Advanced Text-Mining Algorithms enhance Substructure Exploration in Untargeted Metabolomics Data

**Justin J.J. van der Hooft1\*, Joe Wandy2, Michael P. Barrett1, Karl E.V. Burgess1, Simon Rogers1,2\***

1 Glasgow Polyomics, University of Glasgow, Glasgow, United Kingdom

2 School of Computing Science, University of Glasgow, Glasgow, United Kingdom

**\*Correspondence:** Justin J.J. van der Hooft, Glasgow Polyomics, Level 2, Room 235, Wolfson Wohl Cancer Research Centre, Garscube Estate, Glasgow, G61 1HQ, United Kingdom.

justin.vanderhooft@glasgow.ac.uk

**\*Correspondence:** Simon Rogers, Room S123, School of Computing Science, Sir Alwyn Williams Building, Glasgow, G12 8QQ, United Kingdom.

Simon.Rogers@glasgow.ac.uk

Keywords: data mining, metabolomics, mass spectrometry, high-resolution, fragmentation, Latent Dirichlet Allocation (LDA), visualization, software, substructures, text-mining, pattern recognition.

# Abstract

The potential of untargeted metabolomics to answer important questions across the life sciences is hindered due to a paucity of computational tools that enable extraction of key biochemically relevant information. Available tools focus on using mass spectrometry fragmentation spectra to identify molecules whose behavior suggests they are relevant to the system under study. Unfortunately, fragmentation spectra cannot identify molecules in isolation, but require authentic standards or databases of known metabolite fragments. Fragmentation spectra are, however, replete with information pertaining to the biochemical processes present; much of which is currently neglected. Here we present an analytical workflow that exploits all fragmentation data from a given experiment to extract biochemically-relevant features in an un-supervised manner. We demonstrate that an algorithm originally utilized for text-mining, Latent Dirichlet Allocation, can be adapted to handle metabolomics datasets. Our approach extracts biochemically-relevant molecular substructures (‘Mass2Motifs’) from spectra as sets of co-occurring molecular fragments and neutral losses. The analysis allows us to identify molecular substructures, whose presence allows molecules to be grouped based on shared substructures regardless of classical spectral similarity. These substructures in turn support putative *de novo* structural annotation of molecules. Combining this spectral connectivity to orthogonal correlations (e.g. common abundance changes under system perturbation) significantly enhances our ability to provide mechanistic explanations for biological behavior.

# 1.0 Introduction

Mass Spectrometry (MS) based metabolomics aims to capture the entire small molecule composition of biological systems. Analysis of MS metabolomics data is challenging as many molecules cannot be identified from their mass alone (e.g. isobaric molecules, and isomers) (1-3). Separation by liquid chromatography prior to MS (LC-MS) can add discriminatory information but does not solve the problem as isomers can exhibit similar chromatographic behavior, and chromatographic retention time is unpredictable.

Fragmentation spectra incorporate considerable information regarding structural identification and have been used to partially overcome this problem (4-6). Most fragmentation tools compare individual fragmentation spectra to reference spectra (5, 7) stored in public databases, e.g. MassBank (8) or HMDB (9), and are thus constrained by the limited number of reference spectra (10-12). Poor identification coverage then results in poor biochemical insight. We propose a method that analyses all acquired fragmentation spectra to expose the biochemistry without relying on metabolite identification. The method is inspired by machine learning techniques developed initially for text processing (13).

The paucity of techniques that share information across fragmentation spectra can be explained by the complexity of fragmentation data (14). One example, “Molecular Networking”, clusters MS1 peaks by their MS2 spectral similarity such that one identifiable metabolite in a cluster facilitates structural annotation of its neighbors (15, 16).  However, spectral features causing the clustering must be extracted manually and only MS2 spectra with high overall spectral similarity are grouped. Another package, MS2Analyzer (17),[1, 10, 11] mines MS2 spectra for specific features defined by the user (i.e., mass fragments and neutral losses). Some will be common to many experiments (e.g. CO or H2O losses), but sample-specific features are easily overlooked. Whilst Molecular Networking requires no user intervention it may fail to group molecules that share small substructures, whilst MS2Analyzer can find all molecules that share a particular set of features provided they are user-specified. Our approach, MS2LDA, which is based on Latent Dirichlet Allocation (LDA) (13), a probabilistic topic model, retains the benefits of both of these approaches whilst losing the shortfalls – it can find relevant substructures based on the co-occurrence of mass fragments and neutral losses, and group the molecules accordingly. LDA was previously adapted to other types of high throughput data (genomics (18), metagenomics (19), and transcriptomics (20)) but never before to exploit the parallels between MS2 fragmentation data and text.

Fragmentation spectra contain recurring patterns of fragments and losses due to the presence of common biological substructures (e.g. a hexose unit, or a CO loss). We assume each observed spectrum is comprised of one or more such substructures, an assumption successfully used in other workflows (6, 21-23); however, no un-supervised strategy exists that finds mass fragmental-based substructures without training data.

Figure 1 demonstrates the parallels between text and fragmentation data. LDA decomposes documents into topics based on co-occurring words, while MS2LDA decomposes fragmentation spectra into blocks of co-occurring fragments and losses, referred to as ‘Mass2Motifs’. Using all of the fragmentation spectra generated by data-dependent mass fragmentation analysis (DDA), MS2LDA learns the conserved substructures (the Mass2Motifs) and the decomposition of the fragmentation spectra into Mass2Motifs.

Our analysis pipeline (Figure 1-B) performs data pre-processing, extracts Mass2Motifs and allows the user to explore the results interactively. Through the analyses of four beer extracts, we show that without labelled training data or metabolite identification, MS2LDA extracts mass patterns indicative of easily identifiable biological substructures, some of which are pathway related. These can aid in the putative *de novo* annotation or functional classification of otherwise unidentifiable molecules. Many more molecules can be annotated in this way than can be identified by comparison with reference spectra. This is particularly useful for hypothesis-generating research. For example, if metabolites with similar changes in MS1 abundance share a pathway-related substructure, MS2LDA provides a hypothesis as to the source of their change in abundance.

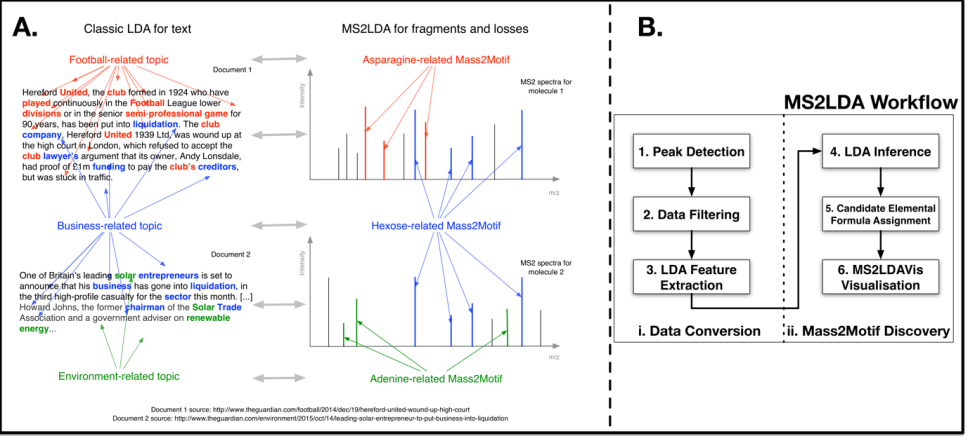


Figure 1. **A)** Analogy between LDA for text MS2LDA. LDA has extracted topics interpreted as ‘football related’, ‘business-related’ and ‘environment related’. MS2LDA extracts sets of concurring mass fragments or losses (Mass2Motifs) that can be interpreted as ‘Asparagine-related’, ‘Hexose-related’ and ‘Adenine-related’. **B)** The MS2LDA workflow.

# 2.0 Material and Methods

## 2.1 Materials

### 2.1.1 Beer Samples

Beer samples, acquired from three different commercial beers and one home-brewed beer, were used as representative complex mixtures of diverse biochemicals. *Beer1* is from a home-brewed bottle of German Wheat Beer (see Section S-1); *Beer2* is from a bottle of ‘Jaw Glyde Ale’ (a Golden/Blond Ale; http://www.jawbrew.co.uk); *Beer3* is from a bottle of ‘Seven Giraffes Extraordinary Ale’ (an IPA style beer; <http://www.williamsbrosbrew.com/beerboard/bottles/seven-giraffes>); *Beer4* is from a bottle of ‘Black Sheep Ale’ (a Golden Bitter Ale; <https://www.blacksheepbrewery.com/beers/15/black-sheep-ale>).

### 2.1.2 Beer Sample Preparation

10 ml of beer was sampled directly after opening and stored at -20 ˚C before extraction. After thawing, i) 200 µL of beer was mixed with 600 µL of methanol/chloroform, ii) sonicated for 5 minutes at room temperature; iii) and centrifuged for 5 minutes (12,000 g) at room temperature. A pooled aliquot of the four extracts was also prepared. The supernatants were stored at -80 ˚C until analysis.

### 2.1.3 Chemicals

HPLC-grade methanol, acetonitrile, and analytical reagent grade chloroform were acquired from Fisher Scientific, Loughborough, UK. HPLC grade H2O was purchased from VWR Chemicals, Fountenay-sous-Bois, France. Formic acid (for MS) and ammonium carbonate were acquired from Fluka Analytical (Sigma Aldrich), Steinheim, Germany.

## 2.2 Methods

### 2.2.1 Analytical Platform

A Thermo Scientific Ultimate 3000 RSLCnano liquid chromatography system (Thermo Scientific, CA, USA) was coupled to a Thermo Scientific Q-Exactive Orbitrap mass spectrometer equipped with a HESI II interface (Thermo Scientific, Hemel Hempstead, UK). Thermo Xcalibur Tune software (v2.5) was used for instrument control and data acquisition.

### 2.2.2 LC Settings

Column temperature was maintained at 25 °C. The HILIC separation was performed with a SeQuant ZIC-pHILIC column (150 x 4.6 mm, 5 µm) equipped with the corresponding pre-column (Merck Sequant, Darmstadt, Germany). A linear LC gradient was conducted from 80% B to 20% B over 15 min, followed by a 2 min wash with 5% B, and 7 min re-equilibration with 80% B, where solvent B is acetonitrile and solvent A is 20 mM ammonium carbonate in water. The flow rate was 300 μL/min, column temperature held at 25 °C, injection volume was 10 μL, and samples maintained at 4 °C in the autosampler [1]. Samples were measured in randomized order (24) (see section S-3).

### 2.2.3 MS and MS/MS Settings

#### Positive Negative Ionization Combined Fragmentation and Full Scan Mode:

Duty cycles consisted of a full scan in positive ionization mode, followed by a TopN data dependent MS/MS (MS2) fragmentation event taking the 10 most abundant ion species not on the dynamic exclusion list, followed by the same two scan events in negative mode. Details can be found in S2.1. MS/MS fragmentation spectra were acquired using stepped higher collision dissociation (HCD) combining 25.2, 60.0, and 94.8 normalized collision energies (NCEs) in one MS2 scan. In full scan mode, the duty cycle consisted of two full scan events.

#### Positive or Negative Ionization Separate Fragmentation modes:

See S2.2. Duty cycles consisted of one full scan (MS1) event and one Top10 MS/MS (MS2) fragmentation event.

## 2.3 MS2LDA

The complete MS2LDA workflow is illustrated in Figure 1-B and described below (for full details see section S-4). All codes and Jupyter notebooks to run the workflow are provided. Data, in the form of .mzXML (full scan) and .mzML (fragmentation) files, is pre-processed using XCMS (25) and MzMatch (26) for peak detection and RMassBank (27) for detecting MS1-MS2 pairs, before matrix formation by aligning MS2 features across different fragmentation spectra. The resulting matrix has MS2 features (fragments and losses) as rows and columns corresponding to the MS1 peaks. The values in the matrix are the MS2 feature intensities which are subsequently turned into integer counts by normalizing and rounding such that the most intense feature for each MS1 peak has value 100.

LDA is implemented in Python and uses collapsed Gibbs sampling for inference (28) (see Section S4.2), the output of which is a set of Mass2Motifs and assignments of Mass2Motifs to each MS1 peak. In addition we provide an optional elemental formula assignment step (see Section S4.3; (29-31)) to assign candidate elemental formulae to the MS2 features and MS1 peaks. On a typical laptop (Intel Core i5, 4GB RAM) running the workflow for one sample takes around 3 hours.

The LDA output can be explored interactively in the MS2LDAvis module which has been customized from LDAvis (32). We used MS2LDAvis to inspect Mass2Motifs with degree ≥10 (i.e. that were present in ten or more spectra) and structurally characterized them at varying levels of confidence (see Tables S1 and S2) through expert knowledge and matching of the Mass2Motif spectra to reference spectra in MzCloud ([www.mzcloud.org](http://www.mzcloud.org/)).

# 3.0 Results

The MS2LDA workflow was independently applied to 4 beer extracts. After pre-processing, we were left with around 1,000 MS peaks in both positive and negative ionization mode (see section S5.1). 300 Mass2Motifs were extracted for each data file and checked for biochemical relevance. 30-40 Mass2Motifs in each of the positive ionization mode files were structurally annotated (see Supporting Table S1) and diverse biochemically relevant substructures found included histidine, phenylalanine, adenine, hexose-units, and structural features such as water or carboxyl group loss.

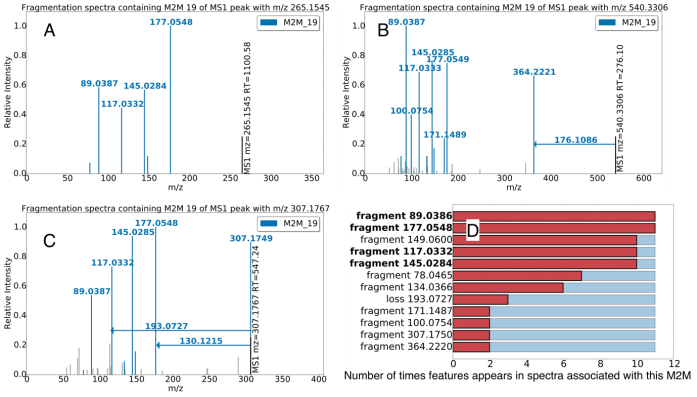
The degree of Mass2Motifs (the number of spectra in which they occurred) varied from 1 to over 200 spectra (see Section S5.2), demonstrating the ability of MS2LDA to extract both generic and specific structural features. The number of Mass2Motifs within each spectrum also varied (around 600 spectra in each file consisted of one Mass2Motif, 300 of two, 50 of three, and 20 of four or more). Across the four files, an average of 70% of spectra (Table S4) include at least one annotated Mass2Motif, demonstrating the power of MS2LDA for data reduction – i.e. annotating just 30-40 of the discovered Mass2Motifs provides biochemical insight into 70% of the spectra. For comparison, we matched the spectra to the MassBank and NIST libraries (see Section S5.5) and obtained hits for only 25% and 6% of the MS2 spectra for NIST and MassBank, respectively, at a threshold of 90% normalized scores. Although the methods are not mutually exclusive, gaining biochemically relevant insights from 70% of the spectra presents a clear advantage over spectral matching.

## 3.1 Automatic, Unsupervised, Chemical Substructure Discovery

Mass2Motifs cover a diverse set of biochemical features, including amino acid related (i.e. histidine, leucine, tryptophan, and tyrosine), nucleotide related (i.e. adenine, cytosine, and xanthine), and other molecules such as cinnamic acid, ferulic acid, ribose and N-acetylputrescine. Mass2Motifs related to the same substructure or structural feature were consistently found across multiple beers (e.g. hexose-related Mass2Motifs were present in all positive ionization mode files with degrees from 58 to >100 (see Tables S1 and S2)). Differences in degree and absence of some Mass2Motifs across the extracts show that MS2LDA also captures variability in metabolic composition of beer.

An example of ferulic acid (a compound present in cereals, an ingredient of beer) is given in Figure 2. Three of the eleven spectra that include Mass2Motif 19 are shown. Conserved mass fragments are clearly visible across the three spectra with the most conserved highlighted in Figure 2D. Unlike existing software, e.g. MS2Analyzer (17), our method is unsupervised and has no need for prior knowledge about fragments of interest. It is of note that the neutral loss of the complete ferulic acid moiety was also included by MS2LDA, demonstrating that both fragments and losses can be present in a motif. It is able to extract a biochemically relevant pattern present in just 11 of the >1000 spectra, despite the fact that the individual spectra are quite different (see Section 3.3).

Positive ionization mode fragmentation spectra generally provide larger sets of conserved fragments but some Mass2Motifs e.g. those related to phosphate and sulfate groups (fragments at 78.9593 ([PO3]-) and 79.9575 ([SO3]-) m/z, respectively) were more easily identifiable in negative mode; hence an argument to use both ionization modes. Interestingly, three positive mode Mass2Motifs pointed to the highly similar aromatic substructures of phenylethene, cinnamic acid (cinnamate), and phenylethyleneamine (i.e., [phenylalanine – CHOOH]), demonstrating separation of very similar yet functionally different substructures (see S5.4).

Figure 2. Three spectra, from the beer3 positive ionization mode file, each of which includes Mass2Motif 19, annotated as the plant derived ferulic acid substructure. A-C highlight mass fragments and neutral losses (arrows originating at the precursor ions) included in Mass2Motif 19 (fragments not explained by Mass2Motif 19 are light grey). The histogram in D shows how common each fragment / loss is in the 11 instances of motif 19 found in the dataset. The abundant fragments with an m/z of 177.0545, 145.0284, 117.0332, and 89.0386 Da are most consistently present. It is of note that the neutral loss of 176.1086 and the protonated fragment of 177.0575 both relate to the complete ferulic acid substructure.

## 3.2 Structurally Annotated Mass2Motifs Validated in Authentic Standards

Reference molecules present in the beer extracts can be identified based on chromatographic co-elution and exact mass. As their identity is known, we can validate our structurally annotated Mass2Motifs. Of the 45 reference molecules we could identify, 38 included one or more annotated Mass2Motif, despite the fact that the original Mass2Motifs annotation was made independently of reference molecule characterization. 32 contained Mass2Motifs corresponding to known biochemical features.

Figure 3 shows examples with fragmentation spectra colored by Mass2Motif. The spectra for phenylalanine (Figure 3A) and histidine (Figure 3B) share Mass2Motif 262, indicating the presence of a free (underivatized) carboxylic acid group. The loss of CHOOH (Mass2Motif 262) is in fact a common characteristic for many other underivatized amino acids and free organic acids and was associated with 10 of the 18 amino acids structures matched from the standards (the remaining 8 prefer alternative fragmentation routes – e.g. see the amine loss in tryptophan, Figure 3C). The other Mass2Motifs (115, 241) in Figures 3A and 3B are related to phenylalanine and histidine, respectively (all Figure 3 related Mass2Motifs can be found in section S5.3). Finally, Figure 3D is the MS2 spectrum of adenosine, which consists of an adenine molecule conjugated to a ribose sugar molecule. The two associated Mass2Motifs (156, 220) represent these two biochemically relevant structural features (i.e., adenine substructure and a loss corresponding to a ribose sugar).

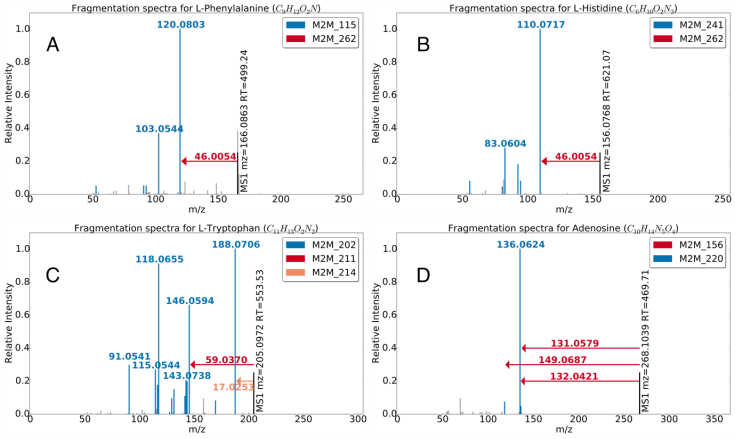
Spectra can include multiple Mass2Motifs. In each of Figures 3A to 3D, we observe two or more Mass2Motifs. We know of no other method that can do this without training spectra consisting of known structures, or *a priori* knowledge of interesting feature combinations. Multiple Mass2Motifs can also explain the same feature in one spectrum, i.e. the fragments 110.0717 (C5H8N3, [M+H]+) and 120.0803 (C8H10N, [M+H]+) in Figures 3A and 3B are explained by Mass2Motifs 241 and 115 and also by the 46.0054 loss (CHOOH) of Mass2Motif 262. This demonstrates the manner in which MS2LDA decomposes molecules into their constituent building blocks, allowing for *de novo* metabolite annotation (see Section 3.3).

Figure 3. Mass2Motif spectra of identified metabolites A) L-histidine, B) L-phenylalanine, C) L-tryptophan, and D) adenosine. Characterized motifs (see Table 1) are indicated by color.

## 3.3 Mass2Motifs Aid *de-novo* Metabolite Annotation

On average 70% of the fragmented MS1 features are explained by at least one structurally annotated Mass2Motif (see Section S5.2) and can therefore be automatically classified. For comparison, we performed spectral matching using the NIST MS/MS database for small molecules (http://chemdata.nist.gov/mass-spc/msms-search/) and MassBank (8) (see Section S5.5) on 7 of the metabolites annotated via the ferulic acid Mass2Motif. Only 1 returned a ferulic acid related hit, in spite of the clear presence of ferulic acid in all spectra (see e.g. Fig. 2). The Mass2Motif itself can be represented as a spectrum and be subjected to spectral matching, resulting in trans-ferulic acid as the best hit (hinting at the possibility of automatic Mass2Motif annotation). Spectra that are explained by the Mass2Motifs related to histidine, tyrosine, and tryptophan in Table S7 were also subjected to spectral matching. From 39 metabolites annotated by MS2LDA, 7 resulted in correct hits with another 8 producing structurally related hits. These results demonstrate the annotative power of MS2LDA, with which annotations can be made by matching only small portions of the spectra and therefore allowing annotation (classification) of molecules not present in databases. In summary, our method, adapted from text mining rather than developed explicitly for mass spectral data is able to annotate approximately three times as many metabolites as spectral matching. In addition, MS2LDA can annotate and group spectra based on neutral losses (e.g. the loss of CHOOH), which is not possible with spectral matching.

Figure 4 highlights the advantages of our novel topic-based modelling, showing a subset of the network produced by MS2LDAvis (a bipartite graph with nodes consisting of Mass2Motifs and parent ions, see Section S4.4) consisting of molecules that include two Mass2Motifs (ferulic acid and ethylphenol). All but one molecule belongs to just one of the Mass2Motifs but one belongs to both (the fragments belonging to each Mass2Motif are clearly visible). The presence of both Mass2Motifs in this molecule allows us to putatively annotate it as feruloyltyramine (314.1386 m/z; [C18H20NO4]+) despite spectral matching producing no relevant hits (see Table S8). For comparison, the output of Molecular Networking (15, 33, 34) is shown on the right of Figure 4 (see also Section S5.6). This produces clusters interpretable as ferulic acid and ethylphenol related, but as each molecule belongs to one cluster, feruloyltyramine is assigned to the ethylphenol cluster and its relationship with ferulic acid is lost. Allowing each spectra to include multiple Mass2Motifs thus gives far greater potential in making *de novo* structural annotations of molecules.

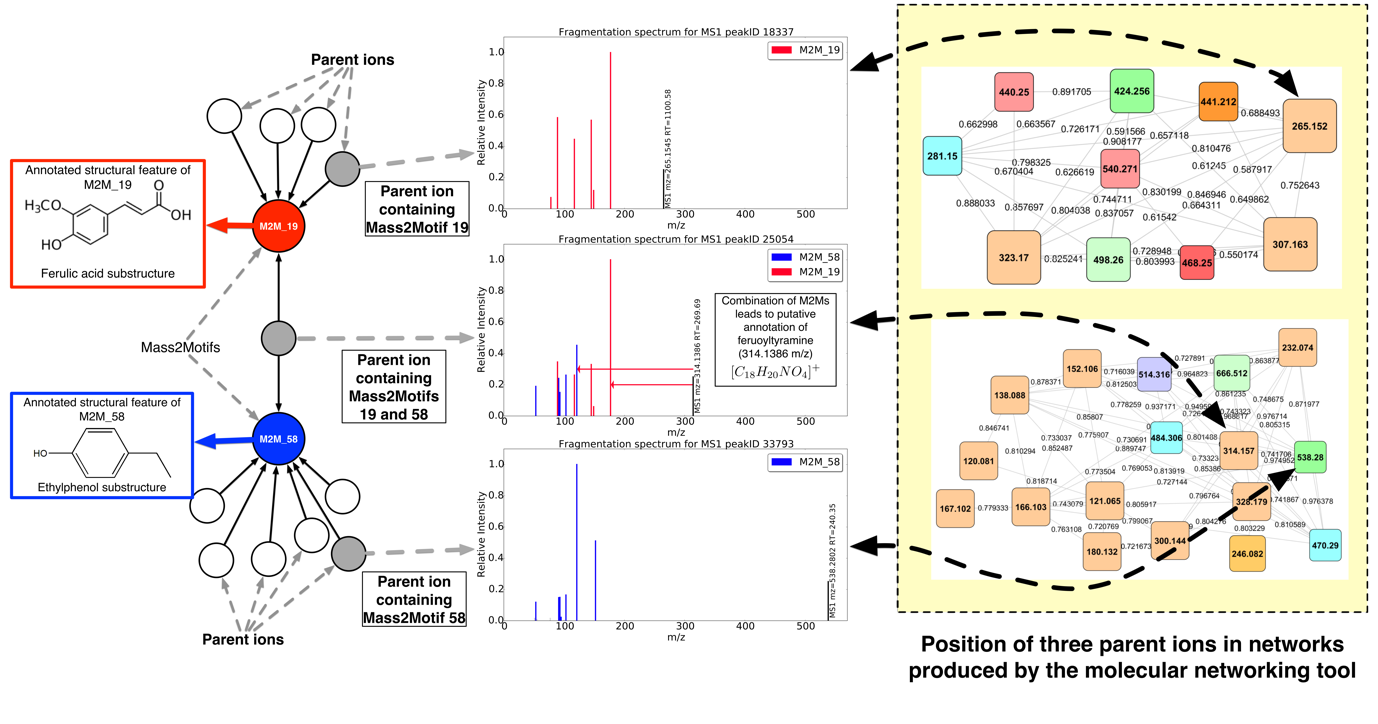


Figure 4. Mass2Motifs 19 and 58 were found to be representative of ferulic acid and ethylphenol, respectively. 11 and 42 MS1 features in the beer3 data set were explained by those two Mass2Motifs, respectively. Of those, one was explained by both Mass2Motifs, aiding in its annotation as feruloyltyramine (314.1386 m/z; [C18H20NO4]+). On the right of the plot, we show the clusters containing these three MS1 features created using the molecular networking tool (15) (top: ferulic acid, bottom: tyramine (ethylphenol)). The coloring of the nodes is dependent on their presence across the different beer extracts and the node size is proportional to the number of unique beer extracts the node fragmentation spectra are found in. The compound containing both Mass2Motifs is forced into the ethylphenol cluster, losing its relationship with ferulic acid.

## 3.4 Differential Expression of Mass2Motifs Reveals Biochemical Changes Across Samples

Being able to annotate more metabolites is beneficial when investigating the changes in metabolite intensity across multiple samples. As MS2LDA groups metabolites in a biochemically relevant manner, we can go a step further and consider the differential expression of Mass2Motifs in a manner similar to approaches taken in transcriptomics where it is common to consider the shared differential expression (DE) of a group of related transcripts as indicative of their contribution to a common aspect of cellular biology (35). Using PLAGE (36), we assessed the DE of each Mass2Motif (see S5.7) based on the intensity changes of the relevant MS1 peaks between beers 2 and 3. Figure 5 shows MS1 intensities of metabolites including two Mass2Motifs with high PLAGE scores (note that for a high PLAGE score changes in expression do not need to be in the same direction). As an example of the kind of biochemical insight this provides, in Beer3, the free guanine (Figure 5a) is more abundant whereas in Beer2, guanine-conjugates are more abundant. Similarly, the molecules associated to the pentose Mass2Motif (Figure 5B) show DE between the extracts.. We investigated whether or not similar outcomes can be achieved with with spectral similarity clustering. With this approach the 12 pentose-related metabolites were distributed across 10 clusters rather than appearing in a single grouping, hiding the potentially relevant correlated intensity change (more examples are provided in S5.7).

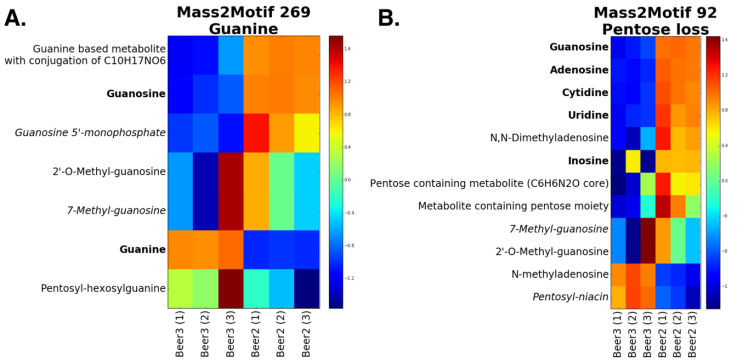


Figure 5. Log fold change heat-maps for the **A)** guanine and **B)** pentose loss Mass2Motifs. Each row is an annotated MS1 peak and columns represent samples. Bold names could be matched to a reference compound.

# 4. Discussion

MS2LDA was inspired by the idea that conserved fragments and neutral losses can be indicative of metabolite substructures and the implied parallel with topic modelling of text (spectra = documents, fragments / losses = words). No alternative tools exist that allow un-supervised substructure mining from MS fragmentation data whilst also allowing for multiple structural features or substructures to be present within one metabolite. MS2LDA can group molecules that share substructures without high similarity across their entire MS2 spectra. It reduces complex fragmentation data sets into metabolites explained by one or more patterns of concurring mass fragments or neutral losses -- Mass2Motifs.

MS2LDA relies on reliable matching of MS1 peaks to MS2 spectra and works best for complex mixtures where a large number of metabolites are fragmented and information-rich MS2 spectra are available (e.g. generated by ramped or stepped collision energy). High-resolution MS fragmentation can differentiate mass fragments and neutral losses even at low mass range of 50 -70 m/z (see section S5.9).

Manual annotation of many Mass2Motifs is straightforward and the structural features or substructures can be propagated to all connected MS2 spectra. Based on our initial experiments, automated Mass2Motif annotation is promising but requires further refinement as current spectral matching tools cannot match neutral losses (see Section S5.10). In the future, we plan include a database of validated Mass2Motifs. Other techniques developed for text mining (e.g. hierarchical LDA (37)) are also likely to offer benefits in this domain.

Metabolite annotation and identification is a bottleneck in high-throughput metabolomics. MS2LDA can assist by automatically assigning possible substructures to a fragmented LC-MS peak via the Mass2Motifs present in its MS2 spectrum. MS2LDA can thus quickly classify MS1 peaks into functional classes without knowing the complete structure of the metabolite. On average, over 70% of the fragmented metabolites were explained by one or more structurally annotated Mass2Motifs, a massive improvement on results reported in a recent study, again using beer as an exemplar, where only 2-3% of the high-abundance differentially expressed molecular features could be classified (12). In addition, the biochemically relevant metabolite grouping provided by MS2LDA allows us to identify Mass2Motifs that are enriched with metabolites with correlated intensity variation. For pathway related Mass2Motifs, this is similar to computing the differential expression of metabolites involved in a pathway without prior matching of metabolites to the pathway via identification.

# 5. Conclusions

We have introduced MS2LDA, a workflow that simplifies fragmentation data by exploiting parallels between fragmentation data and text documents. Our workflow performs all analysis steps: the preparation of a MS1-MS2 matrix, LDA, and the graphical visualization of the output. Testing on Beer extracts resulted in numerous informative “Mass2Motifs” that we could annotate with diverse biochemical core structures. This approach is markedly different from other advanced spectral analysis tools as multiple Mass2Motifs can be associated with one metabolite, and determination of the key mass fragments or neutral losses that are part of a conserved structural motif is un-supervised. Our proposed focus on pattern mining on the MS2 fragmentation data alone to aid in identification of functional classes of metabolites is novel and complementary to existing use of fragmentation data (one-by-one comparisons to spectral databases). We anticipate MS2LDA to be particularly useful in research areas such as clinical/pharmaco and nutritional metabolomics, environmental analysis, and natural products research, as it can quickly recognize substructure patterns related to drugs and food-derived metabolites in an un-supervised way.

# 6. Acknowledgements

We thank Dr. Emma Schymanski, Dr. Tony Larson, and Dr. Samuel Bertrandt for helpful discussions and assistance with implementation of R scripts from: RMassBank (ES), xcmsFragments (TL), and the 7 golden rules (SB), as well as Dr. Niels van den Broek for helpful discussions on acquisition of mass fragmentation spectra. JJJvdH was supported by the Wellcome Trust [grant no. 105614/Z/14/Z]. MPB was funded by the Wellcome Trust core grant to the Wellcome Trust Centre for Molecular Parasitology 104111/Z/14/Z. JW was supported by a SICSA PhD studentship. SR was supported by BBSRC [grant no. BB/L018616/1].

# Supporting Information:

Supporting Information is available with this manuscript – the following files can be found online:

Supporting Table S-1 (word document) - Table with Mass2Motifs (MSMs) discover in the four positive ionization mode fragmentation files of the beer extracts

Supporting Table S-2 (word document) - Table with Mass2Motifs (MSMs) discover in the four negative ionization mode fragmentation files of the beer extracts

Supporting Material containing Supporting Sections 1 – 10 (word document)

Cytoscape session files for Molecular Networking results of beer extracts in positive and negative ionization modes.

The data will become available through http://[to-be-filled-in]

# References

1. van der Hooft JJJ, de Vos RCH, Ridder L, Vervoort J, & Bino RJ (2013) Structural elucidation of low abundant metabolites in complex sample matrices. *Metabolomics* 9(5):1009-1018.

2. Kind T & Fiehn O (2006) Metabolomic database annotations via query of elemental compositions: Mass accuracy is insufficient even at less than 1 ppm. *Bmc Bioinformatics* 7:-.

3. Dunn WB*, et al.* (2013) Mass appeal: metabolite identification in mass spectrometry-focused untargeted metabolomics. *Metabolomics* 9(1):44-66.

4. Misra BB & van der Hooft JJJ (2016) Updates in metabolomics tools and resources: 2014–2015. *Electrophoresis* 37(1):86-110.

5. Hufsky F, Scheubert K, & Böcker S (2014) Computational mass spectrometry for small-molecule fragmentation. *TrAC Trends in Analytical Chemistry* 53(0):41-48.

6. Dührkop K, Shen H, Meusel M, Rousu J, & Böcker S (2015) Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *Proceedings of the National Academy of Sciences* 112(41):12580-12585.

7. Ridder L*, et al.* (2013) Automatic Chemical Structure Annotation of an LC–MSn Based Metabolic Profile from Green Tea. *Anal. Chem.* 85(12):6033-6040.

8. Horai H*, et al.* (2010) MassBank: A public repository for sharing mass spectral data for life sciences. *Journal of Mass Spectrometry* 45(7):703-714.

9. Wishart DS*, et al.* (2013) HMDB 3.0—The Human Metabolome Database in 2013. *Nucleic Acids Res.* 41(D1):D801-D807.

10. Jeffryes J*, et al.* (2015) MINEs: open access databases of computationally predicted enzyme promiscuity products for untargeted metabolomics. *Journal of Cheminformatics* 7(1):1-8.

11. Ridder L*, et al.* (2014) In Silico Prediction and Automatic LC–MSn Annotation of Green Tea Metabolites in Urine. *Analytical Chemistry* 86(10):4767-4774.

12. Allen F, Greiner R, & Wishart D (2015) Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification. *Metabolomics* 11(1):98-110.

13. Blei DM, Ng AY, & Jordan MI (2003) Latent dirichlet allocation. *J. Mach. Learn. Res.* 3:993-1022.

14. Garg N*, et al.* (2015) Mass spectral similarity for untargeted metabolomics data analysis of complex mixtures. *International Journal of Mass Spectrometry* 377:719-727.

15. Yang JY*, et al.* (2013) Molecular Networking as a Dereplication Strategy. *Journal of Natural Products* 76(9):1686-1699.

16. van der Hooft JJJ, Padmanabhan S, Burgess KEV, & Barrett MP (2016) Urinary antihypertensive drug metabolite screening using molecular networking coupled to high-resolution mass spectrometry fragmentation. *Metabolomics* Accepted manuscript.

17. Ma Y, Kind T, Yang D, Leon C, & Fiehn O (2014) MS2Analyzer: A Software for Small Molecule Substructure Annotations from Accurate Tandem Mass Spectra. *Analytical Chemistry* 86(21):10724-10731.

18. Xin C, Xiaohua H, Xiajiong S, & Rosen G (2010) Probabilistic topic modeling for genomic data interpretation. *Bioinformatics and Biomedicine (BIBM), 2010 IEEE International Conference on*, pp 149-152.

19. Zhang R, Cheng Z, Guan J, & Zhou S (2015) Exploiting topic modeling to boost metagenomic reads binning. *BMC Bioinformatics* 16(Suppl 5):S2.

20. Rogers S, Girolami M, Campbell C, & Breitling R (2005) The latent process decomposition of cDNA microarray data sets. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on* 2(2):143-156.

21. Sweeney DL (2014) A Data Structure for Rapid Mass Spectral Searching. *Mass Spectrometry* 3(Special\_Issue\_2):S0035-S0035.

22. Scott DR (1994) Pattern recognition/expert system for identification of toxic compounds from low resolution mass spectra. *Chemometrics and Intelligent Laboratory Systems* 23(2):351-364.

23. Hummel J, Strehmel N, Selbig J, Walther D, & Kopka J (2010) Decision tree supported substructure prediction of metabolites from GC-MS profiles. *Metabolomics* 6(2):322-333.

24. Creek DJ*, et al.* (2011) Toward Global Metabolomics Analysis with Hydrophilic Interaction Liquid Chromatography–Mass Spectrometry: Improved Metabolite Identification by Retention Time Prediction. *Analytical Chemistry* 83(22):8703-8710.

25. Smith CA, Want EJ, O'Maille G, Abagyan R, & Siuzdak G (2006) XCMS:  Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification. *Analytical Chemistry* 78(3):779-787.

26. Scheltema RA, Jankevics A, Jansen RC, Swertz MA, & Breitling R (2011) PeakML/mzMatch: A File Format, Java Library, R Library, and Tool-Chain for Mass Spectrometry Data Analysis. *Analytical Chemistry* 83(7):2786-2793.

27. Stravs MA, Schymanski EL, Singer HP, & Hollender J (2013) Automatic recalibration and processing of tandem mass spectra using formula annotation. *Journal of Mass Spectrometry* 48(1):89-99.

28. Griffiths TL & Steyvers M (2004) Finding scientific topics. *Proceedings of the National Academy of Sciences* 101(suppl 1):5228-5235.

29. Kind T & Fiehn O (2007) Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *Bmc Bioinf.* 8:art. no. 105.

30. Böcker S & Lipták Z (2007) A Fast and Simple Algorithm for the Money Changing Problem. *Algorithmica* 48(4):413-432.

31. Böcker S, Letzel MC, Lipták Z, & Pervukhin A (2009) SIRIUS: decomposing isotope patterns for metabolite identification. *Bioinformatics* 25(2):218-224.

32. Sievert C & Shirley KE (2014) LDAvis: A method for visualizing and interpreting topics. *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*:63-70.

33. Watrous J*, et al.* (2012) Mass spectral molecular networking of living microbial colonies. *Proceedings of the National Academy of Sciences* 109(26):E1743–E1752.

34. Nguyen DD*, et al.* (2013) MS/MS networking guided analysis of molecule and gene cluster families. *Proceedings of the National Academy of Sciences* 110(28):E2611-E2620.

35. Tarca AL, Bhatti G, & Romero R (2013) A Comparison of Gene Set Analysis Methods in Terms of Sensitivity, Prioritization and Specificity. *PLoS ONE* 8(11):e79217.

36. Tomfohr J, Lu J, & Kepler TB (2005) Pathway level analysis of gene expression using singular value decomposition. *BMC Bioinformatics* 6(1):1-11.

37. Blei DM, Griffiths TL, Jordan MI, & Tenenbaum JB (2004) Hierarchical topic models and the nested chinese restaurant process. *Advances in neural information processing systems* 16:17.