University of Glasgow | School of Computing Science

# Probabilistic Topic Modelling to Reduce the Noise in Metabolomics

Arijus Pleska

School of Computing Science
Sir Alwyn Williams Building
University of Glasgow
G12 8QQ

Masters project proposal

Date of submission placed here

# Contents

# 1 Introduction

This research proposal suggests an application of machine learning in biology. We intend to study the scans of low level biological entities in order to derive their patterns. To be more specific, we will be tackling a problem of noise. This issue of noisiness arise from the hardware limitation of the high precision data collection devices.

The application domain of this project is metabolomics – a branch of computional biology. In brief, metabolomics is the scientific study of chemical processes which involve small molecules – metamobilites. Ultimately, these chemical processes can be predicted by the features of metabolites. We intend to set the basis of our study on data sets acquired by *mass spectrometry* (MS). By using this technique, we obtain the masses and intensities of metabolites within a sample of interest. This data representation is known as mass spectrum. Note that a different mass indicates different types of metabolites, whereas intensity suggests the level of concentration. Thereby, a mass spectrum is used to distinguish and visualise the observed metabolites. By deriving the formulae of specific metabolites types, we can induce patterns leading to the prediction of chemical processes.

In the terms of machine learning applicability, we are looking into a problem of vision. The metabolites of some biological tissue are sequentially captured within regions of fixed size. Effectively, this process is generating an image: each scan is a pixel, therefore, the collections of scans produce an image. However, each pixel contains a large amount of metabolites, so the visualisation of these in a single image would be rather complex. Thereby, we can choose a specific metabolite type and produce pixels reflecting such metabolite activity. For a rigorous discussion on machine learning applications in metabolomics, the reader can consult the recent survey conducted by Alonso et al. [1]. Going back to the visualisation of metabolites, we can choose a specific combination (structure) and produce an image where activated pixels would suggest the concentration of the chosen metabolite structure. The problem arises upon discovering valid matabolite structures. Therefore, the techniques of machine learning, or more specifically topic modelling, are used to induce the prospective structures.

The objective to discover novel structures is treated as a problem of *unsupervised* machine learning. That is, we have no initial settings in how these structures are expected to look like: the structures will be inferred by machine learning methodology. The inference of unknown structures is studied within *topic modelling*. Further, note that we are concentrating on statistical topic modelling. In other words, we are putting the emphasis on the *probabilistic* models: the structures are inferred from the data instance distributions. To be more specific, we are focusing to enhance a particular probabilistic machine learning model – *latent Dirichlet allocation* (LDA). First of all, the successor LDA models are treated as a state-of-the-art method in tackling semantic-analysis type problems. For more detailed review on the LDA applications, the reader can consult the survey [2] conducted by Blei. Further, Hooft et al. [4] has recently addressed the applicability of LDA in metabolomics. Note that this proposal is directly influenced by the latter prospect: we will attempt to utilise the LDA-like models to reduce the impact of noise in the data sets of the metabolomics domain.

The structure of this proposal is given as follows: in Section 2 we present the issue of noise in Metabolomics; in section 3 we cover the literature in topic modeling which focuses on the variations of Latent Dirichlet Allocation; in Section 4 we provide the metric of success of the proposed research; finally, in Section 5, we go into details on how the project will be executed.

## 2 Statement of Problem

Currently, MS equipment is incapable to fully represent the metabolite construction. As discussed by Palmer et al. [3], even though the current machinery differentiates metabolites in millidalton precision, we still obtain a significant amount of false positives. Thereby, rather than relying upon the advances of MS equipment, we can investigate the prospect of enhancing the data processing techniques. As described in Section 1, LDA derivatives are promising methods in discovering the hidden structures in metabolomics data sets. For this reason, we raise a hypothesis that the degeneracies of metabolomics data sets can be detected using LDA-like models. Note that currently there is no set approach on how to configure LDA-like models to reduce the impact of noise in metabolomics data sets.

The primary goal of this research is to expand the consensus on whether topic modelling is a valid approach to smooth metabolomics data sets. To be more specific, we set our objectives to the following:

1. Utilise general topic modelling methods in metabolomics;
2. Reduce the impact of noise in metabolomics data sets;
3. Give a basis to a general topic modelling method for error correction.

Briefly, the first objective is expected to utilise the general state-of-the-art topic modelling methodology to distinguish the techniques displaying better performance in metobolomics applications. In other words, we will conduct a survey on LDA-like topic modelling methodology. The second objective is focusing on enhancing the general models for the particular task of error correction in metabolomics. Finally, the third objective can be treated as a general contribution to the field of machine learning, since the discovered techniques might be applicable to other domains.

We are setting a hypothesis that LDA-like topic modelling can be successfully utilised to correct errors in metabolomics data sets. Both cases of proving and disproving this hypothesis would contribute to metabolomics. The successful outcome would lead to the potential applications discussed in the following paragraph, whereas the unsuccessful outcome would set-up additional requirements to tackle the issue of noisiness in metabolomics data sets.

The potential applications of this research can be directly induced from the previously listed objectives. First of all, we would show that the general models or the models applied in other domains can be utilised in metabolomics. Further, the developed error correction methodology would improve the accuracy of the chemical processes prediction. Finally, the generalisation of the developed error correction model might suggest different approaches in tackling error correction in general; also, it would impact the awareness in metabolomics and the field's potential in developing machine learning generalisations.

# 3 Background Survey

The provided literature review on topic modelling is directed towards familiarising with LDA derivatives. Also, the review assesses the methodology of LDA-like models and proposes several implementation variations. The review is structured in progressing order: at the start, we look into the original LDA paper and define the terminology (it will be used throughout the proposal); then, we familiarise with two different techniques used for the inference of LDA-like models; finally, we review LDA derivatives which shows the potential to be utilised in designing the error correction technique.

## 3.1 Terminology

Recall that in topic modelling induces a hidden structure within the studied data sets. We will provide an example providing an intuition in this process. Say, if we were analysing a book, its chapters might address different topics or address the topics at a different extent. Further, the chapters itself would have unique distributions of the topics. Note that we are assuming that we do not know what the topics addressed in the book. In order to derive these hidden topics, a probabilistic topic modelling method would analyse the structure of the words used in the book. In other words, the model would induce patterns which would refer to a particular topic. Regarding the method's performance, it would depend on the method's ability to optimise the inference of the patterns. For example, commonly used words (such as and, or, the, etc.) would be relevant in all topics. Therefore, their semantic impact in defining the topics would be negligible. For this reason, we could dismiss these to improve the time performance of the method.

The terminology used in this proposal is equivalent to the terminology used by Blei [2]. Figure 1 below represents the original LDA model in plate notation. We will use it as a guide to familiarise with the terminology.

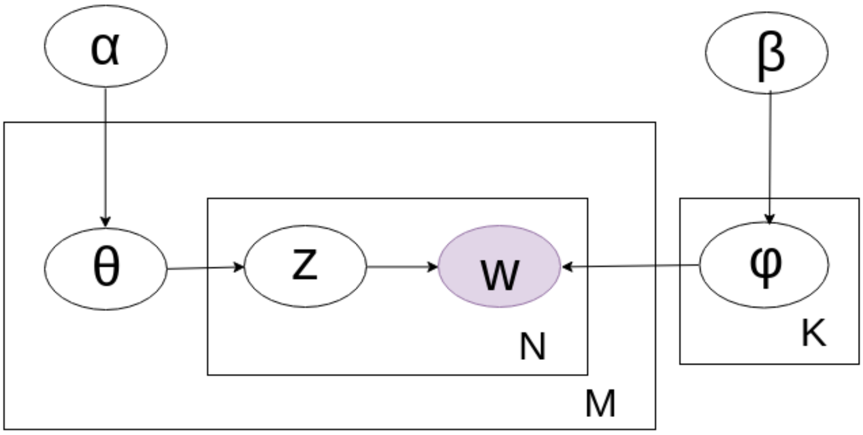

Figure 1: The design of the original LDA model

To start with, we will familiarise with the high level entities of the model. The circles represent a single entity, whereas the plates represent the number of entities. For example, there would $N \times M$ entities of $z$ and $M$ entities of $\theta$. Further, LDA is a three-level hierarchical structure: the smallest entity (grey circle) is defined as a *word*; then a collection of words (inner plate) is defined as a *document*; and a collection of documents (outer plate) is a *corpus*. Note that Figure 1 represents $M$ documents and $N$ words per document, it follows that in the corpus there would be $M \times N$ words. Also, as suggested by Figure 1, let's denote a word by $w$. Then, let's denote the sequence of words in document $d$ by

$$\mathbf{w}_d = \{w_1, w_2, \ldots, w_N\},$$

then the sequence of words in the corpus would be denoted by

$$\mathbf{D} = \{\mathbf{w_1}, \mathbf{w_2}, \ldots, \mathbf{w_M}\}.$$

Further, the grey circle indicates that an *observable variables*, whereas the white coloured circles indicate the *hidden variables.*

The grey circle represents a single word, whereas a collection of words (document) is represented by the inner plate, and the outer plate corresponds to a collection of documents (corpus). In the model, words are the only *observable variables* – this is indicated by the grey colour of the circle, whereas the white coloured circles indicate the *hidden variables.* As suggested by the figure, a *word* is denoted by $w$; a *document* (sequence of words) is denoted by $\mathbf{w} = \{w_1, w_2, \ldots, w_N\}$, where $N$ indicates the number of words in the document; and a *corpus* is denoted by $\mathbf{D} = \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_M\}$, where $D$ is the number of documents in the corpus.

Now, we will introduce the lower level entities – the hidden variables. Recall that latent Dirichlet allocation attempts to induce an arbitrary mixture of topics. The corpus and each document is expected to vary in the proportions of the induced topics; and each document is likely to have a different number of word. Let's denote the number of such arbitrary topics by $K$, and the number of unique words in the corpus by $V$. This notation will simplify the explanation on the remaining entities in Figure 1. To start with, $\beta$ is the hyper-parameter (a matrix of dimension $V \times K$) referring to the prior distribution of the topics over the vocabulary. The hyper-parameter $\alpha$ (a vector of length $K$) refers to the prior distribution of the topics over the documents. Further, the latent variable $\theta$ (a matrix of dimension $D \times K$) is the topic distribution over documents. Finally, $z$ is the latent variable on the topic assignments of each word in every document, effectively, it is a matrix of dimension $M \times V$.

## 3.2 Preliminaries

The relationship between observable and hidden variables, effectively, describes the inference of the underlying topic structure. We will go through an arbitrary example in order to give intuition in the inference process and provide the notation of the variable instances. By looking into Figure 1, we can see that $\alpha_k$ (the prior weight of the $k$th topic in a document) influences $\theta_{:,k}$ (the latent $k$th topic distribution over $M$ documents). It follows that $\theta_{d,k}$ (the probability of the $k$th topic occurrence in the document $d$) influences $z_{d,:}$ (the topic assignment over $V$ vocabulary terms in document $d$). Note that $z_{d,v}$ would refer to the assignment of some specific vocabulary term $v$. Now, by taking $z_{d,v}$ and $\beta_{v,k}$ (the prior weight of the $k$th topic on the vocabulary $v$) we describe the origins of word $w_{d,v}$. The relationship of the terms, effectively, induces the generation of the corpus, this can be expressed as the following joint probability distribution

$$p(\beta, \theta, z, w) = \prod_{k=1}^{K} p(\beta_{:,k}) \prod_{d=1}^{D} p(\theta_{d,:}) \Bigg( \prod_{v=1}^{V} p(z_{d,v}|\theta_{d,:}) p(w_{d,v}|\beta, z_{d,v}) \Bigg). \tag{1}$$

By using Bayes Theorem, we derive the computation of the conditional distribution

$$p(\beta, \theta, z|w) = \frac{p(\beta, \theta, z, w)}{p(w)}. \tag{2}$$

This equation expresses the computation of the posterior – the derivation of the hidden corpus structure given observable variable $w$. The problem arises from inefficient analytical computation of evidence $p(w)$. By reviewing the literature on latent Dirichlet allocation [CITATION], we will also familiarise with the numerical methods used in approximating $p(w)$.

## 3.3 Latent Dirichlet allocation

Latent Dirichlet allocation is a probabilistic topic modelling technique proposed by Blei et al [citation]. Since this subsection is dedicated for the origins of the technique, it will be broader in detail. This will cover aspects addressed in the following subsections covering the advances on latent Dirichlet Allocation.

The initial latent Dirichlet allocation model sets some assumptions for the data which will be processed by the model. First of all, it is assumed that both documents and words in the documents are *exchangeable*. That is, the order in which these will be processed does not matter. More formally, this assumption means that the model follows the *bag-of-words* principle. The next assumption is that the model works on discrete data. Effectively, the original paper does not cover the treatment of continuous features. Another assumption is on the dimensionality of the topic variables: the number of topics is set to be fixed during the complete run of inference.

As discussed in the preliminaries subsection, the process of latent Dirichlet allocation assumes that the corpus can be generated by a probabilistic process. The generation of the corpus is made of three levels: the parameters $\alpha$ and $\beta$ are sampled once; the variable $\theta$ is sampled once per document; and the variables $z$ and $w$ are sampled once per word. The generative algorithm given in the original paper is equivalent to Algorithm 1 below.

**Algorithm 1** Document Generation.

```
1: $N \sim \text{Poisson}(\xi)$
2: $\theta \sim \text{Dir}(\alpha)$
3: for $n \leftarrow 1, N$ do
4:     $z_n \sim \text{Multinomial}(\theta)$
5:     $w_n \sim \text{p}(w_n|z_n, \beta)$
6: end for
```

Note that the algorithm describes the generation of a single document. In order to generate the corpus, we would run the algorithm several times. To start with, in the line 1 we draw the number of words in the document $N$, where $\xi$ is an ancillary variable suggesting the mean for the Poisson Distribution. In the line 2 we draw the topic distribution $\theta$ from the Dirichlet distribution on $\alpha$. Now in the lines 3–6 we generate the words: in the line 4 we draw topic $z_n$ from the multinomial distribution on $\theta$; and, finally, in the line 5 we obtain the word $w_n$ from the conditional probability on $z_n$ and the distribution over the words in the vocabulary $\beta$.

The inference is, effectively, the computation of the evidence for the posterior distribution. The original latent Dirichlet allocation paper uses the method of variational inference. Basically, the authors introduce the variational parameters $\gamma$ and $\phi$ (computation of these set-ups a problem of optimisation). Note that the original paper provides pseudo code for solving the optimisation problem.

Additionally, the authors emphasise the estimations of the parameters $\alpha$ and $\beta$. They suggest a two-step EM procedure utilising the variational parameters $\gamma$ and $\phi$: E-step is the optimisation problem addressed in the previous paragraph; and M-step updates the parameters $\alpha$ and $\beta$. Note that the authors provide an analytical update for $\beta$ and a Newton–Raphson method for updating $\alpha$. Apart from that, the authors also introduce the issue of a sparse corpus. Such corpus would be expected upon having a large vocabulary or a significant amount of documents. The authors suggest addressing this problem by introducing smoothing. They introduce an additional parameter $\eta$ which impacts the estimation of $\beta$.

The conducted experiments display the improvement performance over the predecessor models and suggest the application domain of latent Dirichlet allocation. The first experiment addresses document modelling. The latent Dirichlet allocation model is reported to display balanced hidden topic proportions in the set of test documents. That is, the test set has replicated the proportions of the training set. The second experiment addresses document classification. In order to display better performance compared to a support vector machine (SVM) model, the authors address the dimensionality reduction of the tested documents. This means that the documents would possess a lower amount of features. It follows that the latent Dirichlet allocation model has successfully reduced the number of features and displayed an improvement in accuracy compared to the support vector machine model.

The review on the initial latent Dirichlet allocation model has set-up a basis for introducing the successor variations. These will be provided in the following sections. We have introduced the initial methodology for the following reasons: (1) the novel models will display the relaxation of the initial assumptions; (2) the provided document generation methodology will be compared to the document generation in batches; (3) we will cover different inference and parameter estimation approaches. Also, note that the paper is relatively old in time, and the experiments do not display the state-of-the-art performance. Rather than that, they were introduced to address the application domains of the model.

## 3.4 Finding scientific topics

The paper on "Finding scientific topics" [CITATION] suggests an alternative approach to the inference in latent Dirichlet allocation. To be more specific, the paper provides a readable introduction on how to apply a Markov chain Monte Carlo (MCMC) algorithm as a sampling-based method for the inference in topic modelling. Further, the authors discuss the method's application settings in vision and conduct an experiment on deriving topics from the abstracts of papers published in "Proceedings of the National Academy of Sciences" (PNAS). Finally, the authors compare the performance of the suggested sampling-based inference method to variational inference methods.

The inference method using a Markov chain Monte Carlo algorithm can be referred by *Gibbs sampling.* The authors claim that the method provides a first-order approximation in deriving the topics on the given corpus. That is, we can establish quantitative reasoning on the document correlation in terms of content and content's change over time. Recall that by $\theta$ we denote the topic distribution over documents and by $\phi$ we denote the topic distribution over words. In Gibbs sampling, these values can be obtained by examining the posterior distribution. Essentially, the variables are sequentially sampled from their respective distributions. The process draws topic assignment $z_i$ and keeps a track of this assignment with respect to each word and document. Note that $z_i$ is drawn from the following distribution:

$$P(z_i = j | z_{-i}, w) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + T\alpha}. \tag{3}$$

It follows that this distribution is utilise the previously mentioned topic assignments. For a brief familiarisation with the notation used in the previous expression, note the following: $n_j^d$ and $n_j^w$ are the counts of topic $j$ assignments in document $d$ and word $w$; $W$ is the number of words; and $T$ is the number of topics. Ultimately, the left side of the expression is $\phi$ and the right side is $\theta$. That is,

$$\begin{aligned}
\phi_j &= \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)}}, \\
\theta_j &= \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + T\alpha}.
\end{aligned}$$

This process of topic assignment would be executed for every word in a document and for all documents in the corpus. The procedure would be repeated until the distributions displayed fluctuations below some set threshold. Notice that the procedure does not require direct assignment of $\theta$ and $\phi$. If needed, the values of these can be sampled upon converge by using the previously provided representations.

The authors have executed Gibbs sampling in the domain of vision. They have represented an image as a document and each pixel as a word. This means that they were able to recover the underlying basis from which the images could be constructed. The performance was compared to the latent Dirichlet allocations models which use variational inference. Note that we define the performance in terms of perplexity, which indicates the uncertainty of the word assignment to a topic. It appears that, with relatively small data sets, Gibbs sampling displays superior performance. Further, the authors have investigated the performance upon varying the number of topics. The results indicate that there exists a trade-off between the model being able to cover the significant underlying topics (low number of topics) and start deriving irrelevant branches (high number of topics).

The main experiment was carried in the PNAS data set, note that the processed papers were published from 1991 to 2001. By having a varying notion of time, the authors were able to investigate the prevailing topics. Thus, establish a sense in dynamic topic modelling. Even though the model was not adjusted to smooth the time-series, from the derived results we can deduce the varying relevance of underlying science domains.

We have reviewed an alternative method for the inference in latent Dirichlet allocation. The idea behind the inference process was explained by considering the primary aspects: the notion of counts and the implicit use of hidden variables. Also, we have introduced the experiment settings its results. This allowed to familiarise with an example of topic modelling application and vision. Further, the experiment on PNAS has emphasised the relevance of time series

## 3.5 Dynamic Topic Models

Now we review the paper on Dynamic Topic Models [CITATION]. Effectively, the dynamic topic model can be treated as an extension of static topic models (such as latent Dirichlet allocation). The main idea behind a dynamic topic model is that it induces the topic evolution the processed corpus. That is, the documents in the corpus evolve over time. In this subsection, we review the key assumptions of dynamic topic models, go over the generative process of documents, review the suggested inference methods, and familiarise with the results of the carried experiment.

We will emphasise the assumptions given on the dynamic topic models. First of all, even though we are discussing a model of time series suggesting the use of continuous variables, the data remains categorical. The distinction between continuous and categorical data is that continuous data is strictly numeric and infinite, whereas categorical data can take only particular values. The second assumption relaxes the ex-changeability on documents. That is, the documents are expected to form a sequence. Or in other words, the corpus is divided into time-slices. Note that in such time-slice documents are exchangeable.

The generative process of the models updates the parameters $\alpha$ and $\beta$. This means that the topic distributions over documents and words change over time. The visualisation of the process is given in Figure 2 below.

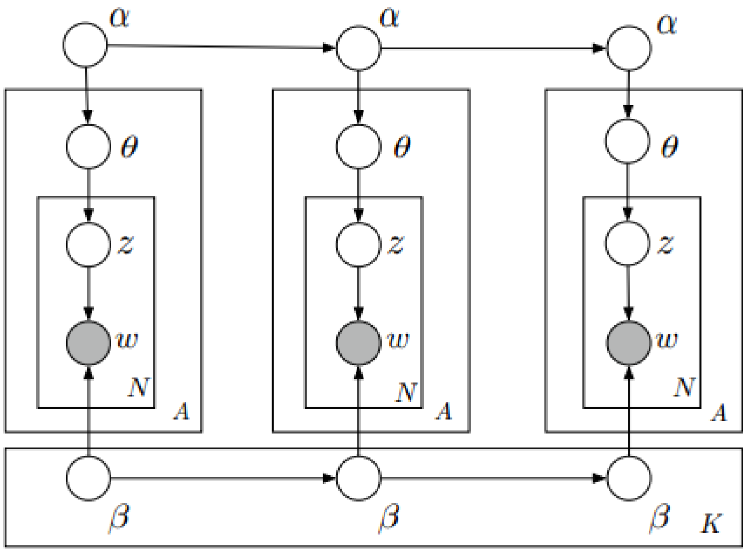

Figure 2: Dynamic topic model. COPYRIGHT.

Notice that each time-slice is equivalent to the model shown in Figure 1. Also, the parameters $\alpha$ and $\beta$ are induced by their predecessors in the previous time-slice. More formal representation of the generative process is given in Algorithm 2.

**Algorithm 2** Dynamic document generation.

```
$\beta_t | \beta_{t-1} \sim \mathcal{N}(\beta_{t-1}, \sigma^2)$
$\alpha_t | \alpha_{t-1} \sim \mathcal{N}(\alpha_{t-1}, \delta^2)$
for $d \leftarrow 1, D$ do
    $\eta_d \sim \mathcal{N}(\alpha_t, a^2)$
    for $n \leftarrow 1, N$ do
        $z_{d,n} \sim \text{Multinomial}(\pi(\eta_d))$
        $w_{t,d,n} \sim \text{Multinomial}(\pi(\beta_t | z_{d,n}))$
    end for
end for
```

The given algorithm describes the generative process for the time-scale $t$. It starts by drawing the parameters $\alpha_t$ and $\beta_t$ from the Gaussian distribution with the means given by the previous parameter values. For now, we will not discuss the meaning of the variance values. Further, for each document in the time-scale, we carry the following procedure: we draw the topic distribution in a document $\eta_d$ from the Gaussian distribution with the mean given by $\alpha_t$; then, for each word in the document, we draw the topic $z_{d,n}$ from the Multinomial distribution on the parametrised $\eta_d$, and, finally, we draw the word $w_{t,d,n}$ from the Multinomial distribution on the parametrised $\beta_t$ which is conditioned on $z_{d,n}$. By the use of *parametrisation*, we are expressing the result of the Gaussian distributions into a format suitable for Multinomial distribution. This mapping is expressed as

$$\pi(x)_n = \frac{\exp{(x_n)}}{\sum_{i=1}^{n} \exp{(x_i)}}. \tag{4}$$

The authors suggest using variational methods to approximate the inference posterior. It is claimed that stochastic simulation would be unable to scale with large data sets. Two techniques of variational methods are provided: Kalman Filtering and Wavelet Regression. The implementation details of these are provided in a brief manner suggesting the computation of the parameter $\beta_t$.

The dynamic topic model is evaluated by conducting an experiment on the journals of *Science*. The experiment is expected to establish the time series on the topic development with the articles. Indeed, the experiment is able to produce results displaying the rises and falls of the topics indicating scientific fields. Note that the experiment is performed using both variational inference methods.

The review on the dynamic topic models has familiarised us with more realistic generative process: the documents are induced by time series. Further, we have provided the visualisation of the model and introduced the algorithm for the generative process. Also, we have captured the techniques used for inference. Finally, we have familiarised with the experiment settings.

# 4 Proposed Approach

## 4.1 Overview

## 4.2 Approach

## 4.3 Risk Management

# 5 Work Plan

## 5.1 Schedule

## 5.2 Deliverables

# References

[1] Arnald Alonso, Sara Marsal, and Antonio Julià. Analytical methods in untargeted metabolomics: state of the art in 2015. *Frontiers in bioengineering and biotechnology*, 3:23, 2015.

[2] David M Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

[3] Andrew Palmer, Prasad Phapale, Ilya Chernyavsky, Regis Lavigne, Dominik Fay, Artem Tarasov, Vitaly Kovalev, Jens Fuchser, Sergey Nikolenko, Charles Pineau, et al. Fdr-controlled metabolite annotation for high-resolution imaging mass spectrometry. *Nature Methods*, 2016.

[4] Justin Johan Jozias van der Hooft, Joe Wandy, Michael P Barrett, Karl EV Burgess, and Simon Rogers. Topic modeling for untargeted substructure exploration in metabolomics. *Proceedings of the National Academy of Sciences*, page 201608041, 2016.