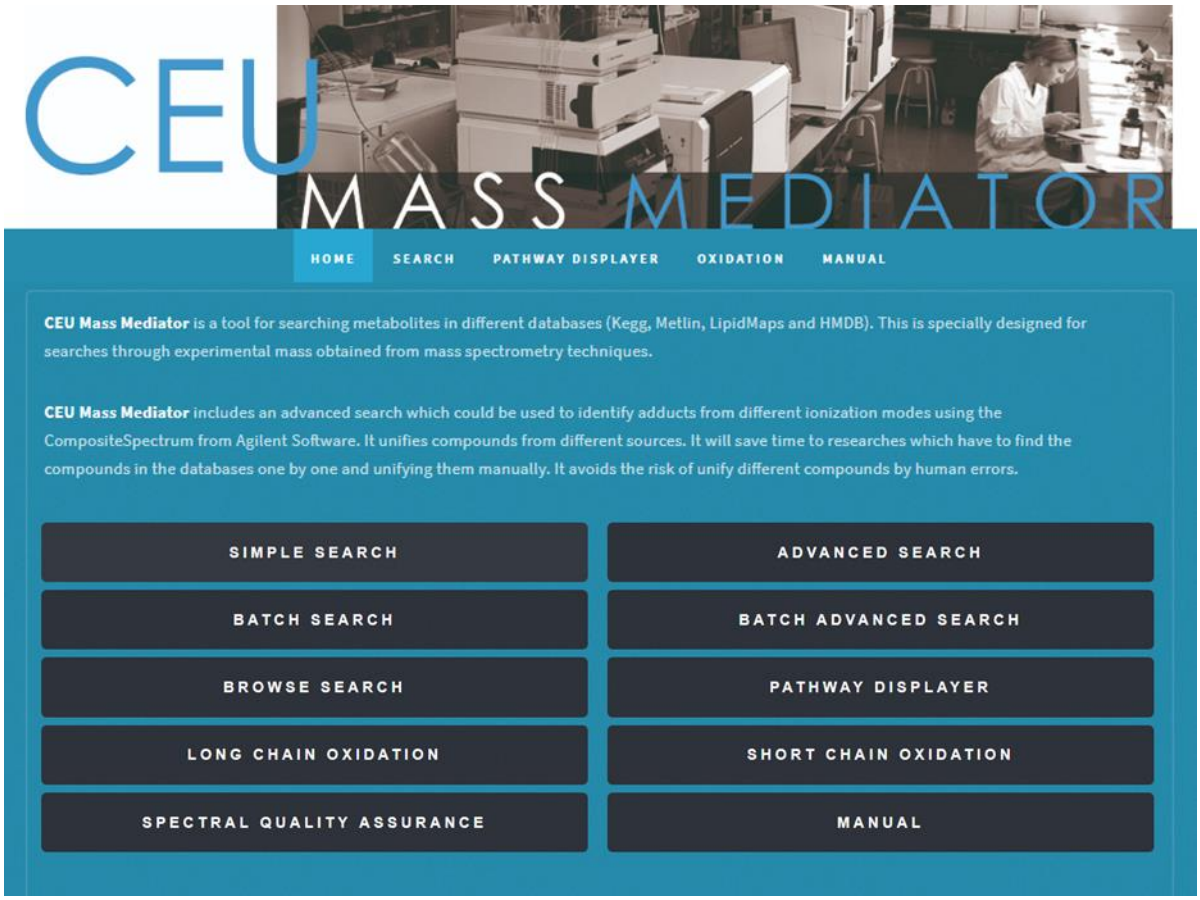


Figure 12 CMM main page.

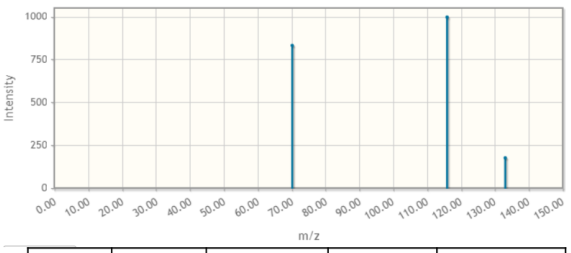
The search results are shown as a list on the screen and paginated depending on the number of features. The results can be exported to a downloadable file in Excel format to facilitate the processing of information by the chemical analyst. Each unified compound has a link to the web of the metabolomic database in which it was found. This link allows the access to more detailed information about the compound that is not stored internally in the CMM system. A typical query of the significant masses from MS data of a single experiment can return thousands of results [57].

3 Functionalities design and implementation

3.1 MS/MS based search

One of the CEMBIO's requirements was the integration of an MS/MS based search functionality into CMM. This feature was developed following a set of steps. First, it was necessary to create an entity-relationship model for MS/MS data and to integrate HMDB MS/MS spectral data into CMM database. Next, the spectral matching algorithms' design needed to be developed and validated.

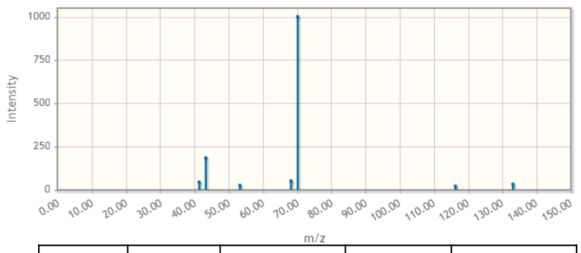
MS/MS spectral matching is not a trivial task since the acquired spectra depends on the machines and the acquisition techniques. The following figures (Figure 13 and Figure 14) show four different MS/MS from the same compound, Ornithine. Each spectrum has its acquisition information bellow. Figure 13 illustrates two MS/MS spectra obtained by different ionisation voltages. The molecule's fragmentation increases with the increment of the voltage. The spectra from Figure 14 were obtained with distinct ionisation modes. From these MS/MS spectra it is appreciable that the voltages, ionisation modes and instruments (each of the spectra present in the Figure 13 and Figure 14 correspond to a different instrument) condition the fragmentation pattern, making the identification and reproducibility more complex.



Voltage	Voltage level	Instrument	Ionization mode	Spectra type
10	low	Quattro_QQQ	Positive	Experimental

Peaks

m/z	Intensity
71.133	834.95
115.811	1000
132.962	179.61

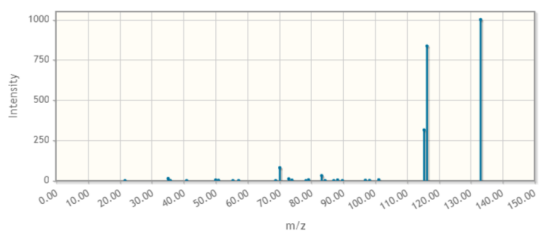


Voltage	Voltage level	Instrument	Ionization mode	Spectra type
40	high	Quattro_QQQ	Positive	Experimental

Peaks

m/z	Intensity
132.994	45.11
116.162	184.78
70.11	26.49
68.159	52.17
53.221	1000
43.271	22.42
41.35	33.15

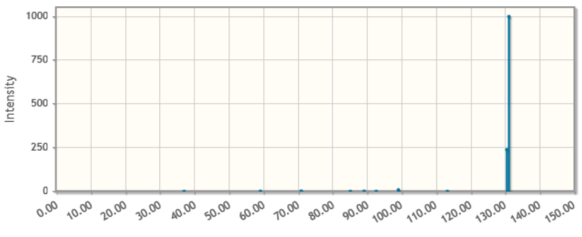
Figure 13 Experimental MS/MS from Ornithine (mass 132.08988). The spectra were obtained with different voltages (10V and 40V).



Voltage	Voltage level	Instrument	Ionization mode	Spectra type
10	-	LC-ESI-QQ (API3000, Applied Biosystems)	Positive	Experimental

Peaks

m/z	Intensity
35.1	10
70.1	80
72.9	10
83.2	30
96.9	3
98.2	2
115.3	315
116.2	836
133	1000



Voltage	Voltage level	Instrument	Ionization mode	Spectra type
10	-	LC-ESI-QQ (API3000, Applied Biosystems)	Negative	Experimental

Peaks

m/z	Intensity
37	1
59.1	2
70.9	3
85.1	1
89.1	2
92.6	1
99	10
113.2	1
130.4	240
131	1000

Figure 14 Experimental MS/MS from Ornithine (mass 132.08988). Different ionisation modes (*m/z* of the precursor ion in positive mode: 133, *m/z* of the precursor ion in negative mode: 131) and 10V.

3.1.1 Entity-relationship model

During the internship period, the MS/MS entity-relationship model was developed according to CEMBIO's feedback. Some of its members reviewed the attributes that we decided to consider for the subsequent spectral matching. Those attributes were extracted from a set of MS/MS spectra XML files from Human Metabolome Database web page [45].

Figure 15 shows a XML file from a MS/MS fragmentation from HMDB. It contains a set of peaks and zero or more references. Within the MS/MS attributes, the relevant ones for the spectral matching are the msms_id, instrument-type, peak-counter, collision-energy-level, collision-energy-voltage, ionisation-mode, predicted and database-id. The instrument type is the separation technique, the ion source and/or the mass analyser employed to obtain the spectrum and the peak counter is the number of peaks contained in the spectrum. The collision-energy-level and the collision-energy-voltage provide qualitative (low, medium or high) and quantitative (from 10 to 60) information respectively about the voltage used in the MS/MS spectrum acquisition. A low voltage corresponds to the range between 10 to 20 volts, a medium voltage from 20

to 40 volts and a high voltage from 40 to 60 volts. The ion source can apply either a positive or a negative ionisation to the sample, therefore, the ionisation-mode attribute indicates this value. The spectrum was obtained experimentally or predicted by *in-silico* fragmentation: the parameter predicted can be true or false and it indicates how the spectrum was obtained. Finally, the database-id is the HMDB id, that allows to relate each MS/MS with its corresponding compound in CMM database.

Regarding the peaks, their relevant attributes are the peak_id, the intensity, which is the ion abundance, mass-to-charge that corresponds to the *m/z* and the msms_id. The reference relevant variables are the reference_id, pubmed-id, ref-text and the msms_id and they are the publications where the spectrum fragmentation was published.

Each compound has from zero to many MS/MS. The corresponding entity-relationship model for the MS/MS data is illustrated in Figure 16. Figure 16 was extracted from MySQL WorkBench by automatic generation. Figure 17 shows the analogous UML model for the MS/MS data.

In Figure 17 we can see an intermediate entity between MSMS and compound entities called MSMSCompound. This intermediate entity inherits from MSMS and it is used to store the putative annotations. It contains data from the MS/MS and from the compound taken from CMM database. Moreover, it contains the match score, which is calculated and it will be explained at 3.1.3.

```
<ms-ms>
  <id>606810</id>
  <notes>Predicted by CFM-ID</notes>
  <sample-concentration nil="true"/>
  <solvent nil="true"/>
  <sample-mass nil="true"/>
  <sample-assessment nil="true"/>
  <spectra-assessment nil="true"/>
  <sample-source nil="true"/>
  <collection-date nil="true"/>
  <instrument-type nil="true"/>
  <peak-counter>5</peak-counter>
  <created-at>2017-10-04T17:32:59Z</created-at>
  <updated-at>2017-10-18T11:03:33Z</updated-at>
  <mono-mass nil="true"/>
  <energy-field nil="true"/>
  <collision-energy-level>high</collision-energy-level>
  <collision-energy-voltage>40</collision-energy-voltage>
  <ionization-mode>Positive</ionization-mode>
  <sample-concentration-units nil="true"/>
  <sample-mass-units nil="true"/>
  <predicted>true</predicted>
  <structure-id>623616</structure-id>
  <splash-key>splash10-00jj-0000943000-ac62f010f4ce079b9979</splash-key>
  <database-id>HMDB0097792</database-id>
  <references>
    <reference>
      <id>410722</id>
      <spectra-id>606810</spectra-id>
      <spectra-type>MsMs</spectra-type>
      <pubmed-id nil="true"/>
      <ref-text>Allen F, Greiner R, Wishart D: Competitive fragmentation modeling
      of ESI-MS/MS spectra for putative metabolite identification. Metabolomics.
      2015 11(1):98-110. doi: 10.1007/s11306-014-0676-4.</ref-text>
      <database nil="true"/>
      <database-id nil="true"/>
    </reference>
  </references>
  <ms-ms-peaks>
    <ms-ms-peak>
      <id>21425027</id>
      <ms-ms-id>606810</ms-ms-id>
      <mass-charge>712.64551</mass-charge>
      <intensity>15.0</intensity>
      <molecule-id nil="true"/>
    </ms-ms-peak>
        ...
    <ms-ms-peak>
      <id>21425031</id>
      <ms-ms-id>606810</ms-ms-id>
      <mass-charge>523.47263</mass-charge>
      <intensity>90.0</intensity>
      <molecule-id nil="true"/>
    </ms-ms-peak>
  </ms-ms-peaks>
</ms-ms>
```

Figure 15 MS/MS spectrum file from HMDB [45].

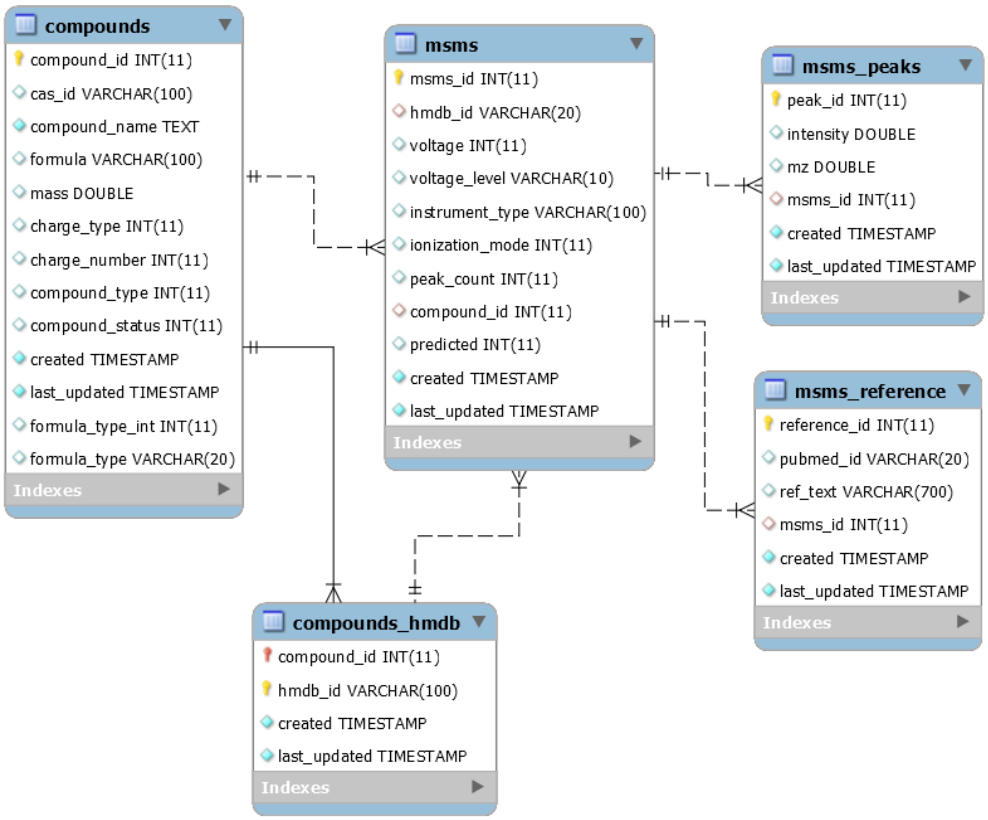


Figure 16 Entity-relationship model for MS/MS data in CMM database.

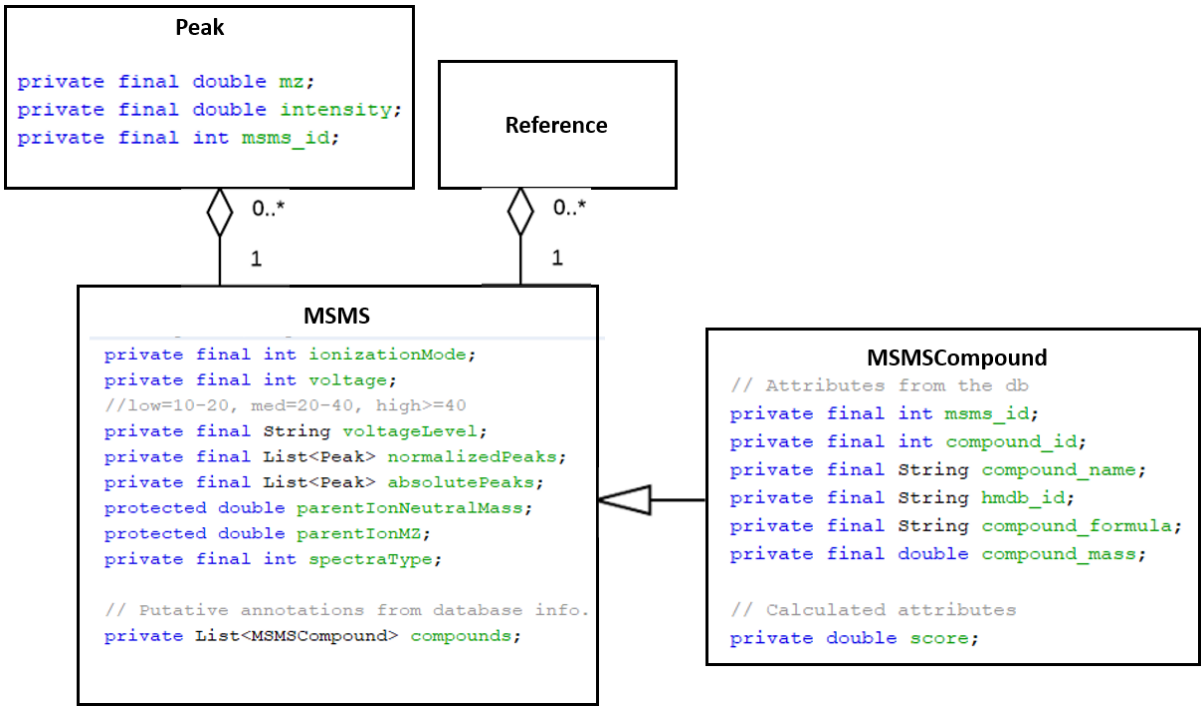


Figure 17 UML model for MS/MS search functionality.

3.1.2 Database insertion

Once the relevant attributes were selected, the insertion procedure in CMM database should be implemented. The insertion code was programmed in Java and the data was inserted into a MySQL database (CMM database). This data integration was developed during a previous internship period. Nevertheless, the HMDB database was updated during the development of this project and it was necessary to insert the data again for this thesis with some modifications. Before HMDB's update, there were only 5,912 MS/MS spectra, all of them experimentally acquired. Some spectra files are duplicated and we aimed to insert only one. Therefore, 5,780 MS/MS were inserted. After the update of HMDB 4.0, all HMDB's spectra files were loaded, 520,604 in total. But those files were from MS, NMR and MS/MS spectra. Our interest was focused in the last ones. Moreover, there exist duplicates, as it was explained before. Therefore, 457,614 MS/MS experimental and predicted spectra were inserted in CMM database. We have around 20x times more spectra after the HMDB 4.0 update compared to HMDB 3.0.

Regarding the improvements needed, some MS/MS attributes were inserted as integers to provide efficiency. The ionisation mode is 1 if positive and 2 if negative. The predicted value is 1 if it was calculated by *in-silico* fragmentation and 0 if it was acquired experimentally.

As previously mentioned, each compound has one or more MS/MS. Normally a compound may have three MS/MS for each ionisation mode, each one obtained by a different voltage level (low, medium, high). Nevertheless, since there are several ionisation instruments (see Table 5) and voltage values (see Table 6), the same compound can have more than six MS/MS due to the possible voltages and instrument type combinations.

Table 5 Different instrument types in CMM database.

INSTRUMENT TYPES	
Quattro_QQQ	EI-B (SHIMADZU QP-1000)
EI-B (HITACHI RMU-6L)	LC-ESI-ITTOF (Shimadzu LC20A-IT-TOFMS)
EI-B (HITACHI RMU-6M)	EI-B (HITACHI M-80A)
CI-B (HITACHI M-80)	FD-B (Unknown)
LC-ESI-QQ (API3000, Applied Biosystems)	LC-APPI-QQ (API2000)
LC-ESI-QTOF (UPLC Q-Tof Premier, Waters)	EI-B (JEOL JMS-DX-300)
EI-B (HITACHI RMU-7M)	EI-B (JEOL JMS-HX-100)
LC-ESI-IT (LC/MSD Trap XCT, Agilent Technologies)	EI-B (HITACHI RMU-6D)
EI-B (HITACHI M-80)	EI-B (JEOL JMS-06-H)
EI-B (HITACHI M-68)	EI-B (HP 5970)
EI-B (HITACHI M-52)	LC-ESI-ITFT (LTQ Orbitrap XL, Thermo Scientific)
LC-ESI-ITFT (LTQ Orbitrap XL, Thermo Scientfic)	LC-ESI-QIT (4000Q TRAP, Applied Biosystems)
EI-B (HITACHI M-80B)	LC-ESI-QTOF (ACQUITY UPLC System, Waters)
CE-ESI-TOF (CE-system connected to 6210 Time-of-Flight MS, Agilent)	LC-ESI-ITFT (LTQ Orbitrap XL Thermo Scientific)
EI-B (JEOL JMS-01-SG-2)	ESI-TOF
EI-B (HITACHI RMU-6E)	DI-ESI-Hybrid FT
MALDI-TOF (Voyager DE-PRO, Applied Biosystems)	DI-ESI-qTof
LC-ESI-QQ (UPLC Waters, Quattro Ultima Pt Micromass)	DI-ESI-Ion Trap
FAB-EBEB (JMS-HX/HX 110A, JEOL)	LC-ESI-Ion Trap
EI-B (Unknown)	LC-ESI-qTof
EI-B (JEOL JMS-D-3000)	LC-ESI-Hybrid FT
EI-B (HITACHI M-60)	DI-ESI-Q-Exactive Plus
EI-B (MX-1303)	LC-ESI-QIT
LC-ESI-ITTOF (LCMS-IT-TOF)	LC-ESI-ITFT
EI-B (JEOL JMS-AX-505-H)	LC-ESI-QFT
EI-B (VARIAN MAT-44)	LC-ESI-QQ
EI-B (JEOL JMS-D-300)	LC-ESI-IT
CI-B (JEOL JMS-D-300)	LC-ESI-ITTOF
EI-B (SHIMADZU LKB-9000B)	Linear Ion Trap
CI-B (HITACHI M-60)	LC-ESI-TOF
EI-B (HITACHI RMU-7L)	ESI-ITFT
CI-B (FINNIGAN-MAT 4500)	APCI-ITFT
CI-B (Unknown)	LC-APPI-QQ
EI-B (HITACHI RMU-7)	ESI-QTOF
EI-EBEB (JMS-HX/HX 110A, JEOL)	

Table 6 Different voltages in CMM database.

VOLTAGE	
0	30
5	35
6	40
10	45
20	50
25	55
26	60

Duplicates removal

The elimination of duplicates during the MS/MS data insertion supposed a challenge during the internship. Some compounds from the HMDB database were not found in the CMM database. Therefore, the compound_id of some MS/MS were set to zero. The compound_id is the identifier of each compound. The problem was occurring due to the existence of HMDB ids synonyms in the Human Metabolome Database. This means that, for example the HMDB0000020 entry could be accessed also by the HMDB0000453 entry. Many duplicates were not in the CMM database, hence their compound was not found. The problem was solved by applying some criteria before the insertion (see Figure 18):

● If the compound_id is not zero, check if the MS/MS is already inserted in the database. If it is not, insert it. Otherwise, the spectrum information is already in the database, so do nothing.

● If the compound id is zero, it means that it is from a HMDB synonym absent in our database. Therefore, we look for its synonyms. Then, that synonyms are searched against the CMM database. When a valid synonym is found, we check if the MS/MS information is already in the database. If it is not, insert it. Otherwise, do nothing. If we do not find a synonym, a synonym log will be written, and that MS/MS will not be inserted in the database.

The duplicates are spectra containing the same information. Same msms_id, collision energy, instrument, ionisation mode, peaks and references, but they differ in

their HMDB id. During the insertion four logs were written, one for HMDB0014330, and three for HMDB0032262. Their corresponding pages in HMDB did not exist during the insertion, so it means that the compounds have been deleted from HMDB.

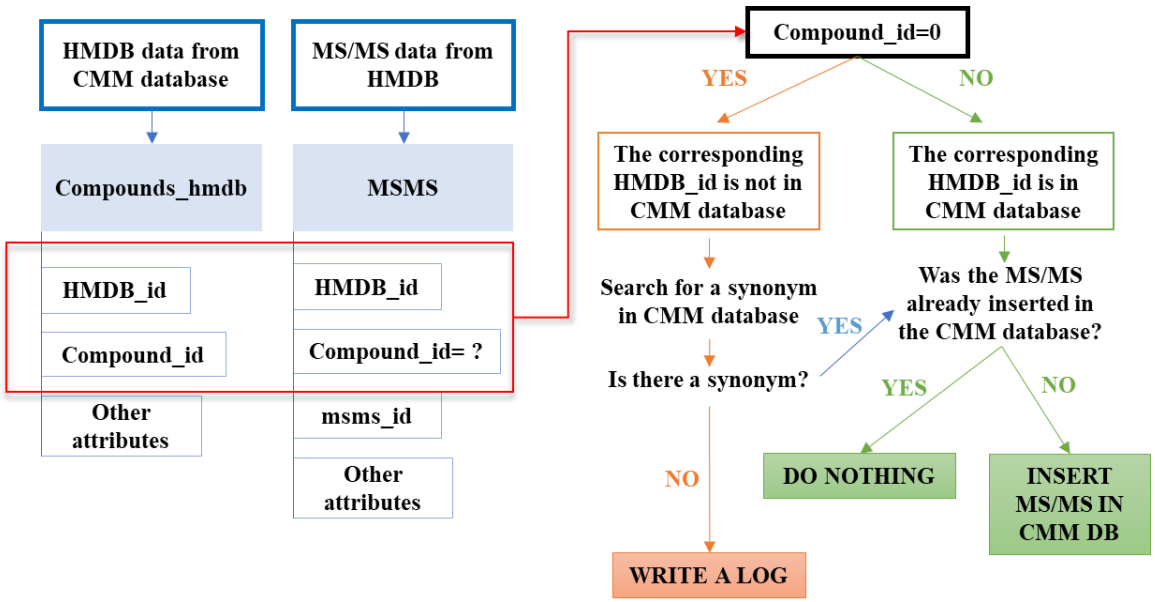


Figure 18 Duplicates removal work-flow.

3.1.3 Spectral matching algorithms

After the data insertion, the next step was the development of the algorithms to perform the MS/MS based identification. After a bibliographic study, the methods that better fitted our purposes were the ones from MetFrag and MyCompound (as mentioned in 2.2). The matching algorithms developed during the internship were not working properly, they needed to be modified.

The spectral matching algorithm work-flow performed during the internship is the following:

1. The precursor ion mass, which is provided as an input from the user, and the precursor ion mass tolerance are used to retrieve a set of candidates.

2. Each candidate has several MS/MS associated, because it can be obtained by different voltages, instruments and ionisation modes, leading to different spectra fragmentation patterns. All the different MS/MS spectra from the

candidates are considered for the matching along with the experimental spectrum introduced by the user. The matching is done by the *m/z* of each peak within a tolerance also provided by the user.

3. Once we have the set of matched spectra, a score function is executed. We computed three different score functions.

   ● A weighted dot product between the matched peaks (MetFrag approach, Equation 1). This product weights the intensity with 0.6 and the *m/z* ratio with 3.

Equation 1 MetFrag score (approach).

$$MetFrag\ score = \sum_{i}^{Number\ of\ matched\ peaks} (inputPeak\_mz_i * libraryPeak\_mz_i)^3 + \\ +(inputPeak\_intensity_i * libraryPeak\_intesity_i)^{0.6}$$

   ● A simple dot product (MyCompoundID approach, Equation 2).

Equation 2 MyCompound score (approach).

$$MyCompound\ score = \sum_{i}^{Number\ of\ matched\ peaks} (inputPeak_i * libraryPeak_i) = \\ = (inputPeak\_mz_i * libraryPeak\_mz_i) + (inputPeak\_intensiy_i * libraryPeak\_intesity_i)$$

   ● The inverse of the Euclidean distance among the matched peaks.

Equation 3 Euclidean distance score.

$$Euclidean\ distance\ score = \sum_{i}^{Number\ of\ matched\ peaks} \frac{1}{\sqrt{(inputPeak_i - libraryPeak_i)^2}} = \\ = \frac{1}{\sqrt{(inputPeak\_mz_i - libraryPeak\_mz_i)^2 + (inputPeak\_intensity_i - libraryPeak\_intesity_i)^2}}$$

The three scores disposed normalised in the output. Moreover, MetFrag and MyCompoundID scores are divided by the dot product of the input peaks over

themselves to penalise the absence of some input peaks in the matched library spectrum. This dot product is weighted in MetFrag's approach. The complete equations are illustrated in Equation 4 and Equation 5.

Equation 4 MyCompounID score penalised.

$$MyCompound\ penalised = \frac{\sum_{i}^{Number\ of\ matc\ hed\ peaks} (inputPeak_i * libraryPeak_i)}{\sum_{j}^{Number\ of\ input\ peaks} (inputPeak_j * inputPeak_j)}$$

Equation 5 MetFrag score penalised.

$$MetFrag\ penalised = \frac{\sum_{i}^{Number\ of\ matc\ hed\ peaks} (inputPeak\_mz_i * libraryPeak\_mz_i)^3 + (inputPeak\_intensity_i * libraryPeak\_intesity_i)^{0.6}}{\sum_{j}^{Number\ of\ input\ peaks} (inputPeak\_mz_j * inputPeak\_mz_j)^3 + (inputPeak\_intensity_j * inputPeak\_intesity_j)^{0.6}}$$

Nevertheless, the scoring obtained needed to be improved since the algorithms were not working properly. The matching pipeline also needed a couple of modifications. Instead of considering all the MS/MS spectra from the candidates, now we just consider those spectra that fit the experimental characteristics (ion mode, voltage and type of spectra -predicted and/or experimental-). Moreover, the scoring functions were not performed over all the matched peaks, the loop was not coded properly.

After the bugs were fixed, the MS/MS search pipeline was very similar (see Figure 19). Once the input data is introduced: precursor ion mass, precursor ion tolerance, peaks (*m/z*, intensity pairs), peaks' *m/z* tolerance, ionisation mode, voltage, and spectra type, the search starts. First, the candidates are searched by querying the precursor mass within a tolerance against CMM database. Those candidates are a list of MSMSCompound objects containing information about a compound with the wanted mass and its corresponding MS/MS fulfilling the input data. Then the candidates are scored according to the scoring function. It only scores those peaks within the specified tolerance. At this moment, the scoring function employed is the one from MetFrags's

approach. MyCompoundID's approach and the Euclidean distance algorithms are also available in the project.

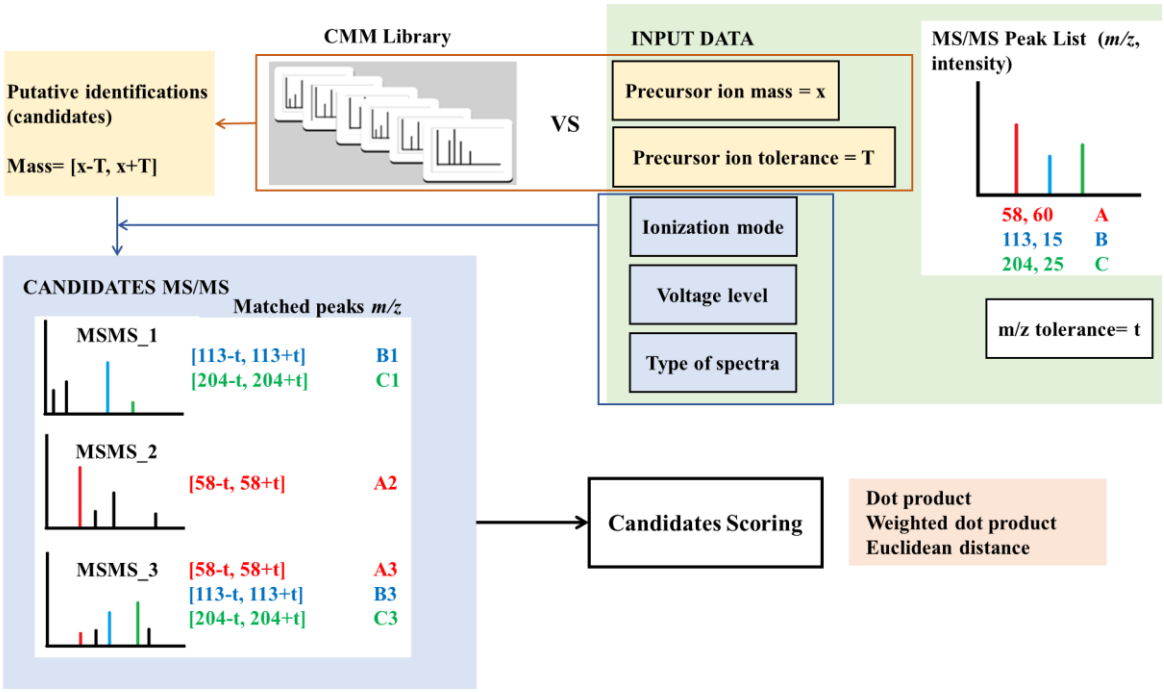


Figure 19 MS/MS based search pipeline.

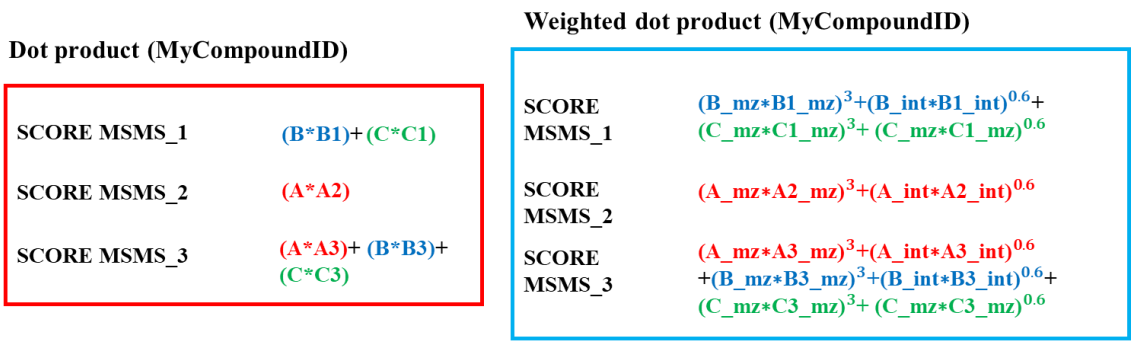


Figure 20 Scoring functions examples.

Figure 20 illustrates some score examples (not penalised) analogous to Figure 19 to facilitate the understanding of the formulas. In Figure 19 the input spectrum "MSMS" has tree peaks; A, B, C. There are three candidate spectra from the database/library; "MSMS_1", "MSMS_2" and "MSMS_3". These candidates have peaks in common with "MSMS", the matched peaks are: B1, C1 in MSMS_1; A2 in MSMS_2; and A3, B3, C3 in MSMS_3. Each peak is formed by a *m/z* and an intensity,

which are referred to as peak_mz and peak_int respectively. For example, A's *m/z* and intensity are expressed as A_mz and A_int. Figure 20 represents the Equation 1 in blue and the Equation 2 in red.

3.1.4 Front-end

The MS/MS search feature was integrated in CMM web tool. The front-end is formed by all the input parameters. It is shown in Figure 21:

[1] Precursor ion mass (*m/z*): the mass to search in CMM (Da).

[2] MS/MS peak list: a set of peaks (*m/z*, intensity) from the mass spectrum. Intensities can be introduced as absolute of relative. It is mandatory to introduce just one *m/z* and its correspondent intensity per line, in that order and separated by a blank space. Figure 22 illustrates how to insert the peaks' input with absolute or relative intensities.

[3] Precursor ion tolerance: the mass difference allowed between the EM and the precursor ion mass. It can be specified in ppm or Da. The default value is 0.1 Da.

[4] *m/z* peaks tolerance: the tolerance for peaks' *m/z* matching (spectral matching). It can be specified in ppm or Da. The default value is 0.5 Da.

[5] Ionisation mode: the ionisation mode applied when performing MS/MS. The default value is positive.

[6] Ionisation voltage: the ionisation voltage applied when performing MS/MS. The default value is low.

[7] Type of spectra: the type of spectra against which the search is performed. The type of spectra can be experimental (MS/MS data obtained from real metabolites) and/or predicted (MS/MS data obtained through *in-silico* fragmentation performed by HMDB). The default value is experimental.

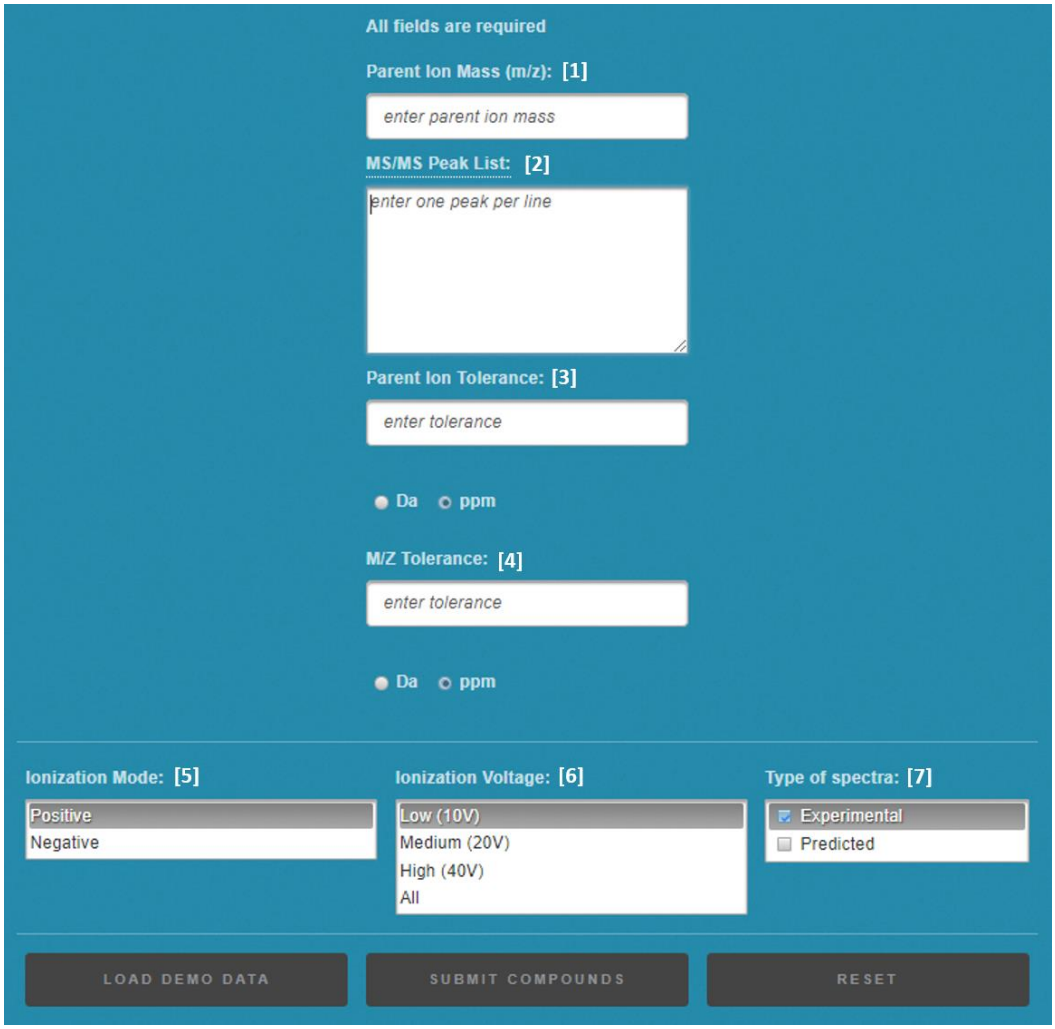


Figure 21 MS/MS search interface.

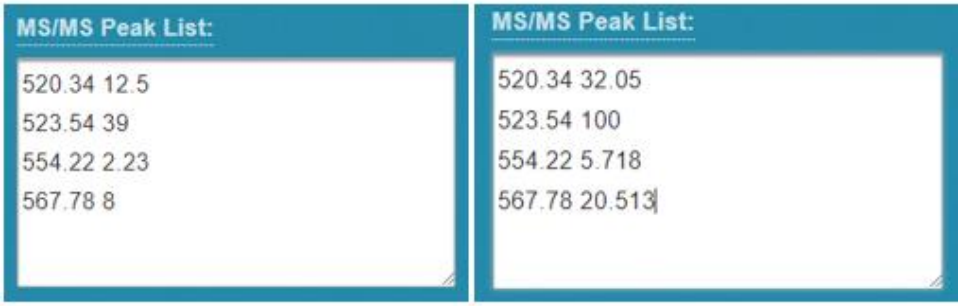


Figure 22 Peak input. Absolute intensities (left) and relative intensities (right).

3.2 New data model

Due to CEMBIO's interest in having a LC-MS search feature in CMM, it was necessary to modify CMM data model. The previous model did not facilitate features' grouping by RT. Therefore, to ease the implementation of the functionality, the model from simple, batch, advanced and batch advanced searches have been adapted.

3.2.1 MS search refactoring

Since the LC-MS search implementation required a new data model, the available CMM searches needed a refactoring. Moreover, more data is being considered to complete compound's information. The database contains more tables with information about different classifications that the compounds belong to. The previous entity-relationship model is visible in Figure 23 and the actual entity-relationship model is visible in Figure 24. The size of the actual model does not fit in a single page, therefore Figure 24 is split in two pages. The model is divided by a red line to ease tracking the image.

In the previous model, the entity **compounds_identifiers** saves each compound's unique identifier, named **inchi_key**. The entity named **compounds_cas** saves information about those compounds provided by the American Chemical Society official's API. That entity is not related with the compounds in the model, it only saves the CAS information. There is an entity for the possible reactions where the different compounds are involved, which is **compounds_reactions_kegg**.

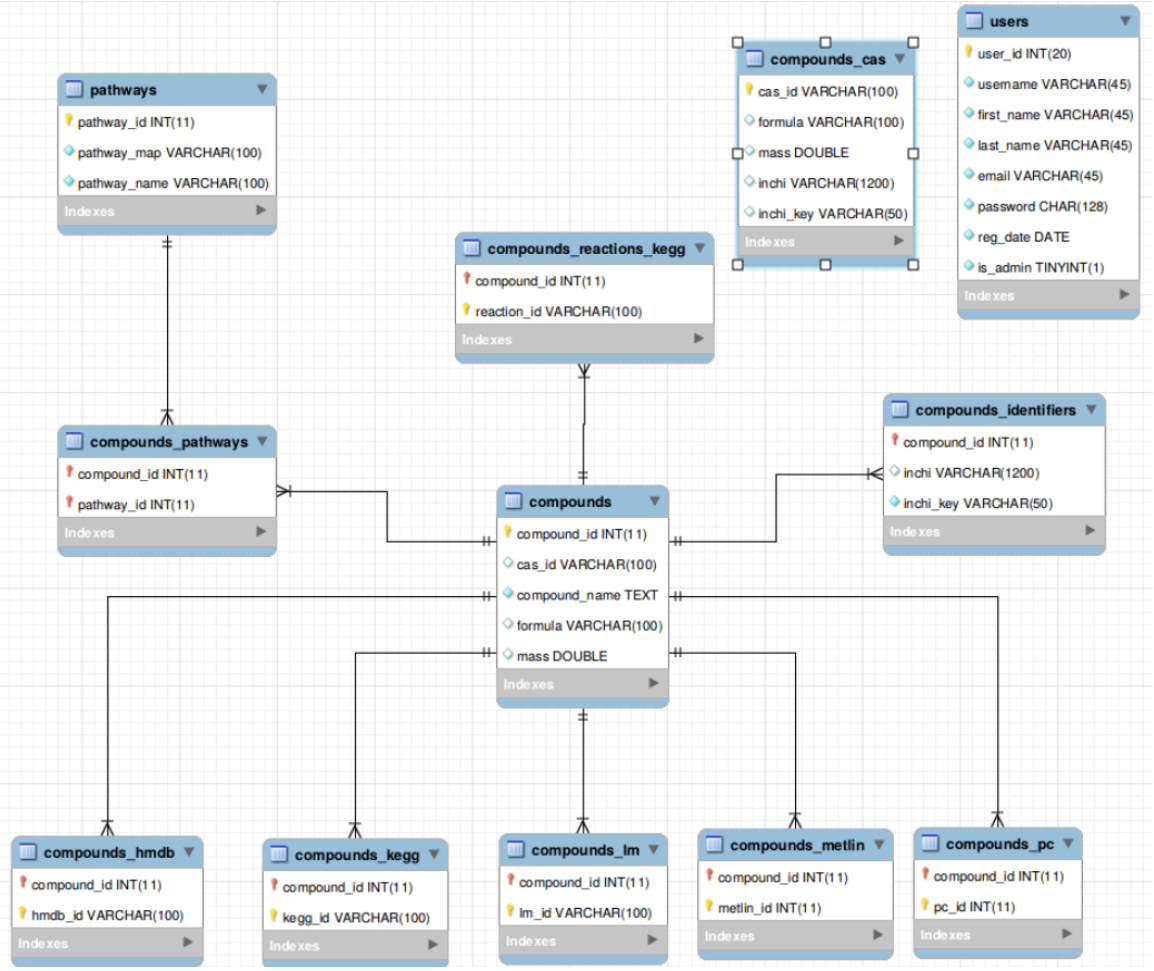
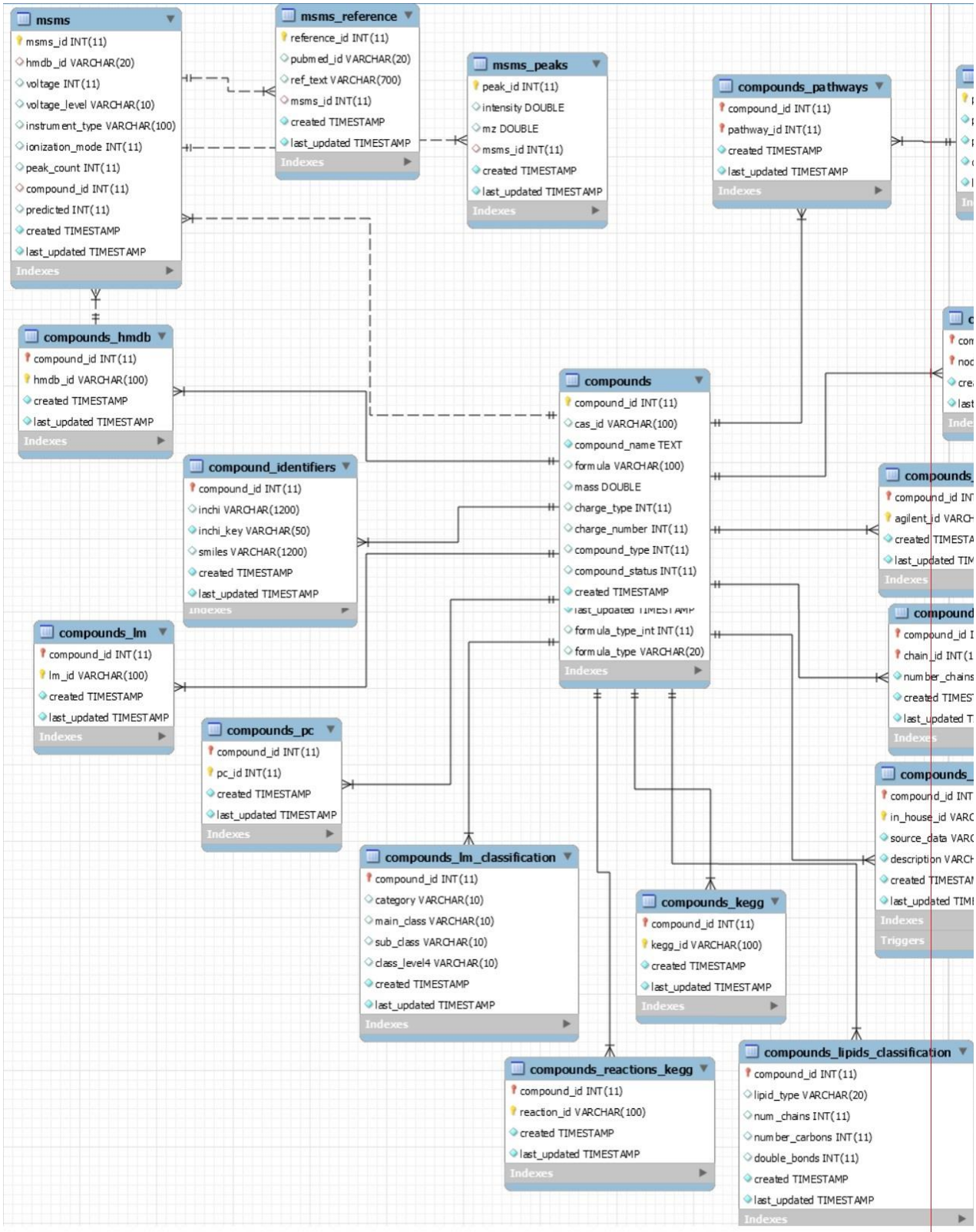


Figure 23 Entity-relationship model from the previous application.

In the new data model, the compounds are the center of the entity-relationship model, as in the previous one. Nevertheless, the number of entities and relations have increased considerably. The MS/MS establish a N:1 relation with the compounds, and the MS/MS are assigned to the compounds according to their HMDB_id. The compounds have different classifications assigned, used for a different purpose than the project presented here.


msms
msms_id INT(11)
hmdb_id VARCHAR(20)
voltage INT(11)
voltage_level VARCHAR(10)
instrument_type VARCHAR(100)
ionization_mode INT(11)
peak_count INT(11)
compound_id INT(11)
predicted INT(11)
created TIMESTAMP
last_updated TIMESTAMP
Indexes
msms_reference
reference_id INT(11)
pubmed_id VARCHAR(20)
ref_text VARCHAR(700)
msms_id INT(11)
created TIMESTAMP
last_updated TIMESTAMP
Indexes
msms_peaks
peak_id INT(11)
intensity DOUBLE
mz DOUBLE
msms_id INT(11)
created TIMESTAMP
last_updated TIMESTAMP
Indexes
compounds_pathways
compound_id INT(11)
pathway_id INT(11)
created TIMESTAMP
last_updated TIMESTAMP
Indexes
compounds_hmdb
compound_id INT(11)
hmdb_id VARCHAR(100)
created TIMESTAMP
last_updated TIMESTAMP
Indexes
compounds
compound_id INT(11)
cas_id VARCHAR(100)
compound_name TEXT
formula VARCHAR(100)
mass DOUBLE
charge_type INT(11)
charge_number INT(11)
compound_type INT(11)
compound_status INT(11)
created TIMESTAMP
formula_type_int INT(11)
formula_type VARCHAR(20)
Indexes
compound_identifiers
compound_id INT(11)
inchi VARCHAR(1200)
inchi_key VARCHAR(50)
smiles VARCHAR(1200)
created TIMESTAMP
last_updated TIMESTAMP
Indexes
compounds_lm
compound_id INT(11)
lm_id VARCHAR(100)
created TIMESTAMP
last_updated TIMESTAMP
Indexes
compounds_pc
compound_id INT(11)
pc_id INT(11)
created TIMESTAMP
last_updated TIMESTAMP
Indexes
compounds_lm_classification
compound_id INT(11)
category VARCHAR(10)
main_class VARCHAR(10)
sub_class VARCHAR(10)
class_level4 VARCHAR(10)
created TIMESTAMP
last_updated TIMESTAMP
Indexes
compounds_kegg
compound_id INT(11)
kegg_id VARCHAR(100)
created TIMESTAMP
last_updated TIMESTAMP
Indexes
compounds_reactions_kegg
compound_id INT(11)
reaction_id VARCHAR(100)
created TIMESTAMP
last_updated TIMESTAMP
Indexes
compounds_lipids_classification
compound_id INT(11)
lipid_type VARCHAR(20)
num_chains INT(11)
number_carbons INT(11)
double_bonds INT(11)
created TIMESTAMP
last_updated TIMESTAMP
Indexes

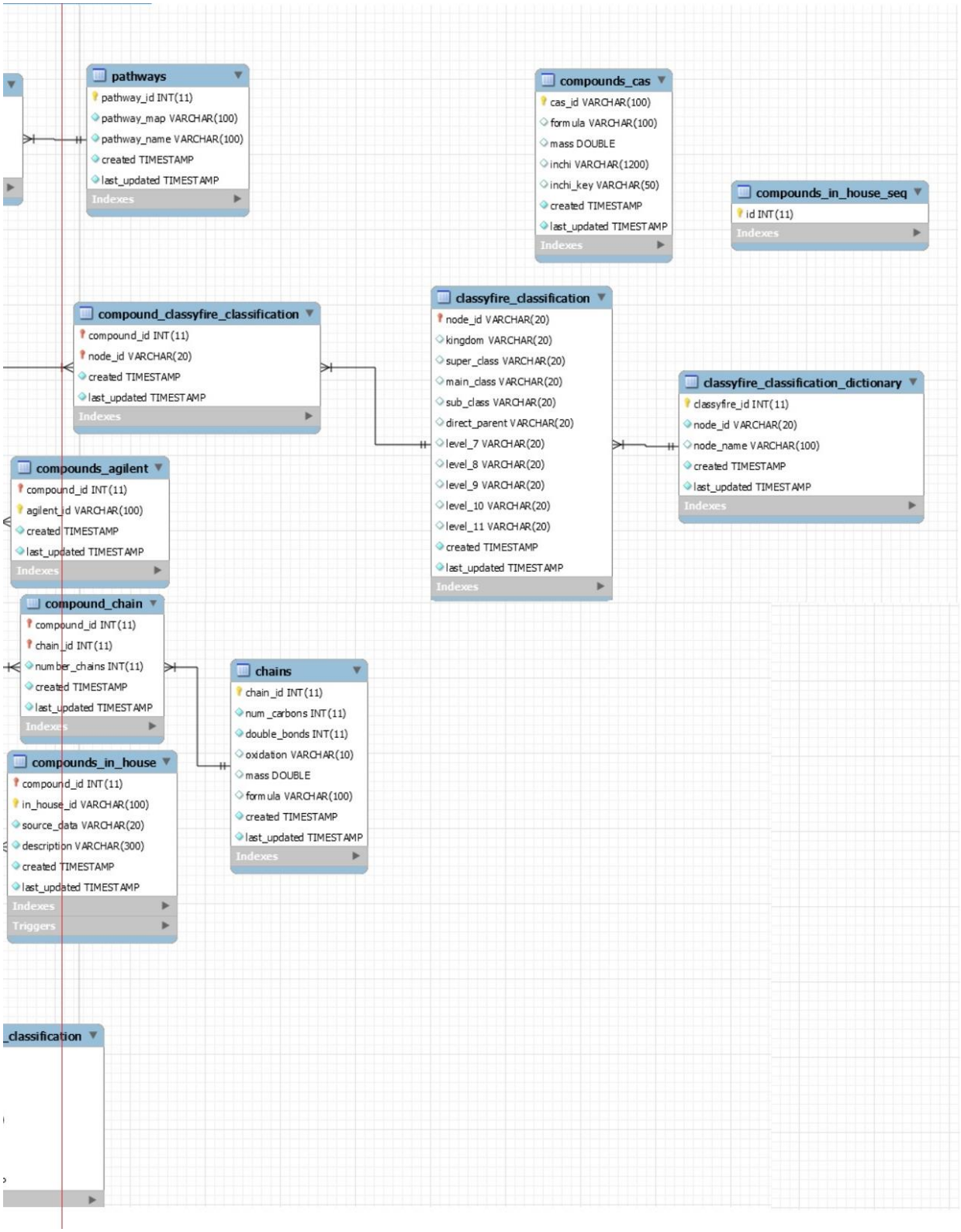


Figure 24 New entity-relationship model.

Figure 25 illustrates the new UML model developed for simple and batch searches, where the red squared fields are the data model modifications. The entity experiment has two sets of features, the significant features and the set of all features. As previously explained in the Chapter 0, a feature is a three-dimensional signal, composed by EM, RT and CS. But in these simple searches is only composed by EM. Each feature contains a list of compounds grouped by adduct (list of compoundsLCMSGroupByAdduct objects). Each compoundsLCMSGroupByAdduct is formed by the feature's EM, RT, CS, an adduct and a list of compounds with the feature's EM according to the given adduct. For example, if the EM is 301.0073 and the adduct $[M+H]^+$, the list of compounds will be with a mass of 300 (mass - $[H]^+$ mass → 301.0073 - 1.0073) within a tolerance window. The list of compounds is a list of CompoundLCMS objects. This entity inherits from Compound and contains feature's and compound's attributes. In the simple search, an experiment with a single feature is created, whilst in the batch search, the experiment can be formed by one of more features.

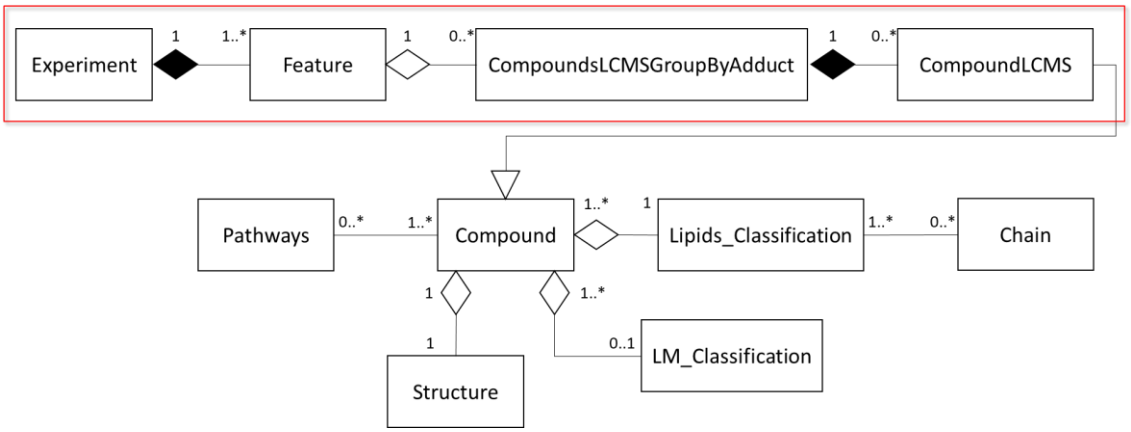


Figure 25 New data model for LC-MS search. Used in simple and batch searches.

The difference between this data model and the previous one is that here the features are grouped by their RT within a tolerance established by the user. The features with a similar RT may have been formed from the same signal, and this grouping tries to look first if there is any type of relation between different adducts, and second tries to search if some of the signals may correspond to an in-source fragment of a higher *m/z* signal. The in-source fragmentation is a phenomenon that happens when an ion is

fragmented directly in the ionisation source. Therefore, the fragments have the same RT than the precursor ions.

The Figure 26 shows the new data model for advanced and batch advanced searches. An experiment contains a set of features grouped by RTs (list of FeaturesGroupByRT objects). Each FeatureGroupByRT is formed by a RT and a list of features. A feature is a three-dimensional signal. But its only mandatory field is the EM. In other words, in advanced searches we can have a feature with EM, with EM and RT, with EM and CS, and with the three of them. The model keeps as the simple one from this point. A Feature contains a list of compounds grouped by adduct. Each compound grouped by adduct contains a list of compoundsLCMS which is a class that inherits from Compound. In the advanced search the experiment is formed by a single FeaturesGroupByRT containing a single feature. Otherwise, the advanced batch search can be composed by one or more FeatureGroupByRT, each containing one or more features.

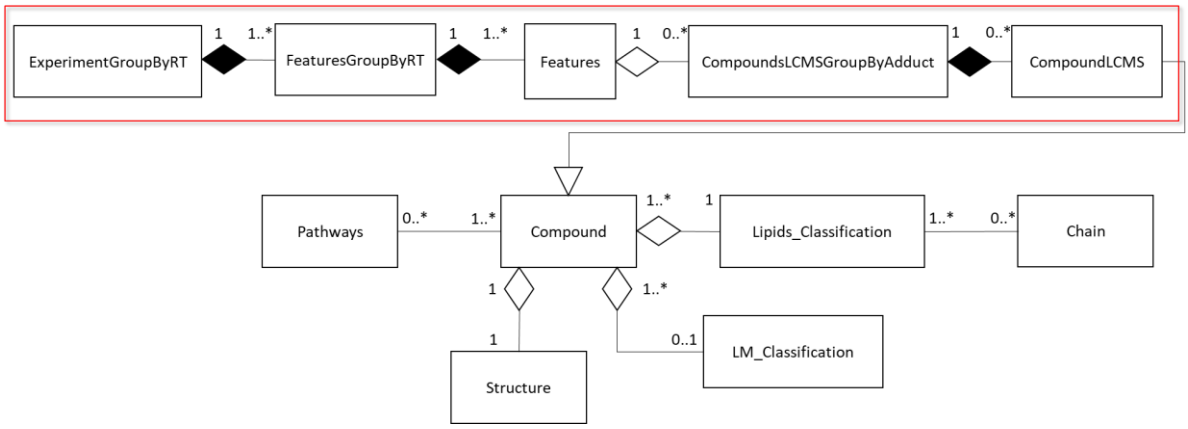


Figure 26 New data model for LC-MS search for advanced and batch advanced searches.

3.3 In-source fragments detector

The annotation of a LC-MS feature is a key task in metabolomic analysis. Within a set of features, we aim to determine if the unidentified ones may be fragments of another features' annotations with the same RT (which means they are derived from the same analyte). Once the new data model was designed (see Figure 27), we could proceed with the algorithm design, development and validation for the in-source fragmentation.

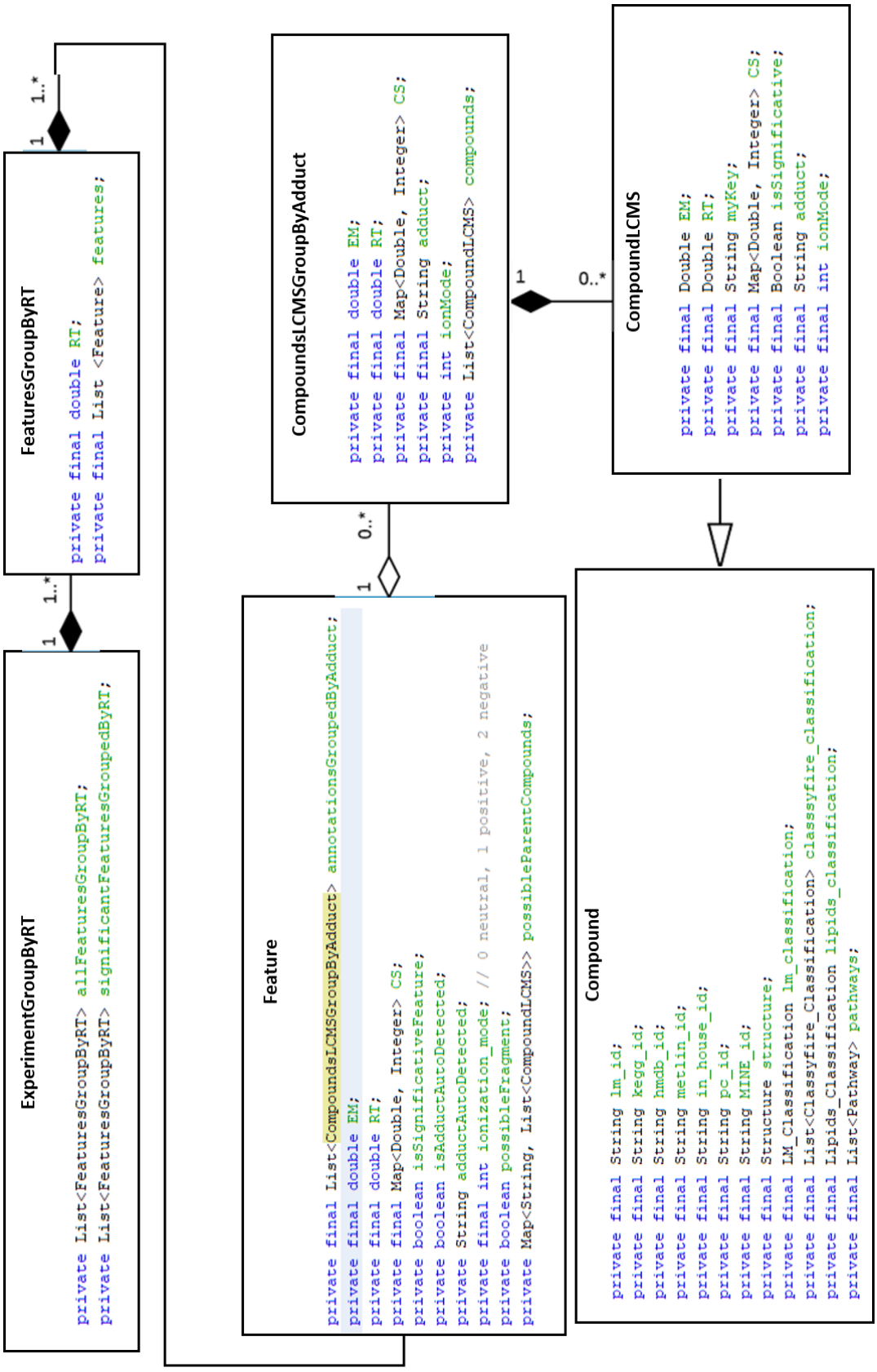


Figure 27 LC-MS Search data model.

The fragments detection pipeline (see Figure 28) starts grouping the features by RT, within a window tolerance introduced by the user. When these groups are formed, the algorithm tries to automatically detect if each feature corresponds to a different adduct arising from the same signal. This detection can be done with the information provided by the CS or by the relationship among features within the same retention time window. Then, the putative annotations corresponding to different adducts are assigned to each feature. At this point, the algorithm marks the features as possible fragments of the ones with a greater EM. Finally, CMM looks into the features marked as possible fragments. Each putative fragment has a different ionisation adduct hypothesis. Depending on the ionisation adduct hypothesis, the precursor ion should be ionised consequently (look Table 7 and Table 8). Then the algorithm considers the fragmentation pattern (MS/MS data) to check if the EM of the possible fragment has been detected in a MS/MS spectrum of the putative precursor ions. Each step will be explained in more detail in the following subsections.

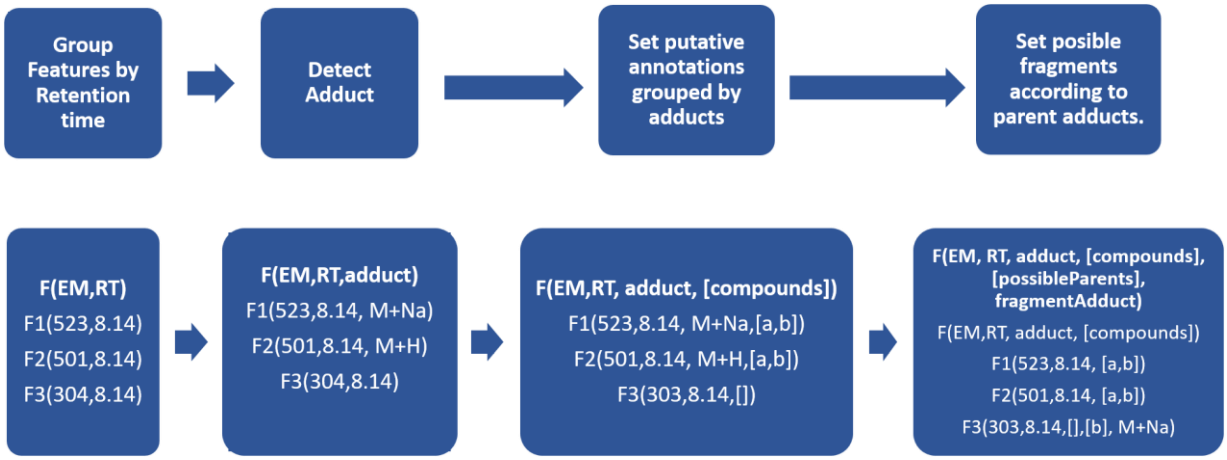


Figure 28 LC-MS search work-flow.

Table 7 Fragments' adducts and corresponding precursor ions. Positive ionisation mode.

POSITIVE IONISATION MODE	
Fragment ion	Precursor ions
M+H	M+H, M+2H, M+3H, M+H-H2O, M+H+NH4, M+H+HCOONa, M+2H+Na, M+H+2K, M+H+2Na, M+H+Na, M+H+K, 2M+H, 2M+H-H2O, M+ACN+2H, M+2ACN+2H, M+3ACN+2H, M+CH3OH+H, M+ACN+H, M+IsoProp+H, M+DMSO+H, M+2ACN+H, M+IsoProp+Na+H, 2M+ACN+H
M+2H	M+2H, M+2H+Na, M+ACN+2H, M+2ACN+2H, M+3ACN+2H
M+Na	M+Na, 2M+Na, M+2H+Na, M+H+2Na, M+3Na, M+H+Na, M+2Na, M+2Na-H, M+ACN+Na, M+IsoProp+Na+H, 2M+ACN+Na
M+K	M+K, M+H+2K, M+H+K, M+2K-H, 2M+K
M+NH4	M+NH4, M+H+NH4, 2M+NH4
M+H-H2O	2M+H-H2O, M+H-H2O
M+H+NH4	M+H+NH4
M+H+HCOONa	M+H+HCOONa
M+3H	M+3H
M+2H+Na	M+2H+Na
M+H+2K	M+H+2K
M+H+2Na	M+H+2Na
M+3Na	M+3Na
M+H+Na	M+2H+Na, M+H+2Na, M+H+Na
M+H+K	M+H+2K, M+H+K
M+ACN+2H	M+3ACN+2H, M+2ACN+2H, M+ACN+2H
M+2Na	M+3Na, M+2Na, M+H+2Na, M+2Na-H
M+2ACN+2H	M+3ACN+2H, M+2ACN+2H
M+3ACN+2H	M+3ACN+2H
M+CH3OH+H	M+CH3OH+H
M+ACN+H	M+ACN+2H, M+ACN+H, M+2ACN+2H, M+3ACN+2H, M+2ACN+H, 2M+ACN+H
M+2Na-H	M+2Na-H
M+IsoProp+H	M+IsoProp+Na+H, M+IsoProp+H
M+ACN+Na	2M+ACN+Na, M+ACN+Na
M+2K-H	M+2K-H
M+DMSO+H	M+DMSO+H
M+2ACN+H	M+3ACN+2H, M+2ACN+2H, M+2ACN+H
M+IsoProp+Na+H	M+IsoProp+Na+H

Table 8 Fragments' adducts and corresponding precursor ions. Negative ionisation mode.

NEGATIVE IONISATION MODE	
Fragment ion	Precursor ions
M-H	M-H, M+FA-H, M-H-H2O, M-H+HCOONa, 2M-H, M+Hac-H, M+TFA-H, 2M+FA-H, 2M+Hac-H, 3M-H, M-2H, M+Na-2H, M+K-2H, M-3H,
M+Cl	M+Cl
M+FA-H	2M+FA-H, M+FA-H
M-H-H2O	M-H-H2O
M-H+HCOONa	M-H+HCOONa
M+H-H2O	2M+H-H2O, M+H-H2O
M-3H	M-3H
M-2H	M+Na-2H, M-2H, M+K-2H
M+Na-2H	M+Na-2H
M+K-2H	M+K-2H
M+H+2K	M+H+2K
M+Hac-H	M+Hac-H
M+Br	M+Br
M+TFA-H	M+TFA-H

3.3.1 Group features by RT

The features will have a RT only when the user has previously introduced this information. If the user does not introduce the RT, the search about relationships between different features within the RT window cannot be used. In this case, the features are handled independently. Therefore, they will not be grouped together, since there is no relation between them to start any hypothesis about if they come from the same analyte.

To group a set of features by their RT within a tolerance window, the features are sorted by descending RT. The first retention time *rt* will determine the first window, with a width *w*. Since the *rt* is in the middle of the window, the features with RTs within [*rt - w/2, rt + w/2*] will be grouped together in the same FeaturesGroupByRT object. The *rt* value will be updated with the retention time of the next feature that does not fit in the window. When *rt*=0, that single feature will be alone in its own FeaturesGroupByRT object.

3.3.2 Adduct detection

The first adduct autodetection is performed over the information provided by the feature's CS. The CS is a clustering of signals made by Agilent that may contain different adducts. The relation between their masses can be used to calculate which is the original mass whose alterations produced the observed CS. However, any clustering algorithm can be handled to present the information in the way of the CS from Agilent. The CS format is (m/z1, intensity1), (m/z2, intensity2), …, (m/zN, intensityN).

After the adduct detection from the CS, a second autodetection is performed. Within the same set of features grouped by RT, adducts can be detected from the relationship among the features' EMs. Both detections have a similar pipeline. When performing this autodetection, there are three possible situations:

a) More than one feature within the same RT (FeaturesGroupByRT) group have an adduct detected from the CS.

b) Just one feature from the FeaturesGroupByRT has the adduct detected from the CS.

c) No adducts were detected in any of the FeaturesGroupByRT from the CS.

If more than one feature has the adduct autodetected (a), the neutral mass of each feature is calculated. All the features without adduct autodetected will be compared against each of these neutral masses, assuming different hypothetical adducts. Figure 29 illustrates an example where two different features have the adduct detected and, based on their neutral masses, the rest of adducts are determined.

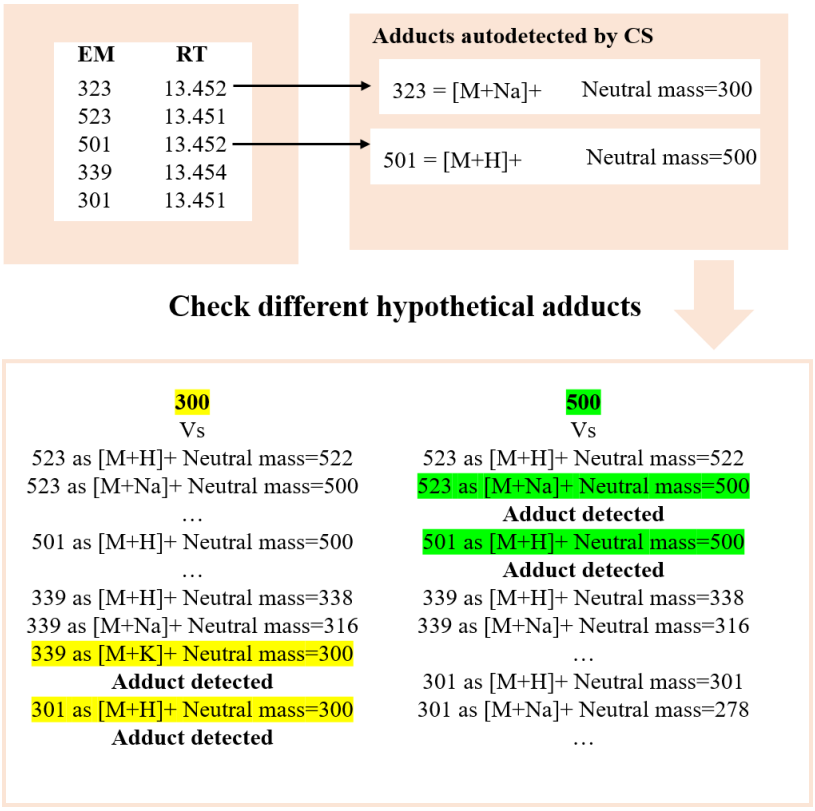


Figure 29 Adduct detection from multiple features in the same RT group.

If just one feature has the adduct detected (b), its neutral mass is calculated and compared against the neutral masses from the rest of the features assuming different hypothetical adducts. It works analogously to (a) but with a single neutral mass.

Finally, if no feature has the adduct detected (c), the detection will be made based on features relation, similar to the adduct detection from the CS. Instead of having one hypothetical loop as in (a) and (b), here there will be two hypothetical adducts loops. In other words, for each feature, different adducts will be assumed, and each corresponding neutral mass will be compared with the neutral masses of the rest of features, assuming different hypothetical adducts.

3.3.3 Group annotations by adduct

Each feature has a list of compounds grouped by adduct (list of CompoundsLCMSGroupByAdduct objects). Each compounds group contains an adduct and a list of compounds. If a feature's adduct is autodetected, its annotations grouped by adduct will consist on a single adduct (the detected one) and a set of annotations

(compounds) with the corresponding neutral mass. For example, if a feature with mass 501 has $[M+H]^+$ as adduct autodetected, its annotations grouped by adduct will be a list of compounds with neutral mass 500 and adduct $[M+H]^+$.

On the other hand, when a feature's adduct is not autodetected, several annotations groups for different adducts will be created. The user can introduce the adducts of interest in CMM interface, which will be explained in more detail in section 3.3.5. If the user selects $[M+H]^+$ and $[M+Na]^+$ adducts, a feature without adduct autodetected and mass 328 would have two annotations groups, one for adduct $[M+H]^+$ and the other for $[M+Na]^+$. The first one would be integrated by compounds with a mass around 327 and the second one by compounds with a mass around 305. The annotations are obtained by querying the neutral mass within a window tolerance against CMM database.

3.3.4 In-source fragmentation search

Once the features are grouped by their RT, some adducts are autodetected if possible, and the annotations grouped by adduct are assigned, the fragments detection can start. From a set of features grouped by their RT, the process begins sorting them by ascending EM. Taking the smallest feature, it can be a fragment of the rest of features with a greater mass. Each feature, from the lightest to the heaviest, is compared against the ones with larger mass and its possibility to be a fragment is checked.

The feature marked as possible fragment is assumed to be different hypothetical adducts. Nevertheless, the adduct of the precursor ion must be capable to form that fragment ion adduct (see Table 7 and Table 8). The hypothetical neutral mass of the possible fragment is calculated and a proton is added or subtracted according to the ionisation mode, since the MS/MS are formed by *m/z* instead of neutral masses. If the fragment and the precursor adducts are coupled, then the precursor's MS/MS from its annotations are checked. When the hypothetical fragment *m/z* matches a compound's MS/MS peak, it means that the feature is a possible fragment of the compound with the hypothetical adduct.

Figure 30 illustrates the possible fragments assignation work-flow with an example analogous to the one in Figure 28.

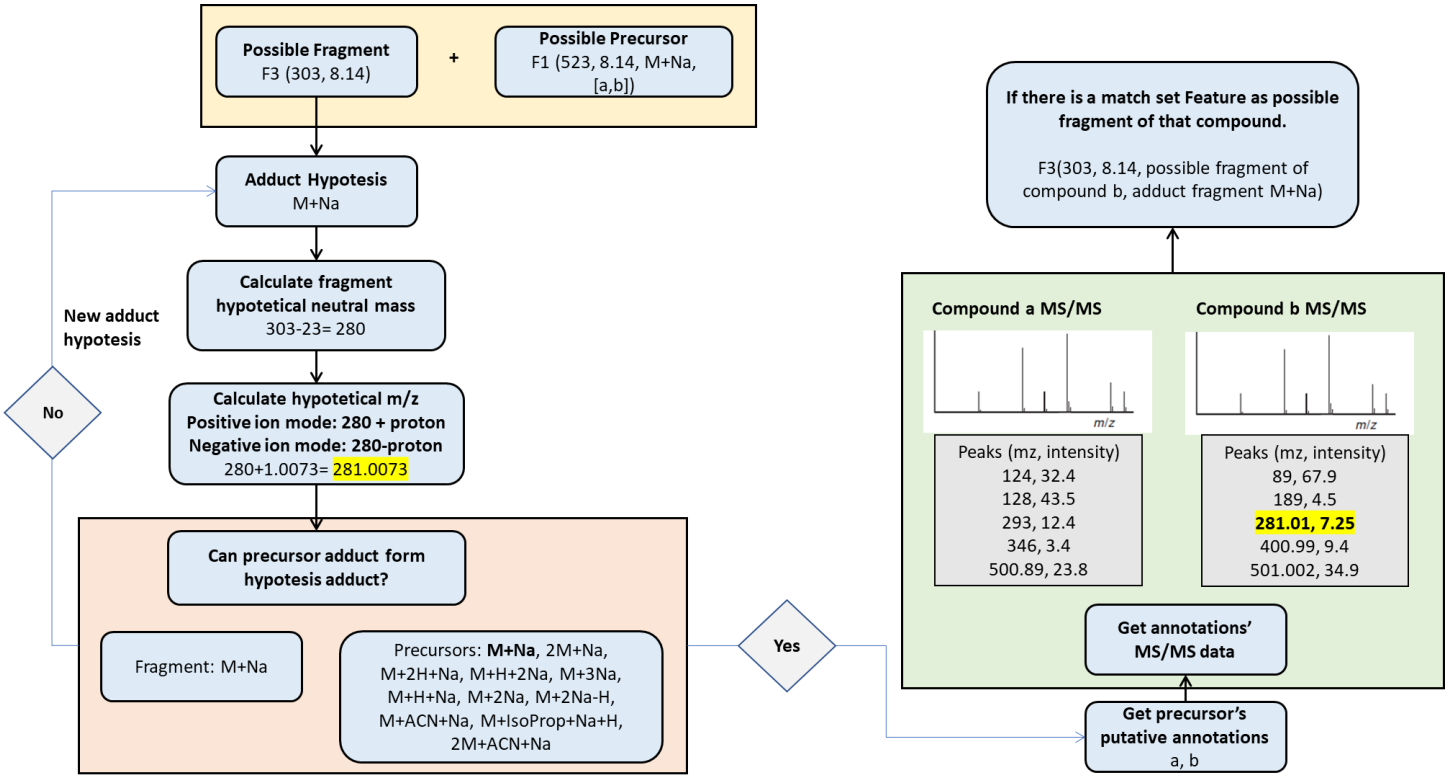


Figure 30 Set Fragments work-flow.

3.3.5 Front-end

The LC-MS search interface is shown in Figure 31. The fields are explained hereunder [58]. They are almost the same as in the advanced batch search. The only mandatory field are the EMs of the significant compounds. The RTs, CS and non-significant features are employed to apply knowledge over the significant ones.

[1] Significant experimental masses (EM): masses (Da) identified as different among the experimental groups during statistical analysis.

[2] Retention time (RT): the units used do not matter since RTs are used for checking relationships between different putative annotations. The RTs introduced here correspond to the EMs introduced in field [1] in the same order.

[3] Composite spectra (CS): spectra created by the summation of all co-eluting *m/z* ions that are related, including isotopes, adducts and dimers formed by the same compound. The CSs introduced here correspond to the EMs introduced in field [1] in the same order.

[4] All experimental masses (EM): all masses (statistically significant and non-significant) found in a particular data set. Statistically non-significant masses provide evidence for supporting or refuting the putative annotations, but are not returned among the results of the query.

[5] All retention times (RT): the RTs introduced here correspond to the EMs introduced in field [4] in the same order.

[6] All composite spectra (CS): the CSs introduced here correspond to the EMs introduced in field [4] in the same order.

[7] RT window with: is the tolerance applied to the features' retention times in order to group them by RT. By default, it is 0.05 min.

[8] Tolerance: tolerance allowed for the putative annotations regarding the statistically significant EM defined as relative (ppm) or absoluted (mDa) value.

[9] Chemical alphabet: possible elements of the putative annotations. This option restricts the returned annotations to only those fulfilling the chosen option. The available options are CHNOPS, CHNOPS + Cl, and all elements. Compounds with deuterium can be filtered or added.

[10] Modifiers: mobile phase modifier used. Depending on this modifier, the adduct formation may change.

[11] Databases: search is performed against databases selected by the user: Kegg, HMDB, LipidMaps, Metlin and/or MINE.

[12] Metabolites: types of metabolites to search. The user can filter the results based on the metabolite type. It may be used for excluding peptides, looking only for

lipids or performing a query over all types of metabolites. CMM considers as lipids the compounds present in LipidMaps.

[13] Masses mode: the user introduces the EM as neutral or *m/z*. Neutral mass search offers three possibilities: true neutral mass search or positive/negative mass search.

[14] Ionisation mode: the user indicates whether the masses were obtained in positive or negative mode. Depending on the ionisation mode, the possible adducts differ.

[15] Adducts: the possible adducts formed when running the experiment. The user may choose between different adducts in negative or positive mode. The list of possible adducts in negative and positive modes are shown in Table 9: All the possible alterations of the mass of the original metabolite (M) given by the selected adducts will be searched by CMM.

Table 9 Possible adducts.

Ionisation Mode	Adducts
Positive	M+H, M+2H, M+Na, M+K, M+NH4, M+H-H2O, M+H+NH4, 2M+H, 2M+Na, M+H+HCOONa, 2M+H-H2O, M+3H, M+2H+Na, M+H+2K, M+H+2Na, M+3Na, M+H+Na, M+H+K, M+ACN+2H, M+2Na, M+2ACN+2H, M+3ACN+2H, M+CH3OH+H, M+ACN+H, M+2Na-H, M+IsoProp+H, M+ACN+Na, M+2K-H, M+DMSO+H, M+2ACN+H, M+IsoProp+Na+H, 2M+NH4, 2M+K, 2M+ACN+H, 2M+ACN+Na
Negative	M-H, M+Cl, M+FA-H, M-H-H2O, M-H+HCOONa, 2M-H, M-3H, M-2H, M+Na-2H, M+K-2H, M+Hac-H, M+Br, M+TFA-H, 2M+FA-H, 2M+Hac-H, 3M-H

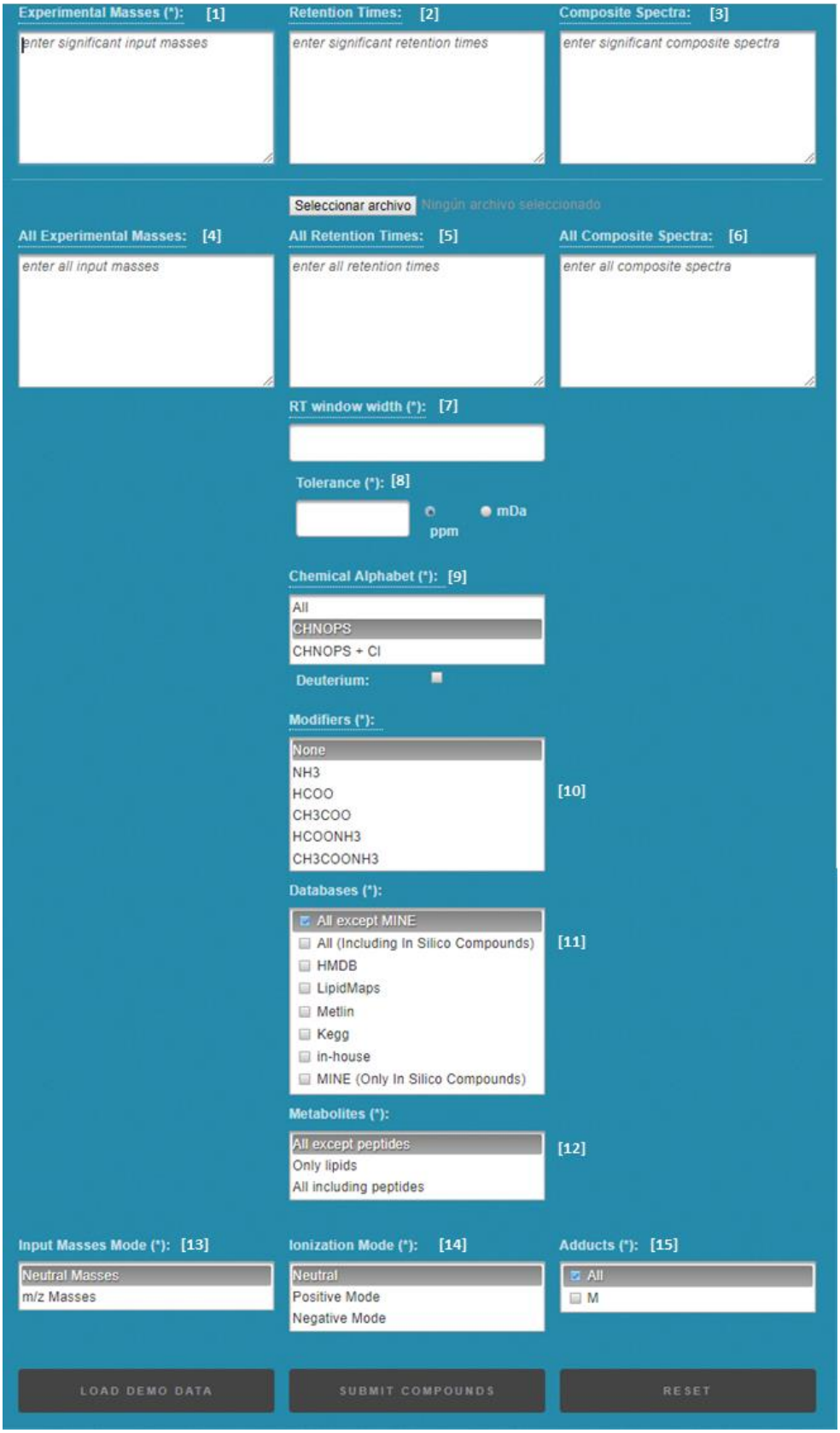


Figure 31 LC-MS search interface.

4 RESULTS

4.1 MS/MS search

4.1.1 MS/MS search algorithms performance

To study the performance of the developed algorithms, four .txt files with 30 metabolites each one was created. These metabolites were taken from MassBank. The first one, a file containing similar spectra to the ones in CMM database named "ToTest.txt". The spectrum is similar when it has been obtained in similar conditions, i.e. same collision energy, ionisation mode, instrument, etc. Secondly, another file containing spectra without the precursor ion peak, named "ToTestLackParentIon.txt". The third file contains spectra with residues. This means that there were peaks from previously analysed compounds contaminating the spectrum. It is called "ToTestWithResidues.txt". The last file "ToTestDifferent.txt", contains spectra that are not in CMM database and were obtained with different instruments and ionisation voltages, i.e. in CMM database we can have a MS/MS from Adenosine obtained by Quattro_QQQ and in "toTestDifferent.txt" there is a MS/MS from Adenosine obtained by LC-ESI-TOF.

The results are indicated as the percentages of hits. To perform the test, all the ionisation voltages were considered, the precursor ion tolerance was 0.1 Daltons and the *m/z* tolerance was 0.5 Daltons. The percentages of hits from ToTest.txt, ToTestLackParentIon.txt, ToTestWithResidues.txt and ToTestDifferent.txt are illustrated in Table 10, Table 11, Table 12 and Table 13 respectively. The top 1 indicates the percentages of the correct identifications against the highest scored annotation. The top 5 is the percentage of the correct identifications against the five best ranked annotations. The "Appears" row indicates the percentage of times that the correct identification has appeared in the whole set of ranked putative annotations.

Table 10 Percentage of hits with spectra from ToTest.txt file.

ToTest.txt						
Experimental				**Experimental and predicted**		
Rank	MetFrag's approach	MyCompoundID's approach	Euclidean Distance	MetFrag's approach	MyCompoundID's approach	Euclidean Distance
Top 1	53.33%	53.33%	60%	43.33%	26.66%	36.66%
Top 5	76.67%	83.33%	83.33%	76.66%	63.33%	66.66%
Appears	83.33%	83.33%	83.33%	83.33%	83.33%	83.33%

Table 11 Percentage of hits with spectra from ToTestLackParentIon.txt file.

ToTestLackParentIon.txt						
Experimental				**Experimental and predicted**		
Rank	MetFrag's approach	MyCompoundID's approach	Euclidean Distance	MetFrag's approach	MyCompoundID's approach	Euclidean Distance
Top 1	46.66%	46.66%	50%	43.33%	23.33%	23.33%
Top 5	66.66%	66.66%	66.66%	63.33%	40%	36.66%
Appears	66.66%	66.66%	66.66%	66.66%	66.66%	66.66%

Table 12 Percentage of hits with spectra from ToTestWithResidues.txt file.

ToTestWithResidues.txt						
Experimental				**Experimental and predicted**		
Rank	MetFrag's approach	MyCompoundID's approach	Euclidean Distance	MetFrag's approach	MyCompoundID's approach	Euclidean Distance

Top 1	56.66%	53.33%	53.33%	43.33%	23.33%	20%
Top 5	76.66%	76.66%	76.66%	70%	56.66%	56.66%
Appears	76.66%	76.66%	76.66%	76.66%	76.66%	76.66%

Table 13 Percentage of hits with spectra from ToTestDifferent.txt file.

ToTestDifferent.txt						
Experimental				**Experimental and predicted**		
Rank	MetFrag's approach	MyCompoundID's approach	Euclidean Distance	MetFrag's approach	MyCompoundID's approach	Euclidean Distance
Top 1	36.66%	33.33%	30%	30%	13.33%	16.66%
Top 5	70%	66.66%	66.66%	63.33%	46.66%	46.66%
Appears	70%	70%	70%	73.33%	73.33%	73.33%

A comparison of the algorithms' performance *vs* real software tools has been done to test the suitability of the integrated feature in CMM. Each of the 30 MS/MS data from "ToTest.txt" were searched in HMDB, MetFrag and MyCompoundID. In HMDB we could establish two comparisons, querying over experimental data and querying over experimental and predicted data. However, MetFrag uses a different approach and it does not allow to perform searches over experimental spectra. It makes an *in-silico* fragmentation of the putative annotations (MS^1) to match them against the input spectra. On the other hand, MyCompoundID searches against experimental spectra and *in-silico* fragmented spectra, but for this testing only its experimental search was considered. Therefore, MyCompoundID tool and HMDB are included in Table 14, where the performance of the different MS/MS search methods against experimental spectra is illustrated. MetFrag and HMDB are visible in Table 15, where the search was done over predicted and experimental spectra.

The filters used for each search were the same used to test CMM MS/MS search algorithms; 0.1 Daltons as the precursor ion tolerance and 0.5 Daltons as the *m/z* tolerance. But in MetFrag the tolerances cannot be set in Daltons, therefore, 100 ppm were used as the precursor ion tolerance.

Table 14 Algorithms *vs* real software over experimental MS/MS.

ToTest.txt					
Experimental					
Rank	MetFrag's approach	MyCompoundID's approach	Euclidean Distance	HMDB tool	MyCompoundID tool
Top 1	53.33%	53.33%	60%	63.33%	23.33%
Top 5	76.67%	83.33%	83.33%	76.66%	66.66%
Appears	83.33%	83.33%	83.33%	86.66%	86.66%

Table 15 Algorithms *vs* real software over experimental and predicted MS/MS.

ToTest.txt					
Experimental and predicted					
Rank	MetFrag's approach	MyCompoundID's approach	Euclidean Distance	HMDB tool	Metfrag tool
Top 1	43.33%	26.66%	36.66%	43.33%	23.33%
Top 5	76.66%	63.33%	66.66%	63.33%	46.66%
Appears	83.33%	83.33%	83.33%	70%	86.66%

4.1.2 MS/MS search output

The output from CMM MS/MS based search provides a list of ranked metabolites' identifications. The demo's data (see Figure 32) result is illustrated in Figure 33 with MetFrag's approach, in Figure 34 with MyCompoundID's approach and in Figure 35 with the inverse of the Euclidean distance. This demo data is the same as in HMDB LC-MS/MS search feature. HMDB's output can be seen in Figure 36. Every tool return L-Glutamine as first identification. Nevertheless, the rest of the putative identifications are ranked differently in each procedure.

By inserting Quercentin's MS/MS spectrum from MassBank database (see Figure 37), the output was similar. The three approaches returned Quercentin as the highest scored putative identification (see Figure 38, Figure 39 and Figure 40), but the rest were ranked differently for each algorithm. The results returned by HMDB are in Figure 41.

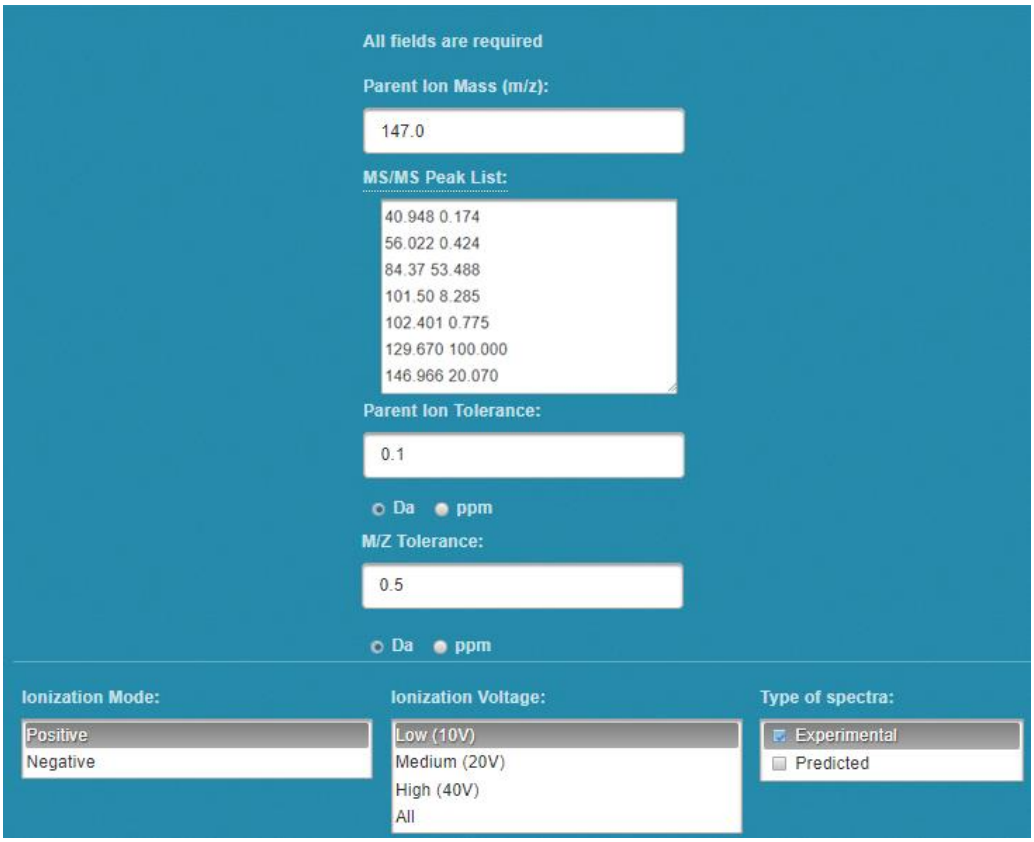


Figure 32 CMM MS/MS based search interface. Demo data.

Results			
Name	Formula	Mass	Score
L-Glutamine	C5H10N2O3	146,0691	0,9329
2-methyl-glutaric acid	C6H10O4	146,0579	0,9129
D-Glutamine	C5H10N2O3	146,0691	0,8690
Coumarin	C9H6O2	146,0368	0,5771
Phenylpropiolic acid	C9H6O2	146,0368	0,5771
Methylglutaric acid	C6H10O4	146,0579	0,3340

Figure 33 CMM MS/MS based search. Demo data result (MetFrag's approach).

Results			
Name	Formula	Mass	Score
L-Glutamine	C5H10N2O3	146,0691	0,8207
D-Glutamine	C5H10N2O3	146,0691	0,6923
2-methyl-glutaric acid	C6H10O4	146,0579	0,6225
Methylglutaric acid	C6H10O4	146,0579	0,3555
Coumarin	C9H6O2	146,0368	0,2795
Phenylpropiolic acid	C9H6O2	146,0368	0,2795

Figure 34 CMM MS/MS based search. Demo data result (MyCompoundID's approach).

Results			
Name	Formula	Mass	Score
L-Glutamine	C5H10N2O3	146,0691	0,1109
D-Glutamine	C5H10N2O3	146,0691	0,0700
Coumarin	C9H6O2	146,0368	0,0125
Phenylpropiolic acid	C9H6O2	146,0368	0,0125
Methylglutaric acid	C6H10O4	146,0579	0,0102
2-methyl-glutaric acid	C6H10O4	146,0579	0,0074

Figure 35 CMM MS/MS based search. Demo data result (Euclidean Distance).

Name/CAS Number	Weight/Formula	Fit(%)	RFit(%)	Purity(%)
L-Glutamine (HMDB0000641)	146.1445	1.00	1.00	1.00
D-Glutamine (HMDB0003423)	146.1445	1.00	0.98	0.99
Methylglutaric acid (HMDB0000752)	146.1412	0.56	0.63	0.66
Coumarin (HMDB0001218)	146.145	0.79	0.36	0.50
Phenylpropiolic acid (HMDB0002359)	146.1427	0.57	0.36	0.49
2-Methylglutaric acid (HMDB0000422)	146.1412	N/A	0.54	0.50

Figure 36 HMDB LC-MS/MS search. Demo data result.

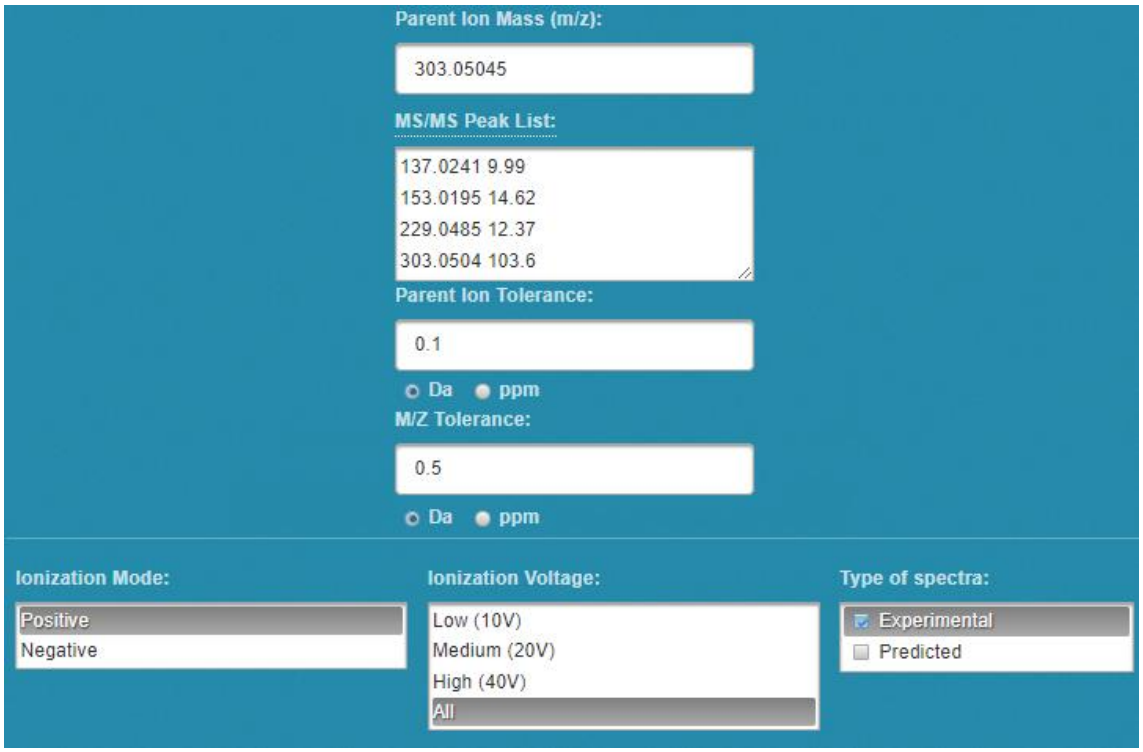


Figure 37 CMM MS/MS based search interface. Quercentin MS/MS data.

Results

Name	Formula	Mass	Score
Quercetin	C15H10O7	302,0427	1,0000
Morin	C15H10O7	302,0427	0,9997
Tricetin	C15H10O7	302,0427	0,9997
5,7,3'-Trihydroxy-4'-methoxyflavanone	C16H14O6	302,0790	0,9995
Hesperetin	C16H14O6	302,0790	0,9995
Ethacrynic acid	C13H12Cl2O4	302,0113	0,9918
Ellagic acid	C14H6O8	302,0063	0,9781

Figure 38 CMM MS/MS based search. Quercentin data result (MetFrag's approach).

Results

Name	Formula	Mass	Score
Quercetin	C15H10O7	302,0427	1,0000
Morin	C15H10O7	302,0427	0,9962
5,7,3'-Trihydroxy-4'-methoxyflavanone	C16H14O6	302,0790	0,9511
Tricetin	C15H10O7	302,0427	0,9468
Hesperetin	C16H14O6	302,0790	0,9396
Ethacrynic acid	C13H12Cl2O4	302,0113	0,8449
Ellagic acid	C14H6O8	302,0063	0,7763

Figure 39 CMM MS/MS based search. Quercentin data result (MyCompoundID's approach).

Results

Name	Formula	Mass	Score
Quercetin	C15H10O7	302,0427	1,0000
Ellagic acid	C14H6O8	302,0063	0,0004
Hesperetin	C16H14O6	302,0790	0,0000
Ethacrynic acid	C13H12Cl2O4	302,0113	0,0000
5,7,3'-Trihydroxy-4'-methoxyflavanone	C16H14O6	302,0790	0,0000
Morin	C15H10O7	302,0427	0,0000
Tricetin	C15H10O7	302,0427	0,0000

Figure 40 CMM MS/MS based search. Quercentin data result (Euclidean Distance).

Name/CAS Number	Weight/Formula	Fit(%)	RFit(%)	Purity(%)
Quercetin (HMDB0005794)	302.2357	1.00	1.00	1.00
Morin (HMDB0030796)	302.2357	1.00	1.00	1.00
Ellagic acid (HMDB0002899)	302.194	1.00	0.99	1.00
Hesperetin (HMDB0005782)	302.2788	1.00	0.96	0.98
5,7,3'-Trihydroxy-4'-methoxyflavanone (HMDB0030746)	302.282	1.00	0.81	1.00
Ethacrynic acid (HMDB0015039)	303.138	1.00	0.74	1.00
Tricetin (HMDB0029620)	302.2357	0.84	0.76	0.84

Figure 41 HMDB LC-MS/MS search. Quercentin MS/MS data.

4.2 LC-MS search output

CMM LC-MS search output grouping the features by their RT is a pair of nested lists. The outer list corresponds to the different retention times while the inner list consists on; first the set of annotations grouped by adduct and secondly the possible precursor ions only if the feature has been detected as a possible fragment. An example has been executed to show the functionality of the feature. The input data for the example is in Figure 42. The adducts selected were $[M+H]^+$ and $[M+Na]^+$ to avoid obtaining a complex result for the explanation. The input masses mode was *m/z* with positive ionisation mode. Part of the output is illustrated in Figure 43 and Figure 44. The nested lists and sections from those figures are explained hereunder:

[1] Outer list: consist on the different retention times. Since in the input we had two distinct retention times, this outer list only has two tabs.

[2] Inner list: consist on the different features grouped within a retention time. In the example, since we are in RT 18.842525 there are three tabs for the three features with that retention time.

[3] Annotations grouped by adduct: the first part of the inner list illustrates the different annotations grouped by the selected adducts of a feature. In Figure 43 three putative annotations with adduct M+Na from feature with mass 192.0743 can be appreciated.

[4] Possible precursor ions: the second part of the inner list illustrates the possible precursor ions of the actual feature and its corresponding adduct. In Figure 44 it is visible that the feature with mass 90.021938 does not have annotations but is a possible fragment of two different annotations of the feature from Figure 43. The feature of the precursor ion mass is inside the red square.

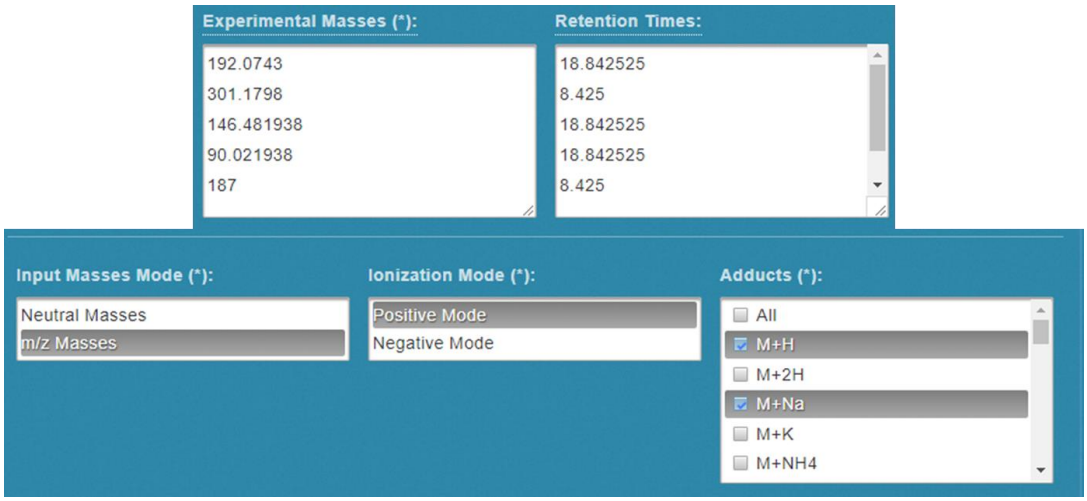


Figure 42 Input data for LC-MS search.

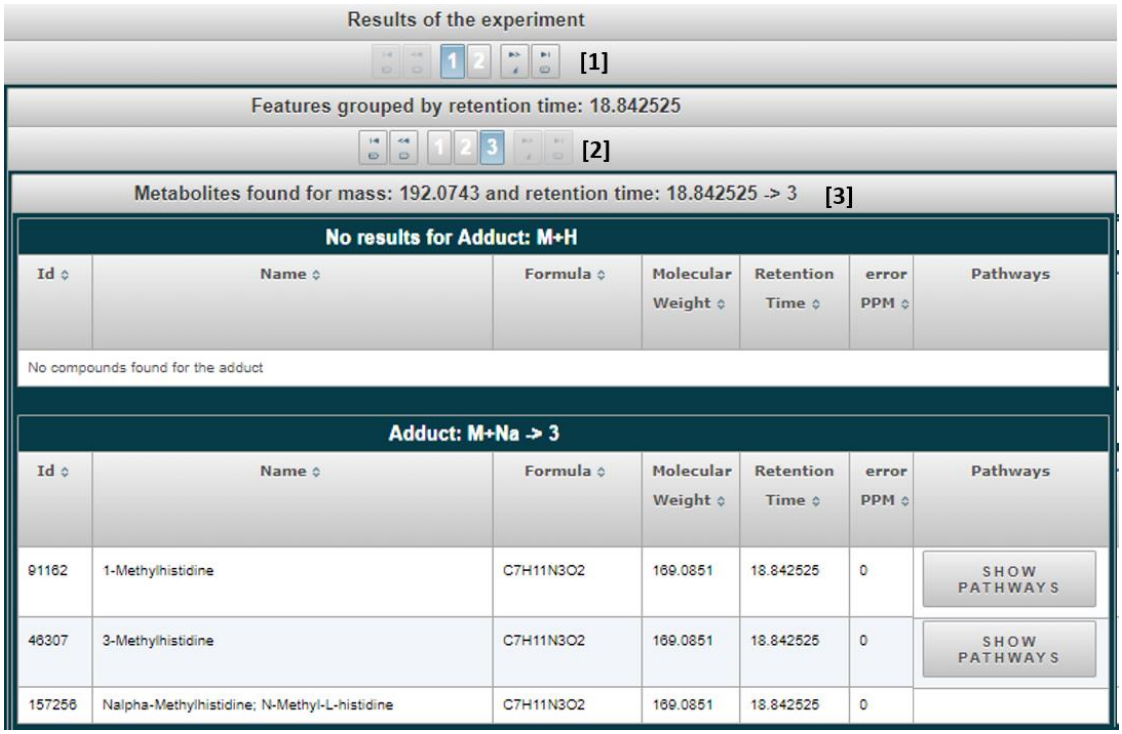


Figure 43 LC-MS output. Annotations grouped by adduct from feature with mass 192.0743.

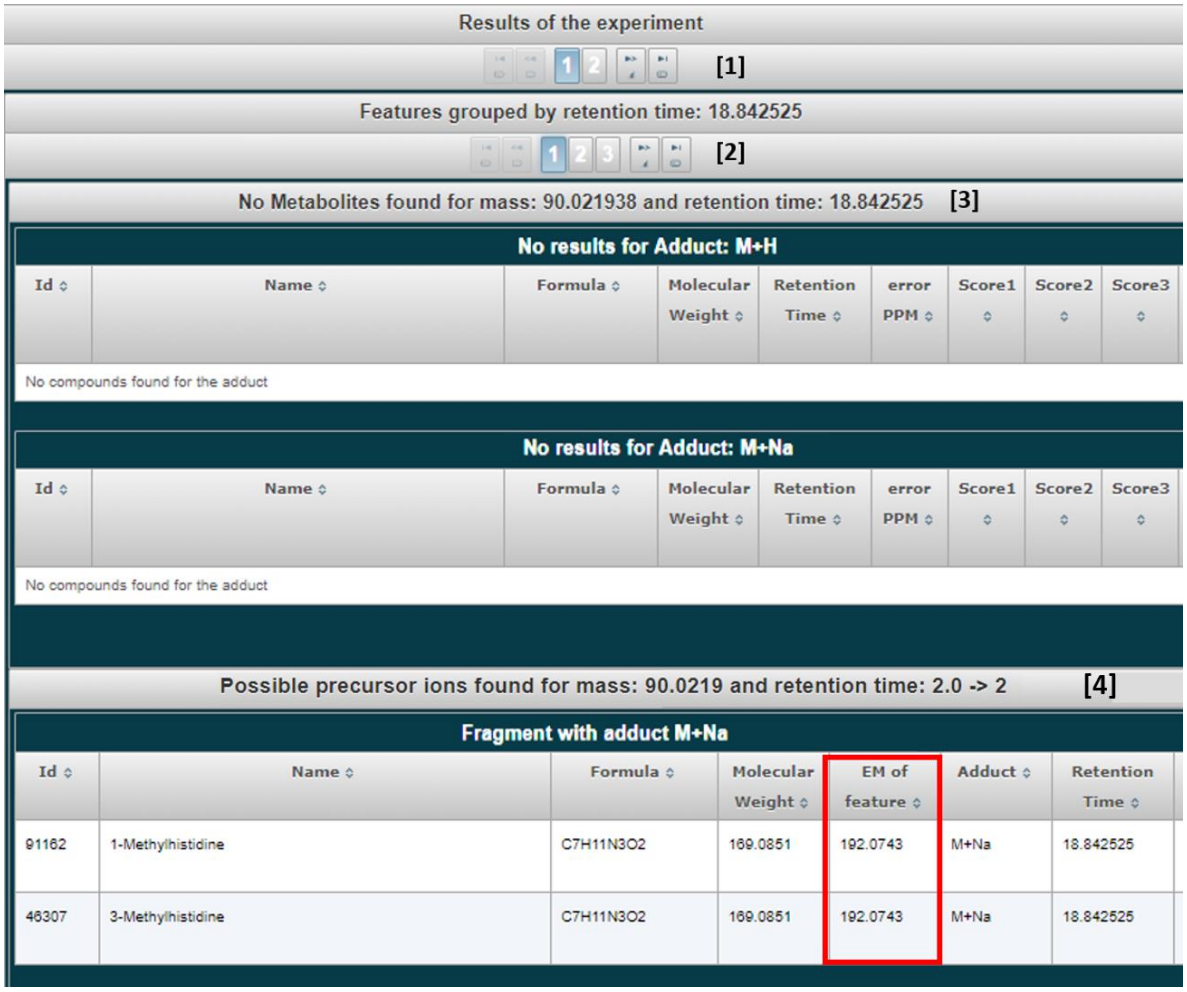


Figure 44 LC-MS output. Possible precursor ions of the feature with mass 90.021938 considered a fragment.

4.3 Execution time

While testing the LC-MS search capacity, we noticed that the execution was extremely slow. The compounds for the different features are retrieved from the CMM database. Each compound has one or more lipids classification, a lipid maps classification, a set of pathways, a structure and a set of classifier classifications. Each of these compound's attributes are filled by accessing to the database, which has a high computational cost.

4.3.1 Execution time results

The execution time for features creation was measured to identify the major cause of the problem. The LC-MS based search feature was tested with 100, 1,000 and 1,400 masses. When submitting 5,000 masses an error occurred, avoiding the time measurement and obtaining results. The measured times were the time employed accessing to the compounds' data from the database (JDBC time) and the time to create the features' annotations. These times are illustrated graphically in Figure 45.

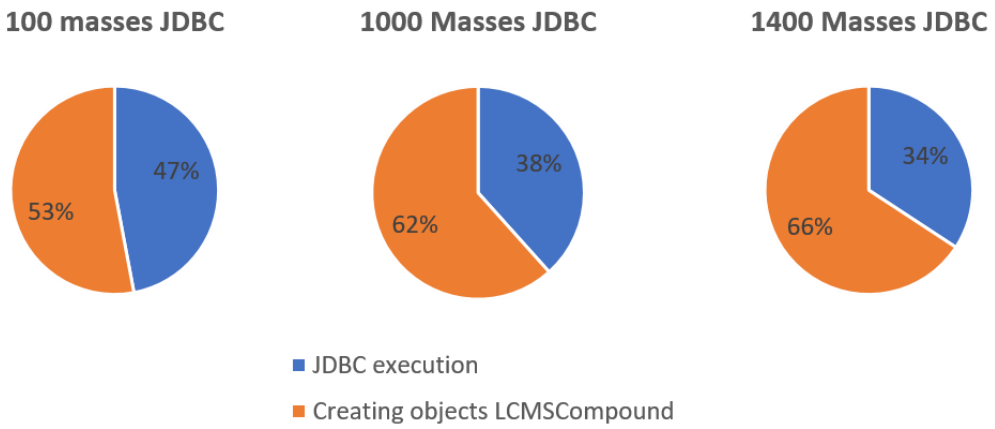


Figure 45 Execution time for features creation in LC-MS search.

Figure 45 shows the critical step in the execution: the LCMSCompound objects creation. To focus on this problem, we proceed to measure the different steps within the objects creation. The time creation for the lipids classification, lipid maps classification, structure, classyfire classifications and pathways creation, each with their corresponding database retrieval, was measured separately to reach a deeper view of the problem. In Figure 46 it is appreciable that the more time-consuming objects creations were the classyfire classifications, the lipids classification and the structure.

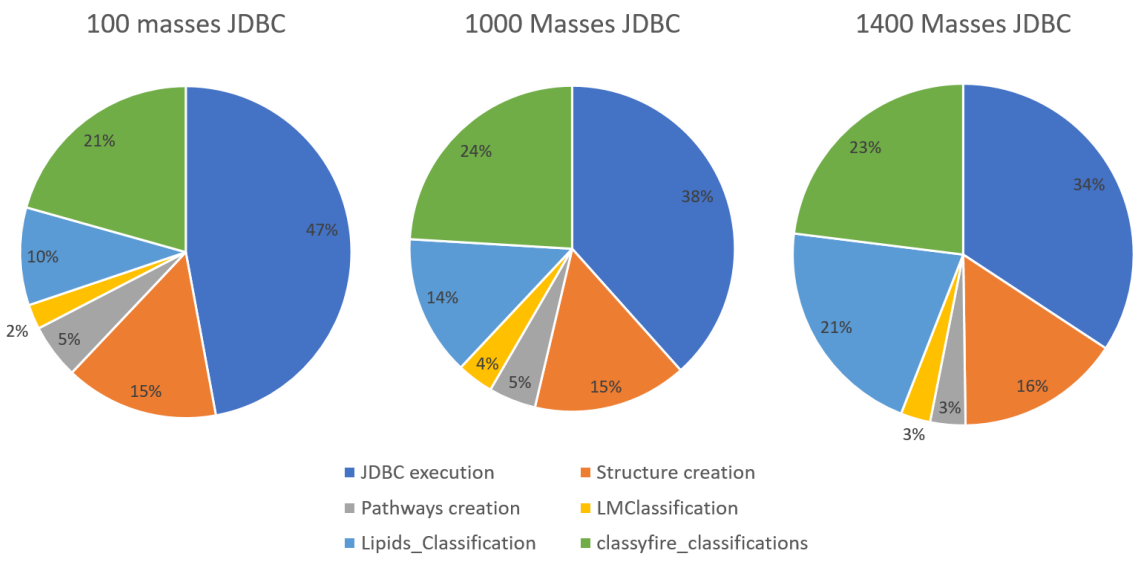


Figure 46 Execution time for creating features split by attributes.

4.3.2 Performance improvement

After the analysis of the processing time, we noticed that the cause of the slowness was on the creation of the features' annotations. The issue was due to the access to different tables in the database to fill the attributes of each compound. This data access is an undesirable loss of efficiency that can be solved by the creation of a view containing all the information required for the compounds creation.

Therefore, a view containing the compounds and its corresponding classifications data (in the cases where the relation was N:1) was created (see Figure 48). This view, named as compounds_view do not have the classyfire classification data since it is not used in the LC-MS search. Moreover, the pathways information is not contained in compounds_view neither since the pathways-compounds relation is many to many whilst LM_Classification, Lipids_Classification and structure is of many compounds to one classification. Therefore, for the pathways creation the database access has been kept the same. The execution time results after this improvement are visible in Figure 47 and a comparison between the times before and after the creation of the view is presented and discussed at 5.3.

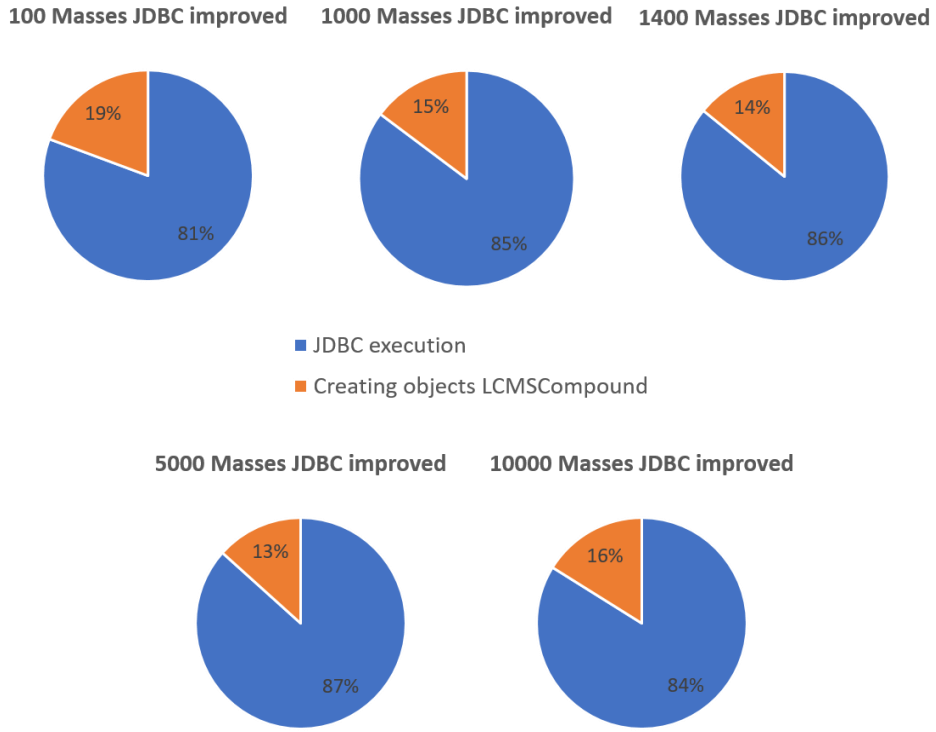


Figure 47 Execution time for features creation after the creation of a view.

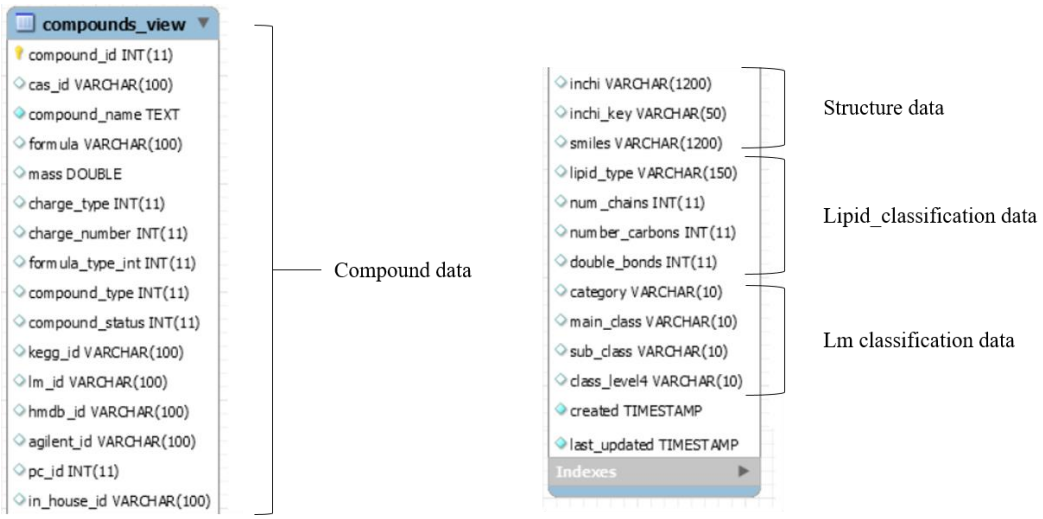


Figure 48 Compounds_view definition.

5 DISCUSSION

5.1 MS/MS search algorithms results

We calculated the average percentages of correct identifications for each scoring approach applied to CMM MS/MS search. These percentages were obtained from the tables of 4.1.1. Figure 49 illustrates the percentage of compounds identified from each algorithm when searching against the experimental spectra whilst the percentage of compounds identified when searching against experimental and predicted spectra is shown in Figure 50.

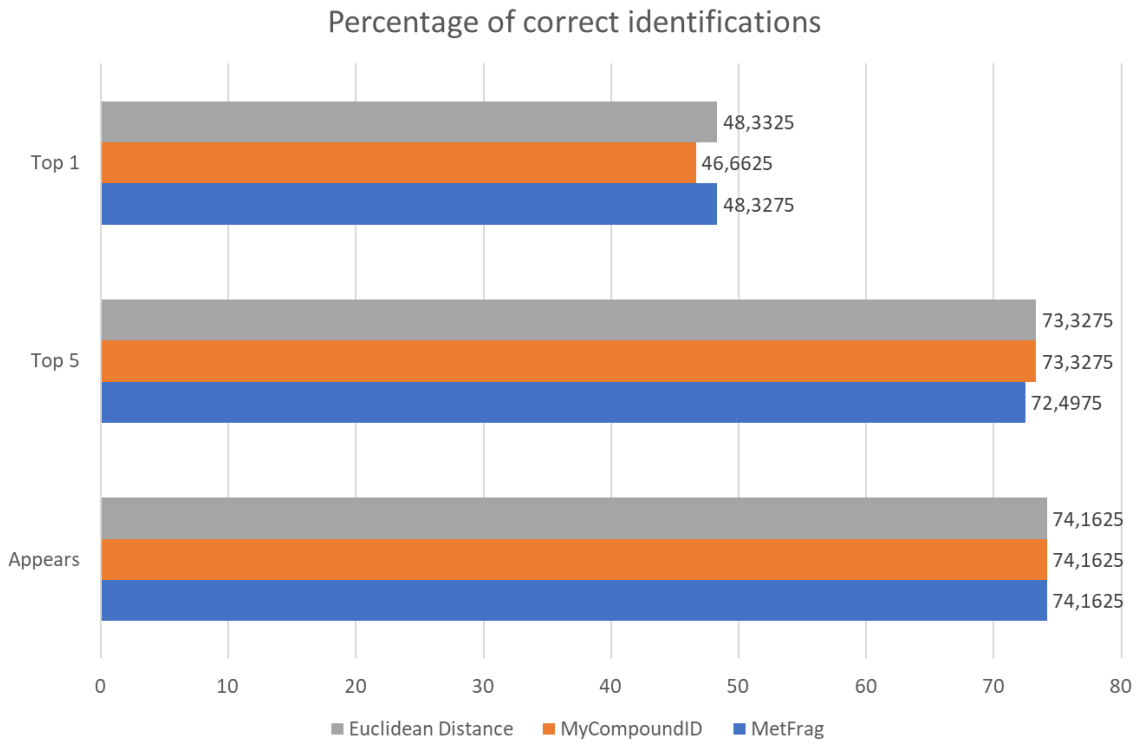


Figure 49 MetFrag, MyCompoundID and Euclidean Distance approaches. Percentage of compounds identified using only experimental spectra.

Figure 49 illustrates that the three procedures provide a very similar percentage of correct identifications. Both MetFrag and Euclidean Distance approaches get the correct metabolite as the first ranked almost half of the times. MyCompoundID is the one with worst percentage of correct identifications at top 1 hits, with less than a 2% of difference over the other two. Regarding the percentage of matches within the top 5

identifications, MetFrag approach has a lower rate of correct identifications of 1% compared to the Euclidean Distance and MyCompoundID approaches. We do not consider this difference significant. Finally, it is appreciable that the correct identification is present within the ranked output the 74.1625% of the times for every approach. From this plot it is reasonable to say that there are not significant differences between the three approaches for the metabolite identification based on MS/MS data when searching against experimental data.

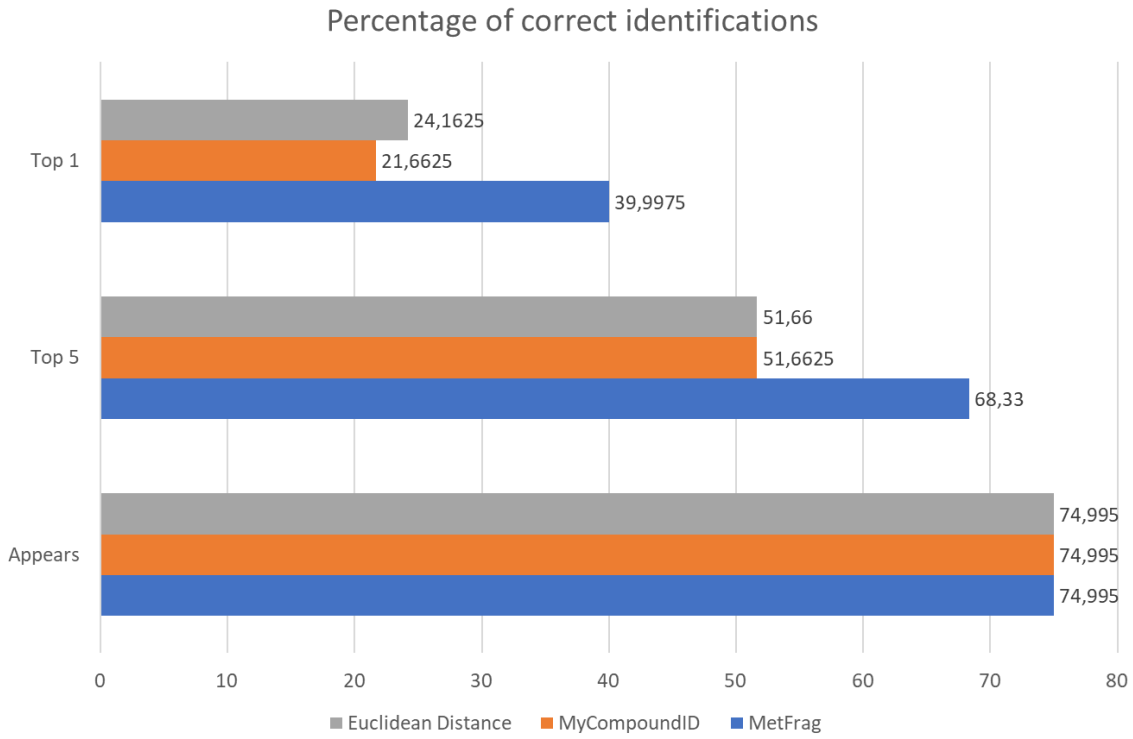


Figure 50 MetFrag, MyCompoundID and Euclidean Distance approaches. Percentage of compounds identified using experimental and predicted spectra.

Nevertheless, by visualising Figure 50 the number of correct identification using different approaches changes. When doing spectral matching against experimental MS/MS the behavior of the three procedures is very similar. However, when searching over experimental and predicted data, MetFrag's approach obtains better results than the other ones. For the top 1 identifications, MetFrag's approach is the one with a higher percentage of correct identifications, 39.9975%, whilst MyCompound and Euclidean distance obtained a 21.6625% and a 24.1625% of hits respectively. The presence of the correct identification in the top 5 shows also that MetFrag's approach has identified

correctly 17% more compounds than the other algorithms. The three approaches contain the correct identification within their ranked annotations almost 75% of the times, but the score assigned by each one is different.

It makes sense that MetFrag's method is the one that offers a higher percentage of correct identifications over both experimental and predicted spectra. As it was already mentioned in 3.1.3, MetFrag's approach assigns a weight of three to the *m/z* dot product whilst the other approaches assign the same weight to the *m/z* and to the intensity. When using a greater number of spectra (experimental and predicted), there is a higher probability of comparing against spectra with very similar fragmentation patterns, but MetFrag assigns a heavier weight to the precursor ion *m/z* leading to a higher number of correct identifications.

After seeing the behavior against the predicted and the experimental spectra, this remarkable difference between approaches made us consider the MetFrag's method as the one that suits better in CMM for the MS/MS based search. For this reason, it is the one that is currently implemented in the tool. Nevertheless, the scoring method can be easily changed at CMM modifying the call to the different functions created.

The developed methods were also compared against some available tools. The 30 spectra from "ToTest.txt" were introduced in HMDB, MyCompoundID and MetFrag MS/MS searches. These results are illustrated graphically in Figure 52 and Figure 53. They show the matching results against only the experimental data and against experimental and predicted data respectively. Figure 52 illustrates the percentages of identifications over experimental spectra for the developed approaches, HMDB tool and MyCompoundID tool. MetFrag is not included since it does not search over experimental data, it performs an *in-silico* fragmentation. Figure 53 shows the percentages of identifications when searching against experimental and predicted spectra for the developed approaches and MetFrag tool. Here, MetFrag is included, but MyCompoundID is not because for this testing its fragmentation functionality was not used to obtain results. Moreover, there was a failure in HMDB that did not allow to obtain valid results for its MS/MS based search over experimental and predicted spectra. The error is illustrated in Figure 51. Therefore, it was not included in Figure 53 neither. To avoid confusion while explaining the results, we will add the suffix "tool"

when talking about available tools and "approach" when referring to CMM developed algorithms.

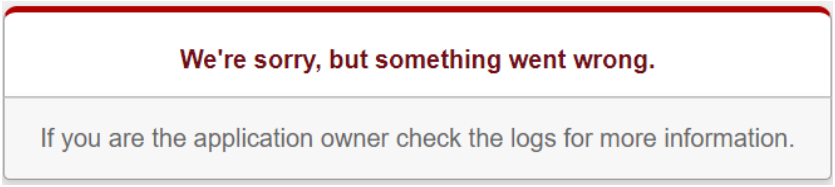


Figure 51 HMDB's error message.

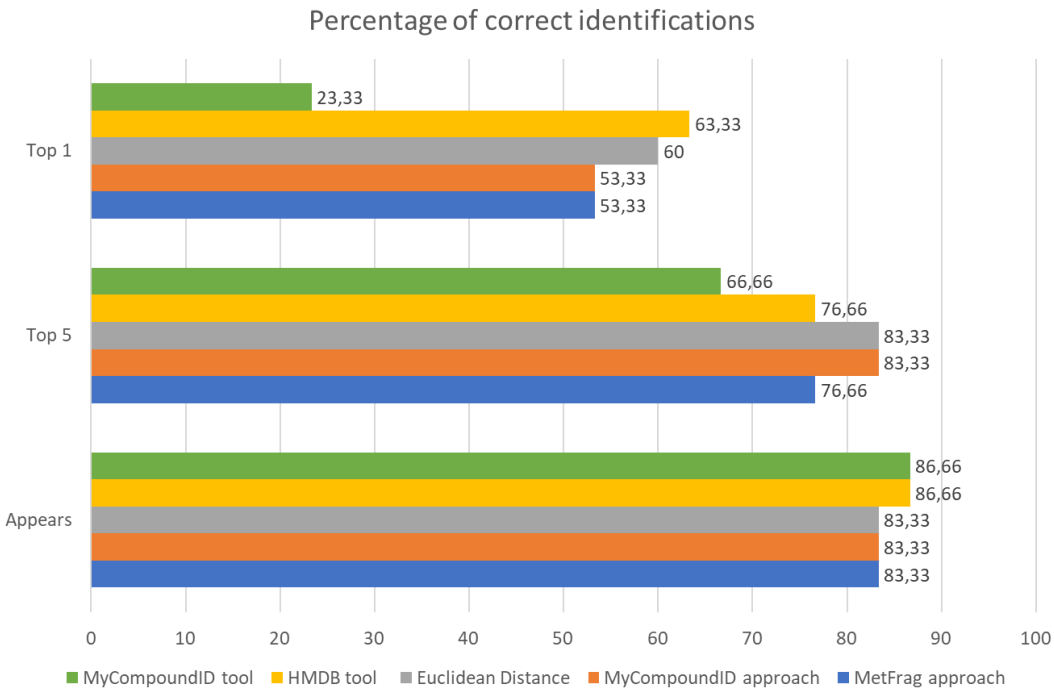


Figure 52 Percentage of correct annotations of CMM approaches and other available tools using only experimental spectra.

In Figure 52 we can notice that the highest percentage of correct annotations ranking as top 1 is achieved by HMDB tool with a 63.33% of correct identifications. However, the behavior of the developed methods is not excessively far from the HMDB tool. The Euclidean distance provides a 3% lower percentage of correct identifications while MetFrag's and MyCompoundID's approaches obtain a 10% lower percentage.

It is noticeable that MyCompoundID tool has a lower number of correct annotations than the developed approaches. This difference is probably due to the recent

HMDB database update (see [44]). MyCompoundID has not been updated since 2013 as well as its database.

The percentage of correct annotations in the top 5 shows that MyCompoundID tool's results improve with a 66.66% of hits. HMDB is still the external tool with a higher precision with a 76.66 % of compounds correctly identified in the top 5. Nevertheless, it is overcome by the developed methods. MyCompoundID's approach and the Euclidean distance obtain an 83.33% of hits followed by MetFrag's approach with a 76.66%.

Finally, the two tools deliver the correct annotation within the whole output list 86.66% of the times. On the other hand, the three developed approaches have a 3% lower precision than the external tools tested.

For the identification using both experimental and predicted spectra we included MetFrag tool since it performs an *in-silico* fragmentation over experimental compounds from different sources, therefore it does not use experimental spectra. The source database for the search of the precursor ions can be specified. For this project, the HMDB database was used as a source for the precursor ions in MetFrag tool, in order to make the comparison in the most similar conditions as possible. The results of querying against the predicted and the experimental spectra are visible in Figure 53. Unfortunately, HMDB had to be deleted from the comparison due to their availability problems during the last month. As it was previously explained, the spectra search was not available, so it has not been possible to compare the precision of the MS/MS search against the other methods.

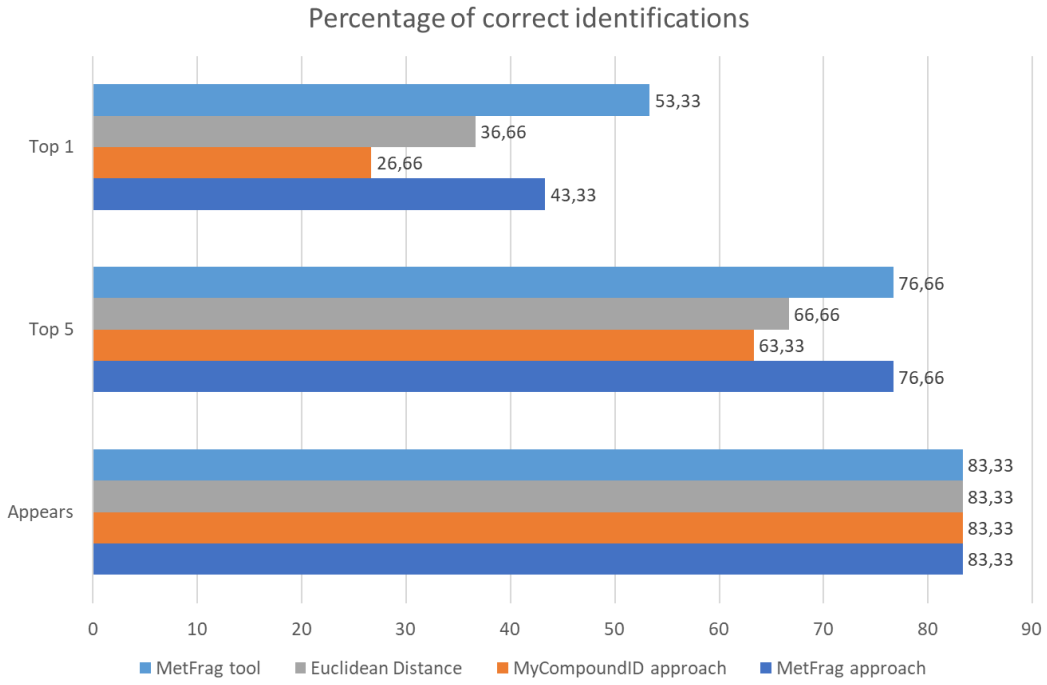


Figure 53 Percentage of correct annotations by CMM and other available tools using experimental and predicted spectra.

It can be seen in Figure 53 that the best first ranked identification percentage is achieved by the MetFrag tool with a 53.33% of correct identifications. It is followed by MetFrag's approach with a 43.33% of hits. The Euclidean Distance's approach identified correctly a 36.66% of the metabolites and MyCompoundID's approach did a correct annotation over 26.66% of the metabolites.

Within the top 5 category, the highest number of metabolites identified is reached by MetFrag's approach and MetFrag tool with a 76.66% of them. The Euclidean Distance's algorithm correctly identifies a 66.66% of the metabolites. MyCompoundID's approach got a 63.33% of correct identifications. Finally, the correct identification appears in the output the 83.33% of the times for every approach.

In Figure 52 and Figure 53 we can see that the developed approaches are not far of the results obtained in the already available tools. Nevertheless, we do believe that some improvements in MS/MS based search metabolite identification can still be reached.

5.2 MS/MS search output

To prove graphically that the developed approaches are appropriate for CMM MS/MS based search, the following figures (Figure 54, Figure 55, Figure 56 and Figure 57) illustrate the first ranked output spectrum against the input spectrum previously mentioned at 4.1.2. The spectra comparison was taken from HMDB web tool.

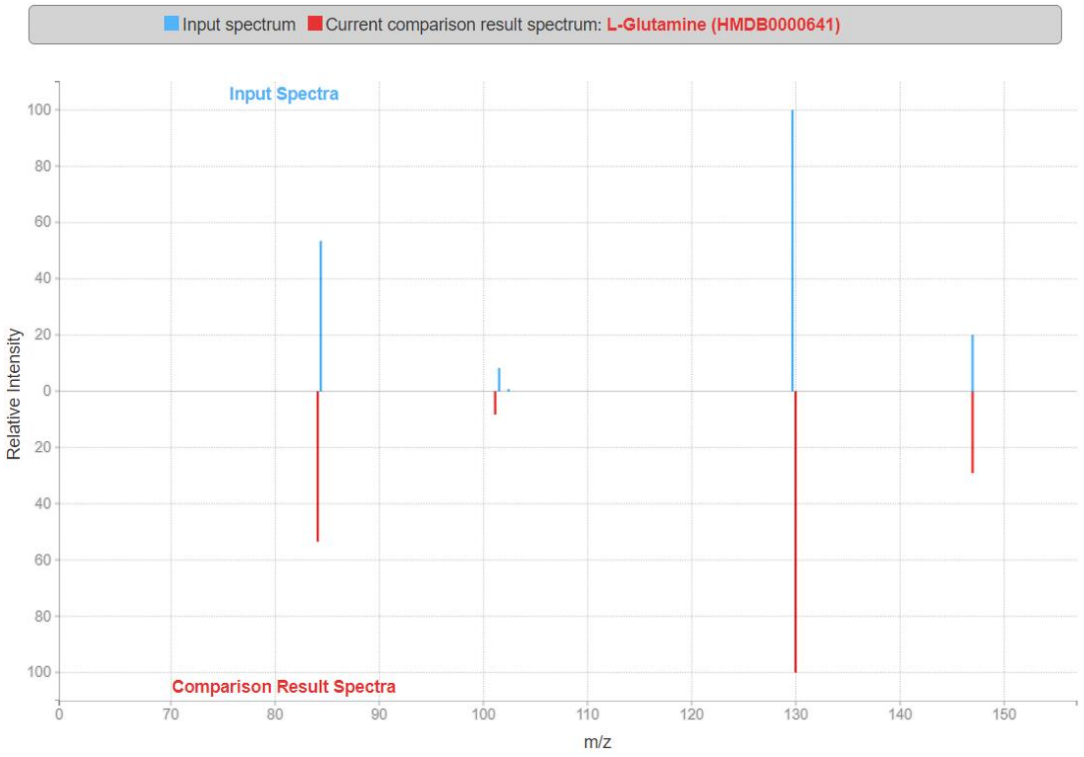


Figure 54 Input spectrum is the demo spectrum from CMM (same as HMDB demo data) in blue, compared against L-Glutamine database spectrum (red).

Figure 54 illustrates the input spectrum (blue) and the database spectrum (red). It can be appreciated the similarity between them. This input spectrum is correctly identified by every approach (MetFrag approach, MyCompoundID approach, Euclidean distance and HMDB MS/MS based search) with the demo data as input. The second best scored identification was D-Glutamine for every method except for MetFrag approach, where it was 2-Methylglutaric acid. The comparisons of the input spectrum against both metabolite's MS/MS are illustrated in Figure 55 and Figure 56 respectively.

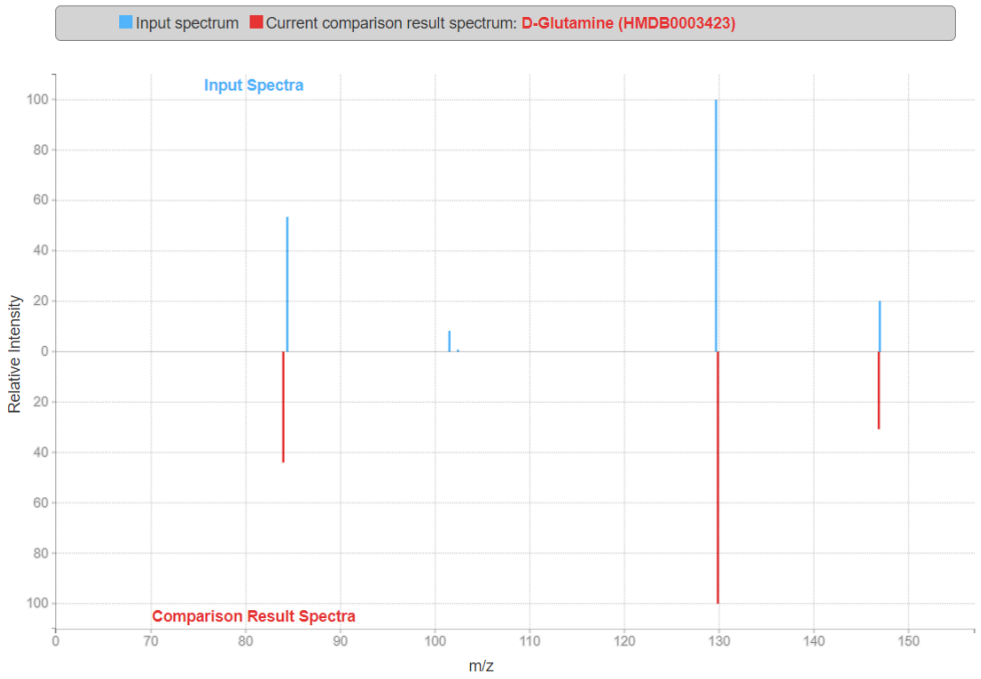


Figure 55 Input spectrum is the demo spectrum from CMM (same as HMDB demo data) in blue, compared against D-Glutamine database spectrum (red).

In Figure 55 we can also appreciate a significant similarity between the input spectrum and D-Glutamine spectrum. We can visualise three overlapping peaks. Nevertheless, this database spectrum does not have a peak near *m/z*=100 whilst L-Glutamine spectrum does. The spectra from Figure 56 compares the input data against 2-Methylglutaric acid database MS/MS. This spectrum is not as similar as the two previous ones, but it has four peaks matched with the input spectrum, one more that D-Glutamine. Since MetFrag weights the matched *m/z* by three, 2-Methyglutaric acid overcomes D-Glutamine score in this approach because they have three and four peak matches respectively. Figure 56 shows that the *m/z* near to 100 is present in the 2-Methyglutaric spectrum. The second example is illustrated in Figure 57, where the Quercentin spectrum from MassBank was the input. We can see a very likely match between the input spectrum and the matched database spectrum.

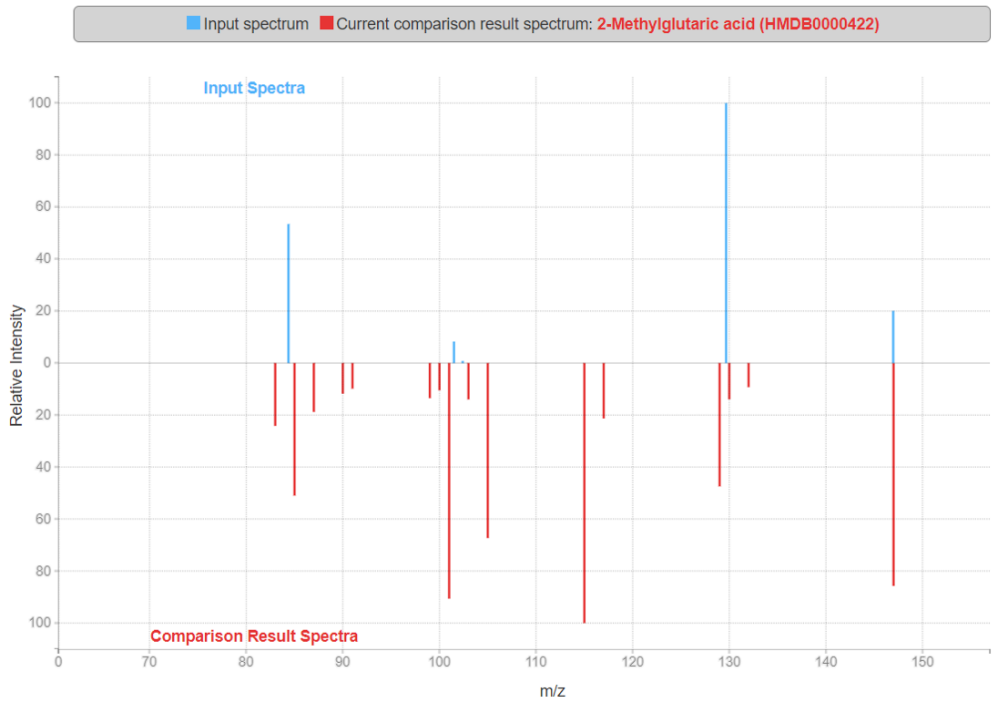


Figure 56 Input spectra is the demo spectrum from CMM (same as HMDB demo data) in blue, compared against 2-Methylglutaric acid database spectrum (red).

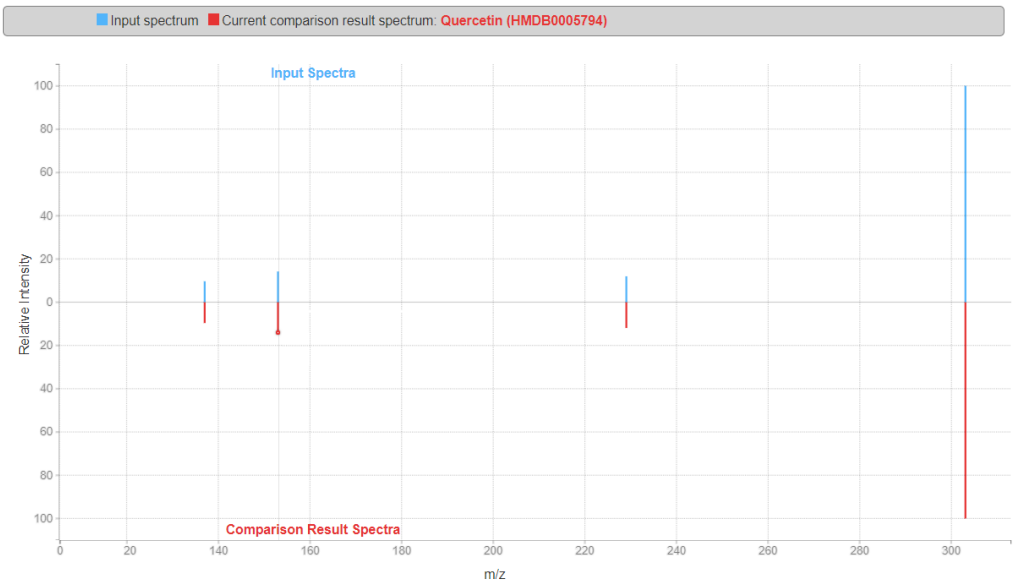
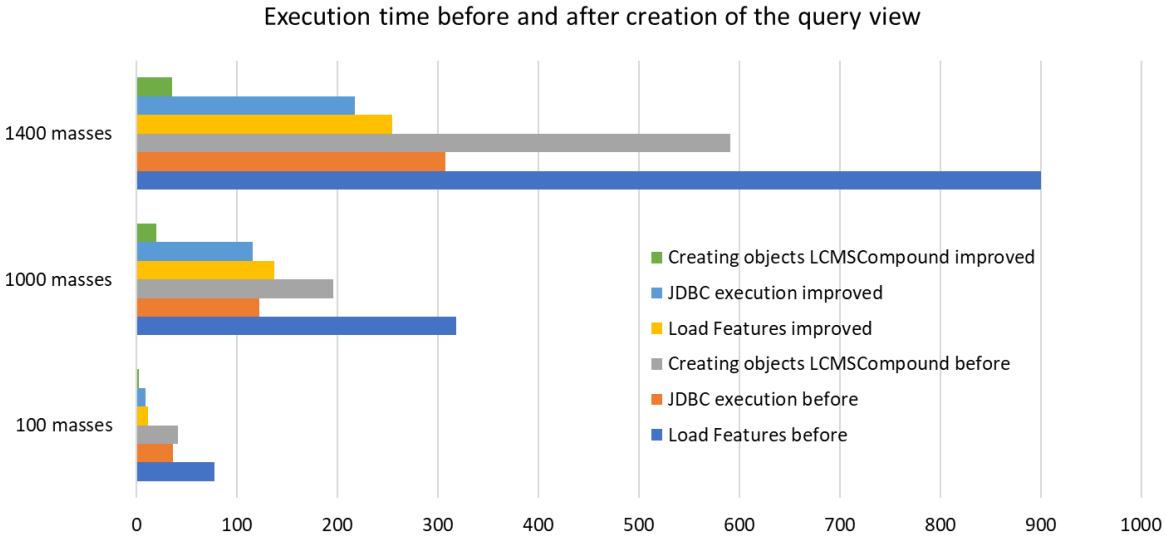


Figure 57 Input spectrum from Quercentin, extracted from MassBank database (blue) *vs* Quercentin database spectrum (red).

5.3 LC-MS execution time

The execution time for CMM LC-MS search was notably reduced by the creation of a view containing all the relevant information from the database to fill the putative annotations. The graph illustrated in Figure 58 shows the execution times in seconds, before and after the improvement, for different tasks.



	100 masses	1000 masses	1400 masses
Creating objects LCMSCompound improved	2,259	20,168	35,929
JDBC execution improved	9,411	116,154	217,614
Load Features improved	11,802	136,926	254,263
Creating objects LCMSCompound before	41,256	196,247	591,209
JDBC execution before	36,621	122,02	307,554
Load Features before	78,036	318,63	900,512

Figure 58 Execution times for loading features (seconds) before and after the creation of the query view.

It is remarkable the difference after the creation of the view. The process for **loading the features** is the summation of the queries execution (**JDBC execution**) plus the objects creation (**Creating objects LCMSCompound**). The features loading has been reduced by almost half. The objects creation became the less time-consuming step, with a significant reduction. JDBC execution time also decreased after the creation of the view.

After solving the execution critical steps, the capacity for the JDBC search was tested. Before the code refactoring, the database accesses were performed with JPA. To test the limit of masses that the batch simple and batch advanced searches could handle

using JPA queries, an excel with 1,400 real EMs was loaded gradually measuring the execution times. Figure 59 shows the results, where the units are in seconds. The advanced search failed when inserting 1,400 masses, therefore, the execution time could not be measured, and it means also that the users could not search an experiment with this number of features, which is not uncommon in metabolomic studies. The execution time for querying 1,400 features in batch simple search or 900 features in batch advanced increases until almost 10 minutes.

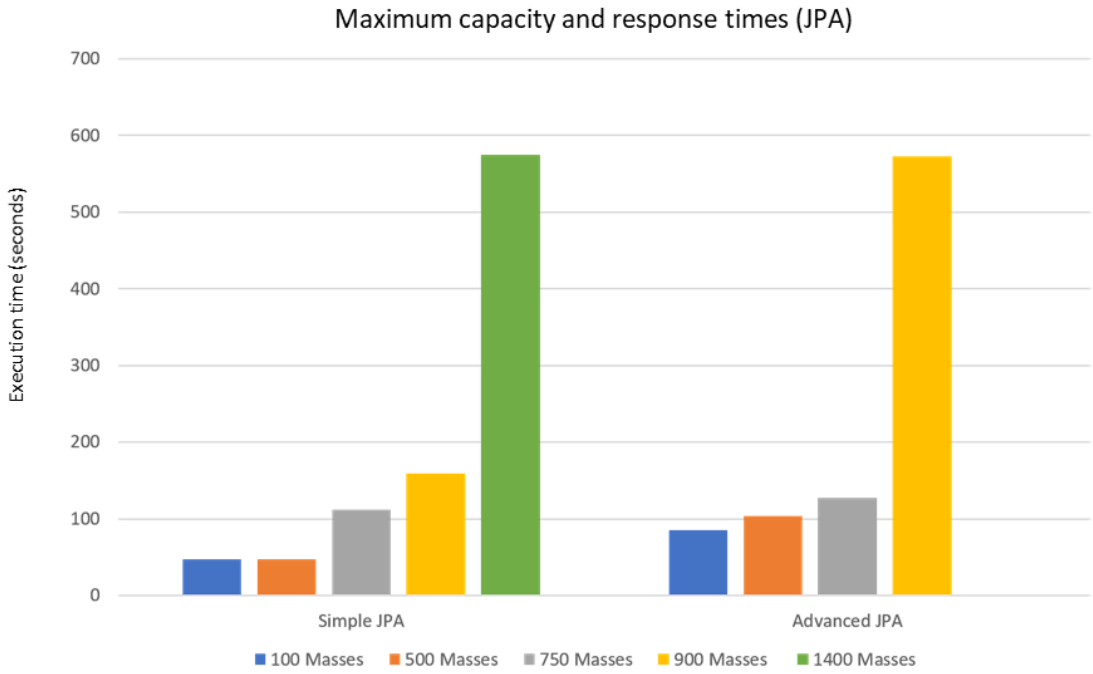


Figure 59 JPA capacity, batch simple and batch advanced searches.

The capacity of the refactored code was also measured. Figure 60 illustrates the execution times for 1,000, 1,400, 5,000 and 10,000 masses with a tolerance of 10 ppm. The 1,000 and 1,400 EMs were taken from the excel file, whilst the 5,000 and 10,000 were generated randomly, since real data from an experiment with more than 1,400 features were not available. The program failed when we tried to load 20,000 random masses. The time for creating the objects is lower in JDBC in comparison with the queries' execution. The capacity of the JDBC queries is higher than the capacity of the JPA queries due to the persistence layer implemented by JPA.

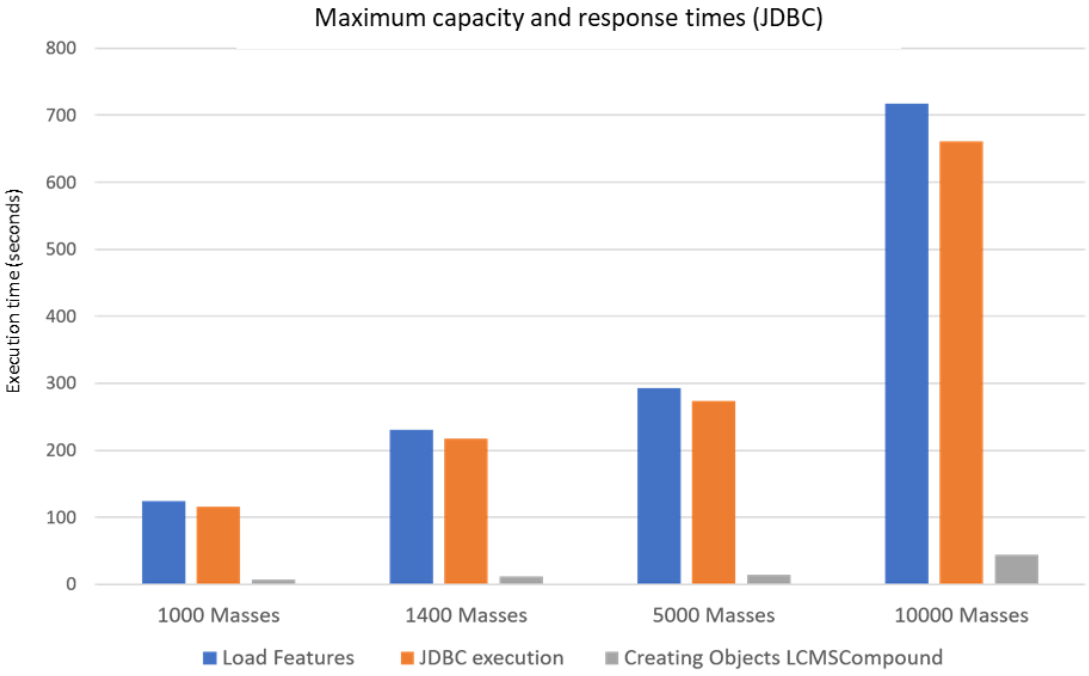


Figure 60 Maximum capacity JDBC LC-MS search.

6 CONCLUSIONS

The present project arises from the collaboration with San Pablo CEU's development team in the project of construction and development of the CMM. This project allows chemical analysts the unified access to different metabolomic databases. Attending CEMBIO's requirements, the present work seeks to solve the following objectives:

- Implement and integrate a MS/MS based search feature in CMM.
- Code refactoring for allowing the grouping of features based on their RT.
- Implement and integrate an in-source fragments detector based on MS/MS patterns.

For the MS/MS based search implementation three different approaches were created. They have a small lower precision than the available external tools tested. Therefore, these three algorithms have been implemented in CMM and they can be tested when they are more suitable. This new feature provides CMM a higher confidence level (see Figure 11, right) for metabolites identification; from level 3, where the metabolites are identified based on their mass accuracy, to level 2, where the metabolites are identified from their MS/MS spectra. Moreover, CEMBIO's members will not require to perform the spectral matching by visual inspection since this feature provides an automated match.

The code refactoring allowed the implementation of the in-source fragments detector functionality. The new UML model (Figure 25, Figure 26) gave to CMM the capability to group features by their RT, making possible a reduction of the complexity of the input data and the application of expert knowledge. Furthermore, this refactoring changed the database information access from JPA to JDBC, reducing the execution times and increasing the number of features to be searched in order to allow the chemists to search and annotate all the features from the same experiment.

Nevertheless, CMM is a project that is constantly evolving since it gives real support to a field of research that is constantly growing. The future developments are subjected to CEMBIO's needs, but some future lines can be highlighted:

- Test both MS/MS search and LC-MS search with real data. Both features were tested in this project with data obtained from different metabolomic repositories (for example, MassBank). To obtain a more accurate verification of their correct functionality, they can be tested with reference standards.

- Improve MS/MS search algorithms with new approaches and/or modify the developed ones. Many of the existing algorithms perform a dot product approach for spectral matching, but each one considers different variables and uses different weights. They can be modified to use different scoring functions to increase the precision.

- Create a MS/MS spectrum comparison viewer for MS/MS search feature. The MS/MS search output is a list of ranked putative identifications. To ensure the researcher that the given identifications are possible, a graphical comparison of the input spectrum *vs* the database spectrum can be very helpful, since the researchers can rapidly check how similar are the experimental and the database spectra.

- Implement search services for other techniques (GC-MS, CE-MS, NMR) to support the integration of them for a better coverage of the metabolomic experiment.

- Apply structural knowledge of molecules for adducts rules, since their structure can give some clues about what adducts are possible to appear and which ones are impossible.

7 REFERENCES

[1] Emily S Boja, Christopher R Kinsinger, Henry Rodriguez, Pothur Srinivas, and on behalf of Omics Integration Workshop Participants. Integration of omics sciences to advance biology and medicine. *Clinical Proteomics* 2014 11:45. (doi: 10.1186/1559-0275-11-45)

[2] Vailati-Riboni M., Palombo V., Loor J.J. What Are Omics Sciences? In: Ametaj B. (eds) Periparturient Diseases of Dairy Cows. Springer, Cham CRC Press, 2017 (ISBN: 978-3-319-43031-7).

[3] Web page of the free medical dictionary. -ome. (n.d.) *Medical Dictionary*. (2009).https://medical-dictionary.thefreedictionary.com/-ome (Accessed: May 2018)

[4] Web page International Service for the acquisition of agri-biotech applications. Pocket K No. 15: *'Omics' Sciences: Genomics, Proteomics, and Metabolomics* (2006) http://www.isaaa.org/resources/publications/pocketk/15/default.asp

[5] Alberto Gil de la Fuente. Diseño Validación e implementación de una herramienta para la identificación de metabolitos. *Trabajo Fin de Máster en Ingeniería Informática* Universidad Complutense Madrid. (2016)

[6] Danuta Dudzik, Cecilia Barbas-Bernardos, Antonia García, Coral Barbas. Quality assurance procedures for mass spectrometry untargeted metabolomics. a review. *Journal of Pharmaceutical and Biomedical Analysis 147 (2018) 149–173* (http://dx.doi.org/10.1016/j.jpba.2017.07.044)

[7] Silas G. Villas-Bôas, Ute Roessner, Michael A.E.Hansen, Jorn Smedsgaard, Jens Nielsen. Melabolome Analysis. An Introduction. 2006 (ISBN: 9780470105504)

[8] Wishart D. et al. The Human Metabolome Database – A Major Update for 2018. *Metabolomics* 2018

[9] Web page *Saccharomyces* GENOME DATABASE (SGD). http://www.yeastgenome.org (Accessed May 2018).

[10] Web page Twitter Matej Oresic https://twitter.com/matejoresic/status/1011623684174331906 (Accesed Juliy 2018)

[11] Ren, Sheng, et al. Computational and statistical analysis of metabolomics data. *Metabolomics*, 2015, vol. 11, no 6, p. 1492-1513

[12] Warwick B Dunn. Current trends and future requirements for the mass spectrometric investigation of microbial, mammalian and plant metabolomes. *IOP PUBLISHING* Phys. Biol. 5 (2008) 011001 (24pp) (doi:10.1088/1478-3975/5/1/011001)

[13] Wetmore, Diana R., et al. METABOLOMIC PROFILING REVEALED BIOCHEMICAL PAHTWAYS AND BIOMARKERS ASSOCIATED WITH PATHOGENSIS IN CYSTIC FIBROSIS CELLS. *Journal of Biological Chemistry*, 2010, p. jbc. M110. 140806. (ISSN: 1083-351X)

[14] Griffin, Julian L., and John P. Shockcor. Metabolic profiles of cancer cells. *Nature reviews cancer*, 2004, vol. 4, no 7, p. 551 (ISSN: 1474-1768)

[15] Kaddurah-Daouk, Rima, and K. Ranga Rama Krishnan. Metabolomics: a global biochemical approach to the study of central nervous system diseases. *Neuropsychopharmacology*, 2009, vol. 34, no 1, p. 173. (ISSN: 1740-634X)

[16] Wang-Sattler, Rui, et al. Novel biomarkers for pre-diabetes identified by metabolomics. *Molecular systems biology*, 2012, vol. 8, no 1, p. 615. (ISSN: 1744-4292)

[17] Griffin, Julian L., et al. Metabolomics as a tool for cardiac research. *Nature Reviews Cardiology*, 2011, vol. 8, no 11, p. 630. (ISSN: 1759-5010)

[18] W.J. Griffiths, K. Karu, M. Hornshaw, G. Woffendin and Y. Wanga. Metabolomics and metabolite profiling: past heroes and future developments. *W.J. Griffiths et al., Eur. J. Mass Spectrom*. 13, 45–50 (2007) 45 (ISSN: 1469-0667)

[19] Xiao, Jun Feng; Zhou, Bin; Ressom, Habtom W. Metabolite identification and quantitation in LC-MS/MS-based metabolomics. *TrAC Trends in Analytical Chemistry* (2012), vol. 32, p. 1-14. (https://doi.org/10.1016/j.trac.2011.08.009)

[20] Li, Yutai. Identification of metabolites in LC–MS-based metabolomics. Identification and Data Processing Methods in Metabolomics (2015) 48-65. (Book ISBN: 978-1-910420-28-7)

[21] Eric Milgram and Anders Nordstrom. Asms metabolomics workshop: Current topics in metabolomics, 2009.

[22] Hill, Dennis W., et al. Mass spectral metabonomics beyond elemental formula: chemical database querying by matching experimental with computational fragmentation spectra. *Analytical chemistry* (2008), vol. 80, no 14, p. 5574-5582. (DOI: 10.1021/ac800548g)

[23] Web page ThoughtCo. https://www.thoughtco.com/definition-of-isotopes-and-examples-604541 (Accessed June 2018)

[24] Brown, Marie, et al. Automated work-flows for accurate mass-based putative metabolite identification in LC/MS-derived metabolomic datasets. *Bioinformatics*, 2011, vol. 27, no 8, p. 1108-1112.

[25] Sobbot, Frank, et al. Subunit Exchange of Multimeric Protein Complexes Real-Time Monitoring of Subunit Exchange Between Small Heat Shock Proteins by Using Electrospray Mass Spectrometry. *Journal of Biological Chemistry*, 2002, vol. 277, no 41, p. 38921-38929.

[26] Emwas, Abdul-Hamid M. The strengths and weaknesses of NMR spectroscopy and mass spectrometry with particular focus on metabolomics research. *Metabonomics: Methods and Protocols*, 2015, p. 161-193.

[27] Dunn, Warwick B. Current trends and future requirements for the mass spectrometric investigation of microbial, mammalian and plant metabolomes. *Physical biology*, 2008, vol. 5, no 1, p. 011001.

[28] Meier, René, et al. Bioinformatics can boost metabolomics research. Journal of biotechnology, 2017, vol. 261, p. 137-141. (https://doi.org/10.1016/j.jbiotec.2017.05.018)

[29] Martinez Alcázar, Maria Paz. MS Basis: GENERAL THINGS TO TAKE IN ACCOUNT. Metabolomics Course. CEMBIO (2018)

[30] Han, Xuemei; Aslanian, Aaron; Yates III, John R. Mass spectrometry for proteomics. *Current opinion in chemical biology* (2008), vol. 12, no 5, p. 483-490. (https://doi.org/10.1016/j.cbpa.2008.07.024)

[31] Pérez García, Carmen. Fac.Farmacia.Univ. CEU-San Pablo. Lesson 18. Proteomic Techniques. *GENOMICS AND PROTEOMICS. BIOMEDICAL ENGINEERING* (2017)

[32] De Hoffmann, Edmond. Mass spectrometry. Principles and applications. *Kirk-Othmer Encyclopedia of Chemical Technology* (2000). (ISBN: 0470033118)

[33] Dunn, Warwick B., et al. Mass appeal: metabolite identification in mass spectrometry-focused untargeted metabolomics. *Metabolomics* (2013), vol. 9, no 1, p. 44-66. (ISSN: 1573-3890)

[34] Yin, Peiyuan; Xu, Guowang. Current state-of-the-art of nontargeted metabolomics based on liquid chromatography–mass spectrometry with special emphasis in clinical applications. *Journal of Chromatography A*, 2014, vol. 1374, p. 1-13.

[35] De Souza, Leonardo Perez, et al. From chromatogram to analyte to metabolite. How to pick horses for courses from the massive web-resources for mass spectral plant metabolomics. *GigaScience* (2017). (ISSN: 2047-217X)

[36] Joanna Godzien, Alberto Gil de la Fuente, Abraham Otero, Coral Barbas; Chapter 16: Metabolite identification and annotation; CAC Vol 82: *Data Analysis for Omic Sciences: Methods and Applications*. Editors: Joaquim Jaumot, Roma Tauler and Carmen Bedia- Science Direct, ISBN: 978-0-44464-044-4 – in press – release day: October 2018

[37] Everett, Jeremy R. A new paradigm for known metabolite identification in metabonomics/metabolomics: metabolite identification efficiency. *Computational and structural biotechnology journal*, 2015, vol. 13, p. 131-144.

[38] Stashenko, Elena; Martínez, Jairo René. Gas chromatography-mass spectrometry. In Advances in Gas Chromatography. InTech (2014). (http://dx.doi.org/10.5772/57492)

[39] Lei, Z., Huhman, D. V., & Sumner, L. W. Mass spectrometry strategies in metabolomics. *Journal of Biological Chemistry* (2011), vol. 286, no 29, p. 25435-25442. (ISSN: 1083-351X)

[40] Kumar, Kalaimani Jayaraj; Vijayan, Venugopal. An Overview of Liquid Chromatography-Mass Spectroscopy Instrumentation (2014). (DOI: 10.5530/phm.2014.2.2)

[41] Coulter, B. Introduction to capillary electrophoresis. *Beckman Coulter*. (1991).

[42] Johnson, Sean R., and Bernd Markus Lange.Open-access metabolomic databases for natural product research: present capabilities and future potential. *Frontiers in bioengineering and biotechnology* (2015), vol. 3, p. 22. (https://doi.org/10.3389/fbioe.2015.00022)

[43] Pöhö, Päivi, and Tuulia Hyötyläinen. Mass Spectrometric Detection for Chromatography. *Chromatographic Methods in Metabolomics; The Royal Society of Chemistry:* London, UK, 2013, p. 43-63. (ISBN: 978-1-84973-727-2)

[44] Wishart, David S., et al. HMDB 4.0: the human metabolome database for 2018. *Nucleic acids research*, 2017, vol. 46, no D1, p. D608-D617. (ISSN 1362-4962)

[45] Web page of the Human Metabolome Database (HMDB). Version 4.0. http://www.hmdb.ca/w/databases (Accessed: May 2018)

[46] Kale, Namrata S., et al. MetaboLights: An Open-Access Database Repository for Metabolomics Data. Current protocols in bioinformatics, 2016, p. 14.13. 1-14.13. 18. (https://doi.org/10.1002/0471250953.bi1413s53)

[47] Web page of Lipid Maps. LIPID Metabolites and Pathways Strategy. Lipidomics Gateway. www.lipidmaps.org/about/about_consortium.html. (Accessed: May 2015)

[48] Fahy, Eoin, et al.LIPID MAPS online tools for lipid research. *Nucleic acids research*, 2007, vol. 35, no suppl_2, p. W606-W612. (ISSN 1362-4962)

[49] Web page of KEGG: Kyoto Encyclopedia of Genes and Genomes. http://www.genome.jp/kegg/. (Accessed: May 2018)

[50] Horai, Hisayuki, et al.MassBank: a public repository for sharing mass spectral data for life sciences. Journal of mass spectrometry, 2010, vol. 45, no 7, p. 703-714. (https://doi.org/10.1002/jms.1777)

[51] Spicer, Rachel, et al. Navigating freely-available software tools for metabolomics analysis. *Metabolomics*, 2017, vol. 13, no 9, p. 106.

[52] Alonso, Arnald; Marsal, Sara; Julia, Antonio. Analytical methods in untargeted metabolomics: state of the art in 2015. *Frontiers in bioengineering and biotechnology*, 2015, vol. 3, p. 23.

[53] Kuhl, Carsten, et al. CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. Analytical chemistry, 2011, vol. 84, no 1, p. 283-289.

[54] Web page of MetFrag. MetFrag Home. http://c-ruttkies.github.io/MetFrag/ (Accessed June 2018).

[55] Web page of My Compound ID. MCID Software. http://mcid.chem.ualberta.ca/analysis (Accessed June 2018)

[56] Web page OMICTOOLS. CAMERA Metabolite identification: MS-Bases untargeted metabolomics. https://omictools.com/camera-tool (Accessed June 2018)

[57] Andrés Esteban Fernández. INTEGRACIÓN DE CONOCIMIENTO EXPERTO EN UN BUSCADOR DE METABOLITOS PARA LA CRIBA DE RESULTADOS. *Final degree project. Informatic Engineering. Universidad Politécnica de Madrid* (UPM) (2017)

[58] Web page of CEU Mass Mediator. Manual. http://ceumass.eps.uspceu.es/manuals.xhtml. (Accessed: May 2018)