



CEU MASS MEDIATOR USER'S MANUAL

Version 3.0, October 14th, 2018



1. Introduction.....	2
1.1. System Requirements.....	2
2. Peak search.....	4
2.2. Simple Search	5
2.3. Advanced Search	7
2.4. Batch Search	9
2.5. Batch Advanced Search	10
2.6. Annotations rules	13
2.7. Submit menu	14
2.8. Result List	15
3. Browse Search	18
3.1. Regular expressions.....	18
3.2. Escape characters	20
3.3. Browse search results.....	20
4. Oxidised Lipids	21
4.1. Long chain oxidised lipids.....	21
4.2. Short chain oxidised lipids.....	22
5. Spectra Quality Controller	25
5.1. Partial scores	25
5.2. Overall score.....	29
6. Pathway Displayer	32
6.1. File structure.....	32
6.2. Result list for pathways	33
7. MS/MS Search	34
7.1. Results MS/MS Search.....	36
8. LC/MS Search grouped by RT	39
8.1. Result LC-MS Search.....	45
9. RESTful API	47
9.1. Peak search.....	47
9.1.1 Batch search	47
9.1.2 Batch advanced search.....	51
10. Manual	55

1. Introduction

Ceu Mass Mediator (CMM) is an on-line tool for aiding researchers when performing metabolite annotation. CMM integrates compounds from different sources (HMDB, LipidMaps, KEGG, Metlin and compounds from in-house libraries developed at CEMBIO -Centro de Excelencia en Metabolómica y Bioanálisis, <http://www.metabolomica.uspceu.es>) based on the IUPAC International Chemical Identifier (InChI).

The versions used for its database are:

- HMDB: Files downloaded from <http://www.hmdb.ca/downloads>. Version 4.0. Updated on 10/05/2018.
- KEGG: Rest API (<http://www.kegg.jp/kegg/rest/keggapi.html>). Updated on 10/05/2018.
- LipidMaps: Files downloaded from <http://www.lipidmaps.org/resources/downloads/index.html>. Updated on 10/05/2018.
- Metlin. Links obtained from other sources (PubChem, KEGG, HMDB, LipidMaps). Updated on 10/05/2018.
- In-House Library. Created at CEMBIO. Updated on 31/03/2018.
- MINE: JavaScript API (<https://github.com/JamesJeffryes/MINE-API>). Updated on 24/01/2018.

Furthermore, CMM scores the putative annotations using three types of rules, explained in detail in section 2.6.

This manual describes the available features in CMM. These features are shown in Figure 1 and described in chapter 2 and 3.

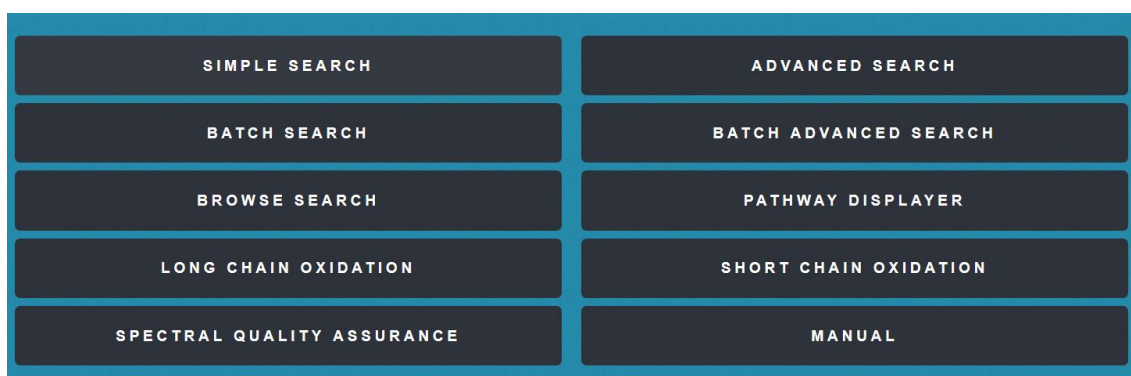


Figure 1 Main menu of Ceu Mass Mediator

1.1. System Requirements

CMM is a J2EE application, and it may be accessed through any web browser which supports JavaScript. CMM does not use Adobe Flash player neither popups. It has been tested in the next browsers:

- i. Mozilla Firefox 50
- ii. Google Chrome 45
- iii. Internet Explorer 11
- iv. Opera 42

2. Peak search

Peak Search allows the user to find metabolites based on the neutral or the m/z mass within a certain tolerance (ppm or mDa -default tolerance: 10 ppm-). CMM provides researchers with different types of search, depending on what information they want to use for performing the metabolite annotation and depending on whether they want to look for multiple compounds or just one. Some of them include a field called Composite Spectrum (CS) which refers to the set of signals which has been previously assigned to the same feature based on a previous pre-processing of the data.

Composite Spectrum

This section explains how to create the CS based on different signals arising from the same feature.

The next list, where x is m/z , y is the intensity, z is the charge, and s is the adduct and/or the isotope, shows different signals arising from the same feature corresponding to glutamic acid.

1. $x=295.1136$ $y=7002.5$ $z=1$ $s=2M+H$
2. $x=296.1166$ $y=845.9$ $z=1$ $s=2M+H+1$
3. $x=297.1184$ $y=161.8$ $z=1$ $s=2M+H+2$
4. $x=148.0610$ $y=100212.0$ $z=1$ $s=M+H$
5. $x=149.0640$ $y=6052.8$ $z=1$ $s=M+H+1$
6. $x=150.0655$ $y=972.1$ $z=1$ $s=M+H+2$
7. $x=186.0169$ $y=1822.0$ $z=1$ $s=M+K$
8. $x=170.0492$ $y=67582.0$ $z=1$ $s=M+Na$
9. $x=171.0460$ $y=4075.2$ $z=1$ $s=M+Na+1$
10. $x=172.0474$ $y=655.5$ $z=1$ $s=M+Na+2$
11. $x=74.5339$ $y=192535.0$ $z=2$ $s=M+2H$
12. $x=75.0354$ $y=11667.6$ $z=2$ $s=M+2H+1$
13. $x=75.5361$ $y=1867.6$ $z=2$ $s=M+2H+2$

This list can be represented as the following CS:

(295.1136,7002.5), (296.1166,845.9), (297.1184,161.8), (148.0610,100212.0),
(149.0640,6052.8), (150.0655,972.1), (186.0169,1822.0), (170.0492,67582.0),
(171.0460,4075.2), (172.0474,655.5), (74.5339,192535.0), (75.0354,11667.6),
(75.5361,1867.6),

where the first number corresponds to the m/z , and the second to the intensity. Each pair of values corresponds to one signal, being either a particular adduct or its isotope(s). Nevertheless, this clustering process sometimes fails, and ions are split into separate features. CMM makes a processing (see section 2.6) to detect different features arising from the same signal.

The following search modes are available:

2.2. Simple Search

Simple search enables the user to find metabolites through the m/z or the neutral mass. Query parameters are specified in the form shown in Figure 2.

Experimental Mass: [1]
Tolerance (ppm): [2]
Databases: [3]
Metabolites: [4]
Input Mass Mode: [5]
Ionization Mode: [6]
Adducts: [7]

Figure 2 Simple search interface

[1] **Experimental Mass (EM):** Mass to search in CMM (Da).

[2] **Tolerance:** Tolerance allowed for the putative annotations regarding the EM (ppm or mDa).

[3] **Databases:** The putative annotations should be present in the databases chosen by the user (Kegg, HMDB, LipidMaps, Metlin or MINE).

[4] **Metabolites:** Metabolite types to search. The user can filter the results based on the metabolite type. It may be used for excluding peptides, look only into lipids or perform a query over all type of metabolites.

[5] **Masses mode:** The user introduces the EM in neutral or m/z mode. If the user is working with neutral masses, CMM performs searches over positive or negative mode based on the hypothesis of the neutral mass calculated as $[M-H]^-$ or $[M+H]^+$. That means that the EM will correspond to the m/z obtained in the mass spectrometer with the addition or subtraction of the mass of the hydrogen (H).

[6] Ionization mode: The user wants to perform searches over a mass obtained in positive or negative mode. Depending on the ionization mode, the possible adducts formed differ.

[7] 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 Figure 3 and Figure 4. All the possible alterations of the mass of the original metabolite (M) given by the selected adducts will be searched by CMM. The complete list of adducts is:

Ionisation mode	Adducts
Positive mode	M+3H, M+2H+Na, M+H+2K, M+H+2Na, M+3Na, M+2H, M+H+NH ₄ , M+H+Na, M+H+K, M+ACN+2H, M+2Na, M+2ACN+2H, M+3ACN+2H, M+H, M+Na, M+K, M+NH ₄ , M+H-H ₂ O, 2M+H, 2M+Na, M+H+HCOONa, 2M+H-H ₂ O, M+CH ₃ OH+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+NH ₄ , 2M+K, 2M+ACN+H, 2M+ACN+Na, M+H-2H ₂ O, M+NH ₄ -H ₂ O, M+Li, 2M+2H+3H ₂ O
Negative mode	M-3H, M-2H, M-H ₂ O-H, M-H, M+Na-2H, M+Cl, M+K-2H, M+FA-H, M+Hac-H, M+Br, M+TFA-H, 2M-H, 2M+FA-H, 2M+Hac-H, 3M-H, M-H+HCOONa, M+F

The screenshot shows a software interface with two main sections: 'Ionization Mode:' and 'Adducts:'. In the 'Ionization Mode:' section, there are three radio buttons: 'neutral', 'Positive Mode', and 'Negative Mode'. The 'Negative Mode' radio button is selected. Below these radio buttons, there is a text label: 'calculation of new m/z from neutral mass based on selected adducts'. In the 'Adducts:' section, there is a list of adducts with checkboxes next to them. The 'All' checkbox is selected. The other adducts listed are M-H, M+Cl, M+HCOO, M-H-H₂O, M-H+HCOONa, and 2M-H.

Figure 3 Adducts to search in negative mode

The screenshot shows a software interface with two main sections. On the left, under 'Ionization Mode:', there is a list with 'neutral', 'Positive Mode' (highlighted), and 'Negative Mode'. Below this list, a text label reads 'calculation of new m/z from neutral mass based on selected adducts'. On the right, under 'Adducts:', there is a list of adduct types, each with a checkbox. The 'All' checkbox is checked. The other adducts listed are M+H, M+Na, M+K, M+NH4, M+H-H2O, M+H-HCOONa, M+2H, 2M+H, 2M+Na, and 2M+H-H2O.

Figure 4 Adducts to search in positive mode.

The only type of knowledge that may be applied in simple search corresponds to the ionization rules. Depending on the metabolite type, some adducts are expected to be formed, some others are possibly present, and some others are not expected to appear. For more information, look into section 2.6.

2.3. Advanced Search

Advanced search enables the user to find metabolites through the m/z or the neutral mass including some extra query parameters that are not available in the simple search. In this section all the parameters are explained (see Figure 5).

[1] Experimental Mass (EM): Mass to search in CMM (Da).

[2] Tolerance: Tolerance allowed for the putative annotations regarding the EM (ppm or mDa).

[3] Retention Time (RT): Amount of time spent by a compound on the column after it has been injected. It is an integer or a real number. The units used do not matter since it is used for checking relations between different putative annotations.

[4] Composite Spectrum (CS): Spectrum created by summation of all co-eluting m/z ions that are related, including isotopes, adducts and dimers. It is used by CMM to calculate relations between them and automatically find which adduct corresponds to the peak, when more than one adduct is present in the CS; i.e., to calculate which is the mass of the original molecule whose alterations have given rise to the observed CS.

[5] Chemical Alphabet: Possible elements of the putative annotations. CHNOPS (carbon, hydrogen, nitrogen, oxygen, phosphorus, sulphur), CHNOPS + Cl (chlorum), all elements. Compounds with deuterium can be filtered or added.

[6] Modifiers: Mobile phase modifier used. Depending on this modifier, the adduct formation may change. They are considered in the adduct formation rules (see section 2.6).

Experimental Mass (*): [1]

Tolerance -ppm- (*): [2]

Retention Time: [3]

Composite Spectrum: [4]

Chemical Alphabet (*): [5]

 CHNOPS
 CHNOPS + Cl

Modifiers (*): [6]

 NH3
 HCOO
 CH3COO
 HCOONH3
 CH3COONH3

Databases (*): [7]
☒ All except MINE
☐ All (Including In Silico Compounds)
☐ Kegg
☐ HMDB
☐ LipidMaps
☐ Metlin
☐ MINE (Only In Silico Compounds)

Metabolites (*): [8]

 Only lipids
 All including peptides

Input Mass Mode (*): [9]

 m/z Masses

Ionization Mode (*): [10]

 Positive Mode
 Negative Mode

Adducts (*): [11]
☒ All
☐ M

Figure 5 Advanced search interface

[7] Databases: The putative annotations should be present in the databases chosen by the user (Kegg, HMDB, LipidMaps, Metlin and/or MINE).

[8] Metabolites: Metabolite types to search. The user can filter the results based on the metabolite type. It may be used for excluding peptides, look only into lipids or perform a query over all type of metabolites.

[9] Masses mode: The user introduces the EM in neutral or m/z mode. If the user is working with neutral masses, CMM performs searches over positive or negative mode based on the hypothesis of the neutral mass calculated as $[M-H]^-$ or $[M+H]^+$. That means that the EM will

correspond to the m/z obtained in the mass spectrometer with the addition or subtraction of the mass of the hydrogen (H).

[10] Ionization mode: The user wants to perform searches over a mass obtained in positive or negative mode. Depending on the ionization mode, the possible adducts formed differ.

[11] 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 Figure 3 and Figure 4. All the possible alterations of the mass of the original metabolite (M) given by the selected adducts will be searched by CMM.

The knowledge that may be applied in advanced search corresponds to the ionization rules. Depending on the metabolite type, some adducts are expected to be formed, some others are possibly present, and some others are not expected to appear. For more information, look into section 2.6. However, the rules about adduct formation and lipid elution time cannot be applied since they are based in the relations between different peaks, and advanced search only accepts one peak.

2.4. Batch Search

Batch search enables the user to find metabolites through the m/z or the neutral masses. Query parameters are specified in the form shown in Figure 6. The list of EMs can be uploaded in a .txt, .csv, .xls or .xlsx file. The header of the EM should be called "masses", and all the values in this column will be handled as decimal floating point.

[1] Experimental Masses (EM): Masses to search in CMM (Da).

[2] Tolerance: Tolerance allowed for the putative annotations regarding the EM (ppm or mDa).

[3] Databases: The putative annotations should be present in the databases chosen by the user (Kegg, HMDB, LipidMaps, Metlin or MINE).

[4] Metabolites: Metabolite types to search. The user can filter the results based on the metabolite type. It may be used for excluding peptides, look only into lipids or perform a query over all type of metabolites.

[5] Masses mode: The user introduces the EM in neutral or m/z mode. If the user is working with neutral masses, CMM performs searches over positive or negative mode based on the hypothesis of the neutral mass calculated as $[M-H]^-$ or $[M+H]^+$. That means that the EM will correspond to the m/z obtained in the mass spectrometer with the addition or subtraction of the mass of the hydrogen (H).

[6] Ionization mode: The user wants to perform searches over a mass obtained in positive or negative mode. Depending on the ionization mode, the possible adducts formed differ.

The screenshot displays the 'Batch search interface' with the following components:

- Experimental Masses:** A text input field labeled 'enter input masses' with a '[1]' indicator.
- Tolerance (ppm):** A numeric input field containing '10' with a '[2]' indicator.
- Databases:** A dropdown menu with '[3]' indicator, showing options:
 - ☒ All except MINE
 - ☐ All (Including In Silico Compounds)
 - ☐ Kegg
 - ☐ HMDB
 - ☐ LipidMaps
 - ☐ Metlin
 - ☐ MINE (Only In Silico Compounds)
- Metabolites:** A dropdown menu with '[4]' indicator, showing options:
 - All except peptides
 - Only lipids
 - All including peptides
- Input Masses Mode:** Labeled '[5]', with a dropdown menu showing 'Neutral Masses' and 'm/z Masses'.
- Ionization Mode:** Labeled '[6]', with a dropdown menu showing 'neutral', 'Positive Mode', and 'Negative Mode'.
- Adducts:** Labeled '[7]', with a dropdown menu showing ☒ 'All' and ☐ 'M'.

Figure 6 Batch search interface

[7] 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 Figure 3 and Figure 4. All the possible alterations of the mass of the original metabolite (M) given by the selected adducts will be searched by CMM.

2.5. Batch Advanced Search

Batch advanced search enables the user to find metabolites through the m/z or the neutral masses query parameters explained in section 2.3. In addition, it has three input fields devoted to biomarker discovery experiments. The experimental masses corresponding to non-significant features together with its corresponding RT and CS may be introduced to provide evidences that support or refute the putative annotations. However, the putative annotations of the compounds introduced in all experimental masses field, but not included in significant experimental masses, are not returned in the result list. The list of EM, RT and CS for significant and all features can be uploaded in a .txt, .csv, .xls or .xlsx file. The header of the EM should be called "masses", the header of the RT should be called "RT" and the header of the CS should be called "CS". Values from masses and RT will be handled as decimal floating point. Values in CS should follow the format specified in the section Composite Spectrum.

Figure 7 shows the fields of the batch advanced search. The only mandatory field regarding to the features obtained in the mass spectrometer are the experimental masses of the significant compounds. RT, CS and non-significant experimental masses are optional fields that will be used by CMM for applying knowledge based on the rules explained in section 2.6. The more information the user provides in the form, the more evidence can be used for supporting or refuting the putative annotations.

[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 experimental masses introduced in field **[1]** in the same order.

Even if RTs were not used for supporting annotations, they will be automatically reported for all the annotations, which simplifies further revision since RTs do not have to be added manually.

[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.

CMM takes advantage of the grouping of signals corresponding to the same feature. It automatically detects the target experimental mass and adduct calculating differences between the m/z listed in the CS. This avoids the need of to manually calculate which adduct corresponds to each feature. The goal of this step is the identification of the true mass of the compound M that generated all the signals in the CS. If this detection is successful, only the mass of M will be searched in the database, ignoring the rest of the masses' alterations. The CSs introduced here correspond to the experimental masses 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 experimental masses introduced in field **[4]** in the same order.

[6] All Composite Spectra (CS): The CSs introduced here correspond to the experimental masses introduced in field **[4]** in the same order.

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

[8] 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.

Experimental Masses (*): [1] <input type="text" value="enter input masses"/>	Retention Times: [2] <input type="text" value="enter Retention Times"/>	Composite Spectrums: [3] <input type="text" value="enter Composite Spectrum"/>
All Experimental Masses: [4] <input type="text" value="enter all input masses"/>	All Retention Times: [5] <input type="text" value="enter Retention Times"/>	All Composite Spectrums: [6] <input type="text" value="enter Composite Spectrum"/>
Tolerance (*): <div> <input type="text" value="10"/> <div> <input checked="" type="radio"/> ppm <input type="radio"/> mDa </div> </div> [7]		
Chemical Alphabet (*): <div> <input checked="" type="text" value="All"/> <input type="text" value="CHNOPS"/> <input type="text" value="CHNOPS + Cl"/> </div> [8]		
Modifiers (*): <div> <input checked="" type="text" value="None"/> <input type="text" value="NH3"/> <input type="text" value="HCOO"/> <input type="text" value="CH3COO"/> <input type="text" value="HCOONH3"/> <input type="text" value="CH3COONH3"/> </div> [9]		
Databases (*): <div> <input checked="" type="checkbox"/> All except MINE <input type="checkbox"/> All (Including In Silico Compounds) <input type="checkbox"/> Kegg <input type="checkbox"/> HMDB <input type="checkbox"/> LipidMaps <input type="checkbox"/> Metlin <input type="checkbox"/> MINE (Only In Silico Compounds) </div> [10]		
Metabolites (*): <div> <input checked="" type="text" value="All except peptides"/> <input type="text" value="Only lipids"/> <input type="text" value="All including peptides"/> </div> [11]		
Input Masses Mode (*): [12] <div> <input checked="" type="text" value="Neutral Masses"/> <input type="text" value="m/z Masses"/> </div>	Ionization Mode (*): [13] <div> <input checked="" type="text" value="neutral"/> <input type="text" value="Positive Mode"/> <input type="text" value="Negative Mode"/> </div>	Adducts (*): [14] <div> <input checked="" type="checkbox"/> All <input type="checkbox"/> M </div>

Figure 7 Batch advanced search interface

[9] Modifiers: Mobile phase modifier used. Depending on this modifier, the adduct formation may change. They are considered in the adduct formation rules (see section 2.6). Options available are: NH_3^+ , HCOOH , CH_3COOH , HCOONH_4^+ , and $\text{CH}_3\text{COONH}_4^+$.

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

[11] 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.

[12] 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. The second and third options are available considering the fact that often the neutral mass obtained during data re-processing does not correspond to $[\text{M}+\text{H}]^+$ or $[\text{M}-\text{H}]^-$.

This is because these ions are used as default ones by many reprocessing software when only a single adduct is detected. However, some compounds, due to their chemical properties, do not form such ions. Consequently, the neutral mass assigned by the software is wrong. To overcome this, when choosing the option positive or negative for neutral mass mode, CMM turns the neutral mass to m/z and performs searches across the databases using this m/z instead of the neutral mass.

[13] Ionization mode: The user indicates whether the masses were obtained in positive or negative mode. Depending on the ionization mode, the possible adducts differ.

[14] 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 Figure 3 and Figure 4. All the possible alterations of the mass of the original metabolite (M) given by the selected adducts will be searched by CMM.

Batch advanced search process all information provided (significant EM are mandatory, RT, CS and non-significant EM are optional) for scoring the putative annotations based on the rules explained in section 2.6.

2.6. Annotations rules

Ceu Mass Mediator scores the putative annotations based on expert knowledge. This knowledge applied is especially devoted to lipids using Liquid Chromatography. It uses 143 rules divided in three main types:

1. Propensity of particular adducts formation depending on the lipid class, ionisation mode and mobile phase modifier used. Lipids belonging to particular class may always form some adducts in certain experimental conditions, whereas they may form others in different conditions. The mobile phase modifier used is indicated manually by the user. For example, phosphocholine

in negative mode primarily form $[M+HCOO]^-$ or $[M+CH_3COO]^-$ depending on the modifier used ($HCOO^-$ or CH_3COO^-); they may also form $M+Cl^-$ with lower intensity; and they never form $M-H^-$. Lipid classes used in these rules are: PC, LPC, PE, LPE, PI, PG, PS, LPS, PA, MG, DG, TG, CER, SM and CE according to the LipidMaps classification.

- Relationship between signals of different adducts from the same compound (Lynn et al., 2015). We only expect certain types of adducts when others are present. For example, glycerophosphoethanolamines (PE) may form $M+Na^+$ adduct, but only when $M+H$ adduct is also formed in higher abundance. If an experimental mass (738.5044 Da) is compatible with a $M+Na^+$ adduct of PE(34:2), but the adduct $M+H^+$ (716.5225 Da) is not present in the whole data matrix, CMM decreases the score of the annotation of PE(34:2) for experimental mass 738.5044 Da and adduct $M+Na^+$.
- Relative RT based on the lipid class and the length and number of double bounds in the lipid carbon chains (Godzien et al., 2016). For example, RT of LPG(18:0) must be greater than RT of LPG(16:0); and RT of LPG(18:0) must be greater than RT of LPG(18:2).

CMM calculates a score for each of these three rule types (χ_1, χ_2, χ_3) and then it integrates them by computing their weighted geometric mean:

$$\chi = \exp\left(\frac{\sum_{i=1}^3 \omega_i \cdot \ln \chi_i}{\sum_{i=1}^3 \omega_i}\right)$$

where ω_i is the weight of each score and χ_i is the punctuation for each score. $\omega_1 = 1$, $\omega_2 = 1$ and $\omega_3 \in [0, 2]$. ω_3 depends on the number of rules applied for lipid elution time. This is the only rule type that can be triggered a variable number of times for the same annotation, depending on how many other lipid annotations with which the retention time of the annotation to be scored can be compared with. The more rules have been triggered, the more evidence supporting or refuting the annotation would have been gathered, the more weight this evidence should have on the final score. Internally all $\chi_i \in [0, 1]$, corresponding 0 with a completely refuted annotation, 1 with an annotation for which all the possible evidence is available and it is positive, and the value of 0.5 with an annotation for which there is no evidence (neither refuting nor supporting) but the annotation's mass matches the query parameters. However, scores are multiplied by 2 in the user interface because our experience has shown us that it is more intuitive to the researchers to see a final score in the interval $[0, 2]$.

2.7. Submit menu

Once the user has performed any type of query explained in sections 2.2, 2.3, 2.4, and 2.5, the query is sent to the server when the button submit compounds (See [2] of Figure 8)

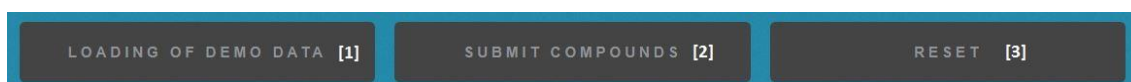


Figure 8 Submit compounds menu

- [1] **Loading demo data:** Demo data is loaded. User data is lost.
- [2] **Submit compounds:** Submit query with the filled fields by the user.
- [3] **Reset:** Clears the fields to start again filling the query parameters and input fields.

2.8. Result List

Once the user has performed any type of query explained in sections 2.2, 2.3, 2.4, and 2.5, a list of results is returned by CMM. Figure 9 shows an example of a result list.

- [1] **Compound Id:** CMM Id.
- [2] **Name:** Name of the putative annotation compound.
- [3] **Formula:** Formula of the putative annotation compound.
- [4] **Molecular weight:** Molecular weight of the putative annotation compound.
- [5] **Retention time:** Retention time introduced by the user for the experimental mass (see [18]).
- [6] **Error PPM:** Difference in parts per million (ppm or mDa) between the molecular weight and the corresponding experimental mass ([18]) and its corresponding adduct ([19]).
- [7] **Score 1:** Score for ionization rules (see item 1 of section 2.6). The code colour is structured in four ranges.
 - [0, 0.5) is red and means that this annotation is very likely wrong.
 - [0.5, 1) is orange and means that this annotation is likely wrong.
 - [1, 1.5) is yellow and means that this annotation is likely right.
 - [1.5, 2] is green and means that this annotation is very likely right.
- [8] **Score 2:** Score for adduct formation rules (see item 2 of section 2.6). The code colour is the same than for score 1 (see [7]).
- [9] **Score 3:** Score for lipid elution order (see item 3 of section 2.6). The code colour is the same than for score 1 (see [7]).
- [10] **Final score:** Integrated score (see section 2.6). The code colour is the same than for score 1 (see [7]).
- [11] **Cas:** CAS Id.
- [12] **KEGG Id:** KEGG ID and its corresponding link.
- [13] **HMDB Id:** HMDB ID and its corresponding link.

[14] LipidMaps Id: LipidMaps ID and its corresponding link.

[15] Metlin Id: Metlin ID and its corresponding link.

[16] PubChem Id: Pub Chemical Id and its corresponding link.

[17] Pathways: Pathways from KEGG where the compound is present and its corresponding link.

[18] Experimental mass: Experimental mass introduced by the user.

[19] Adduct: Corresponding adduct for this table.

[20] Number of hits: Number of hits found for the search corresponding to experimental mass (**[18]**) and its corresponding adduct (**[19]**).

[21] Generate Excel: Button which generates an Excel file with the complete result list (all experimental masses and adducts). This excel file contains the same fields that the on-line interface, the same code colour explained in **[7]**.

GENERATE EXCEL

[21]

Results

[18]

[19]

[20]

11

12

13

14

15

16

17

18

19

20

Metabolites found for mass 495.3352 and adduct M+H -> 9 metabolites found

Id	Name	Formula	Molecular Weight	Retention Time	error PPM	Score1	Score2	Score3	Final Score	Cas	KEGG	HMDB	LipidMaps	Metlin	PubChem	Pathways
[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]	[11]	[12]	[13]	[14]	[15]	[16]	[17]
32773	PC(0.0/16.0)	C24H50NO7P	495.332492	19.46886	5	2.0	2.0	2.0	2.0				LMGP01050074	49340		
32785	PC(16.0/0.0)[rac]	C24H50NO7P	495.332492	19.46886	5	2.0	2.0	2.0	2.0				LMGP01050113	102768		
34165	PE(19.0/0.0)	C24H50NO7P	495.332492	19.46886	5	2.0	2.0	2.0	2.0				LMGP02050028	77694		
32409	PC(O-14.0/2.0)	C24H50NO7P	495.332492	19.46886	5	2.0	2.0	N/A	2.0				LMGP01020019	40048		
101416	PC(O-14.0/2.0)[U]	C24H50NO7P	495.33248947	19.46886	5	N/A	2.0	N/A	2.0					40049		
101417	PC(16.0/0.0)[S]	C24H50NO7P	495.33248947	19.46886	5	N/A	2.0	N/A	2.0					40285		
101418	PC(16.0/0.0)[U]	C24H50NO7P	495.33248947	19.46886	5	N/A	2.0	N/A	2.0					40286		
101419	PC(0.0/16.0)[U]	C24H50NO7P	495.33248947	19.46886	5	N/A	2.0	N/A	2.0					40341		
32744	PC(16.0/0.0)	C24H50NO7P	495.33248947	19.46886	5	2.0	2.0	2.0	2.0			HMDB10382	LMGP01050018	40284	460602	SHOW PATHWAYS

Metabolites found for mass 495.3352 and adduct M+Na -> 1 metabolites found

Id	Name	Formula	Molecular Weight	Retention Time	error PPM	Score1	Score2	Score3	Final Score	Cas	KEGG	HMDB	LipidMaps	Metlin	PubChem	Pathways
0	No compounds found for experimental mass 495.3352 and adduct: M+Na because we detected the adduct based on the composite spectrum. Look results for adduct: M+H		0.0	19.46886	0	N/A	N/A	N/A	N/A							

Metabolites found for mass 495.3352 and adduct M+K -> 1 metabolites found

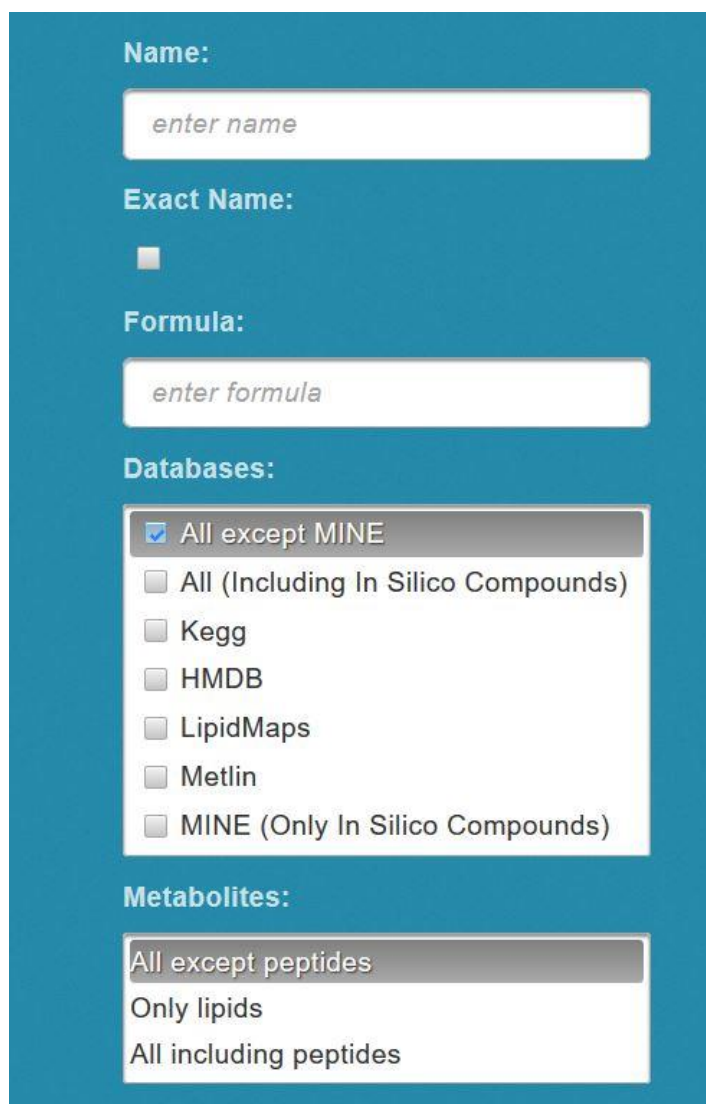
Id	Name	Formula	Molecular Weight	Retention Time	error PPM	Score1	Score2	Score3	Final Score	Cas	KEGG	HMDB	LipidMaps	Metlin	PubChem	Pathways
0	No compounds found for experimental mass 495.3352 and adduct: M+K because we detected the adduct based on the composite spectrum. Look results for adduct: M+H		0.0	19.46886	0	N/A	N/A	N/A	N/A							

Figure 9 Result list interface

3. Browse Search

Browse Search allows the user to find metabolites based on the name or the formula.

It allows to use regular expressions in both name and formula. If the user wants to retrieve only the exact name of some compound, the checkbox Exact name can be selected, and no regular expression will be applied.



The screenshot displays a search interface with a blue background. It contains several sections: 'Name:' with a text input field containing the placeholder 'enter name'; 'Exact Name:' with an unchecked checkbox; 'Formula:' with a text input field containing the placeholder 'enter formula'; 'Databases:' with a list of checkboxes where 'All except MINE' is selected, and others include 'All (Including In Silico Compounds)', 'Kegg', 'HMDB', 'LipidMaps', 'Metlin', and 'MINE (Only In Silico Compounds)'; and 'Metabolites:' with a dropdown menu showing 'All except peptides', 'Only lipids', and 'All including peptides'.

Figure 10 Browse search

3.1. Regular expressions

A regular expression defines a search pattern for strings. In CMM, the search pattern can be anything longer than three characters, including fixed strings or complex expressions, but all the searches are limited by 1,000 compounds due to performance issues.

The pattern defined by the regex may match one, several or no compounds in the databases. Following this, a brief explanation about regular expressions is given below:

Regular Expression	Description	Example
.	Matches any character	PC(...:2. Matches all the compounds containing two characters between PC(and :2.
^regex	Finds regex that must match at the beginning of the line.	^PC(...:2. Matches all the compounds starting with PC(, followed by two characters and :2.
regex\$	Finds regex that must match at the end of the line.	PC(...:2\$. Matches all the compounds finishing with PC(, followed by two characters and :2.
[abc]	Set definition, can match a or b or c.	P[CEG](...:2. Matches all the compounds containing PC(, PE(or PG(followed by two characters and :2.
[abc][de]	Set definition, can match a or b or c followed by either v or z.	P[CEG](...[2-4]. Matches all the compounds containing PC(, PE(or PG(followed by two characters, :, [a number between 2 and 4].
[^abc]	When a ^ appears at the first character inside square brackets, it negates the pattern. This matches any character except a, b or c.	P[^CEG](...:2. Matches all the compounds containing P, any other character than C, E or G followed by (, two characters and :2.
[a-z0-9]	Ranges: matches a letter between a and z and digits from 0 to 9.	P[CEG](...[2-4]. Matches all the compounds containing PC(, PE(or PG(followed by two characters, :, [a number between 2 and 4].
X Y	Finds X or Y, being X and Y regular expressions.	PC(...:2 PE(...:3. Matches all the compounds containing one of the regular expressions.

XY	Finds X followed directly by Y, being X and Y regular expressions.	PC(...:2(*. Matches all the compounds containing PC(, two characters, :2(followed by any number of characters.
----	--	---

For a full explanation of regular expressions, we recommend to read the manual of MySQL (<https://dev.mysql.com/doc/refman/8.0/en/regexp.html>).

3.2. Escape characters

The user should escape some special characters if a regular expression wants to be used. Escaping characters must be done with `\\` before the character to escape. Here is a list of special characters to escape:

- `[!#$%&()*+,\-./:;<=>?@^`{|}~]+`

3.3. Browse search results

An example of the results of the name “PC(20:2” is show in the Figure 11, returning 46 results.

However, if instead of a name we use a regular expression “PC(...:2” to search for all the compounds with any two characters (“.” means any character in regular expressions) instead of 20, the number of results is 207.

Results									
Id	Name	Formula	Molecular Weight	Cas	HMDB	Metlin	LipidMaps	KEGG	P
1602	PC(20:2(11E,14E)/20:2(11E,14E))	C48H88NO8P	837.6248			39786	LMGP01011045		
2210	PC(20:2(11Z,14Z)/12:0)	C40H76NO8P	729.5309			76115	LMGP01011835		
2211	PC(20:2(11Z,14Z)/13:0)	C41H78NO8P	743.5465			76116	LMGP01011836		
2212	PC(20:2(11Z,14Z)/14:0)	C42H80NO8P	757.5622		HMDB0008328	59774	LMGP01011837		5
2213	PC(20:2(11Z,14Z)/14:1(9Z))	C42H78NO8P	755.5465		HMDB0008329	59775	LMGP01011838		5
2214	PC(20:2(11Z,14Z)/15:0)	C43H82NO8P	771.5778		HMDB0008330	59776	LMGP01011839		5
2215	PC(20:2(11Z,14Z)/15:1(9Z))	C43H80NO8P	769.5622			76120	LMGP01011840		
2216	PC(20:2(11Z,14Z)/16:0)	C44H84NO8P	785.5935		HMDB0008331	59777	LMGP01011841		5
2217	PC(20:2(11Z,14Z)/16:1(9Z))	C44H82NO8P	783.5778		HMDB0008332	59778	LMGP01011842		5
2218	PC(20:2(11Z,14Z)/17:0)	C45H86NO8P	799.6091			76123	LMGP01011843		
2219	PC(20:2(11Z,14Z)/17:1(9Z))	C45H84NO8P	797.5935			76124	LMGP01011844		
2220	PC(20:2(11Z,14Z)/17:2(9Z,12Z))	C45H82NO8P	795.5778			76125	LMGP01011845		
2221	PC(20:2(11Z,14Z)/18:0)	C46H88NO8P	813.6248		HMDB0008333	59779	LMGP01011846		5
2222	PC(20:2(11Z,14Z)/18:1(9Z))	C46H86NO8P	811.6091		HMDB0008335	59781	LMGP01011847		5

Figure 11 Result for a regular expression in browse search.

4. Oxidised Lipids

The biological role of oxidised glycerophosphocholines (PCs) is a current topic of research, with new and very important contributions to the health and disease state still being made. Therefore, global, non-targeted metabolomics offers a very good approach to expand current knowledge and to link oxidised glycerophosphocholines (oxPCs) with new biological functions.

This functionality aims to help in the identification of the oxPCs from ESI-LC-MS/MS data. It integrates knowledge about fragmentation of oxPCs in the long and the short chain oxidised lipids. The oxidation and fragmentation process differ depending on the oxidised chain, therefore these oxidations should be handled different.

Information about the oxidation products of oxPCs was added. There are a limited number of oxPCs listed in current databases online, so some of them were added to CMM. Added data contain accurate monoisotopic masses, chemical formulae, as well as systematic and common names of the studied compounds. The database is continuously updated with new oxidation products found in biological experiments.

An indoor-built library for fatty acid chains was created: It covers fatty acids from C3:0 till C36:6 with all intermediate degrees of chain length and unsaturation. The name, the monoisotopic mass and the formula are the descriptors of the fatty acids. The name is given as CX:Y, where X indicates the number of carbons in chains and Y specifies the number of double bonds.

The recognition of oxPCs firstly hypothesised based on the fragmentation patterns observed in biological samples and then confirmed by means of authentic standards. Annotation assumes that an unidentified (for LCh-oxPC) or missing (for SCh-oxPC) fatty acid from the mid-mass region of a negative ionisation mode fragmentation spectrum is oxidised.

Figure 13 illustrates the workflow for identifying oxidised lipids in non-targeted studies. Next two section explains specifically how the recognition of long chain oxidation and short chain oxidation lipids works.

4.1. Long chain oxidised lipids

Long chain oxidised glycerophosphocholines: Input data includes both fatty acids (oxidised and non-oxidised) and precursor ion (Figure 12, panel A). Tolerance for mass matching for precursor and fatty acids can be established and oxidation type can be restricted.

The tool finds the non-oxidised fatty acid checking the matches over the indoor-built library. The oxidised fatty acid mass never matches a non-oxidised fatty acid. In that way, the tool make hypothesis over the oxidised fatty acid experimental mass over the possible oxidations. Then, the tool annotates the precursor ion.

There are two levels of annotation of precursor ion: the first refers to the oxidised form, which is searched against list of oxPCs. This search is restricted to the oxPCs containing a particular, previously annotated native fatty acid (e.g. C16:0) and oxidised one (e.g. C20:4(OH)).

If an oxPC matching the mass of precursor and containing a particular fatty acid is found it is listed in the column corresponding to the oxPC (e.g. PC(16:0/20:4(OH))). However due to the limited numbers of oxPC an alternative search can be performed. The mass of the precursor is deduced and then is searched as a non-oxidised form against the general list of PCs (e.g. PC(16:0/20:4)). In this case the search is also restricted to PCs containing particular fatty acids. As a result, a non-oxidised PC is given, therefore to get the identification of oxPC, the pointed oxidation has to be added to the name of non-oxidised PC. The result of these searches, which includes the name, molecular formula and ppm-error are displayed for each oxidation type in separate bookmarks. In the case of multiple hits, expected NLs are reported either for positive or negative ionisation mode. Although identification of oxidised fatty acid is principally based on the information obtained in negative ionisation mode, for some types of oxidation it is also necessary to use information acquired in positive ionisation mode. Furthermore, m/z of signals arising from a particular neutral loose are calculated for the confirmation of a particular oxidation. Then MS/MS spectra from positive and/or negative mode have to be inspected to confirm or reject the oxidation type proposed by the tool.

4.2. Short chain oxidised lipids

Short chain oxidised glycerophosphocholines: Input data includes non-oxidised (native) fatty acid and precursor ion (Figure 5, panel B). Tolerance for mass matching for precursor and fatty acids can be established and oxidation type can be restricted.

Mass of precursor is subtracted by mass of the adduct (either -H or -HCOO) phosphocholine head group and then by non-oxidized fatty acid. Remain part of the molecule correspond to the oxidized fatty acid. This remained mass is subsequently subtracted by the mass of possible oxidation and then non-oxidised form is searched against fatty acid library. Then results of these searches are reported, which includes the name, molecular formula and ppm-error, displayed for each oxidation type in separate bookmarks. In this case non-oxidised precursor is not searched since it is impossible to deduce the initial length of truncated chain. For this reason, the identity of the molecule is deduced based on oxidation type and annotation of both fatty acids, e.g. PC(16:0/4:0(COOH)). Similarly to LCh-oxPC functionality here also neutral losses or fragments are needed to confirm or reject particular candidate are provided.

In this case, it is only possible to search over oxidised precursor ion, since the oxidation in the short chain modifies the structure of the fatty acids and it is not possible to know from which lipid comes from. Figure 12 panel B) shows the interface for recognition of short chain oxidized fatty acids.

A) Fatty acid 1 -m/z- (*):

Fatty acid 2 -m/z-:

Tolerance for Fatty Acids(*):

☒ ppm
 ☐ mDa

Precursor -m/z for negative mode- (*):

Tolerance for precursor(*):

☒ ppm
 ☐ mDa

Possible oxidations (*):

☒ All
☐ O
☐ OH
☐ OH-OH
☐ OOH

B) All fields are required

Non oxidized fatty acid m/z:

Tolerance for Fatty Acids:

☒ ppm
 ☐ mDa

Precursor -m/z for negative mode-:

Tolerance for precursor(*):

☒ ppm
 ☐ mDa

Possible oxidations (*):

☒ All
☐ COH
☐ COOH

Figure 12 Interface for recognition of long chain and short chain oxidised lipids

Long Chain oxidation

Step 1

$$m/z \text{ of FA1} - \Delta \text{ mass of oxidation} = m/z \text{ of non-oxidized FA}$$

search against FA library → no hits
conclusion: native FA

Step 2

$$m/z \text{ of FA2} - \Delta \text{ mass of oxidation} = m/z \text{ of non-oxidized FA}$$

search against FA library → hits
conclusion: oxidised FA

Step 3

search of m/z of precursor against databases

Step 4

tentative annotation

PC(16:0 / 20:4 (OH))

Short Chain oxidation

Step 1

$$m/z \text{ of precursor} - m/z \text{ of native FA} - m/z \text{ of head group} = m/z \text{ of oxidised FA}$$

search against FA library

Step 2

$$m/z \text{ of oxidised FA} - \Delta \text{ mass of oxidation} = m/z \text{ of non-oxidized FA}$$

search against FA library

Step 3

search of m/z of precursor against databases

Step 4

tentative annotation

PC(16:0 / 5:0 (CHO))

Figure 13 Identification of oxidised lipids

5. Spectra Quality Controller

The spectra quality controller is a tool for analyzing how reliable is the spectrum obtained by the user with the purpose of identifying the primal metabolite in untargeted Metabolomics. Due to the limited number of available authentic standards, tandem mass spectrometry (MS/MS) analysis is a crucial method in metabolite annotation. However, it is very easy to overlook low quality spectra which carry a higher risk of miss-annotation. For this reason, a profound study of hundreds of MS/MS spectra from several biological studies have been performed to establish criteria for evaluation of spectral quality. As a result, a pentagonal-point evaluation system including aspects such as overall intensity of spectra, impact of noise, number of MS/MS scans obtained, co-eluting ions and cross-talk phenomena has been established. These aspects are evaluated giving a partial score between 0 and 1. Then, an overall score is calculated by summation of all previously obtained scores and used to classify spectra as excellent, acceptable or inadequate in a three-tier scoring system.

5.1. Partial scores

The intensity and the noise parameters are closely related: if a spectrum with low intensity has a noise very low, the peaks can be clearly identified. For this reason, the score of the intensity has a relation with the noise score. The intensity of the MS/MS spectra score depends on the average signal of the MS mode. If the signal in MS mode is very high, the intensity of the spectrum should be also high.

The number of scans and the number of samples are also related. When more than one sample has been analysed under MS/MS, the score is the highest.

All the scores are normalised in the range [0,1].

A full description of the partial scores can be read above:

1. Intensity score:

It follows a linear regression between the values specified in the MS/MS intensity lowest score and the MS/MS intensity highest score depending on the average signal intensity in MS analysis:

average intensity in MS analysis	MS/MS intensity lowest score	MS/MS intensity linear regression score	MS/MS intensity highest score
$\leq 10^5$	$\leq 10^2$	$10^2 - 10^3$	$\geq 10^3$
$10^5 - 10^7$	$\leq 10^3$	$10^3 - 10^4$	$\geq 10^4$
$10^7 - 10^8$	$\leq 10^4$	$10^4 - 10^5$	$\geq 10^5$
$\geq 10^8$	$\leq 10^5$	$10^5 - 10^6$	$\geq 10^6$

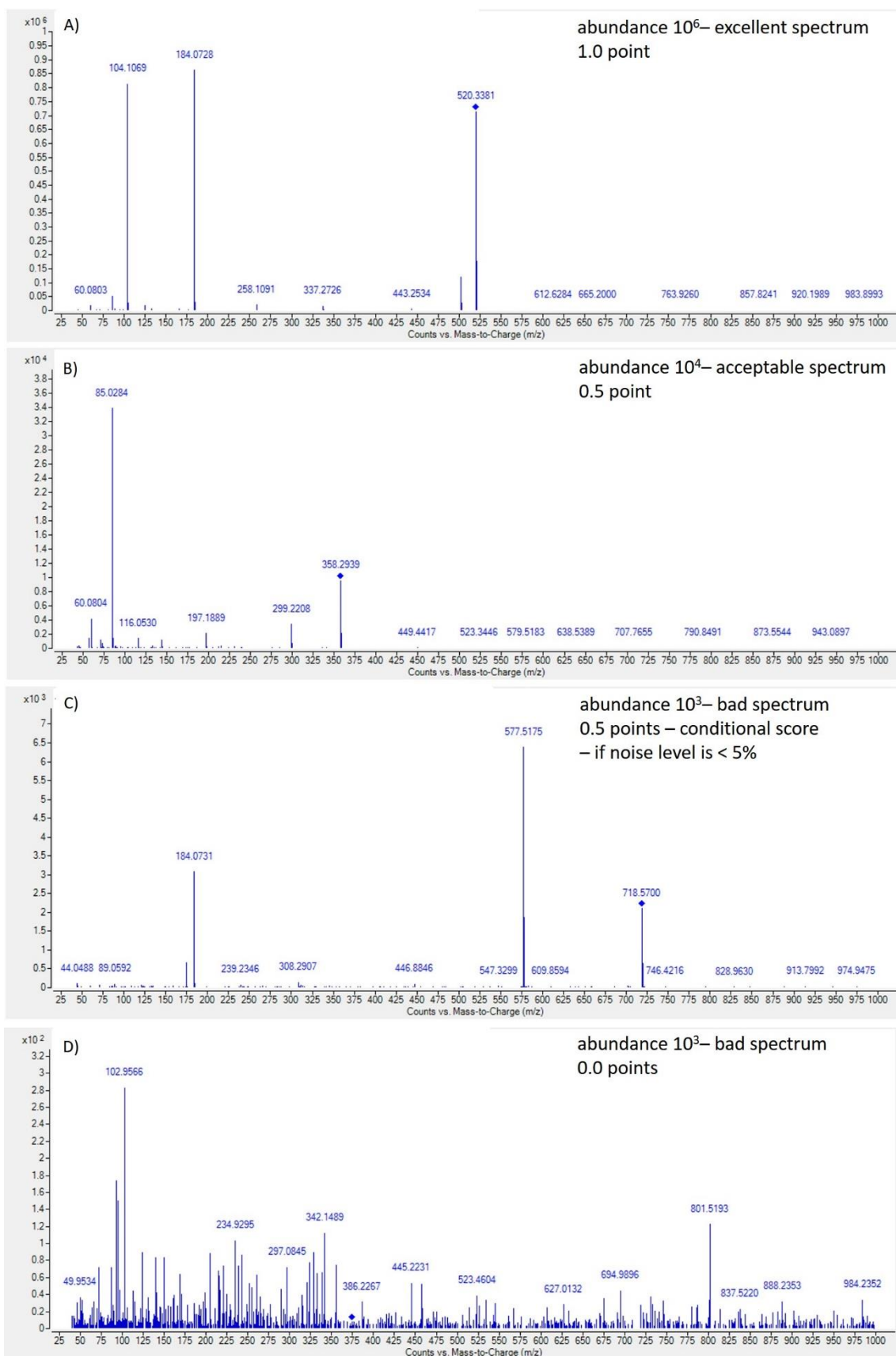


Figure 14 Examples of intensity score for Spectral Quality Controller.

If the average intensity in MS analysis is low, (e.g. $\leq 10^5$ - panel A of Figure 14), then the MS/MS intensity will be good enough for identification if it has an intensity $\geq 10^3$, it will be acceptable if it is between 10^2 - 10^3 (see panel B of Figure 14) and it will be very difficult to identify the peaks formed from the parent ion if it is $\leq 10^2$ (see panel D of Figure 14). However, if the noise is below 5%, the peaks can be distinguishable (see panel C of Figure 14), therefore if the MS/MS intensity score is below 0.5 and above 0.3 and the noise is very low ($\leq 5\%$), the intensity score is set to 0.5.

2. Noise score: a noise percentage below 5 % is considered as excellent for the identification of a compound based on its MS/MS spectrum. In the other hand, a noise percentage above 20 % is high enough to lose important information above the fragments formed by the parent ion and the noise itself, so it is considered as inadequate. A linear regression is calculated if the noise is between 5 and 20 % to set the score.
3. Number of scans score: short methods cause that the width of peaks is reduced, resulting in a lower number of acquisition points across a peak. This characteristic is problematic in MS and it has a higher impact in MS/MS, especially in the cases when MS/MS scans are acquired within intervals of MS scans. To have a reliable MS/MS spectrum at least three MS/MS scans for a single peak are needed. The score is *i*) excellent spectra with more than 5 scans – 1 point; *ii*) acceptable spectra with 5 scans – 0.75 points, 4 scans – 0.5 points, 3 scans – 0.25 points *iii*) inadequate spectra with less than 3 scans – 0 points, *iv*) MS/MS spectra acquired across different analyses (multiple analysis) - 1 point. The last option is particularly devoted to the spectra acquired in data dependent mode, where number of scans is often limited to 1 or 2 and then a particular ion is excluded for certain time to allow other ions to be targeted. Following this stratification such spectra would be classified as an inadequate, therefore to overcome this and, at the same time, to ensure quality, MS/MS acquisition should be repeated. Concordance between different analyses, even for single scan, confirms reliability and consequently the quality of obtained information.
4. Co-elution score: *i*) excellent spectra with no co-elution – 1 point; *ii*) acceptable spectra with co-elution with known molecule and known fragmentation pattern – 0.5 point and *iii*) inadequate spectra with co-elution with an unknown molecule – 0 points. It is unthinkable to distinguish between the fragments of two different parent ions in the same MS/MS analysis if a fragmentation pattern of one of them is not available, making the identification of them almost impossible. If there is co-elution with an unknown compound, then the overall score is automatically 0, since it is not possible to know which peaks observed in MS/MS correspond to each compound.
5. Cross-talk score: The principle of MS/MS analysis is to isolate and target molecule of interest, with the intention to obtain its fragmentation spectrum. Following this, it can be assumed that the collision cell, between different scans, is completely free of any ions preventing overlap of signals originating from

different molecules. However, often the ions from one scan are delayed by fragmentation and are still ongoing when the next scan starts, especially in the case of very short inter-scan delay. This results in the mixture of ions reaching the detector, a phenomenon called cross-talk. It influences both qualitative and quantitative analysis and becomes more problematic when the acquisition rate is increased.

If there is no cross-talk in the MS/MS spectrum, the identification is feasible for this parameter. If the cross-talk has a low intensity, the identification is still possible since the main fragments formed are distinguishable based on their intensity, and a hard cross-talk hinders the identification of the parent ion.

Fragmentation of a particular ion produces a series of its product ions which have smaller m/z than the molecule from which they originate. Consequently, a typical MS/MS spectrum does not have ions with higher mass than the precursor, unless some dimers or clusters were formed. Therefore, one of the easiest ways to detect a cross-talk is to observe the mass region above the targeted mass. The score for the cross-talk parameter is *i*) excellent spectra with no cross-talk – 1 point; *ii*) acceptable spectra with cross-talk where intensities of cross-talk signals are lower than the intensity of product ions – 0.5 point and *iii*) inadequate spectra with cross-talk of intensity comparable or more abundant than intensity of product ions – 0 points.

This approach has been automated in CMM for users who want to check how much they can rely on the MS/MS spectrum from their samples. There are different aspects which join in this analysis (see Figure 15):

[1] Average signal in MS mode: It measures the average intensity of the signals in the MS analysis of the sample. It is a floating-point number between 0 and 1,000,000,000.

[2] Intensity of the MS/MS spectra: It measures the intensity of the MS/MS spectrum according to the average signal of the metabolite analysed in MS. It is necessary to indicate the average signal in MS mode as well as the overall intensity of the MS/MS spectrum. It is a floating-point number between 0 and 1,000,000,000.

[3] Noise: Level of noise detected in the MS/MS spectrum under analysis. It is a floating-point number between 0 and 100, indicated as a percentage. 100 would be a spectrum where no signal is distinguishable due to the noise and 0 would be a clear spectrum where the base signal is low and all the peaks are clearly recognised.

[4] Number of scans: Number of scans acquired during the MS/MS analysis. It is an integer between 0 and 100.

[5] Number of samples: Number of samples used for acquiring the MS/MS spectrum. It is an integer between 0 and 100.

[6] Co-elution: Has the signal of the MS/MS spectra more than one compound? In case yes, do you know what is the other compound? If you know what the other compound is, it is

possible to buy standards and pay attention to the peaks corresponding to the compound under identification study.

[7] Cross-talk: Is there cross-talk from the previous scan? The user can check it if there are signals with an m/z higher than the parent ion. The intensity of the cross-talk is important since a low signal allows the user to identify the peaks corresponding to the parent ion under analysis based on the higher intensity.

Average signal in MS mode:
[1] enter the average signal in MS level

Overall intensity of MS/MS spectra:
[2] enter the overall intensity of MS/MS

Noise (%):
[3] enter the noise level percentage

Number of scans:
[4] enter the number of scans of MS/MS

Number of samples:
[5] enter the number of scans of MS/MS

Co-elution
[6] no co-elution with known compound with unknown compound

Cross-talk
[7] no cross-talk soft cross-talk hard cross-talk

LOAD DATA EXAMPLE RESET

PROCESS

Figure 15 Interface for quality spectra


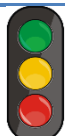
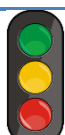
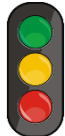
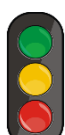
The spectra quality assurance gives information about the partial scores obtained for the parameters explained before, but it also computes an overall score.

5.2. Overall score

Each one of the aspects described above has an important impact on the final quality of the spectrum. Therefore, an overall score is proposed to reflect all characteristics within a single value. This score is obtained by summing all partial scores and consequently ranges from 0 to 5 points. Within this range three levels have been established: *i*) excellent spectra with cumulative score between 5.0 and 3.5 points; *ii*) acceptable spectra with overall score between 3.5 and 2.0 points and *iii*) inadequate spectra with score less than or equal to 2.0 points.

This overall score can be used in two ways: either to discard low quality spectra from the annotation process or to rank identified metabolites according to the confidence of

annotation. A graphical illustration of the scoring system for a sample with an average MS signal intensity between 1,000,000 and 10,000,000 is shown in the next table:

overall intensity	$\geq 10^5$	1.0	Excellent	
	$10^5 - 10^4$	(0-1)	acceptable	
	$\leq 10^4$	0.0	inadequate	
	Noise < 5% and intensity $\leq 5 \cdot 10^4$	(0.5-0.7)	acceptable	
noise	not noticeable < 5%	1	Excellent	
	noticeable 5-20%	(0-1)	acceptable	
	dominating > 20%	0.0	inadequate	
number of scans	> 5	1.0	Excellent	
	5	0.75	acceptable	
	4	0.5		
	3	0.25		
	< 3	0.0	Inadequate	
	multiple analysis (number of samples > 1)	1.0	Excellent	
co-elution	no co-elution	1.0	Excellent	
	co-elution with known metabolite	0.5	acceptable	
	co-elution with an unknown metabolite	0.0	inadequate	
cross-talk	no cross-talk	1.0	Excellent	
	intensity of cross-talk signals lower that product ions	0.5	acceptable	
	intensity of cross-talk signals comparable or more abundant than product ions	0.0	inadequate	
overall score	excellent spectrum	[5.0-3.5]	Excellent	
	high level of confidence of annotation			
	acceptable spectrum	[3.5-2.0]	acceptable	
	medium level of confidence of annotation			
	inadequate spectrum	[2.0-0.0]	inadequate	
low level of confidence of annotation				
	co-elution with an unknown metabolite	0.0		

The user can load some data for example and can reset all the fields. When the user has introduced all the input data, it is necessary to click process. The results obtained for different input parameters with a result for excellent, acceptable and inadequate spectra are shown in Figure 16 Output of the Quality Spectra ControllerFigure 16.

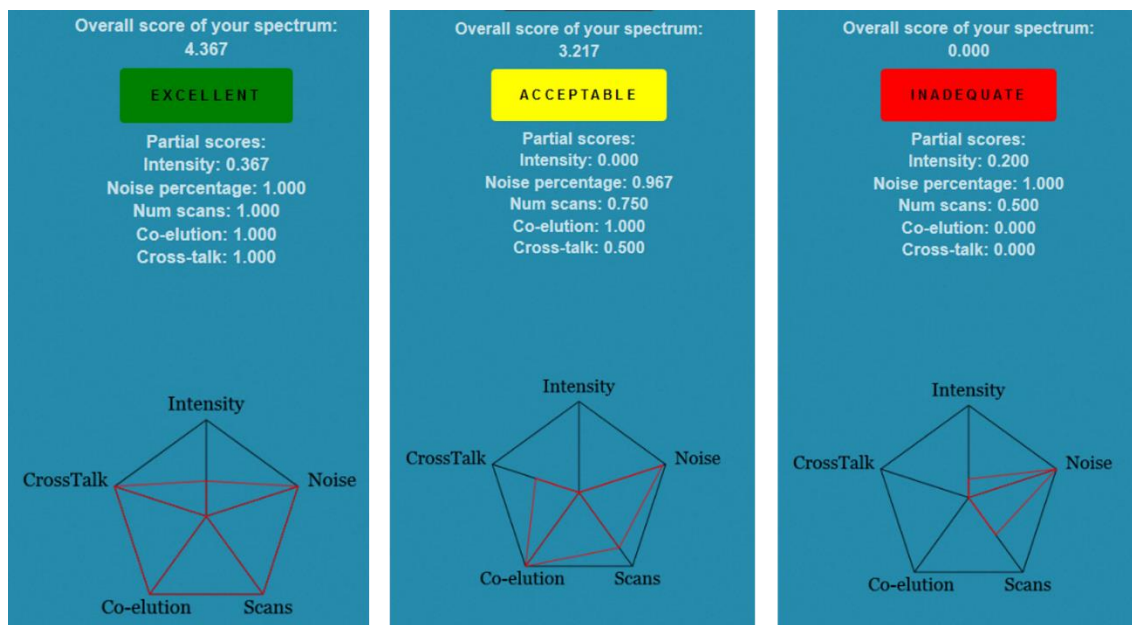


Figure 16 Output of the Quality Spectra Controller

This feature extracts information from a list of already identified compounds in order to perform a rank about the pathways that are more probably affected based on two different parameters: specificity of the compounds and percentage of compounds of the complete pathway from KEGG present in the file.

6.1. File structure

To upload an excel file to be analysed by pathway displayer of CMM, you need to press the button Choose file and, once the file was selected, submit it (see Figure 17). The structure of the file should follow the structure of the downloaded files from the result list (see Figure 18). The header names of lines 1 and 2 should be present in the file, and pathways are listed in subsequent columns after the column 2.

The user should filter the result list until it only contains the annotations corresponding to the identified compounds. If the user has worked with CMM, these annotations have a list of pathways where the compound is present according to KEGG database.

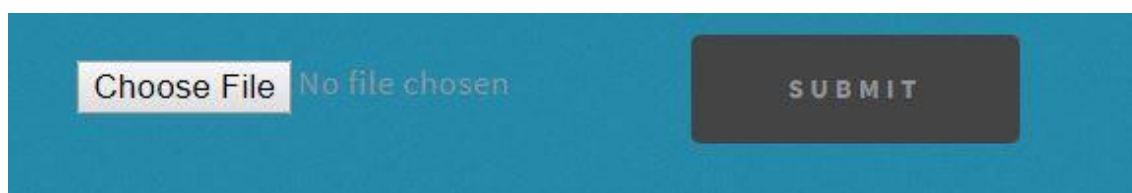


Figure 17 Pathway displayer menu

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
	LIST OF COMPOUNDS	Retention Time	Identifier	Adduct	PPM Error	Molecular Weight	Name	Formula	Score 1	Score 2	Score 3	Final Score	CAS	Keio	HMDB	LipidAtlas	Metlin	PubChem	INChIKey	Pathways	
3	838.5571	27.7552	81616 M-H	M+	216	838.3577	Heme O	C44H58FeN	N/A	N/A	N/A	13739			C1567	HMDB	5044		FFHSPAS	Porphyrin Metabolism	
4	838.5571	27.7552	82502 M-H	M+	647	839.1	1-o-methyl-2-o	C24H30O2	N/A	N/A	N/A			C00311		LMFA6705	439206		INChIKey	Alkaloids metabolism	
5	838.5571	27.7552	17504 M-H	M+	690	839.1363	Lactyl-CoA	C24H40N7O	N/A	N/A	N/A			C00562		LMFA6705	439320		VWMBE0	Propanoate Metabolism	
6	838.5571	27.7552	17471 M-H	M+	690	839.1363	3-hydroxypropionyl-CoA	C24H40N7O	N/A	N/A	N/A			C00566		LMFA6705	440753		BERBFZC	Propanoate Metabolism	
7	838.5571	27.7552	17527 M-H	NH4	493	821.1259	acrylyl-CoA	C24H38N7O	N/A	N/A	N/A			C00689		LMFA6705	439349		POODSOQ	Propanoate Metabolism	
8	838.5571	27.7552	96794 2M-H	H	76	419.3036	Myxalamid S	C25H41N104	N/A	N/A	N/A			C1215			69328	1195399	QADG5H5	Type I polyketide structures	
9	838.5571	27.7552	92133 2M-H	H	78	419.246	13-Deoxyoxalonic	C25H41N104	N/A	N/A	N/A			C2053					SJZCKB	Biosynthesis of L-lysine	
10	838.5571	27.7552	91096 2M-H	H	338	419.1369	Jasmonic acid	C24H21N06	N/A	N/A	N/A			C1968			72495	1216013	AMVSKR	Biosynthesis of L-lysine	
11	838.5571	27.7552	74977 2M-H	H	586	419.0321	5,8-Dichloro-1,4-dipropionylfuran	C21H19C2N2	N/A	N/A	N/A			C1486	HMDB		70366	11954470	BURDIRM	Metabolism of xenobiotics by cytochrome	
12	838.5571	27.7552	88618 2M-Na		215	408.3756	15-cis-4'-c	C30H48	N/A	N/A	N/A			C1614			64111	14019219	UKCFUF8	Carotenoid biosynthesis	
13	838.5571	27.7552	93969 2M-Na		0	408.2876	Bile acid	C24H40O5	N/A	N/A	N/A			C0155				439529	BHOCOFF	Vitamin D metabolism	
14	838.5571	27.7552	2050 2M-Na		0	408.2876	Alcoholic acid	C24H40O5	N/A	N/A	N/A			C1773		LMST0401	160636		YRZLCOO	Secondary bile acid biosynthesis	
15	838.5571	27.7552	2051 2M-Na		0	408.2876	Thiophene-3-ol	C24H40O5	N/A	N/A	N/A			C1766		LMST0401	160636		BHOCOFF	Secondary bile acid biosynthesis	
16	838.5571	27.7552	2059 2M-Na		0	408.2876	Haemocholic acid	C24H40O5	N/A	N/A	N/A			C1766		LMST0401	160636		PSHXEQE	Secondary bile acid biosynthesis	
17	838.5571	27.7552	2060 2M-Na		0	408.2876	Haemocholic acid	C24H40O5	N/A	N/A	N/A			C1765		LMST0401	160636		SLDVIWY	Secondary bile acid biosynthesis	
18	838.5571	27.7552	2065 2M-Na		0	408.2876	Avicholic acid	C24H40O5	N/A	N/A	N/A			C1766		LMST0401	160636		PMFAXGQ	Secondary bile acid biosynthesis	
19	719.5465	27.7563	95253 M-H		140	719.4456	Ethylpropionyl C	C36H65N01	N/A	N/A	N/A			C0661	3			83933	RXPVWNT	Biosynthesis of L-lysine	
20	719.5465	27.7563	28502 M-H		140	719.4456	Ethylpropionyl C	C36H65N01	N/A	N/A	N/A					LMF0400			RXPVWNT	Biosynthesis of L-lysine	

Figure 18 Structure of the Excel file for pathway displayer

Once the excel file is loaded, CMM processes it taking into account two different parameters for ordering the pathways present in the excel file. This order may guide the researcher to focus his hypothesis in these pathways that have compounds more specific (For example, Chlorophyll is only present in pathways related to plants):

1. Specificity: In how many pathways is present the compound? It uses the formula:

$$\text{Specificity} = \text{Min} \left(\frac{1}{\text{number of pathways where the compound has been detected}} \right)$$

Specificity $\in (0,1]$.

- Percentage of the compounds: How many compounds of the pathway are present? It uses the formula:

$$\text{Percentage} = \frac{\text{Number of compounds present in the file in the pathway}}{\text{Total number of compounds present in the pathway}}$$

Percentage $\in (0,1]$.

The final order is determined by specificity and percentage. Specificity is the first parameter and, if the specificity is the same, then the percentage would be taken into account.

6.2. Result list for pathways

When the excel file is processed, CMM returns to the user a list of results with the pathways ordered (see section 6.1). Figure 19 shows an example of a list of pathways present in an excel file ordered using this approach. The results are also available in excel format if the user wants to work with it.

Compounds present in Quinolones											
Experimental mass	Retention Time	Id	Adduct	error PPM	Molecular Weight	Name	Formula	Cas	KEGG	HMDB	Lipid
719.5465	27.7563	94019	2M+Na	880	349.0896	Ulfloxacin, NM 394	C16H16FN3O3S	112984-60-8	C14492		
< >											
Compounds present in Opioid receptor agonists/antagonists											
Experimental mass	Retention Time	Id	Adduct	error PPM	Molecular Weight	Name	Formula	Cas	KEGG	HMDB	Lipid
750.5411	27.7562	94531	M+H-H2O	222	768.3806	Deltorphan C, Deltorphan I	C37H52N8O10	122752-15-2	C18097		
< >											
Compounds present in Aflatoxin biosynthesis											
Experimental mass	Retention Time	Id	Adduct	error PPM	Molecular Weight	Name	Formula	Cas	KEGG	HMDB	Lipid
676.5043	5.7029	92777	2M+H	620	338.0427	Versicolorin A	C18H10O7	6807-96-1	C20583		
719.5465	27.7563	88755	2M+H	864	360.0845	Versiconol	C18H16O8	22268-13-9	C20508		
750.5411	27.7562	92616	2M+H-H2O	498	384.0845	1'-Hydroxyversicolorone; Hydroxyversicolorone	C20H16O8	111975-78-1	C20503		
761.5935	27.7564	92948	2M+Na	809	370.1053	Norsolorinic acid; Norsolorinate; 2-Hexanoyl-1,3,6,8-tetrahydroxy-9,10-anthraquinone	C20H18O7	10254-99-6	C20452		

Figure 19 Results list of the pathway displayer

7. MS/MS Search

This feature allows metabolites' identification through tandem mass spectrometry data (MS/MS). This identification is reached based on the similarity of an input parent ion mass and the parent ion masses of the putative annotations within a tolerance. The MS/MS spectrum patterns from these putative annotations are scored against the input spectrum, which is the set of specified peaks (m/z and intensity couples). The MS/MS spectrum from the putative annotations are extracted from the CMM database. The MS/MS data was implemented from Human Metabolome Database, which includes experimental and *in silico* fragmentation patterns. The parameters to perform the search are illustrated on Figure 20.

The score equation applied over the tandem mass spectra peaks was developed by [MetFrag web tool](#). After testing different parameters for the weights, the best results were achieved with weight 3 for the m/z and 0.6 for the intensities.

$$score = \sum (mz_{library} * mz_{input})^3 + (int_{library} * int_{input})^{0.6}$$

All fields are required

Parent Ion Mass (m/z):

MS/MS Peak List:

Parent Ion Tolerance:

☐ Da
 ☐ ppm

M/Z Tolerance:

☐ Da
 ☐ ppm

Ionization Mode:

Ionization Voltage:

Type of spectra:

☒ Experimental
 ☐ Predicted

Figure 20 MS/MS search interface

[1] **Parent 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 or relative. It is necessary to introduce just one m/z and its correspondent intensity per line, in that order and separated by a blank space. Figure 21 illustrates how to insert the peaks' input with absolute or relative intensities.

MS/MS Peak List:	MS/MS Peak List:
520.34 12.5	520.34 32.05
523.54 39	523.54 100
554.22 2.23	554.22 5.718
567.78 8	567.78 20.513

Figure 21 Peak input. Absolute intensities (left) and relative intensities (right)

[3] Parent ion Tolerance: The mass difference allowed between the experimental mass and the parent ion mass. It can be specified in ppm or Da.

[4] m/z Tolerance: The tolerance for peaks' m/z matching (spectral matching). It can be specified in ppm or Da.

[5] Ionization Mode: The ionization mode applied when performing MS/MS.

[6] Ionization Voltage: The ionization voltage applied when performing MS/MS.

[7] Type of spectra: The type of spectra over 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).

7.1. Results MS/MS Search

The output from this feature when submitting the data provides a list of ranked metabolites' identifications (see Figure 22).

Results						
Spectral Display Tools	Id	HMDB	Name	Formula	Mass	Score
Experimental	147619	HMDB0000841	L-Glutamine	C ₅ H ₁₀ N ₂ O ₃	146,0691	0,8207
Experimental	56002	HMDB0003423	D-Glutamine	C ₅ H ₁₀ N ₂ O ₃	146,0691	0,6923
Experimental	27610	HMDB0000422	2-methyl-glutaric acid	C ₈ H ₁₀ O ₄	146,0579	0,6225
Experimental	27634	HMDB0000752	Methylglutaric acid	C ₈ H ₁₀ O ₄	146,0579	0,3555
Experimental	66673	HMDB0001218	Coumarin	C ₉ H ₆ O ₂	146,0368	0,2795
Experimental	107266	HMDB0002359	Phenylpropionic acid	C ₉ H ₈ O ₂	146,0368	0,2795

Figure 22 MS/MS Search output

The figures hereunder show a comparison among the input spectra (blue) and the spectra from the database (red). The Figure 23 compares the input against L-Glutamine, Figure 24 against D-Glutamine and Figure 25 versus 2-methyl-glutaric acid. The plots were taken from HMDB web tool.

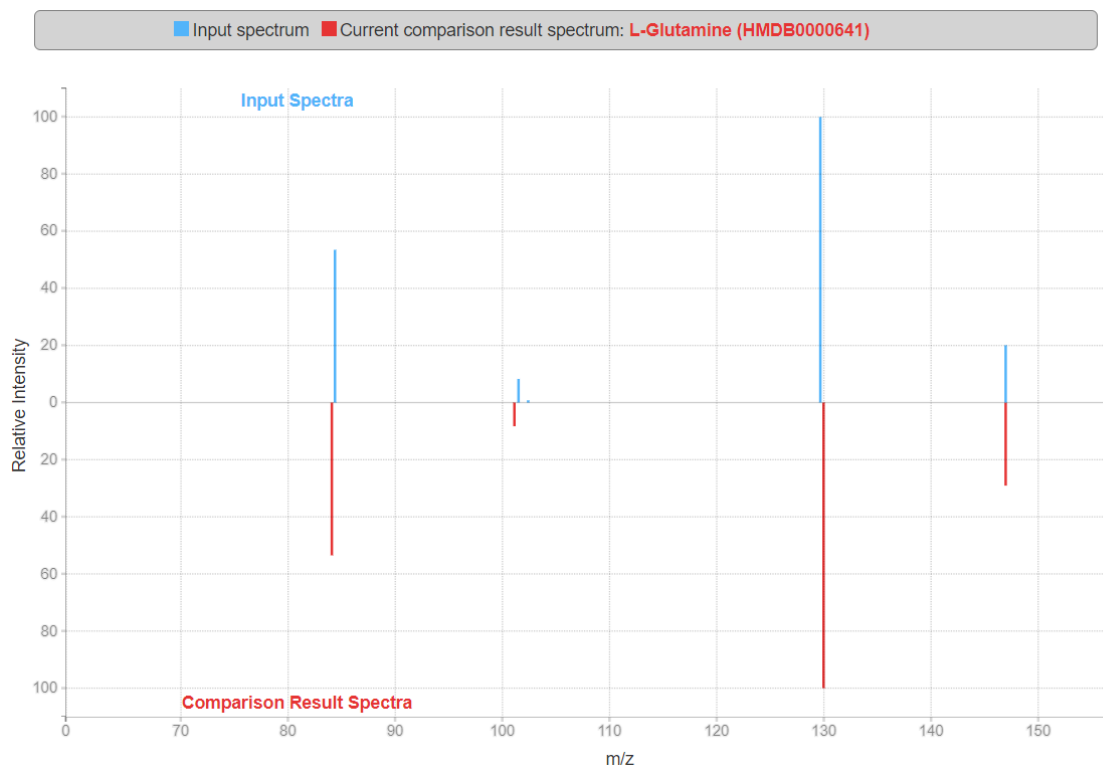


Figure 23 Comparison between the input spectra (blue) and L-Glutamine (red)

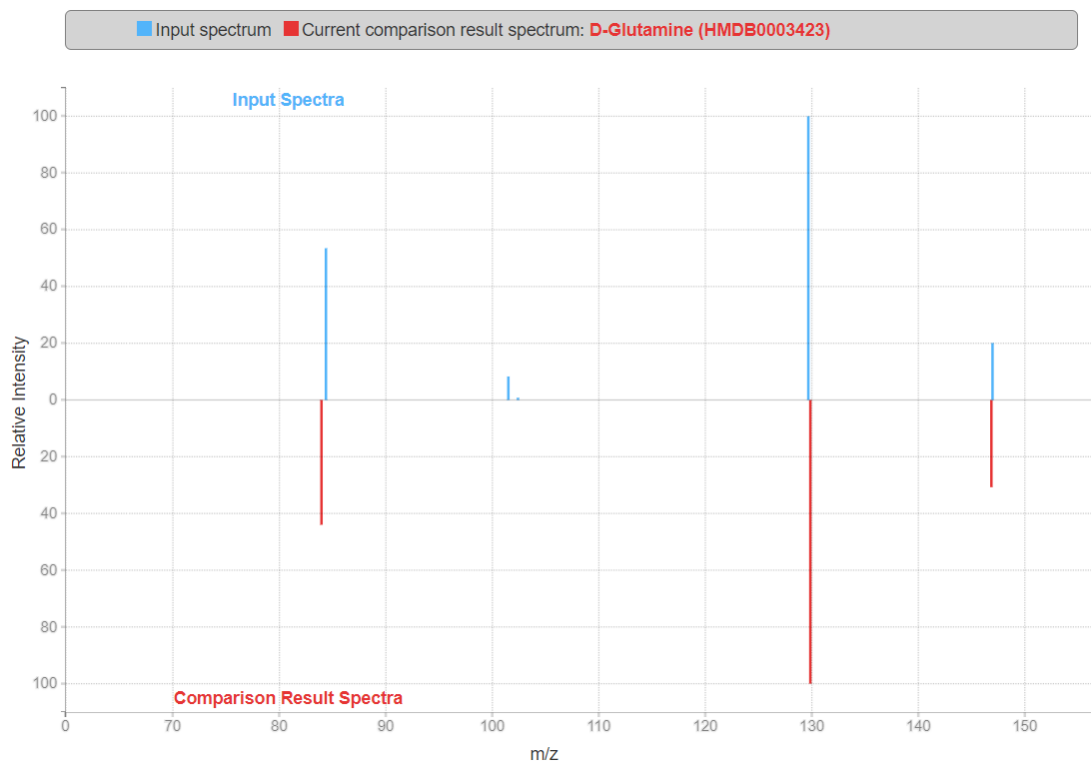


Figure 24 Comparison between the input spectra (blue) and D-Glutamine (red)

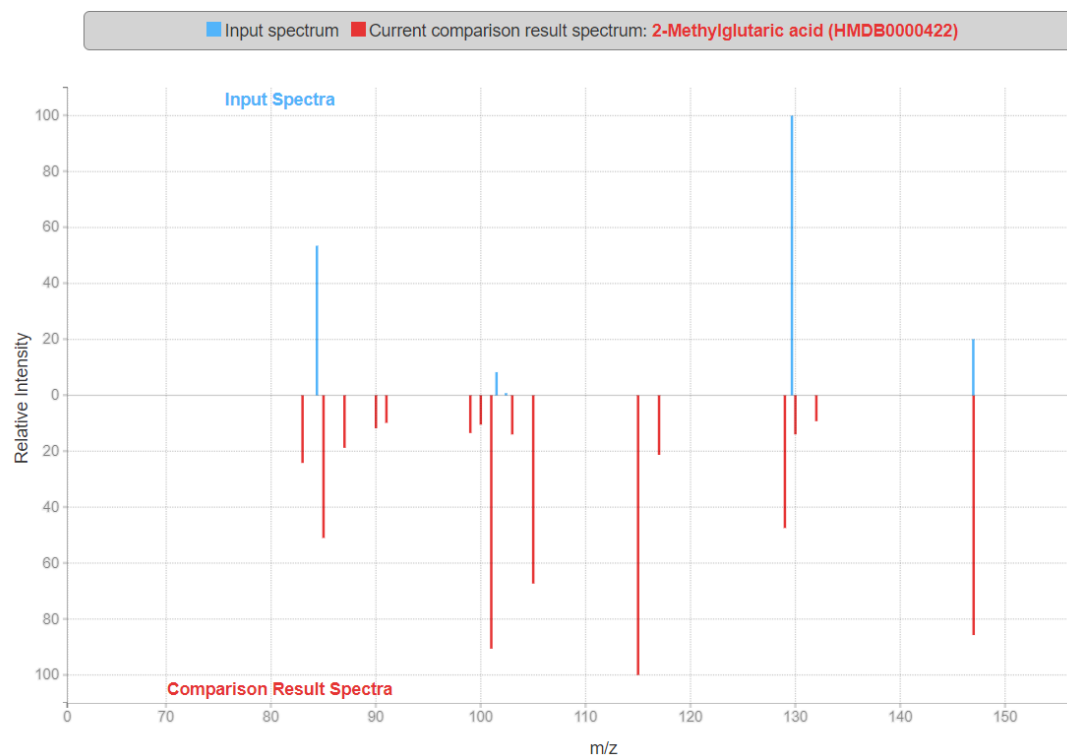


Figure 25 Comparison between the input spectra (blue) and 2-methylglutaric acid (red)

8. LC/MS Search grouped by RT

The coupling of Liquid Chromatography (LC) to Mass Spectrometry (MS) facilitates metabolites identification by reducing the complexity of the data, since it allows separation of the sample over the time. LC is the most used separation technique due to its versatility without losing too much reproducibility.

LC-MS data provides a set of experimental masses (m/z) and retention times. These data pairs are known as features. Those features with the same retention time may be derived from the same analyte. In this case, some of the experimental masses correspond to in-source fragmentation due to the breaking of some parts of the precursor ion.

There are software tools for peak detection as CAMERA that use the peak size and the isotopic profile information for grouping the peaks taking information from the shape of the peak and the isotopic profile. This software tools for peak detection assign the features related to the same group, but in most of the cases, they do not perform an identification of the peaks based on MS/MS data from the putative precursor ions. They group the features and they assign the possible adduct and/or neutral loose suffered by the feature, but the subsequent identification has to be performed.

Due to that, CMM has developed a system where the user can introduce a RT window where all the features will be checked to establish relation between different adducts corresponding to the same ionization mode (positive or negative). This process is called adduct detection. Moreover, it checks if some of the peaks with no relation can correspond to fragments of the features with a higher m/z due to in-source fragmentation.

To do that, it looks into the fragmentation spectra of putative annotations of the compounds with a higher m/z . The MS/MS spectra information of the compounds has been integrated from HMDB (experimental and predicted spectra).

A retention time zero will be assigned for those features without retention time. Features with retention time zero will not be grouped together, as they do not provide information about the elution time into the separation column.

The full process of the LC-MS search grouping the features by their RT is:

1. CMM assigns different groups depending on the tolerance introduced by the user.
2. CMM tries to detect the adduct based on the relationships between peaks within a group.
3. The putative annotations corresponding to different adducts are assigned to each feature. Here CMM marks features as possible fragments of the ones with a greater experimental mass.
4. Iterates the features marked as possible fragments. In this case, each putative fragment has a different ionization hypothesis. Depending on the ionization hypothesis, the possible precursor ion should be ionized consequently (look Table 1 and Table 2). Then, CMM considers the fragmentation pattern (information retrieved from HMDB) to check if the experimental mass of the

possible fragment has been detected in the fragmentation of the putative precursor ions (annotations over the features with a higher m/z).

Table 1 Fragments and the corresponding possible precursor ions. Positive ionization mode.

POSITIVE IONIZATION MODE	
Fragment ion	Precursor ions
M+H	M+H, M+2H, M+3H, M+H-H ₂ O, M+H+NH ₄ , M+H+HCOONa, M+2H+Na, M+H+2K, M+H+2Na, M+H+Na, M+H+K, 2M+H, 2M+H-H ₂ O, M+ACN+2H, M+2ACN+2H, M+3ACN+2H, M+CH ₃ OH+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+NH ₄ , M+H+NH ₄ , 2M+NH ₄
M+H-H₂O	2M+H-H ₂ O, M+H-H ₂ O
M+H+NH₄	M+H+NH ₄
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+CH ₃ OH+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 2 Fragments and the corresponding possible precursors ions. Negative ionization mode.

NEGATIVE IONIZATION MODE	
Fragment ion	Precursor ions
M-H	M-H, M+FA-H, M-H-H ₂ O, 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-H ₂ O
M-H+HCOONa	M-H+HCOONa
M+H-H₂O	2M+H-H ₂ O, M+H-H ₂ O
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
----------------	----------------

The workflow of the entire process is shown on the Figure 26. EM refers to experimental mass and RT to retention time. The final step, the searching over possible in-source fragments, is more detailed on the Figure 27.

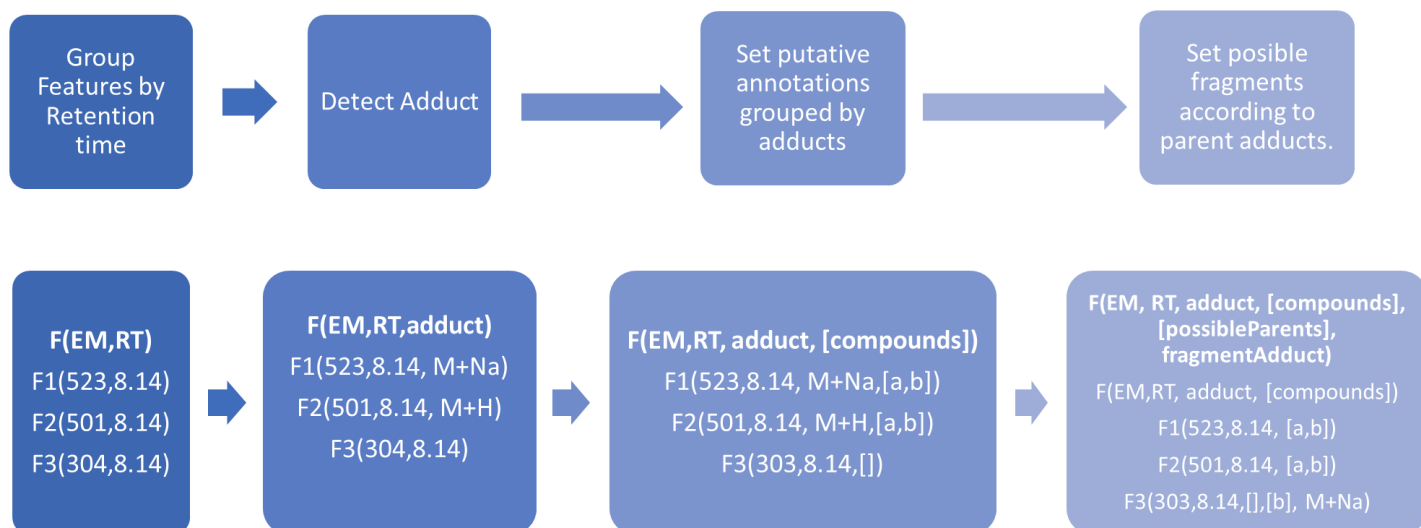


Figure 26 LC-MS search grouping features by RT workflow

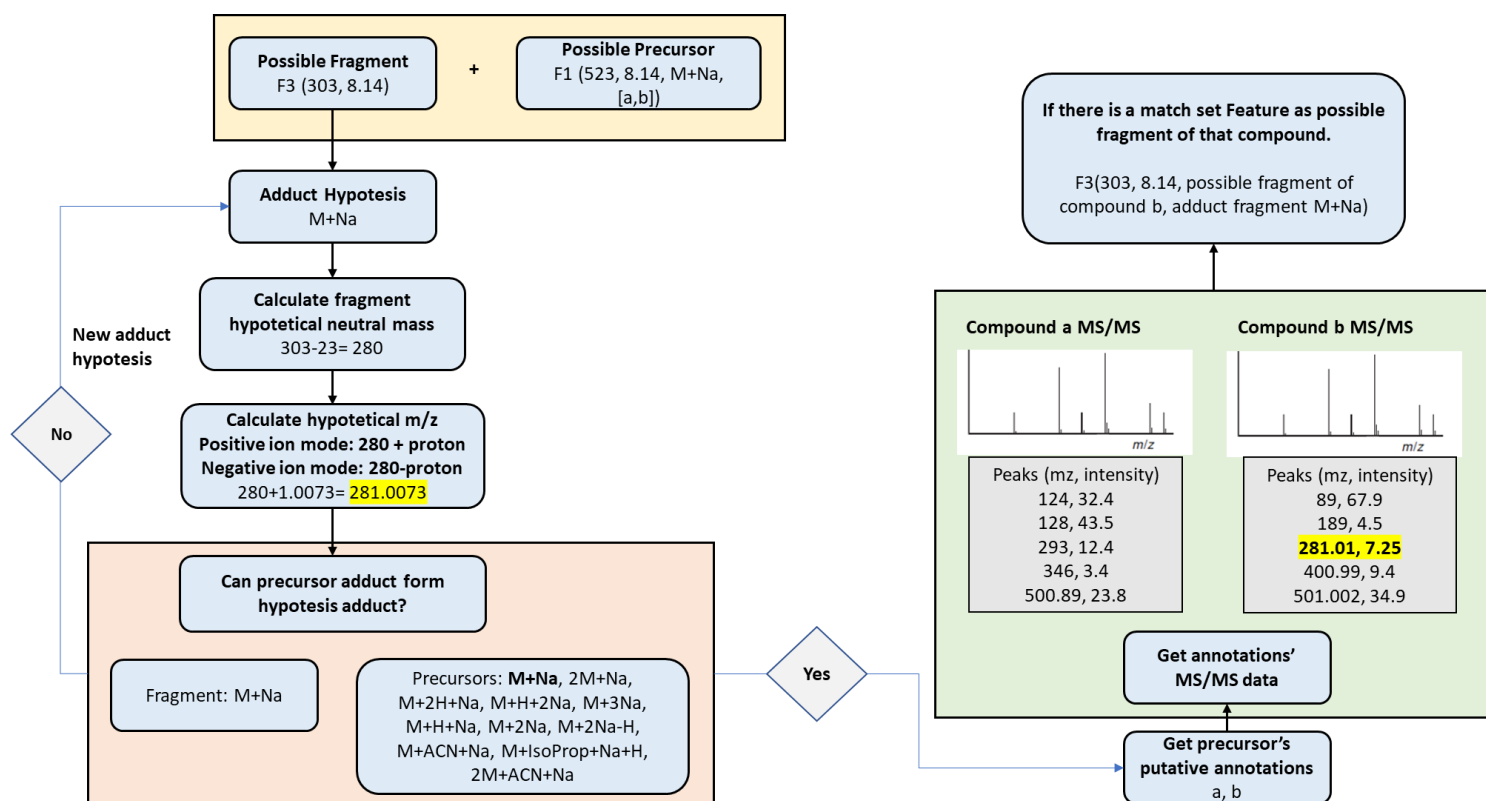


Figure 27 In-source fragmentation search

Figure 28 shows the fields of the LC/MS search grouping the features by their RT. These fields are the same explained in 2.5, except the tolerance introduced to group the features by their RT. The only mandatory field are the experimental masses of the significant compounds, the retention times, composite spectra and non-significant features are employed to apply knowledge.

[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 experimental masses introduced in field **[1]** in the same order.

Even if RTs were not used for supporting annotations, they will be automatically reported for all the annotations, which simplifies further revision since RTs do not have to be added manually.

[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.

CMM takes advantage of the grouping of signals corresponding to the same feature. It automatically detects the target experimental mass and adduct calculating differences between the m/z listed in the CS. This avoids the need of to manually calculate which adduct corresponds to each feature. The goal of this step is the identification of the true mass of the compound M that generated all the signals in the CS. If this detection is successful, only the mass of M will be searched in the database, ignoring the rest of the masses' alterations. The CSs introduced here correspond to the experimental masses 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 experimental masses introduced in field **[4]** in the same order.

[6] All Composite Spectra (CS): The CSs introduced here correspond to the experimental masses 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.

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

The interface is a web-based form for LC-MS data search. It features a blue header and a white body with various input fields and dropdown menus. The form is organized into several sections:

- Experimental Masses (*): [1]**: A text input field with placeholder text "enter significant input masses".
- Retention Times: [2]**: A text input field with placeholder text "enter significant retention times".
- Composite Spectra: [3]**: A text input field with placeholder text "enter significant composite spectra".
- Seleccionar archivo**: A button with the text "Ningún archivo seleccionado".
- All Experimental Masses: [4]**: A text input field with placeholder text "enter all input masses".
- All Retention Times: [5]**: A text input field with placeholder text "enter all retention times".
- All Composite Spectra: [6]**: A text input field with placeholder text "enter all composite spectra".
- RT window width (*): [7]**: A text input field.
- Tolerance (*): [8]**: A text input field with a radio button for "ppm" and a radio button for "mDa".
- Chemical Alphabet (*): [9]**: A dropdown menu with options: "All", "CHNOPS", "CHNOPS + Cl".
- Deuterium:**: A checkbox.
- Modifiers (*):**: A dropdown menu with options: "None", "NH3", "HCOO", "CH3COO", "HCOONH3", "CH3COONH3".
- Databases (*):**: A dropdown menu with options: "All except MINE", "All (Including In Silico Compounds)", "HMDB", "LipidMaps", "Metlin", "Kegg", "In-house", "MINE (Only In Silico Compounds)".
- Metabolites (*):**: A dropdown menu with options: "All except peptides", "Only lipids", "All including peptides".
- Input Masses Mode (*): [13]**: A dropdown menu with options: "Neutral Masses", "m/z Masses".
- Ionization Mode (*): [14]**: A dropdown menu with options: "Neutral", "Positive Mode", "Negative Mode".
- Adducts (*): [15]**: A dropdown menu with options: "All", "M".

At the bottom of the form, there are three buttons: "LOAD DEMO DATA", "SUBMIT COMPOUNDS", and "RESET".

Figure 28 LC-MS search interface

8.1. Result LC-MS Search

The output from this functionality is a couple of nested lists. The outer list corresponds to the different retention times, the inner list consists on; first the set of annotations grouped by adduct and secondly the possible parents if the feature can be a fragment. The result list interface with an example is illustrated on Figure 30. and Figure 31. The input data is shown in Figure 29. The input masses mode was m/z with positive ionization mode and only M+H and M+Na adducts were selected.

[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. Figure 30. **Error! No se encuentra el origen de la referencia.** shows three putative annotations with adduct M+Na from feature with mass 192.0743.

[4] Possible precursor ions: The second part of the inner list illustrates the possible parents of the actual feature and its corresponding adduct. Figure 31 illustrates the feature with mass 90.021938 does not have annotations but is a possibly an in-source fragment from one of the two different annotations of the feature with a EM 192 and a RT within the window established by the user. The “parent” feature mass in inside the red square.

The screenshot displays the input interface for an LC-MS search. It is organized into several sections:

- Experimental Masses (*):** A list box containing the following mass values: 192.0743, 301.1798, 146.481938, 90.021938, and 187.
- Retention Times:** A list box containing the following retention time values: 18.842525, 8.425, 18.842525, 18.842525, and 8.425.
- Input Masses Mode (*):** A dropdown menu with 'Neutral Masses' and 'm/z Masses' (selected).
- Ionization Mode (*):** A dropdown menu with 'Positive Mode' and 'Negative Mode' (selected).
- Adducts (*):** A list box with checkboxes for 'All', 'M+H' (checked), 'M+2H', 'M+Na' (checked), 'M+K', and 'M+NH4'.

Figure 29 Input data for LC-MS search

Results of the experiment																
Features grouped by retention time: 18.842525																
Metabolites found for mass: 192.0743 and retention time: 18.842525 -> 3																
No results for Adduct: M+H																
Id	Name	Formula	Molecular Weight	Retention Time	error PPM	Score1	Score2	Score3	Final Score	Cas	HMDB	Metlin	LipidMaps	KEGG	PubChem	Pathways
No compounds found for the adduct																
Adduct: M+Na -> 3																
Id	Name	Formula	Molecular Weight	Retention Time	error PPM	Score1	Score2	Score3	Final Score	Cas	HMDB	Metlin	LipidMaps	KEGG	PubChem	Pathways
91162	1-Methylhistidine	C7H11N3O2	169.0851	18.842525	0	N/A	N/A	N/A	N/A	332-80-9	HMDB00000001	3741			7020397	SHOW PATHWAYS
48307	3-Methylhistidine	C7H11N3O2	169.0851	18.842525	0	N/A	N/A	N/A	N/A	366-16-1	HMDB00000479	3293		C01152	6971061	SHOW PATHWAYS
157256	Na(pha-Methyl)histidine; N-Methyl-L-histidine	C7H11N3O2	169.0851	18.842525	0	N/A	N/A	N/A	N/A	24886-03-1		65651		C03298	6971273	

Figure 30 LC-MS grouped by RT output. Annotations grouped by adduct from feature with mass 192.0743 and RT 18.8425

Results of the experiment

[1]

Features grouped by retention time: 18.842525

[2]

No Metabolites found for mass: 90.021938 and retention time: 18.842525 [3]

No results for Adduct: M+H

Id	Name	Formula	Molecular Weight	Retention Time	error PPM	Score1	Score2	Score3	Final Score	Cas	HMDB	Metlin	LipidMaps	KEGG	PubChem	Pathways
No compounds found for the adduct																

No results for Adduct: M+Na

Id	Name	Formula	Molecular Weight	Retention Time	error PPM	Score1	Score2	Score3	Final Score	Cas	HMDB	Metlin	LipidMaps	KEGG	PubChem	Pathways
No compounds found for the adduct																

Possible precursor ions found for mass: 90.0219 and retention time: 2.0 -> 2

Fragment with adduct M+Na

Id	Name	Formula	Molecular Weight	EM of feature	Adduct	Retention Time	Cas	HMDB	Metlin	LipidMaps	KEGG	PubChem	Pathways
91162	1-Methylhistidine	C7H11N3O2	169.0851	192.0743	M+Na	18.842525	332-80-9	HMDB00000001	3741			7020397	SHOW PATHWAYS
48307	3-Methylhistidine	C7H11N3O2	169.0851	192.0743	M+Na	18.842525	366-16-1	HMDB00000479	3293		C01152	6971061	SHOW PATHWAYS

Figure 31 LC-MS grouped by RT output. Possible precursor ions of the feature with mass 90.021938 considered a fragment

9. RESTful API

An application program interface (API) for a website is code that allows two software programs to communicate with each other. The API spells out the way to request services from a different program. REST is an architectonic style based on REpresentational State Transfer technology for developing web services oriented to the world wide web (www). REST is simple to use since it is built over the Hypertext Transfer Protocol (HTTP), the protocol used by internet. The REST APIs are independent from the programming language as far as they are able to make HTTP requests.

To facilitate the integration of CMM (CEU Mass Mediator) functionalities into other tools, we have built a RESTful API which exposes two of our services: the batch search service and the batch advanced search service.

The following sections show detailed information on how to invoke these services using the new API.

The solution implemented is a REST API, which is an API that uses HTTP request to GET, PUT, POST and DELETE data. In the case of CMM, the only method implemented is POST, due to the amount of data needed for performing a request: the number of EMs, RTs and composite spectra (CS, groups of signals clustered because they correspond to the same signal like adducts, isotopes, neutral losses, etc.). The number of parameters in the get method would make the request very long and tedious, therefore the requests are accepted only through the POST method. The PUT and DELETE methods are restricted to the administrators of the software tool. The format for exchange the data with CMM services is a JavaScript Object Notation (JSON) included in the body (see SI 2 for the attributes allowed). JSON has become the main format for the data exchange over the internet, and it is commonly supported for most of the standards, tools and technologies. The structure of the JSON will depend on the service requested. In the next sections the different end-points are explained, including the JSON structure, the possible values and one example of an input JSON and the corresponding output JSON.

9.1. Peak search

9.1.1 Batch search

Batch search enables the user to find metabolites through the m/z or the neutral masses. The service is accessed through the following URI: <http://ceumass.eps.uspceu.es/api/v3/batch>

To perform a query, the user must send a **POST** request. This request must include:

- A **Content-type** header set to **application/json**.
- A **request body** with a JSON object that includes all data needed for the query: **masses** to search in CMM, **tolerance** allowed for the putative annotations regarding the masses, **metabolite types** to search, **masses mode**, **ionization mode**, possible **adducts** formed when running the experiment and **databases** that will be included in the search.

The query's attributes, its name, type, default value (the value which will be used if the user does not specify the attribute) and optativity are defined in Table 3. As the value of some attributes is restricted to a range of literals, table \ref{sim:enum} shows the defined enumeration types.

Table 3 Batch Search service – Request - Query

Mandatory	Name	Type	Default value
	masses	array of doubles	-
	tolerance	double (range: [0..100])	10
	tolerance_mode	tolerance_mode_enum	"ppm"
	databases	array of database_enum	"all-except-mine"
	metabolites_type	metabolites_type_enum	"all-except-peptides"
	masses_mode	masses_mode_enum	"mz"
	ion_mode	ion_mode_enum	"positive"
	adducts	array of positive_enum	["M+H", "M+2H", "M+Na", "M+K", "M+NH4", "M+H-H2O"]
		array of negative_enum	["M-H", "M+Cl", "M+FA-H", "M-H-H2O"]
		array of neutral_enum	["M"]

Table 4 Batch Search service - Enumeration types

Name	Values
tolerance_mode_enum	"ppm", "mDa"
database_enum	"all", "all-except-mine", "HMDB", "LipidMaps", "Metlin", "Kegg", "in-house", "mine"
metabolites_type_enum	"all-except-peptides", "only-lipids", "all-including-peptides"
masses_mode_enum	"neutral", "mz"
ion_mode_enum	"neutral", "positive", "negative"
positive_enum	M+3H, M+2H+Na, M+H+2K, M+H+2Na, M+3Na, M+2H, M+H+NH4, M+H+Na, M+H+K, M+ACN+2H, M+2Na, M+2ACN+2H, M+3ACN+2H, M+H, M+Na, M+K, M+NH4, M+H-H2O, 2M+H, 2M+Na, M+H+HCOONa, 2M+H-H2O, 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, M+H-2H2O, M+NH4-H2O, M+Li, 2M+2H+3H2O
negative_enum	M-3H, M-2H, M-H2O-H, M-H, M+Na-2H, M+Cl, M+K-2H, M+FA-H, M+Hac-H, M+Br, M+TFA-H, 2M-H, 2M+FA-H, 2M+Hac-H, 3M-H, M-H+HCOONa, M+F
neutral_enum	"M"

The following example shows a query to the Batch Search service:

```
{
  "metabolites_type": "all-except-peptides",
  "databases": ["hmdb"],
  "masses_mode": "mz",
  "ion_mode": "positive",
  "adducts": ["all"],
  "tolerance": 10.0,
  "tolerance_mode": "ppm",
  "masses": [400.3432, ..., 288.2174]
}
```

If the request contains no errors and is therefore correctly processed, the service returns a set (see Table 5) of putative annotations for the masses submitted. Each putative annotation structure (see Table 6) contains the name of the putative annotation compound, its formula, its molecular weight, the difference between the molecular weight and the corresponding experimental mass (ppm), and references of the compound in external databases.

Table 5 Batch Search service – Response - Results

Name	Type	Default value
results	Array of putative_annotation_object (see Table 6)	-

Table 6 Batch Search Service – Response – Putative annotations

Putative_annotation_object	
Name	Type
identifier	integer
EM	double

name	string
formula	string
adduct	positive_enum
	negative_enum
	neutral_enum
molecular_weight	double
error_ppm	integer
ionizationScore	integer (Range: -2, [0..2])
FinalScore	integer (Range: -2, [0..2])
cas	string
kegg_compound	string
kegg_uri	string
hmdb_compound	string
hmdb_uri	string
lipidmaps_compound	string
lipidmaps_uri	string
metlin_compound	string
metlin_uri	string
pubchem_compound	string
pubchem_uri	string
pathways	array of strings

While some of these attributes are related with score rules, please bear in mind that rules are only applied when using the Batch Advanced Search service. Therefore, when using the batch search, all the putative annotations returned will have a score of -2, which shows that the rules engine has not been used in this type of search. Check the section 9.1.2.

This example shows the results of a successful request:

```
{
  "results": [
    {
      "identifier": 32600,
      "EM": 400.3432,
      "name": "Palmitoylcarnitine",
      "formula": "C23H45NO4",
      "adduct": "M+H",
      "molecular_weight": 399.334858933,
      "error_ppm": 3,
      "ionizationScore": -2,
      "finalScore": -2,
      "cas": "2364-67-2",
      "kegg_compound": "C02990",
      "kegg_uri": "http://www.genome.jp/dbget-bin/www_bget?cpd:C02990",
      "hmdb_compound": "HMDB0000222",
      "hmdb_uri": "http://www.hmdb.ca/metabolites/HMDB0000222",
      "lipidmaps_compound": "LMFA07070004",
      "lipidmaps_uri": "http://www.lipidmaps.org/data/LMSDRecord.php?LMID=LMFA07070004",
      "metlin_compound": "961",
      "metlin_uri": "https://metlin.scripps.edu/metabo_info.php?molid=961",
      "pubchem_compound": "11953816",
      "pubchem_uri": "https://pubchem.ncbi.nlm.nih.gov/compound/11953816",
      "pathways": []
    },
    ...
  ]
}
```

9.1.2 Batch advanced search

Batch advanced search also enables the user (on top of the functionality offered by the batch search service) to find metabolites through the m/z or the {neutral masses query parameters. But, in contrast with the batch search service, it uses additional information devoted to biomarker discovery experiments using LC/MS. The service is accessed through the following URI: <http://ceumass.eps.uspceu.es/mediator/api/v3/advancedbatch>.

To perform a query, the user must send a **POST** request. This request must include:

- A **Content-type** header set to **application/json**.
- A request body with a **JSON** object that includes all data needed for the query. In this case, the query is just an extension of the Batch Search query. Therefore, it must include all attributes described in Table 3 and, on top of that, provide the additional information shown in Table 7: RTs, composite spectra (spectra created by the summation of all co-eluting m/z ions that are related), chemical alphabet (possible elements of the putative annotations), etc.

Table 7 Batch Advanced Search Service – Request – Query – Extra attributes

	Name	Type	Default value
mandatory	chemical_alphabet	chemical_alphabet_enum	“CHNOPS”
	deuterium	boolean	false
	modifiers_type	modifiers_type_enum	“none”
optional	retention_times	array of doubles	empty
	composite_spectra	array of arrays of spectra_object (see Table 8)	empty
	all_masses	array of doubles	empty
	all_retention_times	array of doubles	empty
	all_composite_spectra	array of arrays of spectra_object (see Table 8)	empty

Table 8 Batch Advanced Search Service – Request – Spectra

Spectra_object		
Name	Type	Default value
mz	double	-
intensity	double	-

Table 9 Batch advanced Search Service – Enumeration types

Name	Type
chemical_alphabet_enum	“CHNOPS”, “CHNOPSCL”, “ALL”
modifiers_type_enum	“none”, “NH3”, “HCOO”, “CH3COO”, “HCOONH3”, “CH3COONH3”

The next example shows the JSON structure of a query for the Batch Advanced Search service:

```
{
  "chemical_alphabet": "all",
  "modifiers_type": "none",
  "metabolites_type": "all-except-peptides",
  "databases": ["hmdb"],
  "masses_mode": "mz",
  "ion_mode": "positive",
  "adducts": ["all"],
  "deuterium": false,
  "tolerance": 10.0,
  "tolerance_mode": "ppm",
  "masses": [400.3432, ..., 288.2174],
  "all_masses": [],
  "retention_times": [18.842525, ..., 4.021555],
  "all_retention_times": [],
  "composite_spectra": [
    [
      {
        "mz": 400.3432,
        "intensity": 307034.88
      },
      ...,
      {
        "mz": 311.20145,
        "intensity": 400.03336
      },
      ...
    ]
  ]
}
```

When using the Batch Advance Search service, CMM scores the putative annotations based on expert knowledge. Thus, the response structure of this service contains all attributes already defined in Table 6, plus some other attributes defined in Table 10.

Table 10 Batch Advanced Search service - Response - Putative Annotation - Extra Attributes

Putative_annotation_object – additional attributes	
Name	Type
RT	double
adductRelationScore	integer (Range: -2, [0..2])
RTScore	integer (Range: -2, [0..2])

This example shows the results of a successful request:

```
{
  "results": [
    {
      "RT": 8.144917,
      "adductRelationScore": -2,
      "RTscore": 2,
      "identifier": 111123,
      "EM": 338.2299,
      "name": "MG(0:0/i-12:0/0:0) ",
      "formula": "C15H30O4",
      "adduct": "M+ACN+Na",
      "molecular_weight": 274.214409446,
      "error_ppm": 1,
      "ionizationScore": -2,
      "finalScore": 2,
      "kegg_compound": "",
      "kegg_uri": "",
      "hmdb_compound": "HMDB0072858",
      "hmdb_uri": "http://www.hmdb.ca/metabolites/HMDB0072858",
      "lipidmaps_compound": "",
      "lipidmaps_uri": "",
      "metlin_compound": "",
      "metlin_uri": "",
      "pubchem_compound": "131779644",
      "pubchem_uri": "https://pubchem.ncbi.nlm.nih.gov/compound/131779644",
      "pathways": []
    },
    ...
  ]
}
```

10. R library

Our colleague Yaxoiang Li (<https://github.com/lzyacht>) has created an R library to consume the REST API of CMM (<https://github.com/lzyacht/>). The CMM team acknowledges his efforts for his kindness help.

10.1. CMMR - Ceu Mass Mediator API in R

10.1.1 Installation

```
install.packages(devtools)
devtools::install_github("lzyacht/cmmr")
```

10.1.2 Example

Batch search all result in positive mode:

```
library(cmmr)
df_pos = batch_search_full('all-except-peptides',
                           '["all-except-mine"]',
                           'mz',
                           'positive',
                           '["M+H", "M+Na"]',
                           10,
                           'ppm',
                           system.file("extdata",
"unique_mz.csv", package = "cmmr"))
```

Batch search all result in negative mode:

```
library(cmmr)
df_neg = batch_search_full('all-except-peptides',
                           '["all-except-mine"]',
                           'mz',
                           'negative',
                           '["M-H", "M+Cl"]',
                           10,
                           'ppm',
                           system.file("extdata",
"unique_mz.csv", package = "cmmr"))
```


Advanced batch search:

```
df = advanced_batch_search(  
    chemical_alphabet = 'all',  
    modifiers_type = 'none',  
    metabolites_type = 'all-except-peptides',  
    databases = ['hmdb'],  
    masses_mode = 'mz',  
    ion_mode = 'positive',  
    adducts = ['all'],  
    deuterium = 'false',  
    tolerance = '7.5',  
    tolerance_mode = 'ppm',  
    masses = '[400.3432, 288.2174]',  
    all_masses = '[]',  
    retention_times = '[18.842525, 4.021555]',  
    all_retention_times = '[]',  
    composite_spectra = '[[{ "mz": 400.3432, "intensity":  
307034.88 }, { "mz": 311.20145, "intensity": 400.03336 }]]',  
    cmm_url =  
'http://ceumass.eps.uspceu.es/mediator/api/v3/advancedbatch')
```

11. Manual

This section corresponds to the download of the CMM manual in PDF.

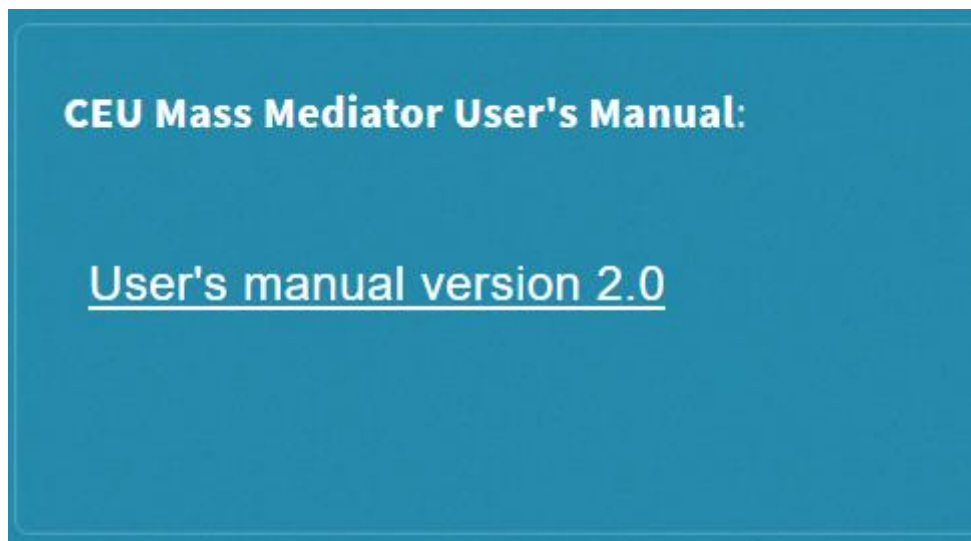


Figure 32 User's manual page