



# **Tandem Mass Spectrometry workflow in EI-MAVEN**

By

**Shubhra Agrawal**

Under the supervision of Dr. Abhishek Jha & Dr.  
Ganesh Bagler

Indraprastha Institute of Information Technology  
Delhi  
2016-2018

# **Tandem Mass Spectrometry workflow in El-MAVEN**

By

**Shubhra Agrawal**

Submitted in partial fulfillment of the requirements  
for the degree of Master of Technology

To

**Indraprastha Institute of Information Technology,  
Delhi  
November, 2018**

## Certificate

---

This is to certify that the thesis titled “**Tandem Mass Spectrometry workflow in El-MAVEN**” being submitted by **Shubhra Agrawal** to the **Indraprastha Institute of Information Technology**, Delhi for the award of **Master of Technology (Computational Biology)**, is an original work of research work carried out by her under my supervision. In my opinion, the thesis has reached the standards fulfilling the requirements of the regulations relating to the degree.

The results contained in this thesis have not been submitted in part or full to any other university or institute for the award of any degree/diploma.

Supervisor: Dr. Ganesh Bagler,  
Indraprastha Institute of Information Technology, New Delhi

Co-supervisor: Dr. Abhishek Jha,  
Elucidata Corporation, Cambridge, MA

## Acknowledgements

---

I sincerely acknowledge and thank my supervisors Dr. Abhishek Jha and Dr. Ganesh Bagler for the opportunity to work on this project as well as their valuable guidance during the course of this project. I would also like to thank Dr. Eugene Melamud and Dr. Phillip Seitzer from Calico who were instrumental in solving some of the most challenging algorithmic problems in this project.

I acknowledge the contributions and support from my teammates at Elucidata, Raghav Sehgal, Sahil Kumar, Rishabh Gupta and Saiful Bari Khan, without whom the quality of results would not be the same.

## Abstract

---

Metabolomics, defined as the study of an organism's entire metabolic profile, is a direct read-out of the physiological changes at the cellular level and has the potential to positively inform drug-target discovery and biomarker identification.

Mass Spectrometry is one of the most popular techniques used to measure the levels of metabolites present in biological samples. Tandem Mass Spectrometry, and more commonly, Data Dependent Acquisition (DDA) has become a trusted technique for metabolite identification and quantification due to its dependence on spectral pattern matching with existing libraries.

Since existing mass spectrometry data processing tools are either vendor-specific or difficult to use, the DDA workflow has been added to El-MAVEN, an open-source mass spectrometry data processing tool, maintained by Elucidata. Spectral matching capabilities have been added as part of the targeted DDA workflow and algorithmic improvements have been made to the untargeted workflow for optimum results.

Additional widgets and features have been added for a better user experience in data curation. The improvements have been validated against known standards using datasets obtained from Elucidata's partner labs.

# Table of Contents

---

1. Introduction	10
1.1 Mass spectrometry in metabolomics	
1.2 Data variation in Mass spectrometry	
1.3 Challenges with existing data processing tools	
2. Mass spectrometry data analysis using El-MAVEN	13
2.1 History of El-MAVEN	
2.2 Improvements over MAVEN	
2.3 LCMS analysis in El-MAVEN	
2.4 Need for Tandem Mass Spectrometry	
3. Data Dependent Acquisition in El-MAVEN	17
3.1 Data Dependent Acquisition method	
3.2 Spectral libraries for DDA	
3.3 Spectra matching	
3.4 Visualizations for DDA method	
3.4.1 Fragmentation spectra widget	
3.4.2 Fragmentation event markers	
3.4.3. Fragmentation event list	
3.4.4 Match Compounds widget	
3.5 Targeted DDA workflow in El-MAVEN	
3.6 Results	
4. Mass slicing for untargeted detection	27
4.1 Challenges in untargeted detection	
4.2 Mass slicing algorithm in MAVEN	
4.3 Understanding the challenges in the original algorithm	
4.3.1 Inconsistency of results across sessions	
4.3.2 Duplicate features	
4.3.3 Incorrect mass resolution	
4.4 Mass slicing 2.0	
4.5 Additional improvements in untargeted pipeline	
4.6 Untargeted DDA workflow in El-MAVEN	
4.7 Results	
5. Conclusion	40
6. References	41

# List of Figures

---

## 2. Mass spectrometry data analysis using El-MAVEN

### 2.3 LCMS analysis in El-MAVEN

Fig. 1 a) Liquid chromatography process separating two components in a mixture

Fig. 1 b) Internal working of a Mass Spectrometer

Fig. 2 a) Raw data with m/z, RT and intensity axes

Fig. 2 b) Chromatogram for a selected m/z range in El-MAVEN across samples where each sample is represented by a different colour

### 3. Data Dependent Acquisition in El-MAVEN

Fig. 3 Spectral library in El-MAVEN with fragment information

Fig. 4 Fragmentation spectra widget with comparison of group and reference spectra

Fig. 5 EIC with fragmentation markers on the x-axis

Fig. 6 Fragmentation Events widget captures information for every MS2 event in a peak group

Fig. 7 Match compound widget for acetoacetate within the selected database. MS2 score and RT deviation are displayed for ease of identification

### 3.5 Targeted DDA workflow in El-MAVEN

Fig. 8 Summary of the complete targeted workflow for DDA datasets in El-MAVEN

### 3.6 Results:

Fig. 9 Distribution of auto-detected peak groups after spectra matching

## 4. Mass slicing for untargeted detection

### 4.2 Mass slicing algorithm in MAVEN

Fig. 10 Original mass slicing algorithm in MAVEN

### 4.3 Understanding the challenges in the original algorithm

Fig. 11 Results with 9580 and 9619 features in 2 different sessions

Fig. 12 Multiple peak groups detected at the same m/z and RT values and their intensity distribution across samples

Fig. 13 a) Peak bubble without EIC

Fig. 13 b) Peak bubbles displaced on the EIC

#### **4.4 Mass Slicing 2.0**

Fig. 14 Data-driven mass slicing algorithm in El-MAVEN

#### **4.5 Additional improvements in untargeted pipeline**

Fig. 15 Peak detection dialog with the options for untargeted detection with annotation and spectral matching turned on

#### **4.6 Untargeted DDA workflow in El-MAVEN**

Fig. 16 Automated untargeted workflow for DDA data in El-MAVEN

Fig. 17 Comparative analysis of peak groups



# List of Tables

---

## 4. Mass slicing for untargeted detection

### 4.7 Results

Table 1. El-MAVEN parameters for the untargeted runs. Default values were used for all other parameters

Table 2. Performance comparison of the untargeted workflow in El-MAVEN v0.9.1 and v0.10.0

# Abbreviations

---

LCMS- Liquid Chromatography Mass Spectrometry

GCMS- Gas Chromatography Mass Spectrometry

m/z- Mass to charge ratio

RT- Retention time

FDA- Food and Drug Administration

NMR- Nuclear Magnetic Resonance

EIC- Extracted Ion Chromatogram

DDA- Data Dependent Acquisition

DIA- Data Independent Acquisition

UI- User Interface

# 1. Introduction

---

## 1.1 Mass spectrometry in metabolomics

Metabolomics involves the identification and quantification of metabolites and small molecules in biological specimens. The past two decades have been dominated by research in the genomics, transcriptomics and proteomics fields, in the hopes that it will help us understand the biological system well enough to predict cellular responses in case of environmental perturbations. While that has proven true to some extent, it is metabolomics that acts as a direct read out of the physiological state of an organism as it is more tissue specific and provides a closer look at the molecular mechanisms in play. It is now being used to understand complex diseases and in discovering new therapeutic targets and biomarkers to diagnose and monitor diseased conditions. In certain cases, metabolomics has acted as an early indicator of diseases [1]. A notable example in this respect is Agios, a pharmaceutical company that has managed to win 2 FDA approvals within 6 years for finding supplemental applications using a multi-omics approach focusing on metabolomics. [2] [3]

NMR and Mass spectrometry have evolved into the two main tools for detecting small molecules [4]. NMR is used for quantitative measurement and although it avoids the extra sample preparation steps required in MS, Mass spectrometry combined with chromatographic techniques has emerged as the more common tool for metabolomics research due to its superior sensitivity. Much of the progress in the field of metabolomics in the past decade is owing to the rapidly improving, sophisticated instruments for detecting and measuring metabolites. Since metabolites, unlike genes, transcripts and proteins, show much more disparity in nature, there is no one technique that can resolve the entire metabolome at the same time. A combination of methods has to be used for extraction, detection, quantification and identification of all metabolites in a biological sample. This, combined with the lack of software that can handle the various data types generated across instruments/methods is a major bottleneck in streamlining and standardizing metabolomic workflows [5].

## 1.2 Data variation in Mass spectrometry

There are a growing number of data acquisition methods for metabolomics as the hardware capabilities get better. In this section, we will discuss the ways in which MS data can vary.

There are 3 major points of differentiation in the setup: 1) Machine vendor, since every vendor has a proprietary format where the data can only be visualized using the software that is shipped with the hardware. There are already more formats available than the number of machine vendors. 2) Chromatography technique used in combination with MS can differ; Liquid chromatography or Gas chromatography could be used based on the volatility of analytes. 3) Acquisition method can vary greatly based on the type of analysis. The methods can differ on the basis of the range of mass covered by the MS, level and type of fragmentation, and ionization.

Each vendor has their own software suite to handle the data from their machines. Since a single lab can have machines from one or more vendors, analysts need to be proficient in a variety of software tools, and there is no consistency in visualization of results, or the processing methods employed.

## 1.3 Challenges with existing data processing tools

The raw data obtained from the MS is three-dimensional and therefore cannot be used for analysis directly. Every MS machine vendor has a proprietary data processing software that is installed along with the hardware to monitor the experiment but is not very efficient for data analysis because of the following reasons:

- Consistency- A metabolomics lab might have instruments from different vendors. The results obtained from each of these software will look different
- Cost- There is a licensing fee associated with vendor applications. An independent user will have to purchase processing software from multiple vendors as these applications tend to only work with that specific vendor format files

- Compatibility- Some of these tools are platform-dependent, for e.g., MultiQuant from SciEx is only available for Windows systems.

There are a number of open-source MS data processing software available today, such as, MAVEN [6], XCMS [7], MzMine [8] etc. They have certain advantages over proprietary software in the case of cost and compatibility, but there are well-documented issues with these applications with respect to false positives and inaccuracy in peak detection [9][10][11]. Many such applications also suffer from poor interactivity on the user interface which results in a long feedback loop where many rounds of analysis are required to obtain accurate peak results [6]. There are also concerns around the speed and reliability of these applications in case of high-throughput datasets where the number of samples is very large.

In order to resolve the issues present in both proprietary and open-source applications, and accelerate the speed at which such analyses can be performed, a popular open-source tool, MAVEN, was extended as El-MAVEN to allow interactive, fast, efficient and reliable analysis of MS and MS/MS datasets in just four steps.

## 2. Mass spectrometry data analysis using El-MAVEN

---

### 2.1 History of El-MAVEN

El-MAVEN is an open-source, vendor-neutral, platform-independent desktop tool with graphical and command line interface, written in C++ [12]. It is a widgets-based application, supported by the Qt Framework, which allows a number of visualization tools within the main application that can be rearranged on the user interface based on user preference. It is an extension of MAVEN, a software developed by the Rabinowitz Lab at Princeton in 2010.

MAVEN was known for its visualizations and flexibility in data processing methods and was adopted by a number of metabolomics labs working with labeled data. The software has been cited over 500 times since 2010. A few years later, the original group stopped maintenance work on the project and since its scope was fairly limited, and it had a number of bugs and stability issues, the Elucidata group was commissioned to resolve the issues in the original software and expand the capabilities of MAVEN.

Since 2018, over a thousand users have processed their mass spectrometry data using El-MAVEN with an active usage of 400 users per month.

### 2.2 Improvements over MAVEN

The core requirements for the development of El-MAVEN was the need to address some stability and usability issues encountered in MAVEN for larger datasets (~100 samples or more). The software would often crash during the analysis leading to data loss and reanalysis of the dataset.

El-MAVEN has been under active development since 2016 by the Elucidata group. In the last 4 years, significant improvements have been made for stability and ease of installation of the software on all platforms with additional support for a wider variety of MS data and other algorithmic improvements. Some of the major improvements include 1) multiprocessing

for faster sample import, 2) crash reporting system with session restore capability to prevent data loss, 3) more streamlined automation of peak detection, 4) new algorithms and finer control over peak detection, 5) enhanced LC-MSMS capabilities, 6) additional visualizations and 7) direct integration with Polly, an online platform for downstream processing of LCMS data.

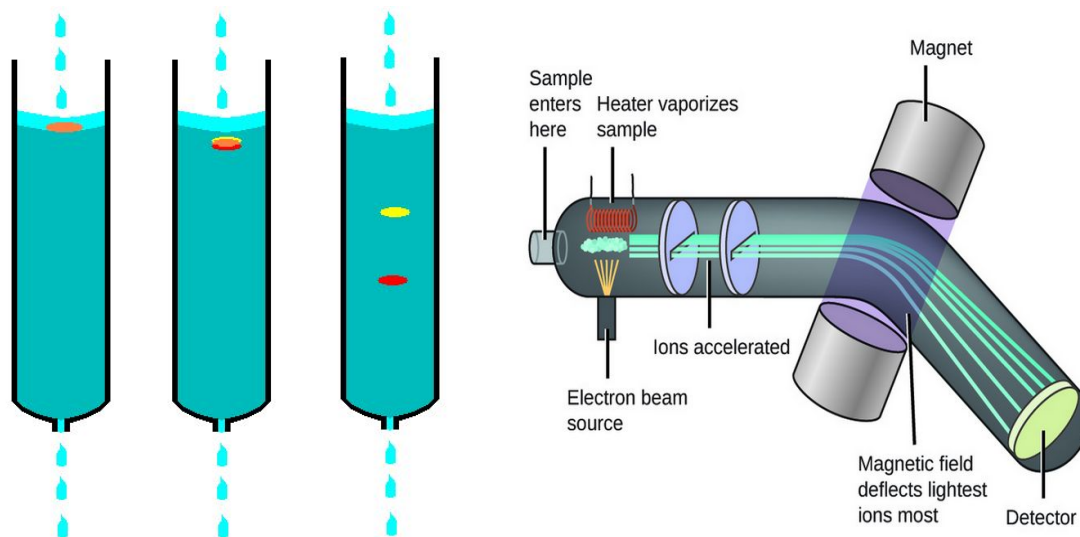
The above improvements make El-MAVEN a strong choice for metabolomic analysis and hence there is a demand to expand the scope of the application to support other types of MS data along with improved metabolite detection and identification.

### **2.3 LCMS analysis in El-MAVEN**

Before discussing the types of data supported in El-MAVEN, it might be helpful to go over the basic LCMS setup and data structure.

The most common setup for mass spectrometry in metabolomics is the LCMS setup where liquid chromatography (LC) technique is combined with Mass spectrometry (MS) to obtain better resolution of data.

LC is a technique to physically separate components of a liquid mixture based on the component's chemical affinity to the stationary matrix in the LC column. Every component takes a certain amount of time to flow through the whole column known as the retention time (RT). The LC directly feeds into the Mass spectrometer.

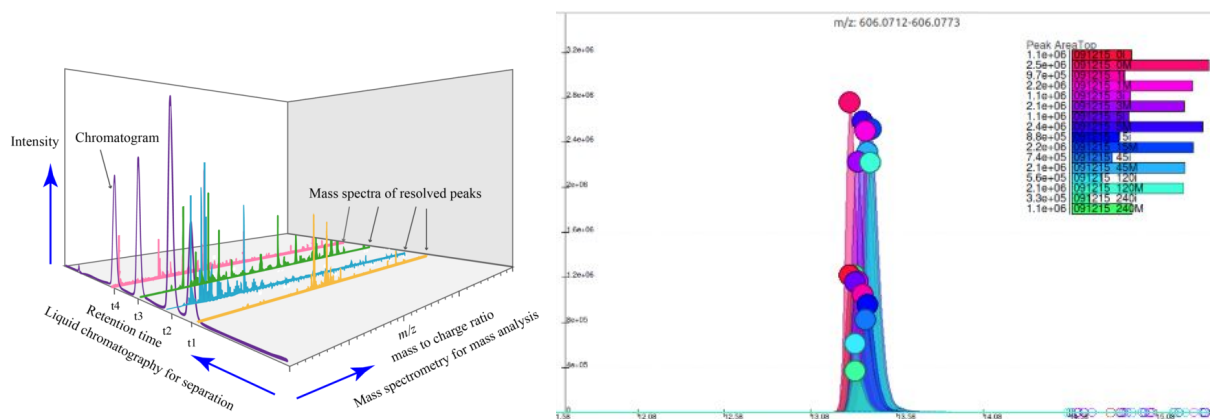


**Fig. 1 a) Liquid chromatography process separating two components in a mixture b) Internal working of a Mass Spectrometer**

A mass spectrometer has three major components. A sample flows through the LC and into the *ionization chamber* where a charge is added to the metabolites using different ionization techniques. The ions flow into the *mass analyzer* where a magnetic force is applied to accelerate them across the chamber. The accelerated ions hit the *detector plate* at an angle, based on the mass/charge ratio of the ion. The ion count (or intensity) is used as a proxy for relative concentration of the metabolite. This data is used to detect the mass profile of a biological sample, while the time axis helps in resolving components with the same mass/charge ratio ( $m/z$ ).

The raw data, therefore, has 3 axes-  $m/z$ , RT and intensity. Once this data is imported into El-MAVEN, the chromatogram can be extracted for any given  $m/z$ . The chromatogram or Extracted Ion Chromatogram (EIC) is a 2-D plot with RT on the X-axis and intensity on the Y-axis. El-MAVEN, unlike some other tools, displays chromatograms across samples in a single visualization where the relative intensity for every sample is shown by a barplot that is easy to interpret.





**Fig 2. a) Raw data with m/z, RT and intensity axes, b) EIC for a selected m/z range in El-MAVEN across samples where each sample is represented by a different colour**

A peak group in a chromatogram is the collection of near-Gaussian peaks at the same m/z and RT across samples and represents the signal from a single molecular species. Given a list of metabolites with their chemical formula, El-MAVEN can run an automated analysis with user-defined filters to generate a list of peak groups that have been found in the dataset, along with any isotopologues that might be present. The RT information can be used to identify metabolites in case standards have been run.

## 2.4 Need for Tandem Mass Spectrometry

A basic LCMS setup has limited capabilities in terms of identification of metabolites. Since there can be multiple molecular species with the same m/z, RT information has to be obtained from running known standards in order to identify a particular metabolite. For better identification of metabolites, Tandem Mass spectrometry (LC-MSMS) is employed where the ions are fragmented in a second MS step giving rise to a mass spectra that is characteristic of the molecular structure of that metabolite [13].

There are a number of ways to perform an LC-MSMS experiment and El-MAVEN supports some of them already. Since DDA and DIA methods are becoming more popular, efforts have been made to add functionalities that can ease the process of analyzing DDA data within El-MAVEN. These improvements will be discussed in detail in Chapter 3.

## 3. Data Dependent Acquisition in EI-MAVEN

---

In the previous section, we discussed the many challenges present in most MS data processing tools and the ways in which EI-MAVEN offers an improved, efficient platform for MS data processing. In this chapter, we will discuss the DDA workflow and the engineering effort required to implement the functionality.

### 3.1 Data Dependent Acquisition method

DDA has become the standard acquisition method for LC-MSMS experiments where both precursor and fragmentation information is collected for n-most intense metabolites. There is one MS1 full-scan to identify the n-most intense m/z and then a series of MS1 and MS2 scans running in parallel to capture the complete spectra for the selected parent ions. This method generates clean mass spectra that can be used for metabolite identification using available spectral libraries.

Following feature additions were done in EI-MAVEN to add complete support for DDA data analysis:

- Parser for spectral libraries
- Calculating average spectra for peak groups
- Automated spectra matching against the library
- Visualisation for fragmentation spectra

### 3.2 Spectral libraries for DDA

Since fragmentation patterns depend on the chemical structure of the metabolite, the fragmentation spectra is a reliable metric for metabolite identification. Apart from the metabolite name and other physical properties, a spectral library stores the collision energy used for fragmentation of the parent ion, fragment masses (m/z) observed after the event and the relative intensities of every fragment in the spectra.

These libraries are often sourced from core metabolomics facilities that have a high data throughput and many are made publicly available on databases like MoNA [14] and NIST.

A new parser was written for common spectral library formats like .MSP and .mgf which are both user-readable formats with enough flexibility to handle variations coming from different sources.

name	m/z
(2-AMINOETHYL)PHOSPHONATE-20.0,...	126.031456
Collision Energy	200.000
Formula	C2H8NO3P
Fragments	62.96;77.98;78.96;79.97;94.99;106.01;106.99;124.02
(R,R)-TARTARIC ACID-20.0,50.0,100.0	151.023712
Collision Energy	200.000
Formula	C4H6O6
Fragments	59.01;72.99;75.01;87.01;103.00;105.02;131.00;149.01
(R)-MALATE-20.0,50.0,100.0	135.028809
Collision Energy	200.000
Formula	C4H6O5
Fragments	59.01;71.01;72.99;87.01;89.02;115.00;133.01

Fig 3. Spectral library in El-MAVEN with fragment information

Since fragmentation also depends on the polarity of the ion, care should be maintained to use the correct library for a given dataset. Keeping this in mind, a popup was added in El-MAVEN in case users import a library with a different polarity than the dataset in use.

### 3.3 Spectra matching

El-MAVEN relies on peak groups to reduce the amount of curation required by the user. The same principle is applied during spectral matching. An average fragmentation pattern of each group is calculated that will ultimately be used for matching against the spectral library.

Following steps are taken to create an average fragmentation pattern:

- Fetch all MS2 scans across all peaks in a group
- Use the scan with the most fragments as a seed for average spectra
- All fragments and their intensities are added to this spectra

- In case two or more fragments fall within the user-defined mass resolution, their intensities are averaged out to create the group's average spectra

A scoring algorithm has been put in place to assign scores to signify the quality of the fragmentation pattern match. The algorithm assumes a hypergeometric distribution for how frequently fragments match between the group and library spectra. The null hypothesis in this case is that any matches between the two spectra are completely random [15]. The probability of k random matches when there are m fragments in the group spectra and n fragments in the library spectra can be calculated by the following formula:

$$P(k) = \left[ \frac{C_k^m \times C_{n-k}^{N-m}}{C_n^N} \right]$$

N is the total number of possible fragments. Taking the negative logarithm of this probability is reported as the score which represents that the probability of k random matches is  $e^{-score}$ . Therefore, a higher score indicates a better match.

Every peak group is assigned a hypergeometric score based on the quality of its spectra match against the metabolite in the library. For automated peak detection, users can set a threshold for the match score to filter out unreliable matches. This significantly reduces the number of candidates per metabolite compared to the LCMS method.

Apart from the Match score, users can also set the minimum number of fragments that need to match before reporting a candidate metabolite during automated peak detection.

### 3.4 Visualizations for DDA method

In order to assist the user in making the most informed identification, several UI elements were added to El-MAVEN.

#### 3.4.1 Fragmentation spectra widget

A new widget was added to El-MAVEN for side-by-side comparison of fragmentation patterns for the detected group and the recorded metabolite in

the library. Clicking on a peak group opens the fragmentation spectra widget for verification of spectra matching as shown in Fig. 4.



Fig 4. Fragmentation spectra widget with comparison of group and reference spectra

The number of matching fragments have been highlighted in blue in the reference spectra whereas the smaller fragments that do not find a match in the group spectra are marked red. When manually verifying the matches, the distribution of red and blue fragments should be taken into consideration along with the relative intensity of the matches and mismatches.

The widget also displays the purity of the calculated spectra. The purity of a spectra denotes whether or not there were co-eluting parent ions in the MS1 scan, in which case the spectra is representative of fragments created from all the co-eluting ions instead of the parent in question. In cases of high matches and mismatches, purity can provide valuable information for identification.

### 3.4.2 Fragmentation event markers

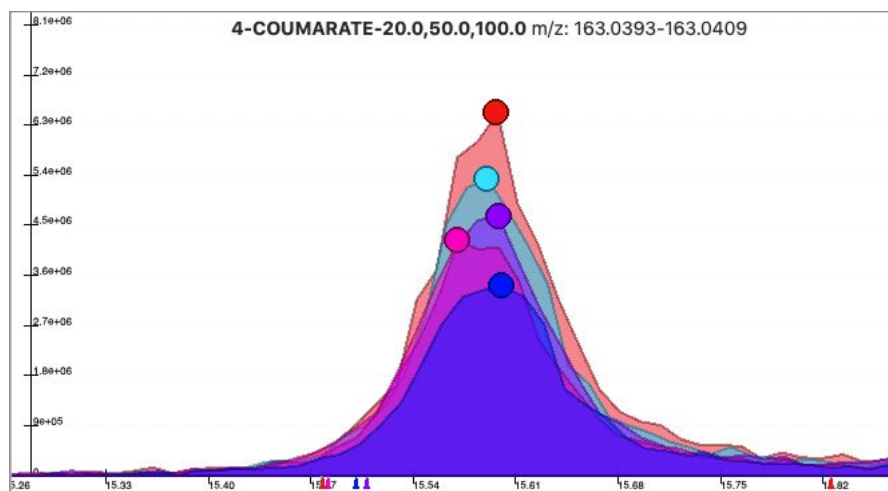


Fig 5. EIC with fragmentation markers on the x-axis

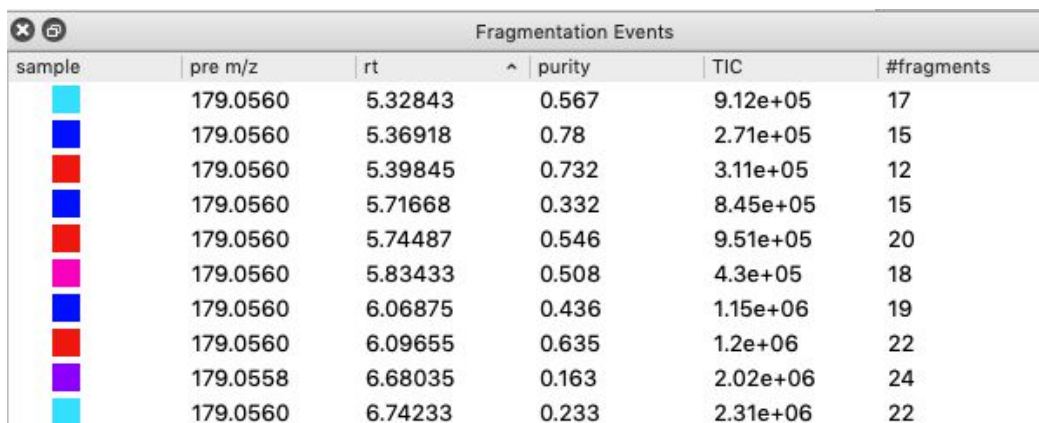
Fig. 5 depicts an EIC with the fragmentation markers displayed on the x-axis to denote the RT at which a particular MS2 scan was recorded. Since the group spectra is created by averaging all spectra that fall under this peak group, the position of the fragmentation markers can also help with identity verification when spectra match is unclear.

For the cleanest spectra, the fragmentation markers should be closer to the highest intensity of the peak. This ensures that the fragments were created in abundance and are expected to have high purity. Clicking on a particular marker opens the Fragmentation spectra widget where the individual MS2 spectra is displayed against the library spectra. This is useful if the peak group covers a large RT range where two different groups of markers exist, lowering the match score for this metabolite.

The markers can also provide important information about any technical issues with the DDA run.

### 3.4.3. Fragmentation event list

Apart from the markers, another widget was added to track all fragmentation events. Fig. 6 shows the information displayed in the Fragmentation event list for the selected peak group.



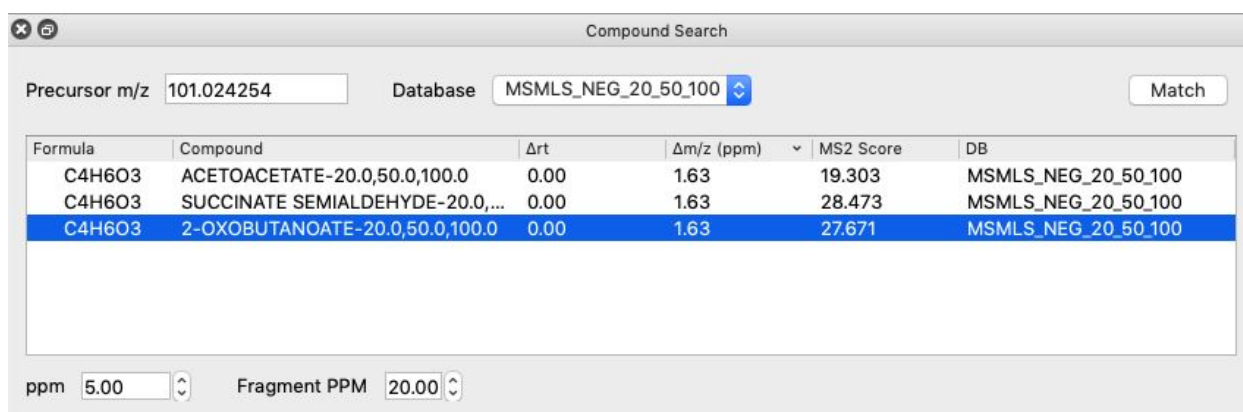
sample	pre m/z	rt	purity	TIC	#fragments
	179.0560	5.32843	0.567	9.12e+05	17
	179.0560	5.36918	0.78	2.71e+05	15
	179.0560	5.39845	0.732	3.11e+05	12
	179.0560	5.71668	0.332	8.45e+05	15
	179.0560	5.74487	0.546	9.51e+05	20
	179.0560	5.83433	0.508	4.3e+05	18
	179.0560	6.06875	0.436	1.15e+06	19
	179.0560	6.09655	0.635	1.2e+06	22
	179.0558	6.68035	0.163	2.02e+06	24
	179.0560	6.74233	0.233	2.31e+06	22

Fig 6. Fragmentation Events widget captures information for every MS2 event in a peak group

The widget provides a birds eye view of the complete list of fragmentation events within a peak group, their individual purity and number of fragments. This is a great tool to get a summary of the data before performing automated peak detection.

### 3.4.4 Match Compounds widget

While the previously mentioned widgets and tools help identify which candidate peak group is the correct match for a metabolite, there is a possibility of assigning the same peak group to multiple metabolites which could lead to errors downstream.



Formula	Compound	Δrt	Δm/z (ppm)	MS2 Score	DB
C4H6O3	ACETOACETATE-20.0,50.0,100.0	0.00	1.63	19.303	MSMLS_NEG_20_50_100
C4H6O3	SUCCINATE SEMIALDEHYDE-20.0,...	0.00	1.63	28.473	MSMLS_NEG_20_50_100
C4H6O3	2-OXOBUTANOATE-20.0,50.0,100.0	0.00	1.63	27.671	MSMLS_NEG_20_50_100

Fig 7. Match compound widget for acetoacetate within the selected database. MS2 score and RT deviation are displayed for ease of identification

The Match Compounds widget was introduced to resolve such issues. Fig 7 shows the properties tracked by the widget. Selecting a peak group from a list

would update the widget with other possible annotations of this signal. Information like the deviation from expected RT for each metabolite, as well as the MS2 score is crucial in curation of data. While the parent ion for both metabolites could be the same, the fragmentation patterns would differ and this information can be used for correct identification.

### 3.5 Targeted DDA workflow in El-MAVEN

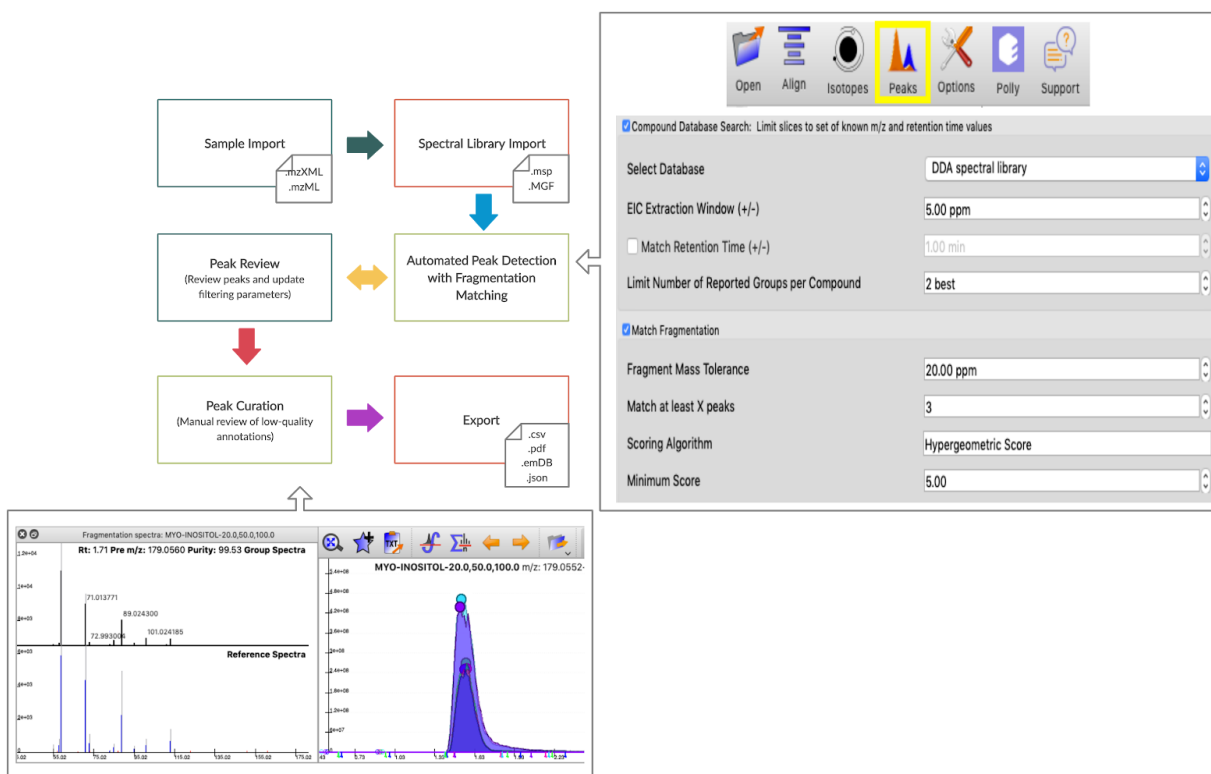


Fig 8. Summary of the complete targeted workflow for DDA datasets in El-MAVEN

El-MAVEN v0.7.0 and above have the complete targeted workflow for DDA data in place for public use. It is a 6 step process as summarized in Fig 8. Some of the important steps are:

- **Automated peak detection:** If a spectral library is provided, El-MAVEN can generate a list of detected peaks within a few seconds. The parameters for detection are usually kept liberal so as to review the peaks. Users can set whether they want to perform spectral matching or not, how many fragments should match between the two spectra, what is the mass resolution for MS2 scans and the threshold for the match



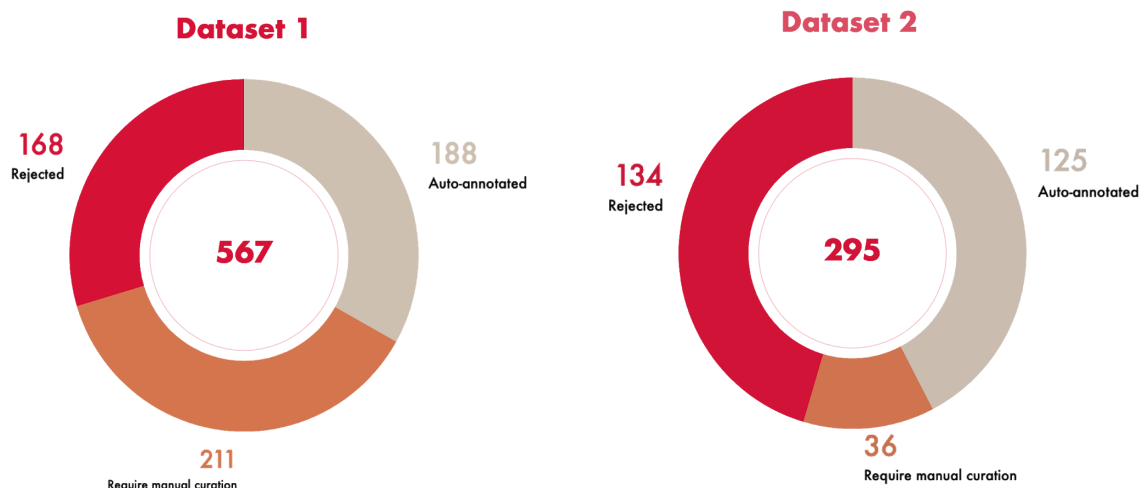
score. If a detected peak group qualifies all these parameters, it is added to a Peak Table for review.

- **Peak review:** Once the Peak Table has been generated with liberal parameters, users can go through the peak groups to observe general trends for the dataset for eg. the range of MS2 match scores for perfect spectra match, the range of intensities for different mass ranges etc. A second round of detection is usually performed after updating the appropriate parameters.
- **Peak curation:** Once the Peak Table has been generated using the desired thresholds, it is recommended that users go through the peaks to resolve any dispute with the software's annotations. Problems like multiple groups being assigned to the same metabolite and vice-versa are resolved at this stage using the Match Compound Widget and the Fragmentation spectra widget. Peak groups where the identification is deemed incorrect by the user can be deleted from the table, or marked as bad.

The complete workflow can take from 10 minutes to an hour depending on the size of the dataset and the number of metabolites in the library. The number of peak groups assigned to a single metabolite goes down by a significant number once the spectral matching is taken into account, hence reducing the amount of time spent on manual curation of the data as well as the error rate in annotation as compared to the general LCMS workflow.

### 3.6 Results

The objective of adding spectra matching capabilities to El-MAVEN was to narrow down the number of annotations per feature while improving the chances for identification of metabolites. In order to evaluate the functionality, 2 DDA datasets were processed in El-MAVEN v0.10.0, once using the general LCMS workflow at default parameters and once with the spectral matching option at default parameters and a match score cut-off of 20. The cut-off was decided on the basis of manual review of 5 spectra matches from each dataset.



**Fig 9. Distribution of auto-detected peak groups after spectra matching**

Fig 9 shows the number of peak groups detected using the LCMS workflow and the distribution after spectra matching was used. About a third are rejected once the fragmentation information is used to identify metabolites. Often, multiple peak groups are mapped to the same metabolite. Since a metabolite can only have one valid peak group, if it is present in a dataset, manual review of the candidate groups is required before exporting the data for downstream analysis. Selecting the candidate with the highest match score is another alternative, further easing the curation process.

The default parameters for the DDA workflow include a minimum threshold for the number of fragment matches as 3, hence, any metabolite with less than 3 fragments in the spectra is also excluded. Additionally, since the match score is directly proportional to the number of matching fragments between the two spectra, the score tends to be low for spectra with fewer fragments. Since the null hypothesis states that the two spectra match purely by chance, fewer fragments inspire less confidence in the validity of the match and hence the group is rejected. For the above mentioned datasets, rejected metabolites were reviewed and it was found that only about 5-6% metabolites were rejected because of this. Since the app is equipped with features to highlight and add missing metabolites, this has been accepted as a fair trade-off for partial automation of the workflow.

Another interesting observation was that spectra with more than 8-10 fragments had a high score ( $>80$ ) even when more than 50% fragments did not match between the two spectra. There is a clear scope for improvement in

match score calculation. Adjustments can be done by penalizing missing fragments while allowing for additional fragments in the observed spectrum in case of low purity spectra.

While the DDA workflow has its limitations as described above, it allows for a lot more automation compared to basic LCMS with little trade-off in terms of accuracy. The widgets added as part of the workflow are crucial for manual review of annotations and provide actionable information about the annotations for the user to make an informed choice.

We have discussed the feature additions done in El-MAVEN to support the targeted DDA workflow, where the user has a list of all metabolites that are suspected to be in the samples and how to identify those metabolites. Chapter 4 will discuss the method used for detecting signals without the help of a metabolite list which is popular for exploratory experiments. We will also discuss the improvements made in the untargeted algorithm as part of this thesis.

## 4. Mass slicing for untargeted detection

---

Untargeted metabolomics is the method of comparing the metabolome of two cohorts to find significant differences in their metabolic profile to understand the differences in biological conditions [16]. The untargeted method seeks to detect all metabolites within a certain mass range as opposed to the targeted method discussed in previous chapters where only a specific list of known metabolites is queried. This method is especially useful for discovery of biomarkers of disease conditions.

### 4.1 Challenges in untargeted detection

Since peak detection largely depends on the mass ranges provided for extracting out the chromatograms from the three dimensional raw data, untargeted peak detection presents a unique challenge since there is no list of masses to query against unlike targeted methodology. This leads to detection of a large number of features since in-source fragments, naturally abundant isotopes and adducts are also picked up. Additionally, untargeted datasets tend to be heavier than targeted LCMS data and need high computation power to process the data in a reasonable time.

A number of algorithms have been developed to solve this challenge to variable degrees. The centWave algorithm used by XCMS and mzMine2 is one of the more popular methods for detection of peaks in untargeted data but is prone to a high number of false positives [17]. MAVEN has a mass slicing algorithm based on dynamic binning of masses followed by peak detection, but the volume of peaks detected has been a major deterrent in popularization of the software for untargeted processing. El-MAVEN has retained the mass slicing algorithm from its predecessor, MAVEN. However, efforts have been made to resolve the issues present in the original algorithm and have been outlined in this chapter.

In the following sections, we will discuss MAVEN's mass slicing algorithm followed by the improvements made in El-MAVEN and how the results compare with the earlier version.

## 4.2 Mass slicing algorithm in MAVEN

Mass slicing is the process of finding regions of interest (called slices) in the  $m/z$ -rt space across samples. Instead of using a fixed bin size on the  $m/z$  axis, the mass slicing algorithm in MAVEN starts with as many slices as the number of observations in the dataset and then goes on to merge them based on their overlap in the  $m/z$ -rt space. This is supposed to ensure that slices are only created around detected signals and to prevent fragmentation of a peak into different bins. The size of a slice is determined by the mass resolution of the instrument, measured in parts per million and the expected peak width based on the runtime and scan rate of the mass spectrometer run.

Fig 10. details the original mass slicing algorithm used before peak detection. The input to the algorithm is the mass resolution and expected peak width for the experiment. The output of the mass slicing algorithm is a list of slices that will be used by the peak detection algorithm to extract peaks across samples. A slice is defined as the  $m/z$  and RT bounds that represents a region of interest.

In short, MAVEN iterates over all observations across samples and either creates a new slice for that observation or merges the slice to an existing one that matches it most closely. The decision to merge slices is based on the shortest euclidean distance between the centres of two slices from a list of slices of similar  $m/z$  ranges.

Following are some of the major issues observed in the results

- Inconsistency in results across sessions
- Duplicate and/or highly overlapping features
- Bugs/inconsistencies in peak visualisation
- Higher number of low quality peaks compared to XCMS centWave
- No annotation of known metabolites

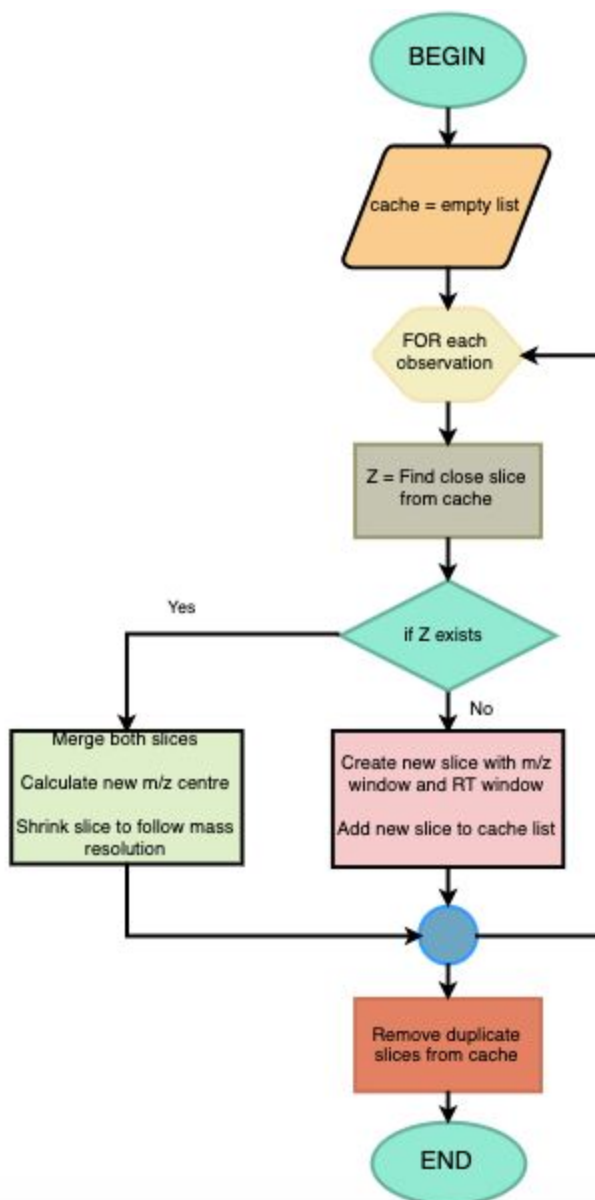


Fig 10. Original mass slicing algorithm in MAVEN

### 4.3 Understanding the challenges in the original algorithm

The mass slicing algorithm has recently been updated to resolve the issues in the original algorithm mentioned above. We will discuss some of these issues, their root cause and the resolution implemented in the new algorithm.

#### 4.3.1 Inconsistency of results across sessions

The first and most important requirement was to fix the inconsistency of results across sessions. It was observed that running the algorithm multiple times within the same session gave identical results but running it in a new session resulted in minor differences in the number of features detected as well as the mean m/z for some features, while all parameters were kept consistent.

#	ID	Observed m/z	Expected m/z	rt
9580	259.015381@10.19	259.0154	NA	10.19
9579	75.113518@7.29	75.1135	NA	7.29
9578	701.627625@3.33	701.6276	NA	3.33
9577	213.026901@5.68	213.0269	NA	5.68
9576	405.098877@9.30	405.0989	NA	9.30
9575	292.907593@18.55	292.9076	NA	18.55
9574	712.550293@3.15	712.5503	NA	3.15
9573	188.175537@9.39	188.1755	NA	9.39

Group Validation Status: Good=0 Bad=0 Total=9580

#	ID	Observed m/z	Expected m/z	rt
9619	259.015381@10.19	259.0154	NA	10.19
9618	75.113518@7.29	75.1135	NA	7.29
9617	701.627625@3.33	701.6276	NA	3.33
9616	213.026901@5.68	213.0269	NA	5.68
9615	405.098877@9.30	405.0989	NA	9.30
9614	712.550293@3.15	712.5503	NA	3.15
9613	188.175537@9.39	188.1755	NA	9.39
9612	718.054871@8.50	718.0549	NA	8.50

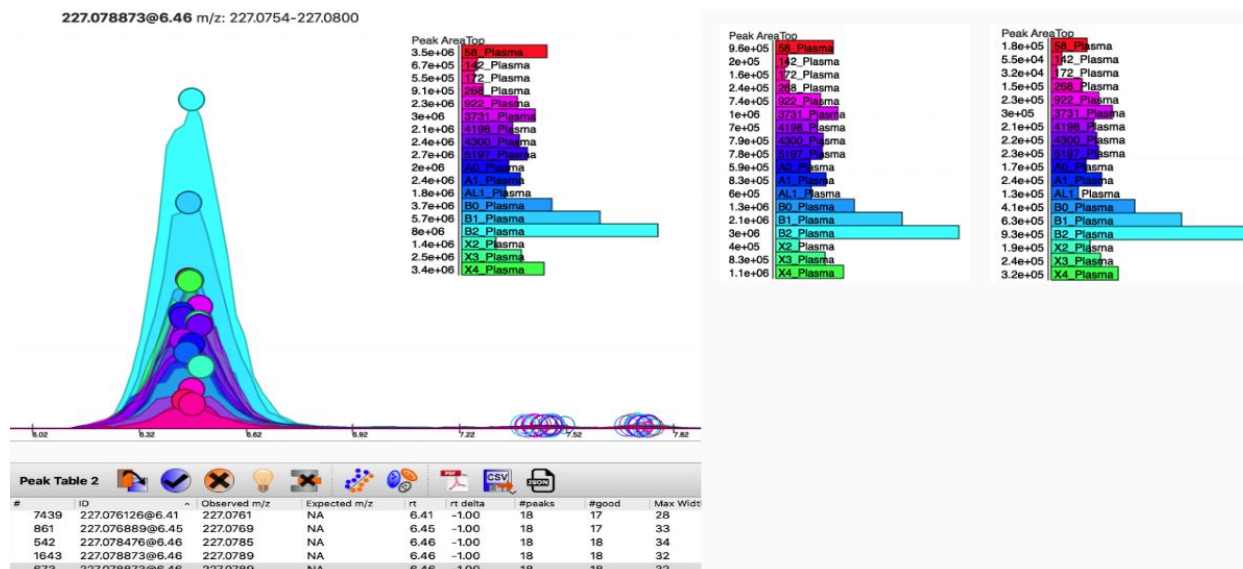
Group Validation Status: Good=0 Bad=0 Total=9619

Fig 11. Results with 9580 and 9619 features in 2 different sessions

After debugging, the root cause was found to be the algorithm's dependence on the order of samples during file import. Since the creation of any new slice depends on its mergeability with existing slices, the order in which the algorithm iterates over samples creates a bias towards the first sample. In order to remove this bias, the latest version of the algorithm creates slices for every observation across samples and sorts it by the m/z before the merging step. This ensures that the results are consistent across systems and sessions as long as the data and the parameters stay constant.

#### 4.3.2 Duplicate features

As many as ~10 duplicates or highly overlapping peak groups could be found for 50% of reported features at default settings. This led to an overestimation of features detected for every dataset while simultaneously making curation very cumbersome for users. These duplicate groups could also have different intensities for some samples, leading to inaccurate curation.



**Fig 12. Multiple peak groups detected at the same m/z and RT values and their intensity distribution across samples**

This could be explained by a bug in the merging step where instead of merging a new slice with its closest existing slice, El-MAVEN was picking the slice with the narrowest m/z range and merging it with that. Combine that with twice the time resolution used for creating slices, the algorithm ended up with highly overlapping slices, leading to multiple occurrences of the same feature.

In the new version, merging of slices is done twice. Once on the basis of m/z-RT overlap between slices and once on the basis of peak overlap. This will be explained further in the coming section.

### 4.3.3 Incorrect mass resolution

UI inconsistencies were often observed for untargeted peak tables where the EIC would be missing, or the peak tops would be displaced. Increasing the mass resolution window on the main UI would usually correct the error.



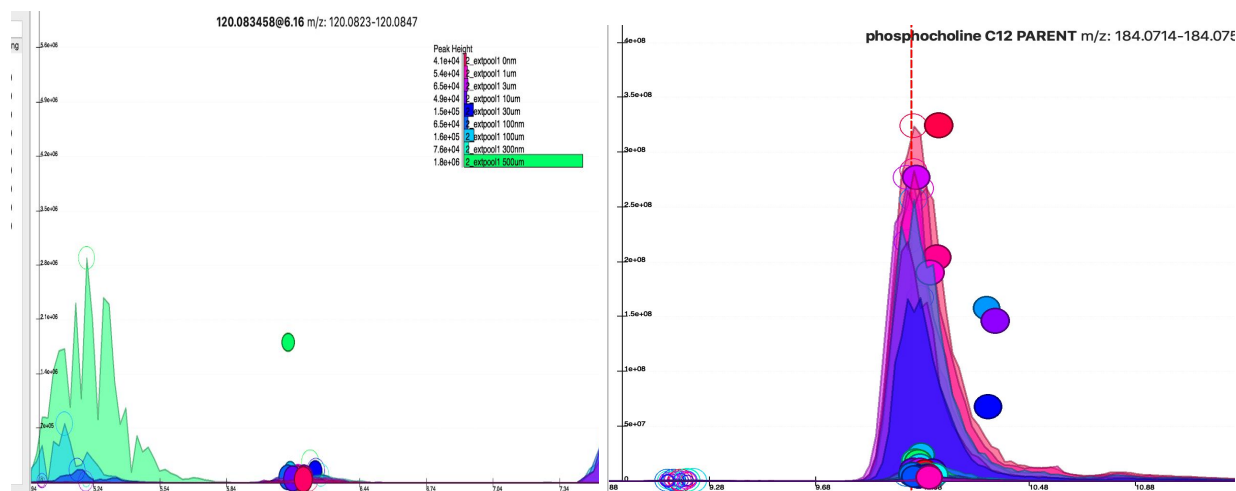


Fig 13. a) Peak bubble without EIC b) Peak bubbles displaced on the EIC

In the mass slicing algorithm, the slices are adjusted after every merge step to maintain the m/z window according to the resolution set by the user. However, the adjustment is done in a data-independent manner, on the basis of the m/z centre of the slice. During visualization, El-MAVEN uses the same mass resolution window around the group m/z to display the EIC. Since the group m/z is based on the high intensity m/zs across samples, and not the slice centre, the EIC on display tends to be somewhat different from the real EIC used for peak detection.

In the latest version, multiple changes have been done to prevent this incongruity. The m/z window for merged slices is adjusted on the basis of the highest intensity and the final slice bounds are retained after peak detection. Instead of recalculating the m/z window from the group m/z, the EIC is extracted using the original bounds so as to prevent any confusion.

#### 4.4 Mass slicing 2.0

The original mass slicing algorithm has been modified to resolve the issues mentioned above. Fig 14 is a depiction of the improved algorithm.

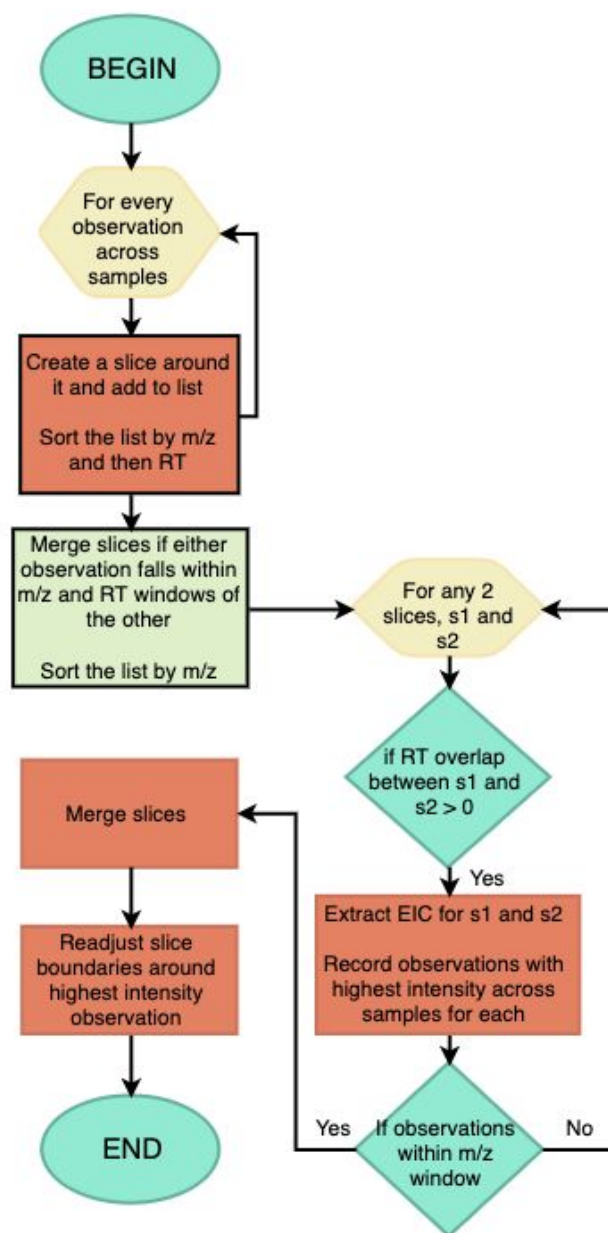


Fig 14. Data-driven mass slicing algorithm in EI-MAVEN

The mass slicing 2.0 algorithm fixes a major oversight in the original algorithm- data-blindness. The creation and merging of slices was originally based on every observation and their subsequent overlap in the  $m/z$ -RT space without considering the signal strength of merging slices.

The new algorithm starts with creating slices around every observation across samples. Once the slices are sorted by  $m/z$ , there are 3 major steps before the final list of slices required for peak detection.

- Reducing slices: Any 2 slices that have their centre lie within the bounds of the other slice, are merged together. Since the slice boundaries are defined by the mass and time resolution, and the centres are just the original observations, observations that fall within each others' boundaries are merged. This is close to the merge step in the original algorithm.
- Merge slices: In this step, a more relaxed (10x) mass resolution is used for better results. The EIC is extracted for all slices that have an RT overlap and fall within the relaxed mass resolution bounds. All pairs of slices, where the highest intensity of the EIC lies within the original mass resolution of each other, are merged. This is done to prevent the fragmentation of a real signal into multiple peaks.
- Adjust slices: Since a number of slices are merged together and expanded, the centre of the slice might no longer correspond to the peak top. The slice boundaries are recalculated around the highest intensity observation in the EIC.

The new algorithm has been successful in reducing the number of overlapping or fragmented peak groups to close to zero with some increase in the runtime of the algorithm. The tradeoff is acceptable since accuracy of data is more important than the speed, especially since the effective speed reduction is less than 2x.

#### **4.5 Additional improvements in untargeted pipeline**

Another challenge with the current untargeted pipeline within El-MAVEN was the lack of annotation. While an untargeted run is usually done for exploratory analysis, especially where the user seeks to find novel compounds that are relevant to their study, there are still a list of known metabolites that are present in the samples. Since the output of the untargeted peak detection in El-MAVEN is represented in  $m/z@RT$  format, this mapping of peak groups to known metabolites was happening outside of El-MAVEN, using downstream scripts.

To ease the process of annotation, a new option has been added to the untargeted detection pipeline, which allows the user to select a compound

database or spectral library for annotation of features. Users can decide whether they want to perform spectral matching and isotope detection or not. The spectral matching option works the same way for targeted and untargeted groups. If the same feature matches 2 or more metabolites in the spectral library, the feature is duplicated and shown as an option for both metabolites.

**Fig 15. Peak detection dialog with the options for untargeted detection with annotation and spectral matching turned on**

The untargeted pipeline in El-MAVEN also allows users to filter out only fragmented peaks, i.e. the high intensity peak groups that have MS2 information.

Using the annotation option creates a Peak Table with a list of annotated and unannotated peak groups. Since this excludes duplicates to a great extent and also reduces the “unknown” peak groups, the manual curation effort is made easier. Combine that with statistical approaches to find highly varying features between cohorts, the curation pool becomes even smaller and more practical.

## 4.6 Untargeted DDA workflow in El-MAVEN

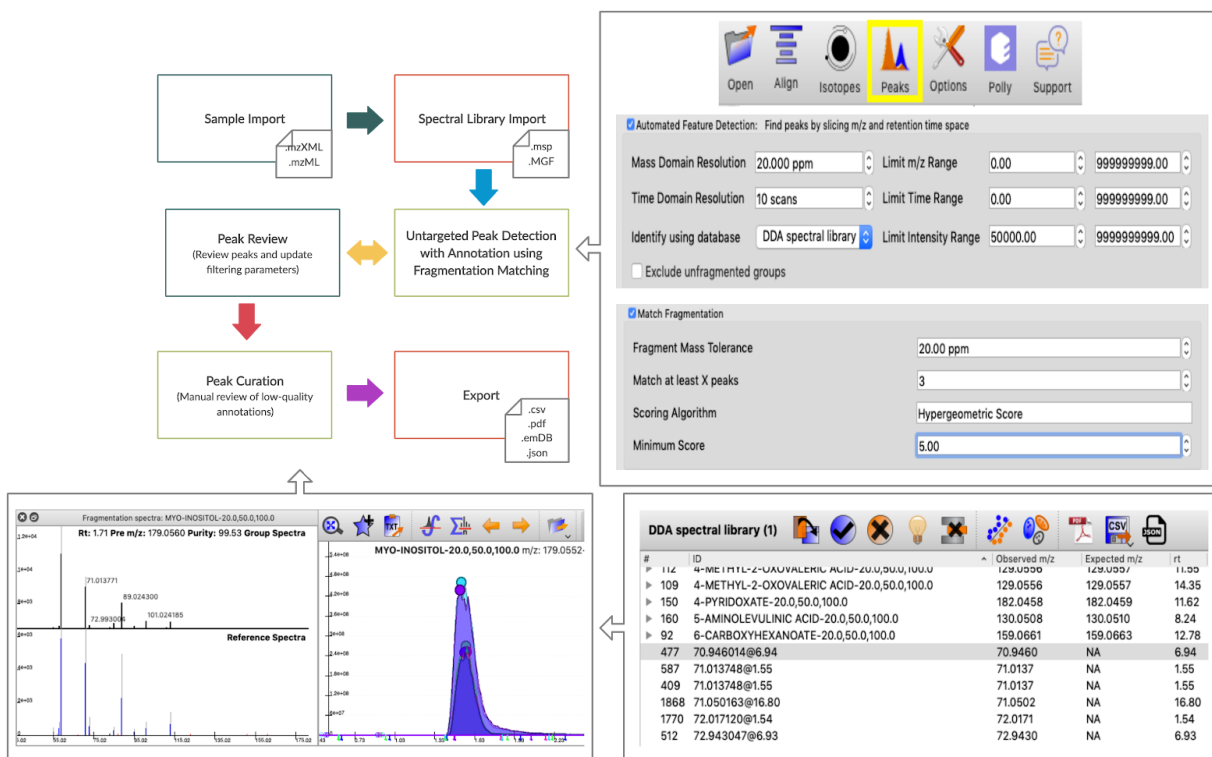


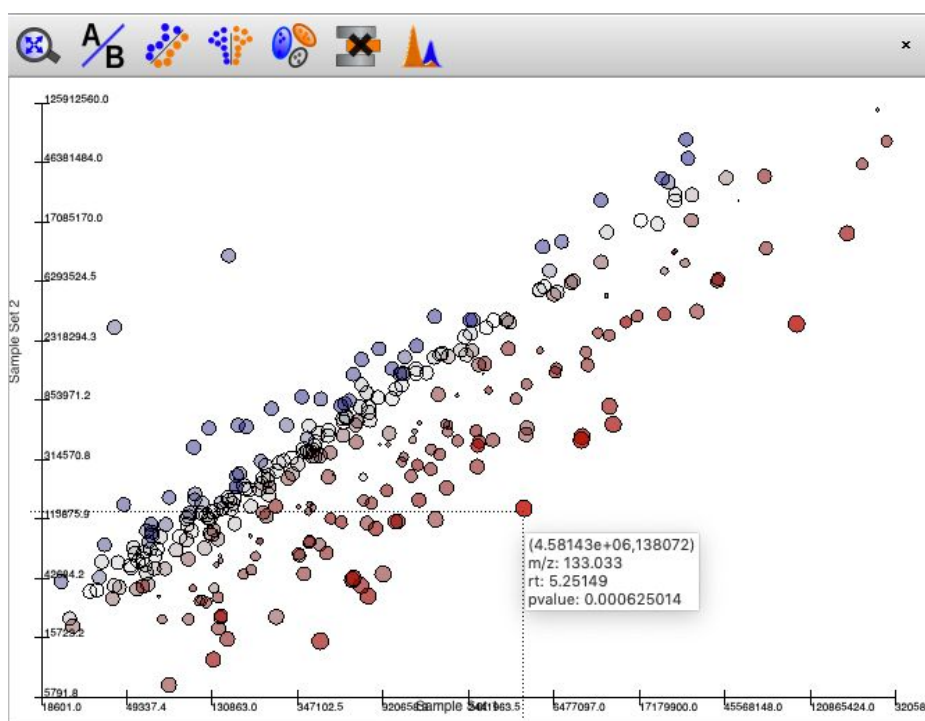
Fig 16. Automated untargeted workflow for DDA data in EI-MAVEN

The major steps for the untargeted DDA workflow are depicted in Fig 16. EI-MAVEN v1.0 and above will have the complete targeted workflow for DDA data in place for public use. Some of the important steps are:

- **Untargeted peak detection:** This option uses the mass slicing algorithm to find all relevant signals in the data, irrespective of a compound list. The input is the mass and time resolution expected for the dataset, usually defined by the instrument and the experimental setup. Some filters for m/z, time and intensity range are also provided for more narrowed search. If the user has selected a spectral library and selected the Match Fragmentation option, the detected features are queried against the selected library just like the targeted workflow. The features that don't find a match in the selected library are reported in the m/z@RT format in the Peak Table.
- **Peak review:** Once the Peak Table has been generated with liberal parameters, users can go through the peak groups to observe general trends for the dataset for eg. the range of MS2 match scores for perfect spectra match, the range of intensities for different mass ranges etc. A second round of detection is usually performed after updating the appropriate parameters.

- **Peak curation:** Once the Peak Table has been generated using the desired thresholds, it is recommended that users go through the peaks to resolve any dispute with the software's annotations using the Match Compound and Fragmentation Spectra widgets. Peak groups where the identification is deemed incorrect by the user can be deleted from the table, or marked as bad. For unidentified peak groups, statistical measures are used to identify features that vary significantly between two cohorts. The selected peak groups can then be exported in various formats for downstream processing.

Basic statistical analysis is supported in El-MAVEN through the Statistics widget in the Peak Table toolbar. If sample cohorts are defined in the sample widget, users can select a pair of cohorts in the Statistics widget to visualize their data and find any and all peak groups that show high variation between two cohorts. These groups can be examined further to identify unknown compounds in the sample set.



**Fig 17. Scatter plot of peak groups. The axes represent the average cohort intensity for a group. The opaqueness and size of bubbles depict the significance and fold change respectively**

The complete workflow can take from 10 minutes to an hour depending on the size of the dataset and the number of metabolites in the library. The number of

peak groups assigned to a single metabolite goes down by a significant number once the spectral matching is taken into account, hence reducing the amount of time spent on manual curation of the data as well as the error rate in annotation as compared to the general LCMS workflow.

## 4.7 Results

The improved slicing algorithm combined with the annotation and mass tolerance lock should contribute to higher accuracy in picking all relevant metabolic species in the dataset and possibly identifying the known compounds.

In order to evaluate the improvements in the slicing method, the untargeted workflow was run in EI-MAVEN v0.9.0 and v0.10.0 on a curated LCMS dataset with known retention times. The results were then compared on a set of metrics to measure the improvements in duplicate reduction and better slicing. The parameters for the runs are summarized in Table 1.

<b>Mass resolution</b>	<b>Time resolution</b>	<b>Min intensity for mass slices</b>	<b>Min intensity for groups</b>	<b>Min quality for groups</b>
10 ppm	10 scans	10000	5000 (10% peaks)	0.5 (33% peaks)

Table 1. EI-MAVEN parameters for the untargeted runs. Default values were used for all other parameters

The results of the comparison in Table 2 were found to be consistent with the objective. The new version is able to reduce the number of duplicate or highly overlapping features from 56% of all detected features to 18% of all detected features, sharply bringing down the total number of detected features to a third of its size.

The time complexity of the algorithm has increased for the new algorithm due to addition of multiple merge steps to reduce overlaps. Multi-processing was introduced at appropriate points in the algorithm to reduce the time taken for the run.

The number of annotations remain the same in both runs. On manual

inspection of the remaining 16 targets, it was found that the features were low-intensity and could not pass the user set threshold and were rejected for the same.

	<b>El-MAVEN v0.9.1</b>	<b>El-MAVEN v0.10.0</b>
<b>Number of features</b>	29922	9968
<b>Number of duplicates</b>	16860 (56%)	1800 (18%)
<b>Number of standard metabolites detected (79)</b>	63	63
<b>Redundancy among standards</b>	228 extra groups	15 extra groups
<b>Time taken to process</b>	~10 min	~6 min

**Table 2. Performance comparison of the untargeted workflow in El-MAVEN v0.9.1 and v0.10.0**



## 5. Conclusion

---

Data dependent acquisition is becoming a popular method for tandem mass spectrometry. The improvements made in El-MAVEN to support this method, in a targeted and untargeted way, have been validated using standard data. The targeted pipeline has been shown to reduce mis-annotations by a third for 2 datasets. It also provides more information for curation of data through the different widgets added as part of the pipeline.

The untargeted algorithm was practically unusable with tens of thousands of features detected for every dataset. While there is still scope for improvement, the workflow is now released in beta phase and provides a reasonable number of detected features that require curation. Combined with auto-annotation using spectral matching, the workflow can add a lot of value to an analyst's life.

Comparison of the untargeted algorithm with XCMS has been done in the past where El-MAVEN was able to detect features that XCMS did not report. Our hypothesis is that El-MAVEN is able to detect all features in a given dataset and reports every feature that qualifies user set thresholds. The major drawback was that the data output was way too large and ambiguous due to the presence of overlapping features that should have been one whole feature. Since the new algorithm resolves those issues to a large extent, we expect similar or superior performance in El-MAVEN compared to other alternatives. Further validations will be performed to verify this hypothesis.

## 6. References

---

1. Clish, Clary B. "Metabolomics: an emerging but powerful tool for precision medicine." *Cold Spring Harbor molecular case studies* vol. 1,1 (2015): a000588. doi:10.1101/mcs.a000588
2. "FDA Grants Approval of TIBSOVO® , the First Oral, Targeted Therapy for Adult Patients with Relapsed/Refractory Acute Myeloid Leukemia and an IDH1 Mutation" Press Release, Agios Pharmaceuticals, 20 July 2018, <https://investor.agios.com/news-releases/news-release-details/fda-grants-approval-tibsovor-first-oral-targeted-therapy-adult>.
3. "FDA Grants Approval of IDHIFA® , the First Oral Targeted Therapy for Adult Patients with Relapsed/Refractory Acute Myeloid Leukemia and an IDH2 Mutation." Press Release, Agios Pharmaceuticals, 1 Aug. 2017, <https://www.businesswire.com/news/home/20170801006281/en/>.
4. Roessner, Ute, and Jairus Bowne. "What Is Metabolomics All about?" *BioTechniques*, vol. 46, no. 5, Apr. 2009, pp. 363–365., doi:10.2144/000113133.
5. Matsuda, Fumio. "Technical Challenges in Mass Spectrometry-Based Metabolomics." *Mass Spectrometry*, vol. 5, no. 2, 23 Nov. 2016, doi:10.5702/massspectrometry.s0052.
6. Clasquin, Michelle F., et al. "LC-MS Data Processing with MAVEN: A Metabolomic Analysis and Visualization Engine." *Current Protocols in Bioinformatics*, 1 Mar. 2012, doi:10.1002/0471250953.bi1411s37.
7. Tautenhahn, Ralf et al. "XCMS Online: a web-based platform to process untargeted metabolomic data." *Analytical chemistry* vol. 84,11 (2012): 5035–9. doi:10.1021/ac300698c
8. Pluskal, Tomáš, et al. "MZmine 2: Modular Framework for Processing, Visualizing, and Analyzing Mass Spectrometry-Based Molecular Profile Data." *BMC Bioinformatics*, vol. 11, no. 1, 23 July 2010, doi:10.1186/1471-2105-11-395.
9. Coble, Jamie B., and Carlos G. Fraga. "Comparative Evaluation of Preprocessing Freeware on Chromatography/Mass Spectrometry Data for Signature Discovery." *Journal of Chromatography A*, vol. 1358, 2014, pp. 155–164., doi:10.1016/j.chroma.2014.06.100.
10. Myers, Owen D., et al. "Detailed Investigation and Comparison of the XCMS and MZmine 2 Chromatogram Construction and Chromatographic

- Peak Detection Methods for Preprocessing Mass Spectrometry Metabolomics Data.” *Analytical Chemistry*, vol. 89, no. 17, 2017, pp. 8689–8695., doi:10.1021/acs.analchem.7b01069.
11. Myers, Owen D., et al. “One Step Forward for Reducing False Positive and False Negative Compound Identifications from Mass Spectrometry Metabolomics Data: New Algorithms for Constructing Extracted Ion Chromatograms and Detecting Chromatographic Peaks.” *Analytical Chemistry*, vol. 89, no. 17, 2017, pp. 8696–8703., doi:10.1021/acs.analchem.7b00947.
  12. Agrawal, Shubhra, et al. “El-MAVEN: A Fast, Robust, and User-Friendly Mass Spectrometry Data Processing Engine for Metabolomics.” *High-Throughput Metabolomics Methods in Molecular Biology*, 2019, pp. 301–321., doi:10.1007/978-1-4939-9236-2\_19.
  13. Mittal, Rama Devi. “Tandem mass spectroscopy in diagnosis and clinical research.” *Indian journal of clinical biochemistry : IJCB* vol. 30,2 (2015): 121-3. doi:10.1007/s12291-015-0498-9
  14. Horai, Hisayuki, et al. “MassBank: a Public Repository for Sharing Mass Spectral Data for Life Sciences.” *Journal of Mass Spectrometry*, vol. 45, no. 7, 2010, pp. 703–714., doi:10.1002/jms.1777.
  15. Sadygov, Rovshan G., and John R. Yates. “A Hypergeometric Probability Model for Protein Identification and Validation Using Tandem Mass Spectral Data and Protein Sequence Databases.” *Analytical Chemistry*, vol. 75, no. 15, 2003, pp. 3792–3798., doi:10.1021/ac034157w.
  16. Alonso, Arnald, et al. “Analytical Methods in Untargeted Metabolomics: State of the Art in 2015.” *Frontiers in Bioengineering and Biotechnology*, vol. 3, 2015, doi:10.3389/fbioe.2015.00023.
  17. Li, Zhucui, et al. “Comprehensive Evaluation of Untargeted Metabolomics Data Processing Software in Feature Detection, Quantification and Discriminating Marker Selection.” *Analytica Chimica Acta*, vol. 1029, 9 Feb. 2018, pp. 50–57., doi:10.1016/j.aca.2018.05.001.

## Supplementary Data

---

### Targeted DDA validation table

	Dataset 1	Dataset 2
Number of peak groups in LCMS run	567 135 have multiple groups	295 51 have multiple groups
Number of metabolites with one group	121	180
Number of metabolites with more than 1 group	135	51
<b>After Spectral Matching</b>		
Number of groups rejected by fragmentation matching	168 (-42.1%)	134 (-46%)
Number of metabolites with more than 1 group	74	17
Possible false negatives	14	15