# 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 - if space permits!

Notes: contents of BBSRC review form - this is what reviewers are asked to comment on…

Scientific Excellence:

1. Clarity of hypothesis, aims, and objectives
2. Strengths and weaknesses of the experimental design
3. Feasibility of the work programme given the track-records of the applicants

Strategic relevance

1. Relevance to industry and other stakeholders
2. Relevance to BBSRC strategy

Economic and Social Impact

Timeliness and promise

Value for money

Staff training potential

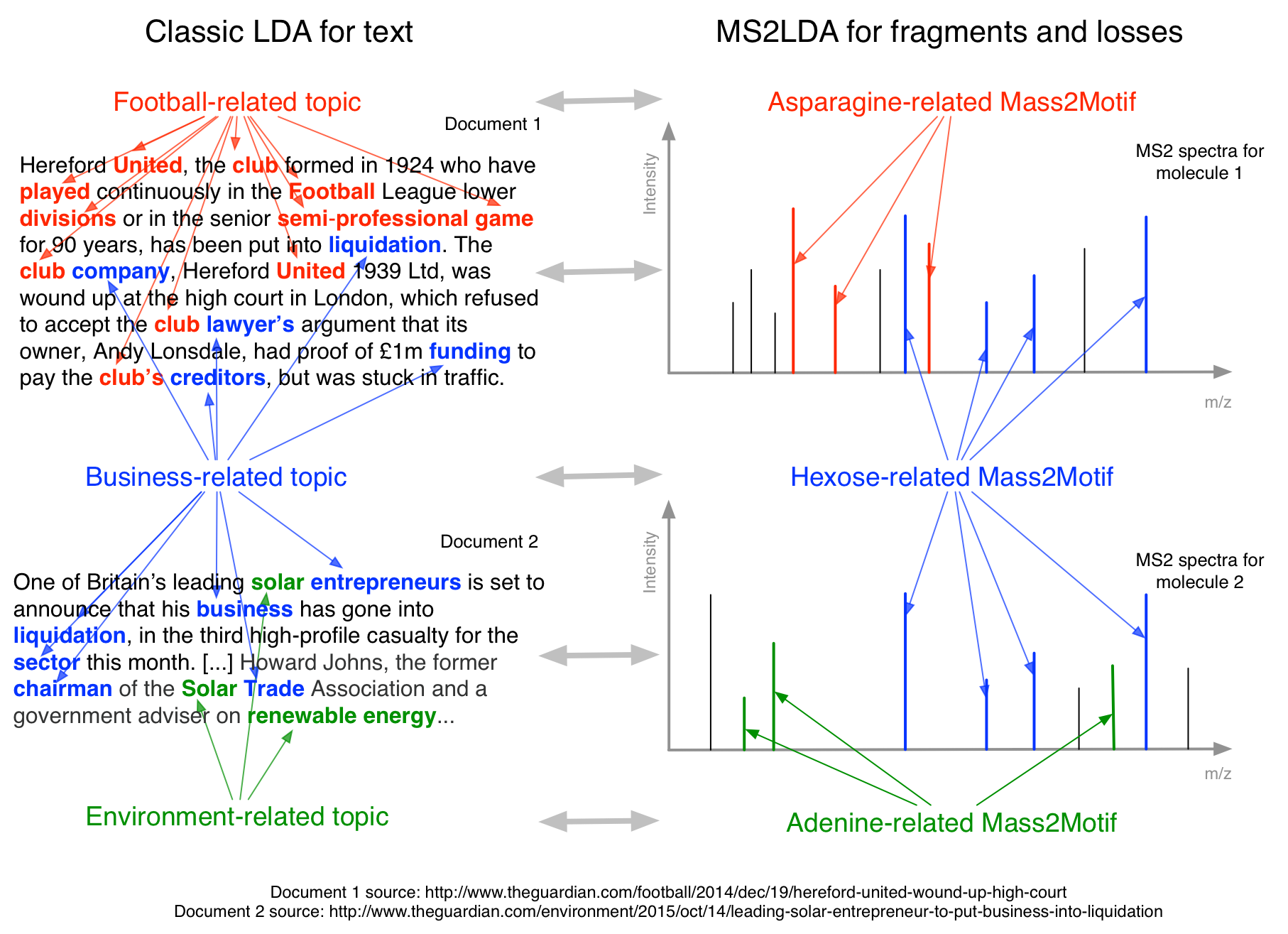
## Background

The ability to measure changes in the metabolite composition of complex samples under various perturbation conditions is crucial to many areas of medicine and the life sciences, across both academia and industry. Changes in the metabolome are often 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 investigate the links between phenotypic traits and variation in the underlying biochemistry. 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 metabolites that have been measured. The identification step represents a severe bottleneck in the metabolomics analysis pipeline — only analysing molecules that can be identified results in vast quantities of useful data being discarded, low coverage of the underlying biochemical processes and ultimately hindering the extraction of biological understanding from these datasets. For example…. There is a need for advanced computational approaches that can better make use of this data. This proposal addresses this shortcoming through ***the development of a suite of computational tools inspired by text-mining algorithms that bypass the identification bottleneck by performing analysis in fragment space****.* As such, this proposal falls within the BBSRC responsive mode priority area ‘*Data driven Biology’*.

**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 metabolites 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 the more precise definitions of these terms as used by 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.

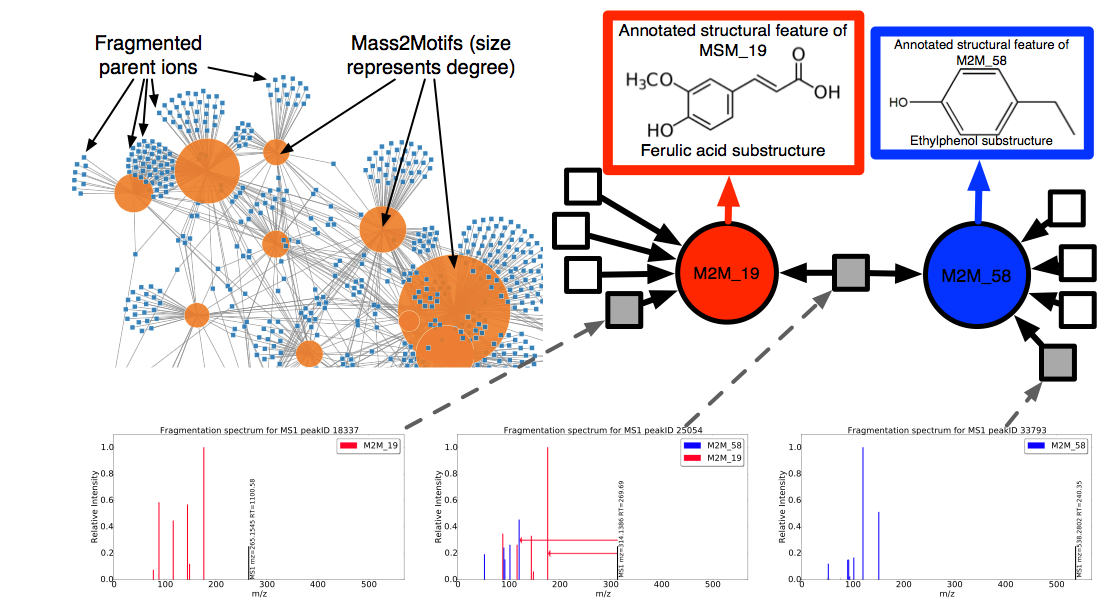
**Holistic analysis of Fragmentation Data.** Very few approaches have looked beyond using fragmentation spectra only for identification. Those that do have in common that they extract *groups* of metabolites with *similar* spectra. To produce groups of similar spectra requires a useful definition of *similarity*. 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, 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. histidine). The spectra returned have a clearly defined biochemical similarity 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!



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

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 an 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 (a real example is shown in Figure ZZ).

**MS2LDA.** By combining expertise in Machine Learning (SR) and LC-MS/MS data collection and analysis (JvdH) we have developed a prototype analysis pipeline called MS2LDA. MS2LDA is based upon Latent Dirichlet Allocation (LDA), a hugely successful text processing algorithm. LDA decomposes a corpus of documents into a set of *topics*. Each document is assumed to consist of multiple topics. The ability to index documents via a concise set of topics makes exploration of the corpus much easier. For example, documents can be searched by topics (or combinations of topics) or grouped according to the presence of a particular topic (or combination). LDA has been used in biological domains too (e.g. [][][][]). We see clear parallels between decomposition of text into topics and decomposition of metabolite fragment spectra into biochemical substructures (illustrated in Figure TT): *as documents consist of multiple topics, so molecules can consist of multiple sub-structures*. Using LDA, MS2 spectra can be decomposed into molecular substructures in the same way that documents can be decomposed into topics. MS2LDA, our prototype analysis tool for LDA analysis of DDA data does just this. MS2LDA comprises three steps: 1) data pre-processing, 2) LDA analysis, and 3) interactive visualisation. In the LDA step, substructures are automatically discovered (c.f. MS2Analyzer) as frequently concurring combinations of fragments and neutral losses, known as *Mass2Motifs*. In the same way that topics have made text much easier to explore and analyse, Mass2Motifs improve the potential for biological knowledge extraction from fragmentation data. In particular, MS2LDA offers the following improvements over current use of fragmentation data (an example of MS2LDA output is shown in Figure ZZ):



**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.

1. **It bypasses the identification bottleneck**. Mass2Motifs are small collections of fragments and/or neutral losses that are often straightforward to manually annotate. Mass2Motif annotations can be passed onto all metabolites that include the Mass2Motif in their spectra. Although not as informative as a full identification, this provides a degree of functional (or perhaps pathway, depending on the Mass2Motif in question) information sufficient for hypothesis generation. In our test cases (complex beer extracts), 70% of metabolite spectra could be annotated in this way (based on ~30 annotated Mass2Motifs) whilst only ~25% could be identified via database search (NIST, MassBank) at a fairly low threshold (>=75% spectral similarity).
2. **It automatically extracts biochemically relevant groups of fragments and losses**. MS2LDA does not require the user to specify, a-priori, which fragments and losses are relevant (c.f. MS2Analyzer) and does not require structurally labelled training data.
3. **Metabolites are grouped in a biochemically interpretable manner**. Metabolites with low spectral similarity can be grouped based on the inclusion of a Mass2Motif, which might only represent a small portion of their spectra. Annotated Mass2Motifs make groupings interpretable (e.g. ‘a group of histidine-related metabolites’) and because each spectrum can contain multiple Mass2Motifs, metabolites can be placed in multiple groups (c.f. traditional spectral clustering).
4. **Mass2Motifs indicative of changes in metabolite intensity can be identified**. When linked with MS1 data, Mass2Motifs included in metabolites with consistent intensity changes can be highlighted as possible explanations of the intensity changed. For example, the observation that a group of histidine-related metabolites are changing in intensity in a phenotypically interesting manner is exactly the kind of output desired from un-targeted metabolomics experiments. MS2LDA allows the extraction of such hypotheses from **all** spectra, not just those that can be identified via spectral matching (c.f. the current technique of identifying molecules and mapping them to pathways).

## Aims: Improved metabolomics via MS2LDA ‘fragmentome’ analysis

We strongly believe that MS2LDA can help many researchers from academia and industry efficiently extract greater biochemical meaning from their metabolomics experiments. To realise this aim, our objective is to develop, test, and deploy a suite of user-friendly MS2LDA software for this wide potential user base. Our proposed tools 1) Will be **end-to-end,** performing all steps from pre-processing to interactive results visualisation. 2) Will be able to handle **multiple experimental setups** from individual fragmentation files to complex experiments in which all samples have been fragmented. 3) Will be developed in **conjunction with end users** and **regularly tested** by them through the team’s strong links with Glasgow Polyomics. 4) Will be **user-friendly**, and accessible to a wide range of metabolomics researchers. 5) Will be **cross-platform** (both in terms of MS and software). 6) Will **integrate with spectral databases** and support standard **open data formats**. 7) Will be **open source** and freely available to researchers and developers. The following three use case sketches demonstrate the core functionality of the 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.

**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.

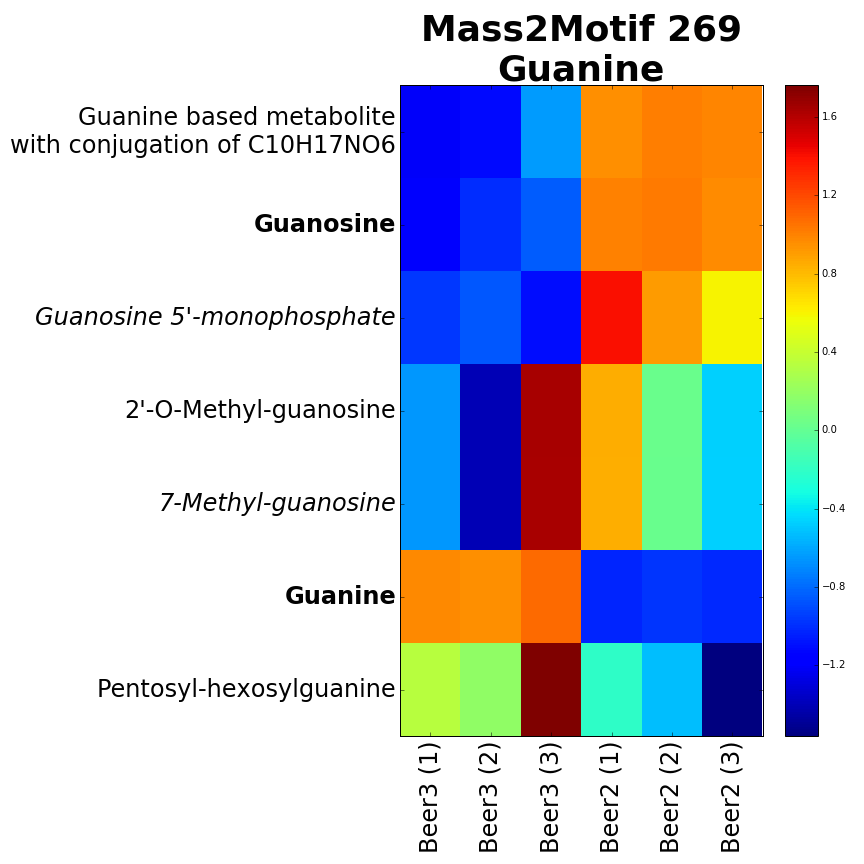


Figure GG. Molecules including a Mass2Motif identified as guanine show clear differential expression across two beer extracts (three replicates of each). Interestingly, guanine is over-expressed in Beer3 whilst conjugates of guanine are more prevalent in Beer2.

## 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.

### Programme of work (not in temporal order — see accompanying GANT chart)

**WP1: A user-friendly MS2LDA workbench (RD, SR, JvdH, Comp Sci PDRA):** Success of this project is dependent on closely working with potential users to ensure the system meets their needs. We will develop two systems, one to enable maximum user involvement in the design, and the other to ensure accessibility to a wide user base. System 1 (WP1.1) will be a Python/Django web application linked to the open source PiMP metabolomics software [E] developed by Glasgow Polyomics. The combination of the rapid development possible in Python/Django/D3, and the large and increasing PiMP user group will allow us to adopt the AGILE Software Engineering methodology and regularly obtain feedback from end-users that can be fed into the subsequent versions. The Python/Django system will be made available as part of PiMP but the barriers to installing this on a local server are likely to be too high for many users. Later in the project we will use the feedback from the rapid Python/Django development cycles to design and build a Java (for portability) standalone application (WP1.2) that will have identical functionality but can be downloaded and run easily on any modern desktop computer. We are under no illusions as to the difficulty of producing a piece of software that is robust and usable by non-computational experts. but within the team, we have the expertise to do this. RD has considerable academic and industrial Software Engineering experience and will lead this WP. RD leads the PiMP development project within Glasgow Polyomics that has recently been successfully released to a large cohort of users, to which SR contributed code, and KB provided significant design input. JvdH has experience of testing metabolomics and mass spectrometry fragmentation related software. As Python and Java are such popular languages we do not feel it will be difficult to hire a CSPDRA competent in both.

**WP2: Efficient MS2LDA implementation (SR, CSPDRA):** Rapid LDA inference via collapsed Gibbs Sampling or Variational Bayes will be at the heart of all MS2LDA analysis, particularly as we scale up to large experiments. LDA has been used on very large text corpora and we will follow the techniques used there to implement an efficient inference engine, making use of the sparse nature of the data (most metabolites do not include most fragments). This will give the new CSPDRA LDA experience that will be beneficial in other WPs.

**WP3: Efficient pre-processing (SR, JvdH, CSPDRA).** The first step of MS2LDA is the pre-processing of the data into a matrix with fragments / losses as rows and metabolites as columns. This requires matching MS1 peaks to MS2 spectra, and matching fragments / losses across spectra (and across files for multi-file LDA, WP4.2). In addition, some metabolites are fragmented more than once (due to multiple high-intensity ion-products) and it would be beneficial to filter these out. As well as our own work in ion-product clustering [], we will investigate alternatives (e.g. [RAMCLUSTR]) with the view to adding one or more method as an optional pre-processing step.

**WP4: Algorithmic development (SR, RD, CSPDRA):** There are three algorithmic directions in which we will extend our MS2LDA system to solve further problems in metabolomics.

* : Linking MS2LDA analyses to MS1 data to, for example, compute the differential expression of Mass2Motifs. This has three parts: 1) extracting fold changes from MS1 data, 2) the matching of MS1 peaks between the MS1 data and fragment data and 3) the subsequent statistical analysis of MS1 peaks corresponding to a Mass2Motif. Part 1 has been solved by many systems (e.g. [][]). We will ensure that our software can import the file formats exported by these systems (typically .csv or .tsv text files containing peak m/z, RT and intensities). Part 2 is similar to MS1 peak RT alignment and we will implement (and compare) three methods — a greedy matching algorithm, a system that implements stable matching (like that used in e.g. SIMA [A] and our previous work on MS1 alignment [B]) and one that uses the imported MS1 peak lists to directly extract the MS2 spectra from the fragment files. Many solutions for part 3 have been developed for transcriptomics (see e.g. []) and we will implement several, starting with PLAGE with which we have already had some success. 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].
* : Multi-file MS2LDA. As highlighted in use case 3, processing multiple files together will allow us to directly compare the changing abundance of Mass2Motifs *shared* across the files. In Multi-file MS2LDA, each file will have its own *distribution* over the shared Mass2Motifs. Comparison of these distributions across files will highlight how the Mass2Motifs (and hence the biochemistry) change across the files / cohorts. We anticipate that this will be particularly useful in clinical applications (and will evaluate it on clinical data, see WPBB). This is a novel extension of LDA that we anticipate being of use in other domains.
* : More sophisticated Mass2Motif models. Mass2Motifs are defined as multinomial probability distributions over fragments and neutral losses. These work well, but could be improved to more realistically reflect the fragmentation process and become more interpretable. We will investigate two alternative models. 1) In hierarchical LDA [] topics exist in a hierarchy from very abstract (perhaps individual words) to very specific (large sets of concurring words). This linking helps data exploration as groups of documents can be extracted from different levels in the hierarchy. We will translate hierarchical LDA to the MS2LDA framework with the expectation that it will help exploration here too. For example, a CO loss is common in many spectra and of interest in itself, but also interesting in concert with other fragments and losses. The CO loss might exist as a top-level Mass2Motif, connected to a cascade of increasingly specific Mass2Motifs, allowing users to dig as deeply as their biochemical interest requites. 2) 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.

**WP5: Storing Annotated Mass2Motifs (SR, RD, JvdH, CSPDRA):** Having gone to the effort of annotating Mass2Motifs it makes sense to use them again. We will enable users to annotate Mass2Motifs and make our MS2LDA software more flexible to allow it to use a combination of standard unknown Mass2Motifs to be learnt from the data and fixed, previously annotated Mass2Motifs. The latter allow rapid annotation of metabolites containing them. This will require a method for storing Mass2Motifs and their annotations. We will develop a database system (SQL due to its flexibility and compatibility with both Python/Django and Java) to store Mass2Motifs, as well as allowing for the import and export of Mass2Motifs. This latter option will require some assessment of the most suitable file format. Fragment spectra are most commonly stored in .msp and .mgf formats, but these will need to be modified slightly to handle neutral losses. The export of fragment-only Mass2Motifs in .msp and .mgf will also be supported to provide an interface with spectral databases and allow for automated Mass2Motif annotation.

**WP6: 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 demonstrating its capabilities. Our strong links with Glasgow Polyomics (KB and RD are full-time Polyomics Staff, JvdH is a Polyomics research fellow, SR is an affiliate member) provide us with access to a large number of collaborators willing to make their data available for this project. This includes fragmentation data as pooled samples are now fragmented as standard in the analysis pipeline. Additionally, we have budgeted for doing DDA analysis on XX samples, that we will obtain from collaborators in various disciplines in Glasgow (i.e. when they pay Glasgow Polyomics to run LC/MS metabolomics analysis, we can offer them DDA of all files at no extra cost). Algorithmic extensions and the Python/Django web application will be continually evaluated with this data. Initially, we already have data for the following 2 test cases:

* : Beer extracts. Using a set of previously analyzed as well as un-analyzed beer extracts we will explore the Mass2Motifs found by performing MS2LDA on each individual file as well as those found (and how they change in abundance and intensity) using multi-file MS2LDA. We have a total of 18 extracts, each of which has been analysed in full-scan (MS1) mode and DDA mode (three technical replicates in each case). Although we are not directly interested in the analysis of beer, we use it as an example of a readily available complex mixture that will highlight the performance of MS2LDA to a wide range of metabolomics specialists.
* : Clinical urine samples. We will make use of two sets of fragmented urine samples from patients who have been given antihypertensive drugs. One set has already been analysed and published by JvdH [] allowing us to compare the findings of MS2LDA with the previous analysis of this data (based on molecular networks). This will provide a test case that directly highlights the complementarity of the two approaches. The second set is previously un-analyzed and we will analyse it separately to assess the consistency in the observed Mass2Motifs.

**WP7: 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. Our core inference algorithms (WP2) 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 (WP1). JvdH will continue working closely with other researchers to help them use the tools, and discover the additional features that would make the tools more appealing to the potential users. As the Python/Django system will be linked to PiMP it will be possible for the users to use 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 other 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.

Development code will be stored in the Glasgow Polyomics git version control repository. 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. This is something that has been used successfully in the PiMP development.

**Risk mitigation**: Interdisciplinary projects such as this come with risks associated with communication. We believe this to be minimal as all team members have worked and published together (SR, RD and KB have worked together for several years, and are currently all investigators on an Innovate UK project [refxxx]). It is unlikely that we will be able to recruit a PDRA with experience in both Machine Learning and metabolomics. It is likely that the project will start with a training period. WP2 and WP3 will provide excellent training in both LDA and LC-MS/MS data. The expertise exists within the team to provide the necessary training (e.g. 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 was able to learn rapidly working alongside JvdH). Of the algorithmic development, we consider WP2, WP3, WP4.1 and WP4.2 to be low risk — our prototype system demonstrates that the system works. WP4.3 is higher risk but if the outcome of this WP is that enhanced Mass2Motif models do not improve the functionality of the system, it is not very detrimental to the project as a whole. When developing software for scientific users there is always the risk that the functionality of the software does not match the users’ expectations. Our proposed use of AGILE Software Engineering techniques minimises this risk by regularly involving users within rapid design and prototype cycles. This risk is further minimised by the expertise present in the team. Finally, there is a risk that we will not find users willing to test the system. Our links with many collaborators at Glasgow Polyomics who are unable to extract the biological knowledge that they know exists within metabolomics data gives us confidence that this is very unlikely.

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.