# A novel tool for mass spectrometry fragmentation data: exploration of the ‘fragmentome’ with text-modelling algorithms

## Track Records [Suggests 1 to 2 pages]

**Dr. Simon Rogers (SR)** is a Senior Lecturer in the School of Computing Science at the University of Glasgow and an affiliate member of Glasgow Polyomics (GP; an -omics data generation and analysis unit within the University). He has worked in the development of Machine Learning and Statistical approaches to -omics data analysis since starting his PhD (University of Bristol), working on transcriptomics [refs], one of the first studies to link proteomic and transcriptomic data [ref] and, more recently, on metabolomics [Refs. i.d. Wandy et al. etc]. His work on computational metabolomics includes some of the first work on assigning probabilistic confidence values to metabolite identification [ref,ref]. He has also applied Machine Learning approaches to the problem of peak alignment across files [ref] and investigated how best to combine information from multiple peaks to determine differential expression [ref,ref]. His affiliate status at GP brings him regularly into contact with metabolomics researchers. In December 2016 he was one of only 2 UK attendants at the invitation only Dagstuhl Seminar on Computational Metabolomoics. He became a Lecturer in 2009 and was promoted to Senior Lecturer in 2015. Prior to obtaining a permanent post he worked as a Post Doctoral Researcher under Prof. Mark Girolami at the University of Glasgow. He is currently funded by EPSRC, BBSRC, and Innovate UK and has also worked on industrial projects funded by Nokia as well as industrial consultancy. He is an Associate Editor for PLoS One and has reviewed for many leading journals and conferences in Machine Learning and Computational Biology (NIPS, ICML, Bioinformatics, BMC Bioinformatics, PLoS Computational Biology) as well as research councils (EPSRC, BBSRC, MRC, Wellcome Trust, Academy of Finland). He is the co-author of a popular Machine Learning textbook [ref] (second edition to be published in Summer 2016).

**Dr. Karl Burgess (KB)** is etc

**Dr. Rónán Daly (RD)** is etc

**Dr. Justin van der Hooft (JvdH)** started his academic career in Wageningen (The Netherlands) where he obtained a BSc and MSc in Molecular Sciences. During his MSc, he carried out two theses working with fluorescence spectroscopy and photoacoustic spectroscopy. An MSc internship performed at the group of Prof. Jerzy Jaroszewksi (now Dan Staerk) in Copenhagen, Denmark, offered opportunities in mass spectrometry (MS) and nuclear magnetic resonance (NMR) based analyses of biological extracts. During this time he fully elucidated structures of several complex natural products from medicinal plants and this stimulated a fascination with the processes required to identify structures of metabolites and led to his entering a PhD program at the University of Wageningen with a project on Systematic Metabolite Identification and Annotation. His PhD project focused on the development of analytical workflows combining liquid chromatography (LC), MS, and NMR. The generated data and identified structures were then used by other partners within the project to develop software tools assisting in the annotation and identification process of metabolites in complex biological extracts. His first post-doc contract brought him to Glasgow, UK, where he analysed the bioavailability of epicatechin, a major abundant flavonoid in chocolate, in the group of Prof. Alan Crozier, before moving to Glasgow Polyomics, where JvdH has been involved in different projects on metabolite annotation of urine and bacterial metabolomics data sets based on mass spectrometry fragmentation data sets. His work has been published in international journals and he has been awarded several travel awards to present his work at international conferences. JvdH is currently chair of the Early-Career Members Network (EMN) of the Metabolomics Society and is actively involved in the Metabolite Identification and Society Strategy Task groups.

**The research team and Glasgow Polyomics.** SR and JvdH started working together in Summer 2015 when it became apparent that their complementary expertise in molecular fragmentation via Mass Spectrometry and Machine Learning had potential to fill unmet analysis needs within metabolomics. Since then, they have successfully worked together to design, build and evaluate the prototype fragmentation analysis system (MS2LDA) on which this proposal is based. SR, RD and KB have collaborated extensively in the past, most recently to produce a Bayesian statistical system for metabolite identification [REF] and, more recently, in developing PiMP, a web-based metabolomics analysis suite (demonstrating clear evidence of the ability to build large systems for non-computational end users). As metabolomics and software managers at GP respectively, KB and RD provide excellent access to both metabolomics researchers and the software platforms that they are using. GP supports many researchers through the measurement and analysis of different omics data [numbers?]. As such, strong links with GP give us access to large datasets for the development of software and a large user-base for evaluation and deployment.

## Executive Summary

Some short summary explaining what we’re doing, for the short of attention…

## Background

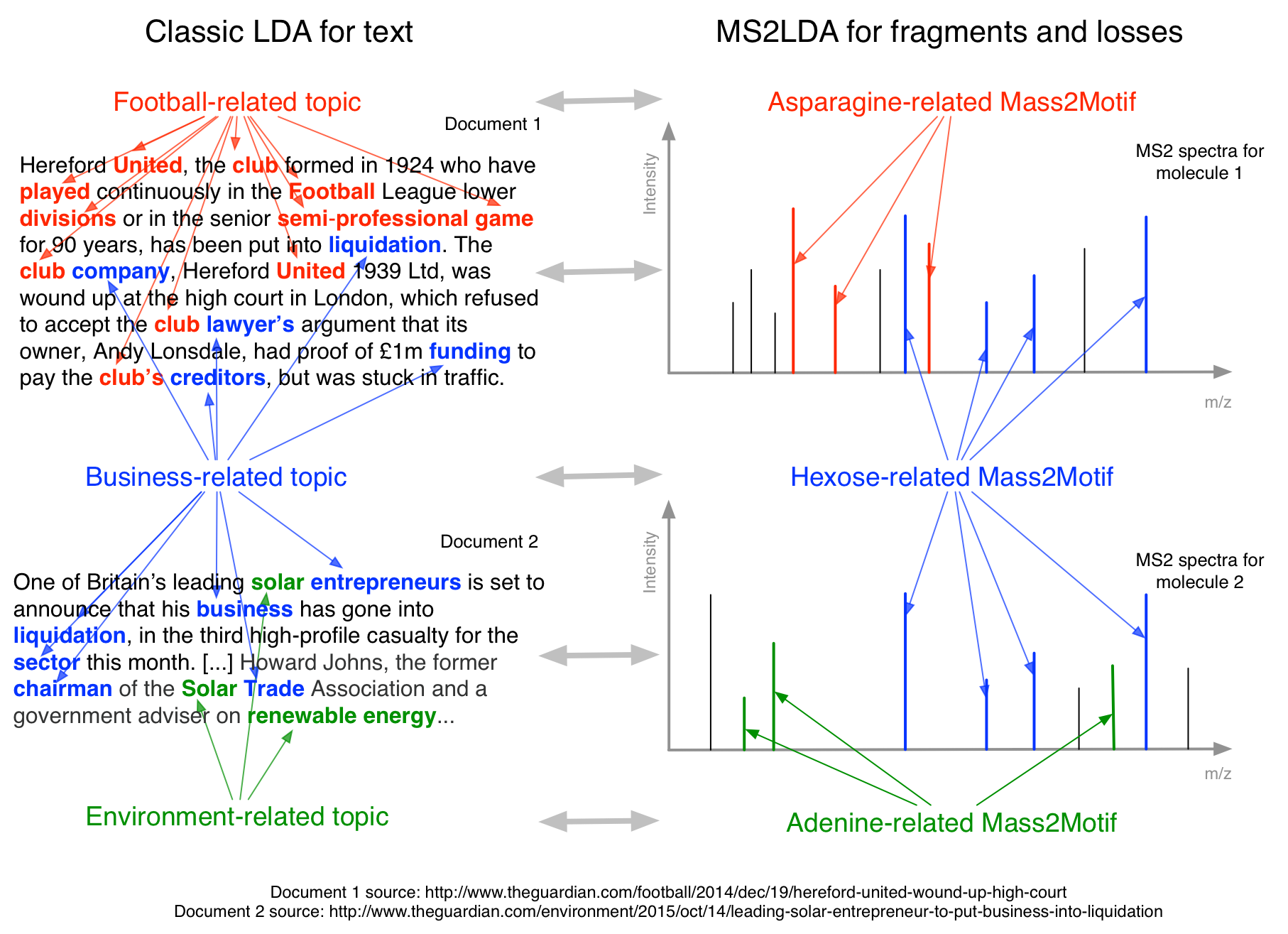
**Metabolomics.** The ability to measure changes in the metabolite composition of complex samples under various perturbation conditions is beneficial to many areas of medicine and the life sciences. Changes in the metabolome are more closely linked to phenotypic changes than those in the transcriptome and proteome [REF to Rainer paper]. The goal of a typical untargeted metabolomics study is to extract potential relationships between phenotypic variation and biochemical changes. For example, a successful experiment might result in a list of differentially expressed metabolites or a list of pathways enriched with differentially expressed metabolites. The standard paradigm for analysing metabolomics data relies on the ability to identify the molecules that have been measured. For reasons discussed below, the identification step is a bottleneck in the metabolomics analysis pipeline — only analysing molecules that can be identified through traditional means results in vast quantities of useful data being discarded. This proposal addresses this problem *through the development of a suite of computational tools that bypass the identification bottleneck by performing analysis in fragment space, inspired by algorithms from the domain of text processing.*

**LC-MS data.** Liquid-Chromatography coupled to Mass Spectrometry (LC-MS) is one of the most widely used experimental platforms for high throughput metabolomics. The separation phase (liquid chromatography; LC) prior to mass spectrometry (MS) allows complex samples to be analysed as it simultaneously separates molecules with very similar masses and overcomes the problem of ion suppression by reducing the number of molecules competing for charge in any given MS scan. Each ion measured is characterised by its mass per charge (m/z), its chromatographic retention time (RT) and its intensity (abundance) that might change across different samples. Identification or annotation (see [] for more precise definitions of these terms within the metabolomics community) of molecules from these measured ions is very challenging and, in some cases, impossible. Firstly, there will be many more ions than measurable molecules due to contaminant molecules and the fact that each metabolite can potentially result in the formation of many ions (different adducts, isotopes, in-source fragmentation etc). Secondly, although the mass resolution of modern MS is incredibly high, mass is not sufficient to reliably identify many molecules (due to isobaric molecules, having nearly identical masses, and isomers, having the same elemental formulas and thus the same masses). RT offers an extra dimension that can aid in annotation but is notoriously variable and hard to predict. For these reasons, gas-type fragmentation experiments (used widely in proteomic MS) are becoming increasingly popular in metabolomics.

**LC-MS/MS fragmentation.** Mass Spectra of fragmented ions can be used to identify molecules. This is done by comparing an ion’s fragmentation spectrum against a database of reference metabolite spectra. These reference spectra are obtained by fragmenting synthesised molecules, or, increasingly via in-silico fragmentation of known structures. A common method of MS fragmentation is *Data Dependent Acquisition (DDA)*. In DDA, the top e.g. 5 or 10 most intense peaks in the MS scan for a particular RT are automatically isolated and fragmented. The resulting data is more complex than standard LC-MS as is consists of both ‘standard full-scan MS spectra’ (containing peaks measured from un-fragmented ions (MS1 peaks)) as well as fragmentation spectra (containing MS2 peaks) for the most intense MS1 ions. These data sets can be large — a typical DDA dataset for a single experimental sample can consist of more than a thousand fragmented MS1 peaks and several thousand fragment peaks. Whilst more reliable than identification from just MS1 data (m/z, RT), identification from fragment spectra is limited to those molecules for which reference spectra exist (a relatively small proportion of molecules for even the most well-studied organisms). Spectral databases are growing, but only with known molecules — we cannot populate databases with molecules of which we have no knowledge. Fragmentation is also equipment specific which will be particularly problematic when new MS technology is developed (it will take time to populate databases with spectra measured on the new technology). In summary, fragmentation-based identification is a key tool in the analysis of metabolomics data but does not overcome the identification bottleneck and is unlikely to do so in the foreseeable future.

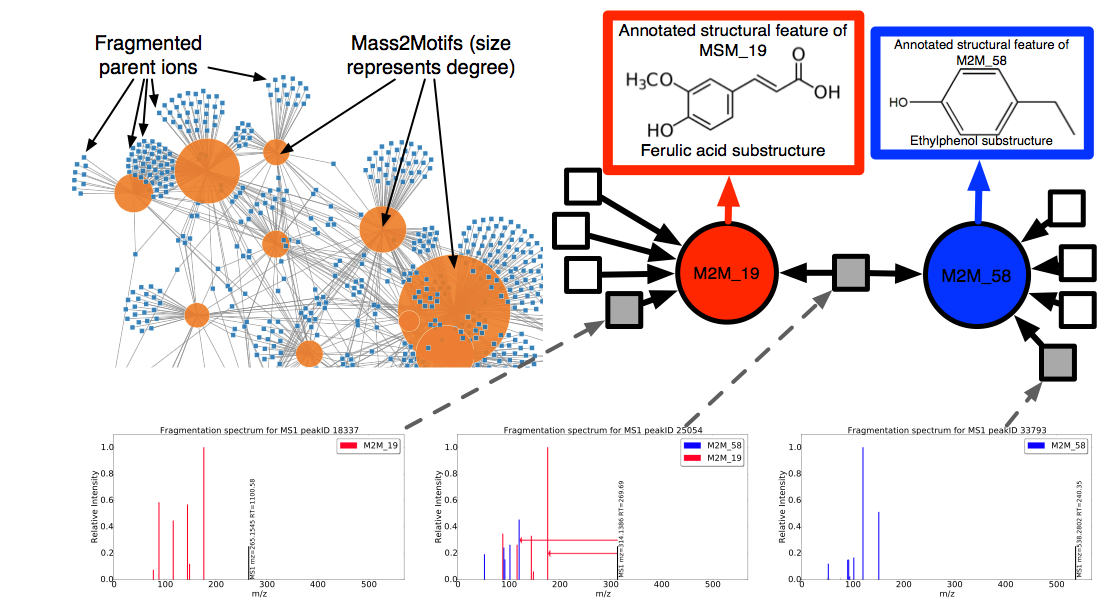
**Holistic analysis of Fragmentation Data.** Very few approaches have looked beyond using fragmentation spectra only for identification. Those that do exist typically have in common that they extract *groups* of *similar* spectra (i.e. groups of molecules). For example, the amino acid histidine displays several characteristic mass fragments upon fragmentation and related metabolites present in a sample can be grouped together based on those shared fragments (Figure YY-C). Molecules grouped in this way can be putatively annotated (e.g. as *histidine-related*) and can be explored for consistent differential expression in relevant MS1 data (e.g. differential expression of our group of *histidine-related* molecules provides an potential biochemical hypothesis to explain the phenotypic variations across the groups of samples).

**Grouping MS2 spectra - defining similarity.** To produce groups of similar spectra we must define what we mean by *similar*. Two recently published tools: MS2Analyser and Molecular Networking do this rather differently. MS2Analyser allows the user to extract all spectra that include a set of user-defined fragments and/or neutral losses. For example, the user might wish to extract all spectra that include a CO or H2O loss (a neutral loss is defined as the difference between the fragmented ion and a resulting mass fragment) or fragments indicative of a particular molecule substructure / moiety (e.g. the aforementioned histidine example). The spectra returned potentially have a clearly defined biochemical relationship but specifying interesting sets of fragments and/or losses is challenging in experiments where the goal is to *uncover* the biochemical changes. For example, which fragments and losses should we look for when analysing the effects of a new therapeutic? The lack of knowledge on the type of metabolites that are affected by the treatment, or the lack of known key characteristic fragments for the metabolites (drugs) of interest is one of the reasons for performing the experiment! Molecular Networking, in contrast, takes a purely data-driven approach, grouping spectra according to their overall level of spectral similarity (defined by the vector-based cosine score — the same scoring system used in the majority of database search routines). This has the advantage of requiring no user intervention. However, the measure of similarity can be too crude and the groupings are not as immediately interpretable as those derived from a specific subset of fragments and/or losses. For example, consider two molecules that share a ethylphenol substructure. The fragments derived from this substructure might represent only a small portion of their respective MS2 spectra resulting in a low overall spectral similarity. Consequently these molecules might not be grouped. However, it is clear that they are chemically related and, if both molecules are differentially expressed, their similarity might point to the biochemical changes present within the experiment.



**Figure** **TT**. The relationship between LDA for decomposing text into topics (left) and MS2LDA for decomposing fragment spectra into substructures (right).

**MS2LDA.** Recently, we presented a proof-of-concept study motivated by the parallels we observed between DDA mass spectrometry metabolite fragmentation data and text documents. Like Molecule Networking, this method does not require user-specified groups of fragments and masses but, like MS2Analyzer, it produces groups of molecules with clear biochemical relationships. Text analysis is a domain rich with sophisticated analysis methods, the porting of which to other domains is likely to be very productive. Traditionally text documents could be grouped together to aid with search and recommendation based on global similarity (c.f. Molecular Networking) or selected based on the presence of particular words or phrases (MS2Analyzer). A drawback of using global similarity in text documents is the underlying assumption that each document is about only one *topic*, which is hardly ever the case*.* This unrealistic assumption for text documents was overcome through the introduction of Latent Dirichlet Allocation (LDA) which assumed that documents were combinations of topics and could group documents in a more meaningful manner. Based on a corpus of text documents LDA is able to learn the topics (groups of co-occurring words) and label each document with the topics that it consists of.



**Figure ZZ**. Results from an MS2LDA analysis of a beer sample. The network on the top left is one way to visualise the output of MS2LDA. The small squares are the fragmented parent ions, which are connected to the Mass2Motifs (large circles). It is clear that many molecules consist of >1 Mass2Motif and that some Mass2Motifs appear in a large number of molecules. In the top right, we show a zoom of one part of the network consisting of two Mass2Motifs (red and blue) that we have been able to annotate as indicative of a ferulic acid and ethyl ethylphenol and parent ions that consist of one or other or both Mass2Motifs.The central ion consists of both Mass2Motifs allowing us to putatively annotate it as XXXXX. In tools that allow molecules to be in only one group this molecule would be forced into one or the other and this annotation would not be possible.

We consider there to be clear parallels between text documents and MS2 spectra, depicted in Figure TT. As documents consist of multiple topics, so molecules can consist of multiple sub-structures. MS2 spectra can be decomposed into molecular substructure in a manner analogous to the decomposition of documents into topics. We have developed MS2LDA, a prototype analysis tool for LDA analysis of DDA data. In MS2LDA, substructures are automatically discovered as frequently co-occurring combinations of fragments and neutral losses (known as *Mass2Motifs*). Molecules can be meaningfully grouped if their spectra share a Mass2Motif, even if this Mass2Motif only explains a small portion of their spectra, and no neutral losses or fragments have to be specified by the user. Figure ZZ shows example output from MS2LDA analysis of a beer sample, highlighting some of the strengths of the method.

In summary, MS2LDA overcomes some of the main problems in fragment-based analysis:

* 1. **It does not rely on molecular identification**. Information from all fragment spectra can be used, not just those that can be matched to reference spectra. In our experience Mass2Motifs are much easier to identify/annoate than molecules. In our analysis of beer extracts, we found that over 70% of measured spectra included one or more easily annotatable Mass2Motifs and could therefore be given some degree of annotation. Less than 10% of the same spectra could be matched to the MassBank database at a conservative score threshold of 75%.
  2. **It extracts biochemically relevant substructures** in an unsupervised manner.
  3. **Molecules are grouped according to biochemical similarity** even if they lack a high degree of global similarity. Molecules can include multiple Mass2Motifs and can therefore belong to multiple groups, reflecting the diverse biochemistry of some molecules.

## Aims: A software platform for MS2LDA analysis of the fragmentome

Our proof-of-concept study demonstrates the potential of MS2LDA but our implementation exists as a set of scripts. Here we propose the development of a **suite of user-friendly computational tools** that will bring the complete MS2LDA pipeline (and extensions described below) to a large user group. Our proposed tools:

1. Will be **data-centric**: they require no database of reference spectra (although they will be able to interact with, and integrate results from these databases).
2. Will be **unsupervised:** they do not require access to a large corpus of labeled training data.
3. Will identify the **chemical relationships** between fragment spectra without relying on molecular identification.
4. Will use **all fragmented molecules** in the DDA data, and not just those for which reference spectra exist in a database.
5. Will be able to highlight how **Mass2Motifs are linked to change in peak intensity** and the changes in **prevalence of particular chemical structures** across cohorts.
6. Will be able to **putatively annotate molecules** based on their Mass2Motifs.
7. Will be **user-friendly**, and accessible to a wide range of metabolomics researchers.
8. Will be **cross-platform** (both in terms of MS and software).

The following three use case sketches demonstrate the core functionality of the proposed software:

**Use case 1 – Standard MS2LDA:** A user has performed DDA analysis on a single urine sample. Their goal is to better understand the metabolite content of this sample. Comparing spectra to a database results in high confidence hits for ~15% of the fragmented molecules leaving many unknown. They run their data through the MS2LDA software and are presented with an interactive visualization. This allows them to rapidly explore the most prevalent substructures and the molecules in which they are present, as well as highly-preserved fragment patterns of potential drug-derived metabolites. The software allows the user to annotate Mass2Motifs manually, or through a database search. If they wish, users can perform MS2LDA with a set of manually curated Mass2Motifs derived from previous analyses. Using this latter method, the user discovers that many molecules contain a clopidogrel substructure indicating administration of the drug and its subsequent metabolism by the human body (see also Fig YY-B).

**Use case 2 – linking MS2LDA to MS1 data:** A user is performing an analysis of the metabolomic makeup of samples from two cohorts. As well as MS1 data, they perform DDA analysis on pooled samples. As in a normal analysis, a peak-level comparison is performed on the MS1 files, identifying fold changes of the MS1 peaks. MS2LDA analysis is performed on the DDA file. The software allows the MS1 peaks to be matched to their respective MS2 spectra. Mass2Motifs can now be ranked according to *their differential expression*, computed from the fold changes of the molecules associated with them (using e.g. PLAGE, a group-based statistical approach developed for transcriptomics data sets where transcripts can often be grouped into one gene family). Mass2Motifs with significant differential expression are highlighted and represent excellent starting points for exploration and hypothesis generation. For example, in our comparison of two beer extracts, a Mass2Motif annotated as guanine was found to be significantly enriched with DE molecules leading to a hypothesis as to the differing biochemistry of the two extracts, as shown in Figure GG.

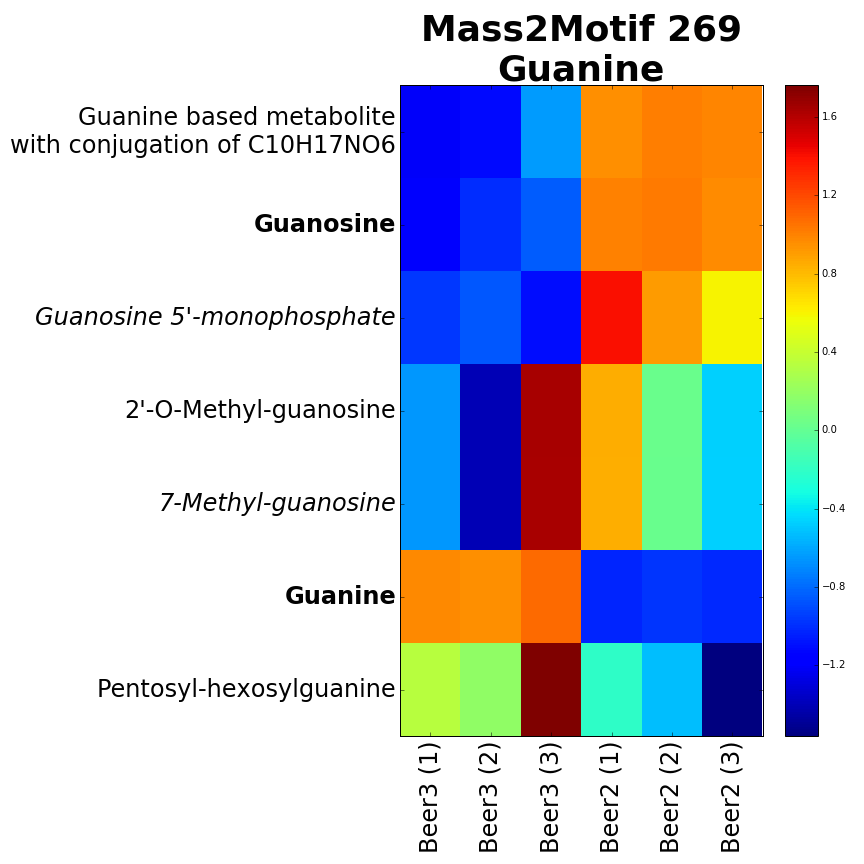


Figure GG. Molecules including a Mass2Motif identified as guanine show clear differential expression across two beer extracts (three replicates of each). Interestingly, guanine is over-expressed in Beer3 whilst conjugates of guanine are more prevalent in Beer2.

**Use case 3 – Multi-file MS2LDA:** A user is performing an analysis of the metabolomic makeup of urine samples from two (or more) cohorts. The user has performed DDA fragmentation on **all samples** in the study. Multi-file MS2LDA (details below) is used to analyse all data simultaneously, automatically decomposing the entire set of files into shared Mass2Motifs, and identifying how these Mass2Motifs change across the cohorts. Key characteristic fragments of losartan (see also Fig YY-A) are present at different levels across different urine samples and provide a quick insight in differences in metabolism between the different patients. Moreover, the user finds the presence of a food-related Mass2Motif to be correlated to the number of losartan metabolites in the urine. This information would not be as readily accessible in any other analysis platform and suggests useful hypotheses for subsequent verification.

## Programme and Methodology

### Objectives

Our primary objective is the development, evaluation and deployment of a set of algorithms implemented within a user-friendly and accessible software system that allows metabolomics researchers to more efficiently explore the biochemical makeup of their MS fragmentation data both in individual files and comparatively across experimental cohorts. The following work packages (WPs) describe the pieces of work that will be performed to meet these objectives, and the deliverable outcomes at each stage against which the work should be assessed.

### Programme of work

**WP1: Efficient MS2LDA implementation (SR and PDRA based in Comp. Science (CSPDRA)).** To enable the system to scale to large datasets an efficient MS2LDA implementation will be required. We will follow the techniques used within the text domain (LDA has been used on very large corpora [??]) to improve the efficiency of the inference algorithms ensuring that all developed software is efficient, scalable, and usable for metabolomics researchers. This and all other codes developed will be open source and made available through code sharing resource (e.g. github) and institutional repositories.

**WP2: Algorithmic development (SR, RD, CSPDRA):** There are three directions in which we will extend our prototype MS2LDA system, described below. All algorithmic developments will be assessed in the test case development package (WP5):

1. Developing the framework to allow MS2LDA analyses to be linked to MS1 fold changes. Use case 2 and the guanine example in Figure GG, demonstrate the benefit for being able to link Mass2Motifs to MS1 data. This requires three parts: extracting fold changes from MS1 data, the matching of MS1 peaks between the MS1 data and fragment data and the subsequent statistical analysis of the MS1 expression for all peaks corresponding to a Mass2Motif. The first part has been solved by many systems, most of which support the export of a .csv file and we do not plan to re-implement this here. For the second part, we will combine two methods — one that implements a stable matching system (like that used in e.g. SIMA [A] and our previous work on MS1 alignment [B]) and one that uses the resulting MS1 peak lists to actually extract the MS2 spectra prior to MS2LDA analysis. For the final part, differential expression of Mass2Motifs can then be computed in a manner analogous to that used for computing the expression change of sets of transcripts (various methods will be compared, see e.g. []). In addition, we have previously developed Bayesian ANOVA methods for groups of variables that were shown to be particularly effective for MS1 data [X,Y].
2. Multi-file MS2LDA (see also user case 3). As highlighted in user case 3, processing multiple files together will allow us to directly compare the changing abundance of Mass2Motifs that are *shared* across the files. Multi-file MS2LDA will decompose each file intosharedMass2Motifs. Each file will have its own *distribution* over the Mass2Motifs. Comparison of these distributions across files will highlight how the Mass2Motifs (and hence the biochemistry) change across the files. We anticipate that this will be particularly useful in clinical applications (and will evaluate it on clinical data, see WPBB). This is novel for LDA in general and not just MS2LDA and we anticipate it being of interest in other domains (and will publish accordingly).
3. More sophisticated Mass2Motif models. A CO loss will appear in many very different molecules. In standard MS2LDA it will likely appear as both a single Mass2Motif (including just the CO loss), in very specific Mass2Motifs (including the CO loss and other fragments / losses), and other Mass2Motifs at intermediate levels of specificity all of which are considered to be completely separate entities. The same issue has been overcome in the text domain via Hierarchical LDA [C] where topics become more specific (typically include more words) the further you move down the hierarchy. In MS2LDA hierarchical Mass2Motifs will enable a more parsimonious decomposition of the fragmentation data and make interpretation of the MS2LDA output faster. We will translate Hierarchical LDA into the MS2LDA framework allowing users the choice between Hierarchical Mass2Motifs and the original Mass2Motifs (based on *bag-of-words* models). We will also consider Mass2Motif models inspired by recent work in in-silico fragmentation. CFM-ID [D] uses a Markov Chain model to predict how molecules will fragment. LDA has been extended to topics built from Markov Chains in the domain of XXX [C; Mark’s paper] and we will explore the possibility of similar extensions here.

**WP3: Annotated Mass2Motif databases (SR, RD, JvdH, CSPDRA):** One of the benefits of MS2LDA is that it is unsupervised. However, having gone to the effort of annotating Mass2Motifs (see e.g. the guanine example above), it makes sense to use them again. We will make our MS2LDA software more flexible to allow it to use a combination of both fixed and unknown Mass2Motifs. Pre-annotated Mass2Motifs will allow rapid annotation of molecules containing them, leaving researchers more time to spend on the previously unexplored areas of the data. This work will take two forms — a local database (built into the standalone Java application, see WP4) to which users can add their own annotated Mass2Motifs, and a central SQL database curated and maintained by JvdH. To avoid reproducing resources available elsewhere we will assess the feasibility of using existing databases as sources of Mass2Motifs. For example, databases with exact mass fragment spectra (e.g. [XX]) can be used, where relevant, to provide input Mass2Motifs. Where this is possible we will provide the software to allow such links.

**WP4: A user-friendly MS2LDA workbench (RD, SR, JvdH, CSPDRA):** Success of this project is dependent on closely working with potential users to ensure a system that meets their needs, and allows them to get the most out of the MS2LDA method. We propose maximising user involvement in development and final user base by developing two systems. System 1 will be a Python/Django web application linked to the open source PiMP software [E] already developed by Glasgow Polyomics and incorporating a visualisation module in, e.g., D3. The combination of the speed of developing in Python/Django/D3, and the large and increasing user group using PiMP will allow us to rapidly prototype our developments and obtain feedback from real users. It will also allow us to link to MS1 data already stored in the PiMP system. The Python/Django system will be made available as part of PiMP but the barriers to installing such a system on a local server are likely to be too high for the average metabolomics user. For this reason, in the latter stages of the project we will use the feedback from the rapid Python/Django development cycles to design and build a standalone application (written in Java for portability) that will have the same functionality but can simply be downloaded and run by potential users on any standard PC. We are under no illusions as to the difficulty of producing a piece of software that is robust and usable by non-computational experts. However, within the team, we have the expertise to do this. RD has considerable Software Engineering expertise and will lead this software engineering WP. RD leads the PiMP development project within Glasgow Polyomics that has recently been successfully released to a large cohort of users within Glasgow. SR also contributed code to PiMP, and KB provided substantial design inputs. JvdH has experience with testing metabolomics and mass spectrometry fragmentation related software.

**WP5: Test case evaluation (JvdH, KB, RD):** A key part of dissemination of this system will be the existence of a set of strong test cases that demonstrate the capabilities of MS2LDA. Our strong links with Glasgow Polyomics (KB, RD, JvdH) provide us with access to a large number of collaborators willing to make their data available for this project. Algorithmic extensions and the Python/Django web application will be continually evaluated with this data. In addition, to ensure that we have access to as much fragmentation data as possible (fragmentation is not done routinely for all samples), we have budgeted for the analysis of XX samples by DDA fragmentation. We will perform the following four test cases to demonstrate the potential of the new MS2LDA tools:

1. Validation of Mass2Motifs found by multifile-MS2LDA and comparison to those found by MS2LDA. Here we will use a set of previously analyzed as well as un-analyzed beer extracts to extract Mass2Motifs. We define success as being able to find – unsupervised – the same biochemically relevant cores as MS2LDA found. We will compare the fragment/loss composition of mass2Motifs to see how that changes upon using multiple files. In addition, we also have access to MS1 data for these beer extracts allowing us to demonstrate the ability to link MS2LDA with MS1 data.
2. How does the performance of multifile-LDA compare to drug metabolites found by molecular networking and manual analysis of urine extracts containing antihypertensives? Here we will make use of a dataset already analysed by JvdH [F] consisting of fragmented urine samples from patients who have been given antihypertensive drugs. We will compare the findings of this analysis with the previous analysis of this data (based on molecular networks). This will provide a test case that directly highlights the differences between two approaches. In addition, we have access to a previously un-analyzed set of fragmented urine extracts which we will analyse separately to assess the consistency in the observed Mass2Motifs.
3. Other test cases will be assessed as other data become available through collaborators. In particular, we will be interested to demonstrate MS2LDA on various MS platforms allowing us to compare Mass2Motifs extracted from the same samples on different platforms.

**WP6: Evaluation on other data types (All):** The core focus of our proposed work is for the analysis of DDA metabolomics fragmentation data. However, we hypothesise that MS2LDA-type methods could be useful for other types of fragmentation data. Indeed, our core algorithm will be developed in such a way as to not be restricted to data from any particular protocol by ensuring it can work with multiple open data formats (mzML, mzXML, etc). Through our local, national, and international collaborators, we will investigate the potential of MS2LDA in other fragmentation domains. The following have already promised data on which MS2LDA (and the proposed extensions) can be evaluated: JACQUES, Prof. Andy Jones (University of Liverpool; proteomics), LIPIDS?, Dr. Emma Schymanski (ETH Zurich; natural products)?

**WP7: Dissemination (All):** End-user engagement is crucial to the success of the project. Through our connections with Glasgow Polyomics (particularly KB, head of metabolomics, and RD, head of sotware) we will have a clear line of communication to potential end users in Glasgow. This will be used for continued evaluation of the software (WP4) and working closely with researchers (JvdH) to help them use the tools, and discover the additional features that would make the tools more appealing to the potential users. Once the Python/Django system is ready with the basic MS2LDA functionality (i.e. use cases 1 and 2) we will begin to recruit local users from the researchers that use GP (pooled fragmentation files are collected as standard in the GP metabolomics data collection protocol). As the system will be linked to PiMP (which is used by researchers performing metabolomics at GP) it will be possible for the users to experiment MS2LDA with a very low barrier to entry. As the system becomes more mature based on this initial feedback, the PiMP/Django MS2LDA system and the standalone Java application will be included in the metabolomics training courses offered by GP, reaching a wider audience of academic and industrial researchers.

## Project Management

The day-to-day management of the project will be performed by SR who will also act as the line manager for the CSPDRA. JvdH will be line managed by KB, within his role as the metabolomics manager at Glasgow Polyomics. The CSPDRA and SR will be based in the School of Computing Science, and KB, RD, and JvdH at GP. The entire team will meet fortnightly at GP (SR, KB, RD, and JvdH already meet regularly to discuss ongoing projects). Within their current roles, RD and KB are continually in contact with metabolomics experimentalists within the University and further afield. They will feed back input from these stake-holders at the regular meetings.

Four times throughout each year, we will hold ‘coding’ days at GP, in which demonstration of the software will be provided (to the project team and other parties), issues with the software will be discussed and resolved, and plans made for the subsequent development.

Interdisciplinary projects such as this can come with risks associated with communication. In this instance, we believe this risk is minimal as all named members of this project have worked together extensively in the past on metabolomics problems (SR, RD and KB have worked together for several years, and are currently all investigators on an Innovate UK project [refxxx]).

We will make every effort to recruit a CSPDRA with experience in both Machine Learning and metabolomics. This may not be possible (very few people have this combination of skills), increasing the risk that the CSPDRA will take some time to get to grips with the metabolomics element. However, the expertise exists within the team to provide the necessary training (indeed, one of SR’s PhD students did much of the work on the prototype MS2LDA system, starting with no knowledge of MS fragmentation data, but able to learn rapidly working alongside JvDH).

/notes – alternative narrative

1. Fragmentation data is currently used almost entirely for identification.
2. This ignores the richness present in DDA (or DIA?) type data
3. Better fragment-based analysis could answer key questions in clinical (?) metabolomics that can’t currently be answered well by standard approaches
   1. And standard approaches have hit a wall based on the small numbers of compounds for which we have pure reference spectra for, and these go out of date as soon as new technology comes along
4. We have developed a prototype system that decomposes large fragmentation datasets into conserved sets of fragments and losses that we have demonstrated represent biochemically relevant substructures
5. This is an improvement on the alternatives because stuff
   1. Discovery of chemistry
   2. Uses all fragmented ions, not just the ones that match to a DB
   3. Substructures can be indicative of drugs / pathways
   4. Provides a more chemically meaningful definition of similarity between molecules
   5. Allows for multiple annotations of molecules
6. In this proposal we want to develop this prototype into a tool that can be used for large, untargeted clinical metabolomics studies.

/notes

to cover:

1. MS-based metabolomics data is important but we are unable to use it to its full potential due to analysis bottlenecks that remain to be solved. We can align and extract features and pull out those who’s intensities change in an interesting manner, but at this point we hit a brick wall because…
2. Structural identification is virtually impossible for all but a small number of compounds without fragmentation information.
   1. Compound ID is now done via comparison of fragmentation spectra with reference libraries for known compounds.
   2. This approach is fundamentally flawed (limited). It relies upon possession of a reference spectra which are available only when a pure compound has been run in isolation **on the equipment of interest**. This is particularly problematic when new technology is developed. E.g. change of capillary.
   3. There have been some advances in predicting spectra from structure (e.g. CFM-ID, Grimm) but it’s hard to be sure of the quality of the obtained spectra **and** they require a large training set ideally **on the equipment of interest.**
   4. All of this requires that we somehow know what we’re looking for. In the case of novel drugs, this will not necessarily be the case
3. It is also worth mentioning that individual changing metabolites don’t tell us an awful lot. We would instead like to know why we see the changes we do. I.e. we might typically try and map identified metabolites to pathways.
4. Can we make progress if we put full metabolite identification to one side for a while?
   1. There is a great deal of information in a LC-MS DDA file that we lose if we consider each spectra individually. For example, the presence of a shared set of fragmentation peaks within the spectra of severally differentially expressed compounds might be indicative of a shared substructure which in turn may give a (bio)chemical (?) hypothesis as to why they are changing.
   2. A recently developed tool (Molecular Networking) groups compounds according to similarity of fragmentation spectra. This allows us to look at whether such groups are enriched with DE metabolites [DOES IT?] but the spectra only cluster together if their spectra are globally similar. If they share, for example, a functional subunit but differ elsewhere in the molecule they will not be grouped, and it may be this subunit that is of interest.
   3. Describe MS2LDA as a prototype system.
5. In this proposal we describe how we intend to create a user-friendly, easily accessible MS2LDA system that can provide previously unavailable insight and aid in the creation of testable hypothesis across a wide range of fragmentation data types and experimental designs.

The following three use cases describe ways in which our platform will enable analysis of LC-MS DDA data. Use cases for other data types are described in ‘Case Studies’.

Example Use Case 1 – Standard MS2LDA

Example Use Case 2 – MS2LDA on a pooled file for a case-control experiment

Example Use Case 3 – Multifile LDA

The large-scale characterization of the metabolome in complex samples through the use of Liquid Chromatography coupled to Mass Spectrometry (LC/MS) is becoming more popular across many areas of medical and life science research. Changes in the metabolome can be more closely related to phenotype and as such provide a key view of the operation of an organism in healthy or diseased states, or in response to a treatment.

The standard approach to generation and analysis of LC/MS data consists of running each sample (e.g. a urine sample from each patient in a clinical trial) through the LC/MS instrumentation, matching peaks between the runs, statistically determining which metabolites are significantly changing across different samples, and finally identifying those metabolites. The major bottleneck in this procedure is the structural annotation and identification stage. In many cases, the measured mass of a molecule is insufficient to unambiguously identify it. Orthogonal information (obtained through the LC/MS), e.g. capillary retention time (RT) is useful but very difficult to predict (for a molecule - equipment pair) and can only provide limited assistance. Regardless of the information used, identification involves comparing the data observed with records in one or more databases. If the molecule is not in the database, it cannot be identified without complementary structural information, e.g., from nuclear magnetic resonance spectroscopy (NMR) data).

One aid in identification is the use of mass spectrometry fragmentation ( also reffered to as tandem MS or MS/MS) where molecules detected in the MS are isolated and then fragmented. A mass spectrum of the generated molecular fragments is subsequently measured. The manner in which a molecule fragments is dependent on its chemical structure and therefore its fragmentation spectrum is indicative for its identity. Fragment spectra can be used as ‘fingerprints’ of metabolites and searched across spectral databases of known molecules to provide identifications that are considered more reliable than those just based on the obtained mass and retention time from a LC-MS experiment without fragmentation experiment.

Identification via fragmentation is again limited by the set of metabolites for which we have database entries - a subset of the complete metabolome biased by the very fact that they are easily accessible, abundantly present, and identifiable. Furthermore, to achieve reliable identification, database entries also have to be on the same equipment, and with the same fragmentation process and energy (both influence the observed spectra). Exotic and previously unseen metabolites can still not be easily identified based on spectral matching. This is a problem - in many applications of metabolomics (particularly clinical applications where new therapeutic treatments are being investigated), it is very likely that metabolites of interest will not have been previously identified and catalogued. This hampers biological and clinical research - we need tools that are not biased and restricted to those things we already know.

The ‘classical’ use of fragmentation data information is by a one-to-one comparison of MS2 spectra to spectral databases consisting of experimental and/or theoretical fragmentation spectra, thereby ignoring the chemical relationships existing between metabolites in the biological extract.

Mass spectrometry fragmentation experiments generate large quantities of complex data that is hard to interpret. Modern mass spectrometers can generate more than 6000 fragmentation spectra within the course of a 20 minute LC run. This complexity motivates the development of scalable procedures tailored to the unique statistical properties of the data (REF to machine learning review]. In particular, methods that allow us to focus on the relevant information within the data without prior knowledge are needed to integrate fragmentation analysis with larger-scale metabolomics experiments.

To overcome these limitations, we propose a fundamentally different approach to the analysis of high-throughput metabolomics data based on performing the key analysis steps in the fragment space rather than the metabolite space. We believe that this analysis is more relevant to the scientific questions that the experiments are designed to answer, and is not restricted to the small subset of metabolites for which identification is possible.

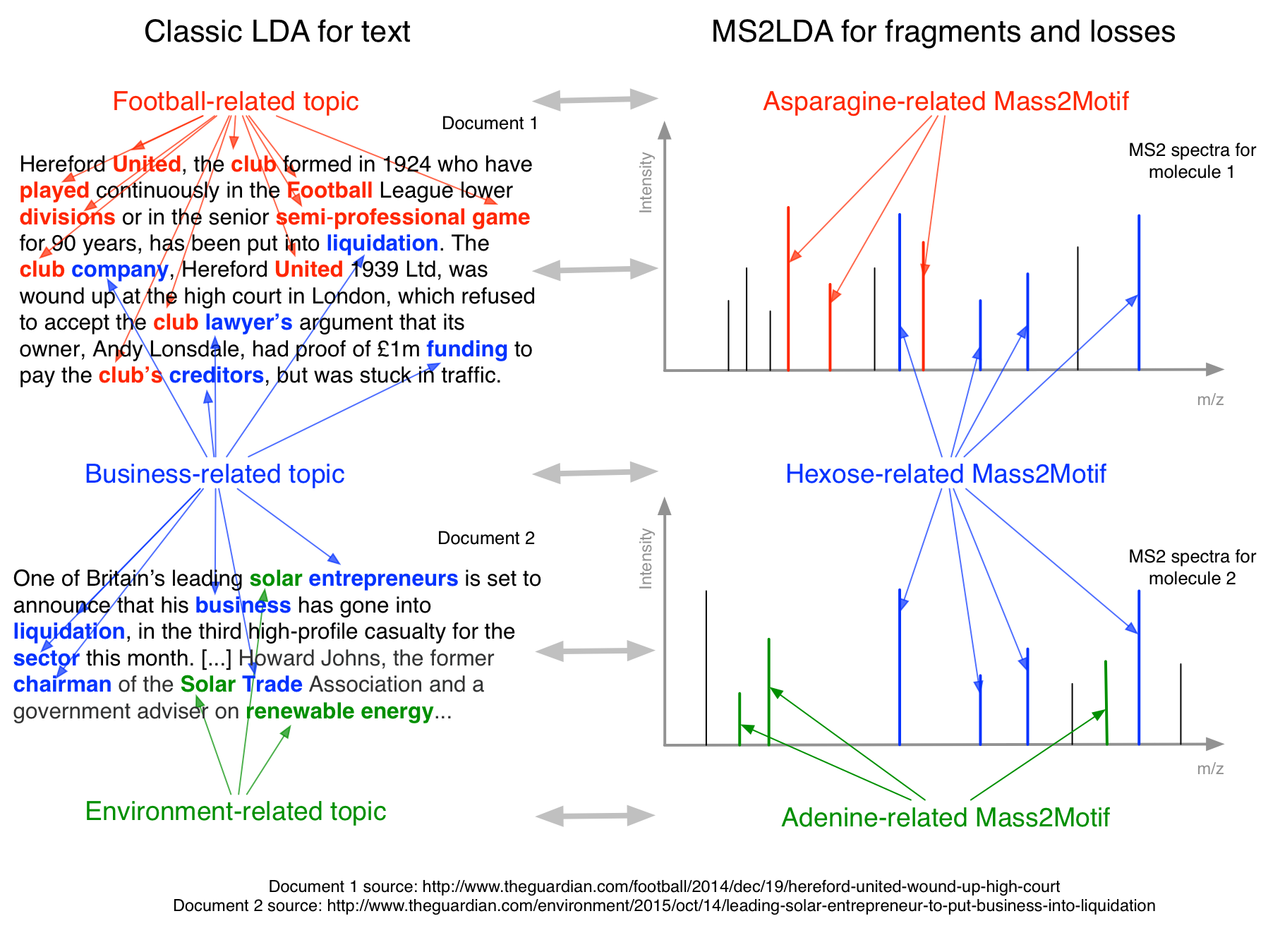


Figure 4 Cartoon showing the relationship between LDA for text and MS2LDA for fragmentation spectra

Our proposed method includes changes to how the data are both generated and analysed.

Current data collection protocols are focused on the collection of MS1 data, with fragmentation data being collected less often (perhaps in pooled samples) and only used for identification. We propose the routine use of data-dependent (DDA) fragmentation data (the most abundant ions in each scan are fragmented) with instrument settings that promote information-rich fragmentation spectra

Statistical models, inspired by text modelling are used to find groups of fragments (and neutral losses) that occur regularly throughout the data (we call these mass2motifs). These mass2motifs are indicative of functionally relevant molecular substructures (see proof of concept study described below). Molecules sharing a mass2motif are therefore potentially functionally related. The proposed workflow requires no a-priori knowledge of the relevant molecules and substructures in a particular experiment, removing any bias.

Perform statistical analysis on the change in presence (and abundance) of mass2motifs rather than individual metabolites. Significant changes in mass2motifs that are functionally relevant are more biologically relevant than changes in individual metabolites as they provide a hypothesis as to the cause of the change. For example, when a patient takes a drug, we would expect to see increased presence of metabolites related to the drug in, say, urine samples taken after drug consumption. These metabolites would likely share a mass2motif and would be collected together in our analysis. Analysis of the change in a mass2motif also increases statistical power over analysis of individual metabolites as it effectively looks at the change in a group of metabolites. We stress that the grouping of metabolites is based purely on the fragment patterns present in the data and is therefore not biased towards known metabolites or metabolite groups (e.g. pathways).

Our analysis doesn’t require identification of metabolites but it is likely that, once interesting mass2motifs have been extracted we may wish to identify metabolites that share them. This can be done in the standard manner (although now we will be identifying a smaller set of metabolites) and the mass2motifs themselves can be very helpful in doing this by so-called substructure constraints: for one elemental formula many atomic configurations are possible, usually in the range of 100s to 1000s, but given one or more substructures that have to be present in the metabolite, this will reduce considerably to 10s or even less. Fragments and losses are generally straightforward to identify (there is often only one feasible combination of atoms that can describe their low masses). In many cases this can lead to identification of the mass2motif itself. Each metabolite ……?

## Programme and Methodology

Programme and methodology

Identify the overall aims of the project and the individual measurable objectives against which you would wish the outcome of the work to be assessed. This should refer to the objectives set out in the proposal form (Objectives section).

Detail the methodology to be used in pursuit of the research and justify this choice

Explain why the proposed project is of sufficient timeliness and novelty to warrant consideration for funding

Describe the programme of work, indicating the research to be undertaken and the milestones that can be used to measure its progress. The detail should be sufficient to indicate the programme of work for each member of the research team. Explain how the project will be managed

References should appear in a list at the end of the case for support and be linked to relevant text by, for example, sequential numbering and superscript reference numbers embedded in the body of the document. Only one publication should be listed for each number. Within the list of references, URL links to relevant publications or online resources are permissible. The case for support should be a self-contained description of the proposed work with relevant background, and should not depend on additional information. Applicants must not include URLs to web resources in order to extend their case for support. The inclusion of such URLs will result in your application being withdrawn for correction. Peer reviewers are advised to base their assessment solely on the information contained within the application, and instructed not to access external links.

What we are going to do (how do we judge success??):

1. [CS PDRA & SR & RD]Software [success = it existing and usable and people using it]
   1. New, efficient implementation of MS2LDA ready for big (e.g. clinical) data sets
   2. Web front-end including viz that can handle different experiment types and data types
   3. Incorporate into PiMP (other tools too?...lots of work. Standalone + PiMP probably enough (and we’ve kind of already done it)). E.g. for pooled files and MS2LDA we should use PiMP as the platform to extract the DE MS1 peaks.
   4. Automatic DB searching of Mass2Motifs and fragmented metabolites
   5. Aim to include as many spectral databases/data as we possibly can?
   6. Library of annotated topics
2. [JUSTIN & KARL]Investigating / demonstrating utility. Via case studies – how do we define success?
   1. LC-MS DDA. Drugs/urine, what else? Demonstrate diversity. Both individual file MS2LDA, MS2LDA with pooled data, Multi-file LDA. (in house – Glasgow data). Detecting drugs (think Mike was suggesting keeping that up our sleeves), viewing the effect of drugs
   2. Jacques data (letter of support)
   3. Lipids (Birmingham)
   4. Proteins (riskiest, but worth trying…Andy Dowsey?)
3. [CS PDRA & SR & RD]Algorithm development
   1. Multifile LDA. Low risk, but needs new code and data to try it out on.
   2. Hierarchical LDA – topics become hierarchical. I think this would be very cool. Might provide us with some insight into the hierarchical nature of fragmentation (Canada data?)
   3. Advanced topic definitions – CFM-ID??? [high risk]
   4. Incorporating text from e.g. patient records – META DATA
   5. Putative ID from substructure ID? i.e. using LDA to spread identifications….molecular networking does this.

### Project aims

### Methodology

### Programme of work

Innovation

We are proposing a completely new and innovative method for LC/MS analysis of large metabolomics experiments. This is based on a shift from performing analysis in the metabolite space to analysis in the fragment space. Global fragment analysis has not been performed before, with fragments only being used to identify individual metabolites.

Whilst there has been work that uses Machine Learning (ML) to analyse fragmentation data (see e.g. Bocker CSI), we are the first group to propose the use of advanced text processing techniques within metabolomics analysis. And the first to propose the joint analysis of the complete fragmentation dataset (previous work has looked at each spectrum in isolation). Aside from recent ML approaches, much metabolomics analysis is performed using ad-hoc methods with little formal statistical justification. Crucially, these methods are scalable (they have been used on text corpora far larger than current metabolomic datasets), and using a larger set of fragmentation spectra will only result in better defined Mass2Motifs

Our proposed methods are unbiased. They do not rely upon already identified metabolites, instead using patterns of fragment and loss co-existence within the dataset. This is a highly innovative step forward from current approaches. Standard analysis in the metabolite space requires a database of fragment patterns that are highly equipment specific. If a metabolite is not in the database it cannot be identified. When new equipment is developed, there is a long delay before databases are updated with new standard profiles. Our proposed method can be used immediately on data from new equipment.

Data reduction: focus on relevant patterns AND data size (Mbs)

Visualization: completely novel way of looking at fragmentation files!

Sustainability

This proposal concerns the initial development, testing and deployment of our proposed methods. We anticipate this creating a new research area within the University and the UK with three strands:

The text analysis literature includes many extensions to the algorithms we have converted to this domain. Many of these can be used to improve our initial algorithms. There will also be novel metabolomic-specific extensions that can be explored. We anticipate a subsequent proposal (to e.g. EPSRC) to look exclusively at the statistical and machine learning developments that can improve our basic mass2motif model.

We anticipate our primary user base to be researchers involved in therapeutic development and assessment. Although we believe our initial model will be usable across a wide range of problems, the nature of the LC/MS data collection and statistical analysis means that neither can be considered ‘black box’ systems that can be operated by anyone. Therefore, any serious attempt to apply these techniques to new disease domains will require expertise in the data collection, algorithmic processing, and results interpretation. We see many potential application areas across the University (and beyond) that would be eligible for funding to cover these staff and consumables costs from both RCUK and charities.

As mentioned above, the mass2motifs themselves can also be used as an aid towards metabolite identification. We will pursue this as a research direction in its own right, initially investigating how the use of the substructures implied by particular mass2motifs can aid in identification (i.e,, substructure constraints). The novelty here lies in the use of automatic substructure discovery, which removes the machine specific properties of substructure and fragment patterns available from databases. This leverage of the information present globally within a dataset has not been used for identification before and could provide useful constraints on potential identifications (metabolites sharing a mass2motif are likely to have structures that share a substructure and the confidence in particular identifications ought to reflect this).

EF assignment? We can still do some nice stuff in that field – by ‘combining the SIRIUS approach with our EF assigner….’

Impact

Improving the process through which drugs are developed and tested will have impact across clinical and medical sciences and, subsequently, the public. Glasgow has key clinical strengths in the analysis of disease X, disease, Y and disease Z responsible for XXX million deaths per year. We believe our approach will reduce the time taken to identify suitable drug targets and understand their effect through trials. Reducing this time will save lives and improve quality of life for those suffering from disease.

We also believe that there is utility in our approach for clinical diagnosis and assessing patient conditions. For example, the presence of a mass2motif linked to a particular drug will allow doctors to objectively assess whether the patient is taking medication that they ought to be (a key problem in disease YYY). Similarly, changes in the metabolome explainable via mass2motifs linked to other conditions (diseases, smoking, alcohol, etc) can be easily removed from analysis, revealing only the changes of interest.

What do we want

A PDRA for technical things.

Justin time.

Money to measure samples existing for other projects.

Notes

Q: Why not just create a fragment x sample matrix?

A: because we cannot combine fragment intensities across parent ions due to the differing ionisation affinity. Also, some framgents are part of several mass2motifs; and that would be hard to extract from such a matrix.

Q: Isn’t it time consuming to identify all of these mass2motifs (presumably they need to be identified ultimately).

A: Yes, but in some cases we can use already known ones (we’ll maintain a DB) and in other cases it will be quicker than identifying all of the metabolites. Also, the number of mass2motifs is much smaller than the number of metabolite features in the data.

Q: Should we add things about text from patient records?

A: Maybe.

Q: What if people don’t want to do their entire analysis as DDA - it’s a bit risky and we know that the parent ion detection and intensities won’t be as good.

A: We can do many of the steps through a standard analysis and a pooled fragment file. We lose the change in number of molecules with particular mass2motif, but we can still assess the differential expression of mass2motifs. There will be fewer mass2motifs as we won’t be able to match as many MS1 peaks to MS2 scans, and there won’t be as many because we simply won’t fragment as many peaks in the pooled sample as we would across all samples.

Je-S Instructions

BBSRC recommend that you use typefaces Arial, Helvetica or Verdana and a strict minimum font size of 11 must be used for the entire Case for Support, Justification of Resources and CVs (excluding text on diagrams and the use of mathematical symbols). A minimum of single line spacing and standard character spacing must be used. Margins must not be less than 2cm. Applications will be checked for faults by BBSRC Administrative staff soon after the closing date to ensure that relevant aspects of the application are legible and comply with the formatting rules. Any component(s) of an application which do not meet these rules will be returned for amendment before being validated for peer review. A late response in amending returned elements of the application will result in the application being withdrawn from the round.

Case For Support and Track Record

The page limit for the combined track record and case for support is maximum 8 sides of A4.

Proposals exceeding the 8 page limit, or not adhering to the specified format, will not be considered. More specific instructions relating to format of the Case for Support may be detailed for some strategic programmes. These will be set out in the relevant call for proposals.

Previous research track record (suggested one to two pages within the overall eight page limit) should:

Provide a summary of the results and conclusions of your recent work in the technological/scientific area which is covered by the research proposal. Include reference to both BBSRC funded and non-BBSRC funded work. Details of past collaborative work with industry and/or with other beneficiaries should be given.

Indicate where your previous work has contributed to the UK's economic competitiveness or to improving the quality of life.

Outline the specific expertise available for the research at the host organisation and that of any associated organisations.

The information provided should relate to all applicants involved in the project.

Case for support which (suggested up to seven pages within the overall eight page limit) should provide a description of the proposed research and its content.  Lists of references and illustrations should be included in the page limit and should not be submitted as additional documents or as an annex. The description should include the following sections:

Background

Introduce the topic of research and explain its academic and wider context

Demonstrate a knowledge and understanding of past and current work in the subject area both in the UK and abroad

Note: Justification of Resources should be completed as a separate item. See below:

CVs

Proposals without CVs for applicants and named research staff, or with CVs that exceed the 2 page limit, will not be considered.

CVs are required for all named applicants and named research staff only. These must be no more than 2 sides of A4 per person and should be submitted as attachment type CV. The CV should include details of

Employment history (give dates and details of position held including the nature of your current employment)

Qualifications (state subject, class of degree with university dates)

Patents

Most recent publications, within the last 5 years, in refereed journals relevant to the project.

Please note that any lists of publications should be included within the CV and not submitted as a separate document. Separate lists of publications and other unsolicited documents will not be included in the peer review process.

Letters of Support

Letters of support should be submitted as attachment type ‘Letter of Support’ with no limitation on page length. Letters of support must be included to confirm an active collaboration or contribution to a project in terms of resources or expertise, and may be included where a statement from a third party is necessary to enable the informed assessment of a proposal. Applicants are asked to note that members of an institution which has provided a letter of support will not in general be used as referees for that proposal. Therefore, including more than a few carefully chosen letters can be detrimental to the peer review process. Letters of support are required for the following types of application;

New Investigator scheme – a letter of support from the Head of Department describing the financial contribution from the institution to the start-up of the applicant’s laboratory.

Applications with project partners – a letter of support from each project partner, confirming its support for the research, confirming any financial contributions to be made, and outlining the expected benefits to the organisation.

Proposal Cover Letter

Inclusion of a cover letter is optional. Letters should be submitted as attachment type ‘Proposal Cover Letter’ with no limitation on page length.

Applicants may use the cover letter to list reviewers that they would prefer BBSRC do not approach, but BBSRC reserves the right to make the final selection.

Facility Form

If facility access is being requested (Primarily TGAC and HECToR). Failure to include the required forms will result in withdrawal of the proposal.

Final/Interim report

Principal Investigators and Co-Investigators on an application must submit an interim or final report on any related BBSRC research grant that they have held or completed in the last twelve months (excluding those under six months old and training grants) on which they have been the Principal Investigator. The interim report should use the form on the [BBSRC website](http://www.bbsrc.ac.uk/nmsruntime/saveasdialog.aspx?lID=1581&sID=440) and the summary report section should be a maximum of two sides of A4. Submit as ‘Final/Interim Report’

Diagrammatic workplan

The workplan is mandatory with a maximum one side of A4.  Submit as ‘Workplan’.

Justification of Resources

The Research Council guidance notes for the completion of the Justification of Resources attachment in JeS.  Details are available [here](https://je-s.rcuk.ac.uk/Handbook/pages/GuidanceonCompletingaStandardG/CaseforSupportandAttachments/JustificationofResourcesCrossC.htm).  
  
Pathways to Impact

The attachment should be a maximum of 2 sides of A4, minimum sans serif font size 11 and entitled “Pathways to Impact”.

The Pathways to Impact attachment should:

be project-specific and not generalised,

be flexible and focus on potential outcomes,

Researchers should be/are encouraged to:

identify and actively engage relevant users of research and stakeholders at appropriate stages,

articulate a clear understanding of the context and needs of users and consider ways for the proposed research to meet these needs or impact upon understandings of these needs,

outline the planning and management of associated activities including timing, personnel, skills, budget, deliverables and feasibility,

include evidence of any existing engagement with relevant end users.

The Proposal must include A Pathways to Impact attachment. If you are unable to provide one, the attachment document Type should be used to fully justify why this is not possible.

The RCUK position and expectations document on the RCUK website outlines considerations that funding recipients would be expected to undertake: http://www.rcuk.ac.uk/innovation/policies/

Further background information on Pathways to Impact is available on the RCUK website: <http://www.rcuk.ac.uk/innovation/impacts/>

Head of Department statement

The Head of Department statement is generally optional, but must be included in certain circumstances (see the BBSRC Grants Guide).  Statements should be submitted as attachment type ‘Head of Department Statement’ with no limitation on page length.

Data Management Plan

Please include a statement on data sharing as attachment type ‘Data Management Plan’.  A maximum of one side of A4 is allowed for this and must not be used for any other purpose. This statement must clearly detail how you will comply with BBSRC’s published Data Sharing Policy, including concise plans for data management and sharing as part of research grant proposal, or provide explicit reasons why data sharing is not possible or appropriate.  
  
The policy, and detailed guidance notes, can be viewed at <http://www.bbsrc.ac.uk/web/FILES/Policies/data-sharing-policy.pdf>   
Comprehensive data sharing plans will be expected, in particular, in the “data sharing areas” highlighted in the policy. More succinct plans may be appropriate for applications outside of these areas.  
  
Data sharing plans may include details of:

Data areas and data types - the volume, type and content of data that will be generated e.g. experimental measurements, records and images;

Standards and metadata - the standards and methodologies that will be adopted for data collection and management, and why these have been selected;

Relationship to other data available in public repositories;

Secondary use - further intended and/or foreseeable research uses for the completed dataset(s);

Methods for data sharing - planned mechanisms for making these data available, e.g. through deposition in existing public databases or on request, including access mechanisms where appropriate;

Proprietary data - any restrictions on data sharing due to the need to protect proprietary or patentable data;

Timeframes - timescales for public release of data;

Format of the final dataset.

Applicants may claim justifiable costs associated with data sharing activities, which should be captured in the application proforma and in Justification of Resources statement.

Important - This page should be used only for the statement on data sharing. Any information included other than that relating to data sharing statement requirements, as prescribed above, will result in your application being rejected. Only one statement is required per project.

Please note that preliminary data and descriptions of the proposed work belong in the Case for Support and should not be included in the data sharing statement.

Some text from the BBSRC [their highlights]

Data driven biology

Background

This priority falls under the enabling theme 'Exploiting New Ways of Working'.

World-class bioscience is critically dependent on new computational technologies, methodologies and resources. This priority aims to encourage research that will yield the next-generation of these 'new ways of working'. Projects should focus on underpinning and enabling one of our strategic research priorities (agriculture and food security, industrial biotechnology and bioenergy, bioscience for health) or have potential generic utility across one or more broad areas of the biosciences.

Aim

The data driven biology priority aims to encourage the development of the bioinformatics tools and computational approaches that are required to extract value and generate new biological understanding from the huge volume and diversity of bioscience data now available and so underpin and enable biological research as it continues to evolve as a data intensive discipline.

Scientific scope

The complexity and scale of biological data is continually increasing and this places demands on the ability of biologists to manage and analyse data. Innovative computational approaches are needed for the integration, analysis and interpretation of new and repurposed biological data to enable bioscientists to gain value and scientific leads from the enormous quantities and diversity of data available.

For a project to address the data driven biology priority a significant focus of the work must involve the initiation or further development of advanced computational tools, resources or methodologies relevant to our remit. Projects may develop entirely new applications, employ cutting-edge computational methods to better exploit data resources, or provide innovative functionality and improvements to an existing computational tool or resource.

Under this priority, examples of broad data driven research challenges that projects might address include:

Integration, interrogation and analysis of large or complex datasets such as those generated by multiple 'omics technologies

Investigating links between phenotypic traits and variation in biological systems or processes

Extracting quantitative information from large or complex image sets

Supporting knowledge discovery from biological data, for example: developing platforms for data sharing and integration, or new data visualisation approaches

The data driven biology priority also seeks to encourage exploitation of advanced computing technologies and approaches, for example: semantic computing, high performance computing, cloud computing and text-mining. Activities that support the maturation of the biological data landscape, such as the development of community data standards, ontologies and data management tools, or enhancement and maturation of existing research software, are also incorporated within the priority.

Data driven biology complements the technology development priority by providing a focus on the computational tools, resources and methods that are essential to derive maximum value from bioanalytical or biological-based technologies. Projects that combine computational approaches with the development of data-generating bioanalytical or biological technologies, for example to enhance analysis or automate metadata generation and manipulation, are also covered by this priority.

Requirements

Proposals in data driven biology (informatics tools development) should describe how they will fulfil (an) unmet need(s) in the biosciences. It is expected that new informatics tools and resources developed under this priority area will be designed, as much as possible and practical, with end users in mind. Evidence of end-user engagement may be provided in support of applications.

Many of the most exciting advances in biology are likely to occur at the interface with other disciplines through truly multidisciplinary approaches. Proposals involving strong multidisciplinary partnerships between bioscientists and researchers in the physical sciences, engineering and information technology disciplines are therefore particularly welcome.

Projects focused primarily on the use of an existing tool or minor developments of existing tools do not fall within this priority area.

Data sharing

Proposals should comply with our data sharing policy (see related links). Proposals developing informatics tools should make such tools available to the wider user and developer community with as few restrictions as possible, ideally using open source best practices (e.g. Creative Commons or Open Source Initiative recommended licences). However, we recognise that, at times, the creators' intellectual property rights may need to be protected before any sharing takes place. Such protection should not unduly delay the release of any data or tools arising from BBSRC funding.

Pathways to impact

It is expected that proposals in the area of 'data driven biology' will provide tools, resources and methodologies of potential application to broad communities in the biosciences. As well as enabling world class bioscience proposals they may have particular relevance to one or more of our other Council-wide strategic priorities.