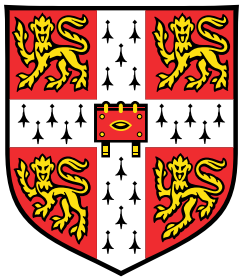
Bayesian methods for spatial proteomics



Oliver McKenzie Crook

Darwin College
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

September 2020

I would like to dedicate this thesis to those that gave me life and those that let me stand upon their shoulders.

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 60,000 words excluding appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Oliver McKenzie Crook
September 2020

Summary

Bayesian Methods for Spatial Proteomics
Oliver M. Crook

Proteins are biomolecules that govern the biochemical processes of the cell. Correct cellular function, therefore, depends on correct protein function. For a protein to function as intended, there need to be sufficient copies of that protein, it should be correctly folded into its tertiary structure and ought to be in proximity of its interaction partners, amongst many other requirements. For a protein to be in the proximity of its interaction partners, whether those be other proteins, RNA or metabolites, it needs to be localised to the required compartment. Cells from all organisms display sub-cellular compartmentalisation, though to vastly differing degrees. *E. coli*, for example, has remarkably simple sub-cellular organisation, whilst the apicomplexan *Toxoplasma gondii* has a vast number of specialised organelles.

In seminal experiments, Christian De Duve showed that upon biochemical fractionation of the cell, proteins co-fractionated if they were localised to the same organelle. These experiments led to the discovery of two organelles: the lysosome and the peroxisome, for which Christian De Duve was awarded the Nobel prize. Upon the advent of mass-spectrometry, these experiments were refashioned into high-throughput techniques with the development of Localisation of Organelle Proteins by Isotope Tagging (LOPIT) and Protein Correlation Profiling (PCP). Now these techniques have been redeveloped and a typical experiment can accurately measure thousands of proteins per experiments, whilst also providing information on (at least) a dozen subcellular compartments.

To analyse spatial proteomics data, they are first annotated with marker proteins, which are proteins with a priori known unambiguous localisations. Typical analysis proceeds by training a machine learning classifier to assigned proteins with unknown localisations to one of the compartments based on the spatial proteomics data. However, this framework holds back spatial proteomics from answering more complex questions. The first challenge is that proteins are not necessarily localised to a single compartment and so there is uncertainty associating a protein with an organelle. There is also uncertainty associated with the experiment itself, for example, reproducing the biochemical fraction and the stochastic nature of mass spectrometric quantitation. Two chapters of my thesis are dedicated to alleviating this problem by developing a Bayesian model for spatial proteomics data, with dedicated software. These approaches perform competitively with state-of-the-art classification algorithms whilst Markov-chain Monte Carlo algorithms are employed to sample from the posterior distribution of localisation probabilities. This is the basis for quantifying uncertainty in protein-organelle associations.

This Bayesian approach has several limitations, for example it still relies on marker proteins. This precludes analysis of poorly annotated non-model organisms using spatial proteomics

techniques. A chapter of my thesis is dedicated to this challenge with a motivating application to the *T. gondii* sub-proteome. Following on from this in a separate chapter, I develop a semi-supervised Bayesian model that reduces the reliance on marker proteins. The application to *T. gondii* constitutes a massive knowledge expansion revealing localisation of thousands of proteins to complex specialised niches. I also analyse the relative redundancy of the organelle sub-proteomes and the selective pressure of the host-adaptive response, revealing previously unknown insights.

The semi-supervised Bayesian approach makes use of the principle of over-fitted mixtures, currently used for data clustering, by extending it to model spatial proteomics data. Reanalysis of spatial proteomics data reveals new annotations in all datasets and allows interrogation of previously overlooked organelles. Another limitation of the approaches, thus far, is the parametric assumptions made by the Bayesian approach. One chapter is dedicated to placing the analysis of spatial proteomics in the semi-supervised Bayesian non-parametric context.

In the final chapters of thesis, I summarise the modern questions that spatial proteomics seeks to answer, including deciphering multi-localisation, change in localisation and the effect of post-translation modifications on subcellular localisation. I carefully define these problems and motivate further Bayesian models. I develop a Bayesian model to analyse differential localisation experiments; that is, spatial proteomics concerned with changes in localisation. This approach improves over currently ad-hoc methods applied to such data. I conclude with the limitations of our approach and potential solutions to the other methods.

Acknowledgements

I am thankful to my supervisors Prof Laurent Gatto, Dr Paul Kirk and Prof Kathryn Lilley for their continued and unquestionable support. I am thankful to all members of the Cambridge Center for Proteomics and MRC Biostatistics Unit. I am grateful for discussion with Claire Mulvey who introduced me to spatial proteomics. I would like to mention Konstantin Barylyuk and Ross Waller for excellent discussions and including me in their Toxoplasma project. I am grateful to John Shin and Sean Munro of the MRC Laboratory of Molecular Biology for extensive discussions. I have gratitude to all the PhD students in the Cambridge Center for Proteomics and MRC Biostatistics Unit for laughter and support. I am indebted to Gerry Tonkin-Hill for much caffeine-fuelled friendship. Finally, I want to thank my friends and family for support in all aspects of life.

Preface

The work presented in this thesis is either published or in preparation for submission to a scientific journal. The details are:

- Chapter 2 is published in PLOS Computational Biology.
- Chapter 3 is published in F1000 Research.
- Chapter 4 is in press at Cell Host and Microbe.
- Chapter 5 is in press at PLOS Computational Biology.
- Chapter 6 is in review at The Annals of Applied Statistics.
- Chapter 7 is in preparation for a biological sciences journal and parts are published in Proteomics.

A stylistic "we" is used throughout this thesis for the purpose of clarity.

Table of contents

List of figures

List of tables

Abbreviations

AIC Akaike information criterion
APEX ascorbic acid peroxidase
BANDLE Bayesian analysis of differential localisation experiments
BCC Bayesian consensus clustering
BIC Bayesian information criterion
CC cellular compartment
ChIP chromatin immunoprecipitation
CRISPR clustered regularly interspaced short palindromic repeats
CyTOF cytometry by time of flight
DOM dynamic organeller maps
DP Dirichlet processes
EB empirical Bayes
EGF epidermal growth factor
ELISA enzyme-linked immunosorbent assay
EM expectation-maximisation or electron micrograph
EP expectation propagation
ER endoplasmic reticulum
ESI electrospray ionization
FDR false discovery rate
FWER family-wise error rate
GMM Gaussian mixture model
GO gene ontology
GP Gaussian process
GRF Gaussian random field
GWAS genome wide association study
HCMV human cytomegalovirus
HDBSCAN hierarchical density-based spatial clustering of applications with noise
HMC Hamiltonian Monte-Carlo
HPA human protein atlas
hpi hours post infection
hyperLOPIT hyperplexed localisation of organelle proteins by isobaric tagging
IMC inner membrane complex
indel insertion-deletion
iTRAQ isobaric tag for relative and absolute quantitation
KL Kullback-Leibler

KNN k-nearest neighbours
KO knock-out
KS Kolmogorov-Smirnov
LC-SPS-MS3 liquid chromatography synchronous precursor selection mass spectrometry
LOPIT localisation of organelle proteins by isobaric tagging
LOPIT-DC localisation of organelle proteins by isobaric tagging after differential centrifugation
MAP *maximum a posteriori*
MCMC Markov-chain Monte-Carlo
MDI multiple dataset integration
mESC mouse embryonic stem cells
MH Metropolis-Hastings
MR movement reproducibility
mRNA messenger ribonucleic acid
MS mass-spectrometry
MSE mean square error
NIW normal inverse Wishart
OOP object-oriented programming
PC penalised complexity
PCA principal component analysis
PCP protein correlation profiling
PCR polymerase chain reaction
PEAR posterior expected adjusted Rand index
PM plasma membrane
PSM posterior similarity matrix or peptide spectrum match
PTM post-translational modification
PV parasitophorous vacuole
RJ reversible-jump
RKHS reproducing kernel Hilbert space
RPPA reverse phase protein arrays
SILAC stable isotope labeling using amino acids in cell culture
SNP single nucleotide polymorphism
sgRNA single guide ribonucleic acid
SVM support vector machine
TAGM T augmented Gaussian mixture
TGN *trans*-Golgi network
TL transfer learning
TMT tandem mass tag

tp,fp,tn,fn true positive, false positive, true negative, false negative
t-SNE t-distributed stochastic neighbor embedding
VI variational inference
WLLN weak law of large numbers

Chapter 1

Introduction

1.1 The post genomic era

The human genome was sequenced almost two decades ago, in what was a marvel for biological sciences [255]. The human genome has illuminated our understanding of human development, physiology, medicine and evolution. What followed is commonly referred to as the GWAS (genome-wide association study)-era [148]. This period saw, one study after another, the link between genetic variability and disease phenotype unravel [193]. However, these associations tell us little of the function of the gene products of interest [447].

The number of proteoforms, specific molecular forms of a protein arising from a specific gene product, vastly outnumber the number of genes because of alternative splicing and post-translational modifications (PTMs) [3]. A major goal of biology is to determine the function of all these proteoforms, though it is not clear that all of them are, in fact, functional. This task is also complicated by the observation that each protein may carry out more than one function - they moonlight [231]. Furthermore, whilst the genome is somewhat constant, proteomes can differ from one cell to another even amongst the same cell type, there is tissue specific variability and proteomes are not static with respect to time.

The desire for high-throughput deconvolution of protein function has led to intense biotechnology development [20]. There are many lenses with which we can look at protein function, for example via protein abundance, via protein interactions, via protein structure, via subcellular localisation, via thermal stability. Each assay provides a different perspective on the proteome, eventually allowing us to pinpoint protein function.

The central advancement of these biotechnologies is the invention of electrospray ionization [127] and the Orbitrap mass analyzer [211], which celebrate their 31st and 21st anniversaries at the time of writing. These advancements, along with others, spurred more interest in proteomics [2]. Now a mainstay of the biological community, mass spectrometry-based proteomics experiments can now measure thousands of proteoforms per experiment. These

advancements allow highly sophisticated functional proteomic experiments [3]. Though, as with the sequencing of the human genome, these advancements come with the challenging task of how to analyse the resultant data.

1.2 Proteomics

Proteomics is the study of proteins on a system-wide scale. Practitioners of proteomics are interested in all aspects of proteins and the deconvolution of protein complexity. Two-dimensional gel electrophoresis was possibly the first technique that could be considered proteomics [347, 9]. Many original analyses used the transcriptome as a proxy for understanding the proteome [179]. However, this provides an unsatisfactory picture because not all mRNA is translated, proteins are degraded at different rates and there is an increasing appreciation for the role of post-transcriptional regulation [23, 277].

There are many approaches to proteomics and proteins can be detected using immunoassays or mass spectrometry. Other common techniques used by biochemists include enzyme-linked immunosorbent assay (ELISA) [1], western blotting, reverse phase protein arrays (RPPA) and, even earlier, the use of Edman degradation allowed proteins to be sequenced [260]. However, mass spectrometry with the use of electrospray ionization and the Orbitrap coupled with nano liquid chromatography has been a driving force in increasing the throughput of proteomics [20].

1.3 Mass spectrometry

At its core mass spectrometry is a method to measure the mass-to-charge ratio of ions. Typically, the readout from a mass spectrometer is the mass spectrum - the intensity plotted as a function of mass-to-charge ratio. Typically, ions are identified by comparing this mass spectrum to *in silico* fragmentation patterns [268]. The analyte must first be ionized and in most biological samples this is performed by electrospray ionization (ESI) [2].

The electrospray disperses the liquid analyte by generating an aerosol [128]. More precisely, the electrospray emits a jet of liquid droplets, which are subject to high voltage. One can usually observe the so-called Taylor cone - a cone of liquid at the edge of the electrospray capillary. A fine jet of droplets emanates from the cone and the solvent rapidly evaporates. This process causes the liquid droplets to become progressively more charged. Finally, a phase transition occurs at the so-called Rayleigh limit (the theoretical maximum amount of charge a liquid droplet can hold), at which the droplet dissociates leaving a stream of positively charged ions [480].

In brief, the ions enter the mass spectrometer through a quadrupole [279]. Using oscillating electric fields, only the ions in a certain mass-to-charge ratio range are passed through the system,

which is comprised of a series of chambers in vacuum. After entering the mass spectrometer, ions are transmitted, captured and fragmented in a variety of different devices including quadrupoles, hexapoles and ion traps. In some mass-spectrometers, these ions then enter further chambers one of which contains the Orbitrap [211]. The Orbitrap contains a spindle-like electrode which holds the ions in orbital motions around the central spindle. This is due to the balancing of the electrostatic attraction of the ions to the electrode with the inertia of the ions themselves. Hence, the ions trace out elliptical trajectories around the electrode. Using electrostatics a quadro-logarithmic potential is generated, resulting in the ions moving back and forth along the central spindle. Viewed in three-dimensions, one would observe a helical like motion around the central electrode [290]. This motion is harmonic and only depends on the ions mass-to-charge ratio [290]. The angular frequency is governed by the well known equation $\omega = \sqrt{\frac{k}{m/z}}$, where k is the force constant of the potential. This process generates a waveform or *image current* that can be measured. The Fourier transform of the image current can be converted into the mass spectrum [295].

1.4 Mass-spectrometry based proteomics workflows

In a typical proteomics workflow the quantities of interest: proteins or proteoforms, are not directly measured. Indeed, proteins are first proteolytically digested to peptides using an enzyme. For example, a trypsin digest generally cleaves proteins at the C-terminal side of the residues lysine and arginine, except when either is bound to a proline on their C-terminal side [392]. Measurement of (semi)-tryptic peptides is a surrogate for the protein from which the peptides have been derived. To perform peptide identification the peptides are first fragmented. This ion fragmentation creates a series of nested fragments, the masses of which are measured and search engines are employed to achieve identification [444, 48, 78].

For quantitative proteomics a number of different methods are used. In the label-free strategy, commonly referred to as LFQ, peptide quantitation is given by the integral under the spectral peak, which is assumed linearly proportional to the concentration of the protein in the sample. For isobaric tagging methods such as tandem mass tags (TMT) peptides are tagged using a chemical tag. Each tag is isobaric but the reporter group is sample specific. Thus, when the tag is fragmented from the peptide inside the mass-spectrometer a unique reporter ion signature is observed in the low m/z area of the mass spectrum. An *in vivo* strategy is to use stable isotope labelling using amino acids in cell culture (SILAC). In this approach, cells metabolically incorporate heavy or light amino acids from their growth media. Thus, in the heavy sample all peptides are heavier, by a known amount, than their lighter counterparts. This difference can be differentiated in a mass-spectrometer and the ratio of the peak intensities

in the mass spectrum is assumed to reflect the abundance ratio for the two peptides. See Pappireddi et al. [354] for a recent summary and review.

1.5 Spatial proteomics

Compartmentalisation and localisation are ways of life. Biological organisms display compartmentalisation in a multi-scale fashion. Humans have organs: the heart, for example, has several chambers and each heart cell is subdivided into complex organelles and subcellular niches. Proteins are distributed amongst these subcellular niches in accordance with their function. Thus determining a protein's subcellular localisation is a key part in the process of pinpointing a protein's function. Subcellular localisation can be studied either using imaging approaches or mass spectrometry based methods, each of which can be further subdivided [286].

1.5.1 Fluorescent microscopy

Imaging based spatial proteomics are a set of methods that allow for the visualization of proteins *in situ* [455]. These approaches do not require cell lysis and can obtain single cell information. This allows the visualisation of cell-to-cell variability in protein subcellular localisation. The visualisation of proteins themselves however is a daunting task, requiring either an antibody to the target protein or by expressing a fluorescent protein fusion [72, 264]. The process of generating antibodies and genetically modified proteins limits the throughput of imaging approaches [286]. The process is time intensive and also expensive. There are also uncertainties around the specificity of the antibody to the target protein - with frequent cross reactivity with nuclear proteins observed [407, 430]. In all, imaging based approaches are useful, however they will currently remain in low throughput. Lack of reproducibility in antibody-based localisation has contributed to a reproducibility crisis [16].

1.5.2 Proximity labelling

Mass-spectrometry based spatial proteomics can be performed using proximity labelling [42]. Here, bait proteins are tagged with an enzyme, such as ascorbate peroxidase (APEX) [460] or a biotin ligase (BioID) [398]. These enzymes catalyse the production of activated biotin, which results in the biotinylation of accessible lyseine residues on proteins in close proximity. It is also possible to use protein engineering to target the tagged protein to an organelle of interest. In addition, the use of multiple baits can provide additional information and reduce false positives [286]. Mass-spectrometry is used to identify the proteins with increased biotinylation in comparison to the background. However, again the tagging may result in artefacts, the information per bait is minimal, the method is low in throughput and relies on accessible lysines in close proximity [460].

1.5.3 Subcellular fractionation coupled to mass spectrometry

High-throughput mass spectrometry based methods that provide a holistic view of the spatial proteome are possible by coupling subcellular fractionation with mass spectrometry [135, 118]. The key idea rests on observations made by Christian de Duve in a series of experiments [100, 95, 97, 98]. De Duve developed the following technique. First, cells are gently lysed in a fashion that maintains the integrity of the organelles. Then cellular content is fractionated using centrifugation, with each fraction differentially enriched for different organelles. Using this principle de Duve was able to localize enzymes to the cellular structure with which they are associated, by correlating the relative enzyme enrichment with known organelle properties. At the time this was performed by measuring the activity of the enzymes of interest in each fraction. These experiments led de Duve to discover the lysosome and peroxisome, along with cataloguing the properties of a large number of enzymes [96, 94, 99, 101].

Modern inceptions of this experiment begin in much the same way. After gentle cell lysis, the cellular content is then fractionated using either density gradient centrifugation or differential centrifugation. Depending on the experimental design, the fractions are then collected and possibly multiplexed using isobaric labelling reagents [118]. These samples are then subject to quantitative analysis using a mass spectrometer. This method can also be performed label-free with the caveat of excessive missing values [135]. An overview of the approach is given in figure 1.1.

Whilst providing a cell-wide view and being high-throughput, this approach requires extensive data analysis and relies on marker proteins with localisation known prior to any experimentation [154]. This thesis will focus on the challenging task of statistical and machine learning analysis of this flavour of spatial proteomics data. More detail on the experimental approach is introduced throughout the thesis as it becomes necessary.

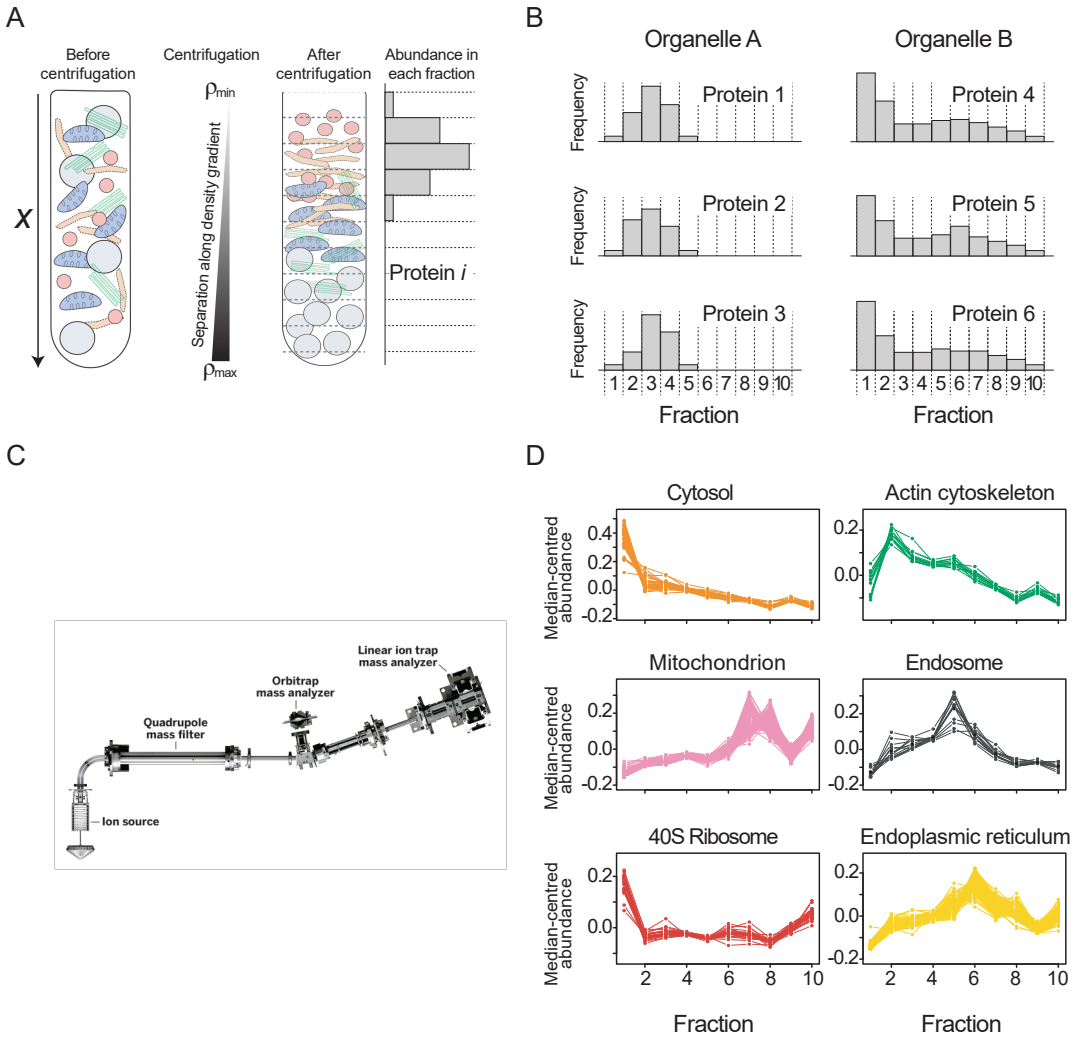


Fig. 1.1 An overview of profiling spatial proteomics methods. (A) After gentle cell lysis, cellular content is loaded onto a preformed iodixanol density gradient. The tube is then subject to centrifugation, typically at $10^6 g$ for 8 hours. After centrifugation organelles have migrated to their buoyant densities and proteins localised to these organelles will be more abundant in that part of the density gradient. (B) Discrete fractions are collected along the density gradient. Proteins localised to the same organelle share characteristic distributions across the fractions. (C) After multiplexing, fractions are analysed by mass-spectrometry. (D) Proteins with *a priori* known localisation are annotated. Proteins from the same sub-cellular niche share the same (median-centered) abundance profiles.

1.6 Statistical inference

There is no one approach to statistical inference. Though likelihood and frequentist approaches are straightforward and informative, they do have limitations. When we bring together different datasets and desire the quantification of uncertainty the Bayesian paradigm is often more amenable to answer the question of interest [232]. Furthermore, in the presence of many latent variables the Bayesian framework offers a clear approach. The development of a generative model for the data also allows tools for model criticism and simulation to be used [29]. The thesis concerns itself mostly with the use of Bayesian tools to analyse spatial proteomics data. Bayesian modelling and computation is revisited throughout this thesis for clarity and completeness.

1.7 Thesis outline and contributions

The next two chapters of this thesis focus on the development of a Bayesian model for spatial proteomics data. We revisit current approaches used to analyse the data and provide a thorough background of Bayesian inference tools to lay the foundation for later chapters. We then develop a Bayesian mixture model for analysing spatial proteomics data and demonstrate is utility by comparing it with other methods. Reanalysis of a mouse stem cell dataset demonstrates the information gain by employing a Bayesian model. The chapter ends by discussing some limitations of the approach.

Chapter 3 discusses a software implementation of the method provided in chapter 2. We provide a completely reproducible analysis of spatial proteomics data. A walk-through is provided so that those not versed in Bayesian methodology can still use sophisticated methods to analyse their data. Furthermore, we develop more visualisations and include more in depth discussion on prior choices in our Bayesian model.

Chapter 4 applies our Bayesian model to a challenging spatial proteomics dataset on *Toxoplasma gondii*. We introduce the cellular biology of *T. gondii* and why the organism is important to study. We demonstrate that our Bayesian analysis of this dataset constitutes a massive knowledge expansion. We map genomic features onto our spatial proteomics dataset and reveal spatial heterogeneity in these features. We conclude with the limitations of our analysis, laying the foundation for later chapters.

The limitations discussed in chapter 2,3 and 4 motivate further extensions of our original model. One of these limitation is the reliance on annotation and markers. In chapter 5, we develop a semi-supervised Bayesian model that can also uncover additional phenotypes within the data without markers. This method relies on a technique called *overfitted mixtures*. This opens up spatial proteomics to organisms that have little or no annotation.

In chapter 6, we develop non-parametric Bayesian approaches and contrast these with parametric alternatives presented in chapter 2. This chapter mostly focuses on the statistical challenge of deriving a model that is closer to the mechanisms that generate the data. This chapter also develops functional data analysis tools that are important for more advanced models. We also introduce methods to alleviate computation in these complex models.

Having extensively discussed the allocation problem, we turn to the dynamic question: which proteins change localisation upon perturbation of the subcellular environment? We introduce the concept of *differential localisation* in chapter 7 and argue that these experiments provide an opportunity to revolutionise our understanding of cell biology. Building on previous chapters, we develop a semi-supervised integrative Bayesian mixture model and show that it outperforms current ad-hoc approaches in the literature. We provide an extensive case study on human cytomegalovirus infection.

We conclude by summarising the contributions of this thesis and outline research directions for the future.

Chapter 2

A Bayesian mixture modelling approach for spatial proteomics

This chapter introduces computational spatial proteomics and commonly used machine learning algorithms that are applied to such data. We highlight the limitations of such machine learning algorithms and present a new Bayesian model for spatial proteomics data. The material presented here is an edited version of Crook et al. [83].

2.1 Motivation

2.1.1 Abstract

The analysis of the spatial sub-cellular distribution of proteins is of vital importance to fully understand context-specific protein function. Some proteins can be found with a single location within a cell, but up to half of proteins may reside in multiple locations, can dynamically re-localise, or reside within a compartment of unknown function. These considerations lead to uncertainty in associating a protein to a single location. Currently, mass spectrometry (MS) based spatial proteomics relies on supervised machine learning algorithms to assign proteins to sub-cellular locations based on common gradient profiles. However, such methods fail to quantify uncertainty associated with sub-cellular class assignment. Here we reformulate the framework on which we perform statistical analysis. We propose a Bayesian generative classifier based on Gaussian mixture models to assign proteins probabilistically to sub-cellular niches, thus proteins have a probability distribution over sub-cellular locations, with Bayesian computation performed using the expectation-maximisation (EM) algorithm, as well as Markov-chain Monte-Carlo (MCMC). Our methodology allows proteome-wide uncertainty quantification, thus adding a further layer to the analysis of spatial proteomics. Our framework is flexible, allowing many different systems to be analysed and reveals new modelling opportunities for spatial proteomics.

We find our methods perform competitively with current state-of-the-art machine learning methods, whilst simultaneously providing more information of biological significance. We highlight several examples where classification based on the support vector machine is unable to make any conclusions, while uncertainty quantification using our approach provides biologically intriguing results. To our knowledge this is the first Bayesian model of MS-based spatial proteomics data.

2.2 Introduction and literature review

Spatial proteomics is an interdisciplinary field studying the localisation of proteins on a large-scale. Where a protein is localised in a cell is a fundamental question, since a protein must be localised to its required sub-cellular compartment to interact with its binding partners (for example, proteins, nucleic acids, metabolic substrates) and carry out its function [171]. Furthermore, mis-localisations of proteins are also critical to our understanding of biology, as aberrant protein localisation has been implicated in many pathologies [350, 283, 259, 102, 69], including cancer [240, 393, 257, 419] and obesity [423].

Sub-cellular localisations of proteins can be studied by high-throughput mass spectrometry (MS) [154]. MS-based spatial proteomics experiments enable us to confidently determine the sub-cellular localisation of thousands of proteins within in a cell [68], given the availability of rigorous data analysis and interpretation [154].

In a typical MS-based spatial proteomics experiment, cells first undergo lysis in a fashion which maintains the integrity of their organelles. The cell content is then separated using a variety of methods, such as density separation [119, 68], differential centrifugation [220], free-flow electrophoresis [356], or affinity purification [194]. In LOPIT [118, 119, 400] and *hyper*LOPIT [68, 324], cell lysis is proceeded by separation of the content along a density gradient. Organelles and macro-molecular complexes are thus characterised by density-specific profiles along the gradient [98]. Discrete fractions along the continuous density gradient are then collected, and quantitative protein profiles that match the organelle profiles along the gradient, are measured using high accuracy mass spectrometry [324]. LOPIT-DC is a variant of this workflow where sub-cellular compartments are fractions based on differential centrifugation strategies [159].

The data are first visualised using principal component analysis (PCA) and known sub-cellular compartments are annotated [45]. Supervised machine learning algorithms are then typically employed to create classifiers that associate un-annotated proteins to specific organelles [155], as well as semi-supervised methods that detect novel sub-cellular clusters using both labelled and un-labelled features [43]. More recently, a state-of-the-art transfer learning (TL) algorithm has been shown to improve the quantity and reliability of sub-cellular protein

assignments [44]. Applications of such methods have led to organelle-specific localisation information of proteins in plants [119], *Drosophila* [448], chicken [185], human cell lines [43], mouse pluripotent embryonic stem cells [68] and cancer cell lines [455].

Classification methods which have previously been used include partial least squares discriminate analysis [119], K nearest neighbours [181], random forests [349], naive Bayes [341], neural networks [450] and the support vector machine amongst others (see [155] for an overview). Although these methods have proved successful within the field, they have limitations. Typically, such classifiers output an assignment of proteins to discrete pre-annotated sub-cellular locations. However, it is important to note that half the proteome cannot be robustly assigned to a single sub-cellular location [68, 455], which may be a manifestation of proteins in so far uncharacterised organelles or proteins that are distributed amongst multiple locations. These factors lead to uncertainty in the assignment of proteins to sub-cellular localisations, and thus quantifying this uncertainty is of vital importance [246].

To allow us to quantify uncertainty, this chapter presents a probabilistic generative model for MS-based spatial proteomics data. Our model posits that each annotated sub-cellular niche can be modelled by a multivariate Gaussian distribution. Thus, the full complement of annotated proteins is captured by a mixture of multivariate Gaussian distributions. With the prior knowledge that many proteins are not captured by known sub-cellular niches, we augment our model with an outlier component. Outliers are often dispersed and thus this additional component is described by a heavy-tailed distribution: the multivariate Student's t-distribution, leading us to a T Augmented Gaussian Mixture model (TAGM).

Given our model and proteins with known location, we can probabilistically infer the sub-cellular localisation of thousands of proteins. We can perform inference in our model by finding *maximum a posteriori* (MAP) estimates of the parameters. This approach returns the probability of each protein belonging to each annotated sub-cellular niche. These posterior localisation probabilities can then be the basis for classification. In a more sophisticated, fully Bayesian approach to uncertainty quantification, we can additionally infer the entire posterior distribution of localisation probabilities. This allows the uncertainty in the parameters in our model to be reflected in the posterior localisation probabilities. We perform this inference using Markov-chain Monte-Carlo methods; in particular, we provide an efficient collapsed Gibbs sampler to perform inference.

We perform a comprehensive comparison to state-of-the-art classifiers to demonstrate that our method is reliable across 19 different spatial proteomics datasets and find that all classifiers we considered perform competitively. To demonstrate the additional biological advantages our method can provide, we apply our method to a *hyper*LOPIT dataset on mouse pluripotent embryonic stem cells [68]. We consider several examples of proteins that were unable to be assigned using traditional machine-learning classifiers and show that, by considering the full

posterior distribution of localisation probabilities, we can draw meaningful biological results and make powerful conclusions. We then turn our hand to a more global perspective, visualising uncertainty quantification for over 5,000 proteins, simultaneously. This approach reveals global patterns of protein organisation and their distribution across sub-cellular compartments.

We make extensive use of the R programming language [372] and existing MS and proteomics packages [153, 156]. We are highly committed to creating open software tools for high quality processing, visualisation, and analysis of spatial proteomics data. We build upon an already extensive set of open software tools [156] as part of the Bioconductor project [166, 212] and our methods are made available as part of this project. In chapter 3, we focus on the software implementation of our method.

This chapter is organised as follows, we first introduce supervised machine-learning algorithms with a particular focus on the those most applied in spatial proteomics: kernel methods and the K-nearest neighbours algorithm. We then divert to mixture models, first in the context of clustering, and their robust formulations. We proceed to provide the necessary background on Bayesian inference; such that we can formulate Bayesian inference for mixture models. Our methods end with the exposition of a semi-supervised robust Bayesian mixture model for spatial proteomics data. This model is then compared with state-of-the-art approaches, before a detailed case study on mouse pluripotent Embryonic Stem Cells (mESCs) is provided. We conclude the chapter with the limitations of our approach.

2.3 Methods

2.3.1 Previous methods

Kernel Machines

The discussion henceforth follows excellent books by Cristianini et al. [80] and Schölkopf et al. [409], the seminal text of Cortes and Vapnik [76], as well as the review of Smola and Schölkopf [429]. Consider the scenario where we are given data that arise from two classes $\{(x_1, y_1), ..., (x_n, y_n)\} \subset \mathcal{X} \times \{-1, 1\}$. Here $\mathcal{X}$ denotes the abstract space in which the input data live and is typically a subset of Euclidean space. We also briefly restrict ourselves to cases where the possible class labels y_i are either -1 or 1. If there is a hyperplane that separates these two classes, then there exists β, such that $\|\beta\|_2 = 1$, that satisfies $y_i \beta^T x_i > 0$ for all i. Indeed, β is a unit-normal vector to the hyperplane that partitions the two classes. Such a hyperplane may not be unique and so we focus interest on the hyperplane that maximises the

margin between the two classes. This can be formulated as the following optimisation problem:

$$\begin{aligned} &\max_{\beta\in\mathcal{X},M\geq 0} M \\ &\text{subject to: } y_i x_i^T\beta \geq M,\ i=1,...,n. \end{aligned} \tag{2.1}$$

The current assumption that there is a hyperplane separating the two classes is unlikely in practice. In more realistic scenarios, we wish to replace the constraint with a penalty for allowing x_i to be on the wrong side of the margin boundary. A sensible choice of penalty is equal to the distance over the boundary measured in units of M. Thus, the penalty has the form $1 - y_i x_i^T\beta/M$. Now, instead of enforcing $\|\beta\|_2 = 1$, we can rescale such that $\|\beta\|_2 = 1/M$, which eliminates M from our objective function. Hence, we may write the penalty in following form

$$\arg\min_{\beta\in\mathcal{X}} \sum_{i=1}^{n}(1 - y_i x_i^T\beta)_+ + \lambda\|\beta\|_2^2, \tag{2.2}$$

where $(.)_+$ denotes the positive part and λ is a free parameter. Note that we reformulated $\max_{\beta\in\mathcal{X},M\geq 0} M$ as $\min_{\beta\in\mathcal{X}}\|\beta\|_2$. To allow hyperplanes that do not intersect the origin we can replace $x_i^T \to (x_i - b)^T$. Then, with minor algebraic manipulation, we may rewrite the penalty as:

$$\arg\min_{\beta\in\mathcal{X},\mu\in\mathbb{R}} \sum_{i=1}^{n}(1 - y_i(x_i^T\beta + \mu))_+ + \lambda\|\beta\|_2^2. \tag{2.3}$$

This is the typical objective function of the *support vector machine*; however, we wish to generalise further using *kernels*.

Definition 1. *A kernel k is a symmetric map $k : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ such that for all $n \in \mathbb{N}$ and all $x_1, .., x_n \in \mathcal{X}$ the matrix K, with entries $K_{ij} = k(x_i, x_j)$ is positive semi-definite.*

It is useful to note that linear combinations and pointwise products of kernels are also kernels. Some examples of kernels include the following [435]:

$$\begin{aligned} &\text{Linear kernel:} \quad k(x_i, x_j) = x_i^T x_j, \\ &\text{Gaussian kernel:} \quad k(x_i, x_j) = \exp\left(\frac{\|x_i - x_j\|_2^2}{\sigma^2}\right), \\ &\text{Sobelev kernel:} \quad k(x_i, x_j) = \min(x_i, x_j). \end{aligned} \tag{2.4}$$

The following theorem identifies k with a feature map.

Theorem 1. *For every kernel k there exists a feature map ϕ taking values in some inner product space $\mathcal{H}$ such that $k(x_i, x_j) = \langle\phi(x_i), \phi(x_j)\rangle$.*

The proof of this theorem is omitted for brevity, but the important insight is that $\phi(x) : \mathcal{X} \to \mathbb{R}$ is identified as $\phi(x) = k(\cdot, x)$. The space $\mathcal{H}$ from theorem 1 can be studied in more

detail. The completion of $\mathcal{H}$ (by adding the limits of all Cauchy sequences) makes $\mathcal{H}$ a Hilbert space (a complete inner product space). Actually, $\mathcal{H}$ is more than a Hilbert space it is a reproducing kernel Hilbert space (RKHS) [13].

Definition 2. *A Hilbert space $\mathcal{H}$ of function $f : \mathcal{X} \to \mathbb{R}$ is a reproducing kernel Hilbert space (RKHS) if for all $x \in \mathcal{X}$, there exists $k_x \in \mathcal{H}$ such that*

$$f(x) = \langle k_x, f(x) \rangle \text{ for every } f \in \mathcal{H}. \tag{2.5}$$

The function $k : \mathcal{X} \times \mathcal{X} \to \langle$ defined by $k(x, x') = k_{x'}(x)$ is called the *reproducing kernel*. A reproducing kernel can be written as the inner product between two feature maps and so it is a kernel. Furthermore, for any kernel k there is a unique RKHS with reproducing kernel k. An illustrative example is the following, let $\mathcal{H} = \{f : f(x) = x^T\beta, \beta \in \mathbb{R}^p\}$. The norm on this space is $\|f\|^2_{\mathcal{H}} = \|\beta\|^2_2$ and thus $\mathcal{H}$ is the RKHS corresponding to the linear kernel. Thus far, we have overlooked the rather crucial consideration that optimisation of a loss function over $\mathcal{H}$ could be a fruitless endeavour, because $\mathcal{H}$ is potentially infinite dimensional. It is the content of the Representer theorem that overcomes this observation [242, 408].

Theorem 2. *The Representer Theorem*
Let $c : \{-1, 1\}^n \times \mathcal{X}^n \times \mathbb{R}^n \to \mathbb{R}$ be a loss function, and let $J : [0, \infty) \to \mathbb{R}$ be strictly increasing. Let $x_1, .., x_n \in \mathcal{X}$, $y_1, ..., y_n \in \{-1, 1\}$. Furthermore, let $f \in \mathcal{H}$, where $\mathcal{H}$ is an RKHS with reproducing kernel k, and let $K_{ij} = k(x_i, x_j)$. Then $\hat{f}$ minimises:

$$R_1(f) := c(y_1, ..., y_n, x_1, ..., x_n, f(x_1), ..., f(x_n)) + J(\|f\|^2_{\mathcal{H}}) \tag{2.6}$$

over $f \in \mathcal{H}$ if and only if $\hat{f} = \sum_{i=1}^n \hat{\alpha}_i k(\cdot, x_i)$ and $\hat{\alpha} \in \mathbb{R}$ minimises the following over $\alpha \in \mathbb{R}^n$:

$$R_2(\alpha) := c(y_1, ..., y_n, x_1, ..., x_n, K\alpha) + J(\alpha^T K\alpha). \tag{2.7}$$

The proof of this theorem is omitted, but its implications are somewhat remarkable. Viewing the Theorem in its reverse implication tells us that optimising R_1 is not hopeless, since it is equivalent to finding $\hat{\alpha}_i$ that optimise R_2, which is a finite dimensional problem. This is opposed to the infinite dimensional problem of optimising R_1. Now returning to the objective function of the SVM in equation 2.3, we observe that

$$(\hat{\mu}, \hat{f}) = \underset{f \in \mathcal{H}, \mu \in \mathbb{R}}{\arg\min} \sum_{i=1}^{n} (1 - y_i(f(x_i) + \mu))_+ + \lambda \|f\|^2_{\mathcal{H}}, \tag{2.8}$$

where $\mathcal{H}$ is the RKHS corresponding to the linear kernel, is equivalent to this formulation. Now, applying the Representer theorem generalises this to an arbitrary RKHS with kernel k with

corresponding kernel matrix K, with the following objective function

$$(\hat{\mu}, \hat{\alpha}) = \underset{\alpha \in \mathbb{R}^n, \mu \in \mathbb{R}}{\arg\min} \sum_{i=1}^{n} (1 - y_i(K_i^T \alpha + \mu))_+ + \lambda \alpha^T K \alpha. \tag{2.9}$$

The objective function has a free parameter λ, typically referred to as the cost, and the Gaussian Kernel (a popular kernel) has free parameter σ. The parameters are typically chosen using a grid search and k-fold cross-validation. The score for an observation x is given by

$$\sum_{i=1}^{n} \hat{\alpha}_i k(x, x_i) + \hat{\mu}, \tag{2.10}$$

and the class is given by the sign of the score. Frequently, it is desired to obtained "probabilities" from these scores. This is usually performed using logistic regression with scores as input and maximum likelihood estimation is used for inference. The errors of the predictors are usually assumed to follow a centred Laplace distribution. The process is often referred to as Platt scaling or a variant thereof [365, 269].

To this point, we have only considered binary classification. To extend to the multi-class situation, we use a one-vs-one schema. In this framework, a binary classifier is used pairwise on the $\frac{c(c-1)}{2}$ classification problems, where c is the number of classes. A simple vote is used to obtain the predicted class. To compute "probabilities" in the multi-class framework, we use pairwise probabilities computed from the logistic regression approach previously described. Thus, we have estimates of the pairwise class probabilities $r_{ij} = p(y = i | y = i \text{ or } y = j, x)$, but we wish to obtain $p_i = p(y = i|x)$ for $i = 1, ..., c$. Though a number of approaches are available (see Wu et al. [485]), we explain a popular approach which is frequently used because of its stability [485]. Consider the following optimisation problem

$$\begin{aligned} &\min_p \frac{1}{2} \sum_{i=1}^{c} \sum_{j:j \neq i} (r_{ji} p_i - r_{ij} p_j)^2, \\ &\text{subject to: } \sum_{i=1}^{c} p_i = 1, \ p_i \geq 0. \end{aligned} \tag{2.11}$$

This can be re-formulated as

$$\begin{aligned} &\min_p \frac{1}{2} p^T Q p, \\ &\text{where } Q_{ij} = \sum_{s:s \neq i} r_{si}^2 \text{ if } i = j, \text{ otherwise } Q_{ij} = r_{ji} r_{ij}, \end{aligned} \tag{2.12}$$

which is a classical linear-constrained convex quadratic programming problem [141]. A standard approach is to use Lagrange multipliers to solve this problem; however, iterative methods

are more frequently used because of improved numerical stability [141]. There are several criticisms of this process. The first is that the SVM computes a hard margin and so the scores do not contain information on probabilities distant from 0.5. Secondly, the obtained probabilities are not necessarily consistent in the sense that the class which maximises the score will not necessarily maximise the probability. Finally, the probabilities are not produced from a generative probabilistic model and so calibration of the probabilities cannot be criticised from predictive checks. Highly optimised libraries and software are available to implement SVMs [238, 63].

K-Nearest Neighbours

The k-Nearest Neighbours (k-NN) algorithm, first proposed by Fix [134] and Cover and Hart [77], is perhaps the simplest non-parametric classification approach. Suppose we are given data of the form $((x_1, y_1), ..., (x_n, y_n)) \subset \mathcal{X} \times \{1, ..., c\}$, where we begin in the multi-class setting with c possible classes. The k-nearest neighbour algorithm assigns a class to a data point using the following procedure. Firstly, the k nearest neighbours are computed, where nearest is computed with respected to a distance - usually the Euclidean distance. The class labels of these k nearest neighbours are tallied and the most frequent class amongst the tally is the assigned class. The first question to answer is: how to choose k? Firstly, it is preferred that k is odd, since this avoids ties. Though ties can be overcome using random assignment [77]. Then, typically, a grid search is employed and cross-validation used to select the k. Though we will not use this approach here, for completeness we highlight another approach used to select k. Hall et al. [184] derive conditions for the optimal choice of k, in the sense of minimising risk, for data obtained from Poisson and Binomial models. The theory motivates using a bootstrapping procedure to empirically select an optimal k.

As for the SVM, we frequently desire probabilities from the k-NN algorithm. Consider, the following equation [223]

$$p(y_i = j|x_i) = \frac{1}{k} \sum_{l \in \mathcal{N}_i} \mathbb{I}(y_l = j), \tag{2.13}$$

where $\mathcal{N}_i$ denotes the indices of the k nearest neighbours to x_i. The above provides a reasonable probabilistic interpretation of k-NN algorithm. However, this formula results in most classes receiving 0 probability and frequent ties. Alternatively, we could interpret the proportion of neighbours as a non-parametric posterior probability. To avoid non-zero probabilities for classes, we perform Laplace smoothing; that is, the posterior allocation probability is given by

$$p(z_i = j|x_i) = \frac{N_{ij} + \alpha d_j c}{k + \alpha c}, \tag{2.14}$$

where N_{ij} is the number of neighbours belonging to class j in the neighbourhood of x_i, c is the number of classes, K is the number of nearest neighbours (optimised through cross validation) and d_j is the incidence rate of each class in the training set. Finally, $\alpha > 0$ is the pseudo-count smoothing parameter. Motivated by a Bayesian interpretation of placing a Jeffrey's type Dirichlet prior over multinomial counts [232], it is typical to choose $\alpha = 0.5$ [190, 462, 293].

Mixture models for clustering

For pedagogical reasons, we introduce mixture models in the unsupervised/clustering setting. This allows us to seamlessly transition into mixture models for classification and lays the foundation for the various types of semi-supervised mixture models introduced in later chapters.

The following material is well described in a number of articles, books and technical reports (such as Banfield and Raftery [19], McLachlan and Basford [301], McLachlan and Peel [303], Murphy [327], Scrucca et al. [412]). Finite mixture models are of the form,

$$p(\mathbf{x}|\boldsymbol{\pi}, \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k F(\mathbf{x}|\boldsymbol{\theta}_k), \tag{2.15}$$

where K is the number of mixture components, π_k are the mixture proportions, and $F(\mathbf{x}|\boldsymbol{\theta}_k)$ are the component densities. We assume each component density to have the same parametric form, but with component-specific parameters, $\boldsymbol{\theta}_k$.

We suppose that we have a collection of n data points, $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ that we seek to model using Equation (2.15). We associate with each of these data points a latent component indicator variable, $z_i \in \{1, \ldots, K\}$, which indicates which component generated observation $\mathbf{x}_i$. The likelihood of this model is then given by

$$p(X|\pi, \theta) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k F(\mathbf{x}_i|\boldsymbol{\theta}_k) = \prod_{i=1}^{n} \prod_{j=1}^{K} F(\mathbf{x}_i|\boldsymbol{\theta}_{z_i})^{\mathbb{I}(z_i=j)}. \tag{2.16}$$

Though any likelihood is admissible in practice, we focus on the Gaussian case for clarity. In the Gaussian case the form of F is given explicitly as:

$$F(\mathbf{x}_i|\mu_k, \Sigma_k) = (2\pi)^{-(p/2)} \, |\Sigma_k|^{-1/2} \exp\left\{ -\frac{1}{2} (\mathbf{x}_i - \mu_k)^T \Sigma_k^{-1} (\mathbf{x}_i - \mu_k) \right\}. \tag{2.17}$$

To fit this model, one typically uses maximum likelihood estimation. However, the log-likelihood in this scenario depends on the latent (unobserved) variables z_i. The expectation-maximisation (EM) algorithm is a general method for handling scenarios of this type and is guaranteed to reach a local maximum [108, 302]. The expectation-maximisation algorithm iterates between an expectation step, which is taken with respect to the latent variables and a maximisation

step, seeking to maximise the log-likelihood with respect to the estimated values of the latent variables. The algorithm iterates until the change in the log-likelihood is below some tolerance threshold. The expectation-maximisation algorithm for the Gaussian mixture model is [302, 327]

- Expectation Step: Compute for every $i = 1, ..., n$ and $j = 1, ..., K$

$$r_{ij} = \frac{\pi_j F(\mathbf{x}_i|\boldsymbol{\theta}_j)}{\sum_{k=1}^{K} \pi_k F(\mathbf{x}_i|\boldsymbol{\theta}_k)}. \tag{2.18}$$

- Maximisation Step:

$$\begin{aligned} \pi_k &= \frac{\sum_{i=1}^{n} r_{ik}}{\sum_{j=1}^{K}\sum_{i=1}^{n} r_{ij}}, \\ \mu_k &= \frac{\sum_{i=1}^{n} r_{ij}\mathbf{x}_i}{\sum_{i=1}^{n} r_{ik}}, \\ \Sigma_k &= \frac{\sum_{i=1}^{n} r_{ij}(\mathbf{x}_i - \mu_k)^T(\mathbf{x}_i - \mu_k)}{\sum_{i=1}^{n} r_{ik}}. \end{aligned} \tag{2.19}$$

- Compute the log-likelihood $Q(\theta_t)$.
- Repeat for $t \to t + 1$, until $|Q(\theta_t) - Q(\theta_{t-1})| < tolerance$.

A standard issue with this approach is variance collapse [327]. This is when a mixture component is centred exactly on a data point or data points are (nearly) collinear. The eigenvalues of the covariance matrix then shrink to 0, causing the log-likelihood to increase indefinitely, as well as leading to singular covariance matrices which causes estimation issues. One way to handle variance collapse is to monitor the smallest eigenvalue of the covariance matrix of each mixture component. If this value falls below a threshold then reset the covariance matrix to its initial value [412]. However, if one is willing to move beyond the likelihood framework then there is an alternative approach using priors, which we defer momentarily [412].
Until this point, we have assumed the number of components K is given. However, we may wish to perform *model selection* on the number of components. We refrain from saying *infer* K, because this assumes that K is random - whereas K is (currently) fixed. The typical approach to selecting K, is to consider a number of different values for K and choose an "optimal" K using the Bayesian Information Criterion (BIC) [411, 303]. The Bayesian information is so-called because it approximates the Bayes factor under a flat prior [31]. The derivation requires a Laplace approximation and application of the Weak Law of Large Numbers (WLLN) [31]. For model based clustering the BIC is given by

$$\text{BIC } = 2\log p(x|\hat{\theta}, \mathcal{M}) - m\log(n), \tag{2.20}$$

where $\log p(x|\hat{\theta}, \mathcal{M})$ is the maximised log likelihood for the model and the data, m is the number of parameters in model $\mathcal{M}$ and n is the number of data points. It is clear from the above equation that as more parameters are introduced, such as the number of components, the greater the penalty on the model.

To allow the inclusion of prior information and avoid problems associated with variance collapse, we may instead adopt a Bayesian approach and introduce priors on the mixture components. In the Gaussian case, a common and practical choice is the use of a normal-inverse-Wishart prior. That is

$$\begin{aligned} \mu|\Sigma &\sim \mathcal{N}(\mu_0, \Sigma/\lambda_0) \\ \Sigma &\sim \mathcal{IW}(\nu_0, S_0) \\ &\propto |\Sigma|^{\frac{\nu_0+d+1}{2}} \exp\left[-\frac{1}{2}\text{trace}(\Sigma^{-1}S_0^{-1})\right], \end{aligned} \tag{2.21}$$

for each mixture component and where d is the dimension of the data. As a result, maximum likelihood inference is replaced with *maximum a posterori* (MAP) inference. We do not derive the update equations here as a more complex example is presented later. To complete this discussion, we need to specify the hyperparameters. Fraley and Raftery [139] introduce diffusive priors that make minimal assumptions about the data, but they are set semi-empirically as to obtain the correct scale of the data. The hyperparameters are selected as follows

$$\begin{aligned} \mu_0 &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{x_i}, \\ \lambda_0 &= 0.01, \\ \nu_0 &= d+2, \\ S_0 &= \frac{(\text{diag }(\text{var}(X)))}{K^{1/d}}. \end{aligned} \tag{2.22}$$

Each hyperparameter is interpreted in the following way. The prior mean is the mean of the data. Then λ_0 is viewed as the number of observations with data μ_0 which are added to each component-specific mean. This value is small to avoid strong prior influence. The marginal prior distribution (or prior predictive) for a component-specific mean μ is given by a student's t-distribution. This can be observed by recalling that the student's t-distribution arises by marginalisation of the covariance from a normal distribution. Now, to ensure this t-distribution has finite covariance we require that $\nu_0 > d+1$. Thus, the choice presented here is the smallest integer value of ν_0 that ensures a finite covariance matrix. Hence, we have a well defined t-distribution with heavy tails. The empirically chosen scale matrix S_0 is chosen to roughly partition the range of the data into K balls of equal size.

Robust mixture models

Previously, we have assumed that our observations arise out of one of possibly c components/clusters. However, some measurement errors may produce outliers that do not arise from any of these clusters or require their own cluster. The covariance of a cluster can be artificially stretched to accommodate these outliers - this leads to poor inference for the observations that do, in fact, cluster. If these outliers require their own clusters, this can lead to numerical, estimation, and interpretation problems. Neither of these considerations is satisfactory. A number of approaches have been suggested in the literature and the first relies on the idea of spatial Poisson processes [180].

Definition 3. *Spatial Poisson process*
Let $B \subset \mathbb{R}^n$ be a Borel measurable set. Let $N(B)$ denote a point process confined to B. $N(B)$ is called a spatial Poisson process with intensity $\lambda > 0$ if

$$P(N(B) = n) = \frac{(\lambda |B|^n)}{n!} \exp\left(-\lambda |B|\right), \tag{2.23}$$

and for finite $k \geq 1$, given disjoint Borel sets $B_1, ..., B_k$, the number of points arising in $N(B_i)$ has distribution given by

$$P(N(B_i) = n_i, i = 1, ..., k) = \prod_{i=1}^{k} \frac{(\lambda |B_i|_i^n)}{n_i!} \exp\left(-\lambda |B_i|\right). \tag{2.24}$$

The mixture model can be reformulated, using spatial Poisson process:

$$p(\mathbf{x}|\boldsymbol{\pi}, \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k F(\mathbf{x}|\boldsymbol{\theta}_k) + \frac{\pi_0}{V}, \tag{2.25}$$

where V is the hypervolume of the data [19, 412]. An alternative approach, rather than relying on a noise component to model the outliers, is to use a heavy tailed family in the in the likelihood. For example a generalised Student's t-distribution could be used and the modified model becomes [358]

$$p(\mathbf{x}|\boldsymbol{\pi}, \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k G(\mathbf{x}|\boldsymbol{\theta}_k), \tag{2.26}$$

where G denotes the density corresponding to the t-distribution. The outliers in this scenario are those that lie in low density regions of these components. However, Hennig et al. [197] showed that neither of these approaches are breakdown-robust. Loosely, this means that a sequence of points can be added to the data which can arbitrarily drive an estimator from its original value [152, 149, 150]. For example the scale parameter of a component of a Gaussian or Student's t mixture can be driven towards 0. This sequence of points is what we colloquially

refer to as outliers. Coretto and Hennig [75] proposed an alternative approach, which is to introduce a pseudo-model where the noise component is an improper density :

$$p(\mathbf{x}|\boldsymbol{\pi}, \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k F(\mathbf{x}|\boldsymbol{\theta}_k) + \pi_0 \delta, \tag{2.27}$$

where $\delta > 0$. The parameter δ is not considered a formal parameter to be inferred, but a tuning parameter that is to be set based on prior knowledge. Coretto and Hennig [75] suggest a number of data-driven strategies to select δ. This approach, though displaying desirable theoretical properties, requires a complex EM algorithm over a carefully constrained parameter space and is sensitive to initialisation [75]. Furthermore, it fails to be a formal density and thus cannot be used as a generative model [75]. Thus, the interpretation of this pseudo-model is complex and a Bayesian interpretation, for uncertainty quantification, is impossible.

Bayesian mixture models

Thus far, we have discussed mixture models for clustering, their robust formulations and methods to perform inference in these models. However, these approaches have all been optimisation focused. If uncertainty quantification is desired then Bayesian inference is an alternative approach, which is performed by obtaining samples from the posterior distribution of the parameters and latent variables [164]. A number of approaches are available for Bayesian inference in mixture models, notably Markov-chain Monte-Carlo (MCMC) [110, 294, 229], variational inference (VI) [74, 330, 34], expectation-propagation (EP) [315, 314], amongst others [112].

As the prevailing method in the literature, because of its well-studied theoretical properties and relative ease of implementation, we focus on MCMC for Bayesian inference [173, 287, 10, 288, 151, 47]. We provide a brief interlude to discuss the key ideas of MCMC for Bayesian inference (the text follows closely that of Andrieu et al. [10]). The goal of MCMC sampling is to produce samples from the posterior distribution, generically written $p(x|y)$. This is challenging because it typically involves intractable integration problems. Markov-chain Monte-Carlo relies on its namesake: the Monte-Carlo method [309].

The Monte-Carlo method seeks to produce samples $\{x^{(i)}\}_{i=1}^{N}$ from some target density $T(x)$ (usually, but not necessarily, a posterior distribution), where $T(x)$ is defined in some space $\mathcal{X}$. These samples can then be used as an empirical estimator for the density:

$$p_N(x) = \frac{1}{N} \sum_{i=1}^{N} \delta_{x^{(i)}}(x), \tag{2.28}$$

where $\delta_{x^{(i)}}(x)$ is the Dirac measure located at $x^{(i)}$. Then, one can approximate the intractable integral of interest $I(f)$ with finite sums $I_N(f)$ ([180]):

$$I_N(f) = \frac{1}{N}\sum_{i=1}^{N} f(x^{(i)}) \rightarrow_{a.s} I(f) = \int_{\mathcal{X}} f(x)T(x)\,\mathrm{d}x. \tag{2.29}$$

By the strong law of large numbers (SLLN), the estimate $I_N(f)$ converges almost surely (a.s.) to $I(f)$ as $N \rightarrow \infty$. If we also assume that

$$\sigma_f^2 = \mathbb{E}_{T(x)}\left[f^2(x)\right] - I^2(f) < \infty, \tag{2.30}$$

then $\mathrm{var}(I_N(f)) = \frac{\sigma_f^2}{N}$ and a central limit theorem holds ([382, 180])

$$\sqrt{N}(I_N(f) - I(f)) \rightarrow_d \mathcal{N}(0, \sigma_f^2). \tag{2.31}$$

For very few distributions is it simple to obtain samples from $T(x)$. A number of sampling algorithms are available, including rejection sampling [382], importance sampling [399, 168], sequential importance sampling [275, 274], sequential Monte-Carlo samplers [105], as well as auxiliary variable samplers such as slice sampling [334] and Hamiltonian Monte-Carlo [115, 335, 28]. The latter will be introduced later in this thesis. For now, we focus on general MCMC algorithms.

MCMC applies in the general setting where we are interested in some target $T(x)$ from which we cannot draw samples directly, but can evaluate up to some normalising constant. The strategy of MCMC is to obtain samples $x^{(i)}$ from the target, whilst efficiently exploring the state space $\mathcal{X}$. The mechanism of the MCMC algorithm is a Markov-chain [180].

Definition 4. *Let $\mathcal{X}$ be a measurable space and let K be a Markov kernel. A stochastic process (X_n) on $\mathcal{X}$ is called a time homogeneous Markov-chain with Markov kernel K and initial distribution μ if*

$$P(X_0 \in A_0, ..., X_n \in A_n) = \int_{A_0,...,A_n} K(x_{n-1}|A_n)K(x_{n-2}|x_{n-1})...\mu(x_0), \tag{2.32}$$

for any $n \in \mathbb{N}$ and any measurable sets $A_0, ..., A_n$.

We can think of the Markov kernel as how the Markov-chain transitions around the space. The key result for Markov-chain theory is that if the kernel satisfies some technical conditions (such as Harris recurrence) then they admit a unique stationary distribution, as well as desirable convergence properties [396]:

$$p\left(x^{(i+1)}\right) = \int p(x^{(i)})K(x^{(i+1)}|x^{(i)})\,\mathrm{d}x^{(i)}. \tag{2.33}$$

Thus, if the Markov-chain is carefully constructed then the stationary distribution is the target distribution from which we are interested in sampling [456]. Again, under technical assumptions, Ergodic theorems and central limit theorems hold for these samples [449, 456, 22, 151, 160, 227, 311, 345, 383, 386, 384, 385, 388, 389, 395, 396, 406, 449]. Thus we can use $\{x^{(i)}\}$ to construct Markov-chain Monte-Carlo estimators of the quantities of interest [456]. In Bayesian analysis the target stationary distribution of the MCMC algorithm is the posterior distribution [456].

The challenge is to construct a valid Markov kernel with the desired properties. The elegance of the Metropolis-Hastings algorithm is a generic method for construct valid Markov kernels [310, 187, 382]. Let $T(x)$ be the target distribution of interest and let $q(x^*|x)$ be a *proposal* distribution; that is, we sample a candidate value x^* from $q(|x)$. The Metropolis-Hastings step of the Markov chain is to move to x^* with the follow *acceptance* probability:

$$A(x, x^*) = \min\left\{1, \frac{T(x^*)q(x|x^*)}{T(x)q(x^*|x)}\right\}, \tag{2.34}$$

otherwise the Markov chain stays at x. A key insight is that for *any* proposal distribution the Metropolis-Hastings transition kernel defines a valid Markov kernel:

$$K_{MH} = A(x, x^*) \cdot q(x^*|x) + r(x) \cdot \delta_x(x^*), \tag{2.35}$$

where

$$r(x) = \int_{\mathcal{X}} q(x^*|x)(1 - A(x, x^*))\,\mathrm{d}x^*. \tag{2.36}$$

We can see from the construction of the Metropolis-Hastings algorithm that it satisfies so-called *detailed balance*:

$$T(x^*)K_{MH}(x|x^*) = T(x)K_{MH}(x^*|x) \tag{2.37}$$

and this has stationary distribution (or formally invariant measure associated to) $T(x)$. Technical conditions such as aperiodicity and irreducibility ensure convergence of the algorithm [389]. Further technical arguments establish geometric ergodicity [228] and other convergence results [227]. It is often useful to cycle through different kernels to explore different parts of the target distribution in different ways or use mixtures of kernels to allow global and local moves through the target [386, 388]. The choice of proposal distribution strongly affects how well the MH algorithm *mixes* (how many effective samples are produced per unit time) [382]. The *Gibbs sampler* is a special case of the MH for a particular choice of proposal [383].

Suppose we have access to the full conditional distributions $p(x_j|x_{-j})$ for $j = 1, ..., n$, where x_j denotes the j^{th} co-ordinate of x and x_{-j} denotes all but the j^{th} co-ordinate. Then let the proposal distribution of the Gibbs sampler be defined as follows, for $j = 1, .., n$

$$q(x^*|x^{(i)}) = T(x_j^*|x_{-j}^{(i)}) \text{ if } x_{-j}^{(i)} = x_{-j}^*, \tag{2.38}$$

otherwise the proposal distribution is 0. Let us compute the acceptance probability

$$\begin{aligned} A(x^{(i)}, x^*) &= \min\left\{1, \frac{T(x^*)q(x^{(i)}|x^*)}{T(x^{(i)})q(x^*|x^{(i)})}\right\} \\ &= \min\left\{1, \frac{T(x^*)T(x_j^{(i)}|x_{-j}^{(i)})}{T(x^{(i)})T(x_j^*|x_{-j}^*)}\right\} \\ &= \min\left\{1, \frac{T(x_j^*|x_{-j}^{(i)})T(x_{-j}^{(i)})T(x_j^{(i)}|x_{-j}^{(i)})}{T(x_j^{(i)}|T(x_{-j}^{(i)}))T(x_{-j}^{(i)})T(x_j^*|x_{-j}^*)}\right\} \\ &= \min\left\{1, \frac{T(x_{-j}^*)}{T(x_{-j}^{(i)})}\right\} \\ &= 1 \end{aligned} \tag{2.39}$$

Thus proposals from the Gibbs sampler are always accepted. In practice, when only some conditional distributions are available in closed form, we can cycle between MH moves and Gibbs moves.

We provide a brief review of Bayesian inference for finite mixture models (see, for example [261, 110, 139] for more details). We recall that finite mixture models are of the form,

$$p(\mathbf{x}|\boldsymbol{\pi}, \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k F(\mathbf{x}|\boldsymbol{\theta}_k), \tag{2.40}$$

where K is the number of mixture components, π_k are the mixture proportions, and $F(\mathbf{x}|\boldsymbol{\theta}_k)$ are the component densities. We assume each component density to have the same parametric form, but with component-specific parameters, $\boldsymbol{\theta}_k$. We denote the prior for these unknown component parameters by $G_0(\boldsymbol{\theta})$. We suppose that we have a collection of n data points, $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ that we seek to model using Equation (2.40). We associate with each of these data points a component indicator variable, $z_i \in \{1, \ldots, K\}$, which indicates which component generated observation $\mathbf{x}_i$. Given the mixing proportions, the joint prior distribution of these indicators is multinomial with parameter vector $\boldsymbol{\pi} = [\pi_1, \ldots, \pi_K]$,

$$P(z_1, \ldots, z_n|\boldsymbol{\pi}) = \prod_{k=1}^{K} \pi_k^{n_k}, \tag{2.41}$$

where n_k is the number of data points x_i for which $z_i = k$. If we assign the mixture proportions a symmetric Dirichlet prior with concentration parameter α/K, then we may marginalise the π_k in order to yield the following joint distribution for the indicators [327],

$$P(z_1, \ldots, z_n|\alpha) = \frac{\Gamma(\alpha)}{\Gamma(n+\alpha)} \prod_{i=1}^{K} \frac{\Gamma(n_i + \alpha/K)}{\Gamma(\alpha/K)}. \tag{2.42}$$

For Gibbs sampling, we require the conditional priors for a single indicator, z_i, given all of the others, z_{-i}. These are given by [327],

$$P(z_i = k|z_{-i}, \alpha) = \frac{n_{-i,k} + \alpha/K}{N - 1 + \alpha}, \tag{2.43}$$

where $n_{-i,k}$ is the number of observations, excluding $\mathbf{x}_i$, that are associated with component k. If we are given the parameters, $\boldsymbol{\theta}_k$, associated with each of the components then we may combine the above conditional priors with the likelihoods, $F(\mathbf{x}_i|\boldsymbol{\theta}_k)$, in order to obtain the conditional posterior:

$$P(z_i = k|z_{-i}) \propto \frac{n_{-i,k} + \alpha/K}{N - 1 + \alpha} F(\mathbf{x}_i|\boldsymbol{\theta}_k). \tag{2.44}$$

An alternative to integrating out the mixture proportions is to sample them directly from the posterior, which leads to increased posterior variance [160, 60] but can be computationally advantageous. Conjugacy of the Dirichlet prior and multinomial likelihood means that the posterior distribution of the mixing proportions is also Dirichlet,

$$\boldsymbol{\pi}|z_1, ..., z_n, \alpha \sim Dir(\alpha/K + n_1, ..., \alpha/K + n_K). \tag{2.45}$$

In this situation the conditional posterior becomes

$$P(z_i = k|\boldsymbol{\pi}) \propto \pi_k F(\mathbf{x}_i|\boldsymbol{\theta}_k). \tag{2.46}$$

If $G_0(\boldsymbol{\theta})$ is conjugate for $F(\mathbf{x}|\boldsymbol{\theta}_k)$, then we perform Gibbs sampling for $\boldsymbol{\theta}_k$, otherwise we can perform a MH move. In fact, in the case of conjugacy, we can go further and analytically compute the following integral

$$G(\mathbf{x}|H) = \int_{\theta_k} F(\mathbf{x}|\boldsymbol{\theta}_k) G_0(\boldsymbol{\theta}_k|H) \, \mathrm{d}\boldsymbol{\theta}_k. \tag{2.47}$$

That is to say we have marginalised the parameters $\boldsymbol{\theta}_k$. In these cases, we do not need to sample $\boldsymbol{\theta}_k$, but rather update prior hyperparameters at each iteration of the MCMC algorithm. This is frequently referred to as *collapsed Gibbs sampling* [386].

2.3.2 Semi-supervised robust Bayesian mixture models

To summarise thus far, we have introduced mixture models for clustering, typical model fitting and selection methods for mixture models, as well as robust methods. We have also provided a primer on Bayesian inference, Markov-chain Monte-Carlo and Bayesian inference in mixture models. The goal of this section is to introduce a mixture model, which is suitable for spatial proteomics data. The first goal is to formulate a *generative* model so that we can perform uncertainty quantification on quantities of interest. Mixture models are mostly used in the

clustering paradigm, in our scenario we wish to classify proteins to organelles, using labelled and unlabelled data, so our second goal is the development of a *semi-supervised* model. Finally, the robust mixture model in equation 2.27 is a pseudo-model and includes an improper density. This precludes a Bayesian interpretation and so we mimic this strategy whilst maintaining a proper generative model.

We observe N protein profiles each of length L, corresponding to the number of quantified fractions along the gradient density, including combining replicates. For $i = 1, \dots, N$, we denote the profile of the i-th protein by $\mathbf{x}_i = [x_{1i}, \dots, x_{Li}]$. We suppose that there are K known sub-cellular compartments to which each protein could localise (e.g. cytoplasm, endoplasmic reticulum, mitochondria, ...). Henceforth, we refer to these K sub-cellular compartments as *components*, and introduce component labels z_i, so that $z_i = k$ if the i-th protein localises to the k-th component. We denote by X_L the set of proteins whose component labels are known, and by X_U the set of unlabelled proteins. If protein i is in X_U, we desire the probability that $z_i = k$ for each $k = 1, \dots, K$. That is, for each unlabelled protein, we want the probability of belonging to each component (given a model and the observed data).

We initially model the distribution of profiles associated with proteins that localise to the k-th component as multivariate normal with mean vector $\boldsymbol{\mu}_k$ and covariance matrix Σ_k, so that:

$$\mathbf{x}_i | z_i = k \quad \sim \mathcal{N}(\boldsymbol{\mu}_k, \Sigma_k). \tag{2.48}$$

For any i, we define the prior probability of the i-th protein localising to the k-th component to be $p(z_i = k) = \pi_k$. Letting $\boldsymbol{\theta} = \{\boldsymbol{\mu}_k, \Sigma_k\}_{k=1}^K$ denote the set of all component mean and covariance parameters, and $\boldsymbol{\pi} = \{\pi_k\}_{k=1}^K$ denote the set of all mixture weights, it follows (from the law of total probability) that:

$$p(\mathbf{x}_i | \boldsymbol{\theta}, \boldsymbol{\pi}) = \sum_{k=1}^{K} \pi_k f(\mathbf{x}_i | \boldsymbol{\mu}_k, \Sigma_k), \tag{2.49}$$

where $f(\mathbf{x}|\boldsymbol{\mu}, \Sigma)$ denotes the density of the multivariate normal with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ evaluated at $\mathbf{x}$.

Equation (2.49) defines the previously introduced *mixture model*, which posits a *generative* model for the data. Such models are useful for describing populations that are composed of a number of distinct homogeneous subpopulations. In our case, we model the full complement of measured proteins as being composed of K subpopulations, each corresponding to a different organelle or sub-cellular compartment. The literature of mixture model applications to biology is rich and some recent example include applications to retroviral integration sites [245], genome-

wide associations studies [267], single-cell transcriptomics [280] and affinity purification MS proteomics [66].

Though some proteins are well described as belonging to a single component, many proteins multi-localise or might belong to uncharacterised organelles. In order to allow the model to better account for these "outliers" that cannot be straightforwardly allocated to any single known component, we extend it by introducing an additional "outlier component". To do this, we augment our model by introducing a further indicator latent variable ϕ. Each protein $\mathbf{x}_i$ is now described by an additional variable ϕ_i, with $\phi_i = 1$ indicating that protein $\mathbf{x}_i$ belongs to a organelle derived component and $\phi_i = 0$ indicating that protein $\mathbf{x}_i$ is not well described by these known components. This outlier component is modelled as a multivariate T distribution with degrees of freedom κ, mean vector $\mathbf{M}$, and scale matrix V. Thus equation (2.48) becomes

$$\mathbf{x}_i|z_i = k, \phi_i \quad \sim \mathcal{N}(\boldsymbol{\mu}_k, \Sigma_k)^{\phi_i}\mathcal{T}(\kappa, \boldsymbol{M}, V)^{1-\phi_i}. \tag{2.50}$$

Further let $g(\mathbf{x}|\kappa, \mathbf{M}, \mathbf{V})$ denote the density of the multivariate T-distribution so that Equation (2.49) becomes:

$$p(\mathbf{x}_i|\boldsymbol{\theta}, \boldsymbol{\pi}, \phi_i, \kappa, \mathbf{M}, V) = \sum_{k=1}^{K} \pi_k \left(f(\mathbf{x}_i|\boldsymbol{\mu}_k, \Sigma_k)^{\phi_i} g(\mathbf{x}_i|\kappa, \boldsymbol{M}, V)^{1-\phi_i} \right). \tag{2.51}$$

For any i, we define the prior probability of the i-th protein belonging to the outlier component as $p(\phi_i = 0) = \epsilon$.

We can then rewrite equation (2.51) in the following way (by marginalising ϕ_i):

$$p(\mathbf{x}_i|\boldsymbol{\theta}, \boldsymbol{\pi}, \kappa, \epsilon, \mathbf{M}, V) = \sum_{k=1}^{K} \pi_k \left((1-\epsilon)(f(\mathbf{x}_i|\boldsymbol{\mu}_k, \Sigma_k) + \epsilon g(\mathbf{x}_i|\kappa, \boldsymbol{M}, V) \right). \tag{2.52}$$

This mimics the strategy of Hennig et al. [197] (see equation 2.27), whilst remaining a proper density. Indeed, ϵ can be inferred from the data and thus can be interpreted as a regular parameter rather than a pseudo-parameter. Throughout we take $\kappa = 4$, $\mathbf{M}$ as the global mean, and V as half the global variance of the data, including labelled and unlabelled proteins. The reason for formulating the model as in equation (2.51) is because it leads to a flexible modelling framework. Furthermore, ϕ has an elegant model selection interpretation, since it decides whether $\mathbf{x}_i$ is better modelled by the known components or the outlier component. It is important to note that f and g could be replaced by many combinations of distributions and thus could be valuable in modelling other datasets. The choice of parameters for the multivariate

T-distribution was decided so that it mimicked a multivariate normal component with the same mean and variance but with heavier tails to better capture dispersed proteins, which we refer to as outlier proteins throughout the text. The variance of the multivariate T-distribution is designed to be large such that is relatively flat when compared with multivariate Gaussian distributions which describe annotated components. We refer back to the section on robust mixture models for other strategies for modelling outliers in the literature.

2.3.3 Model fitting

We adopt a Bayesian approach toward inferring the unknown parameters, $\boldsymbol{\theta} = \{\boldsymbol{\mu}_k, \Sigma_k\}_{k=1}^K$, $\boldsymbol{\pi} = \{\pi_k\}_{k=1}^K$, and ϵ of the mixture model presented in Equation (2.51). For $\boldsymbol{\pi}$, we take a conjugate symmetric Dirichlet prior with parameter β, so that $\pi_1, \ldots, \pi_K \sim \text{Dirichlet}(\beta)$; and for the component-specific parameters $\boldsymbol{\mu}_k$ and Σ_k we take conjugate normal-inverse-Wishart (NIW) priors with parameters $\{\boldsymbol{\mu}_0, \lambda_0, \nu_0, S_0\}$, so that:

$$\mu_k, \Sigma_k \quad \sim \mathcal{N}\left(\boldsymbol{\mu}_k | \boldsymbol{\mu}_0, \frac{\Sigma_k}{\lambda_0}\right) I\mathcal{W}\left(\Sigma_k | \nu_0, S_0\right). \tag{2.53}$$

We also place a conjugate Beta prior on ϵ with parameters u and v, so that $\epsilon \sim \mathcal{B}(u, v)$. Allowing ϵ to be random allows us to infer the number of proteins that are better described by an outlier component rather than any known component.

The full model, which we henceforth refer to as a T-augmented Gaussian Mixture model (TAGM), can then be summarised by the plate diagram shown in Figure 2.1.

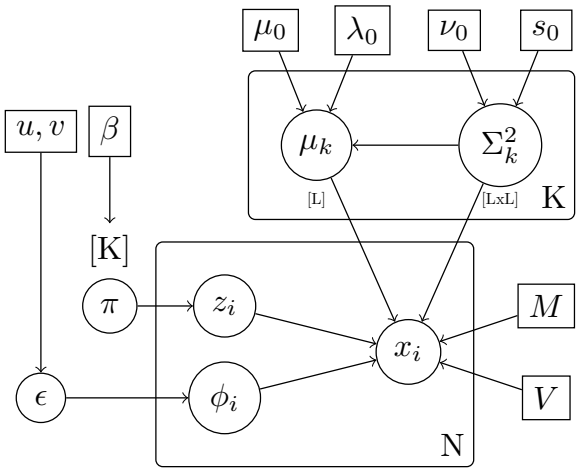


Fig. 2.1 Plate diagram for TAGM model. This diagram specifies the conditional independencies and parameters in our model [327].

To perform inference for the parameters, we make use of both the labelled and unlabelled data. For the labelled data X_L, since z_i and ϕ_i are known for these proteins, we can update the

parameters with their data analytically by exploiting conjugacy of the priors [see, for example, 162]. For the unlabelled data we do not have such information and so in the next sections we explain how to make inferences of the latent variables.

2.3.4 Prediction of localisation of unlabelled proteins

Having obtained the posterior distribution of the model parameters analytically using, at first, the labelled data only, we wish to predict the component to which each of the unlabelled proteins belongs. The probability that a protein belongs to any of the K known components, that is $z_i = k$ and $\phi_i = 1$, is given by (see appendix A.1 for derivations):

$$p(\phi_i = 1, z_i = k|\mathbf{x}_i, \boldsymbol{\theta}, \boldsymbol{\pi}, \epsilon, \kappa, \mathbf{M}, V) = \frac{\pi_k(1-\epsilon)f(\mathbf{x}_i|\boldsymbol{\mu}_k, \Sigma_k)}{\sum_{k=1}^{K}\pi_k\left((1-\epsilon)f(\mathbf{x}_i|\boldsymbol{\mu}_k, \Sigma_k) + \epsilon g(\mathbf{x}_i|\kappa, \mathbf{M}, V)\right)}, \quad (2.54)$$

whilst on the other hand,

$$p(\phi_i = 0, z_i = k|\mathbf{x}_i, \boldsymbol{\theta}, \boldsymbol{\pi}, \kappa, \epsilon, \mathbf{M}, V) = \frac{\pi_k\epsilon g(\mathbf{x}_i|\kappa, \mathbf{M}, V)}{\sum_{k=1}^{K}\pi_k\left((1-\epsilon)f(\mathbf{x}_i|\boldsymbol{\mu}_k, \Sigma_k) + \epsilon g(\mathbf{x}_i|\kappa, \mathbf{M}, V)\right)}. \quad (2.55)$$

Processing of the unlabelled data can be done by performing *maximum a posteriori* (MAP) estimation for the parameters. However, this approach fails to account for the uncertainty in the parameters, thus we additionally explore inferring the distribution over these parameters.

Maximum a posteriori prediction

We use the Expectation-Maximisation (EM) algorithm [108] to find *maximum a posteriori* (MAP) estimates for the parameters [see, for example, 327]. To specify the parameters of the prior distributions, we use the same choices as in the section on mixture modelling. By defining the following quantities:

$$\begin{aligned}
a_{ik} =& p(z_i = k, \phi_i = 1|\mathbf{x}_i), b_{ik} = p(z_i = k, \phi_i = 0|\mathbf{x}_i)\\
w_{ik} =& p(z_i = k|x_i) = a_{ik} + b_{ik},\\
a_k =& \sum_{i=1}^{n} a_{ik}, a = \sum_{k=1}^{K} a_k,\\
b_k =& \sum_{i=1}^{n} b_{ik}, b = \sum_{k=1}^{K} b_k,\\
r_k =& \sum_{i=1}^{n} w_{ik},
\end{aligned} \quad (2.56)$$

we can compute

$$\begin{aligned} \lambda_k =& \lambda_0 + a_k, \\ \nu_k =& \nu_0 + a_k, \\ m_k =& \frac{a_k \bar{\boldsymbol{x}}_k + \lambda_0 \mu_0}{\lambda_k}, \\ S_k^{-1} =& S_0^{-1} + \frac{\lambda_0 a_k}{\lambda_k}(\bar{\boldsymbol{x}}_k - \mu_0)^T(\bar{\boldsymbol{x}}_k - \mu_0) + \sum_{i=1}^{n} a_{ik}(x_i - \bar{\boldsymbol{x}}_k)^T(x_i - \bar{\boldsymbol{x}}_k). \end{aligned} \tag{2.57}$$

Then the parameters of the posterior mode are:

$$\begin{aligned} \hat{\mu}_k =& m_k, \\ \hat{\Sigma}_k =& \frac{1}{\nu_k + D + 2} S_k^{-1}. \end{aligned} \tag{2.58}$$

We note if $\boldsymbol{x}_i$ is a labelled protein then $a_{ik} = 1$ and these parameters can be updated without difficulty. The above equation constitutes a backbone of the E-step of the EM algorithm, with the entire algorithm specified by the following summary:

E-Step: Given the current parameters compute the values given by equations (2.56), with formulae provided in equations (2.54) and (2.55).

M-Step: Compute

$$\epsilon = \frac{u + b - 1}{(a + b) + (u + v) - 2},$$

and

$$\pi_k = \frac{r_k + \beta_k - 1}{N + \sum \beta_k - K},$$

as well as

$$\bar{\boldsymbol{x}}_k = \frac{1}{a_k}\left(\sum_{i=i}^{n} a_{ik}\mathbf{x}_i\right).$$

Finally, compute the MAP estimates given by equations (2.58). These estimates are then used in the following iteration of the E-step. Denoting by Q the expected value of the log-posterior and letting t denote the current iteration of the EM algorithm, we iterate until $|Q(\boldsymbol{\theta}|\boldsymbol{\theta}_t) - Q(\boldsymbol{\theta}|\boldsymbol{\theta}_{t-1})| < \delta$ for some pre-specified $\delta > 0$. Once we have found MAP estimates for the parameters $\boldsymbol{\theta}_{MAP}$, $\boldsymbol{\pi}_{MAP}$ and ϵ_{MAP} we proceed to perform prediction. We plug the MAP parameter estimates into Equation (2.54) in order to obtain the posterior probability of protein i localising to component k, $p(z_i = k, \phi = 1|\mathbf{x}_i, \boldsymbol{\theta}_{MAP}, \boldsymbol{\pi}_{MAP}, \epsilon_{MAP}, \kappa, \mathbf{M}, V)$. To make a final assignment, we may allocate each protein according to the component that has maximal

probability. A full technical derivation of the EM algorithm can be found in the appendix (appendix A.1).

Uncertainty in the posterior localisation probabilities

The MAP approach described above provides us with a probabilistic assignment, $p(z_i = k, \phi = 1|\mathbf{x}_i, \boldsymbol{\theta}_{MAP}, \boldsymbol{\pi}_{MAP}, \epsilon_{MAP}, \kappa, \mathbf{M}, V)$, of each unlabelled protein to each component. However, it fails to account for the uncertainty in the parameters $\boldsymbol{\theta}$, $\boldsymbol{\pi}$ and ϵ. To address this, we can sample parameters from the posterior distribution.

Let $\{\boldsymbol{\theta}^{(t)}, \boldsymbol{\pi}^{(t)}, \epsilon^{(t)}\}_{t=1}^{T}$ be a set of T sampled values for the parameters $\boldsymbol{\theta}$, $\boldsymbol{\pi}$, ϵ, drawn from the posterior. The assignment probabilities can then be summarised by the Monte-Carlo average:

$$p(z_i = k, \phi = 1|\mathbf{x}_i, \epsilon, \mathbf{M}, V) \approx T^{-1} \sum_{t=1}^{T} p(z_i = k, \phi = 1|\mathbf{x}_i, \boldsymbol{\theta}^{(t)}, \boldsymbol{\pi}^{(t)}, \epsilon^{(t)}, \kappa, \mathbf{M}, V).$$

Other summaries of the assignment probabilities can be determined in the usual ways to obtain, for example, interval-estimates. We summarise interval-estimates using the 95% equi-tailed interval, which is defined by the 0.025 and 0.975 quantiles of the distribution of assignment probabilities, $\{p(z_i = k, \phi = 1|\mathbf{x}_i, \boldsymbol{\theta}^{(t)}, \boldsymbol{\pi}^{(t)}, \epsilon^{(t)}, \mathbf{M}, V)\}_{t=1}^{T}$.

Sampling parameter values in our model requires us to compute the required conditional probabilities and then a straightforward Gibbs sampler can be used to sample in turn from these conditionals. In addition, we can bypass sampling the parameters by exploiting the conjugacy of our priors. By marginalising parameters in our model we can obtain an efficient collapsed Gibbs sampler and therefore only sample the component allocation probabilities and the outlier allocation probabilities. The derivations and required conditionals can be found in the appendix (appendix A.2).

2.4 Comparisons

We first concern ourselves with the predictive qualities of our proposed approach. To compare the classification performance of the two above learning schemes (MCMC and MAP estimation) to the K-nearest neighbours (KNN) and the support vector machine (SVM) classifiers.

We use the following standard schema to assess the classification performance of all methods. We split the marker sets for each experiment into a class-stratified training (80%) and test (20%) partitions, with the separation formed at random. The true classes of the test profiles are withheld from the classifier, whilst the algorithm is trained. The algorithm is then assessed on its ability to predict the classes of the proteins in the test partition for generalisation

accuracy. How each classifier is trained is specific to that classifier. The KNN and SVM have hyperparameters optimised using 5-fold cross-validation. This 80/20 data stratification is performed 100 times in order to produce 100 sets of macro-F1 [192] scores and class specific F1 scores [44]. The F1 score is the harmonic mean of the precision and recall, more precisely:

$$\text{precision} = \frac{tp}{tp + fp}, \text{recall} = \frac{tp}{tp + fn}.$$

tp denotes the number of true positives; fp the number of false positives and fn the number of false negatives. Thus

$$\text{F1} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}.$$

High Macro F1 scores indicate that marker proteins in the test dataset are consistently correctly assigned by the classifier. We note that accuracy alone is an inadequate measure of performance, since it fails to quantify false positives and is not adjusted for baseline prediction.

However, a Bayesian generative classifier produces probabilistic assignment of observations to classes. Thus, while the classifier may make an incorrect assignment it may do so with low probability. The F1 score is unforgiving in this situation and will not use this information. To measure this uncertainty, we introduce the quadratic loss (Brier Score) which allows us to compare probabilistic assignments [176]. For these comparisons, we use the probabilistic interpretations of the SVM and K-NN algorithm introduced in previous sections. The quadratic loss is given by the following formula:

$$Q_2 = \sum_{i=1}^{N} \|q_i - p_i\|_2^2, \tag{2.59}$$

where $\|\cdot\|_2$ is the l_2 norm and q_i is the true classification vector and p_i is a vector of predicted assignments to each class. It is useful to note that the corresponding risk function is the mean square error (MSE), which is the expected value of the quadratic loss.

It is desirable to compute these metrics not only for a single dataset but several, so that we can see that the approach is robust across similar but varying experimental designs. For the KNN algorithm, the number of nearest neighbours, is optimised via an additional internal 5-fold cross-validation and the hyperparameters for the SVM, sigma and cost, are also optimised via internal 5-fold cross validation [208].

We test our methods on the following datasets: *Drosophila* [448], chicken [185], mouse pluripotent embryonic stem cells from [68] and [44], the human bone osteosarcoma epithelial (U2-OS) cell line [455], the HeLa cell line of [220], the 3 HeLa cell lines from [202] and 10 primary fibroblast datasets from [24]. These datasets represent a great variety of spatial proteomics experiments across many different workflows.

The two *hyper*LOPIT datasets on mouse pluripotent embryonic stem cells and the U-2 OS cell line use TMT 10-plex labelling and contain the greatest number of proteins. Earlier LOPIT experiments on the *Drosophila* and chicken use iTRAQ 4-plex labelling, whilst another LOPIT mouse pluripotent embryonic stem cell dataset uses iTRAQ 8-plex. The datasets of [220] and [202] employ a different methodology completely - separating cellular content using differential centrifugation (as opposed to along a density-gradient). Furthermore, the methods use SILAC rather than iTRAQ or TMT for labelling. The experiments of [202] were designed to explore the functional role of AP-5 by coupling CRISPR-CAS9 knockouts with spatial proteomics methods. We analysed all three datasets from [202], which includes a wild type HeLa cell line as a control, as well as two CRISPR-CAS9 knockouts: AP5Z1-KO1 and AP5Z1-KO2 respectively.

In addition, we analyse the spatio-temporal proteomics experiments of [24], which uses TMT-based MS quantification. This experiment explored infecting primary fibroblasts with Human cytomegalovirus (HMCV) and the goal of these experiments was to explore the dynamic perturbation of host proteins during infection, as well as the sub-cellular localisation of viral proteins throught the HCMV life-cycle. They produced spatial maps at different time points: $24, 48, 72, 96, 120$ hours post infection (hpi), as well as mock maps at these same time points to serve as a control - this results in 10 different spatial proteomics maps.

In each case, a dataset specific marker list was used, which is curated specifically for the each cell line. We removed "high-curvature ER" annotations from the HeLa dataset [220], as well as the "ER Tubular", "Nuclear pore complex" and "Peroxisome" annotations from the HeLa CRISPR-CAS9 knockout experiments [202] as there are too few proteins to correctly perform cross-validation. Table 2.1 summarises these datasets, including information about number of quantified proteins, the workflow used and the number of fractions.

Figure 2.2 compares the Macro-F1 scores across the datasets for all classifiers and demonstrates that no single classifier consistently outperforms any other across all datasets, with results being highly consistent across all methods, as well as across datasets. We perform a pairwise unpaired t-test with multiple testing correction applied using the Benjamini-Höchberg procedure [25] to detect differences between classifier performance.

In the *Drosophila* dataset only the KNN algorithm outperforms the SVM at significance level of 0.01, whilst no other significant differences exist between the classifiers. In the chicken DT40 dataset only the MCMC method outperforms the KNN classifier at significance level of 0.01, no other significant conclusion can be drawn. In the mouse dataset the MAP based method outperforms the MCMC method at significance level of 0.01, no other significant conclusions can be drawn. In the HeLa dataset all classifiers are significantly different at a 0.01 level. These differences may exist because the dataset does not fit well with our modelling assumptions; in particular, this dataset set has been curated to have a class called "Large Protein Complex", which likely describes several sub-cellular structures. These might include nuclear compartments

MS-based Spatial Proteomics datasets				
Cell line or organism	Workflow	Labelling	Fractions (including combined replicates)	Proteins
Drosophila	LOPIT	iTRAQ	4	888
Chicken DT40	LOPIT	iTRAQ	16	1090
Mouse pluripotent E14TG2a stem cell	HyperLOPIT	TMT	20	5032
HeLa (Itzhak et al.)	Organeller Maps	SILAC	30	3766
HeLa (Hirst et al.)	Organeller Maps	SILAC	15	2046
U-2 OS cell line	HyperLOPIT	TMT	37	5020
Primary Fibroblast	Spatio-Temporal Methods	TMT	6	2196
E14TG2a (Breckels et al.)	LOPIT	iTRAQ	8	2031

Table 2.1 Summary of spatial proteomics datasets used for comparisons

and ribosomes, as well as any cytosolic complex and large protein complexes which pellet during the centrifugation conditions used to capture this mixed sub-cellular fraction. Moreover, the cytosolic and nuclear fraction were processed separately leading to possible imbalance with comparisons with other datasets. Thus, the large protein complexes component might be better described as itself a mixture model or more detailed curation of these data may be required. We do not consider further modelling of this dataset in this chapter. For the U-2 OS all classifiers are significantly different at a significance level of 0.01 except for the SVM classifier and the MCMC method, with the MAP method performing the best. Figure 2.2 shows that for this dataset all classifiers are performing extremely well. In the three Hirst datasets the MAP method significantly outperforms all other methods ($p < 0.01$), whilst in the wild type HeLa and in the CRISPR-CAS9 KO1 there is no significant difference between the KNN and MCMC method. In the CRISPR-CAS9 KO2 the MCMC method outperforms the SVM and KNN methods ($p < 0.01$). In the interest of brevity, the remaining results for the t-tests can be found in tables in appendix A.5.

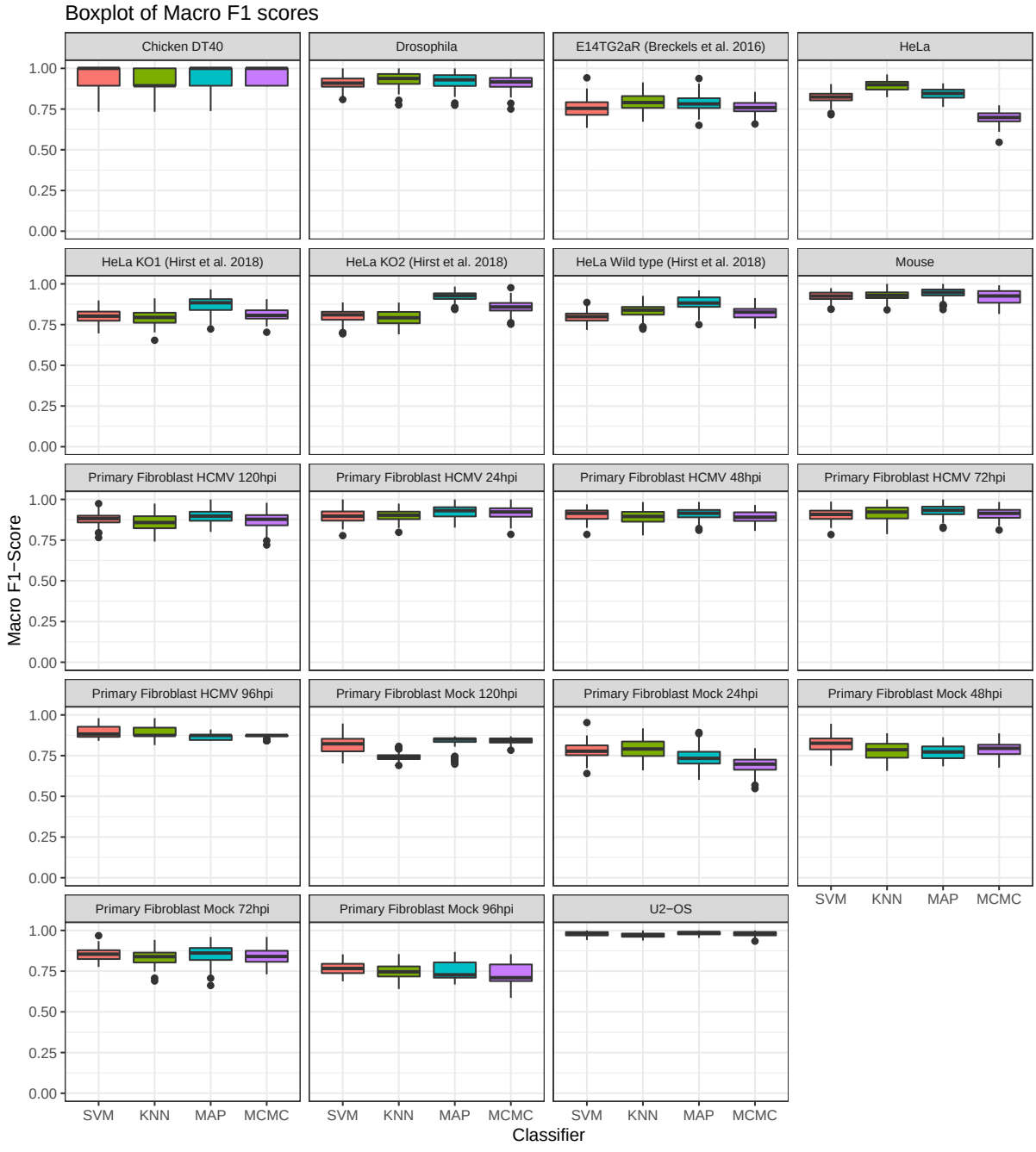


Fig. 2.2 Boxplots of the distributions of Macro F1 scores for all spatial proteomics datasets.

The Macro-F1 scores do not take into account that whilst the TAGM model may misclassify, it may do so with low confidence. We therefore additionally compute the quadratic loss, which allows us to make use of the probabilistic information provided by the classifiers. The lower the quadratic loss the closer the probabilistic prediction is to the true value. We plot the distributions of quadratic losses for each classifier in figure 2.3. We observe highly consistent performance across all classifiers across all datasets. Again, we perform a pairwise unpaired t-test with multiple testing correction.

We find that in 16 out of 19 datasets (all of those except HeLa Wild type, HeLa KO1 and HeLa KO2) the MCMC methods achieves the lowest quadratic loss at a significance level < 0.0001 over the SVM and KNN classifiers. In 6 out of these 16 datasets there is no significant difference between the MCMC and the MAP methods. In the three Hirst datasets in which the MCMC did not achieve the lowest quadratic loss, the SVM outperformed. However, in two of these datasets (HeLa Wild type and KO1) the MAP method and SVM classifier were not significantly different. In the Hirst KO2 dataset there were no significant differences between the MAP and MCMC methods.

In the vast majority of cases, we observe that if the TAGM model, using the MCMC methodology, makes an incorrect classification it does so with lower confidence than the SVM classifier, the KNN classifier and the MAP based classifier, whilst if it is correct in its assertion it does so with greater confidence. Additionally, a fully Bayesian methodology provides us with not only point estimates of classification probabilities but uncertainty quantification in these allocations, and we show in the following section that this provides deeper insights into protein localisation. The lack of stability in the performance of the SVM might draw some concern. A clear limitation of the SVM is the conversion of scores to probabilities. If this conversion is unmerited then this results in poor downstream inferences. Indeed, we note that for the more recent datasets, which usually have more annotated sub-cellular niches, the performance is worse.

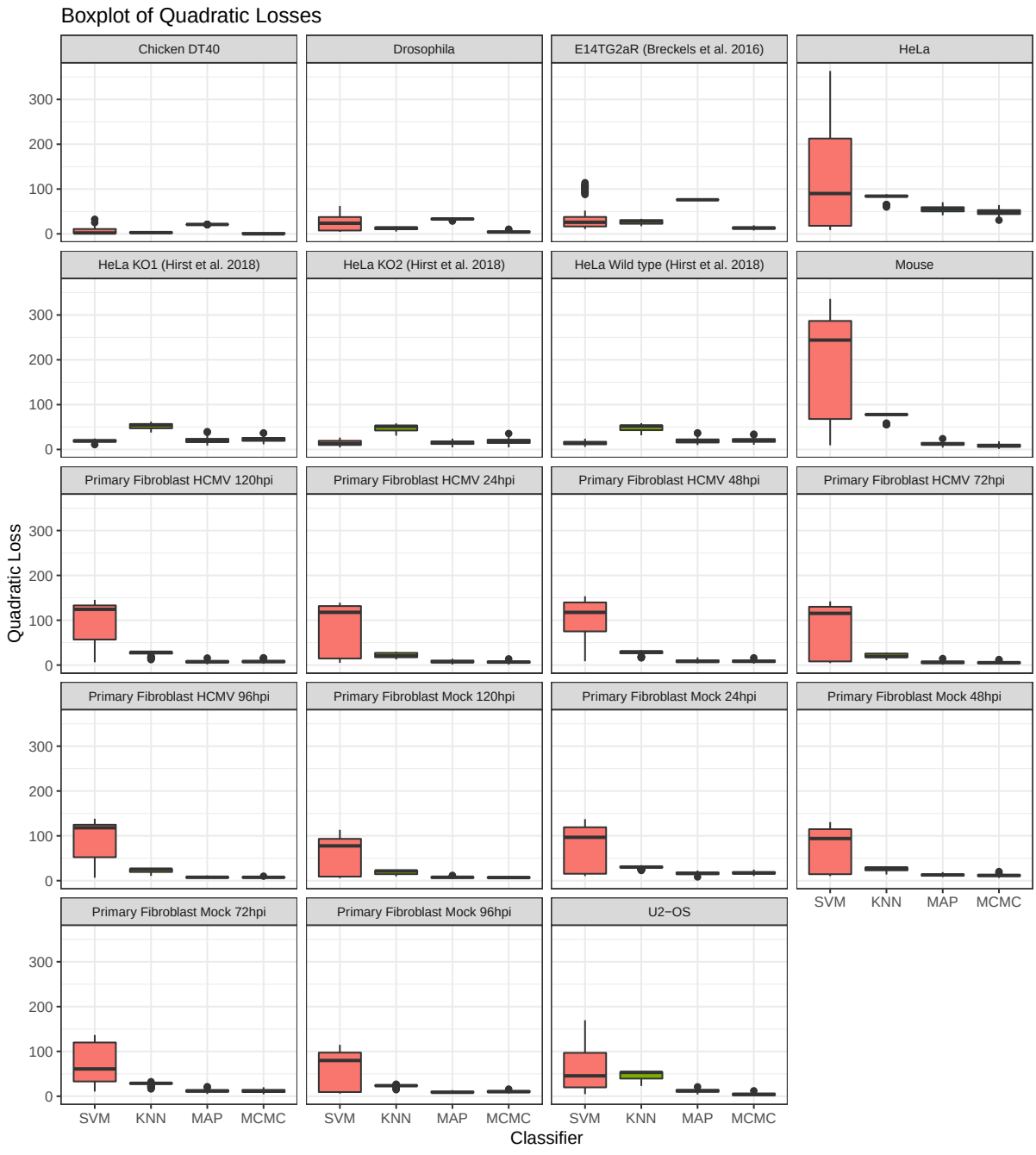


Fig. 2.3 Boxplots of the distributions of Quadratic losses for all spatial proteomics datasets.

Computing distributions of F1 scores and quadratic losses, which can only be done on the marker proteins, can help us understand whether a classifier might have greater generalised performance accuracy. However, we are interested in whether there is a large disagreement between classifiers when prediction is performed on proteins for which we have no withheld localisation information. This informs us about a systematic bias for a particular classifier or whether a classifier ensemble could increase performance. To this end, we examine the SVM and TAGM-MCMC results for the mESC dataset [68] more closely. To maintain a common set of proteins, we set thresholds for each classifier in turn and compare to the other classifier without thresholding. Firstly, we set a global threshold of 0.95 for the TAGM-MCMC and then for these proteins plot a contingency table against the classification results from the SVM. Secondly, we set a 5% FDR for the SVM and then for these proteins plot a contingency table against the classification results from the TAGM-MCMC. We visualise the contingency tables as heat plots in figure 2.4.

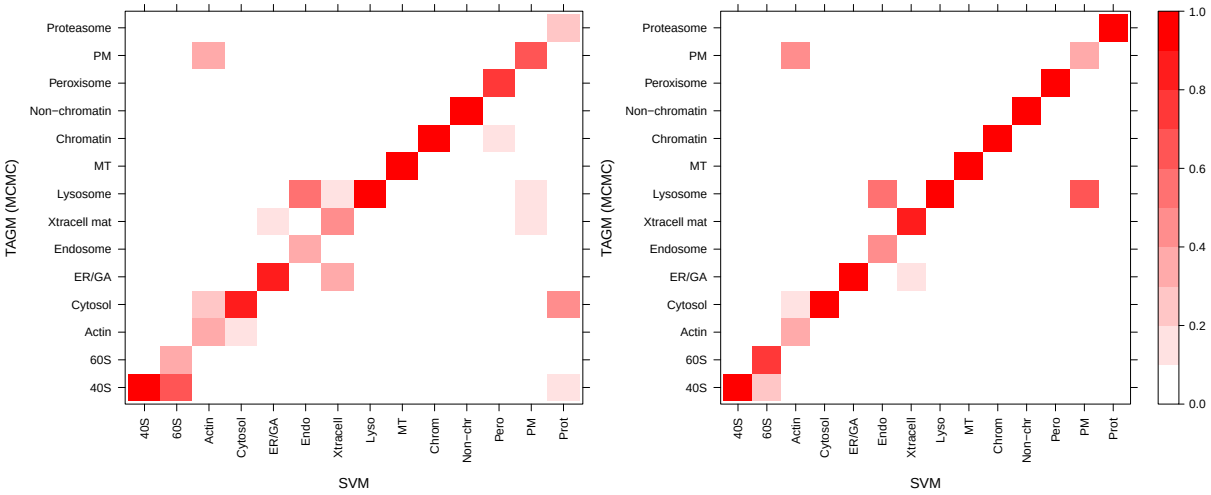


Fig. 2.4 A heatmap representation of a contingency table, where we compare assignment results for proteins with unknown protein localisation using the TAGM-MCMC and SVM on the mESC dataset. The scale ranges from 0 to 1 with values indicating the proportion of assigned proteins to that sub-cellular location. Values along the diagonal represent agreement between classifiers whilst other values represent disagreement. The coherence between the classifiers is very high. (a) In this case we set a probability threshold of 0.95 for the TAGM assignments with no threshold for the SVM. (b) In this case we set a 5% FDR threshold for the SVM and no threshold for the TAGM-MCMC.

In general, we see an extremely high level of coherence between the TAGM and the SVM, with almost all proteins predicted to concordant sub-cellular compartments. Figure 2.4 shows there is some disagreement between assigning proteins to the lysosome and plasma membrane, to the cytosol and proteasome, and between the large and small ribosomal subunits. However,

we have not used the uncertainty in the probabilistic assignments to produce the contingency tables above. In the next sections, we explore examples of proteins with uncertainty in their posterior localisation probabilities. Selecting biologically relevant thresholds is important for any classifier and exploring uncertainty is of vital importance when drawing biological conclusions.

2.5 Case study: mouse pluripotent embryonic stem cells

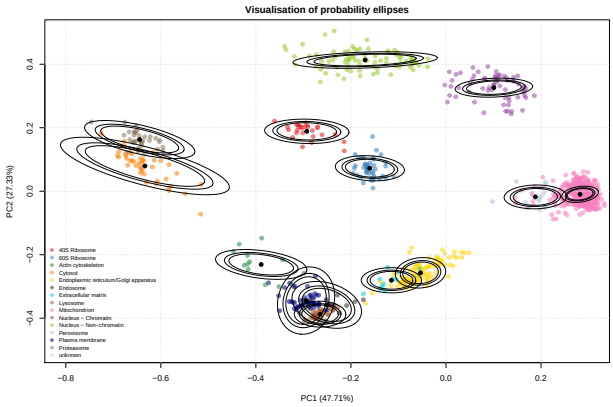
Pluripotency is the ability of a cell to differentiate into multiple germ layers: the endoderm (intestinal tract), mesoderm (muscle, bone, blood) and ectoderm (nervous system) [41, 428]. A cell's potency is toggled by molecular cues; such as, transcriptional regulation [62], epigenetic imprints [353] and gene regulatory networks [132]. However, these processes are yet not fully understood. mES cells are derived from blastocysts that then transition to differentiation to become an ensemble of cell types [41]. Mounting evidence suggest a role for post-transcriptional regulation of pluripotency [404, 51, 109, 420]. The path from self-renewal to differentiation involves dramatic changes to the cell's morphological features, implicating intracellular organisation and compartmentalisation as key factors. Hence, the analysis of the spatial proteome of mESCs is of paramount importance to understand the molecular basis for pluripotency.

Having establish that our method has excellent predictive performance on many data datasets. We wish to model mouse pluripotent embryonic stem cell (E14TG2a) data [68] to demonstrate our approach. This dataset contains quantitation data for 5032 proteins. This high-resolution map was produced using the *hyper*LOPIT workflow [324], which uses a sophisticated sub-cellular fractionation scheme. This fractionation scheme is made possible by the use of Tandem Mass Tag (TMT) 10-plex and high accuracy TMT quantification was facilitated by using a mass spectrometry approach that uses synchronous precursor selection MS3 (SPS-MS3) [298], which reduces well documented issues with ratio distortion in isobaric multiplexed quantitative proteomics [457]. The data resolves 14 sub-cellular niches with an additional chromatin preparation resolving the nuclear chromatin and non-chromatin components. Two biological replicates of the data are concatenated, each with 10 fractions along the density gradient. We defined gold standard organelle markers as those with unambiguous single annotation [155]. A protein marker list for the mESCs was manually curated using information from the UniProt database, the Gene Ontology and the literature, as was performed in [68]. The following section applies our statistical methodology to these data and we explore the results.

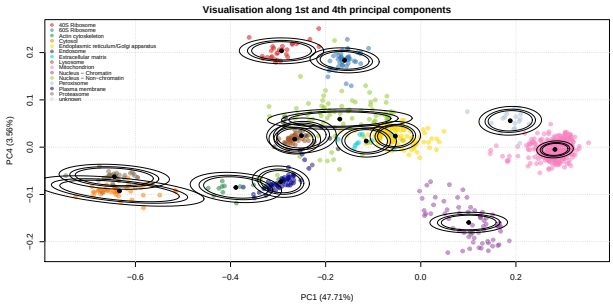
Maximum a posteriori prediction of protein localisation

We first derive MAP estimates for the model parameters of the TAGM model and use these for prediction. Visualisation is important for data analysis and exploration. A simple way to

visualise our model is to project probability ellipses onto a PCA plot, where the ellipse is obtained by evaluating the multivariate Gaussian at θ_{MAP} and then projecting into PCA coordinates. Note that this is different from the posterior distribution of the allocation probabilities which will not, in general, be elliptical. Each ellipse contains a proportion of total probability of a particular multivariate Gaussian density. The outer ellipse contains 99% of the total probability whilst the middle and inner ellipses contain 95% and 90% of the probability respectively. Visualising only the first two principal components can be misleading, since proteins can be more (or less) separated in subsequent principal components. We visualise the first two principal components along with the first and fourth principal components as a representative example.



(a)



(b)

Fig. 2.5 (a) PCA plot of the 1st and 2nd principal components for the curated marker proteins of the mouse stem cell data. The organelles are, in general, well separated. Though some organelles overlap, they are separated along different principal components. The densities used to produce the ellipses are derived from the MAP estimates. (b) Marker resolution along the 1st and 4th principal components show that the mitochondrion and peroxisome markers are well resolved, despite overlapping in the 1st and 2nd component. We also see that the ER/Golgi apparatus markers are better separated from the extracellular matrix markers.

We now apply the statistical methodology described in section 2.3.2, to predict the localisation of proteins to organelles and sub-cellular components. In brief, we produce MAP estimates of the parameters by using the expectation-maximisation algorithm, to form the basis

of a Bayesian analysis (TAGM-MAP). We run the algorithm for 200 iterations and inspect a plot of the log-posterior to assess convergence of the algorithm (see appendix A.3). We confirm that the difference of the log posterior between the final two iterations is less than 10^{-6} and we conclude that our algorithm has converged. The results can be seen in figure 2.6 (left), where the posterior localisation probability is visualised by scaling the pointer for each protein.

Figure 2.6 (right) demonstrates a range of probabilistic assignments of proteins to organelles and sub-cellular niches. We additionally consider a full, sampling-based Bayesian analysis using Markov-chain Monte Carlo (MCMC) to characterise the uncertainty in the localisation probabilities. As explained previously a collapsed Gibbs sampler is used to sample from the posterior of localisation probabilities. The remainder of this chapter focuses on analysis of spatial proteomics in this fully Bayesian framework.

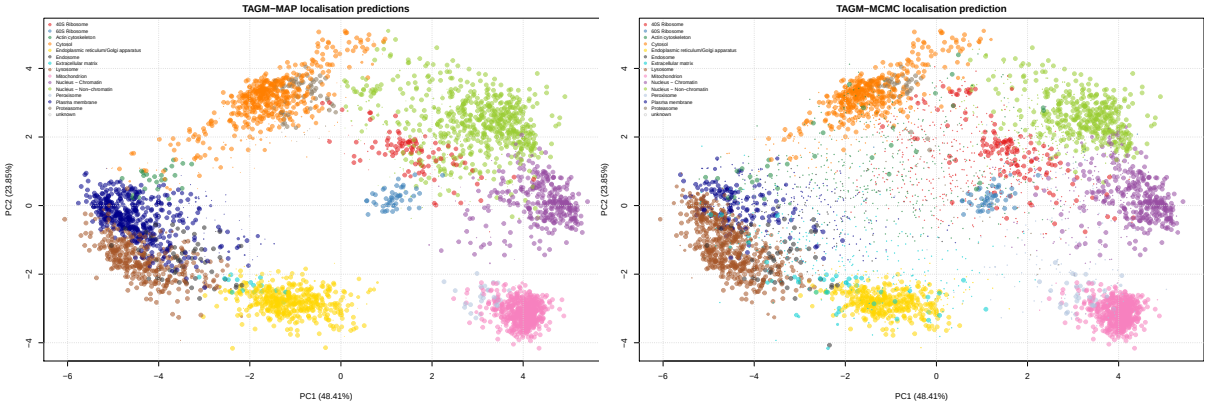


Fig. 2.6 PCA plot of the protein quantitation data with colours representing the predicted class (5032 proteins) illustrating protein localisation predictions using TAGM-MAP (left) and TAGM-MCMC (right) respectively. The pointer size of a protein is scaled to the probability that particular protein was assigned to that organelle. Markers, proteins whose localisations are already known, are automatically assigned a probability of 1 and the size of the pointer reflects this.

Quantifying the uncertainty in the posterior localisation probabilities

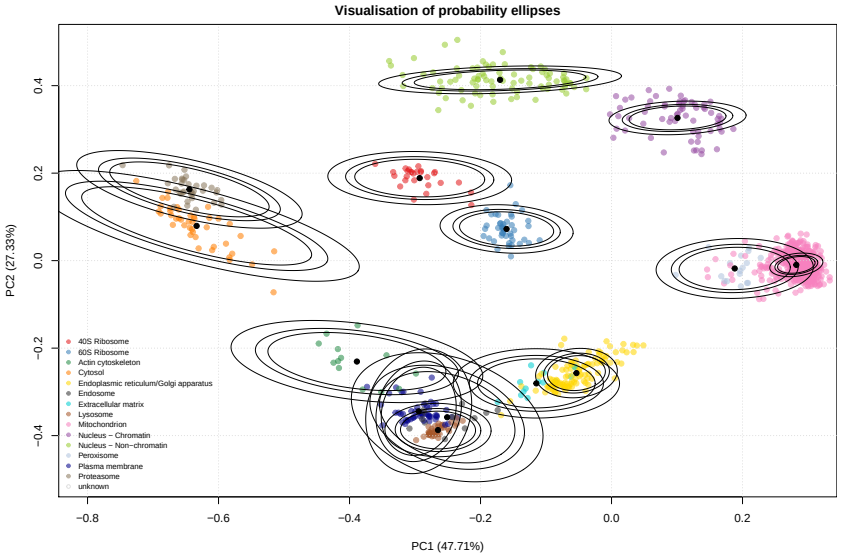
This section applies the TAGM model to the mESC data, by considering the uncertainty in the parameters and exploring how this uncertainty propagates to the uncertainty in protein localisation prediction. In figure 2.7, we visualise the model as before using the first two principal components along with the first and fourth principal component as a representative example. For the TAGM model, we derive probability ellipses from the expected value (Monte-Carlo estimator) of the posterior normal-inverse-Wishart (NIW) distribution.

We apply the statistical methodology detailed in section 2.3.2. Firstly, we perform posterior computation in the Bayesian setting using standard MCMC methods (TAGM-MCMC). We run

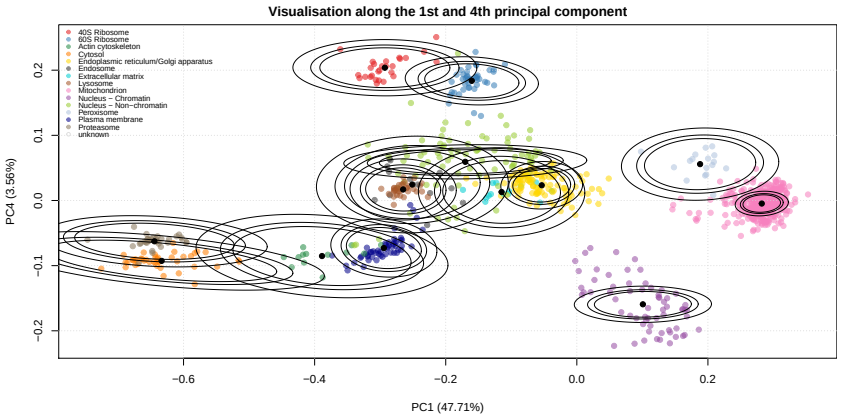
6 chains of our Gibbs sampler in parallel for $15,000$ iterations, throwing away the first $4,000$ iterations for burn-in and retain every 10^{th} sample for thinning. Thus 1,100 sample are retained from each chain. We then visualise the trace plots of our chains; in particular, we monitor the number of proteins allocated to the known components (see appendix A.4). We discard 1 chain because we do not consider it to have converged. For the remaining 5 chains, we further discard the first 500 samples by visual inspection. We then have 600 retained samples from 5 separate chains. For further analysis, we compute the Gelman-Rubin convergence diagnostic [163, 46], which is computed as $\hat{R} \approx 1.05$. Values of $\hat{R}$ far from 1 indicate non-convergence and since our statistic is less than 1.1, we conclude our chains have converged. The remaining samples are then pooled to produce a single chain containing 3000 samples.

We produce point estimates of the posterior localisation probabilities by summarising samples by their Monte-Carlo average. These summaries are then visualised in figure 2.6 (right panel), where the pointer is scaled according to the localisation probabilities of the sub-cellular niche with the largest posterior probability. Monte-Carlo based inference also provides us with additional information; in particular, we can interrogate individual proteins and their posterior probability distribution over sub-cellular locations.

Figure 2.8 illustrates a clear example of the importance of capturing uncertainty. The E3 ubiquitin-protein ligase TRIP12 (G5E870) is an integral part of ubiquitin fusion degradation pathway and is a protein of great interest in cancer because it regulates DNA repair pathways. The SVM failed to assign this protein to any location, with assignment to the 60S Ribosome falling below a 5% FDR and the MAP estimate assigned the protein to the nucleus non-chromatin with posterior probability < 0.95. The posterior distribution of localisation probabilities inferred from the TAGM-MCMC model, shown in figure 2.8, demonstrates that this protein is most probably localised to the nucleus non-chromatin. However, there is some uncertainty about whether it localises to the 40S ribosome. This could suggest a dynamic role for this protein, which could be further explored with a more targeted experiment.



(a)



(b)

Fig. 2.7 (a) Probability ellipses produced from applying the MCMC method. The density is derived from the expected value of the NIW distribution. (b) Probability ellipses visualised along the 1st and 4th principal component, also from the MCMC method.

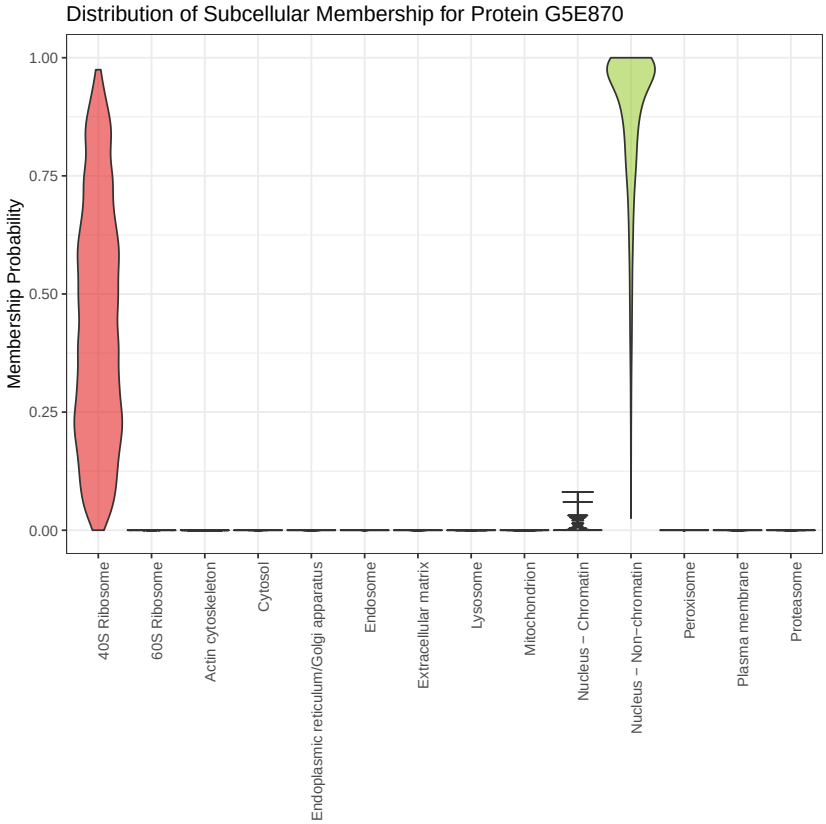


Fig. 2.8 A violin plot visualising the posterior distribution of localisation probabilities of protein E3 ubiquitin-protein ligase (G5E870) to organelles and sub-cellular niches. The most probable localisation is nucleus non-chromatin, however there is uncertainty associated with this assignment.

Enrichment analysis of outlier proteins

In previous sections, we demonstrated that we can assign proteins probabilistically to sub-cellular compartments and quantify the uncertainty in these assignments. Some proteins cannot be well described as belonging to any annotated component and we model this using an additional T-distribution outlier component (see Section 2.3.2).

It is biologically interesting to decipher what functional role proteins that are far away from known components play. We perform an over-representation analysis (hyper-geometric test) of gene ontology (GO) terms to asses the biological relevance of the outlier component [39, 489]. We take 1111 proteins that were allocated to known components with probability less than 0.95. Note that these 1111 proteins exclude proteins that are likely to belong to a known location, but we are uncertain about which localisation. We then perform enrichment analysis against the set of all proteins quantified in the *hyper*LOPIT experiment. We search against the cellular compartment, biological process and molecular function ontologies.

Supplementary figure A.3 shows this outlier component is enriched for cytoskeletal part ($p < 10^{-7}$) and microtubule cytoskeleton ($p < 10^{-7}$). Cytoskeleton proteins are found throughout the cell and therefore we would expect them to be found in every fraction along the density gradient, with no characteristic buoyant density. We also observe enrichment for highly dynamic sub-cellular process such as cell division ($p < 10^{-6}$) and cell cycle processes ($p < 10^{-6}$), again these proteins are unlikely to have steady-state locations within a single component. We also see enrichment for molecular functions such as transferase activity ($p < 0.005$), another highly dynamic process. These observations justify including an additional outlier component in our mixture model, as these proteins are unlikely to be captured by any single component.

Interpreting and exploring uncertainty

Protein sub-cellular localisation can be uncertain for a number of reasons. Technical variations and unknown biological novelty, such as yet uncharacterised functional compartments, can be some of the reasons why a protein might have an unknown or uncertain localisation. Furthermore many proteins are known to reside in multiple locations with possibly different functional duties in each location (referred to as *moonlighting* in the literature) [231]. With these considerations in mind, it is pertinent to quantify the uncertainty in our allocation of proteins to organelles. This section explores several situations where proteins display uncertain localisation and considers the biological factors that influence uncertainty. We later explore and visualise whole proteome uncertainty quantification.

Exportin 5 (Q924C1) forms part of the micro-RNA export machinery of the nucleus, transporting miRNA from the nucleus to the cytoplasm for further processing. It then translocates back through the nuclear pore complex to return to the nucleus. Exportin 5 can then continue to mediate further transport between nucleus and cytoplasm. The SVM was unable to assign a localisation of Exportin 5, with its assignment falling below a 5% FDR to wrongly assign this protein to the proteasome. This incorrect assertion by the SVM was confounded by the similarity between the cytosol and proteasome profiles. Figure 2.9 demonstrates, according to the TAGM-MCMC model, that Exportin 5 most likely localises to the cytosol but there is some uncertainty with this assignment. This uncertainty is reflected in possible assignment of Exportin 5 to the nucleus non-chromatin and this uncertainty is a manifestation of the fact that the function of this protein is to shuttle between the cytosol and nucleus.

The Phenylalanine–tRNA ligase beta subunit protein (Q9WUA2) has an uncertain localisation between the 40S ribosome and the nucleus non-chromatin demonstrated in figure 2.10. This protein was left unclassified by the SVM because its score fell below a 5% FDR threshold to assign it to the 40S ribosome. Considering that this protein is involved in the acylation of transfer RNA (tRNA) with the amino acid phenylalanine to form tRNA–Phe to be used in translation of proteins, it is therefore unsurprising that this protein's stead-state location is ribosomal. Whilst the SVM is unable to make an assignment, TAGM-MCMC is able to suggest an assignment and quantify the uncertainty.

Relatively little is known about the Dedicator of cytokinesis (DOCK) protein 6 (Q8VDR9), a guanine nucleotide exchange factor for CDC42 and RAC1 small GTPases. The SVM could not assign localisation to the ER/Golgi, since its score fell below a 5% FDR. Furthermore, the TAGM-MCMC model assigned this DOCK 6 to the outlier component with posterior probability > 0.95. Figure 2.11 shows possible localisation to several components along the secretory pathway. As an activator for CDC42 and RAC1 we may expect to see them with

similar localisation. CDC42, a plasma membrane associated protein, regulates cell cycle and division and is found with many localisations. Furthermore RAC1, a small GTPase, also regulates many cellular processes and is found in many locations. Thus the steady-state distribution of DOCK6 is unlikely to be in a single location, since its interaction partners are found in many locations. This justifies including an outlier component in our model, else we may erroneously assign such proteins to a single location.

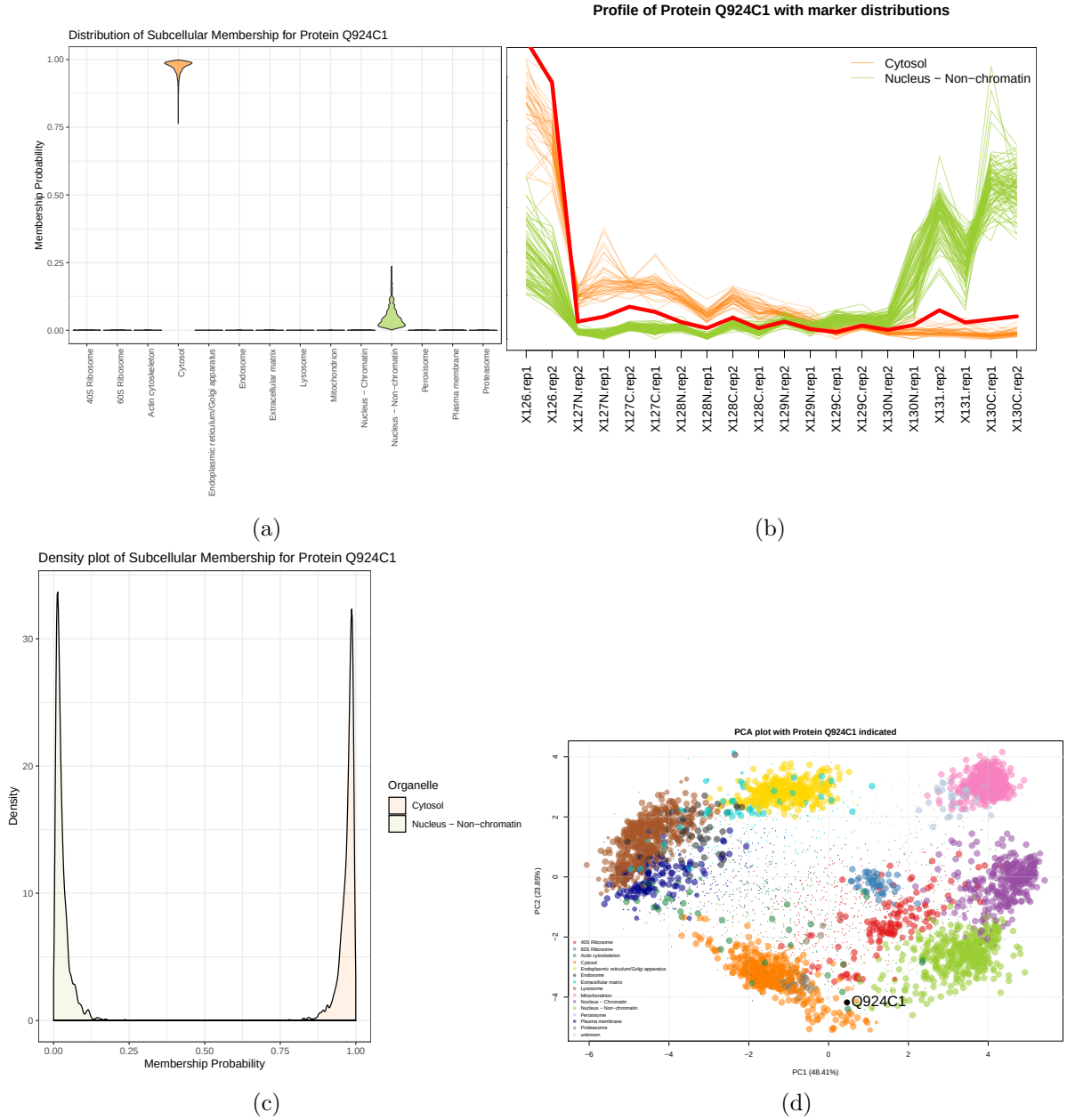


Fig. 2.9 Exportin 5 (Q924C1) showing localisation to the cytosol with some uncertainty about association to the nucleus non-chromatin. (a) The violin plot shows uncertain localisation between these two sub-cellular localisations. (b) The quantitative profile of this protein shows mixed profile between the profiles of the organelle markers. (c) The density plot shows a complex distribution over localisations for this protein. (d) The protein Q924C1 has steady-state distribution between the cytosol and nucleus non-chromatin.

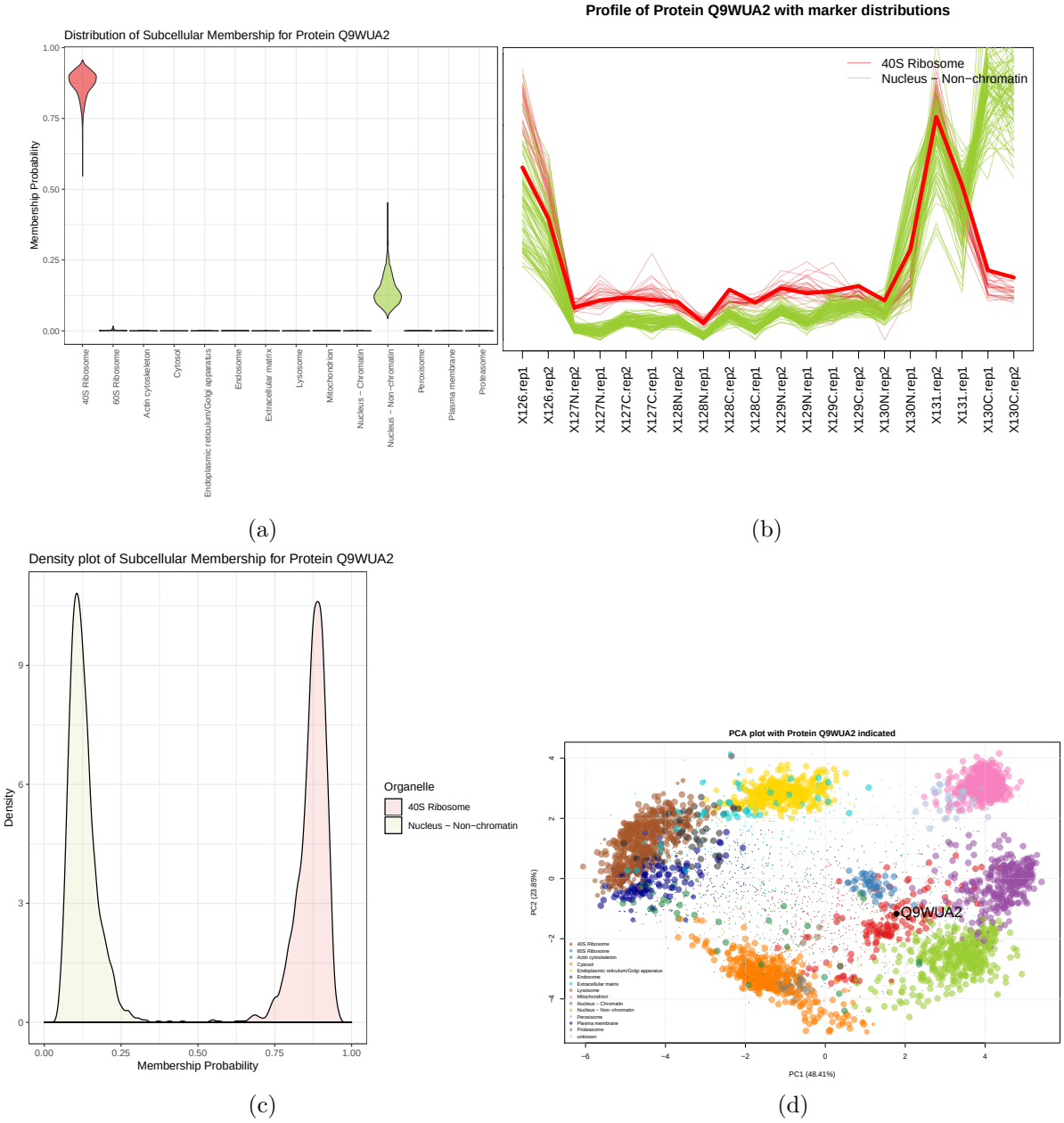


Fig. 2.10 Phenylalanine-tRNA ligase beta subunit protein TRIP12 (Q9WUA2) showing localisation to the 40S Ribosome with some uncertainty about association to the nucleus non-chromatin. (a) The violin plot shows uncertain localisation between these two sub-cellular localisations. (b) The quantitative profile of this protein shows mixed profile between the profiles of the organelle markers. (c) The density plot shows a complex distribution over localisations for this protein. (d) The protein Q9WUA2 has steady-state distribution skewed towards the 40S Ribosome and close to the nucleus non-chromatin.

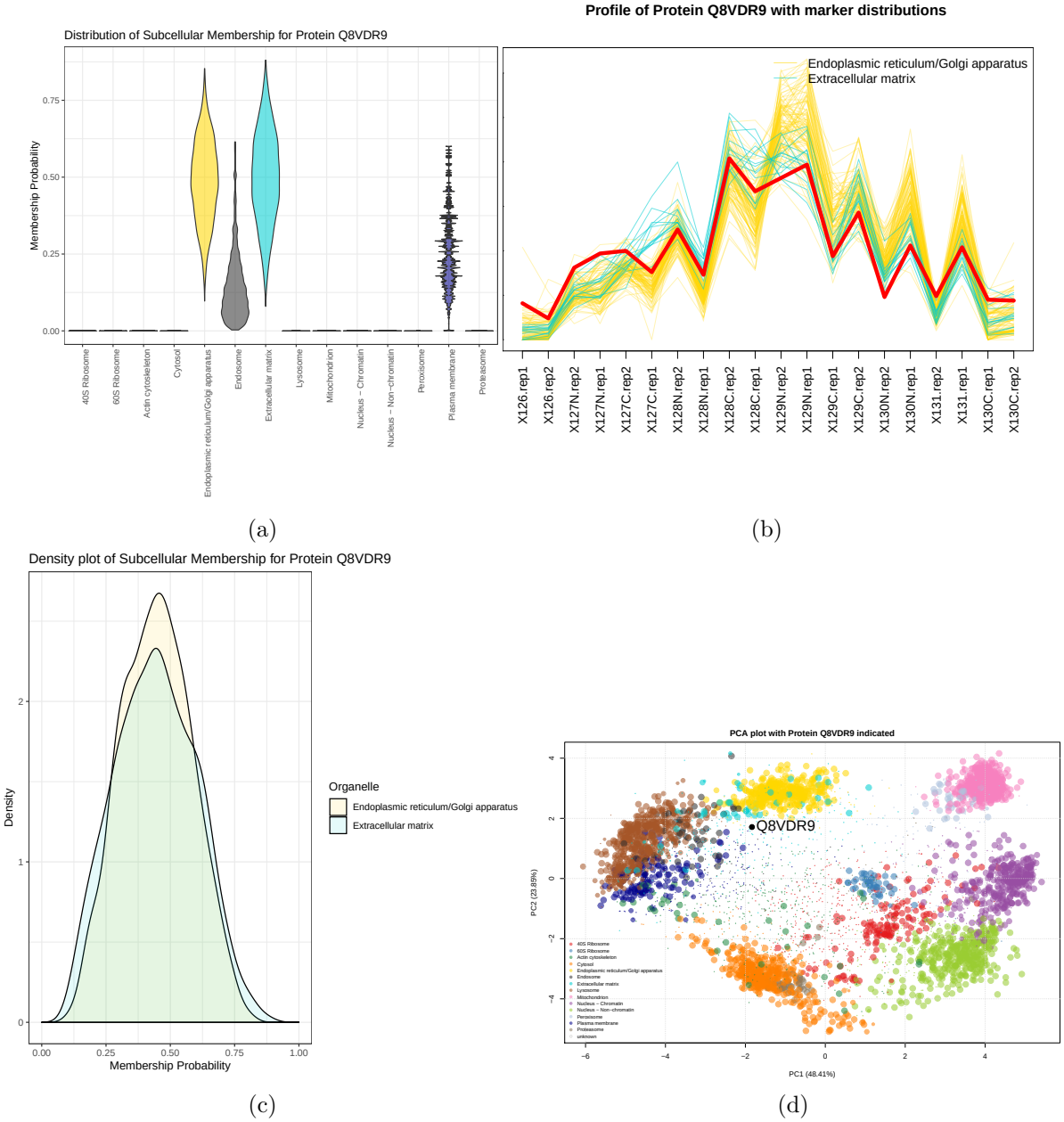


Fig. 2.11 Q8VDR9 showing localisation to the outlier component. (a) The violin plot shows uncertain localisation between several sub-cellular niches. (b) The quantitative profile of this protein shows mixed profile between the profiles of the organelle markers. (c) The density plot shows a similar localisation probabilities for both the ER/Golgi and Extracellular matrix. (d) The protein Q8VDR9 has steady-state distribution in the centre of the plot skewed toward the secretory pathway; in particular, the ER/Golgi and Extracellular matrix components.

Visualising whole sub-cellular proteome uncertainty

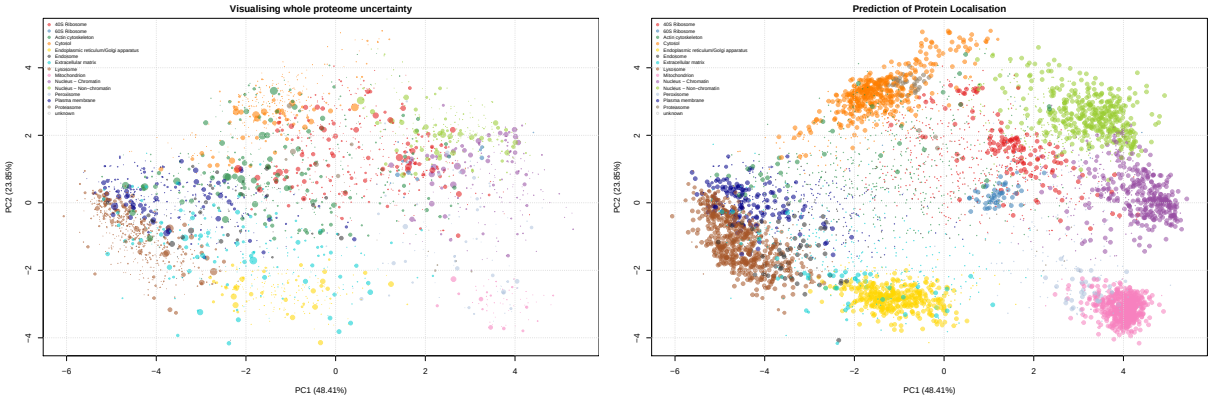
The advantage of the TAGM-MCMC model is its ability to provide proteome wide uncertainty quantification. Regions where organelle assignments overlap are areas where uncertainty is expected to be the greatest, as well as areas with no dominant component. We take an information theoretic approach to summarising uncertainty in protein localisation by computing the Shannon entropy [415] for each Monte-Carlo sample $t = 1, ..., T$ of the posterior localisation probabilities of each protein

$$\left\{ H^{(t)} = -\sum_{k=1}^{K} p_{ik}^{(t)} \log\left(p_{ik}^{(t)}\right) \right\}_{t=1}^{T}, \tag{2.60}$$

where $p_{ik}^{(t)}$ denotes the posterior localisation probability of protein i to component k at iteration t. We then summarise this as a Monte-Carlo averaged Shannon entropy. The greater the Shannon entropy the more uncertainty associated with the assignment of this protein. The lower the Shannon entropy the lower the uncertainty associated with the assignment of this protein. In figure 2.12 panel (a), we visualise the Shannon entropy of each protein in a PCA plot, by scaling the pointer in accordance to this metric. We also note that while localisation probability (of a protein to its most probable location) and the Shannon entropy are correlated, figure 2.12 panel (c), it is by no means perfect. Thus it is important to use both the localisation probabilities and the uncertainty in these assignments to make conclusions.

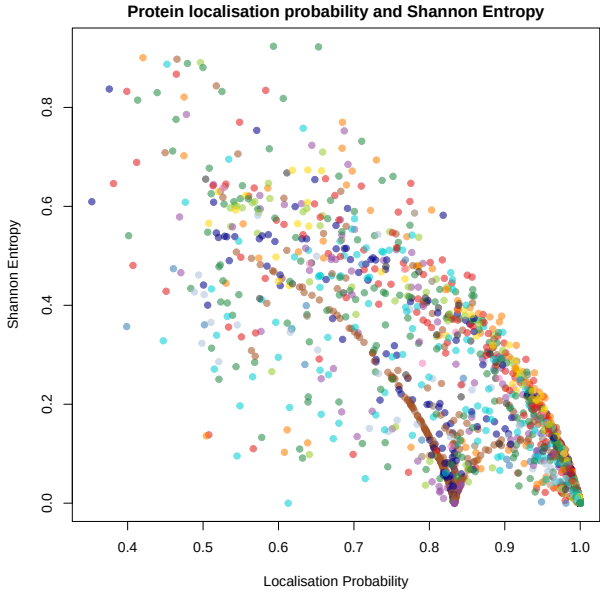
Figure 2.12 demonstrates that the regions of highest uncertainty are those in regions where organelles assignments overlap. The conclusions from this plot are manifold. Firstly, many proteins are assigned unambiguously to sub-cellular localisations; that is, not only are some proteins assigned to organelles with high probability but also with low uncertainty. Secondly, there are well defined regions with high uncertainty, for example proteins in the secretory pathway or proteins on the boundary between cytosol and proteasome. Finally, some organelles, such as the mitochondria, are extremely well resolved. This observed uncertainty in the secretory pathway and cytosol could be attributed to the dynamic nature of these parts of the cell with numerous examples of proteins that traffic in and out of these sub-cellular compartments as part of their biological role. Moreover, the organelles of the secretory pathway share similar and overlapping physical properties making their separation from one another using biochemical fractionation more challenging. Furthermore, there is a region located in the centre of the plot where proteins simultaneously have low probability of belonging to any organelle and high uncertainty in their localisation probability. This suggests that these proteins are poorly described by any single location. These proteins could belong to multiple locations or belong to undescribed sub-cellular compartments. The information displayed in these plots and the

conclusion therein would be extremely challenging to obtain without the use of Bayesian methodology.



(a)

(b)



(c)

Fig. 2.12 PCA plots of the mouse pluripotent embryonic stem cell data, where each point represents a protein and is coloured to its (probabilistically-)assigned organelle. (a) In this plot, the pointer is scaled to the Shannon entropy of this protein, with larger pointers indicating greater uncertainty. (b) In this plot, the pointer is scaled to the probability of that protein belonging to its assigned organelle. (c) We plot the localisation probabilities against the Shannon entropy with each protein.

2.6 Discussion and limitations

This chapter introduced a Bayesian framework, based on Gaussian mixture models, for spatial proteomics that can provide whole sub-cellular proteome uncertainty quantification on the assignment of proteins to organelles. We have demonstrated that such information is invaluable. Performing MAP inference using our generative model provides fast and straightforward approach, which is vital for quality control and early data exploration.

Full sampling-based posterior inference using MCMC provides not only point estimates of the posterior probability that a protein belongs to a particular sub-cellular niche, but uncertainty in this assignment. Then, this uncertainty can be summarised in several ways, including, but not limited to, equi-tailed credible intervals of the Monte-Carlo samples of posterior localisation probabilities. Posterior distributions for individual proteins can then be rigorously interrogated to shed light on their biological mechanisms; such as, transport, signalling and interactions.

As well as the local uncertainty seen by exploring individual proteins, we further explored using a Monte-Carlo averaged Shannon entropy to visualise global uncertainty. Regions of high uncertainty, as measured using this Shannon entropy, reflect highly dynamic regions of the sub-cellular environment. Hence, biologists can now explore uncertainty at different levels and then are able to make quantifiable conclusions and insights about their data. Furthermore, our Bayesian model is interpretable and our inferences are fully conditional on our data, allowing them to be easily modified with changing experimental design.

In addition, we produced competitive classifier performance to the state-of-the-art classifiers. We considered two traditional machine-learning methods: the SVM and KNN classifiers; as well as two classifiers based on our model: a MAP classifier and classification based on MCMC. We compared all methods on 19 different spatial proteomics datasets, across four different organisms. When considering the macro-F1 score as a performance metric, no single classifier outperformed another across all datasets. However, using MCMC based inference our method significantly outperforms the SVM and KNN classifiers with respect to the quadratic loss in 16 out of 19 datasets. This allows us to have greater confidence in our conclusions when they are draw from our Bayesian inferences. Furthermore, using MCMC provides a wealth of additional information, and so becomes the method of choice for analysing spatial proteomics data.

Analysis of a *hyper*LOPIT experiment applied to mouse pluripotent embryonic stem cells demonstrated that the additional layer of information that our model provides is biologically relevant and allows further avenues for additional exploration. Moreover, applying our method to a biologically significant dataset now provides the scientific community with localisation information on up to 4000 proteins for the mouse pluripotent stem cell proteome. Figure 2.13 demonstrates that from an initial input of roughly 1000 marker proteins with *a priori* known location and 4000 proteins with unknown location, SVM and TAGM-MCMC can provide

rigorous localisation information on roughly 2000 proteins. However, our methodology, by also considering uncertainty, allows us to obtain information on another 1000 proteins. Thus, we have augmented this dataset by providing uncertainty quantification on the localisation of proteins to their sub-cellular niches, which had been previously unavailable. We note that our method is general enough to be applied to many MS-based spatial proteomics protocols including: LOPIT, *hyper*LOPIT, protein correlation profiling (PCP) [135], differential centrifugation approaches and spatio-temporal proteomics methods. In our flexible software implementation, all hyperparameters for the priors can be changed if users have precise priors they wish to specify.

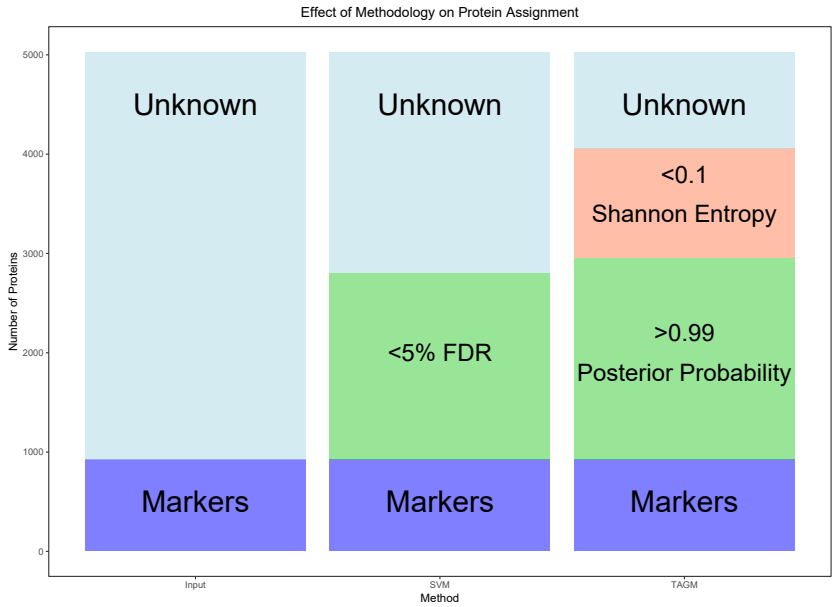


Fig. 2.13 The barplot demonstrates the effect of applying different methodologies on protein assignment when applied the mouse pluripotent embryonic stem cell data. Roughly 2000 proteins are classified using either SVM and TAGM-MCMC; however, TAGM-MCMC can draw additional conclusions about an extra 1000 proteins by quantifying uncertainty.

We have also provided a new set of visualisation methods to accompany our model, which allow us to easily interrogate our data. High quality visualisation tools are essential for rigorous quality control and sound biological conclusions. The methods have been developed in the R statistical programming language and we continue to contribute to the Bioconductor project [166, 212] with inclusion of our methods within the pRoloc package (>= 1.21.1) [156]. The underlying source code used to generate the results and figures of this chapter is available at https://github.com/lgatto/2018-TAGM-paper. The details of our software implementation is the content of the next chapter.

Currently, our model does not integrate localisation information from different data sources, nor does it explicitly model proteins with multiple localisation. However, one (of many)

biological explanations for the uncertainty that we model in the allocation probabilities is provided by multiple localisation. Thus a protein with uncertain allocation to two sub-cellular niches might, in reality, be resident in both. Further chapters explore different sources of uncertainty in more detail.

There are a number of limitations of our approach. Analysis of the outlier component suggests that there are perhaps un-annotated sub-cellular niches within the data. Analysis of spatial proteomics data is heavily reliant of these marker annotations and these are not readily variable for all sub-cellular niches. Furthermore, some organisms to which we wish to apply spatial proteomics have extremely poor annotations. Thus, to enable spatial proteomics as a powerful discovery tool, we need methods to simultaneously assign proteins to organelles and detect un-annotated sub-cellular niches. This is the content of a later chapter.

One of the motivations of our model was that it needed to be robust to outliers. However, robust mixture models in the frequentist literature rely on *pseudo*-models, which are not readily translated into the Bayesian paradigm. Heavy-tailed mixture models and Poisson process based outlier densities lack breakdown robustness properties. The presented model was motivated by the *pseudo*-model in a way that was amenable to the Bayesian framework. However, we did not characterise the theoretical breakdown robustness properties of our model. This would be a substantial endeavour and it is not explored in this thesis.

The t-augmented Gaussian mixture model hints at a possible general strategy for mixture models. Mixture models with mixed parametric components have been explored in the literature, but mostly from the perspective of Bayesian model selection and testing [237]. The properties of *mixed-mixtures* where the different parametric components arise from different families has not been readily explored [292]. Perhaps the most interesting cases are where some parametric components model deviant behaviour. For example, a heavy-tailed log-Gaussian component can be included in a mixture of Gammas or, as we presented, a t-distribution amongst a mixture of Gaussians. There is also scope for exploring discrete models, such as an over-dispersed negative binomial component amongst a Poisson mixture. Characterising the behaviour of these models with respect to robustness and model misspecification would be a further avenue to explore.

Chapter 3

A Bioconductor workflow for the Bayesian analysis of spatial proteomics

This chapter introduces software infrastructure for analysing spatial proteomics data, following on from the methodology described in the previous chapter. Current software is not able to provide bespoke Bayesian analysis for spatial proteomics data. Furthermore, it is unlikely that many practitioners of spatial proteomics are versed in Bayesian analysis so we provide some additional details for users. This is an edited version of Crook et al. [84] and there is significant textual overlap.

3.1 Motivation

3.1.1 Abstract

Knowledge of the subcellular location of a protein gives valuable insight into its function. The field of spatial proteomics has become increasingly popular due to improved multiplexing capabilities in high-throughput mass spectrometry, which have made it possible to systematically localise thousands of proteins per experiment. In parallel with these experimental advances, improved methods for analysing spatial proteomics data have also been developed. In this chapter, we demonstrate using "pRoloc" to perform Bayesian analysis of spatial proteomics data. We detail the software infrastructure and then provide step-by-step guidance of the analysis, including setting up a pipeline, assessing convergence, and interpreting downstream results. In several places we provide additional details on Bayesian analysis to provide users with a holistic view of Bayesian analysis for spatial proteomics data.

3.2 Introduction and literature review

Determining the spatial subcellular distribution of proteins enables novel insight into protein function [83]. Many proteins function within a single location within the cell; however, it is estimated that up to half of the proteome is thought to reside in multiple locations, with some of these undergoing dynamic relocalisation [455]. These phenomena lead to variability and uncertainty in robustly assigning proteins to a unique localisation. Functional compartmentalisation of proteins allows the cell to control biomolecular pathways and biochemical processes within the cell. Therefore, proteins with multiple localisations may have multiple functional roles [231]. Machine learning algorithms that fail to quantify uncertainty are unable to draw deeper insight into understanding cell biology from mass spectrometry (MS)-based spatial proteomics experiments. Hence, quantifying uncertainty allows us to make rigorous assessments of protein subcellular localisation and multi-localisation.

For proteins to carry out their functional role they must be localised to the correct subcellular compartment, ensuring the biochemical conditions for desired molecular interactions are met [171]. Many pathologies, including cancer and obesity are characterised by protein mis-localisations [350, 259, 283, 102, 69, 240, 393, 257, 419, 423]. High-throughput spatial proteomics technologies have seen rapid improvement over the last decade and now a single experiment can provide spatial information on thousands of proteins at once [119, 135, 68, 159]. As a result of these spatial proteomics technologies many biological systems have been characterised [119, 448, 43, 68, 455]. The popularity of such methods is now evident with many new studies in recent years [68, 24, 222, 221, 307, 202, 91, 351, 339].

Mass spectrometry-based spatial proteomic experiments begin with the gentle lysis of a population of cells in a fashion that maintains the integrity of the organelles. To separate cellular content a variety of methods are available, including equilibrium gradient-density separation [68, 324] or differential centrifugation [159]. For example, in hyperLOPIT [324] cell lysis is followed by the separation of subcellular components along a continuous density gradient based on their buoyant density. Discrete fractions along this gradient are then collected, and protein distributions revealing organelle specific correlation profiles within the fractions are achieved using high accuracy MS. Proteins from the dataset are then manually annotated with well-documented single localisations curated from the literature, referred to as organelle markers (see Gatto et al. [155]). A prediction model is then trained from these markers to create a classifier, which assigns proteins with unknown localisation to a sub-cellular niche [155].

Bayesian approaches to machine learning and statistics can provide more insight, by providing uncertainty quantification [162]. In a parametric Bayesian setting, a parametric model is proposed, along with a statement about our prior beliefs of the model parameters. Bayes' theorem tells us how to update the prior distribution of the parameters to obtain the

posterior distribution of the parameters after observing the data. It is the posterior distribution which quantifies the uncertainty in the parameters. This contrasts from a maximum-likelihood approach where we obtain only a point estimate of the parameters.

Adopting a Bayesian framework for data analysis, though of much interest to experimentalists, can be challenging. Once we have specified a probabilistic model, computational approaches are typically used to obtain the posterior distribution upon observation of the data. These algorithms can have parameters that require tuning and a variety of settings, hindering their practical use by those not familiar with Bayesian methodology. Even once the algorithms have been correctly set-up, assessments of convergence and guidance on how to interpret the results are often sparse. This chapter presents a Bayesian analysis of spatial proteomics to elucidate the process for practitioners. Our workflow also provides a template for others interested in designing tools for the biological community which rely on Bayesian inference.

Our model for the data is the t-augmented Gaussian mixture (TAGM) model proposed in the previous chapter of this thesis. Chapter 2 provided a detailed description of the model, rigorous comparisons and testing on many spatial proteomics datasets. In addition, we included a case study of a hyperLOPIT experiment performed on mouse pluripotent stem cells [68, 324]. Revisiting these details of that chapter is not the purpose of this chapter; rather we present how to correctly use the software and provide step-by-step guidance for interpreting the results.

As a brief reminder, the TAGM model posits that each annotated sub-cellular niche can be modelled using a Gaussian distribution. Thus the full complement of proteins within the cell is captured as a mixture of Gaussians. The highly dynamic nature of the cell means that many proteins are not well captured by any of these multivariate Gaussian distributions, and thus the model also includes an outlier component, which is mathematically described as a multivariate student's t distribution. The heavy tails of the t distribution allow it to better capture dispersed proteins. The outlier component is included to avoid allocating proteins which are far from any annotated subcellular niche. These proteins can be interpreted in multiple ways: they could be part of an unannotated subcellular niche, they could reside in multiple locations, they could have highly variable sub-cellular niches or they could have been poorly quantified.

There are two approaches to perform inference in the TAGM model. The first, which we refer to as TAGM MAP, allows us to obtain *maximum a posteriori* estimates of posterior localisation probabilities; that is, the modal posterior probability that a protein localises to that class. This approach uses the expectation-maximisation (EM) algorithm to perform inference [108]. Whilst this is an interpretable summary of the TAGM model, it only provides point estimates. For a richer analysis, we also present a Markov-chain Monte-Carlo (MCMC) method to perform fully Bayesian inference in our model, allowing us to obtain full posterior localisation distributions. This method is referred to as TAGM MCMC throughout the text.

This chapter begins with some details about the R programming language and the Bioconductor project. We highlight some other workflows from which we draw inspiration. From there we provide a brief review of some of the basic features of mass spectrometry-based spatial proteomics data, including our state-of-the-art computational infrastructure and bespoke software suite. We then present each method in turn, detailing how to obtain high quality results. We provide an extended discussion of the TAGM MCMC method to highlight some of the challenges that may arise when applying this method. This includes how to assess convergence of MCMC methods, as well as methods for manipulating the output. We then take the processed output and explain how to interpret the results, as well as providing some tools for visualisation. We conclude with some remarks and directions for the future. Source code for this chapter, including code used to generate tables and figures, is available on [GitHub](https://github.com/ococrook/TAGMworkflow)

S4 Object Orientation in R

Object-oriented programming (OOP) is a programmatic paradigm with the idea of an "object" taking the central role [317]. In class-based OOP languages objects are instances of classes [317]. These classes determine the type of the object. Methods are recipes or procedures than can be applied to particular classes. For example a "show" method, usually intended to produce some short output, produces a different outcome depending on whether it applied to one object or another. Object orientation is somewhat complex in R because it supports multiple OOP frameworks. We refer to Wickham [478] for further details. The S3, S4 and R6 are the most commonly used frameworks. R6, which we mention only briefly for completeness, works based on the encapsulation paradigm and as such methods belong to objects not generics - that is the data and methods are bundled together. Objects in R6 are mutable meaning that the object's state can be modified after it is created. On the other hand S3, the simplest OOP system in R is informal and ad hoc. However, its flexibility makes it quite popular. In the S3 paradigm there is no formal definition of a class, whilst the S4 paradigm is far more formal. Classes, generics and methods make use of precise defining functions and S4 also boasts the slot - accessed via the subsetting operator @. Given we are building on a large body of code, formality and robustness are desirable over flexibility - thus we opt for the S4 paradigm.

The Bioconductor Project and Workflows

The Bioconductor project [166, 212] is an open source and open development software project, originally focused on the analysis of genomic data. The project is design to facilitate reproducible and powerful statistical analysis of biological data. Bioconductor packages are required to have a vignette (task-oriented documentation), unit testing and be written in S4. Packages are required to not rebuild infrastructure or data types that can be meaningfully reused. A

number of fields have substantial contributions from the Bioconductor project, include spatial proteomics [45], microbiome data [55], single-cell RNA sequencing [285, 362], methylation arrays [291], ChIP-seq data [284] and CyTOF data [344].

Workflows in Bayesian Analysis

Workflows in Bayesian analysis are sparse compared to those for biological tailored applications. There are books with detailed code for example see McElreath [299]; however, these often require substantial statistical knowledge and the application is not substantial. There are papers on general good practise for Bayesian workflows [405, 147], though these usually focus on the principal application of Bayesian analysis rather than the application itself. There is a clear need for more workflows detailing analysis on specific applications.

3.3 Getting started and infrastructure

In this workflow chapter, we are using version 1.23.2 of `pRoloc` (Gatto, Breckels, et al. 2014b). The package `pRoloc` contains algorithms and methods for analysing spatial proteomics data, building on the `MSnSet` structure provided in `MSnbase`. The `pRolocdata` package provides many annotated datasets from a variety of species and experimental procedures. The following code chunks install and load the suite of packages require for the analysis.

```
if (!require("BiocManager"))
install.package("BiocManager")
BiocManager::install(c("pRoloc", "pRolocdata"))
```

```
library("pRoloc")
library("pRolocdata")
```

```
##
## This is pRolocdata version 1.22.0.
## Use 'pRolocdata()' to list available data sets.
```

We assume that we have a MS-based spatial proteomics dataset contained in a `MSnSet` structure. For information on how to import data, perform basic data processing, quality control, supervised machine learning and transfer learning we refer the reader to [45]. Here, we start by loading a spatial proteomics dataset on mouse E14TG2a embryonic stem cells [44]. The LOPIT protocol [118, 119] was used and the normalised intensity of proteins from eight iTRAQ 8-plex labelled fraction are provided. The methods provided here are independent of labelling procedure, fractionation process or workflow. Examples of valid experimental

protocols are LOPIT [118], hyperLOPIT [68, 324], label-free methods such as PCP [135], and when fractionation is perform by differential centrifugation [220, 159].

In the code chunk below, we load the aforementioned dataset. The printout demonstrates that this experiment quantified 2031 proteins over 8 fractions.

```
data("E14TG2aR") # load experimental data
E14TG2aR
```

```
## MSnSet (storageMode: lockedEnvironment)
## assayData: 2031 features, 8 samples
##   element names: exprs
## protocolData: none
## phenoData
##   sampleNames: n113 n114 ... n121 (8 total)
##   varLabels: Fraction.information
##   varMetadata: labelDescription
## featureData
##   featureNames: Q62261 Q9JHU4 ... Q9EQ93 (2031 total)
##   fvarLabels: Uniprot.ID UniprotName ... markers (8 total)
##   fvarMetadata: labelDescription
## experimentData: use 'experimentData(object)'
## Annotation:
## - - - Processing information - - -
## Loaded on Thu Jul 16 15:02:29 2015.
## Normalised to sum of intensities.
## Added markers from  'mrk' marker vector. Thu Jul 16 15:02:29 2015
##  MSnbase version: 1.17.12
```

In figure 3.1, we can visualise the mouse stem cell dataset use the `plot2D` function. We observe that some of the organelle classes overlap and this is a typical feature of biological datasets. Thus, it is vital to perform uncertainty quantification when analysing biological data.

```
plot2D(E14TG2aR)
addLegend(E14TG2aR, where = "topleft", cex = 0.6)
```

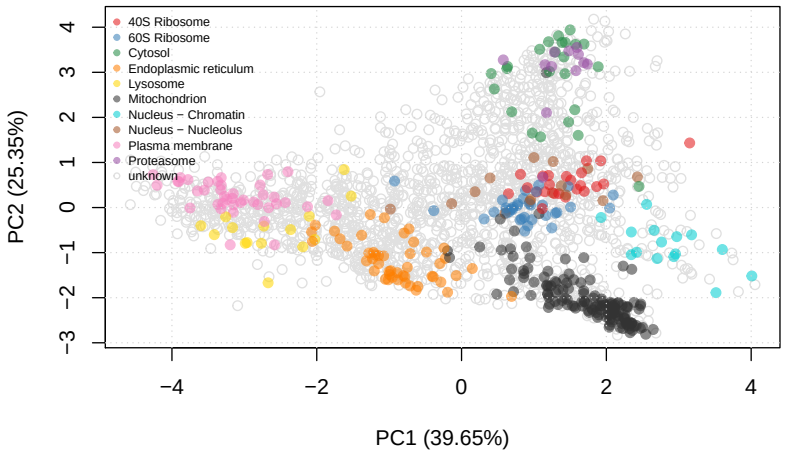


Fig. 3.1 First two principal components of mouse stem cell data.

3.4 Methods: *TAGM MAP*

3.4.1 Introduction to TAGM MAP

We can use *maximum a posteriori* (MAP) estimation to perform Bayesian parameter estimation for our model. The *maximum a posteriori* estimate is the mode of the posterior distribution and can be used to provide a point estimate summary of the posterior localisation probabilities. In contrast to TAGM MCMC (see later), it does not provide samples from the posterior distribution, however it allows for faster inference by using an extended version of the expectation-maximisation (EM) algorithm. The EM algorithm iterates between an expectation step and a maximisation step. This allows us to find parameters which maximise the logarithm of the posterior, in the presence of latent (unobserved) variables. The EM algorithm is guaranteed to converge to a local mode. The code chunk below executes the `tagmMapTrain` function for a default of 100 iterations. We use the default priors for simplicity and convenience, however they can be changed, which we explain in a later section. The output is an object of class `MAPParams`, that captures the details of the TAGM MAP model.

```
set.seed(2)
mappars <- tagmMapTrain(E14TG2aR)
```

```
## co-linearity detected; a small multiple of
##               the identity was added to the covariance
```

```
mappars
```

```
## Object of class "MAPParams"
##  Method: MAP
```

Aside: collinearity

The previous code chunk outputs a message concerning data collinearity. This is because the covariance matrix of the data has become ill-conditioned and as a result the inversion of this matrix becomes unstable with floating point arithmetic. This can lead to the failure of standard matrix algorithms upon which our method depends. In this case, it is standard practice to add a small multiple of the identity to stabilise this matrix. The printed message is a statement that this operation has been performed for these data.

3.4.2 Model visualisation

The results of the modelling can be visualised with the `plotEllipse` function on figure 3.2. The outer ellipse contains 99% of the total probability whilst the middle and inner ellipses contain 95% and 90% of the probability respectively. The centres of the clusters are represented by black circumpunct (circled dot). We can also plot the model in other principal components. The code chunk below plots the probability ellipses along the first and second, as well as the fourth principal component. The user can change the components visualised by altering the `dims` argument.

```
par(mfrow = c(1, 2))
plotEllipse(E14TG2aR, mappars)
plotEllipse(E14TG2aR, mappars, dims = c(1, 4))
```

3.4.3 The expectation-maximisation algorithm

The EM algorithm is iterative; that is, the algorithm iterates between an expectation step and a maximisation step until the value of the log-posterior does not change [108]. This fact can be used to assess the convergence of the EM algorithm. The value of the log-posterior at each iteration can be accessed with the `logPosteriors` function on the `MAPParams` object. The code chuck below plots the log posterior at each iteration and we see on figure 3.3 the algorithm rapidly plateaus and so we have achieved convergence. If convergence has not been reached during this time, we suggest increasing the number of iterations by changing the parameter `numIter` in the `tagmMapTrain` method. In practice, it is not unexpected to observe small fluctuations due to numerical errors and this should not concern users.

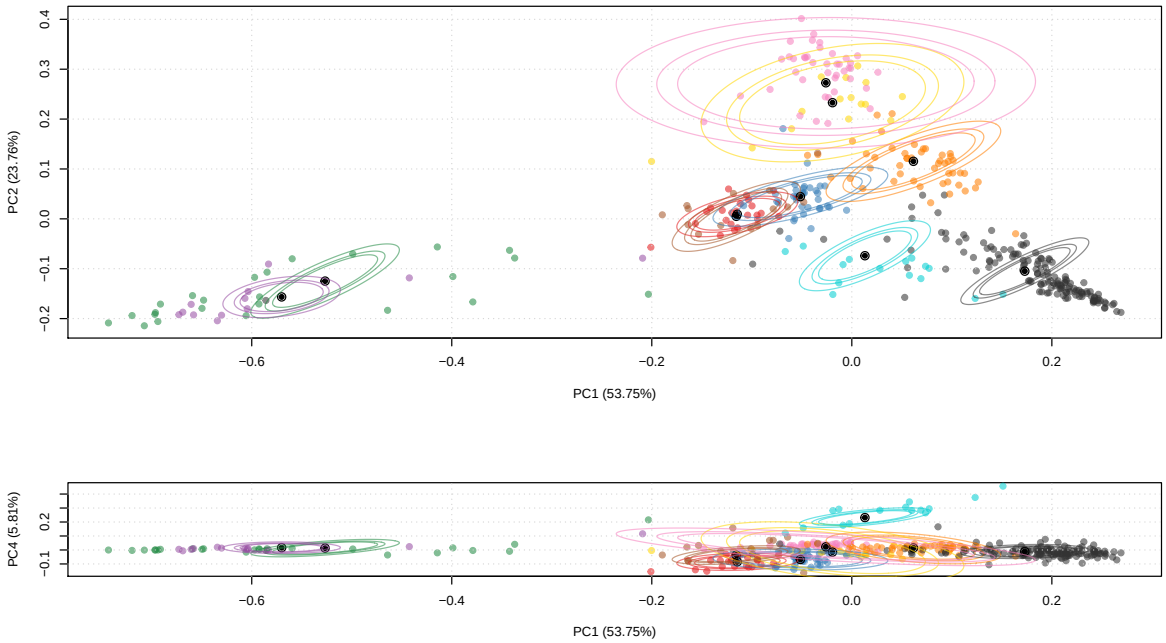


Fig. 3.2 PCA plot with probability ellipses along PC 1 and 2 (left) and PC 1 and 4 (right). The ellipses show the component-conditional densities obtained from the fitted model evaluated at θ_{MAP}

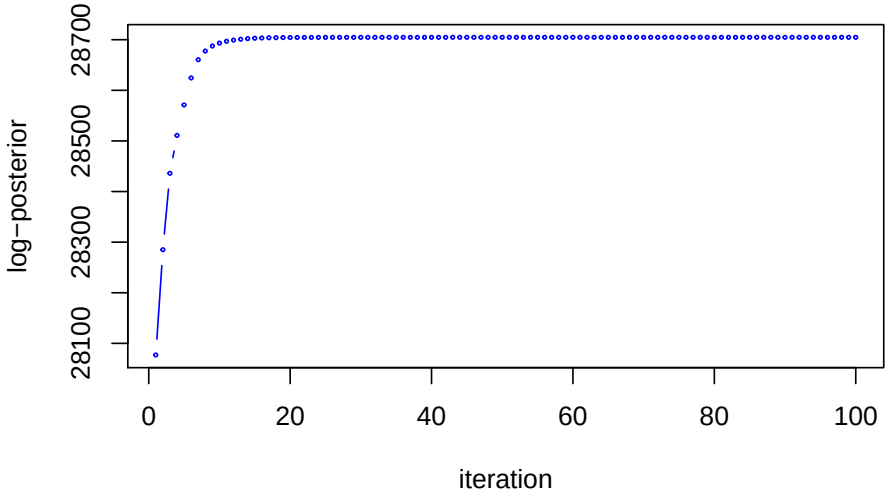


Fig. 3.3 Log-posterior at each iteration of the EM algorithm demonstrating convergence.

```
plot(logPosteriors(mappars), type = "b", col = "blue",
cex = 0.3, ylab = "log-posterior", xlab = "iteration")
```

The code chuck below uses the `mappars` object generated above, along with the `E14RG2aR` dataset, to classify the proteins of unknown localisation using `tagmPredict` function. The results of running `tagmPredict` are appended to the `fData` columns of the `MSnSet`.

```
E14TG2aR <- tagmPredict(E14TG2aR, mappars) # Predict protein localisation
```

The new feature variables that are generated are:

- `tagm.map.allocation`: the TAGM MAP predictions for the most probable protein sub-cellular allocation.

```
table(fData(E14TG2aR)$tagm.map.allocation)
```

```
##
##          40S Ribosome          60S Ribosome               Cytosol
##                    34                    85                   328
## Endoplasmic reticulum             Lysosome         Mitochondrion
##                   284                   147                   341
##   Nucleus - Chromatin   Nucleus - Nucleolus       Plasma membrane
```

```
##                 143                 322                 326
##          Proteasome
##                  21
```

- `tagm.map.probability`: the posterior probability for the protein sub-cellular allocations.

```
summary(fData(E14TG2aR)$tagm.map.probability)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00000 0.06963 0.93943 0.63829 0.99934 1.00000
```

- `tagm.map.outlier`: the posterior probability for that protein to belong to the outlier component rather than any annotated component.

```
summary(fData(E14TG2aR)$tagm.map.outlier)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 0.0000000 0.0002363 0.0305487 0.3452624 0.9249810 1.0000000
```

We can visualise the results by scaling the pointer according to the posterior localisation probabilities. To do this we extract the MAP localisation probabilities from the feature columns of the the `MSnSet` and pass these to the `plot2D` function (figure 3.4).

```
ptsze <- fData(E14TG2aR)$tagm.map.probability # Scale pointer size
plot2D(E14TG2aR, fcol = "tagm.map.allocation", cex = ptsze)
addLegend(E14TG2aR, where = "topleft", cex = 0.6, fcol = "tagm.map.allocation")
```

The TAGM MAP method is easy to use and it is simple to check convergence, however it is limited in that it can only provide point estimates of the posterior localisation distributions. To obtain the full posterior distributions and therefore a rich analysis of the data, we use Markov-Chain Monte-Carlo methods. In our particular case, we use a *collapsed Gibbs sampler* [427].

3.5 Methods: *TAGM MCMC* a brief overview

The TAGM MCMC method allows a fully Bayesian analysis of spatial proteomics datasets. It employs a collapsed Gibbs sampler to obtain samples from the posterior distribution of localisation probabilities, providing a rich analysis of the data. This section demonstrates the advantage of taking a Bayesian approach and the biological information that can be extracted from this analysis.

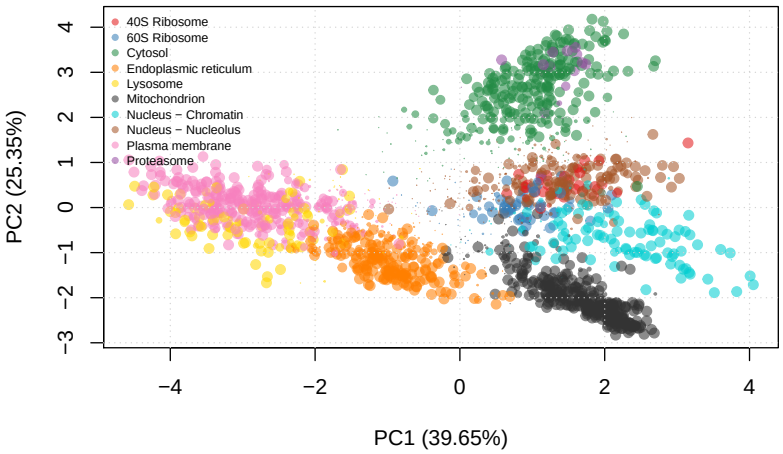


Fig. 3.4 TAGM MAP allocations, where the pointer is scaled according to the localisation probability and coloured according to the most probable subcellular niche.

For those unfamiliar with Bayesian methodology, some of the key ideas for a more complete understanding are as follows. Firstly, MCMC based inference contrasts with MAP based inference in that it *samples* from the posterior distribution of localisation probabilities. Hence, we do not just have a single estimate for each quantity but a distribution of estimates. MCMC methods are a large class of algorithms used to sample from a probability distribution, in our case the posterior distribution of the parameters (Gilks, Richardson, and Spiegelhalter 1995). Once we have sampled from the posterior distribution, we can estimate the mean of the posterior distribution by simply taking the mean of the samples. In a similar fashion, we can obtain estimates of other summaries of the posterior distribution.

A schematic of MCMC sampling is provided in figure 3.5 to aid understanding. Proteins, coloured blue, are visualised along two variables of the data. Probability ellipses representing contours of a probability distribution matching the distribution of the proteins are overlaid. We now wish to obtain samples from this distribution. The MCMC algorithm is initialised with a starting location, then at each iteration a new value is proposed. These proposed values are either accepted or rejected (according to a carefully computed acceptance probability) and over many iterations the algorithm converges and produces samples from the desired distribution. Samples from this distribution are coloured in red in the schematic figure. A large portion of the earlier samples may not reflect the true distribution, because the MCMC sampler has yet to converge. These early samples are usually discarded and this is referred to as burn-in [173]. The next state of the algorithm depends on its current state and this leads to auto-correlation

in the samples. To suppress this auto-correlation, we only retain every r^{th} sample. This is known as thinning. The details of burn-in and thinning are further explained in later sections.

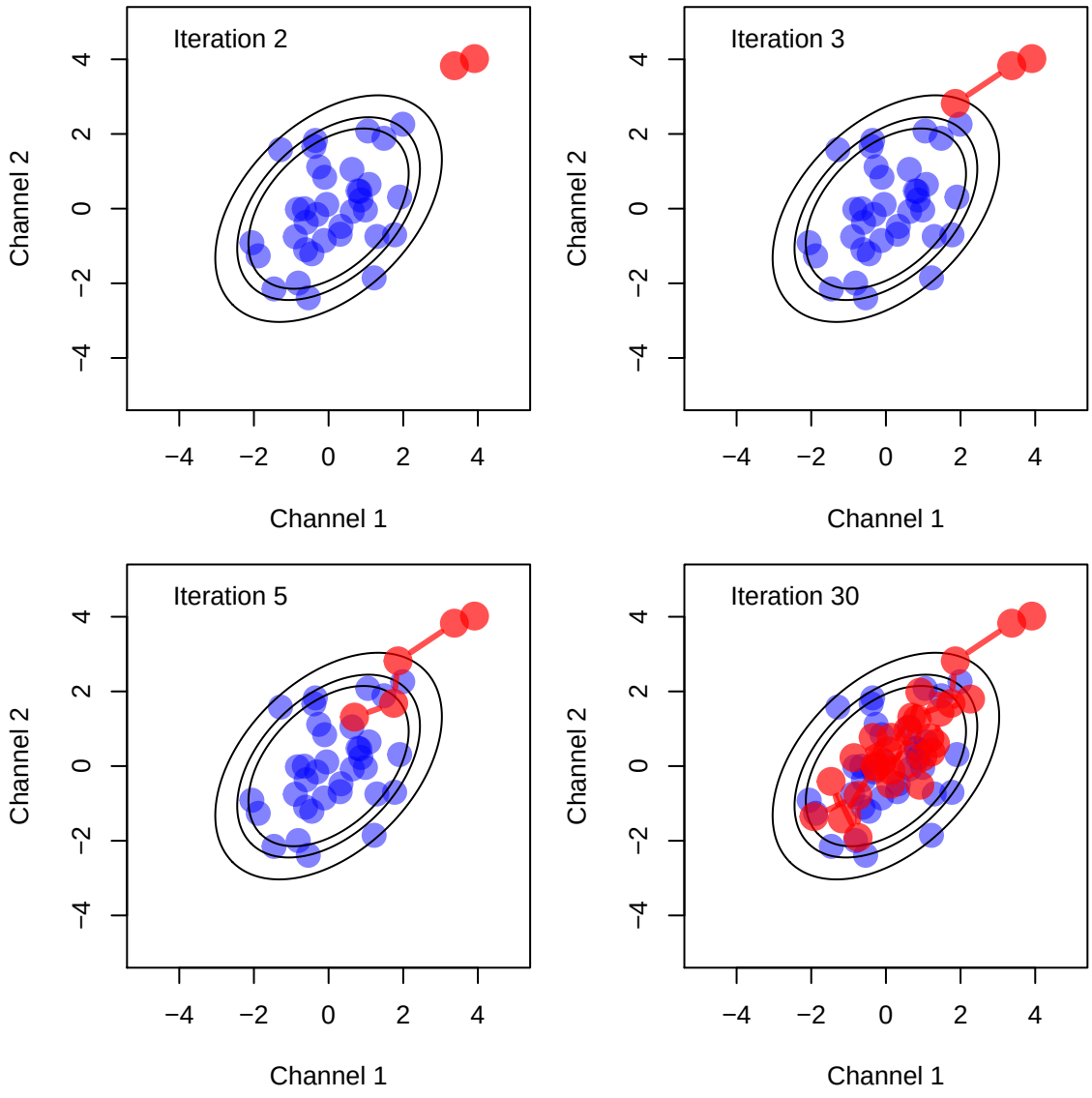


Fig. 3.5 A schematic figure of MCMC sampling. Proteins are coloured in blue and probability ellipses are overlaid representing contours of a probability distribution matching the distribution of the proteins. MCMC samples from this distribution are then coloured in red.

The TAGM MCMC method is computationally intensive and requires at least modest processing power. Leaving the MCMC algorithm to run overnight on a modern desktop is usually sufficient, however this, of course, depends on the particular dataset being analysed. For guidance: it should not be expected that the analysis will finish in just a couple of hours on a medium specification laptop, for example.

To demonstrate the class structure and expected outputs of the TAGM MCMC method, we run a brief analysis on a subset (400 randomly chosen proteins) of the `tan2009r1` dataset from the `pRolocdata`, purely for illustration. This is to provide a bare bones analysis of these data without being held back by computational requirements. We perform a complete demonstration and provide precise details of the analysis of the stem cell dataset considered above in the next section.

```
set.seed(1)
data(tan2009r1)
tan2009r1 <- tan2009r1[sample(nrow(tan2009r1), 400), ]
```

The first step is to run a few MCMC chains (below we use only 2 chains) for a few iterations (we specify 3 iterations in the below code, but typically we would suggest in the order of tens of thousands; see for example the algorithms default settings by typing `?tagmMcmcTrain`) using the `tagmMcmcTrain` function. This function will generate a object of class `MCMCParams`.

```
p <- tagmMcmcTrain(object = tan2009r1, numIter = 3,
burnin = 1, thin = 1, numChains = 2)
```

```
p
```

```
## Object of class "MCMCParams"
## Method: TAGM.MCMC
## Number of chains: 2
```

Information for each MCMC chain is contained within the chains slot. If needed, this information can be accessed manually. The function `tagmMcmcProcess` processes the `MCMCParams` object and populates the summary slot.

```
p <- tagmMcmcProcess(p)
p
```

```
## Object of class "MCMCParams"
## Method: TAGM.MCMC
## Number of chains: 2
## Summary available
```

The summary slot has now been populated to include basic summaries of the MCMC chains, such as organelle allocations and localisation probabilities. Protein information can

be appended to the feature columns of the `MSnSet` by using the `tagmPredict` function, which extracts the required information from the summary slot of the `MCMCParams` object.

```
res <- tagmPredict(object = tan2009r1, params = p)
```

We can now access new variables:

- `tagm.mcmc.allocation`: the TAGM MCMC prediction for the most likely protein sub-cellular annotation.

```
table(fData(res)$tagm.mcmc.allocation)
```

```
##
##  Cytoskeleton            ER         Golgi      Lysosome mitochondrion
##            12            98            23             9            39
## Nucleus    Peroxisome           PM    Proteasome  Ribosome 40S  Ribosome 60S
##      26             3           102            29            30            29
```

- `tagm.mcmc.probability`: the mean posterior probability for the protein sub-cellular allocations.

```
summary(fData(res)$tagm.mcmc.probability)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2972  0.8995  0.9901  0.9080  1.0000  1.0000
```

We can also access other useful summaries of the MCMC methods:

- `tagm.mcmc.outlier` the posterior probability for the protein to belong to the outlier component.
- `tagm.mcmc.probability.lowerquantile` and `tagm.mcmc.probability.upperquantile` are the lower and upper boundaries to the equi-tailed 95% credible interval of `tagm.mcmc.probability`.
- `tagm.mcmc.mean.shannon` a Monte-Carlo averaged Shannon entropy, which is a measure of uncertainty in the allocations.

3.6 Methods: *TAGM MCMC* the details

This section explains how to manually manipulate the MCMC output of the TAGM model. In the code chunk below, we load a pre-computed TAGM MCMC model. The data file `e14tagm.rda` is available online[1] and is not directly loaded into "pRoloc" due to its size. The file itself if around 500mb, which is too large to load directly.

```
load("e14Tagm.rda")
```

The following code, which is not evaluated dynamically, was used to produce the `tagmE14` `MCMCParams` object. We run the MCMC algorithm for 20,000 iterations with 10,000 iterations discarded for burn-in. We then thin the chain by 20. We ran 6 chains in parallel and so we obtain 500 samples for each of the 6 chains, totalling 3,000 samples. The resulting file is assumed to be in our working directory.

```
e14Tagm <- tagmMcmcTrain(E14TG2aR,
numIter = 20000,
burnin = 10000,
thin = 20,
numChains = 6)
```

Manually inspecting the object, we see that it is a `MCMCParams` object with 6 chains.

```
e14Tagm
```

```
## Object of class "MCMCParams"
## Method: TAGM.MCMC
## Number of chains: 6
```

3.6.1 Data exploration and convergence diagnostics

Assessing whether or not an MCMC algorithm has converged is challenging. Assessing and diagnosing convergence is an active area of research and throughout the 1990s many approaches were proposed [169, 163, 387, 46] and these discussions have been refined in recent years (see Vats and Knudson [466], Vehtari et al. [467]. We provide a more detailed exploration of this issue, but readers should bear in mind that the methods provided below are diagnostics and cannot guarantee convergence. We direct readers to several important works in the literature discussing the assessment of convergence. Users that do not assess convergence and base their downstream analysis on unconverged chains are likely to obtain poor quality results.

[1] https://drive.google.com/open?id=1zozntDhE6YZ-q8wjtQ-lxZ66EEszOGYi

We first assess convergence using a parallel chains approach. We find producing multiple chains is beneficial not only for computational advantages but also for analysis of convergence of our chains. As with other authors, we suggest a minimum of 4 chains [467]. This is the default setting in the software. However, in this workflow we run 6 chains to highlight some challenges.

```
## Get number of chains
nChains <- length(e14Tagm)
nChains
```

```
## [1] 6
```

The following code chunks set up a manual convergence diagnostic check. We make use of objects and methods in the package *coda* to perform this analysis [367]. Our function below automatically coerces our objects into *coda* for ease of analysis. We first calculate the total number of outliers at each iteration of each chain and, if the algorithm has converged, this number should be the same (or very similar) across all 6 chains.

```
## Convergence diagnostic to see if we need to discard any
## iterations or entire chains: compute the number of outliers for
## each iteration for each chain
out <- mcmc_get_outliers(e14Tagm)
```

We can observe this from the trace plots and histograms for each MCMC chain (figure 3.6. Unconverged chains should be discarded from downstream analysis.

```
## Using coda S3 objects to produce trace plots and histograms
for (i in seq_len(nChains))
plot(out[[i]], main = paste("Chain", i), auto.layout = FALSE, col = i)
```

Chains 3, 5 and 6 are centred around an average of 153, with rapid back and forth oscillations. Chain 2 should be immediately discarded, since it has a large jump in the chain with clearly skewed histogram. The other two chains oscillate differently with contrasting quantiles to the 3 chains (3, 5 and 6) that agree with one another, suggesting these chains have yet to converge. We can use the *coda* package to produce summaries of our chains. Here is the `coda` summary for the third chain.

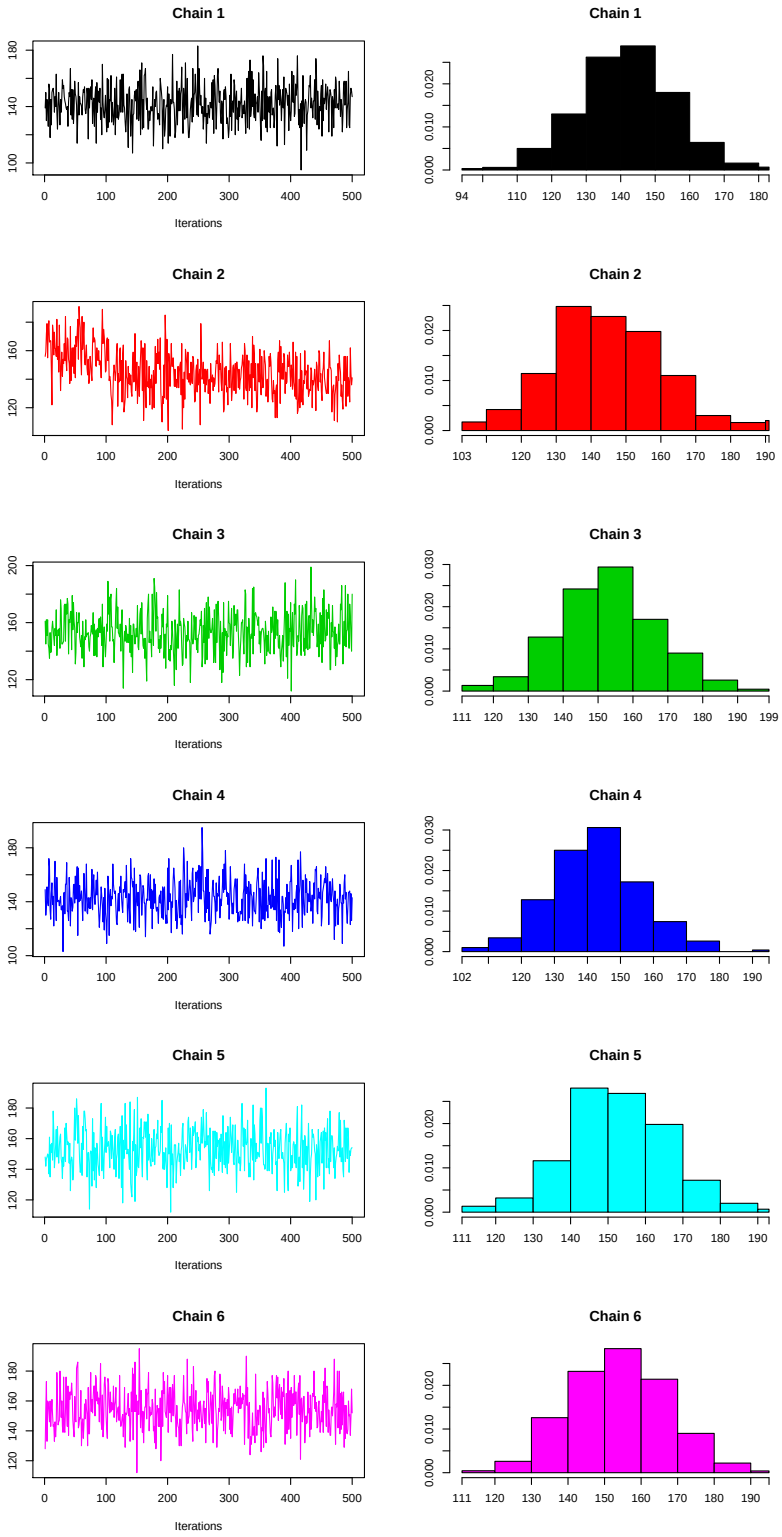


Fig. 3.6 Trace (left) and density (right) of the 6 MCMC chains. 500 iterations were subsampled from the MCMC chains of 20,000 iterations

```
## Chains average around 153 outliers
summary(out[[3]])
```

```
##
## Iterations = 1:500
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 500
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##           Mean             SD       Naive SE Time-series SE
##       153.4520        14.0771         0.6295         0.6820
##
## 2. Quantiles for each variable:
##
##  2.5%   25%   50%   75% 97.5%
##   127   144   153   162   183
```

Applying the Gelman diagnostic

So far, our analysis appears promising. Three of our chains are centred around an average of 153 outliers and there is no observed monotonicity in our output. However, for a more rigorous and unbiased analysis of convergence we can calculate the Gelman diagnostic using the *coda* package [163, 46]. This statistic is often referred to as $\hat{R}$ or the potential scale reduction factor. The idea of the Gelman diagnostics is to compare the inter and intra chain variances. The ratio of these quantities should be close to one. A more detailed and in depth discussion can be found in the references. The *coda* package also reports the 95% upper confidence interval of the $\hat{R}$ statistic. In this case, our samples are approximately normally distributed (see histograms on the right in figure 3.6. The *coda* package allows for transformations to improve normality of the data, and in some cases we set the `transform` argument to apply log transformation. Gelman and Rubin [163] suggest that chains with $\hat{R}$ value of less than 1.2 are likely to have converged, though recent literature suggests considerably smaller values and a thredhold of 1.01 is likely to lead to more stable and reliable results [466, 467].

```
gelman.diag(out, transform = FALSE)
```

```
## Potential scale reduction factors:
##
##      Point est. Upper C.I.
## [1,]       1.14       1.32
```

```
gelman.diag(out[c(1, 3, 4, 5, 6)], transform = FALSE)
```

```
## Potential scale reduction factors:
##
##      Point est. Upper C.I.
## [1,]       1.13       1.31
```

```
gelman.diag(out[c(3, 5)], transform = TRUE) # the upper C.I is 1.01
```

```
## Potential scale reduction factors:
##
##      Point est. Upper C.I.
## [1,]          1       1.01
```

In all cases, we see that the Gelman diagnostic for convergence is < 1.2, but only in the final case is it < 1.01. However, the upper confidence interval is 1.32 when all chains are used; 1.31 when chain 2 is removed and when chains 1, 2 and 4 are removed the upper confidence interval is 1.01 indicating that the MCMC algorithm for chains 3, 5 and 6 might have converged.

We can also look at the Gelman diagnostics statistics for groups or pairs of chains. The first line below computes the Gelman diagnostic across the first three chains, whereas the second calculates the diagnostic between chain 3 and chain 5.

```
gelman.diag(out[1:3], transform = FALSE) # the upper C.I is 1.62
```

```
## Potential scale reduction factors:
##
##      Point est. Upper C.I.
## [1,]       1.22       1.62
```

To assess another summary statistic, we can look at the mean component allocation at each iteration of the MCMC algorithm and as before we produce trace plots of this quantity (figure 3.7).

```
meanAlloc <- mcmc_get_meanComponent(e14Tagm)
```

```
for (i in seq_len(nChains))
plot(meanAlloc[[i]], main = paste("Chain", i), auto.layout = FALSE, col = i)
```

As before we can produce summaries of the data.

```
summary(meanAlloc[[1]])
```

```
##
## Iterations = 1:500
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 500
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##           Mean             SD       Naive SE Time-series SE
##       5.686713       0.059112       0.002644       0.002644
##
## 2. Quantiles for each variable:
##
##  2.5%   25%   50%   75% 97.5%
## 5.552 5.646 5.692 5.728 5.795
```

We already observed that there are some slight differences between these chains, which raises suspicion that some of the chains may not have converged. For example each chain appears to be centred around 5.7, but chains 2 and 4 have clear jumps in their trace plots. To be more precise, we note the jump that occurs are between iteration 100-150 in chain 2 and between iteration 200-250 in chain 4. For a more quantitative analysis, we again apply the Gelman diagnostics to these summaries.

```
gelman.diag(meanAlloc)
```

```
## Potential scale reduction factors:
##
##      Point est. Upper C.I.
## [1,]          1       1.01
```

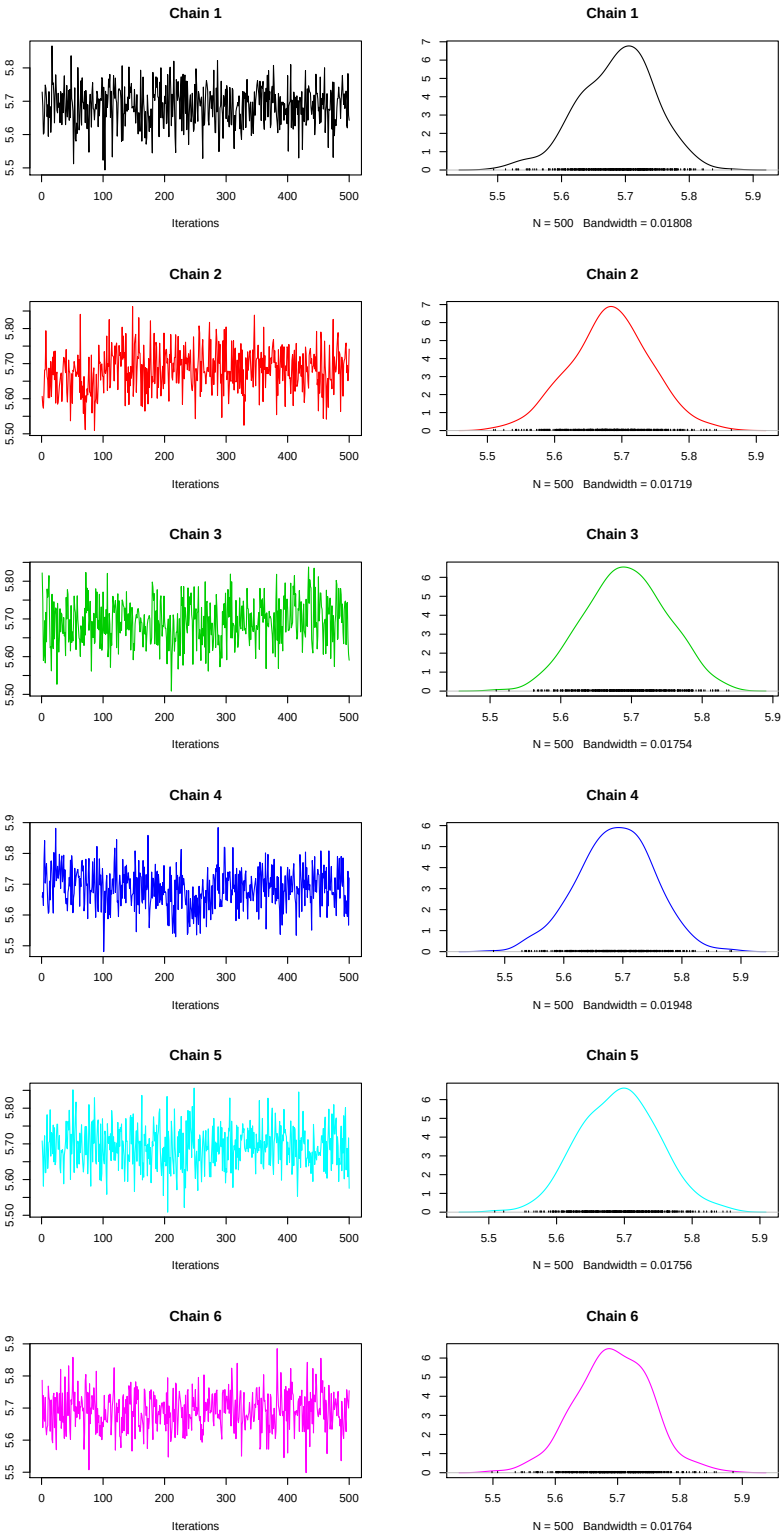


Fig. 3.7 Trace (left) and density (right) of the mean component allocation of the 6 MCMC chains. 500 iterations were subsampled from the MCMC chains of 20,000 iterations.

The above values are close to 1 and so there are no significant differences between the chains. As observed previously, chains 2 and 4 look quite different from the other chains and so we recalculate the diagnostic excluding these chains. The computed Gelman diagnostic below suggest that chains 3, 5 and 6 have converged and that we should discard chains 1, 2 and 4 from further analysis.

```
gelman.diag(meanAlloc[c(3, 5, 6)])
```

```
## Potential scale reduction factors:
##
##      Point est. Upper C.I.
## [1,]          1          1
```

For a further check, we can look at the mean outlier probability at each iteration of the MCMC algorithm and again computing the Gelman diagnostics between chains 3, 5 and 6. An $\hat{R}$ statistic of 1 is indicative of convergence, since it is less than the recommended value of 1.01.

```
meanoutProb <- mcmc_get_meanoutliersProb(e14Tagm)
gelman.diag(meanoutProb[c(3, 5, 6)])
```

```
## Potential scale reduction factors:
##
##      Point est. Upper C.I.
## [1,]          1       1.01
```

Applying the Geweke diagnostic

Along with the Gelman diagnostic, which uses parallel chains, we can also apply a single chain analysis using the Geweke diagnostic [169]. The Geweke diagnostic tests to see whether the mean calculated from the first 10% of iterations is significantly different from the mean calculated from the last 50% of iterations. If they are significantly different, at say a level 0.01, then this is evidence that particular chains have not converged. The following code chunk calculates the Geweke diagnostic for each chain on the summarising quantities we have previously computed.

```
geweke_test(out)
```

```
##           chain 1      chain 2  chain 3    chain 4    chain 5    chain 6
## z.value 0.5749775 8.816632e+00 0.470203 -0.3204500 -0.6270787 -0.7328168
## p.value 0.5653065 1.179541e-18 0.638210  0.7486272  0.5306076  0.4636702
```

```
geweke_test(meanAlloc)
```

```
##           chain 1       chain 2    chain 3    chain 4   chain 5    chain 6
## z.value 1.1952967 -3.3737051063 -1.2232102 2.48951993 0.3605882 -0.1358850
## p.value 0.2319711  0.0007416377  0.2212503 0.01279157 0.7184073  0.8919122
```

```
geweke_test(meanoutProb)
```

```
##           chain 1      chain 2   chain 3    chain 4    chain 5     chain 6
## z.value 0.1785882 1.205500e+01 0.6189637 -0.5164987 -0.2141086 -0.02379004
## p.value 0.8582611 1.825379e-33 0.5359403  0.6055062  0.8304624  0.98102008
```

The first test suggests chain 2 has not converged, since the p-value is less than 10^{-10} suggesting that the mean in the first 10% of iterations is significantly different from those in the final 50%. Moreover, the second test and third tests also suggest that chain 2 has not converged. Furthermore, for the second test chain 4 has a marginally small p-value, providing further evidence that this chain is of low quality. These convergence diagnostics are not limited to the quantities we have computed here and further diagnostics can be performed on any summary of the data.

An important question to consider is whether removing an early portion of the chain might lead to an improvement of the convergence diagnostics. This might be particularly relevant if a chain converges some iterations after our originally specified `burn-in`. For example, let us take the second Geweke test above, which suggested chains 2 and 4 had not converged and see if discarding the initial 10% of the chain improves the statistic. The function below removes 50 samples, known as `burn-in`, from the beginning of each chain and the output shows that we now have 450 samples in each chain. In practice, as 2 chains are sufficient for good posterior estimates and convergence we could simply discard chains 2 and 4 and proceed with downstream analysis with the remaining chains.

```
burn_e14Tagm <- mcmc_burn_chains(e14Tagm, 50)
chains(burn_e14Tagm)
```

```
## Object of class "MCMCChains"
##  Number of chains: 6
```

```
chains(burn_e14Tagm)[[4]]
```

```
## Object of class "MCMCChain"
##  Number of components: 10
##  Number of proteins: 1663
##  Number of iterations: 450
```

The following function recomputes the number of outliers in each chain at each iteration of each Markov-chain.

```
out2 <- mcmc_get_outliers(burn_e14Tagm)
```

The code chunk below computes the Geweke diagnostic for this new truncated chain and demonstrates that chain 4 has an improved Geweke diagnostic, whilst chain 2 does not. Thus, in practice, it may be useful to remove iterations from the beginning of the chain. However, as chain 4 did not pass the Gelman diagnostics we still discard it from downstream analysis.

```
geweke_test(out2)
```

```
##            chain 1      chain 2    chain 3   chain 4   chain 5   chain 6
## z.value -0.1455345 6.379618e+00 -1.6392215 0.3836940 0.1241201 0.6654703
## p.value  0.8842889 1.775298e-10  0.1011671 0.7012053 0.9012202 0.5057497
```

In this section, we have highlighted that assessing convergence is an essential part of Bayesian analysis. As well as the summaries considered here, we recommend that users assess other posterior summaries of the data. Since the best practices for assessing convergence also change overtime, we also suggest searching the literature for current consensus.

3.6.2 Processing converged chains

Having made an assessment of convergence, we decide to discard chains $1, 2$ and 4 from any further analysis. The code chunk below removes these chains and creates a new object to store the converged chains.

```
removeChain <- c(1, 2, 4) # The chains to be removed
e14Tagm_converged <- e14Tagm[-removeChain] # Create new object
```

The `MCMCParams` object can be large and therefore if we have a large number of samples we may want to subsample our chain, known as *thinning*, to reduce the number of samples. Thinning also has another purpose. We may desire independent samples from our posterior distribution but the MCMC algorithm produces auto-correlated samples. Thinning can be applied to reduce the auto-correlation between samples. The code chunk below, which is not

evaluated, demonstrates retaining every 5^{th} iteration. Recall that we thinned by 20 when we first ran the MCMC algorithm.

```
e14Tagm_converged_thinned <- mcmc_thin_chains(e14Tagm_converged, freq  = 5)
```

We initially ran 6 chains and, after having made an assessment of convergence, we decided to discard 3 of the chains. We desire to make inference using samples from all 3 chains, since this leads to better posterior estimates. In their current class structure all the chains are stored separately, so the following function pools all sample for all chains together to make a single longer chain with all samplers. Pooling a mixture of converged and unconverged chains is likely to lead to poor quality results so should be done with care.

```
e14Tagm_converged_pooled <- mcmc_pool_chains(e14Tagm_converged)
e14Tagm_converged_pooled
```

```
## Object of class "MCMCParams"
## Method: TAGM.MCMC
## Number of chains: 1
```

```
e14Tagm_converged_pooled[[1]]
```

```
## Object of class "MCMCChain"
##  Number of components: 10
##  Number of proteins: 1663
##  Number of iterations: 1500
```

To populate the summary slot of the converged and pooled chain, we can use the `tagmMcmcProcess` function. As we can see from the object below a summary is now available. The information now available in the summary slot was detailed in the previous section. We note that if there is more than 1 chain in the `MCMCParams` object then the chains are automatically pooled to compute the summaries.

```
e14Tagm_converged_pooled <- tagmMcmcProcess(e14Tagm_converged_pooled)
e14Tagm_converged_pooled
```

```
## Object of class "MCMCParams"
## Method: TAGM.MCMC
## Number of chains: 1
## Summary available
```

To create new feature columns in the `MSnSet` and append the summary information, we apply the `tagmPredict` function. The `probJoint` argument indicates whether or not to add probabilistic information for all organelles for all proteins, rather than just the information for the most probable organelle. The outlier probabilities are also returned by default, but users can change this using the `probOutlier` argument.

```
E14TG2aR <- tagmPredict(object = E14TG2aR,
params = e14Tagm_converged_pooled,
probJoint = TRUE)
head(fData(E14TG2aR))
```

```
##       Uniprot.ID UniprotName                          Protein.Description Peptides PSMs GOannotation markers.orig        markers   tagm.map.allocation tagm.map.probability
## Q62261     Q62261 SPTB2_MOUSE Spectrin beta chain, brain 1 (multiple isoforms)       42   42      PLM-SKE      unknown        unknown Endoplasmic reticulum         8.165817e-09
## Q9JHU4     Q9JHU4 DYHC1_MOUSE              Cytoplasmic dynein 1 heavy chain 1       33   33          SKE      unknown        unknown   Nucleus - Chromatin         9.996798e-01
## Q9QXS1     Q9QXS1  PLEC_MOUSE                      Isoform PLEC-1I of Plectin       33   33      unknown      unknown        unknown       Plasma membrane         1.250898e-06
## P16546     P16546 SPTA2_MOUSE  Spectrin alpha chain, brain (multiple isoforms)       32   32 PLM-SKE-CYT      unknown        unknown   Nucleus - Chromatin         4.226696e-07
## Q69ZN7     Q69ZN7  MYOF_MOUSE                   Myoferlin (multiple isoforms)       28   28          VES      unknown        unknown       Plasma membrane         9.994502e-01
## P30999     P30999 CTND1_MOUSE             Catenin delta-1 (multiple isoforms)       24   24      PLM-NUC          PLM Plasma membrane       Plasma membrane         1.000000e+00
##       tagm.map.joint.40S Ribosome tagm.map.joint.60S Ribosome tagm.map.joint.Cytosol tagm.map.joint.Endoplasmic reticulum tagm.map.joint.Lysosome tagm.map.joint.Mitochondrion
## Q62261               2.543800e-02               3.905306e-02          1.581542e-01                       1.430889e-01           5.992007e-02                2.133626e-01
## Q9JHU4               8.145897e-06               1.250578e-05          5.064502e-05                       4.582071e-05           1.918793e-05                6.832416e-05
## Q9QXS1               2.543797e-02               3.905301e-02          1.581540e-01                       1.430887e-01           5.991999e-02                2.133624e-01
## P16546               2.543799e-02               3.905304e-02          1.581542e-01                       1.430889e-01           5.992004e-02                2.133625e-01
## Q69ZN7               2.755266e-06               4.229952e-06          1.713015e-05                       1.549838e-05           4.479797e-04                2.310994e-05
## P30999               0.000000e+00               0.000000e+00          0.000000e+00                       0.000000e+00           0.000000e+00                0.000000e+00
##       tagm.map.joint.Nucleus - Chromatin tagm.map.joint.Nucleus - Nucleolus tagm.map.joint.Plasma membrane tagm.map.joint.Proteasome tagm.map.outlier  tagm.mcmc.allocation
## Q62261                       7.280971e-02                     9.016054e-02                  1.859906e-01             1.202229e-02     0.9999999857 Endoplasmic reticulum
## Q9JHU4                       9.997031e-01                     2.887171e-05                  5.955892e-05             3.849844e-06     0.0003202255   Nucleus - Chromatin
## Q9QXS1                       7.280961e-02                     9.016043e-02                  1.859916e-01             1.202228e-02     0.9999987491            Proteasome
## P16546                       7.281009e-02                     9.016050e-02                  1.859905e-01             1.202228e-02     0.9999995462 Endoplasmic reticulum
## Q69ZN7                       7.886235e-06                     9.765556e-06                  9.994703e-01             1.302170e-06     0.0001083130       Plasma membrane
## P30999                       0.000000e+00                     0.000000e+00                  1.000000e+00             0.000000e+00     0.0000000000       Plasma membrane
##       tagm.mcmc.probability tagm.mcmc.probability.lowerquantile tagm.mcmc.probability.upperquantile tagm.mcmc.mean.shannon tagm.mcmc.outlier tagm.mcmc.joint.40S Ribosome
## Q62261             0.5765793                       0.0020296117                       0.9992504           0.201623229      2.547793e-01                4.401228e-10
## Q9JHU4             0.9738206                       0.7594516090                       0.9998822           0.081450206      3.335134e-05                1.936225e-18
## Q9QXS1             0.4957129                       0.0002886457                       0.9947100           0.447665536      6.423799e-01                2.213861e-07
## P16546             0.5214374                       0.0014041362                       0.9946959           0.252833750      2.119112e-01                1.576023e-09
## Q69ZN7             0.9997025                       0.9981794326                       0.9999954           0.002395147      7.274103e-06                3.510523e-22
## P30999             1.0000000                       1.0000000000                       1.0000000           0.000000000      0.000000e+00                0.000000e+00
##       tagm.mcmc.joint.60S Ribosome tagm.mcmc.joint.Cytosol tagm.mcmc.joint.Endoplasmic reticulum tagm.mcmc.joint.Lysosome tagm.mcmc.joint.Mitochondrion
## Q62261               2.778620e-07           2.650861e-12                       5.765793e-01           1.108757e-11                5.020528e-08
## Q9JHU4               1.645727e-21           1.887645e-17                       1.548053e-17           5.577415e-24                2.835919e-22
## Q9QXS1               1.495170e-01           9.062280e-09                       1.768681e-04           1.150706e-04                5.832273e-19
## P16546               3.150122e-06           1.471329e-08                       5.214374e-01           3.687975e-09                4.522032e-08
## Q69ZN7               5.152312e-16           2.063009e-24                       8.397027e-09           2.974966e-04                6.143974e-39
## P30999               0.000000e+00           0.000000e+00                       0.000000e+00           0.000000e+00                0.000000e+00
##       tagm.mcmc.joint.Nucleus - Chromatin tagm.mcmc.joint.Nucleus - Nucleolus tagm.mcmc.joint.Plasma membrane tagm.mcmc.joint.Proteasome
## Q62261                       4.231731e-01                     1.279255e-05                  1.914808e-11             2.345204e-04
## Q9JHU4                       9.738206e-01                     2.617943e-02                  3.514851e-29             7.841425e-11
## Q9QXS1                       7.920397e-03                     1.130580e-05                  3.465462e-01             4.957129e-01
## P16546                       4.776913e-01                     3.448558e-05                  2.489652e-07             8.333595e-04
## Q69ZN7                       4.872032e-21                     7.042891e-30                  9.997025e-01             1.003778e-10
## P30999                       0.000000e+00                     0.000000e+00                  1.000000e+00             0.000000e+00
```

3.6.3 Priors

Introduction

Bayesian analysis requires users to specify prior information about the parameters. This may appear to be a challenging task; however, good default options are often possible. Should expert information or domain specific knowledge be available for any of these priors then the users should provide this, otherwise we have found that the default choices work well in practice.

The priors also provide regularisation and shrinkage to avoid overfitting. Given enough data the likelihood overwhelms the prior and the influence of the prior is weak [162].

Empirical Bayes priors on the mixture components

We place a normal inverse-Wishart prior on the parameters of the multivariate normal mixture components. The normal inverse-Wishart prior has 4 hyperparameters that must be specified. These are: the prior mean `mu0` expressing the prior location of each organelle; a prior shrinkage `lambda0`, which is a scalar expressing uncertainty in the prior mean; the prior degrees of freedom `nu0`; and a scale prior `S0` on the covariance. Together, `nu0` and `S0` specify the prior variability on organelle covariances. The same prior distribution is assumed for the parameters of all multivariate normal mixture components.

An empirical Bayes approach is used to set these priors, which is a pragmatic approach when little prior information is known. The choices for these priors are based on the recommendation by [138]. The prior mean `mu0` is set to be the mean of the data. `lambda0` is set to be 0.01 meaning some uncertainty in the covariance is propagated to the mean, increasing `lambda0` increases shrinkage towards the prior. `nu0` is set to the number of feature variables plus 2, which is the smallest integer value that ensures a finite covariance matrix. The prior scale matrix $S0$ is set to

$$S_0 = \frac{\text{diag}(\frac{1}{n}\sum(X - \bar{X})^2)}{K^{1/D}}, \tag{3.1}$$

and represents a diffuse prior on the covariance. Another good choice, which is often used, is a constant multiple of the identity matrix [412].

Prior on the mixing proportions

The prior on the mixing proportions is the Dirichlet distribution with concentration parameters `beta0` set to 1 for each organelle. Another reasonable choice would be the non-informative Jeffery's prior for the Dirichlet hyperparameter, which sets `beta0` to 0.5 for each organelle. The following discussion assesses the quality and sensitivity of our prior choice. We compute the posterior z-score which assesses how the posterior recovers the assumed true model configuration with small values for the posterior z-score suggesting good calibration [29]. We also compute the posterior shrinkage, which quantifies how much is learnt about a given parameter from the data [29]. Values of the posterior shrinkage close to 1 suggest that the parameter values are strongly informed by the data.

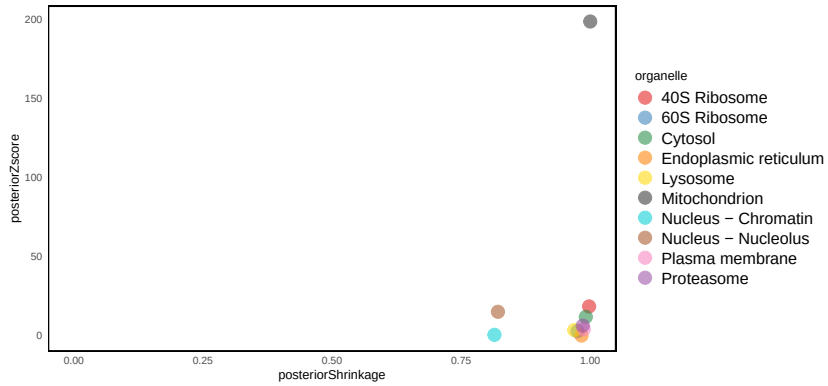


Fig. 3.8 Scatter plot of the posterior shrinkage against the posterior z-score for the mixing proportions of the model

```
mixing_posterior_check(object = E14TG2aR,
params = e14Tagm_converged_pooled[[1]],
priors = e14Tagm@priors)
```

We see that most parameter values concentrate in the lower right hand corner of the plot, which suggests good shrinkage and calibration. However, the parameter for the mitochondrion is located in the top right of the plot suggesting the posterior deviates from the prior. The biological interpretation for this is that the experiment resolved the mitochondrial proteins extremely well and thus allocated many more proteins to this class than perhaps we might have expected. This could be remedied with a more informative prior. If we prefer to use an informative prior, rather than a non-informative prior, it is practical to use information from previous data. To demonstrate this, we consider another experiment on mouse pluripotent stem cells and examine the number of proteins that were allocated to each subcellular niche. The code chunk below extracts this information from another spatial proteomics experiment.

```
data("hyperLOPIT2015")
priordata <- table(fData(hyperLOPIT2015)$final.assignment)
priordata
```

```
##
##                         40S Ribosome                         60S Ribosome                   Actin cytoskeleton                          Cytosol
##                                   48                                   62                                   46                                  339
## Endoplasmic reticulum/Golgi apparatus                            Endosome                  Extracellular matrix                        Lysosome
##                                  426                                   60                                   17                                   80
##                           Mitochondrion                  Nucleus - Chromatin               Nucleus - Non-chromatin                      Peroxisome
##                                  585                                  297                                  396                                   25
##                       Plasma membrane                           Proteasome                              unknown
##                                  392                                   34                                 2225
```

It is clear that the allocations are not uniformly distributed across the classes and that the mitochondrion has more allocations than the other subcellular niches. However, we also do

not have prior information on all the classes. The Dirichlet distribution can be interpreted as specifying the prior relative proportions of the number of proteins allocated to each niche. For the classes where we have no information, we assume equal uniform allocations. First, we compute the number of proteins in this experiment. Then create a vector with proteins allocated equally to each class.

```
N <- nrow(unknownMSnSet(E14TG2aR)) # number of proteins
K <- length(getMarkerClasses(E14TG2aR)) # number of subcellular niches
# uninformative beta0, proteins allocated symmetrically
beta_uninformed <- rep(N/K, K)
names(beta_uninformed) <- getMarkerClasses(E14TG2aR)
```

The code chunk below extracts the data for which we have prior information.

```
shared_info <- intersect(getMarkerClasses(hyperLOPIT2015),
getMarkerClasses(E14TG2aR))
# extracts useful information from other dataset
informativePrior <- priordata[shared_info]
```

We then reweight the prior number of proteins allocated to each class by their relative proportions in the other dataset. We then use this information to create an informative prior.

```
beta_informed <- beta_uninformed
beta_informed[shared_info] <- sum(beta_uninformed[shared_info]) *
 informativePrior/sum(informativePrior)
```

Now, we can check that this prior has captured our beliefs correctly, mainly that the mitchondrion should have more allocations than the other subcellular niches and that distribution is not symmetric. To do this, we simulate 10000 values from the informative prior and compute the expected (prior) number of proteins allocated to each niche.

```
prior_simulation <- colMeans(gtools::rdirichlet(n = 10000,
alpha = beta_informed) * N)
names(prior_simulation) <- getMarkerClasses(E14TG2aR)
prior_simulation
```

```
##        40S Ribosome          60S Ribosome
##            34.72586              44.99799
##             Cytosol  Endoplasmic reticulum        Lysosome
##           245.62171             166.50286           57.99014
```

```
##         Mitochondrion   Nucleus - Chromatin   Nucleus - Nucleolus
##             423.32247             215.53612             166.26266
##       Plasma membrane            Proteasome
##             283.48035              24.55985
```

```
par(mar = c(11.5, 6.5, 0.5, 0.5), mgp = c(10, 1, 0))
barplot(prior_simulation, las = 2, col = "darkgreen", xlab =
 "sub-cellular niche", ylab = "prior expected number of allocations")
```

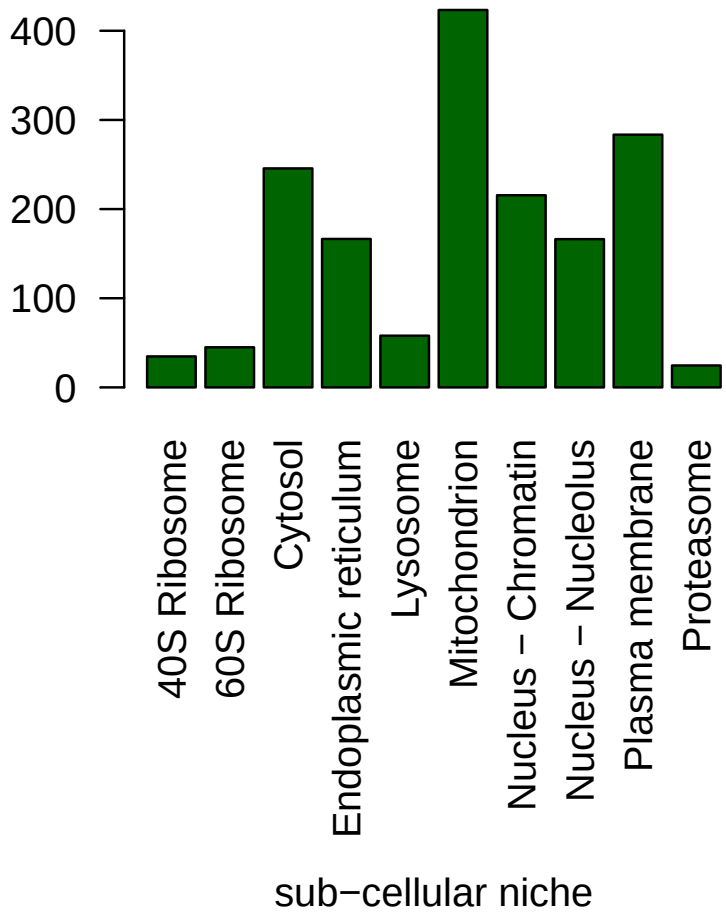


Fig. 3.9 A barplot showing the expected (prior) number of proteins allocated to each niche

It is clear that this prior captures the information that the mitochondrion has more allocations than the other subcellular niches and that the allocations across the classes are not symmetric. It is useful to note that many spatial proteomics datasets can be found in the `pRolocdata` package from which useful information could be extracted.

Prior on the proportion of outlier proteins

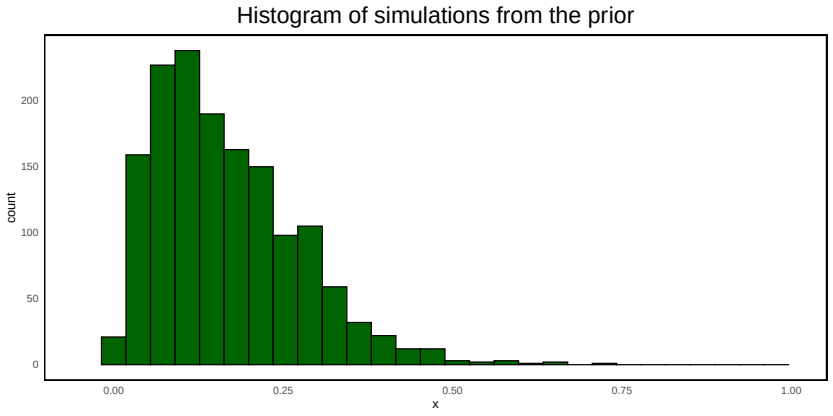
The prior for the proportion of outlier proteins is a $\mathcal{B}(u, v)$ distribution. The default for $u = 2$ and the default for $v = 10$. This represents the reasonable belief that $\frac{u}{u+v} = \frac{1}{6}$ proteins *a priori* might be an outlier and we believe is unlikely that more than 50% of proteins are outliers, which was elicited from expert domain knowledge and analysis of previous datasets. Decreasing the value of v, represents more uncertainty about the number of proteins that are outliers.

To visualise that this prior captures these beliefs, we simulate from the prior and produce a histogram.

```
x <- rbeta(n = 1500, shape1 = 2, shape2 = 10)
gg <- ggplot(data.frame(x), aes(x)) + geom_histogram(fill = "darkgreen",
col = "black") +
      theme_minimal() +
      theme(panel.grid.major =
element_blank(), panel.grid.minor = element_blank(),
panel.border = element_rect(colour = "black", fill = NA, size = 1),
plot.title = element_text(hjust = 0.5, size = 20),
legend.text=element_text(size = 14)) +
      ggtitle(label = "Histogram of simulations from the prior") +
xlim(c(-0.05, 1))
                gg
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 2 rows containing missing values (geom_bar).
```



The probability that more than 50% of the proteins are outliers is small but non-zero. The probability there are fewer than 1% outliers is also small. These quantiles can be used to calibrate the prior beliefs.

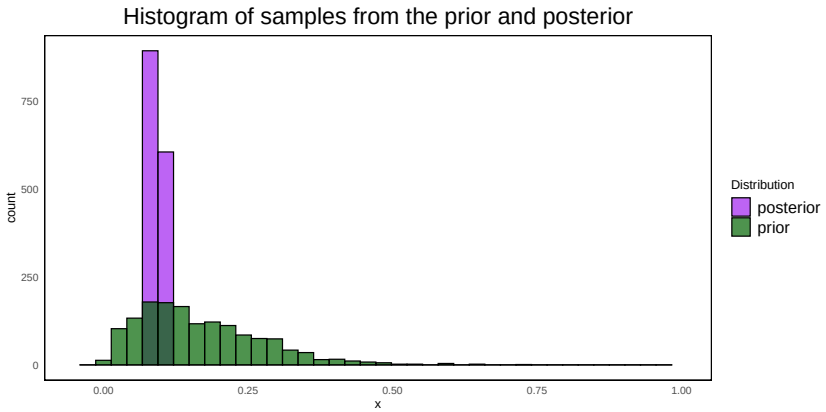
```
pbeta(0.5, shape1 = 2, shape2 = 10, lower.tail = FALSE) # more than 50% outliers
```

```
## [1] 0.005859375
```

Now we turn to the posterior distribution for this quantity of interest and overlay onto the prior.

```
out <- mcmc_get_outliers(e14Tagm_converged_pooled)
propout <- out[[1]]/nrow(unknownMSnSet(E14TG2aR))
df <- data.frame(x = c(x, propout),
y = as.factor(rep(c("prior", "posterior"), each = 1500)))
gg <- ggplot(df, aes(x = x, fill = y)) +
 geom_histogram(alpha = 0.7, col = "black", position = "identity",
bins = 40) +
      theme_minimal() +
      theme(panel.grid.major = element_blank(),
panel.grid.minor = element_blank(),
panel.border = element_rect(colour = "black", fill = NA, size = 1),
plot.title = element_text(hjust = 0.5, size = 20),
legend.text=element_text(size = 14)) + labs(fill = "Distribution") +
      scale_fill_manual(values = c("purple", "darkgreen")) +
      ggtitle(label = "Histogram of samples from the prior and posterior") +
 xlim(c(-0.05, 1))
gg
```

```
## Warning: Removed 4 rows containing missing values (geom_bar).
```



It is clear that the prior and posterior concentrate in the same region, and are thus not in conflict. The variance of the posterior is clearly smaller than that of the prior and so

there is high posterior shrinkage. One could argue that the prior is too diffuse to provide regularisation; however, specifying tighter priors risk biasing the model away from the data generating mechanism.

3.6.4 Analysis, visualisation and interpretation of results

Now that we have a single pooled chain of samples from a converged MCMC algorithm, we can begin to analyse the results. Preliminary analysis includes visualising the allocated organelle and localisation probability of each protein to its most probable organelle, as shown on figure 3.10.

```
layout(matrix(c(1,1,2,2), nrow = 4, ncol = 1, byrow = TRUE))
plot2D(E14TG2aR, fcol = "tagm.mcmc.allocation",
cex = fData(E14TG2aR)$tagm.mcmc.probability,
main = "TAGM MCMC allocations")
addLegend(E14TG2aR, fcol = "markers",
where = "topleft", ncol = 2, cex = 0.6)

plot2D(E14TG2aR, fcol = "tagm.mcmc.allocation",
cex = fData(E14TG2aR)$tagm.mcmc.mean.shannon,
main = "Visualising global uncertainty")
addLegend(E14TG2aR, fcol = "markers",
where = "topleft", ncol = 2, cex = 0.6)
```

We can visualise other summaries of the data including a Monte-Carlo averaged Shannon entropy, as shown in figure 3.10 on the right. This is a measure of uncertainty and proteins with greater Shannon entropy have more uncertainty in their localisation. The Shannon Entropy (and hence uncertainty) is greatest when all localisations are equiprobable and lowest when the probabilities are concentrated on a single localisation. For additional discussion, we refer readers to Crook et al. [83] and Crook et al. [86] and references therein. We observe global patterns of uncertainty, particularly in areas where organelle boundaries overlap. There are also regions of low uncertainty indicating little doubt about the localisation of these proteins.

We are also interested in the relationship between localisation probability to the most probable class and the Shannon entropy. Even though the two quantities are evidently correlated there is still considerable spread. Thus it is important to base inference not only on localisation probability but also a measure of uncertainty, for example the Shannon entropy. Proteins with low Shannon entropy have low uncertainty in their localisation, whilst those with higher Shannon entropy have uncertain localisation. Since multi-localised proteins have uncertain localisation to a single subcellular niche, exploring the Shannon can aid in identifying multi-localised proteins.

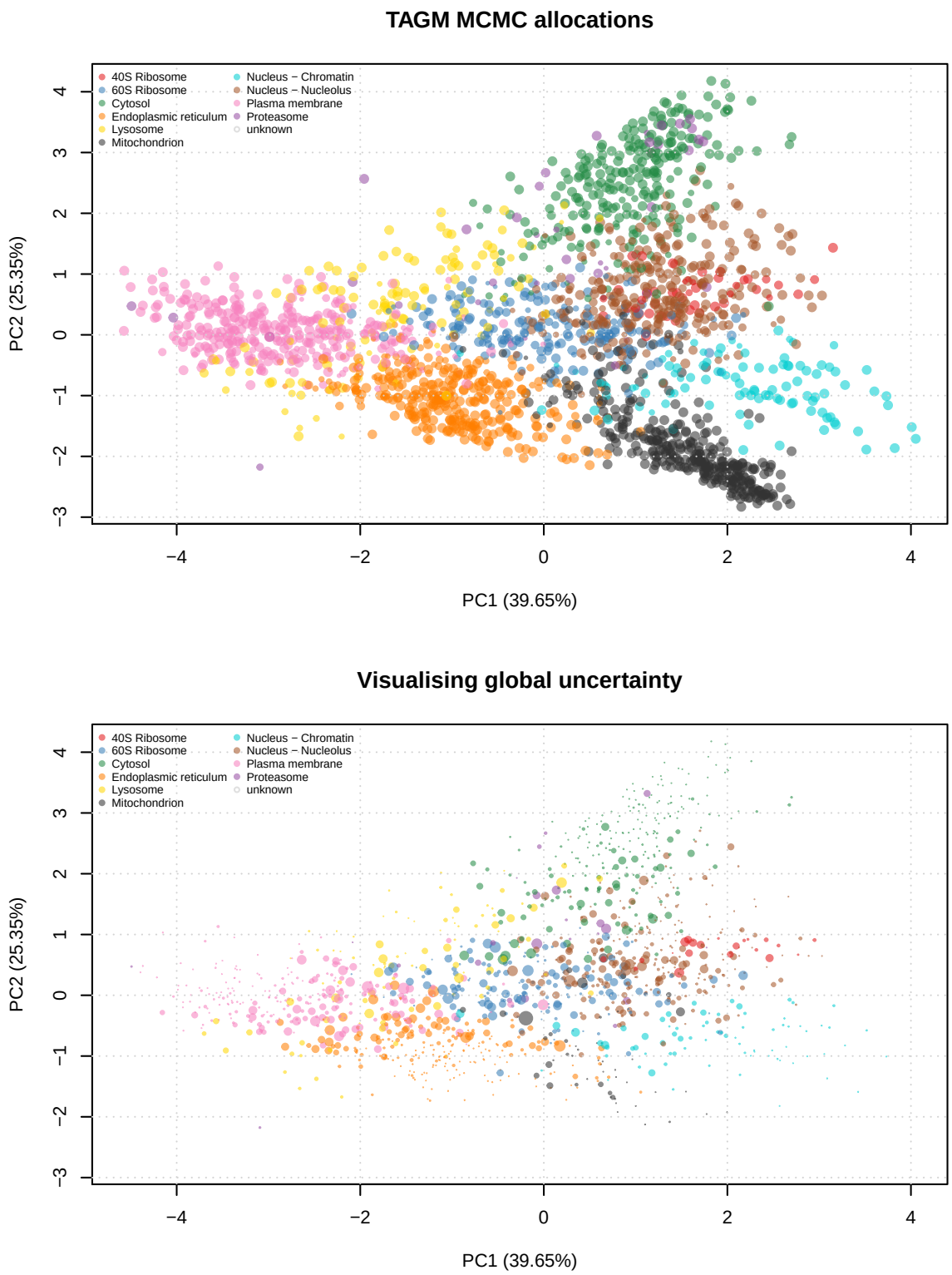


Fig. 3.10 TAGM MCMC allocations. In the upper plot, pointer sizes have been scaled based on allocation probabilities. On the lower plot, the pointer sizes have been scaled based on the global uncertainty using the mean Shannon entropy.

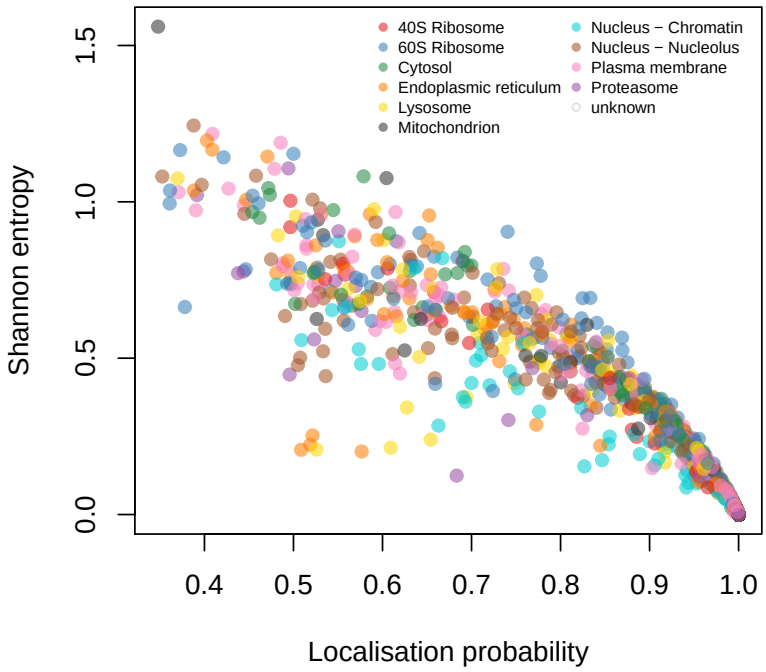


Fig. 3.11 Shannon entropy and localisation probability.

Examples of well characterised multi-localising proteins from the literature are discussed in [83]. The interpretation of uncertain allocations in relation to multi-localisation is further discussed in [83, 86].

```
cls <- getStockcol()[as.factor(fData(E14TG2aR)$tagm.mcmc.allocation)]
plot(fData(E14TG2aR)$tagm.mcmc.probability,
fData(E14TG2aR)$tagm.mcmc.mean.shannon,
col = cls, pch = 19,
xlab = "Localisation probability",
ylab = "Shannon entropy")
addLegend(E14TG2aR, fcol = "markers",
where = "topright", ncol = 2, cex = 0.6)
```

There are further ways in which we can visualise the uncertainty quantified by the Bayesian analysis. For example, we can use the samples from the MCMC algorithm to visualise the uncertainty in the mean localisation of each organelle/niche on a PCA plot. At each iteration of the MCMC, we compute the mean for each organelle as the mean of all associated proteins

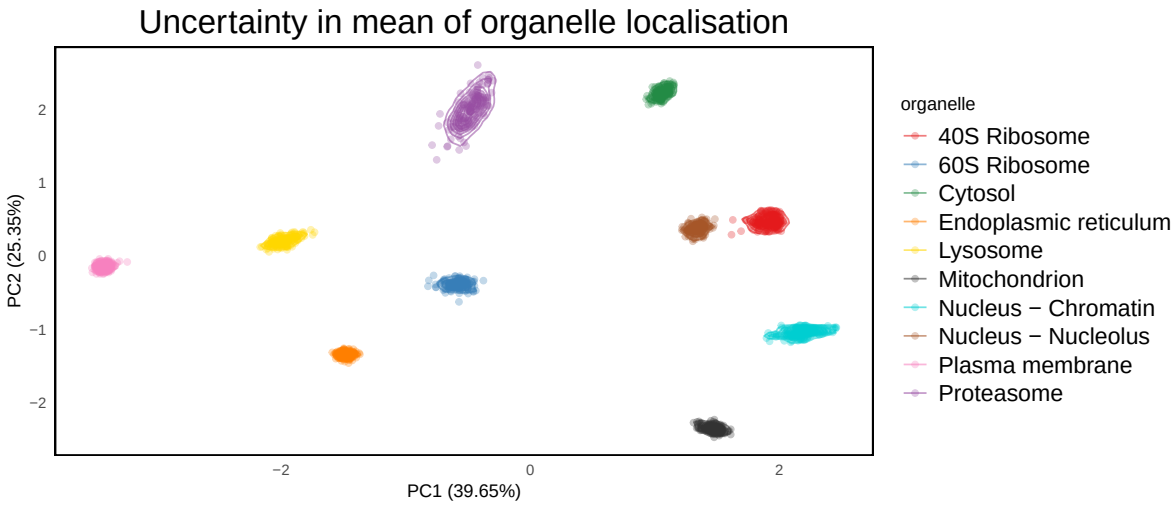


Fig. 3.12 Visualising uncertainty in the mean of each subcellular niche. The pointers correspond to results from different iterations of the MCMC algorithm and are coloured according to the corresponding subcellular niche

to that organelle. These data are then projected on to the PCA plot, having aligned them across the random samples (see [37], as well as [379] and [337] for similar examples).

```
nicheMeans2D(object = E14TG2aR, params = e14Tagm_converged_pooled[[1]],
prior = e14Tagm_converged_pooled@priors)
```

The main quantity of interest is the posterior localisation probability of each protein to each organelle. However, visualising how these probabilities vary in different regions of data space are be challenging, especially with large numbers of proteins. Furthermore, interrogating individual proteins one by one can be cumbersome. Thus, we consider visualising how the probabilities vary across different regions of the PCA plot. To perform this analysis, we first compute the underlying coordinates of the whole data in PC space. That is if $P = XW$ is the PCA decomposition of the data matrix X and W is a matrix whose columns are the eigenvectors of X^TX. Then the i^{th} row of P, P_i, gives the coordinates of measurement x_i in PC space. We can then proceed by linearly interpolating a regular grid in this coordinate system. To obtain the localisation probabilities on this grid, we use a Nadaraya-Watson kernel smoother [329, 474]. Let $Y(v) : \mathbb{R}^p \to \mathbb{R}$ be $C^1(\mathbb{R})$. For each $v_0 \in \mathbb{R}^p$, the Nadaraya-Watson kernel smoother parameterised by λ is

$$Y(v_0) = \frac{\sum_{i=1}^{N} K_\lambda(v_0, v_i) Y(v_i)}{\sum_{i=1}^{N} K_\lambda(v_0, v_i)}, \tag{3.2}$$

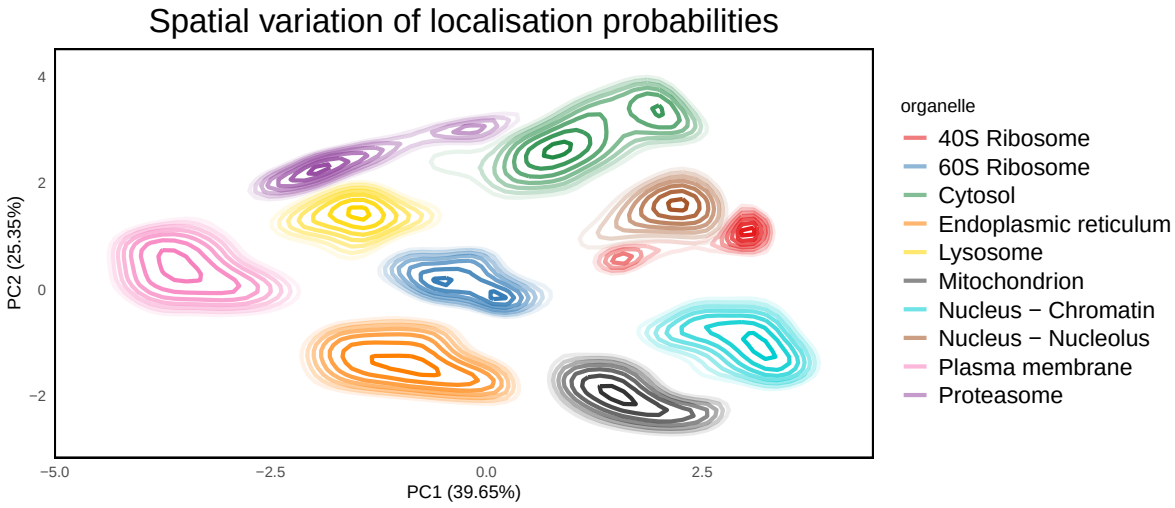


Fig. 3.13 Visualising how the posterior localisation probabilities vary smoothly across different regions of the PCA plot. The colours correspond the different subcellular niches. The inner most contour corresponds to a probability of 0.99 and the following contour to 0.95, with each subsequent contour descreasing in 0.05 increments

for N observed points and $Y(v_j)$ is the observation at v_j. A number of kernels are available and we opt for the Wendland kernel [476]. A fast Fourier transform is used to accelerate computations [424, 434, 175]. A contour plot, in the PC coordinates, of these probabilities is then visualised, where the distribution for each organelle is coloured accordingly (see figure 3.12). The code chunk below produces this plot.

```
spatial2D(object = E14TG2aR)
```

Aside from global visualisation of the data, we can also interrogate each individual protein. As illustrated on figure 3.14, we can obtain the full posterior distribution of localisation probabilities for each protein from the `e14Tagm_converged_pooled` object. We can use the `plot` generic on the `MCMCParams` object to obtain a violin plot of the localisation distribution. Simply providing the name of the protein in the second argument produces the plot for that protein. The solute carrier transporter protein E9QMX3, also referred to as Slc15a1, is most probably localised to plasma membrane in line with its role as a transmembrane transporter but also shows some uncertainty, potentially also localising to other compartments. The first violin plot visualises this uncertainty. The protein Q3V1Z5 is a supposed constitute of the 40S ribosome and has poor UniProt annotation with evidence only at the transcript level. From the plot below it is clear that Q3V1Z5 is a ribosomal associated protein, but its previous localisation

has only been computationally inferred and here we provide experimental evidence of a ribosomal annotation. Thus, quantifying uncertainty recovers important additional annotations.

```
plot(e14Tagm_converged_pooled, "E9QMX3")
plot(e14Tagm_converged_pooled, "Q3V1Z5")
```

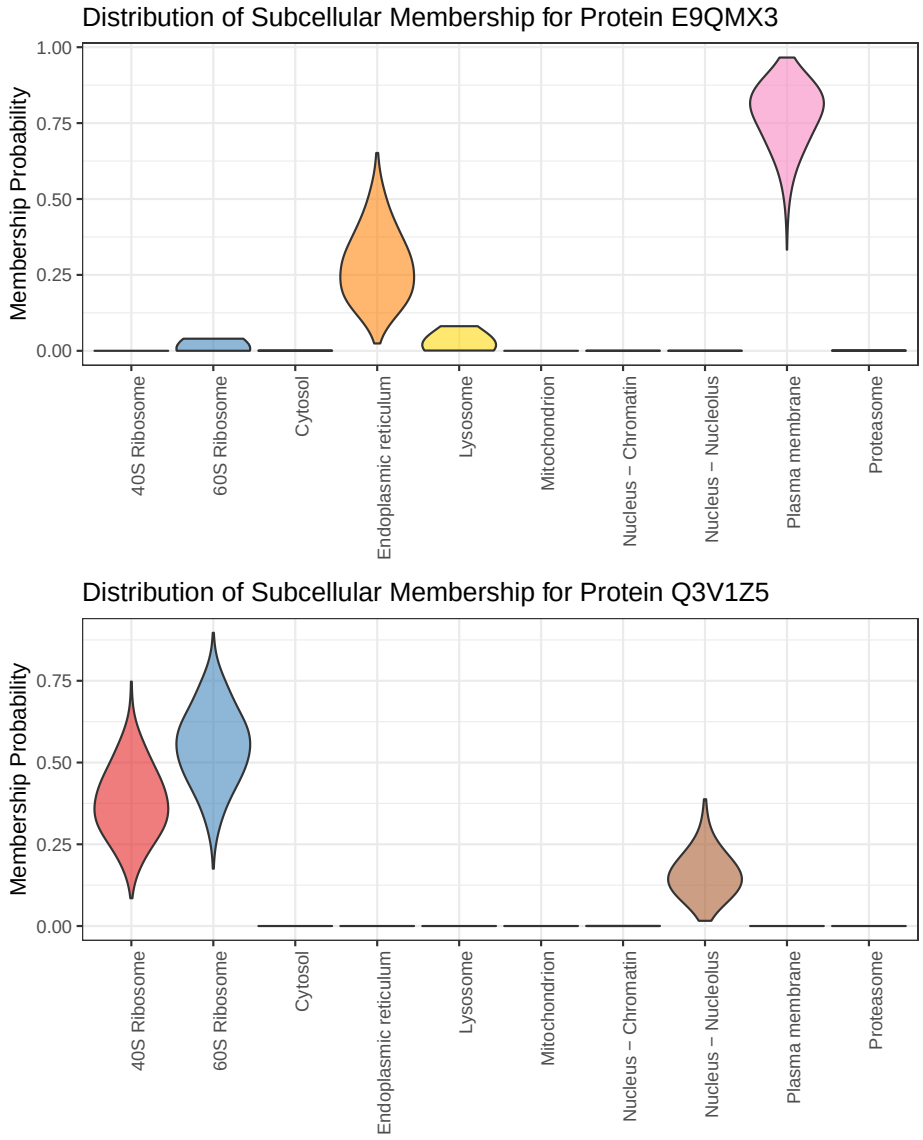


Fig. 3.14 Full posterior distribution of localisation probabilities for individual proteins.

3.7 Discussion and limitations

The Bayesian analysis of biological data is of clear interest to many because of its ability to provide richer information about the experimental results. A fully Bayesian analysis differs from other machine learning approaches, since it can quantify the uncertainty in our inferences. Furthermore, we use a generative model to explicitly describe the data, which makes inferences more interpretable compared to the less interpretable outputs of black-box classifiers such as, for example, support vector machines (SVM).

Bayesian analysis is often characterised by its provision of a (posterior) probability distribution over the biological parameters of interest, as opposed to single point estimate of these parameters. In the case that is presented in this workflow, a Bayesian analysis "computes" a posterior probability distribution over the protein localisation probabilities. These probability distributions can then be rigorously interrogated for greater biological insight; in addition, it may allow us to ask additional questions about the data, such as whether a protein might be multi-localised.

Despite the wealth of information a Bayesian analysis can provide, the uptake amongst cell biologists is still low. This is because a Bayesian analysis presents a new set of challenges and little practical guidance exists regarding how to address these challenges. Bayesian analyses often rely on computationally intensive approaches such as Markov-chain Monte-Carlo (MCMC) and a practical understanding of these algorithms and the interpretation of their output is a key barrier to their use. A Bayesian analysis usually consists of three broad steps: (1) Data pre-processing and algorithmic implementation, (2) assessing algorithmic convergence and (3) summarising and visualising the results. This workflow provides a set of tools to simplify these steps and provides step-by-step guidance in the context of the analysis of spatial proteomics data.

We have provided a workflow for the Bayesian analysis of spatial proteomics using the `pRoloc` and `MSnbase` software. We have demonstrated, in a step-by-step fashion, the challenges and advantages associated with taking a Bayesian approach to data analysis. We hope this workflow will help spatial proteomics practitioners to apply our methods and will motivate others to create detailed documentation for the Bayesian analysis of biological data.

Of course a workflow can always be expanded to provide ever more details on the analysis. Further directions for improving the software suite is perhaps more guidance for users on performing *prior* and *posterior predictive checks*. There is also scope for automatically pulling data from databases about the proteins of interest to alleviate manual literature searches.

Chapter 4

A subcellular atlas of *Toxoplasma* reveals functional context of the proteome

This chapter applies the ideas developed in the previous two chapters to a complex application. The spatial proteome of *Toxoplasma gondii* is poorly characterised and there are few known ground truths, thus it is an excellent use case for uncertainty quantification. The hyperLOPIT experiments were performed by Konstantin Barylyuk (KB) and the gene editing was performed by Ludek Koreny, Huiling Ke and Simon Butterworth. Myself and Konstantin performed the data analysis and this chapter borrows from [21]. Figures are used with kind permission of Konstantin Barylyuk.

4.1 Motivation

4.1.1 Abstract

Spatial proteomics methods are now well established for model organisms [68, 220, 324, 159, 339]. However, these techniques are most promising as a discovery tool in poorly annotated organisms. In this chapter, we demonstrate that hyperLOPIT can be applied to an apicomplexan cell. Apicomplexan parasites are major burdens on human health and food security. *Toxoplasma gondii* is a master biochemist with highly specialised cellular machinery, which allows elaborate modulation of the human host. The evolutionary adaptation of the subcellular landscape comes with extensive proteomic novelty. However, a majority of *T. gondii* proteins are hypothetically predicted from the genome and a majority have no or simply generic functional annotation. The application of hyperLOPIT, here, provides huge knowledge expansion of these parasites. The lack of functional annotations to verify results motivates uncertainty quantification and we

demonstrate that the Bayesian model developed in the previous chapters provides localisation information for thousands of proteins to 26 subcellular niches and sub-niches. We map genomic features onto our spatial atlas and show that compartmental adaptation and evolution is heterogeneous across the landscape.

4.2 Introduction and literature review

The protozoan phylum apicomplexa contains thousands of species of intracellular parasites [468]. Amongst these parasites are *Toxoplasma gondii* and *Plasmodium falciparum*, the primary parasitic agents of Toxoplasmosis and Malaria, respectively [468]. These parasites have been highly successful and infect potentially every vertebrate and most invertebrates [251, 438]. Toxoplasmosis causes chronic infection in 30% of the human population, as well as congenital toxoplasmosis, foetal malformation and abortion, retinochoroiditis and encephalitis [189]. The success of these parasites is, in part, due to their specialised cell compartments [342]. For example, the "apical complex" enables penetration and invasion of animal cells without destruction of the host [210]. This complex includes a set of secretory organelles (micronemes, rhoptries and dense granules) which release biomolecules required for locating, recognising, penetrating and exploiting the host [59, 117, 401, 451, 402]. Furthermore, Toxoplasma contains a complete set of usual eukaryotic compartments [182, 233, 359], as well as a remnant endosymbiotic organelle: a plastid (apicoplast) [248, 394, 136]. The typical ER, Golgi complex, nucleus are all present within Toxoplasma [233], along with a single copy of the mitochondrion [413, 305, 459] . However, Toxoplasma has also developed novel organisation. Upon invasion of the host, the parasite remains within a "parasitophorous vacuole" embellished with parasite proteins [59, 117]. This highly synchronised and specialised organellar biology has made Toxoplasma a master biochemist allowing it unfettered modulation of its host [361, 183].

T. gondii has three infectious stages [116] (see figure 4.1). This involves the tachyzoites, bradyzoites and sporozoites, which form part of the complex life cycle of Toxoplasma [116]. The sporozoites (oocysts) are passed from the definitive feline host via their faecal matter to another host [130]. When ingested by livestock, usually through contaminated water or feed, the oocysts form tissue cysts (bradyzoites) [129]. In uncooked meat, the parasites are passed onto human or feline hosts [116]. Within non-intestinal epithelial cells the resultant tachyzoites rapidly multiply (and can even infect the foetus via the placenta) [116]. The tachyzoites display the most complex compartmentalisation, though are not entirely different from bradyzoites and differ only in their nucleus placement and rhoptry structure [116] (see figure 4.2 and figure 4.3).

Tachyzoites penetrate through the host cell plasma membrane and form a parasitophorous vacuole (PV) [416]. Within the PV a tubulovesicular membranous network develops derived from the posterior of the tachyzoite [416] (see figure 4.4). Apicomplexa boast a surprising

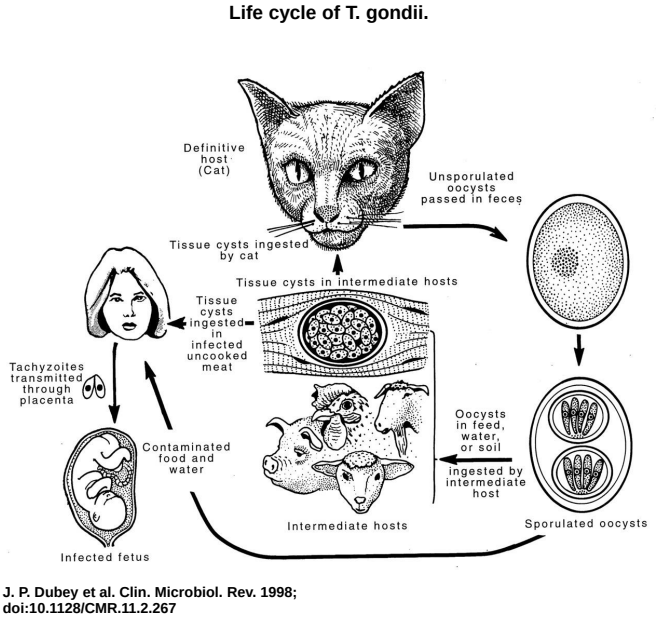


Fig. 4.1 The life cycle of *Toxoplasma gondii*. Toxoplasma transitions from its definitive feline host to intermediate ruminants via infected feed. Ingestion of contaminated meat allows humans to become infected. Figure taken from Dubey et al. [116].

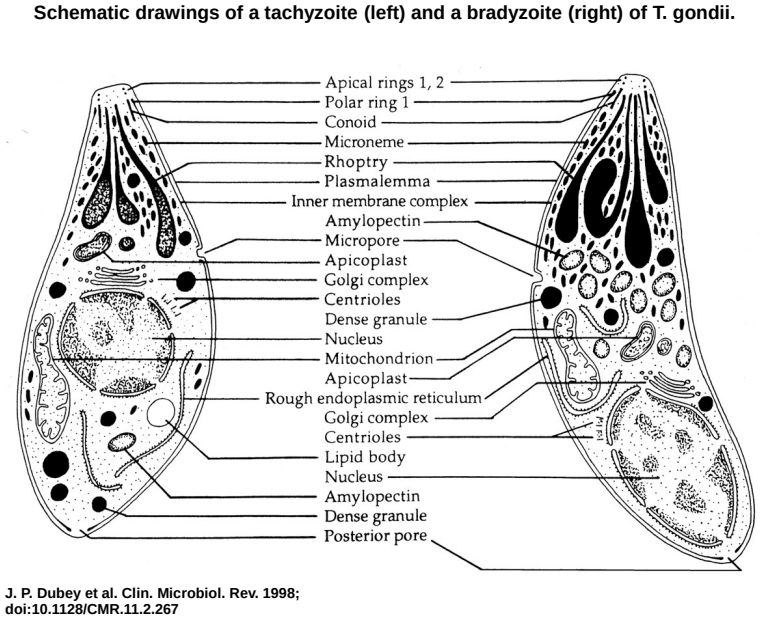


Fig. 4.2 Schematic figures of tachyzoite and bradyzoite *Toxoplasma gondii*. Toxoplasma displays complex subcellular organisation with highly polarised organelle structure. Figure taken from Dubey et al. [116].

strategy for replication: assembling daughter cells *de novo* within the cytoplasm [416, 209]. This contrasts with the familiar events of binary fission in other eukaryotes [7]. This endodyogeny generates progeny within the parent parasite before consuming it [416]. This process requires an extraordinary set of physical and molecular events to occur (see Nishi et al. [342] for more detail), including deriving their own membrane from that of the mother. Fluorescent markers have allowed the study of organellar dynamics in living parasites with detail [439, 440, 191, 209, 359, 121, 461, 464, 186, 210]. For example, the apicoplast has been shown to replicate almost synchronously with nuclear division during early endodyogeny [440]. Nishi et al. [342] use time-lapse microscopy to elucidate how the complex process of endodyogeny and highly polarised organisation of the parasite are compatible. They note that the mitochondrion enters the daughters very late in the cycle whilst the micronemes and rhoptries are generated entirely *de novo*.

The organellar structure of Toxoplasma can be seen in 3 parts (see figure 4.5). Firstly, the bare essential organelles that allow basic molecular processes to be carried out, such as metabolism and intracellular signalling. Secondly, the organelles that are specifically adapted to host mechanisms and, finally, those that allow the transition to parasitism. The huge divergence of these apicomplexan compartments is concordant with a wealth of genomic and proteomic

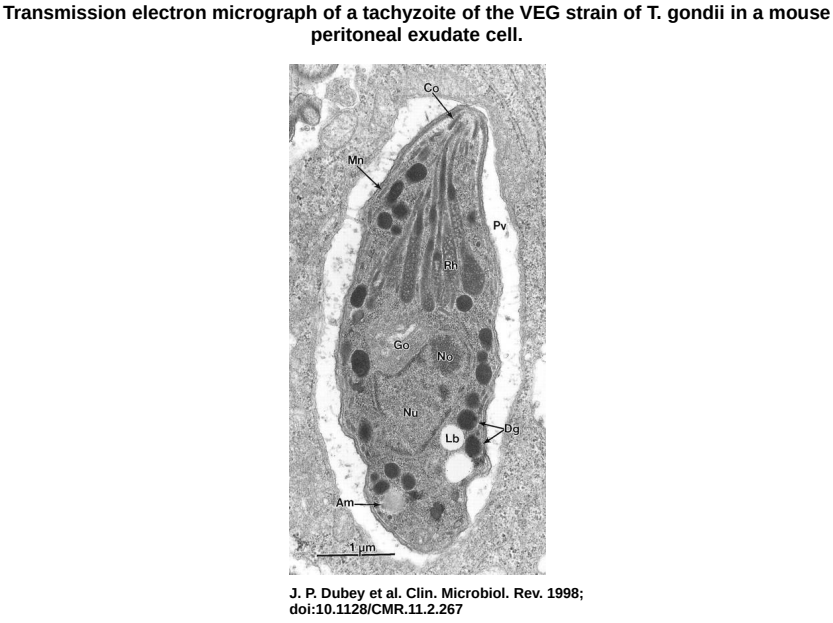


Fig. 4.3 Trasmission electron micrograph of tachyzoite of *T. gondii* within a mouse peritoneal exuldate cell. The complex organelle structure is observed. Annotations are as follows: Am, amylopectin granule; Co, conoid; Dg, electron-dense granule; Go, Golgi complex; Mn, microneme; No, nucleolus, Nu, nucleus; Pv, parasitophorous vacuole; Rh, rhoptry. Figure taken from Dubey et al. [116].

novelty [442]. Unique apicomplexan proteins are frequently annotated as hypothetical and each lineage, such as Toxoplasma, possesses its own set of adapting proteins [482]. The polarity of the subcellular organisation means that protein compartmentalisation and function are tightly knit. Despite this and decades of effort to unravel the distribution of parasite proteins – the spatial proteome remains poorly characterised [483].

Recent advances in genome-wide tools have allowed the screening and testing of the importance of Toxoplasma's extensive protein repertoire [247, 258, 53, 421, 422, 469]. However, these datasets lack a spatial context for superior interpretation. In this chapter, we explore the application of hyperLOPIT and Bayesian modelling to characterise the spatial proteome of *Toxoplasma gondii*. We provide unrivalled insight into the organisation of an apicomplexan cell and derive cellular atlases of genomic features.

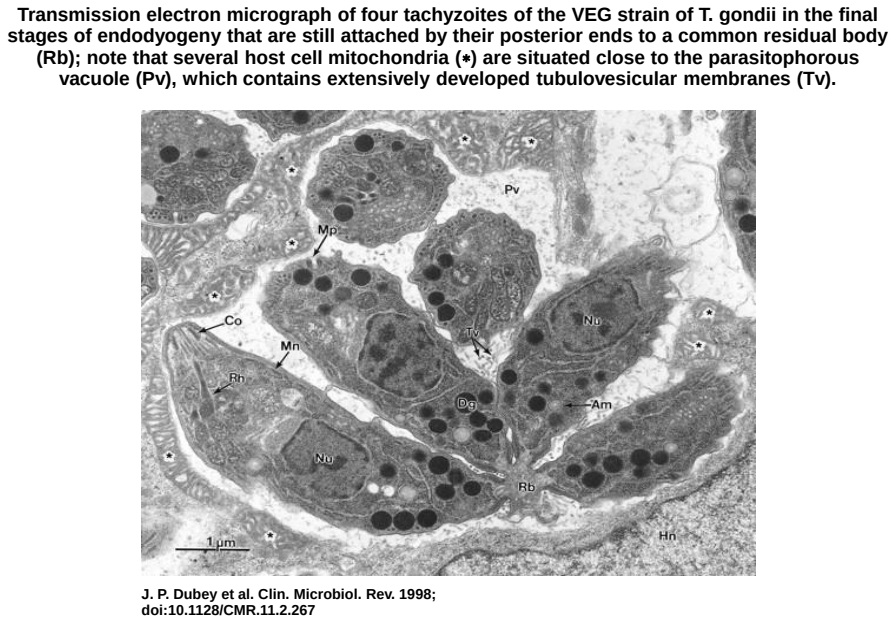


Fig. 4.4 Transmission electron micrograph of four tachyzoites of Toxoplasma in the final stages of endodyogeny that are attached by their posterior ends to a common residual body (Rb); note that several host cell mitochondria (*) are situated close to the parasitophorous vacuole (PV), which contains extensively developed tubulovesicular membranes (TV). Am, amylopectin granule; Co, conoid; Dg, electron-dense granule; Hn, host cell nucleus; Mn, microneme; Mp, micropore; Nu, nucleus; Rh, rhoptry. Figure taken from Dubey et al. [116].

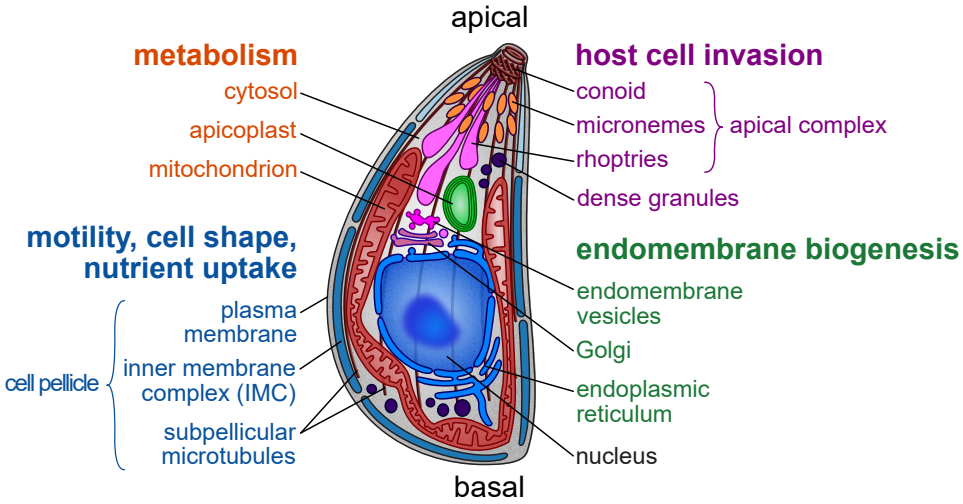


Fig. 4.5 Tachyzoite of *Toxoplasma gondii*, where the compartments are divided up with accordance to their function. Image used by permission of KB.

4.3 Methods and datasets

This section describes the processes by which the data where generated and auxiliary datasets that augment our findings.

4.3.1 Adapting the hyperLOPIT protocol

The hyperLOPIT protocol [68, 324] was adapted for spatial proteomics of *Toxoplasma gondii* tachyzoites (see figure 4.6). Tachyzoites were cultured by serial passaging in a culture of human foreskin fibroblasts. The tachyzoites boasts a cell pellicle that cannot be disrupted by hypotonic lysis. Optimisation of cell disruption, with validation by western blotting, identified nitrogen cavitation [472] as an effective means of cell lysis. 10 billion cells were pressurised to 2,000 psi then cavitated in two cycles. Poorly dispersed cell material was returned to a subsequent cavitation cycle via differential centrifugation. The membranous compartments were then separated from soluble cytosolic material using ultracentrifugation with density-barriers of 6% and 25% iodixanol. A linear gradient was then used to fractionate the remaining material (see figure 4.6). Fractions were sampled along these gradients and peptides were labelled with TMT10plex isobaric tags. Quantification was then performed across all fractions using LC-SPS-MS^3.

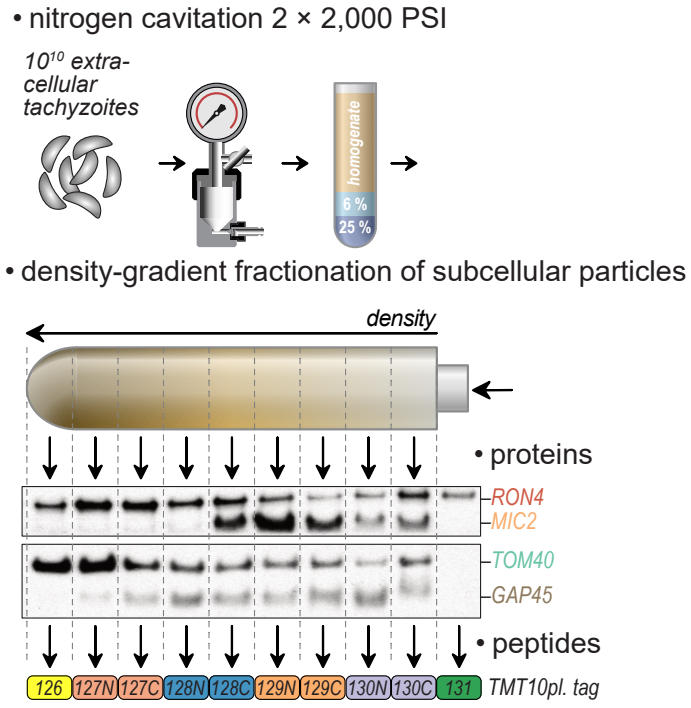


Fig. 4.6 A schematic overview of the ToxoLOPIT protocol. Edited from the original image by permission of KB.

4.3.2 Genomic features

Selection Pressure

Previous genomic data has been collected for 62 geographical isolates of *Toxoplasma gondii* [281]. We wish to analyse single nucleotide polymorphism (SNP) properties across the subcellular compartments. The rate of non-synonymous mutation denoted d_N, which is a nucleotide mutation that results in the amino acid sequence of a protein changing. This contrasts with a synonymous mutation where a base substitution does not lead to change to the amino acid sequence. The rate of synonymous mutations is denoted as d_S. We can compare the ratio d_N/d_S at a particular locus, which provides information on the rate of evolution of that sequence. A gene where synonymous mutations outpace non-synonymous ones, and so $d_N/d_S < 1$, suggests the sequence is being constrained for coding a particular protein. In the case of an elevated ratio, there is positive selection for change.

Genetic Polymorphism

The distribution of SNPs is not homogeneous (uniform) along the genome and, for example, they occur more frequently in the non-coding regions of the genome. There are few SNPs in regions where natural selection is in action and the allele is being "fixed" (elimination of variants). SNP density is computed as the average number of SNPs in a 10kb sequence of the genome. SNP density in a protein coding region provides information on its evolution. The distribution of SNP density across a compartment highlights the evolutionary behaviour of proteins within that compartment.

Functional Redundancy

Genome-wide CRISPR-Cas9 knockout screens allow us to examine the relative redundancy of the proteome across the subcellular landscape [421]. The CRISPR-Cas9 knockout screen works by introducing targeted loss-of-function mutations at specific sites in the genome [414]. Cas9 can be designed to induce DNA double strand breaks at desired genomic loci using a synthetic single guide RNA (sgRNA). When this sgRNA is targeted to coding regions of genes they can create indel mutations, resulting in a frame shift and thus loss-of-function of the allele. Frequently, a phenotype score is reported which is the $\log_2$ fold change for each sgRNA averaged across the top 5 scoring guides. The mean phenotype score is reported across four replicates. Higher scores indicate higher fitness conferring genes.

4.3.3 Validation by gene editing

A CRISPR-Cas9 genomic tagging strategy was used to perform endogenous gene tagging with epitope tags for protein localisation in Toxoplasma. A generic strategy for C-terminal tagging proceeds as follows. Cells are co-transfected with (1) a plasmid (a circular autonomous extrachromosomal DNA molecule) expressing Cas9 endonuclease and a specific gRNA, and (2) a PCR reaction product containing the tag of interest (an epitope or a fluorescent protein) followed by a terminator, a spacer, and a drug resistance cassette (chloramphenicol acetyltransferase, CAT, or dehydrofolate reductase, DHFR). The gRNA sequence is designed to direct Cas9 at a specific site of the gene where it introduces a double-strand DNA break. The tagging construct is flanked by short sequences homologous to the target gene sequence to direct specific integration into the genomic locus via homologous recombination. C-terminal tagging is preferred to avoid disruption of any possible N-terminal signal sequence. Furthermore, as a protein is translated N to C-terminus, if a tag is introduced early in the sequence it is more likely to disrupt protein synthesis. We refer to Barylyuk et al. [21] for precise details.

4.4 Results

4.4.1 Mapping the spatial proteome of *Toxoplasama gondii*

HyperLOPIT was applied to three independent experiments with minor alterations to the cell rupturing and to the density gradients. Each experiment quantified roughly $4,100$ proteins with quantitative measurements for all 10 fractions. $3,832$ proteins (46% of the total proteome) were common to all three datasets providing full profile information for 30 fractions. As a first visualisation of the data, we used t-distributed stochastic neighbour embedding (t-SNE) (see figure 4.7). t-SNE projections reveal the distinct clustering of proteins according to different sub-cellular niches. As a verification of the observed clusters, we applied unsupervised clustering (HDBSCAN) to the untransformed data. A collection of 656 marker proteins were compiled from the literature from previous studies of proteins with unambiguous subcellular localisation (see figure 4.8). Mapping these proteins onto the t-SNE projections reveal that the clustering is according to sub-cellular niche. These clusters represent a total of 23 known apicomplexan compartments and suggest that application of hyperLOPIT affords exquisite resolution of the Toxoplasma tachyzoite.

We observe that we can resolve the major subcellular components of Toxoplasma. Thus, allowing us a substantial insight into the spatial distribution of the proteome and how genomic features may distribute across the different organelles. First, to examine the quality of the clusters, 62 proteins associated with the clusters were epitope-tagged by endogenous gene fusion and immunofluorescnece microscopy was used to determine the protein localisation (see methods).

The 62 proteins were chosen so that they were either completely uncharacterised by previous studies or their current functional annotation was in-conflict with the localisation inferred here. All tested proteins showed subcellular localisation concordant with their hyperLOPIT derived localisation (see figures 4.9, 4.10, 4.11, and 4.12).

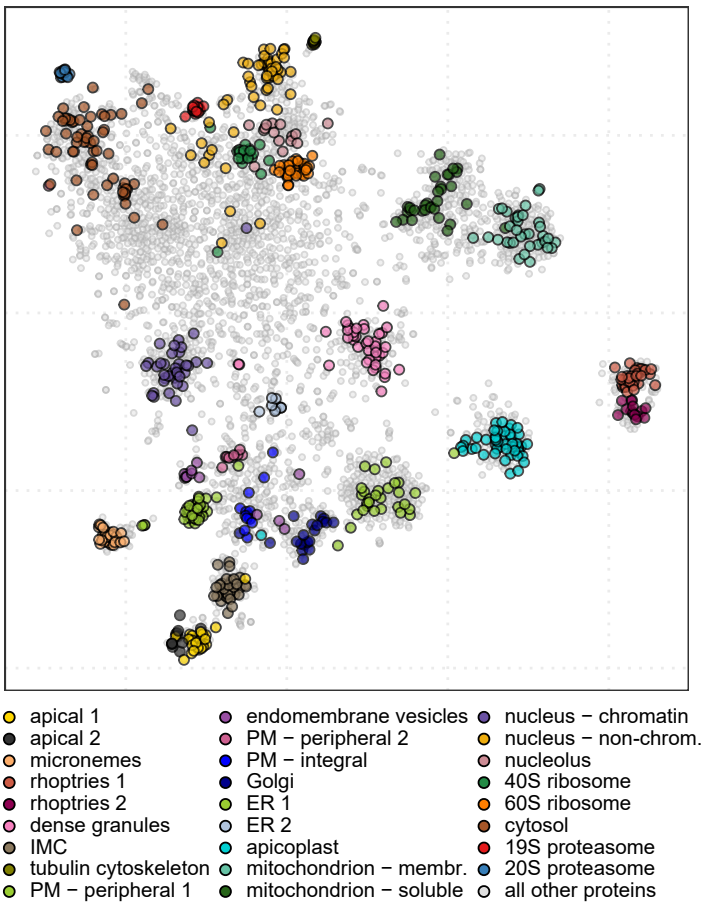


Fig. 4.7 *t*-SNE projection of the hyperLOPIT data. Clustering is observed according to different subcellular components. Edited from the original image by permission of KB.

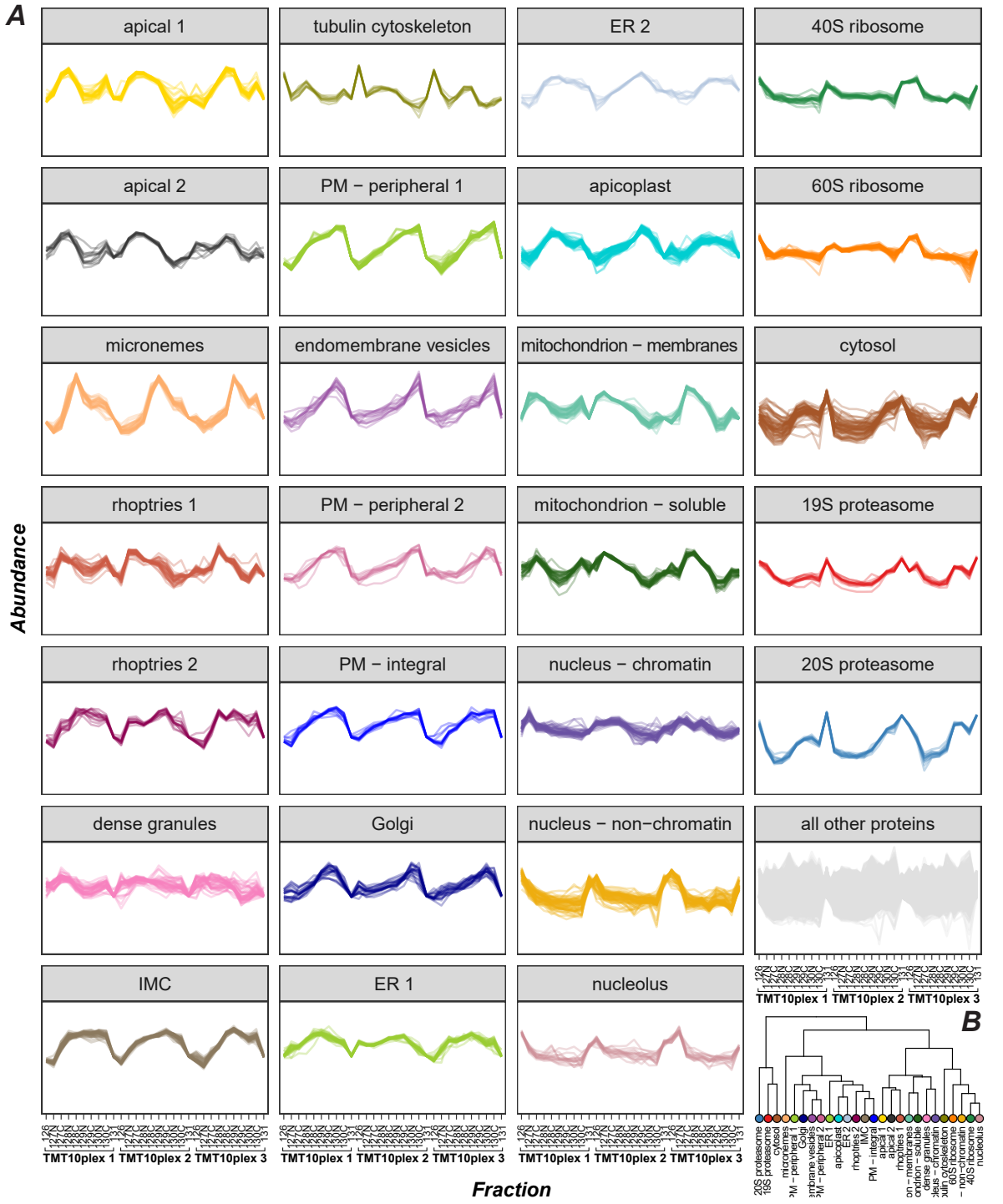


Fig. 4.8 (A) Marker profiles of each subcellular niche, demonstrating characteristic abundance profiles. (B) Hierarchical clustering of the subcellular niches. Edited from the original image by permission of KB.

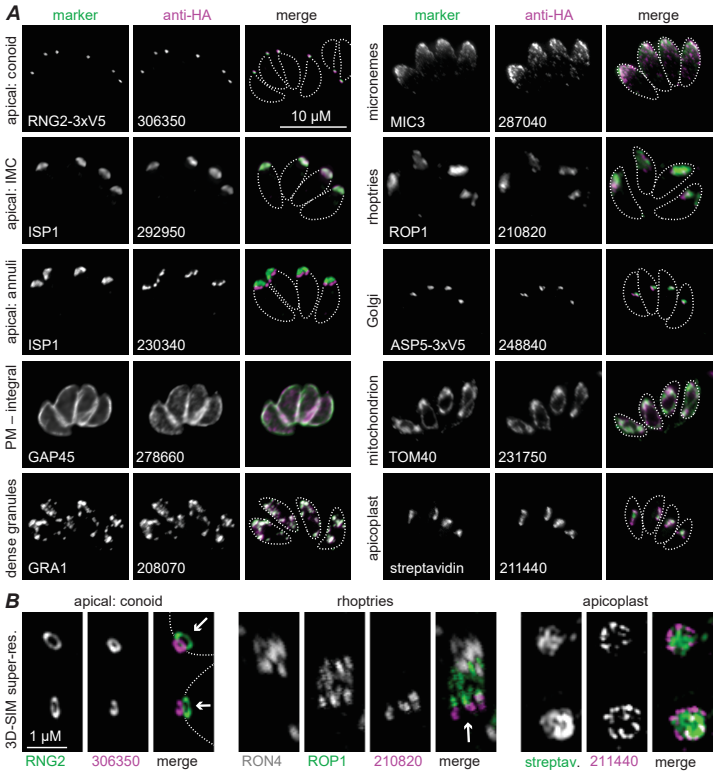


Fig. 4.9 Validated subcellular localisations across the array of subcellular compartments. Edited from the original image by permission of KB.

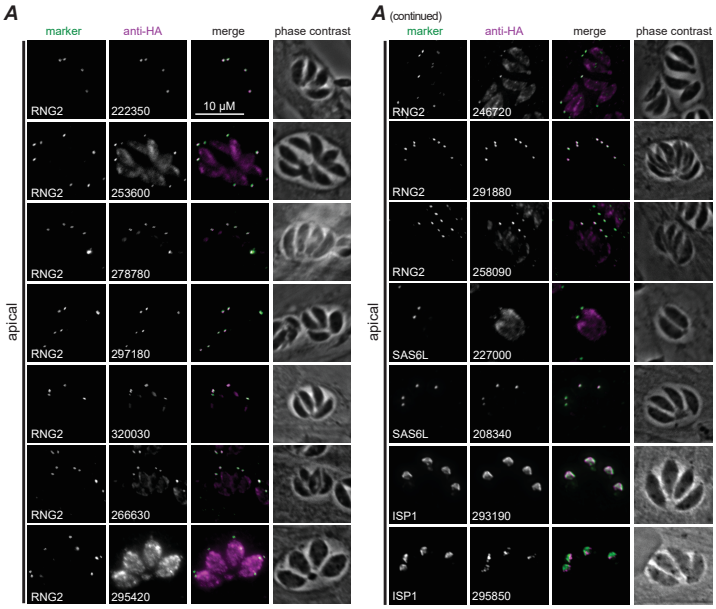


Fig. 4.10 Additional validated localisations of proteins predicted to the apical comeplex (A). Edited from the original image by permission of KB.

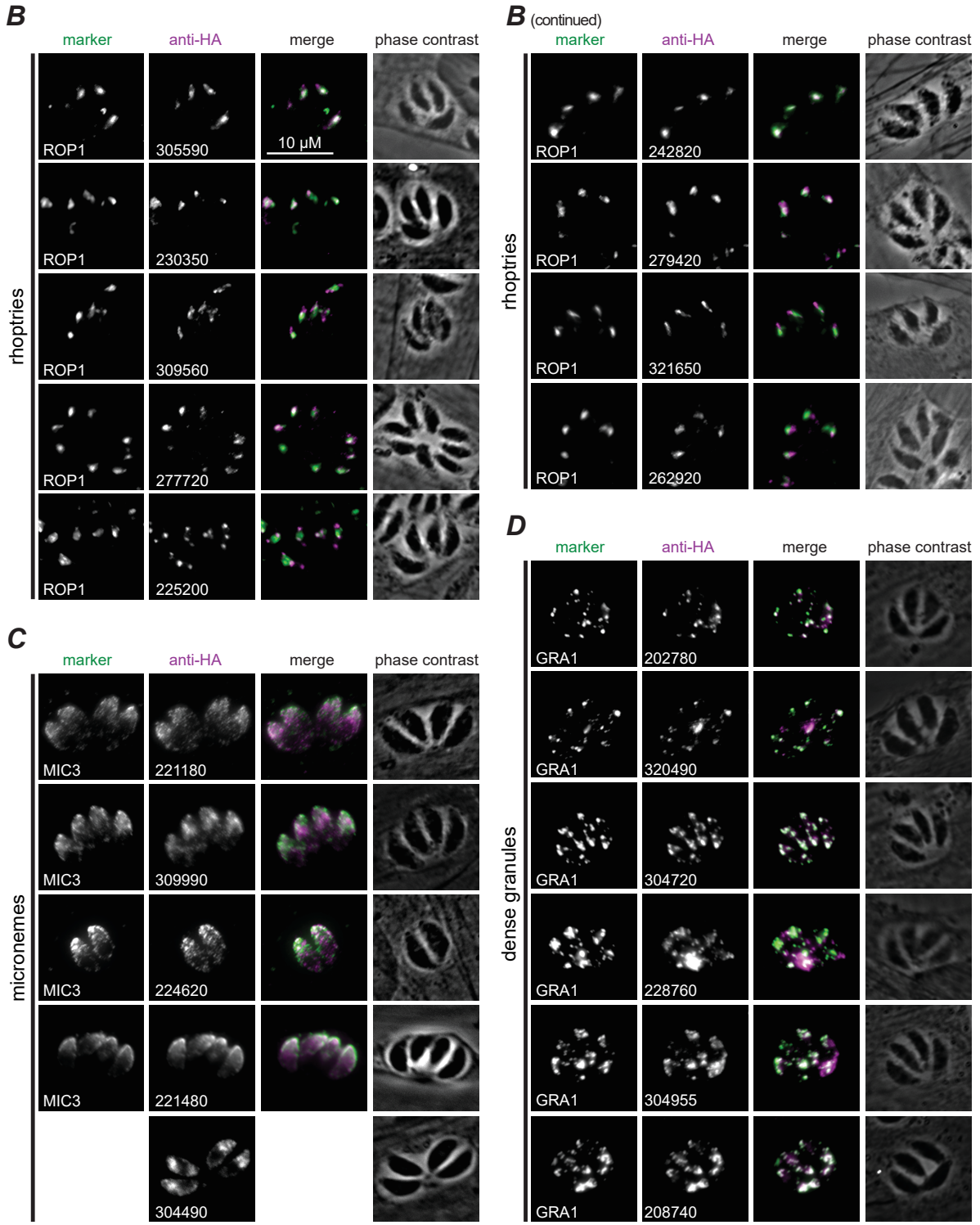


Fig. 4.11 Continued. Validation of proteins with predicted localisation to (B) the rhoptries, (C) micronemes (D) dense granules. Edited from the original image by permission of KB.

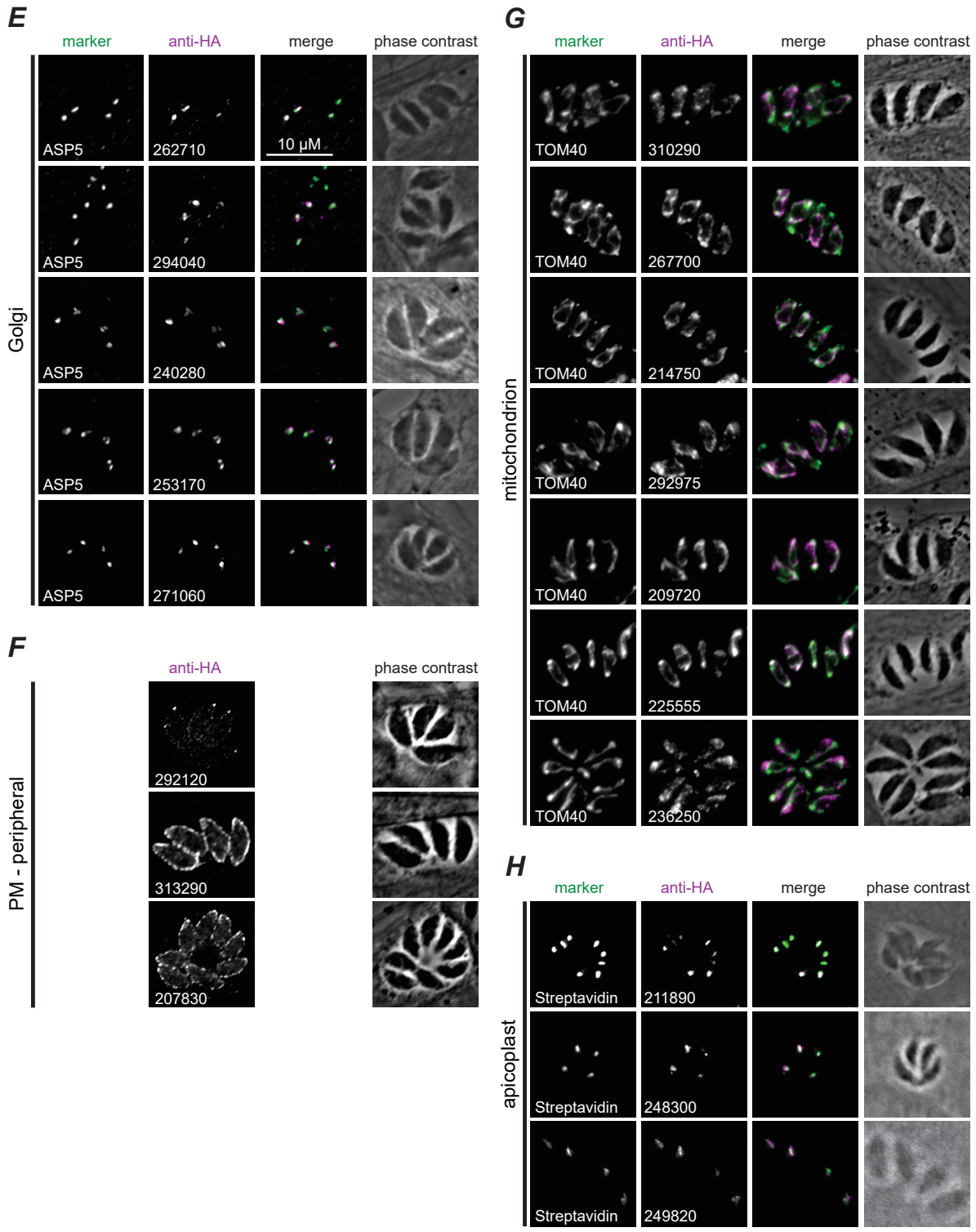


Fig. 4.12 Continued. Validation of proteins with predicted localisation to (E) Golgi, (F) plasma membrane peripheral (G) mitochondrion (H) apicoplast. Edited from the original image by permission of KB.

Having established the reliability of the data, the 62 validated proteins are added to the set of markers to bringing the total to 718 proteins. We now apply the methodology developed in the previous two chapters. We apply TAGM-MAP with default prior settings to compute the posterior localisation probability that each protein belongs to each class. This probability was obtained for the 3,114 remaining proteins. We obtain the probability of every protein belonging to the respective subcellular class and not being an outlier protein. Proteins are considered confident allocations if their localisation probability exceeds 0.99 (see figure 4.13 a). As a result, $1,916$ proteins with previously unknown localisation were allocated to one of 26 subcellular niches (see figure 4.13 c). Roughly 30% of protein are not assigned to any single location.

As previously discussed, determination of a protein to a single localisation overlooks the dynamic behaviour of proteins. A protein may not be well described by any single localisation because it has a dynamic localisation, continually trafficks between different compartments, is localised to an interface between two organelles or exists in genuine multiple pools in different locations. Thus, we wish to quantify the uncertainty in the localisation probabilities using our TAGM-MCMC method. TAGM-MCMC was applied as described in the previous chapter, with some minor modifications. The algorithm was run for $25,000$ iterations for 9 chains in parallel. Then $10,000$ iterations were discarded for burn-in and the chains were thinned by retaining every 20^{th} protein. We discarded 5 chains by visual inspection and assessed convergence using the remaining chains. Using the Gelman-Rubin diagnostic, an $\hat{R} \approx 1.02$ was computed. We concluded our chains were sufficiently well mixed and continued with downstream analysis, pooling the samples from the converged chains.

Most high confidence proteins assigned by TAGM-MAP and TAGM-MCMC are concordant (see figure 4.13 b). However, some proteins display probability distributions of their posterior localisation probability consistent with behaviour across multiple compartments. For example, proteins of the integral plasma membrane, Golgi and endomembrane vesicles show shared probability in some cases (see figure 4.13 b), whilst proteins of the secretory organelles; such as the rhoptries, micronemes and dense granules appear to have well defined single localisations. These results are in agreement with the interpretation that there is significant exchange of proteins between the plasma membrane, Golgi and vesicles, whilst the proteomes of the secretory organelles are mostly static in *Toxoplasma gondii*.

The observed uncertainty in different subcellular localisation follows closely the overlapping of the markers for different subcellular niches. Figure 4.14 shows the marker distributions for different components and, whilst it is clear that niches form distinct clusters, there is significant overlap between some niches. This is typical for subcellular niches that share confounding biochemical properties and cannot be completely separated using subcellular fractionation methods. Improved separation could be achieved with more fractions or an orthogonal method to tease apart these proteins. Furthermore, we observe that the marker proteins for the cytosol

and the nuclear compartments are more diffuse than for other organelles. There are multiple explanations for this observation. The first is that nuclear rupturing means that nuclear proteins are more dispersed along the density gradient. In addition, cytosolic and nuclear proteins have a large dynamic range of abundances. Thus quantitative accuracy for these proteins can differ considerably, resulting in greater variance.

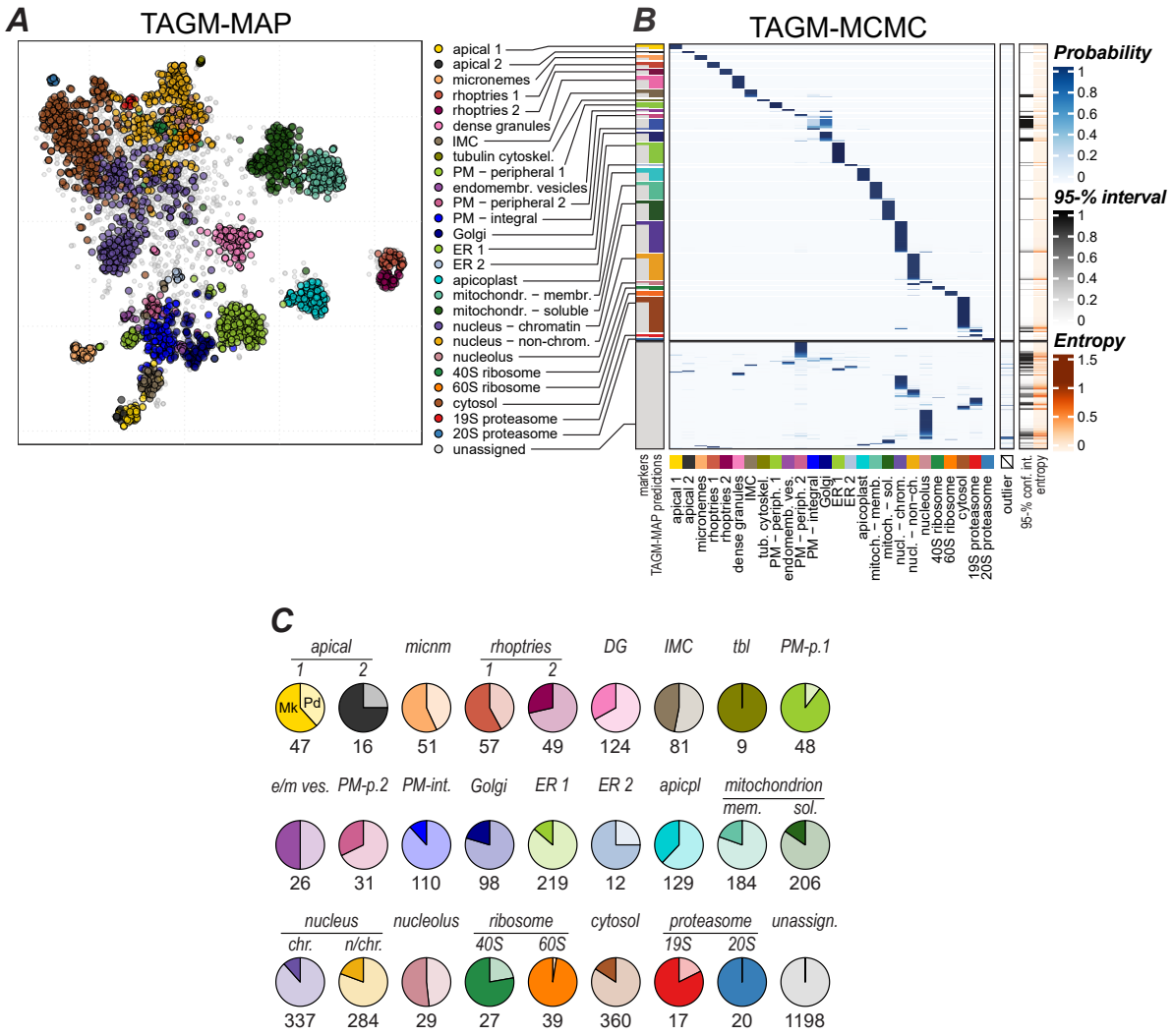
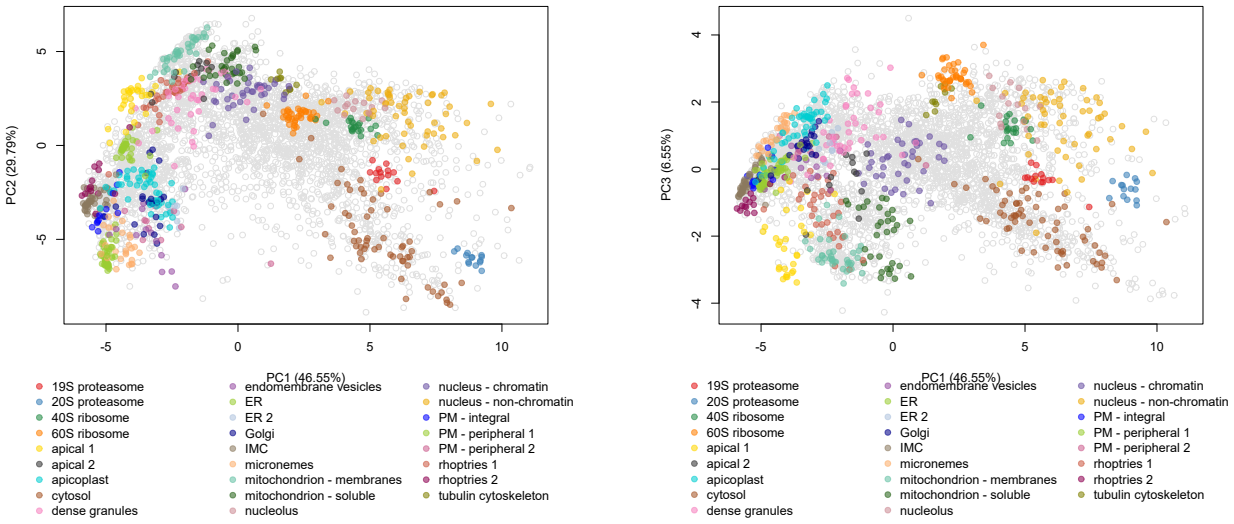
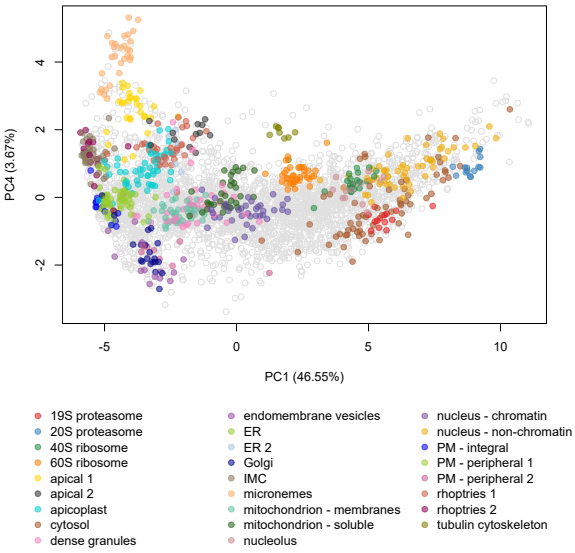


Fig. 4.13 (A) *t*-SNE projection with high confidence TAGM-MAP proteins coloured according to the most probable subcellular niche (b) statistics from TAGM-MCMC displayed as a heatmap, with comparisons to TAGM-MAP on the left colour bar (c) Knowledge expansion for each subcellular niche, showing markers and number of new predicted subcellular locations. Edited from the original image by permission of KB.



(a) (b)



(c)

Fig. 4.14 PCA plots along different PC dimensions. The PCA plots demonstrate a high degree of resolution between the different subcellular niches. Though the secretory and parasitic niches overlap more severely than the components of core cellular machinery. The cytosol and nuclear components are more diffuse than other organelles.

Delving deeper into the characterisation of the uncertainty of proteins' localisation, we project the posterior localisation probabilities into PCA coordinates (see figure 4.15). Figure 4.15 demonstrates a clear spatial pattern of localisation probabilities, ranging from tightly formed clusters to diffuse patterns. This observation adds to the mounting evidence that some subcellular niches are challenging to separate. However, with uncertainty quantification we can, at least, obtain deeper insights - even if the statements are not definitive.

The posterior distribution of the localisation probabilities for those proteins with uncertain localisations can be visualised in a violin plot. These plots allows us to discern in which localisations the proteins are most likely to reside or, potentially, jointly reside. For example 3 uncharacterised toxoplasma proteins are plotted in figure 4.16. These previously hypothetical proteins now have a suggested localisation, despite no precise localisation. One reason that these proteins are challenging to localise is that the Golgi is a particularly transient organelle in terms of protein content [68]. Another, perhaps appealing, interpretation is that these proteins are partially distributed across these organelles. This could either been through genuine multilocalisation, dynamic localisation between the two compartments or the protein being resident in different locations across the cellular population.

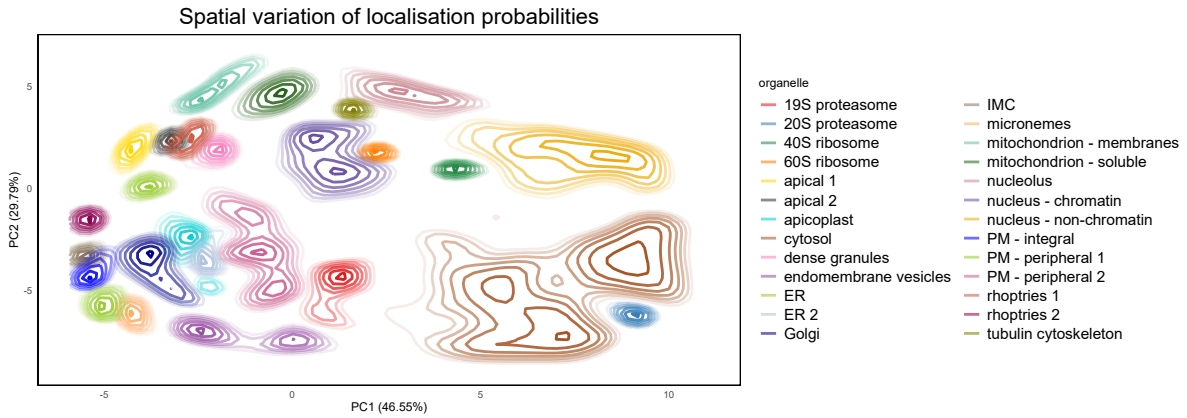


Fig. 4.15 A PCA plot with the posterior localisation probabilities projected as contours (see previous chapter). Subcellular niche display clear separation with contours mostly overlapping for the secretory organelles. The inner contour represent (approximately) a localisation probability of 0.99. The subsequent contour is approximately 0.95 with decreasing increments of 0.05.

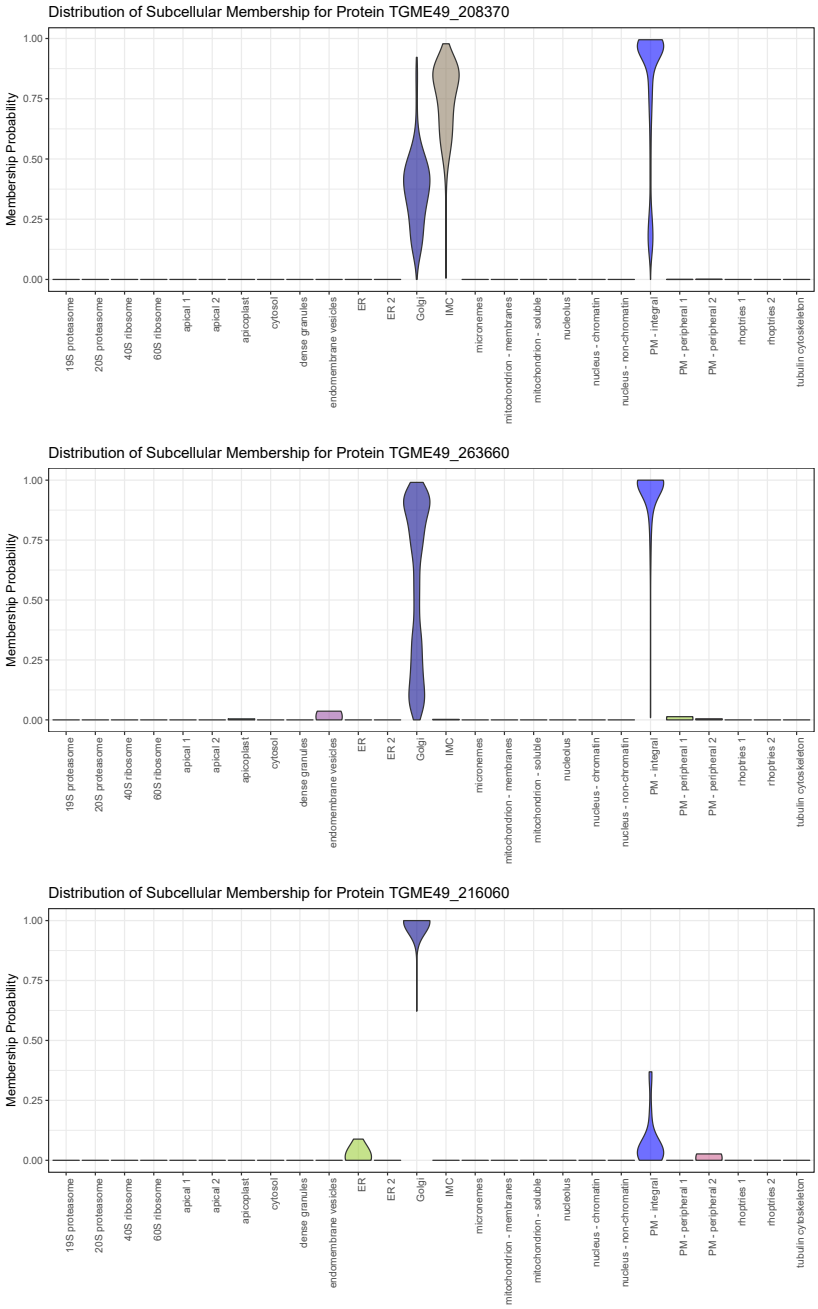


Fig. 4.16 Violin plots of the posterior distribution of localisation probabilities for 3 uncharacterised proteins from *Toxoplasma gondii*. Each of these proteins display uncertain localisation between the Golgi and integral plasma membrane suggesting these proteins are perhaps recycling between these localisations.

4.4.2 HyperLOPIT provides extensive characterisation the subcellular proteome of Toxoplasma

The hyperLOPIT experiments have characterised 26 compartments and sub-compartments of Toxoplasma. Many of the organelles are clearly defined, where class boundaries are well separated form those of neighbouring subcellular niches. This is particularly the case for membrane-bound organelles, such as the mitochondrion, apicoplasts and rhoptries. Though there is also some suggestion of possible disruption of the membranes of these organelles, given that the mitochondrion is associated with two clusters: one enriched for integral membrane proteins and the other is depleted in proteins which are anchored to the membrane.

The inner membrane complex (IMC) is a unique organelle to apicomplexans. The IMC is a crucial element of the invasion machinery of Toxoplasma, allowing the tachyzoite to maintain its distinct crescent shape and subcellular organisation. The separated IMC cluster from the plasma membrane suggest that they have disassociated during cell lysis.

Proteins belonging to the apical region of the parasite are resolved from the IMC, and include proteins involved with the conoid and apical polar ring. These two invasion associated niches are located at the extreme of the cell. The apical proteins resolved as two separate clusters; however, there does not appear to be difference in these clusters based purely on spatial organisation. Rather it appears that these two apical associated niches are enriched for different biochemical properties. The first appears to be enriched in proteins with basic pI (isoelectric point); the second acidic pI. That is to say the two clusters represent the same subcellular niche within the cell but cluster separately as a result of the hyperLOPIT protocol.

Proteins of other classes appear to display sub-organellar resolution with the plasma membrane dividing into two peripheral associated clusters and a plasma membrane cluster enriched for external facing proteins. The ER also separates into two distinct clusters, again one enriched for integral membrane proteins and the other the other soluble proteins. The data also appear to resolve large protein complexes such as the ribosomes and proteosomes, as well as proteins of the cytosol and the numerous sub-niches of the nucleus.

4.4.3 Resolution of subcellular proteomes constitutes massive knowledge expansion

The hyperLOPIT experiments and subsequent Bayesian modelling have assigned a total of $1,916$ proteins to one of the 26 compartments. Of these proteins 795 were annotated simply as “hypothetical” prior to these experiments; 335 where the only annotation was a conserved domain; 256 proteins with generic functional annotations such a “transporter”; 228 where a putative functional annotation was posited. As a result, only 302 of these proteins had some certainty to their functional annotation. Though, despite having functional annotations,

many of these proteins do not have validated subcellular localisations. Dozens to hundreds of proteins are allocated to particular niches, including the organelles that are necessary for parasite function. This section proceeds to annotate the spatial proteome with genomic features to characterise the heterogeneity across the subcellular niches.

Relative redundancy of subproteomes

As described in the methods, we extracted data from a genome-wide CRISPR-Cas9 knockout screen in *T. gondii* [421]. Integration of this screen with its spatial context, provided by the hyperLOPIT datasets, allows us to explore the heterogeneity of dispensable and indispensable protein across subcellular landscape. To perform this analysis, we performed a permutation test by randomly permuting the class labels and computing the mean phenotype score for each niche. Repeating this process 10^6 times allows us to approximate the null distribution of class means. We then ranked the observed values amongst instances from the null distribution, computed approximate p-values, and corrected for multiplicity [93, 363, 25]. There is clearly an uneven spread across the different compartments (see figure 4.17 A). The plasma membrane, dense granules, micronemes and rhoptries and the IMC show a bias towards dispensable proteins $p < 0.01$, indicating functional redundancy. These subproteomes are thus not part of the typical parasitic evolutionary trend for austerity. These niches are responsible for host invasion and thus high overturn is required for continued species prosperity. Meanwhile, the apicoplast, a remnant of a former photosynthetic being, shows a dearth of dispensable proteins $p < 0.01$. Thus, the interpretation that the apicoplast is evolutionary baggage is unsupported and it has become a minimalistic, essential organelle. Unsurprisingly, the proteomes of the basic cell machinery, such as the mitochondrion and ribosome are indispensable because they are necessary for fundamental cell biochemistry.

Selective pressure of the host-adaptive response

The host immune system is constantly at watch for host invasion. Furthermore, *T. gondii* can exploit a variety of warm-blooded hosts suggesting highly successful adaptation. However, such successes and the constant bombardment from the host immune system comes at the cost of huge selective pressure. The magnitude and direction of selection pressure on a protein is characterised by the ratio d_N/d_S, as described in the methods (section 4.3.2). The distribution of d_N/d_S values across the subcellular landscape provide insight into these pressures for each subproteome. Positively-skewed d_N/d_S distributions are those of the plasma membrane, the soluble content of the rhoptries (rhoptries 1) and the dense granules $p < 0.01$. This observed distribution, exemplified in figure 4.17 B, implies a request and tolerance for change in these niches. Given that these niches are the war-zone of the host-pathogen interaction, this adaptation reflects a desire to outpace the host. However, for the integral plasma membrane there is a bias

for purification (against change) $p < 0.01$. This suggest an antagonism between maintaining correct and proper function of the plasma membrane and the exposure of these proteins to the host immune system.

Whilst SNP density (per 10K of coding sequence) is correlated with d_N/d_S and also reports on subproteome evolution, there are some unexpected behaviours. Soluble mitochondrial proteins show enrichment for higher than average SNP densities $p < 0.01$, but no corresponding increase in d_N/d_S (see figure 4.17 B and C). The interpretation here is somewhat challenging, but the enrichment of silent mutations here could suggest implications for metabolic flux control. Thus, modulation of the mitochondrial metabolic processes might be a secondary driver of host-adaptive response. Unsurprisingly, the peripheral plasma membrane proteins stand out amongst the niches with elevated levels of SNP density. Constant selective pressure of this niche has resulted in a high level of redundancy and pressure to adapt. This has strong implications for drug targets as our analysis suggests targeting these proteins would be of little avail.

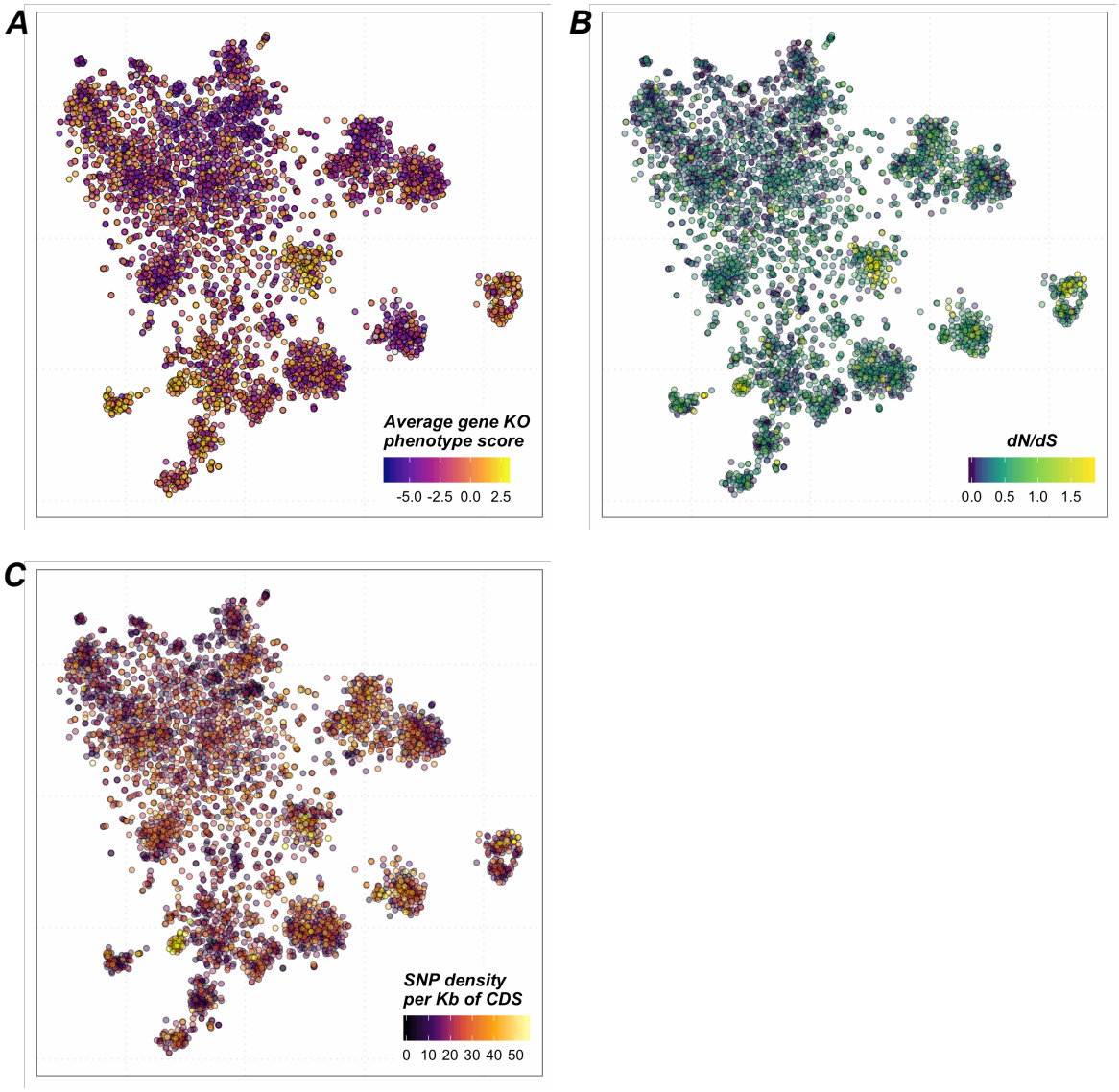


Fig. 4.17 Genomics features of the proteins are displayed in the original t-SNE coordinates of the spatial proteomics data. These quantities clearly have spatial context. Figure used with permission of KB

4.5 Discussion and limitations

This chapter has explored the application of hyperLOPIT and Bayesian analysis to *Toxoplasma gondii* extracellular tachyzoites. This is the first comprehensive and detailed spatial proteome of an apicomplexan cell. Overall we have identified thousands of proteins allocated with high posterior probability to 26 possible subcellular niches. A large proportion of these proteins had essentially no annotation prior to these experiments and analysis. Thus, we have significantly expanded the knowledge of the Toxoplasma proteome.

HyperLOPIT experiments are independent of functional or localisation prediction tools that are derived from sequence motifs or orthologues in model organisms. The approaches here demonstrate that hyperLOPIT is applicable to non model organisms and can provide insight into a significant proportion of the spatial proteome. Furthermore, Toxoplasma has very little prior knowledge and we have highlighted that Bayesian analysis can alleviate many of the challenges by quantifying uncertainty.

There still remain proteins that are not allocated to any particular class. TAGM is unable to model subcellular niches without annotation or those that have insufficient number of proteins to provide a reasonable set of markers. These proteins are usually classified with low probability to one of the classes or as an outlier protein. Furthermore, it is not just the computational approach that has limitations; the experiment and mass-spectrometry are also not perfect. Firstly, for proteins of low abundance MS-based quantitation becomes more challenging and less accurate resulting in distorted abundance profiles. Secondly, perhaps more importantly, hyperLOPIT reports on the steady-state localisation of proteins. Thus proteins that are being constantly recycled between two subcellular niches or in constant dynamic transit between two or more organelles will have composite (not necessarily mixed profiles). Uncertainty quantification can go some way to providing a lens on these proteins, but since there are multiple sources of uncertainty the reasons for this uncertainty will always be a point of interpretation.

HyperLOPIT provides a platform for molecular screens to be placed in their spatial context. Indeed, we have demonstrated that several important genomic features can be mapped onto the spatial data and provide additional insights into apicomplexan function. Future systems-wide studies of Toxoplasma will be able to map their data on the subcellular landscape we have provided and they can examine their data within the context it deserves.

Looking forward, it is clear that hyperLOPIT can be applied to non-model organisms to provide exquisite insights. The natural directions are to apply it to even more poorly annotated organisms, different stages of parasite life cycles and to host cells infected with parasites. Each of these tasks will require new computational tools - some of which are beyond the scope of this thesis. For the proteomes of organisms with poor annotation, whilst there might not be sufficient markers present to perform a supervised analysis, the proteins will still cluster

according to their biochemical fractionation properties. Our Bayesian model TAGM can be extended to allow additional niches to be discovered and this is the content of the next chapter.

Chapter 5

A semi-supervised Bayesian approach for simultaneous protein subcellular localisation and novelty detection

5.1 Motivation

Following on from the previous chapter, we noted that one of the limitations of our Bayesian modelling, thus far, is that it is reliant on marker proteins. For well annotated organisms, such as when working with human cell lines, this is not a substantial limitation. However, as we highlighted in the previous chapter there is desire to apply spatial proteomics methods to non-model organisms. This chapter presents an extension to our Bayesian model to allow additional unannotated subcellular niches to be inferred. This work is an edited version of Crook et al. [81] and there is significant textual overlap.

5.1.1 Abstract

The cell is compartmentalised into complex micro-environments allowing an array of specialised biological processes to be carried out in synchrony. Determining a protein's sub-cellular localisation to one or more of these compartments can therefore be a first step in determining its function. High-throughput and high-accuracy mass spectrometry-based sub-cellular proteomic methods can now shed light on the localisation of thousands of proteins at once. Machine learning algorithms are then typically employed to make protein-organelle assignments. However, these algorithms are limited by insufficient and incomplete annotation. We propose a semi-supervised Bayesian approach to novelty detection, allowing the discovery of additional, previously

unannotated sub-cellular niches. Inference in our model is performed in a Bayesian framework, allowing us to quantify uncertainty in the allocation of proteins to new sub-cellular niches, as well as in the number of newly discovered compartments. In this chapter, we apply our approach across 10 mass spectrometry based spatial proteomic datasets, representing a diverse range of experimental protocols. Application of our approach to *hyper*LOPIT datasets validates its utility by recovering enrichment with chromatin-associated proteins without annotation and uncovers sub-nuclear compartmentalisation which was not identified in the original analysis. Moreover, using sub-cellular proteomics data from *Saccharomyces cerevisiae*, we uncover a novel group of proteins trafficking from the ER to the early Golgi apparatus. Overall, we demonstrate the potential for novelty detection to yield biologically relevant niches that are missed by current approaches.

5.2 Introduction and literature review

In previous chapters, we have demonstrated the importance of characterising the sub-cellular localisation of proteins. Proteins are compartmentalised into sub-cellular niches, including organelles, sub-cellular structures, liquid phase droplets and protein complexes. For some organisms, such as apicomplexans this compartmentalisation can be highly polarised [21]. These compartments ensure that the biochemical conditions for proteins to function correctly are met, and that they are in the proximity of interaction partners [171].

As a brief reminder, a common approach to map the global sub-cellular localisation of proteins is to couple gentle cell lysis with high-accuracy mass spectrometry (MS) [68, 324, 159, 351]. These methods are designed to yield fractions differentially enriched in the sub-cellular compartments rather than purifying the compartments into individual fractions. As such, these spatial proteomics approaches aim to interrogate the greatest number of sub-cellular niches possible by relying upon rigorous data analysis and interpretation [154, 155].

Current computational approaches in MS-based spatial proteomics utilise machine learning algorithms to make protein-organelle assignments (see [155] for an overview). Within this framework, novelty detection, the process of identifying differences between testing and training data, has multiple benefits. For model organisms with well annotated proteomes, novelty detection can potentially uncover groups of proteins with shared sub-cellular niches not described by the training data. Novelty detection can also prove useful in validating experimental design, either by demonstrating that contaminants have been removed or that increased resolution of organelle classes has been achieved by the experimental approach. As we saw from the previous chapter for most non-model organisms, we have little *a priori* knowledge of their sub-cellular proteome organisation. This makes it challenging to curate the marker set (training dataset) from the literature [21]. In these cases, novelty detection can assist in annotating the spatial

proteome. Crucially, if a dataset is insufficiently annotated, i.e. sub-cellular niches detectable in the experimental data are missing from the marker set, then this leads to the classifier making erroneous assignments, resulting in inflated *false discovery rate* (FDR) and uncertainty estimates (where available). Thus, novelty detection is a useful feature for any classifier, even if novel niche detection is not a primary aim.

Previous efforts to discover novel niches within existing sub-cellular proteomics datasets have proved valuable. Breckels et al. [43] presented a phenotype discovery algorithm called *phenoDisco* to detect novel sub-cellular niches and alleviate the issue of undiscovered phenotypes. The algorithm uses an iterative procedure and the *Bayesian Information Criterion* (BIC) [411] is employed to determine the number of newly detected phenotypes. Afterwards, the dataset can be re-annotated and a classifier employed to assign proteins to organelles, including those that have been newly detected. Breckels et al. [43] applied their method on several datasets and discovered new organelle classes in *Arabidopsis* [119] and *Drosophila* [448]. This approach later successfully identified the trans-Golgi network (TGN) in *Arabidopsis* roots [181].

This thesis, thus far, has demonstrated the importance of uncertainty quantification in spatial proteomics studies. In chapter 2 and 3, we proposed a generative classification model and took a Bayesian approach to spatial proteomics data analysis by computing probability distributions of protein-organelle assignments using Markov-chain Monte-Carlo (MCMC). These probabilities were then used as the basis for organelle allocations, as well as to quantify the uncertainty in these allocations. On the basis that some proteins cannot be well described by any of the annotated sub-cellular niches, a multivariate student's t-distribution was included in the model to enable outlier detection. The proposed T-Augmented Gaussian Mixture (TAGM) model was shown to achieve state-of-the-art predictive performance against other commonly used machine learning algorithms (see chapter 2). Furthermore, the model has been successfully applied to reveal unrivalled insight into the spatial organisation of *Toxoplasma gondii* (see chapter 4).

This chapter explores an extension to TAGM to allow simultaneous protein-organelle assignments and novelty detection. One assumption of the existing TAGM model is that the number of sub-cellular niches is known. Here, we design a novelty detection algorithm based on allowing an unknown number of additional sub-cellular niches, as well as quantifying uncertainty in this number.

Quantifying uncertainty in the number of clusters in a Bayesian mixture model is challenging and many approaches have been proposed in the literature (see for example Ferguson [131], Antoniak [11], Richardson and Green [380]). Here, we make use of asymptotic results in Bayesian analysis of mixture models [397]. The principle of *overfitted mixtures* allows us to specify a (possibly large) maximum number of clusters. As shown in Rousseau and Mengersen [397] these components empty if they are not supported by the data, allowing the number of

clusters to be inferred. Kirk et al. [243] previously made use of this approach in the Bayesian integrative modelling of multiple genomic datasets. In our application, some of the organelles may be annotated with known marker proteins and this places a lower bound on the number of sub-cellular niches. Bringing these ideas together results in a semi-supervised Bayesian approach, which we refer to as Novelty TAGM. Table 5.1 summarises the differences between the current available machine-learning methods for spatial proteomics.

In this chapter, we begin by reviewing a number of classical and Bayesian approaches to inferring or performing model selection on the number of clusters in a mixture model. This motivates our extension to TAGM to novelty detection which we refer to as Novelty TAGM. We apply Novelty TAGM to 10 spatial proteomic datasets across a diverse range of protocols, including *hyper*LOPIT [68, 324], LOPIT-DC [159], Dynamic Organellar Maps (DOM) [220] and spatial-temporal methods [24]. Application of Novelty TAGM to each dataset reveals additional biologically relevant compartments. Notably, we detect 4 sub-nuclear compartments in the U-2 OS *hyper*LOPIT dataset: the nucleolus, nucleoplasm, chromatin-associated, and the nuclear membrane. In addition, an endosomal compartment is robustly identified across *hyper*LOPIT and LOPIT-DC datasets. Finally, we also uncover collections of proteins with previously uncharacterised localisation patterns; for example, vesicle proteins trafficking from the ER to the early Golgi in *Saccharomyces cerevisiae*.

MS-based Spatial Proteomics Computational Methods for Prediction and Novelty Detection							
Method	Localisation prediction	Uncertainty in protein localisation	Outlier detection	Novelty detection	Uncertainty in number of novel phenotypes	Uncertainty in allocation to new phenotypes	Integrative
Supervised Machine Learning (as reviewed in [155])	✓	✗	✗	✗	✗	✗	✗
Correlation Profiling [135, 252]	✓	✗	✗	✗	✗	✗	✗
Transfer Learning [44]	✓	✗	✗	✗	✗	✗	✓
Mclust (as used in [351])	✗	✗	✓	✓	✗	✗	✗
PhenoDisco [43]	✗	✗	✓	✓	✗	✗	✗
TAGM [83]	✓	✓	✓	✗	✗	✗	✗
Novelty TAGM (This chapter)	✓	✓	✓	✓	✓	✓	✗

Table 5.1 Examples of computational methods for spatial proteomics datasets for prediction and novelty detection.

5.3 Methods

5.3.1 Previous methods

The perennial question of unsupervised clustering is how to choose, infer or deduce the number of clusters [137]. Indeed, research focusing on this question can, at least, be traced back to the middle of the last century [454] and advances are still being made [145]. Some approaches claim to avoid the choosing of the number of clusters, such as hierarchical clustering but frequently this is replaced with a more obscure question: where to cut the tree [473]? A full survey of the literature covering this topic is beyond the scope of this chapter and we focus on a few key ideas, especially in the context of mixture models.

Frequentist approaches

The most popular approach for selecting the number of clusters in a mixture model is using the BIC (see discussion in chapter 2). The preference for the BIC over the AIC, given below, is because the AIC tends to produce spurious clusters because the number of parameters in the model is not sufficiently penalised [6]:

$$\text{AIC} = 2m - 2\log p(x|\hat{\theta}, \mathcal{M}). \tag{5.1}$$

Variations on the theme of criteria, one may choose a number of different approaches, such as the integrated complete likelihood [32] or the singular BIC [114]. The singular BIC was developed because the dimensionality penalty on the BIC is too large [114]. Alternatively, one might consider a classic approach of using the likelihood ratio test statistic λ for a model with k clusters, $\mathcal{M}_k$, against a model with $k+1$ clusters, $\mathcal{M}_{k+1}$. Given that these models are nested one might hope to apply Wilk's theorem that $-2\log(\lambda)$ is chi-squared distributed under the null hypothesis. However, in mixture models it is typical that the variance component of the expanded model is essentially zero, which violates the regularity assumption of Wilk's theorem. This does not negate, however, the use of a likelihood ratio test. It simply implies that p-values should not be obtained from the asymptotic chi-squared distribution. One way to circumvent this issue to use a parametric bootstrap [300]. It is then simple to obtain the appropriate order statistic to approximate the required p-value.

Bayesian approaches

Typically, a Bayesian approach to selecting an appropriate number of components in the mixture model would compute the Bayes factors between a model with k components or k' components:

$$\text{BF}_{k,k'} = \frac{p(x|\mathcal{M}_{k'})}{p(x|\mathcal{M}_k)}. \tag{5.2}$$

The argument for the Bayes factor is seen through the lens of the marginal likelihood:

$$p(x|\mathcal{M}_k) = \int_{\theta_k} p(x|\theta_k, \mathcal{M}_k) p(\theta_k|\mathcal{M}_k) \, \mathrm{d}\theta_k, \tag{5.3}$$

because it automatically penalises more complex models via the prior probability $p(\theta_k|\mathcal{M}_k)$. In fact, one need not necessarily select one single best mixture model but average over them using Bayesian model averaging (see [85] in the mixture model context). However, for nested models the expanded model is only penalised at a polynomial rate [61]. The implication is that Bayesian models still tend to produce spurious clusters despite the automatic penalisation usually attributed as "Occam's razor". This has led some other to use other criteria such as the *pseudo*-marginal likelihood [471].

Another approach is to consider a single large number of components, say k^* and discard unoccupied components. Rousseau and Mengersen [397] demonstrate that if the prior on the components weights $p(\pi|\mathcal{M}_{k^*}) = \mathrm{Dir}(\pi; \beta)$ is set such that $\max_j \beta_j < d/2$, where d is the dimension of the data, then the posterior distribution of π concentrates at 0 for unnecessary components at rate $n^{-1/2}$. In essence, for some choices of β, which includes the Jeffrey prior $\beta = 0.5$, the spurious components "empty". The authors advocate for even stronger shrinkage $\beta \approx n^{-1}$, empirically [465]. More elaborate non-local priors have also been considered [145].

A seminal approach by Richardson and Green [380], in the context that k is still fixed but unknown, is using Bayesian inference to mix over k. Richardson and Green [380] use reversible-jump MCMC (RJMCMC), which allows mixture components to be split or combined at each iteration of the MCMC algorithm with a carefully computed acceptance probability. A prior is placed directly on k, for example from the Poisson family. The challenge with this approach is that the parameter dimension changes at every iteration, making inferences of important quantities challenging.

Bayesian non-parametric approaches

The Bayesian non-parametric approach to determining the number of clusters is different to a traditional Bayesian perspective. In a Bayesian parametric approach there is assumed to be a true fixed number of clusters and the posterior contracts onto this value as the number of observations grow. Meanwhile, in a Bayesian non-parametric setting, the model is allowed to expand as more observations are obtained. The Dirichlet process mixture model is, perhaps, the simplest non-parametric mixture [130, 11]. The Dirichlet process (DP) is the infinite dimensional extension of the Dirichlet distribution and has a single concentration parameter β.

A typical DP mixture model is specified as follows

$$\begin{aligned} G &\sim DP(\beta), \\ \theta_i|G &\sim G, \\ x_i|\theta_i &\sim F(\theta_i), \end{aligned} \tag{5.4}$$

for some parametric distribution F. The Dirichlet process inherits the conjugacy property from the Dirichlet distribution and a Polya urn scheme makes computations straightforward [33]. Having made $i-1$ observations, the clustering property of the Dirichlet process is that given a new observation x_i, it has prior probability $\beta/(\beta+i-1)$ of belonging to a new cluster and $n_k/(\beta+i-1)$ prior probability of belonging to cluster k. Here, n_k denotes the number of observations such that $\theta_i = \theta_k$. This is frequently referred to has the rich gets richer property of the DP. The distribution of the number of clusters K, clearly depends on the number of observations. Briefly, the expectation and variance are:

$$\begin{aligned} \mathbb{E}[K|n] &= \sum_{i=1}^{n} \frac{\beta}{\beta+i-1} \approx \beta \log(1+\frac{n}{\beta}) \\ \mathbb{V}[K|n] &= \sum_{i=1}^{n} \frac{\beta(i-1)}{(\beta+i-1)^2} \approx \beta \log(1+\frac{n}{\beta}). \end{aligned} \tag{5.5}$$

The results follow by simple application of the harmonic series (see Teh et al. [452]). Clearly, the concentration parameter β directly controls the number of clusters and the number of clusters clearly grow logarithmically in n. It might then be sensible to infer β, for example placing a gamma prior on β [125]. Some authors have argued that a single parameter to control mean and variance is overly restrictive and have proposed generalisations based on the marginals following a two parameter Poisson-Dirichlet distribution [364, 219]. Despite being a flexible approach, we do not consider the Bayesian non-parametric approach for this chapter for two reasons. Firstly, the increase in computation is burdensome and secondly, we do not believe spatial proteomics applications warrant that the number of clusters growing as more observations are added: we do not expect to observe new organelles as we increase the number of proteins measured.

phenoDisco

Breckels et al. [43] proposed *phenoDisco* to perform novelty detection in subcellular proteomics datasets. The approach is a semi-supervised extension of mclust [412], a finite frequentist mixture modelling approach that uses the BIC for model selection.

The method proceeds by first computing the first two principal components of the data. Then we select one of the k organelle classes at random. After that, we cluster the data for this class along with the unlabelled data using Gaussian mixture modelling (GMM) with the BIC

used to select the number of clusters. If proteins cluster with class k they are considered to be candidates of that class. Then for each candidate of class k perform the following steps. The data points belonging to class k are modelled themselves using a GMM and the log-likelihood is a obtained.

Samples are then drawn from this distribution and combined with the observed data. As an aside, this can be seen as a parametric bootstrap sample combined with the observed data. The GMM is then recomputed on this new data, then the log likelihood and log likelihood ratio statistic are computed. This process is repeated 500 times, essentially mimicking a parametric bootstrap approximation of the likelihood ratio distribution.

The log likelihoods for the original cluster along with the candidates (one at a time) are computed and compared to the approximated likelihood ratio distribution. Candidates are rejected if they rank in the tail of the likelihood ratio distribution. Accepted candidates are merged with class k and the process is repeated until all classes have been considered. Once this process is finished, the proteins not merged with any class but which clustered together during the algorithm are considered as new phenotypes.

The process is repeated 100 times, where a new phenotype is declared if, over those 100 repetitions, the proteins consistently cluster together. A minimum group size is specified by the user.

5.3.2 Extending TAGM to allow novelty detection

The goal of this section to use the principle of overfitted mixture to allow TAGM to not only assigned proteins to an annotated organelle but to detect unannotated subcellular niches. Figure 5.1 gives an idea of the approach. Let us briefly remind ourself of the TAGM model.

Let N denote the number of observed protein profiles each of length L, corresponding to the number of quantified fractions. The quantitative profile for the i-th protein is denoted by $\mathbf{x}_i = [x_{1i}, \dots, x_{Li}]$. In chapter 2, the model was formulated such that there are K known sub-cellular compartments to which each protein could be localised (e.g. cytosol, endoplasmic reticulum, mitochondria, ...). We introduce component labels z_i, so that $z_i = k$ if the i-th protein localises to the k-th component. To fix notation, we denote by X_L the set of proteins whose component labels are known, and by X_U the set of unlabelled proteins. If protein i is in X_U, we seek to evaluate the probability that $z_i = k$ for each $k = 1, \dots, K$.

The distribution of quantitative profiles associated with each protein that localises to the k-th component is modelled as multivariate normal with mean vector $\boldsymbol{\mu}_k$ and covariance matrix Σ_k. However, many proteins are dispersed and do not fit this assumption. To model these "outliers", we introduced a further indicator variable ϕ. Each protein $\mathbf{x}_i$ is then described by an additional variable ϕ_i, with $\phi_i = 1$ indicating that protein $\mathbf{x}_i$ belongs to an organelle-derived component and $\phi_i = 0$ indicating that protein $\mathbf{x}_i$ is not well described by these known

components. This *outlier component* is then modelled as a multivariate T distribution with degrees of freedom κ, mean vector $\mathbf{M}$, and scale matrix V. Thus the model can be written as:

$$\mathbf{x}_i|z_i = k, \phi_i \quad \sim \mathcal{N}(\boldsymbol{\mu}_k, \Sigma_k)^{\phi_i}\mathcal{T}(\kappa, \boldsymbol{M}, V)^{1-\phi_i}. \tag{5.6}$$

Let $f(\mathbf{x}|\boldsymbol{\mu}, \Sigma)$ denote the density of the multivariate normal with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ evaluated at $\mathbf{x}$, and similarly let $g(\mathbf{x}|\kappa, \mathbf{M}, \mathbf{V})$ denote the density of the multivariate T-distribution. For any i, the prior probability of the i-th protein localising to the k-th component is denoted by $p(z_i = k) = \pi_k$. Letting $\boldsymbol{\theta} = \{\boldsymbol{\mu}_k, \Sigma_k\}_{k=1}^K$ denote the set of all component mean and covariance parameters, and $\boldsymbol{\pi} = \{\pi_k\}_{k=1}^K$ denote the set of all mixture weights, it follows that:

$$p(\mathbf{x}_i|\boldsymbol{\theta}, \boldsymbol{\pi}, \phi_i, \kappa, \mathbf{M}, V) = \sum_{k=1}^{K} \pi_k \left(f(\mathbf{x}_i|\boldsymbol{\mu}_k, \Sigma_k)^{\phi_i} g(\mathbf{x}_i|\kappa, \boldsymbol{M}, V)^{1-\phi_i} \right). \tag{5.7}$$

For any i, we set the prior probability of the i-th protein belonging to the outlier component as $p(\phi_i = 0) = \epsilon$, where ϵ is a parameter that we infer. Equation (5.7) can then be rewritten in the following way:

$$p(\mathbf{x}_i|\boldsymbol{\theta}, \boldsymbol{\pi}, \kappa, \epsilon, \mathbf{M}, V) = \sum_{k=1}^{K} \pi_k \left((1-\epsilon)(f(\mathbf{x}_i|\boldsymbol{\mu}_k, \Sigma_k) + \epsilon g(\mathbf{x}_i|\kappa, \boldsymbol{M}, V) \right), \tag{5.8}$$

As in chapter 2, we fix $\kappa = 4$, $\mathbf{M}$ as the global empirical mean, and V as half the global empirical variance of the data, including labelled and unlabelled proteins. To extend this model to permit novelty detection, we specify the maximum number of components $K_{max} > K$. Our proposed model then allows up to $K_{novelty} = K_{max} - K \geq 0$, new phenotypes to be detected. Equation 5.8 can then be written as

$$\begin{aligned} p(\mathbf{x}_i|\boldsymbol{\theta}, \boldsymbol{\pi}, \kappa, \epsilon, \mathbf{M}, V) = &\sum_{k=1}^{K} \pi_k \left((1-\epsilon)(f(\mathbf{x}_i|\boldsymbol{\mu}_k, \Sigma_k) + \epsilon g(\mathbf{x}_i|\kappa, \boldsymbol{M}, V) \right) \\ &+ \sum_{k=K+1}^{K_{max}} \pi_k \left((1-\epsilon)(f(\mathbf{x}_i|\boldsymbol{\mu}_k, \Sigma_k) + \epsilon g(\mathbf{x}_i|\kappa, \boldsymbol{M}, V) \right), \end{aligned} \tag{5.9}$$

where, in the first summation, the K components correspond to known sub-cellular niches and the second summation corresponds to the new phenotypes to be inferred. The parameter sets are then augmented to include these possibly new components; that is, we redefine $\boldsymbol{\theta} = \{\boldsymbol{\mu}_k, \Sigma_k\}_{k=1}^{K_{max}}$ to denote the set of all component mean and covariance parameters, and

$\boldsymbol{\pi} = \{\pi_k\}_{k=1}^{K_{max}}$ denotes the set of all mixture weights. Relying on the principle of over-fitted mixtures [397], components that are not supported by the data are left empty with no proteins allocated to them. We find setting $K_{novelty} = 10$ is ample to detect new phenotypes. Note that we have to choose $\max_j \beta_j < d/2$, which is satisfied in all examples by setting $\beta_j = 0.5$ for every j. Importantly, we cannot use the weakly informative prior on π suggested in chapter 3, since $\min_j \beta_j > d/2$.

Bayesian inference and convergence

The MCMC algorithm used in chapter 2 is insufficient to handle inference of unknown phenotypes. As before, a collapsed Gibbs sampler approach is used, but a number of modifications are made. Firstly, to accelerate convergence of the algorithm half the proteins are initially allocated randomly amongst the new phenotypes. Secondly, the parameters for the new phenotypes are simply proposed from the prior. Otherwise, the same default prior choices are used.

Handling label switching

Bayesian inference in mixture models suffers from an identifiability issue known as *label switching* - a phenomenon where the allocation labels can flip between runs of the algorithm [380, 436]. This occurs because of the symmetry of the likelihood function under permutations of these labels. We note that this only occurs in unsupervised or semi-supervised mixture models. This makes inference of the parameters in mixture models challenging. In our setting the labels for the known components do not switch, but for the new phenotypes label switching must occur. One standard approach to circumvent this issue is to form the so-called *posterior similarity matrix* (PSM) [142]. The PSM is an $N \times N$ matrix where the $(i,j)^{th}$ entry is the posterior probability that protein i and protein j reside in the same component. More precisely, if we let S denote the PSM and T denote the number of Monte-Carlo iterations then

$$S_{ij} = P(z_i = z_j | X, \boldsymbol{\theta}, \boldsymbol{\pi}, \kappa, \epsilon, \mathbf{M}, V) \approx \frac{1}{T} \sum_{t=1}^{T} \mathbb{I}(z_i^{(t)} = z_j^{(t)}), \tag{5.10}$$

where $\mathbb{I}$ denotes the indicator function. The PSM is clearly invariant to label switching and so avoids the issues arising from the *label switching* problem.

Visualising patterns in uncertainty

To simultaneously visualise the uncertainty in the number of newly discovered phenotypes, we use a heatmap representation of this quantity.

Summarising posterior similarity matrices

To summarise the PSMs, we take the approach proposed by Fritsch et al. [142]. A general strategy to summarise a PSM into a clustering is as follows. First, we propose some loss function or measure of similarity between two clusterings $L(z^*, z)$. We then wish to find the allocation vector $\hat{z}$ that minimises the expected loss with respect to the true clustering

$$\hat{z} = \arg\min_{z^*} E[L(z^*, z)|X, \boldsymbol{\theta}, \boldsymbol{\pi}, \kappa, \epsilon, \mathbf{M}, V]. \tag{5.11}$$

Fritsch et al. [142] propose to use the adjusted Rand index (AR) [374, 213], a measure of cluster similarity, as the utility function. Then we find the allocation vector $\hat{z}$ that maximises the expected adjusted Rand index with respect to the true clustering z. Formally, we write

$$\hat{z} = \arg\max_{z^*} E[AR(z^*, z)|X, \boldsymbol{\theta}, \boldsymbol{\pi}, \kappa, \epsilon, \mathbf{M}, V], \tag{5.12}$$

which is known as the Posterior Expected Adjusted Rand index (PEAR). One obvious pitfall is that this quantity depends on the unknown true clustering z. However, this can be approximated from the MCMC samples:

$$PEAR \approx \frac{1}{T}\sum_{t=1}^{T} AR(z^*, z^{(t)}). \tag{5.13}$$

The space of all possible clustering over which to maximise is infeasibly large to explore. Thus we take an approach taken in Fritsch et al. [142] to propose candidate clusterings over which to maximise. Using hierarchical clustering with distance $1 - S_{ij}$, the PEAR criterion is computed for clusterings at every level of the hierarchy. The optimal clustering $\hat{z}$ is the allocation vector which maximises the PEAR.

Uncertainty quantification

We may be interested in quantifying the uncertainty in whether a protein belongs to a new sub-cellular component. Indeed, it is important to distinguish whether a protein belongs to a new phenotype or if we simply have large uncertainty about its localisation. The probability that protein i belongs to a new component is computed from the following equation:

$$P(z_i \in \{K+1, ..., K_{max}\}|X) = 1 - P(z_i \in \{1, ..., K\}|X), \tag{5.14}$$

which we approximate by the following Monte-Carlo average:

$$1 - \frac{1}{T}\sum_{t=1}^{T} P(z_i^{(t)} \in \{1, ..., K\}|X) = 1 - \frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{K} P(z_i^{(t)} = k|X), \tag{5.15}$$

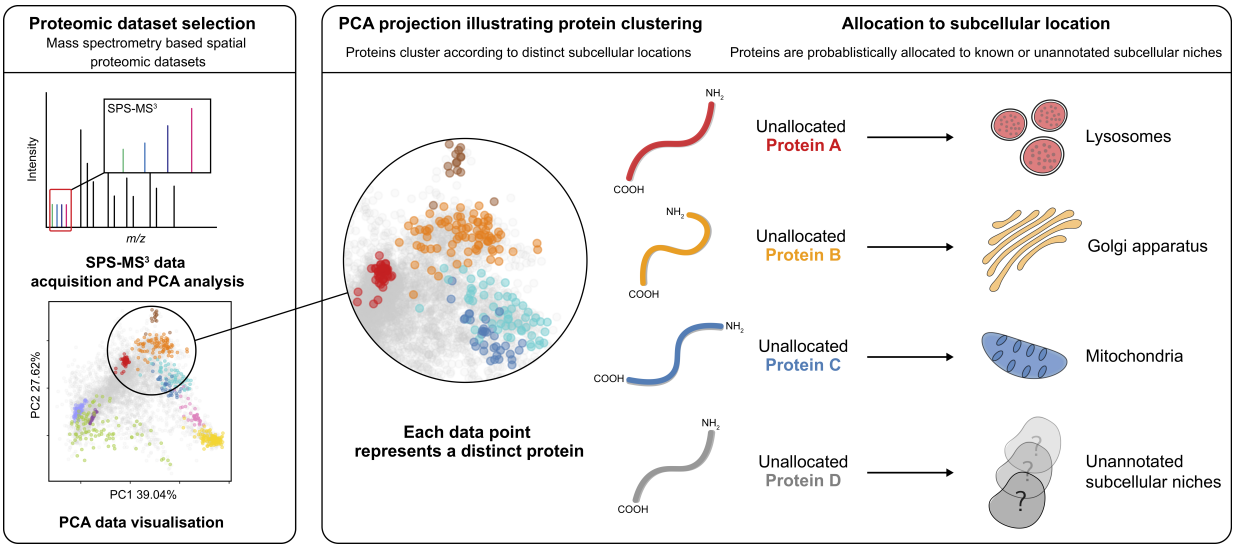


Fig. 5.1 An overview of novelty detection in subcellular proteomics.

where T is the number of Monte-Carlo iterations. Throughout, we refer to equation 5.15 as the *discovery probability*.

Applying the model in practice

Applying Novelty TAGM to spatial proteomics datasets consists of several steps. After having run the algorithm on a dataset and assessing convergence, we proceed to explore the ouput of the method. We explore *putative phenotypes*, which we define as newly discovered clusters with at least 1 protein with discovery probability greater than 0.95.

5.3.3 Validating computational approaches

In a supervised framework the performance of computational methods can be assessed by using the training data, where a proportion of the training data is withheld from the classifier to be used for the assessment of predictive performance. In an unsupervised or semi-supervised framework we cannot validate in this way, since there is no "ground truth" with which to compare. Thus, we propose several approaches, using external information, for validation of our method.

Artificial masking of annotations to recover experimental design

Removing the labels from an entire component and assessing the ability of our method to rediscover these labels is one form of validation. We consider this approach for several of the datasets; in particular, chromatin enrichment was performed in two of the *hyper*LOPIT

experiments, where the intention was to increase the resolution between chromatin and non-chromatin associated nuclear proteins [68, 324, 455]. As validation of our method we hide these labels and seek to rediscover them in an unbiased fashion.

The Human Protein Atlas

A further approach to validating our method is to use additional spatial proteomic information. The Human Protein Atlas (HPA) [455, 441] provides confocal microscopy information on thousands of proteins, using validated antibodies. When we consider a dataset for which there is HPA annotation, we use this data to validate the novel phenotypes for biological relevance.

Gene Ontology (GO) term enrichment

Throughout, we perform GO enrichment analysis with FDR control performed according to the Benjamini-Höchberg procedure [25, 14, 489]. The proteins in each novel putative phenotype are assessed in turn for enriched Cellular Component terms, against the background of all quantified proteins in that experiment.

Robustness across multiple MS-based spatial proteomics datasets

On occasion some cell lines have been analysed using multiple spatial proteomics technologies [159]. In these cases, the putative phenotypes discovered by Novelty TAGM are compared directly. If the same phenotype is discovered in different proteomic datasets we consider this as robust evidence for sufficient resolution of that phenotype.

5.3.4 Datasets

In this section, we provide a brief description of the datasets used in this chapter. We analyse *hyper*LOPIT data, in which sub-cellular fractionation is performed using density-gradient centrifugation [118, 119, 324], on pluripotent mESCs (E14TG2a) [68], human bone osteosarcoma (U-2 OS) cells [455, 159], and *S. cerevisiae* (bakers' yeast) cells [339]. The mESC dataset combines two 10-plex biological replicates and quantitative information on 5032 proteins. The U-2 OS dataset combines three 20-plex biological replicates and provides information on 4883 proteins. The yeast dataset represents four 10-plex biological replicate experiments performed on *S. cerevisiae* cultured to early-mid exponential phase. This dataset contains quantitative information for 2846 proteins that were common across all replicates. Tandem Mass Tag (TMT) [453] labelling was used in all *hyper*LOPIT experiments with LC-SPS-MS^3 used for high accuracy quantitation [457, 298]. Beltran et al. [24] integrated a temporal component to the LOPIT protocol. They analysed HCMV-infected primary fibroblast cells over 5 days, producing control and infected maps every 24 hours. We analyse the control and infected maps

24 hours post-infection, providing information on 2220 and 2196 proteins respectively. In a comparison with *phenoDisco*, we apply Novelty TAGM to a dataset acquired using LOPIT-based fractionation and 8-plex iTRAQ labelling on the HEK-293 human embryonic kidney cell line, quantifying 1371 proteins [43].

Our approach is not limited to spatial proteomics data where the sub-cellular fractionation is performed using density gradients. We demonstrate this through the analysis of DOM datasets on HeLa cells and mouse primary neurons [220, 221], which quantify 3766 and 8985 proteins respectively. These approaches used SILAC quantitation with differential centrifugation-based fractionation. We analyse 6 replicates from the HeLa cell line analyses in [220] and 3 replicates from the mouse primary neuron experiments in [221]. [202] also used the DOM protocol coupled with CRISPR-cas9 knockouts in order to explore the functional role of AP-5. We analyse the control map from this experiment. Finally, we consider the U-2 OS data which were acquired using the LOPIT-DC protocol [159] and quantified 6837 proteins across 3 biological replicates. In favour of brevity, we do not consider protein correlation profiling (PCP) based spatial proteomics datasets in this chapter, though our method also applies to such data [135, 254, 253] and other sub-cellular proteomics methods which utilised cellular fractionation [351].

5.4 Results

Motivated by the need for novelty detection methods which also quantify the uncertainty in the number of clusters and the assignments of proteins to each cluster, we developed Novelty TAGM. Our proposed methodology allows us to interrogate individual proteins to assess whether they belong to a newly discovered phenotype. To demonstrate the value of this approach, we applied Novelty TAGM to a diverse set of spatial proteomics datasets.

5.4.1 Validating experimental design in *hyper*LOPIT

Initially, we validated Novelty TAGM in a setting where we have a strong *a priori* expectation for the presence of an unannotated niche. For this we used a human bone osteosarcoma cell (U-2 OS) *hyper*LOPIT dataset [455] and an mESC *hyper*LOPIT dataset [68]. These experimental protocols used a chromatin enrichment step to resolve nuclear chromatin-associated proteins from nuclear proteins not associated with chromatin. Removing the nuclear, chromatin and ribosomal annotations from the datasets, we test the ability of Novelty TAGM to recover them.

Human bone osteosarcoma (U-2 OS) cells

For the U-2 OS dataset, Novelty TAGM reveals 9 putative phenotypes, which we refer to as phenotype 1, phenotype 2, etc... These phenotypes, along with the uncertainty associated with them, are visualised in figure 5.2. We consider the HPA confocal microscopy data for validation

[455, 441]. The HPA provides information on the same cell line and therefore constitutes an excellent complementary resource. This *hyper*LOPIT dataset was already shown to be in strong agreement with the microscopy data [455, 159]. Proteins in phenotypes $3, 4, 5$ and 8 have a nucleus-related annotation as their most frequent HPA annotation, as well as differential enrichment of nucleus-related GO terms (figure 5.2). Phenotype 3 validates the chromatin enrichment preparation (figure 5.2 panel (c)) and phenotype 4 reveals a nucleoli cluster, where nucleoli and nucleoli/nucleus are the 2^{nd} and 3^{rd} most frequent HPA annotations for proteins belonging to this phenotype. For phenotype 5, the most associated term is nucleoplasm from the HPA data and this is further supported by GO analysis (figure 5.2 panel (c)). Phenotype 8 demonstrates further sub-nuclear resolution and has *nuclear membrane* as its most frequent HPA annotation and has corresponding enriched GO terms (figure 5.2 panel (c)). In addition, phenotypes 1 and 2 are enriched for *ribosomes* and *endosomes* respectively.

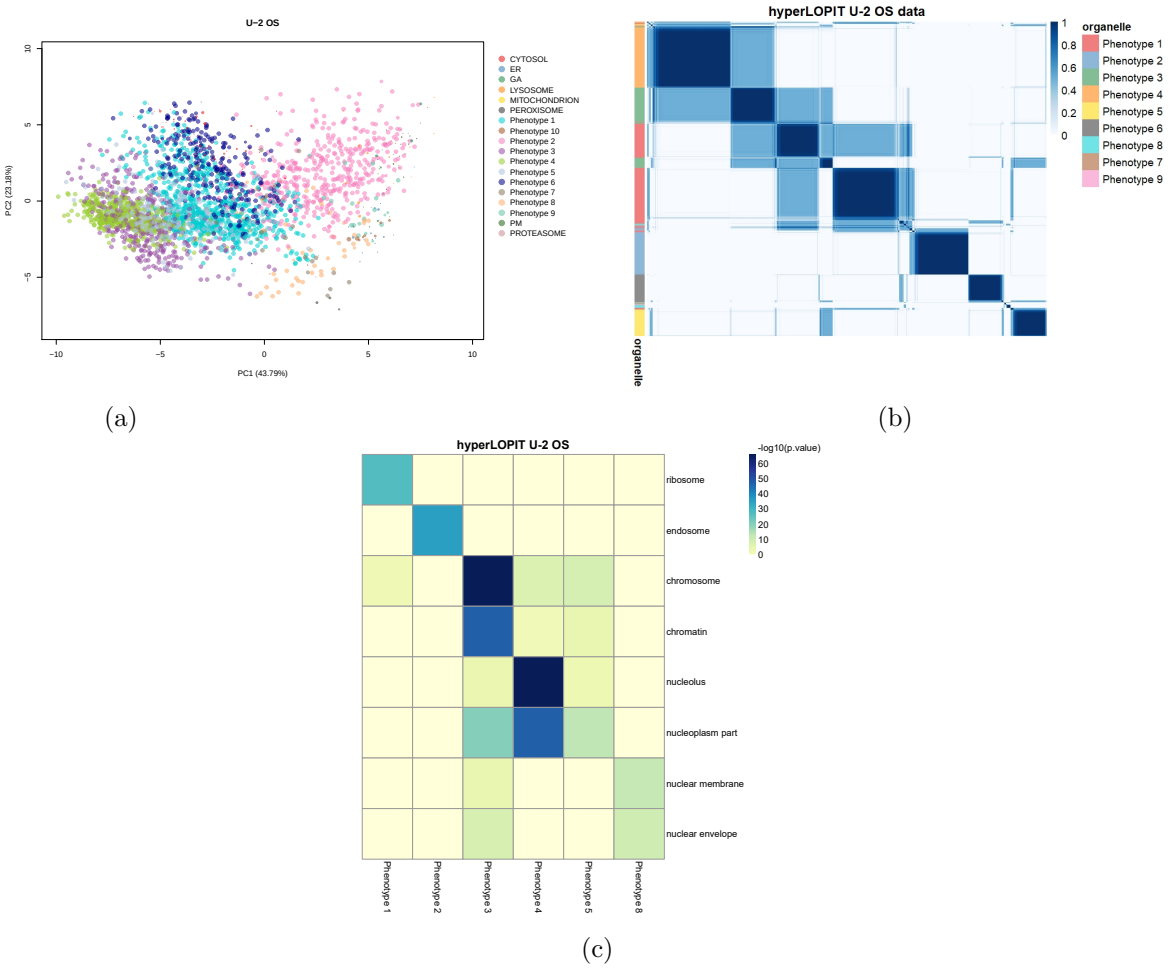
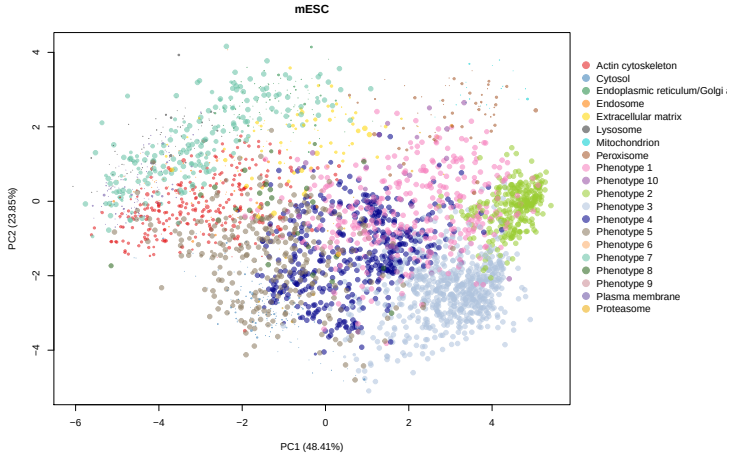


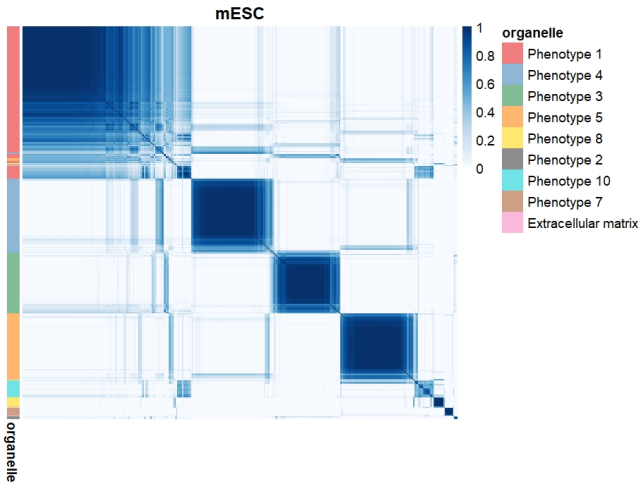
Fig. 5.2 (a) PCA plot of the *hyper*LOPIT U-2 OS cancer cell line data. Points are scaled according to the discovery probability with larger points indicating greater discovery probability. (b) Heatmaps of the posterior similarity matrix derived from U-2 OS cell line data demonstrating the uncertainty in the clustering structure of the data. We have only plotted the proteins which have greater than 0.99 probability of belonging to a new phenotype and probability of being an outlier less than 0.5 for the U-2 OS dataset to reduce the number of visualised proteins. (c) Tile plot of discovered phenotypes against GO CC terms to demonstrate over-representation, where the colour intensity is the -$\log_{10}$ of the p-value.

mESC chromatin enrichment validation

For the mESC dataset, Novelty TAGM reveals 8 new putative phenotypes. Novelty TAGM recovers the masked annotations with phenotype 2 having the enriched terms associated with chromatin, such as *chromatin* and *chromosome* ($p < 10^{-80}$). Phenotype 3 corresponds to a separate nuclear substructure with enrichment for the terms *nucleolus* ($p < 10^{-60}$) and *nuclear body* ($p < 10^{-30}$). Thus, in the mESC dataset Novelty TAGM confirms the chromatin enrichment preparation designed to separate chromatin and non-chromatin associated nuclear proteins [324]. In addition, phenotype 4 demonstrates enrichment for the ribosome annotation ($p < 10^{-35}$). Phenotype 1 is enriched for *centrosome* and *microtubule* annotations ($p < 10^{-15}$), though observing the PSM in figure 5.3 we can see there is much uncertainty in this phenotype. This uncertainty quantification can then be used as a basis for justifying additional expert annotation.



(a)



(b)

Fig. 5.3 (a) PCA plot of the *hyper*LOPIT mESC dataset. Points are scaled according to the discovery probability. (b) Heatmaps of the posterior similarity matrix derived from mESC data demonstrating the uncertainty in the clustering structure of the data. We have only plotted the proteins which have greater than 0.99 probability of belonging to a new phenotype and probability of being an outlier less than 0.95 for the mESC dataset to reduce the number of visualised proteins.

5.4.2 Uncovering additional sub-cellular structures

Having validated the ability of Novelty TAGM to recover known experimental design, as well as uncover additional sub-cellular niches resolved in the data, we turn to apply Novelty TAGM to several additional datasets.

U-2 OS cell line revisited

We first consider the LOPIT-DC dataset on the U-2 OS cell line [159]. Again, we removed the nuclear, proteasomal, and ribosomal annotations. Novelty TAGM reveals 10 putative phenotypes (figure 5.4).

In a similar vein to the analysis performed on the *hyper*LOPIT U-2 OS dataset, we initially use the available HPA data to validate these clusters [455]. Phenotypes $3, 5, 7$ and 9 display nucleus-associated terms as their most frequent HPA annotation. Clear differential enrichment of phenotypes with GO Cellular Component terms is evident from figure 5.4 panel (e). This analysis reveals *nucleolus*, *ribosome*, *proteasome* phenotypes. Furthermore, a *chromatin* phenotype is also resolved. Notably, this is the first evidence for sub-nuclear resolution in this LOPIT-DC dataset. Phenotype 6 represents a cluster with mixed *plasma membrane* and *extracellular matrix* annotations and this is supported by HPA annotation with vesicles, cytosol, and plasma membrane being the top three annotations. An extracellular matrix-related phenotype was not previously known in these data and might correspond to exocytic vesicles containing ECM proteins. Furthermore, phenotype 8 is significantly enriched for *endosomes*, again a novel annotation for this data. In addition, 107 of the proteins in this phenotype are also localised to the endosome-enriched phenotype presented in the U-2 OS *hyper*LOPIT dataset (section 5.4.1). Thus, we robustly identify new phenotypes across different spatial proteomics protocols. Hence, we have presented strong evidence for additional annotations in this dataset, beyond the original analysis of the data [159]. In particular, although a separate chromatin enrichment preparation was not included in the U-2 OS LOPIT-DC analysis and the original authors did not identify sufficient resolution between the nucleus and chromatin clusters in this dataset, Novelty TAGM could, in fact, reveal a chromatin-associated phenotype in the U-2 OS LOPIT-DC data. In addition, we have joint evidence for an endosomal cluster in both the LOPIT-DC and *hyper*LOPIT datasets. Finally, through the discovery probability and by using the PSMs we have quantified uncertainty in these proposed phenotypes, enabling more rigorous interrogation of these datasets.

Saccharomyces cerevisiae

Novelty TAGM uncovers 8 putative phenotypes in the yeast *hyper*LOPIT data [339]. Four of these phenotypes have no significant over-represented annotations. Figure 5.4 panel (f)

demonstrates that the remaining four phenotypes are differentially enriched for GO terms. Firstly, a mixed *cell periphery* and *fungal-type vacuole* phenotype is uncovered along with a *kinetochore* phenotype, and a *cytoskeleton* phenotype. Phenotype 8 represents a joint Golgi and ER cluster with several enriched GO terms. Indeed, most of the proteins in this phenotype have roles in the early secretory pathway that involve either transport from the ER to the early Golgi apparatus, or retrograde transport from the Golgi to the ER [52, 217, 352, 487], (also reviewed in [106]). To be precise, 11 out of the total 20 proteins in this cluster are annotated as core components of COPII vesicles and 6 associated with COPI vesicles. The protein Ksh1p (Q8TGJ3) is further suggested through homology with higher organisms to be part of the early secretory pathway [477]. The proteins Scw4p (P53334), Cts1p (P29029) and Scw10p (Q04951) [57], as well as Pst1p (Q12355)[355], and Cwp1p (P28319) [486], however, are annotated in the literature as localising to the cell wall or extracellular region. It is therefore possible that their predicted co-localisation with secretory pathway proteins observed here reflects a proportion of their lifecycle being synthesised or spent trafficking through the secretory pathway. The protein Ssp120p (P39931) is of unknown function and has been shown to localise in high throughput studies to the vacuole [487] and to the cytoplasm in a punctate pattern [215]. The localisation observed here may suggest that it is therefore either part of the secretory pathway, or trafficks through the secretory organelles for secretion or to become a constituent of the cell wall.

HCMV-infected fibroblast cells

We apply Novelty TAGM to the dataset corresponding to the HCMV-infected fibroblast cells 24 hours post infection (hpi) [24], and discover 9 putative additional phenotypes (demonstrated in figure 5.5). Phenotype 2 contains a singleton protein and phenotypes $4, 6, 7, 8$ and 9 are not significantly enriched for any annotations. However, phenotype 3 is enriched for the *mitochondrial membrane* and *mitochondrial envelope* annotations ($p < 10^{-4}$); this is an addition to the already annotated mitochondrial class, indicating sub-mitochondrial resolution. Phenotype 1 is a mixed ribosomal/nuclear cluster with enrichment for *nucleoplasm* ($p < 10^{-5}$) and the *small ribosomal subunit* ($p < 10^{-4}$), which is distinct from phenotype 5 which is enriched for the *large ribosomal subunit* ($p < 10^{-10}$). This demonstrates unbiased separation of the two ribosomal subunits, which was overlooked in the original analysis [24].

Fibroblast cells without infection

Novelty TAGM reveals 7 putative phenotypes in the control fibroblast dataset [24]. Phenotypes $2, 4, 5, 6$ and 9 have no significantly enriched Gene Ontology terms (threshold $p = 0.01$). However, we observe that phenotype 3 is enriched with the *large ribosomal subunit* with significance at level $p < 10^{-7}$. Phenotype 1 represents a mixed *peroxisome* ($p < 10^{-2}$) and *mitochondrion* cluster ($p < 10^{-2}$), an unsurprising result since these organelles possess similar

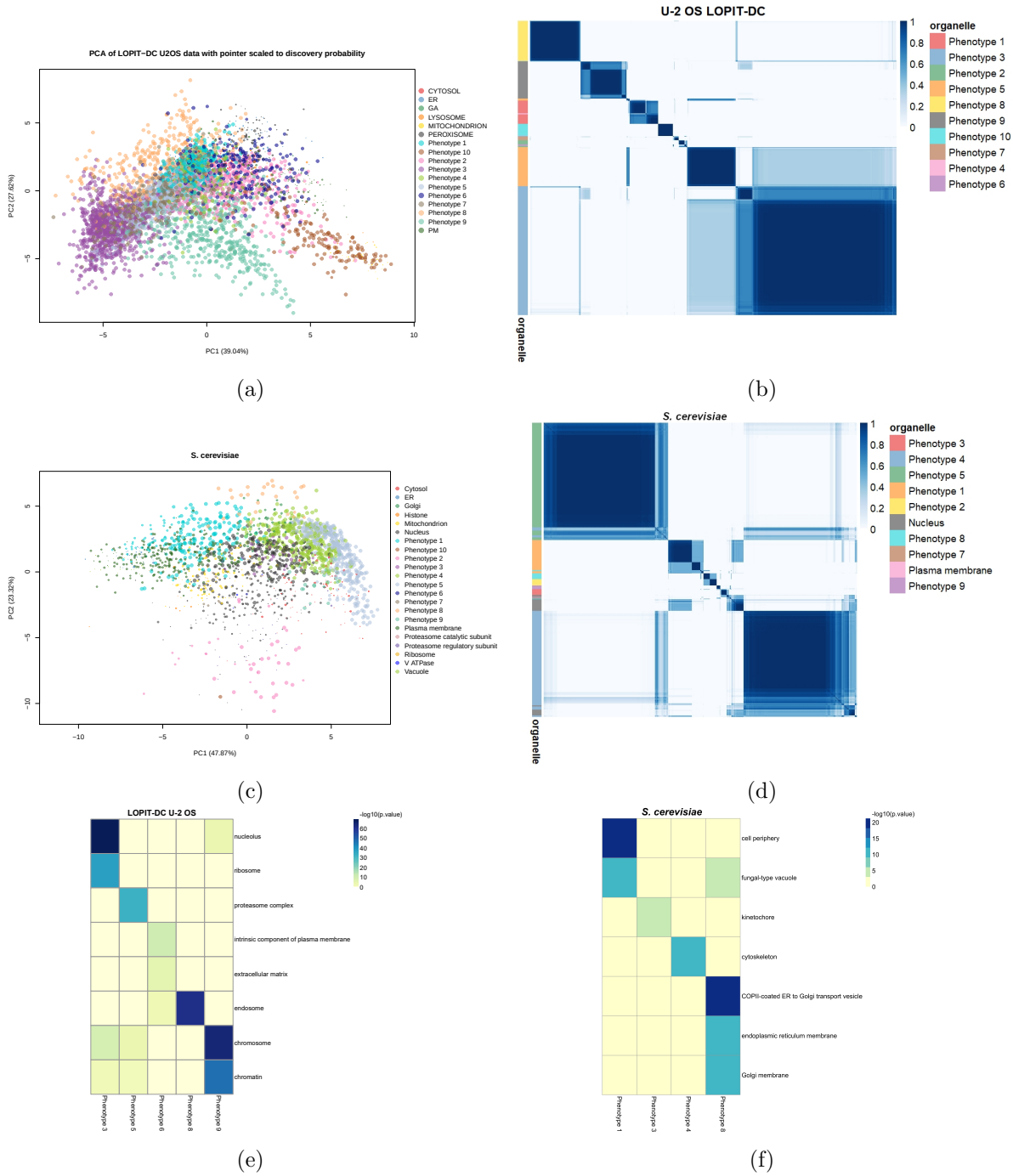


Fig. 5.4 (a, c) PCA plots of the LOPIT-DC U-2 OS data and the *hyper*LOPIT yeast data. The points are scaled according to the discovery probability. (b, d) Heatmaps of the posterior similarity matrix derived from the U-2 OS and yeast datasets demonstrating the uncertainty in the clustering structure of the data. We have only plotted the proteins which have greater than 0.99 probability of belonging to a new phenotype and probability of being an outlier less than 0.95 (10^{-5} for LOPIT-DC to reduce the number of visualised proteins). (e, f) Tile plots of phenotypes against GO CC terms where the colour intensity is the -$\log_{10}$ of the p-value.

biochemical properties and therefore similar profiles during density gradient centrifugation-based fractionation [159, 104]. The differing number of confidently identified and biologically relevant phenotypes discovered between the two fibroblast datasets could be down to the differing levels of structure between the two datasets. Indeed, it is evident from figure 5.5 that we see differing levels of clustering structure in these datasets.

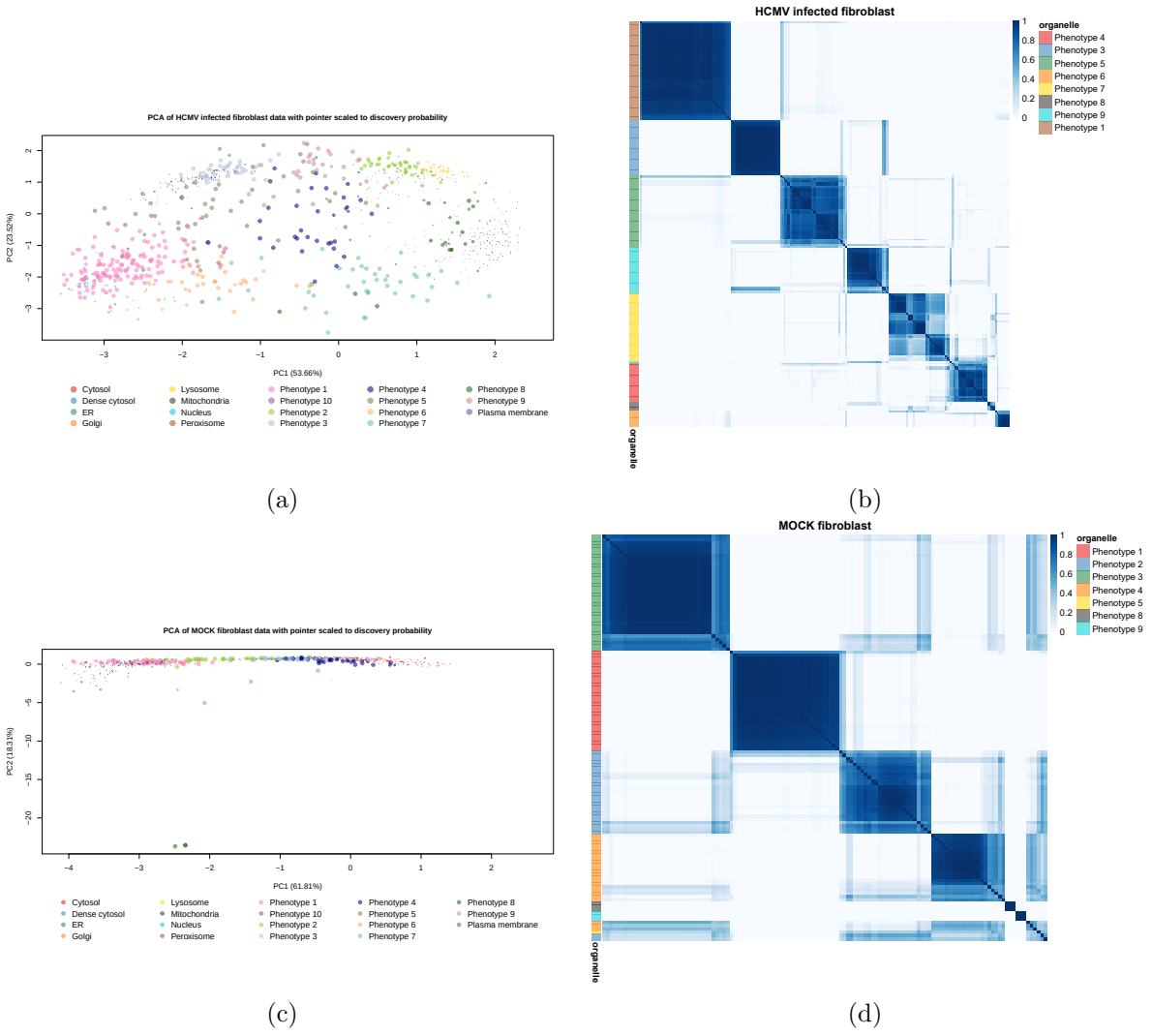


Fig. 5.5 (a, c) PCA plots of the HCMV-infected fibroblast data 24 hpi and the mock fibroblast data 24 hpi. The points are coloured according to the organelle or proposed new phenotype and are scaled according to the discovery probability. (b, d) Heatmaps of the posterior similarity matrix derived from the infected fibroblast data and mock fibroblast data demonstrating the uncertainty in the clustering structure of the data. We have only plotted the proteins which have greater than 0.99 probability of belonging to a new phenotype and probability of being an outlier less than 0.95.

5.4.3 Refining annotation in organellar maps

The Dynamic Organellar Maps (DOM) protocol was developed as a faster method for MS-based spatial proteomic mapping, albeit at the cost of lower organelle resolution [220, 157]. The three datasets analysed here are two HeLa cell lines [220, 202] and a mouse primary neuron dataset [221]. All three of these datasets have been annotated with a class called "large protein complexes". This class contains a mixture of cytosolic, ribosomal, proteasomal and nuclear sub-compartments that pellet during the centrifugation step used to capture this mixed fraction [220]. We apply Novelty TAGM to these data and remove this "large protein complexes" class, to derive more precise annotations for these datasets.

HeLa cells (Itzhak et. al 2016)

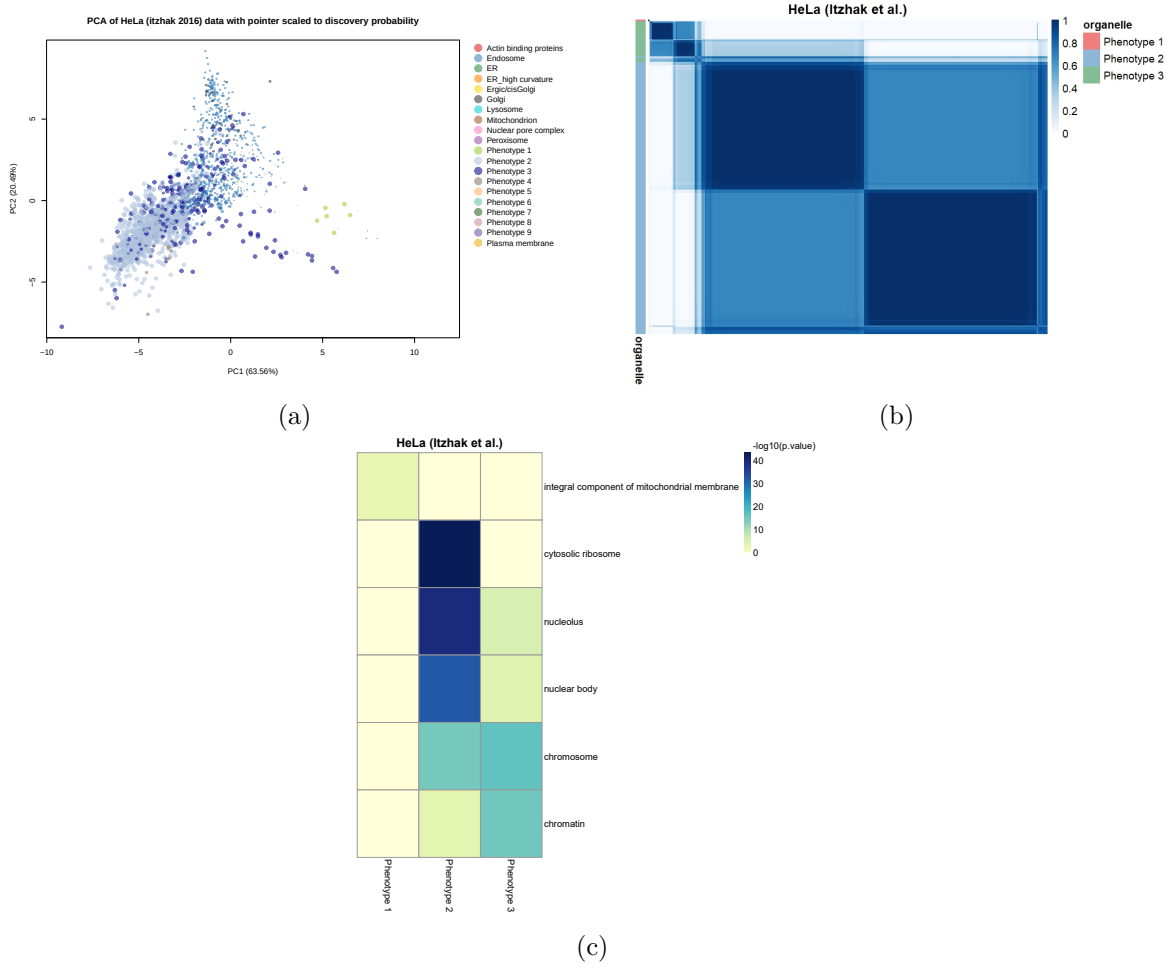
The HeLa dataset of [220] has 3 additional phenotypes uncovered by Novelty TAGM. Figure 5.6 panel c shows a *mitochondrial membrane* phenotype, distinct from the already annotated mitochondrial class. Phenotype 2 represents a mixed cluster with nucleus-, ribosome- and cytosol-related enriched terms. The final phenotype is enriched for *chromatin* and *chromosome*, suggesting sub-nuclear resolution. Furthermore, as a result of quantifying uncertainty, we can see that there are potentially more sub-cellular structures in this data (figure 5.6). However, the uncertainty is too great to support these phenotypes.

Mouse primary neurons

The mouse primary neuron dataset reveals 10 phenotypes after we apply Novelty TAGM. However, 8 of these phenotypes have no enriched GO annotations. This is likely a manifestation of the dispersed nature of this dataset, where the variability is generated by technical artefacts rather than biological signal. Despite this, Novelty TAGM is able to detect two relevant phenotypes: the first phenotype is enriched for *nucleolus* ($p < 0.01$); the second for *chromosome* ($p < 0.01$). This suggests additional annotations for this dataset.

HeLa cells (Hirst et. al 2018)

The HeLa dataset of [202], which we refer to as HeLa Hirst, reveals 7 phenotypes with at least 1 protein with discovery probability greater than 0.95. However, three of these phenotypes represent singleton proteins. Phenotype 1 reveals mixed cytosol/ribosomal annotations with the terms *cytosolic ribosome* ($p < 10^{-30}$) and *cytosolic part* ($p < 10^{-25}$) significantly over-represented. There are no further phenotypes with enriched annotations (threshold $p = 0.01$), except phenotype 2 which represents a mixed extracellular structure/cytosol cluster. For example, the terms *extracellular organelle* ($p < 10^{-13}$) and *cytosol* ($p < 10^{-10}$) are over-represented.



(a) (b)

(c)

Fig. 5.6 (a) PCA plots of the HeLa data. The pointers are scaled according to their discovery probability. (b) Heatmaps of the HeLa Itzhak data. Only the proteins with discovery probability greater than 0.99 and outlier probability less than 0.95 are shown. The heatmaps demonstrate the uncertainty in the clustering structure present in the data. (c) Tile plot of phenotypes against GO CC terms where the colour intensity is the -$\log_{10}$ of the p-value.

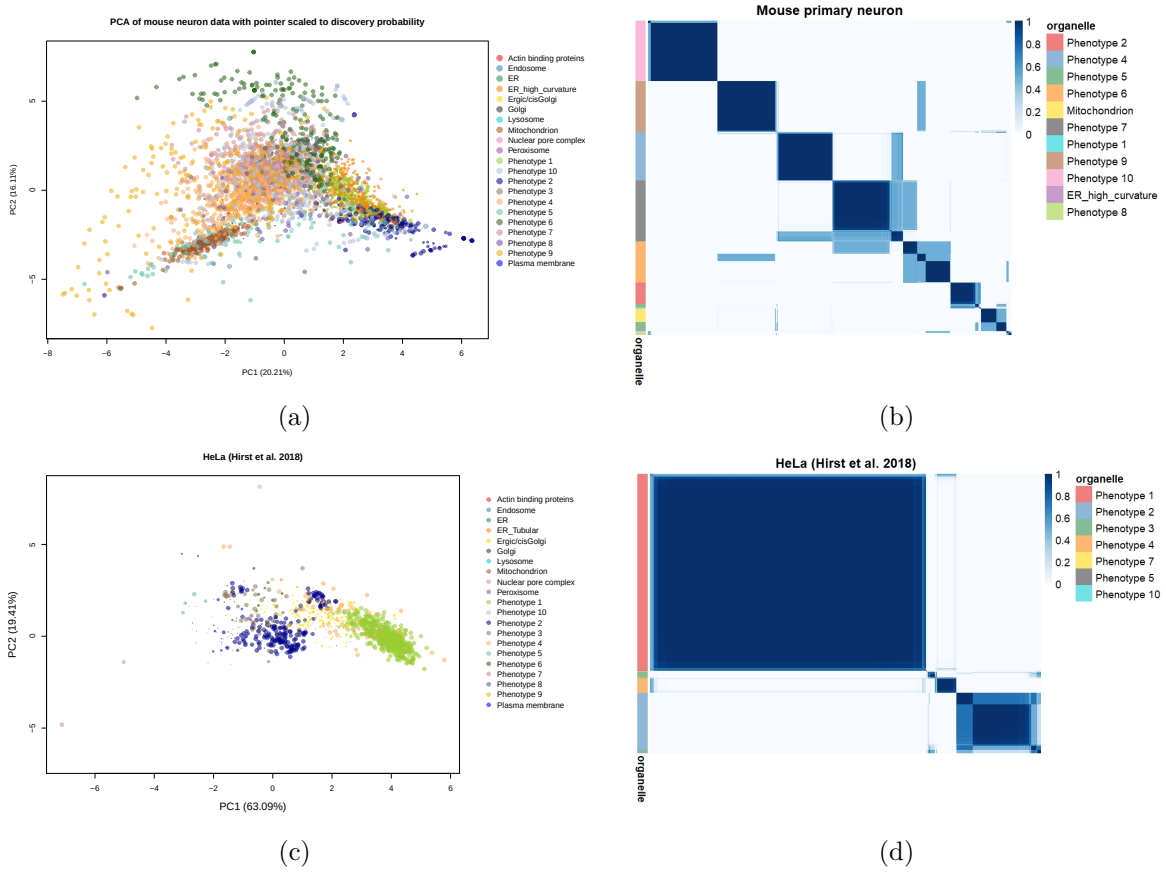
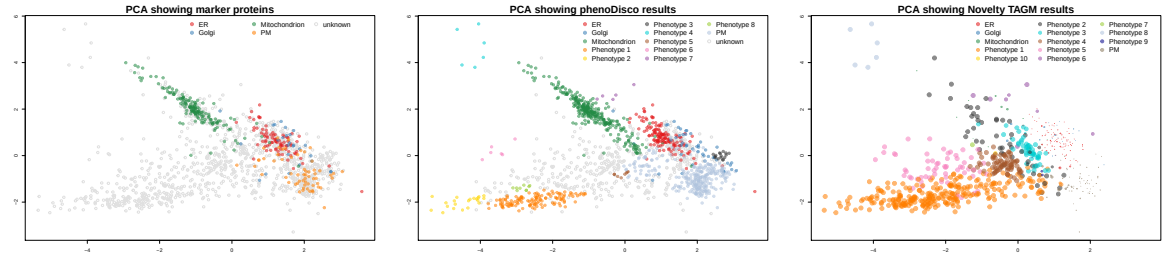


Fig. 5.7 (a),(c) PCA plots of the mouse primary neuron data and HeLa Hirst data. The pointers are scaled according to their discovery probability. (b),(d) Heatmaps of the mouse neuron data and HeLa Hirst data. Only the proteins whose discovery probability is greater than 0.99 and outlier probability less than 0.95 (10^{-2} for the mouse primary neuron dataset to reduce the number of visualised proteins) are shown. The heatmaps demonstrate the uncertainty in the clustering structure present in the data.

5.5 Comparison between Novelty TAGM and *phenoDisco*

Next, we compare an already available novelty detection algorithm, *phenoDisco* [43], with Novelty TAGM. Despite both methods performing novelty detection, the algorithms are quite distinct. The first major difference is that Novelty TAGM is a Bayesian method that performs uncertainty quantification. Novelty TAGM quantifies the uncertainty in both the number of newly identified phenotypes and whether individual proteins should belong to a new phenotype. On the other hand, *phenoDisco* uses the *Bayesian Information Criterion* (BIC) to select just a single clustering, without taking into account the uncertainty in the number of phenotypes, and does not provide an estimate of individual protein-to-phenotype allocation uncertainty. Another difference is the input to both methods; Novelty TAGM uses the data directly, whereas *phenoDisco* takes the top principal components (by default, the first two) as input. *PhenoDisco* also requires an additional parameter - the minimum group size. This parameter can be challenging to specify, since there is a trade-off between identifying functionally relevant phenotypes of different sizes and picking up small spurious protein clusters. Furthermore, *phenoDisco* struggles to scale to many of the datasets presented in this manuscript, because it requires iteratively refitting models and building of an outlier test statistic.

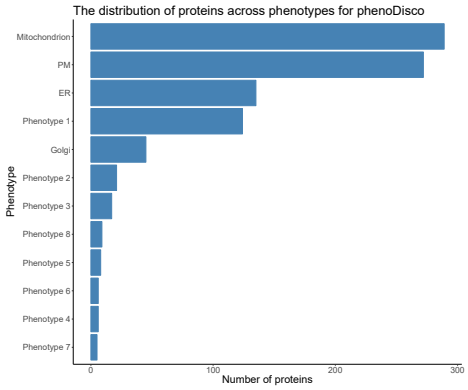
To demonstrate the differences between the two approaches, we apply *phenoDisco* and Novelty TAGM to the HEK-293 spatial proteomics dataset interrogated by [43]. The PCA plots in figure 5.8 reveal broad similarities in the location of the discovered phenotypes. Novelty TAGM provides more information than *phenoDisco*; for example, we can scale the pointer size to the discovery probability. We note that both methods reveal 8 putative phenotypes in the data. Figure 5.8 (panels d and e) reveals the distribution of proteins across these phenotypes. We conclude that both approaches are able to discover small and large clusters, with both methods identifying phenotypes with a few proteins, but also phenotypes with greater than 100 proteins. Figure 5.8 (panel f) shows that both methods find the same number of phenotypes; however, not all of these phenotypes are functionally enriched. For *phenoDisco*, four of the phenotypes had at least 1 significant Gene Ontology term, whereas this was true for five of the Novelty TAGM phenotypes. Figure 5.8 (panel g) characterises the protein overlap between the two approaches. We see that both methods are in broad agreement, with most of the disagreement attributed to cases where one method assigns a protein as unknown whilst the other allocates to it a phenotype or organelle. For example, Novelty TAGM associates *phenoDisco* phenotype 3, which is a lysosome-enriched phenotype, with the plasma membrane (albeit with low probability). On the other hand, Novelty TAGM phenotypes 2 and 3, enriched for chromatin and ribosome respectively, are associated with the mitochondria by *phenoDisco*. This demonstrates the ability of Novelty TAGM to derive more biologically meaningful phenotypes.



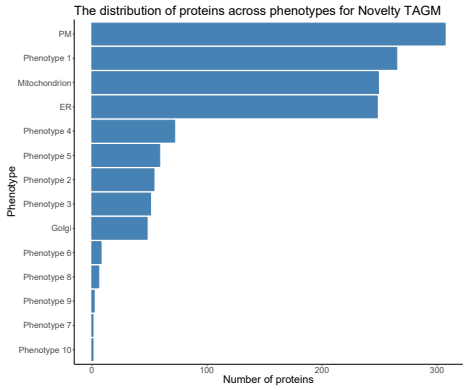
(a)

(b)

(c)



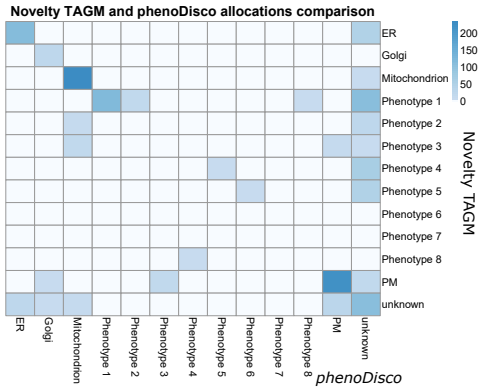
(d)



(e)

	Phenotypes	Enriched
phenoDisco	8	5
Novelty Tagm	8	6

(f)



(g)

Fig. 5.8 (a) PCA plot showing marker proteins for the HEK-293 dataset. (b) PCA plot with phenotypes identified by *phenoDisco*. (c) PCA plot with phenotypes identified by Novelty TAGM with pointer size scaled to discovery probability. (d, e) Barplots showing the number of proteins allocated to different phenotypes by *phenoDisco* and Novelty TAGM respectively. (f) A table demonstrating the number of phenotypes with functional enrichment for both methods and the number of phenotypes discovered. (g) A heatmap showing the overlap between *phenoDisco* and Novelty TAGM allocations.

5.6 Improved annotation allows exploration of endosomal processes

Given the information that the U-2 OS *hyper*LOPIT dataset resolves an endosomal cluster not previously explored, we perform a re-analysis of this dataset focusing on the endosomes. We curate a set of marker proteins for the endosomes and add these annotations to the U-2 OS *hyper*LOPIT dataset. After which, we apply our Bayesian generative classifier TAGM to the data with this additional annotation. Protein allocations to each sub-cellular niche are visualised in the PCA plot of figure 5.9 (panel a). Figure 5.9 (panel c) demonstrates the increased number of proteins that can be characterised by improved annotation of the U-2 OS cell dataset. Furthermore, we examine 7 (of 240) proteins with uncertain endosomal localisation, which can be visualised in each of the violin plots in figure 5.9 (panel d).

All 7 proteins with uncertain assignment to our new endosome cluster are known to function in endosome dynamics. Rab5a and Rab5b (P20339; P61020) are isoforms of Rab5, a small GTPase which is considered a master organiser of the endocytic system, regulating clathrin-mediated endocytosis and early endosome dynamics [425, 484, 493, 381, 308, 64, 158, 262]. RN-tre (Q92738) is a GTPase-activating protein which controls the activity of several Rab GTPases, including Rab5, and is therefore a key player in the organisation and dynamics of the endocytic pathway [256, 158]. KIF16B (Q96L93) is a plus end-directed molecular motor which regulates early endosome motility along microtubules. It is required for the establishment of the steady-state sub-cellular distribution of early endosomes, as well as the balance between PM recycling and lysosome degradation of signal transducing cell surface receptors including EGFR and TfR [203, 58]. Notably, it has been demonstrated that KIF16B co-localises with the small GTPase Rab5, whose isoforms Rab5a and Rab5b we also identified as potentially localised to the endosome and PM in this dataset. ZNRF2 (Q8NHG8) is an E3 ubiquitin ligase which has been shown to regulate mTOR signalling as well as lysosomal acidity and homeostasis in mouse and human cells and has been detected at the endosomes, lysosomes, Golgi apparatus and PM according to the literature [12, 207]. Ykt6 (O15498) is a SNARE (soluble N-ethylmaleimide-sensitive factor attachment protein receptor) protein that regulates a wide variety of intracellular trafficking and membrane tethering and fusion processes. The membrane-associated form of Ykt6 has been detected at the PM, ER, Golgi apparatus, endosomes, lysosomes, vacuoles (in yeast), and autophagosomes as part of various SNARE complexes [111, 445, 144, 304, 446, 296, 272, 488]. In line with this, our results show a mixed sub-cellular distribution for Ykt6 with potential localisation to the endosome and cytosol (figure 5.9, panel d). EHD3 (Q9NZN3) is an important regulator of endocytic trafficking and recycling, which promotes the biogenesis and stabilisation of tubular recycling endosomes by inducing early endosome membrane bending and tubulation [15, 196]. We observe a mixed steady-state

potential localisation to the endosome and PM for EHD3 (figure 5.9, panel d). This is in agreement with EHD3's role in recycling endosome-to-PM transport [331, 332, 167, 54, 196].

Of these 7 proteins with uncertain endosome assignment, only 4 have localisations annotated in HPA (figure 5.9 (b)). The HPA assigns Rab5b to the vesicles which, in this context, include the endosomes, lysosomes, peroxisomes and lipid droplets. Therefore, a more precise annotation is available using Novelty TAGM. Ykt6 is localised to the cytosol, in support of our observations. EHD3 has approved localisation to the plasma membrane, again in agreement with our assignments. KIF16B is assigned to the mitochondrion, which contradicts our findings as well as previously published literature on the localisation and biological role of this protein. We speculate that this disagreement arises from the uncertainty associated with the specificity of the chosen antibody [455]. Thus, Novelty TAGM enables sub-cellular fractionation-based methods to identify proteins in sub-cellular niches which cannot be fully interrogated by immunocytochemistry.

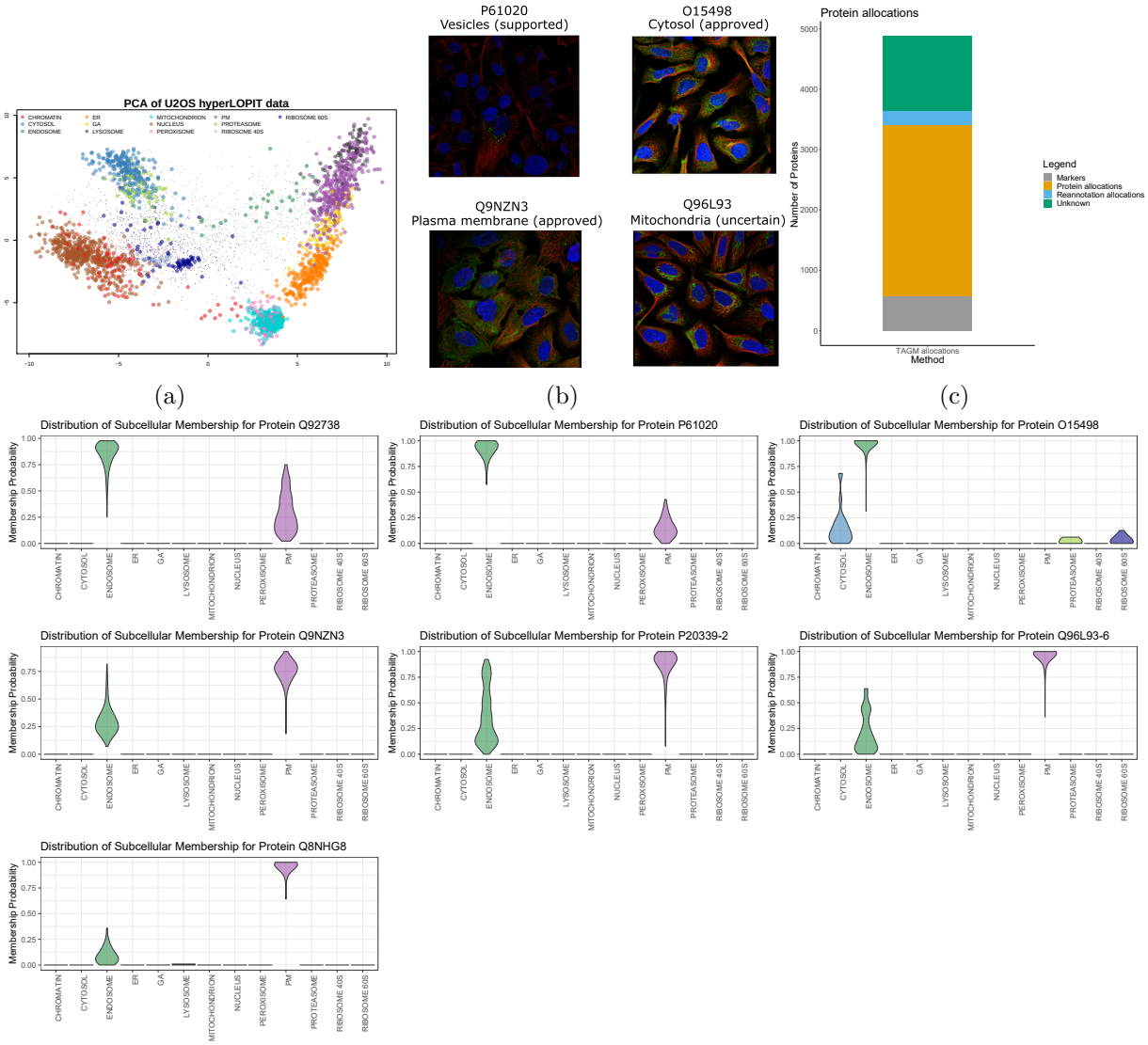


Fig. 5.9 (a) PCA of U-2 OS *hyper*LOPIT data with pointer scaled to localisation probability and outliers shrunk. Points are coloured according to their most probable organelle. (b) Immunofluorescence images and sub-cellular localisation annotation taken from the HPA database (https://www.proteinatlas.org/humanproteome/cell) for the proteins with UniProt accessions P61020 (Rab5b), O15498 (Ykt6), Q9NZN3 (EHD3), and Q96L93 (KIF16B). The nucleus is stained in blue; microtubules in red, and the antibody staining targeting the protein in green. (c) A barplot representing the number of proteins allocated before and after re-annotation of the endosomal class. (d) Violin plots of full probability distribution of proteins to organelles, where each violin plot is for a single protein.

5.7 Discussion and limitations

In this chapter, we developed a semi-supervised Bayesian approach that simultaneously allows probabilistic allocation of proteins to organelles, detection of outlier proteins, as well as the discovery of novel sub-cellular structures. Our method unifies several approaches present in the literature, combining the ideas of supervised machine learning and unsupervised structure discovery. Formulating inference in a Bayesian framework allows for the quantification of uncertainty; in particular, the uncertainty in the number of newly discovered annotations.

To demonstrate the broad applicability of our method, we applied it to 10 different spatial proteomic datasets acquired using diverse fractionation and MS data acquisition protocols and displaying varying levels of resolution revealed additional annotation in every single dataset. Our analysis recovered the chromatin-associated protein phenotype and validated experimental design for chromatin enrichment in *hyper*LOPIT datasets. Our approach also revealed additional sub-cellular niches in the mESC *hyper*LOPIT and U-2 OS *hyper*LOPIT datasets.

Our method revealed resolution of 4 sub-nuclear compartments in the U-2 OS *hyper*LOPIT dataset, which were validated by Human Protein Atlas annotations. An additional endosome-enriched phenotype was uncovered and Novelty TAGM robustly identified an overlapping phenotype in U-2 OS LOPIT-DC data, providing strong evidence for endosomal resolution. Further biologically relevant annotations were uncovered in these, as well as other datasets. For example, a group of vesicle-associated proteins involved in transport from the ER to the early Golgi was identified in the yeast *hyper*LOPIT dataset; resolution of the ribosomal subunits was identified in the fibroblast dataset, and separate nuclear, cytosolic and ribosomal annotations were identified in the DOM datasets.

A direct comparison with the state-of-the-art approach *phenoDisco* demonstrates clear differences between the approaches. Novelty TAGM, a fully Bayesian approach, quantifies uncertainty in both the number of newly discovered phenotypes and the individual protein-phenotype associations - *phenoDisco* provides no such information.

Improved annotation of the U-2 OS *hyper*LOPIT data allowed us to explore endosomal processes, which have not previously been considered with this dataset. We compare our results directly to immunofluorescence microscopy-based information from the HPA database and demonstrate the value of orthogonal spatial proteomics approaches to determine protein sub-cellular localisation. Our results provide insights on the sub-cellular localisation of proteins for which there is no information in the HPA Cell Atlas database.

During our analysis, we observed that the posterior similarity matrices have potential sub-clustering structures. Many known organelles and sub-cellular niches have sub-compartmentalisation, thus methodology to detect these sub-compartments would be desirable. Furthermore, we

have observed that different experiments and different data modalities provide complementary results. Thus, integrative approaches to spatial proteomics analysis are also desired.

Our method is widely applicable within the field of spatial proteomics and builds upon state-of-the-art approaches. The computational algorithms presented here are disseminated as part of the Bioconductor project [166, 212] building on MS-based data structures provided in [153] and are available as part of the pRoloc suite, with all data provided in pRolocdata [156].

There are a number of further directions that our work could take. For example, Breckels et al. [44] develop a multiple kernel SVM method and demonstrate that this can improve classifications. An integrative Bayesian approach to quantify uncertainty is certainly desirable and we could extend TAGM in this direction, for example using multiple dataset integration [243]. Furthermore, new experimental designs are becoming available for spatial proteomics so that data are collected in a control and treatment setting. There are no bespoke methods to analyse these data. In a more statistical direction, we have not explicitly encoded known prior information about the correlation structure present in the data. Indeed, a multivariate Gaussian distribution ignores that the fractions have a particular ordering according to the density gradient. To include this information, we could appeal to Gaussian processes. A more bespoke model for these data are developed in next chapter.

Chapter 6

Semi-supervised non-parametric Bayesian modelling of spatial proteomics

6.1 Motivation

Previous chapters have developed Bayesian approaches in the analysis of spatial proteomics data and we have developed approaches to reduce our reliance on marker proteins. Uncertainty quantification has allowed us to make powerful insights and important advances in the analysis of spatial proteomics data. We now consider a more subtle statistical question: can we develop a model that better reflects the data generating process of the data? In this chapter, we consider functional data analysis approaches to analyse spatial proteomics data. This chapter is an edited version of Crook et al. [86] and there is significant textual overlap.

6.1.1 Abstract

Understanding sub-cellular protein localisation is an essential component in the analysis of context specific protein function. Recent advances in quantitative mass-spectrometry (MS) have led to high resolution mapping of thousands of proteins to sub-cellular locations within the cell. Novel modelling considerations to capture the complex nature of these data are thus necessary. We approach analysis of spatial proteomics data in a non-parametric Bayesian framework, using mixtures of Gaussian process regression models. The Gaussian process regression model accounts for correlation structure within a sub-cellular niche, with each mixture component capturing the distinct correlation structure observed within each niche. Proteins with *a priori* labelled locations motivate using semi-supervised learning to inform the Gaussian process hyperparameters. We moreover provide an efficient Hamiltonian-within-Gibbs

sampler for our model. Furthermore, we reduce the computational burden associated with inversion of covariance matrices by exploiting the structure in the covariance matrix. A tensor decomposition of our covariance matrices allows extended Trench and Durbin algorithms to be applied to reduce the computational complexity of inversion and hence accelerate computation. We provide detailed case-studies on *Drosophila* embryos and mouse pluripotent embryonic stem cells to illustrate the benefit of semi-supervised functional Bayesian modelling of the data.

6.2 Introduction and literature review

Throughout this thesis we have already demonstrated the value of spatial proteomics and the importance of quantification of uncertainty in these experiments. This chapter is more statistical in flavour and considers the challenging task of developing a model that more accurately reflects the data generating process. We begin by revisiting the mechanisms for data generation and an overview of a typical spatial proteomics experiment is provided in Figure 6.1A.

We recall that cells are first gently lysed to expose the cellular content while preserving the integrity of the organelles. The cellular content is then separated using, for example, differential centrifugation [220, 159, 351] or equilibrium density centrifugation [118, 119, 68], among others [356, 194]. After centrifugation, the cellular content is then fractionated, and the abundance of each protein in each fraction is determined experimentally using high accuracy mass-spectrometry. This gives, for each protein, an abundance profile across the fractions.

In the LOPIT (Localisation of Organelle Proteins by Isotope Tagging) [118, 119, 400] and *hyper*LOPIT [68, 324] approaches, cell lysis is proceeded by the separation of sub-cellular components along a continuous density gradient based on their buoyant density. Discrete fractions along this gradient are then collected, multiplexed using tandem mass tags (TMT) [453] and protein distributions revealing organelle specific correlation profiles within the fractions are achieved using synchronous precursor selection mass-spectrometry (SPS-MS^3).

In work that contributed to the discovery of previously unknown organelles and the award of a Nobel prize, de Duve and colleagues [120, 98, 35] observed that proteins belonging to the same organelle possessed very similar abundance profiles (Figure 6.1B). This motivates the following data analysis problem: given the abundance profiles of proteins that are already known to localise to a particular organelle, can we determine which other proteins might also localise to that organelle? In many previous analyses, this problem has been addressed as a black-box classification problem.

The classification approach has a number of major limitations. For example, it implicitly assume that all proteins can be robustly assigned to a primary location, which will often not be the case, since many proteins function in multiple cellular compartments. Other sources of uncertainty include the inherit stochastic processes involved in MS-based quantitation,

as well as each protein's physical properties, which influence how well it is quantified. Post-translation modifications and protein isoforms also add to the challenge of protein quantification. Furthermore, many elements of the experimental procedure are variable and context specific; such as, cell lysis, formation of the density gradients and protein extraction. In addition, organelle integrity maybe disrupted during many of the downstream processing steps. Hence, there are many factors that contribute to the downstream challenge of making protein-niche associations.

We have already developed a generative mixture model of MS spatial proteomics data and, using this model, computed posterior distributions of protein localisation probabilities. However, our model made a number of assumptions that simplified the analysis, but which do not accurately reflect the data generating process. In this chapter, we develop a generative model for the data that is more clearly motivated by the data generating process.

6.2.1 Model development

We proceed by providing a more mathematical description of the data to clarify the modelling task. Let x be the spatial axis along which density gradient separation occurs (see Figure 6.1A), and let $x_1 < x_2$ be two distinct points along x. We assume that the k-th organelle may be characterised by a smooth latent probability density function, $p_k(x)$ (Figure 6.1C), such that, for any protein i that uniquely localises to the k-th organelle, the (unobserved) absolute quantity of protein i in the region $[x_1, x_2]$ after separation is given by:

$$q_k(x_1, x_2) = \int_{x=x_1}^{x_2} p_k(x)\,\mathrm{d}x. \tag{6.1}$$

In a spatial proteomics experiment, quantification occurs in discrete fractions, which we assume to be of approximately the same depth, Δ. Thus, an idealised spatial proteomics experiment would provide us with the quantities $q_k(x_j, x_j + \Delta)$, where $\{x_1, \ldots, x_D\}$ is a grid of spatial coordinates. To simplify notation, we write $q_k(x_j)$ to mean $q_k(x_j, x_j + \Delta)$, i.e. for any protein that uniquely localises to the k-th organelle, $q_k(x_j)$ is the absolute quantity of that protein in the fraction spanning the region from x_j to $x_j + \Delta$.

In practice, current spatial proteomics experiments are unable to determine absolute quantities. We assume that the abundances provided by current spatial proteomics experiments can be expressed as a continuous deterministic function, h, of the absolute quantities, such that the measured abundance, $\mu_k(x_j)$ of protein i in the interval from x_j to $x_j + \Delta$ can be expressed as $\mu_k(x_j) = h(q_k(x_j))$; see Figure 6.1D. Since both h and q_k are unknown, we adopt a functional data analysis approach and treat μ_k as an unknown function to be inferred. We learn μ_k using data from proteins whose localisation to organelle k is already known (see Figure 6.1E), and

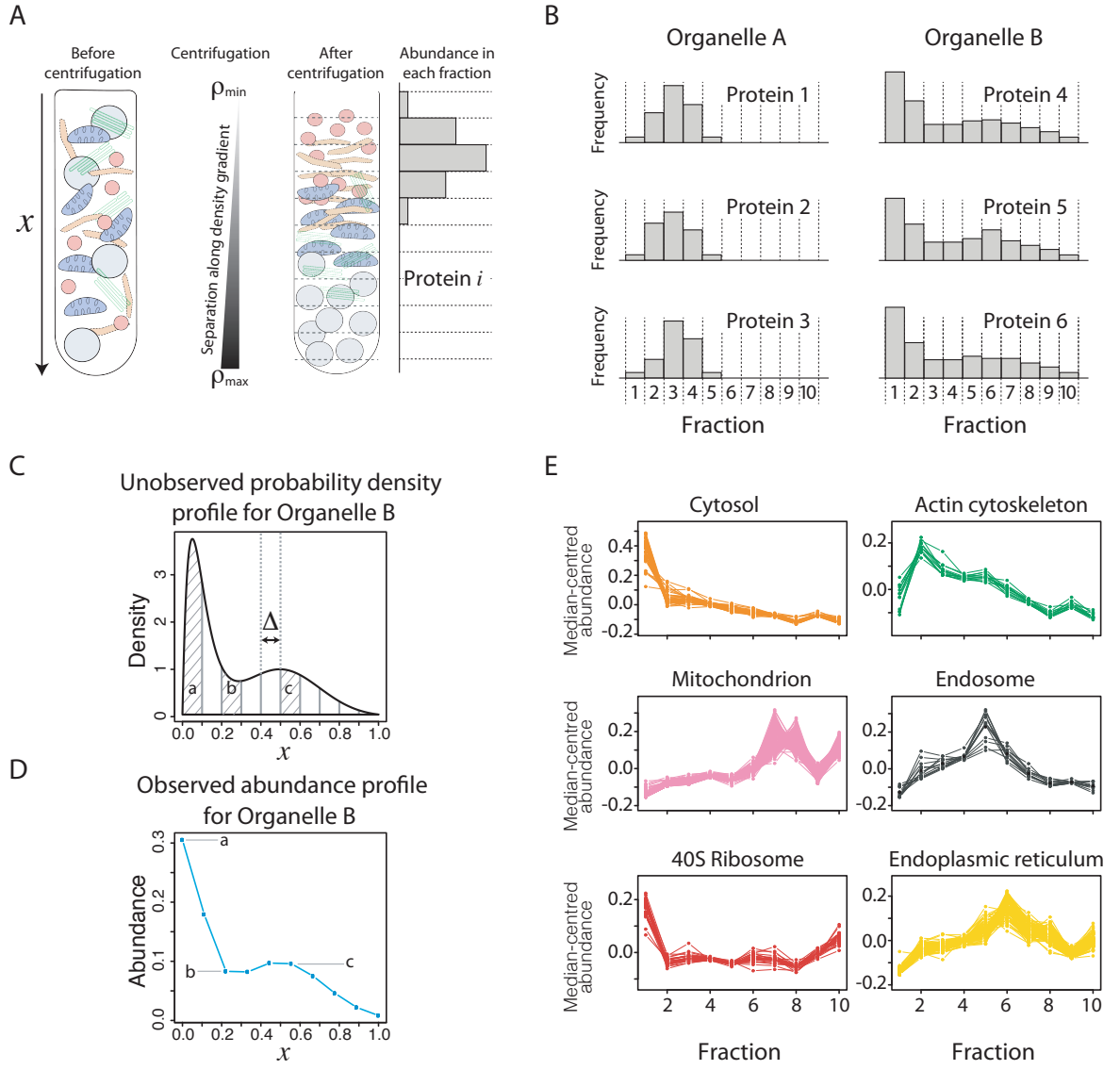


Fig. 6.1 An overview of the experimental design of a spatial proteomics experiment using density-gradient centrifugation. (A) Cellular content is loaded onto a preformed iodixanol density gradient. The tube is then subject to centrifugation, typically at $10^6 g$ for 8 hours. After centrifugation organelles have migrated to their buoyant densities and proteins localised to these organelles will be more abundant in that part of the density gradient. (B) Discrete fractions are collected along the density gradient. Proteins localised to the same organelle share characteristic distributions across the fractions. (C) Organelles are assumed to be characterised by a smooth latent probability density function $p(x)$. Example characteristic probability density shown for organelle B with fractions a, b and c indicated with assumed fixed depth Δ. (D) Observed abundance profile for a protein belonging to Organelle B, after high-accuracy mass-spectrometry. (E) Proteins with *a priori* known localisation are annotated. Proteins from the same sub-cellular niche share the same (median-centred) abundance profiles.

use a semi-supervised approach to further improve the inference of μ_k using data from proteins whose allocations to organelles are unknown a priori (see Section 6.3.6)

6.2.2 Functional data analysis literature

We briefly review functional data analysis tools before introducing those specifically needed for this chapter. Functional data analysis concerns itself with the analysis of data, where the sampled data for each subject is a function [373]. Wang et al. [470] recently reviewed the current major approaches in functional data analysis, including functional principal component analysis [234], functional linear regression [320], functional clustering [225] and functional classification [370]. For classification, the linear discriminant analysis method was extended to the functional setting using splines [224]. Mixture discriminant analysis in the functional setting applied to model bike sharing data was considered by Bouveyron et al. [38], using an functional EM algorithm. Bayesian approaches to functional classification have also been considered; such as, the wavelet based functional mixed model approach [496] and Bayesian variable selection has also been extended to the functional setting [495]. Rodríguez et al. [391] use dependant Dirichlet processes in the non-parametric Bayesian setting to cluster functional data. The Gaussian process approach to analysing functional data in biomedical applications is extensive [205, 276, 433, 235, 195, 458]

We assume each quantitative protein profile can be described by some unknown function, with the uncertainty in this function captured using a *Gaussian process (GP) prior*. Each sub-cellular niche is described by distinct density-gradient profiles, which display a non-linear structure with no particular parametric assumption being suitable. The contrasting density-gradient profiles are captured as components in a mixture of Gaussian process regression models. Gaussian process regression models have been applied extensively and we refer to [375] and [376] for the general theory. In molecular biology and functional genomics the focus of many applications has been on expression time-series data, where sophisticated models have been developed [244, 71, 235, 243, 198]. We remark that many of these applications consider unsupervised clustering problems. In contrast, here we have (partially) labelled data (proteins with location known prior to our experiments) and so we may consider semi-supervised approaches. We explore inference of GP hyperparameters in two ways: firstly, an empirical Bayes approach in which the hyperparameters are optimised by maximising a marginal likelihood; secondly, by placing priors over these GP hyperparameters and performing fully Bayesian inference using labelled and unlabelled data.

A number of computational aspects need to be considered if inference is to be applied to spatial proteomics data. The first is that correlation in the GP hyperparameters can lead to slow exploration of the posterior, thus we use Hamiltonian evolutions to propose global moves through our probability space [115] avoiding random walk nature evident in traditional

symmetric random walk proposals [310, 27]. Hamiltonian Monte-Carlo (HMC) has been explored previously for hyperparameter inference in GP regression [479], and here we show that HMC can be up to an order of magnitude more efficient than a Metropolis-Hastings approach. Furthermore, a particular costly computation in our model is the computation of the marginal likelihood (and its gradient) associated with each mixture component, which involves the inversion of a large covariance matrix - even storage of such a matrix can be challenging. We demonstrate that a simple tensor decomposition of the covariance matrix allows application of fast matrix algorithms for covariance inversion and low memory storage [494].

6.3 Methods

6.3.1 Previous methods

The workhorses of non-linear functional regression are spline [470] and Gaussian processes [375], though splines can be seen as a special case of Gaussian processes for a particular choice of kernel [242]. For completeness, we introduce details on Gaussian processes and we follow [376] and [434] in the discussion that proceeds.

Gaussian processes

A Gaussian Process (GP) is a continuous stochastic process such that any finite collection of these random variables is jointly Gaussian. A Gaussian process is completely specified by its mean function and covariance function. The mean function $m(x)$ and the covariance function $C(x, x')$ of a real function $f(x)$ is defined as

$$m(x) = \mathbb{E}[f(x)] \tag{6.2}$$

$$C(x, x') = \mathbb{E}[(f(x) - m(x))(f(x') - m(x'))] \tag{6.3}$$

and hence we can write

$$f(x) \sim \mathcal{GP}(m(x), C(x, x')). \tag{6.4}$$

Prediction with Gaussian processes

Let us consider a typical modelling scenario, where we observe noisy observations of a regression function $y = f(x) + \epsilon$, where ϵ is i.i.d Gaussian noise with variance σ^2. Under the prior, the covariance of the noisy observations can be written as

$$\text{cov}(y) = C(X, X') + \sigma^2 I, \tag{6.5}$$

where I is the identity of matrix. Hence, we can write the joint distribution of the observed values of y and the function values at test locations (under the prior)

$$\begin{bmatrix} y \\ f(x^*) \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} C(X, X) + \sigma^2 I & C(X, X^*) \\ C(X^*, X) & C(X^*, X^*) \end{bmatrix}\right) \tag{6.6}$$

Once, we have observed data y at locations X, we can derive the predictive distribution at a test locations as

$$f(x^*)|X, y, X^* \sim \mathcal{N}(\bar{f}(x^*),\ \text{cov}(f(x^*))), \quad \text{where} \tag{6.7}$$

$$\bar{f}(x^*) := \mathbb{E}\left[f(x^*)|X, y, X^*\right] = C(X^*, X)\left[C(X, X) + \sigma^2 I\right]^{-1} y \tag{6.8}$$

$$\text{cov}(f(x^*)) = C(X^*, C^*) - C(X^*, X)\left[C(X, X) + \sigma^2 I\right]^{-1} C(X, X^*) \tag{6.9}$$

Examining the equations carefully, we notice that the mean prediction equation is a linear combination of the observations y. Another way to write this equation would be

$$\bar{f}(x^*) = \sum_{i=1}^{n} \alpha_i C(x_i, x^*), \tag{6.10}$$

where we identify $\alpha = \left[C(X, X) + \sigma^2 I\right]^{-1} y$. The is a rather curious observation. We can write the predictive equation as a finite sum, despite the GP being an infinite dimensional object (and hence requiring an infinite number of basis function to represent it). This result is a consequence of the representer theorem (see theorem 2), which we introduced in Chapter 2.

Covariance functions

We have already introduced kernels in the context of support vector machines. A covariance function is a positive semi-definite symmetric kernel. In most cases, the definition of a kernel and covariance function co-inside. The language difference arises because of the use of covariance and covariance matrix in the study of distributions. We introduce two covariance functions for the sake of brevity. The squared exponential function assumes no prior periodicity nor symmetry and the resultant sample paths are smooth:

$$C(x_i, x_j) = a^2 \exp\left(-\frac{\|x_i - x_j\|_2^2}{l}\right). \tag{6.11}$$

The hyperparameter a^2 is a marginal variance and can be seen to control the amplitude of the sample paths, whilst l is a length-scale parameter and controls the decay of correlations. A popular kernel, other than the squared exponential covariance is the Matérn covariance [434]. We use the following parameterisation of the Matérn covariance [271]:

$$C_v(x_i, x_j) = a^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{8\nu}\frac{\|x_i - x_j\|_2^2}{\rho}\right)^{\nu} \mathcal{K}_v \left(\sqrt{8\nu}\frac{\|x_i - x_j\|_2^2}{\rho}\right), \tag{6.12}$$

where Γ is the gamma function and $\mathcal{K}_v$ denotes the modified Bessel function of the second kind of order $\nu > 0$. Furthermore, a and ρ are positive parameters of the covariance. a^2 is interpreted

as a marginal variance, whilst the non-standard choice of $\sqrt{8\nu}$ in the definition of the matérn covariance, allows us to interpret ρ as a range parameter and thus ρ is the distance at which the correlation is 0.1 for any ν. The parameter ν controls the differentiability of the resulting sample paths; such that, $\lceil v \rceil$ is the number of mean-square derivatives. For typical applications, ν is poorly identifiable and fixed [271]. $\nu = 1/2$ recovers the exponential covariance, whereas taking the limit $\nu \to \infty$ one obtains the squared exponential covariance.

6.3.2 Non-parametric Bayesian modelling

Having establish the background on functional data analysis and Gaussian processes, we can now return to our modelling task. In our experiment, we make discrete observations along a continuous density gradient $\mathbf{y}_i = [y_i(x_1), ..., y_i(x_D)]$, where $y_i(x_j)$ indicates the measurement of protein i in the fraction spanning the spatial region from x_j to $x_j + \Delta$ along the density gradient. We assume that protein intensity y_i varies smoothly with the distance along the density-gradient. We then define the following regression model for the measured abundance of protein i as a function of the spatial coordinate x:

$$y_i(x) = \mu_i(x) + \epsilon_i, \tag{6.13}$$

where μ_i is an unknown deterministic function and ϵ_i a noise variable. We assume that $\epsilon_i \sim_{iid} \mathcal{N}(0, \sigma_i^2)$, for simplicity and remark that more elaborate noise models could be chosen but at additional computational cost and greater model complexity. Proteins are grouped together according to their sub-cellular localisation, with all proteins associated with sub-cellular niche $k = 1, ..., K$ sharing the same regression model; that is, $\mu_i = \mu_k$ and $\sigma_i = \sigma_k$ for all proteins in the k-th sub-cellular niche. For clarity, we refer to sub-cellular structures, whether that be organelles, vesicles or large multi-protein complexes, as *components*. Thus proteins associated with component k can be modelled as *i.i.d* draws from a multivariate Gaussian random variable with mean vector $\boldsymbol{\mu}_k = [\mu_k(x_1), ..., \mu_k(x_D)]$ and covariance matrix $\sigma_k^2 I_D$. To perform inference for the unknown function μ_k, as is typical for spatial correlated data [161, 432], we specify a *Gaussian Process* (GP) prior for each μ_k:

$$\mu_k(x) \sim GP(m_k(x), C_k(x, x')). \tag{6.14}$$

The full complement of proteins is then modelled as a finite mixture of Gaussian process regression models. To elaborate, assuming a GP prior for μ_k means that for indices $x_1, ..., x_D$, the joint prior of $\boldsymbol{\mu}_k = [\mu_k(x_1), ..., \mu_k(x_D)]^T$, is multivariate Gaussian with mean vector $\boldsymbol{m}_k = [m_k(x_1), ..., m_k(x_D)]$ and covariance matrix $C_k(i, j) = C_k(x_i, x_j)$. Given no prior belief about symmetry or periodicity in our deterministic function, we assume our GP is centred with

squared exponential covariance function

$$C_k(x_i, x_j) = a_k^2 \exp\left(-\frac{\|x_i - x_j\|_2^2}{l_k}\right). \tag{6.15}$$

6.3.3 Marginalising the unknown function

Having adopted a GP prior with component specific parameters a_k and l_k for each unknown function μ_k, we let observations associated with component k be denoted by $Y_k = \{y_1, ..., y_{n_k}\}$. Our model tells us that

$$Y_k|\boldsymbol{\mu}_k, \sigma_k \sim \mathcal{N}(\boldsymbol{\mu}_k, \sigma_k^2 I_D). \tag{6.16}$$

Then, we can write this as

$$\begin{aligned} Y_k(x_1), ..., Y_k(x_D)|\mu_k, \sigma_k \sim \\ \mathcal{N}(\mu_k(x_1), ..., \mu_k(x_D), ..., \mu_k(x_1), ..., \mu_k(x_D), \sigma_k^2 I_{n_k D}), \end{aligned} \tag{6.17}$$

where $\mu_k(x_1), ..., \mu_k(x_D)$ is repeated n_k times. Our GP prior tell us

$$\mu_k(x_1), ..., \mu_k(x_D), ..., \mu_k(x_1), ..., \mu_k(x_D)|a_k, l_k \sim \mathcal{N}(0, C_k), \tag{6.18}$$

where C_k is an $n_k D \times n_k D$ matrix. This matrix is organised into $n_k \times n_k$ square blocks each of size D. The $(i,j)^{th}$ block of C_k being A_k, where A_k is the covariance function for the k^{th} component evaluated at $\tau = \{x_1, ..., x_D\}$.

$$C_k = \begin{bmatrix} A_k & A_k & \dots & A_k \\ A_k & A_k & \dots & A_k \\ \vdots & \vdots & \ddots & \vdots \\ A_k & A_k & \dots & A_k \end{bmatrix}. \tag{6.19}$$

Letting $\boldsymbol{\theta}_k = \{a_k, l_k, \sigma_k^2\}$, we can then marginalise μ_k to obtain,

$$Y_k(x_1), ..., Y_k(x_D)|\boldsymbol{\theta}_k \sim \mathcal{N}(0, C_k + \sigma_k^2 I_{n_k D}), \tag{6.20}$$

thus avoiding inference of μ_k. Let $Y_k(\tau)$ denote the vector of length $n_k \times D$ equal to $[y_1(x_1), ..., y_1(x_D), \dots, y_{n_k}(x_1), ..., y_{n_k}(x_D)]$. Then we may rewrite equation 2.44 by marginalising μ_k to obtain:

$$P(z_i = k|z_{-i}) \propto \frac{n_{-i,k} + \alpha/K}{K - 1 + \alpha} \int p(\mathbf{y}_i|\mu_k) p(\mu_k|\boldsymbol{\theta_k}, Y_{-i,k}(\tau))\, \mathrm{d}\mu_k, \tag{6.21}$$

where $Y_{-i,k}(\tau)$ is equal to $Y_k(\tau)$ with observation i removed.

6.3.4 Tensor decomposition of the covariance matrix for fast inference

Our covariance matrix has a particularly simple structure allowing us to exploit extended Trench and Durbin algorithms for fast matrix computations [494]. We are interested in the inversion of matrices of the following form

$$C = \begin{bmatrix} A+\sigma^2 I_D & A & \dots & A \\ A & A+\sigma^2 I_D & \dots & A \\ \vdots & \vdots & \ddots & \vdots \\ A & A & \dots & A+\sigma^2 I_D \end{bmatrix}. \tag{6.22}$$

Note that A is a positive symmetric matrix of size $D \times D$ and furthermore it is Toeplitz (constant diagonal and perisymmetric). Let J_n denote an $n \times n$ matrix of ones. It is clear that we can write C in the following form:

$$C = \sigma^2 I_{nD} + B, \tag{6.23}$$

where

$$B = J_n \otimes A, \tag{6.24}$$

and $\otimes$ denotes the Kronecker (tensor) product.

Let e_n denote a column vector of ones of length n. It is easy to see that $J_n = e_n e_n^T$. Trivially, we can write $A = I_D A$ and this leads to the following factorisation

$$\begin{aligned} B =& \left(e_n e_n^T\right) \otimes (I_D A). \\ =& (e_n \otimes I_D)(e_n^T \otimes A), \end{aligned} \tag{6.25}$$

where the second equality follows from the mixed-product property of the Kronecker product. Observing that $e_n \otimes I_D$ is a matrix of size $nD \times D$, and $e_n^T \otimes A$ is matrix of size $D \times nD$. We thus arrive at the following factorisation:

$$C = \sigma^2 I_{nD} + (e_n \otimes I_D) I_D (e_n^T \otimes A), \tag{6.26}$$

which is in the following form

$$\begin{aligned} C &= M + URV \\ M &= \sigma^2 I_{nD},\ U = (e_n \otimes I_D), \\ R &= I_D,\ V = e_n^T \otimes A. \end{aligned} \tag{6.27}$$

Matrices of this form have a simple formula for their inverse (Woodbury Identity):

$$(M + URV)^{-1} = M^{-1} - M^{-1}U(R^{-1} + VM^{-1}U)^{-1}VM^{-1}. \tag{6.28}$$

In our case R is trivially its own inverse and the inversion of M requires only a single computation. Thus the only challenge is to invert $(R^{-1} + VM^{-1}U)$, which we now consider. Consider the following:

$$\begin{aligned} R^{-1} + VM^{-1}U &= I_D + (e_n^T \otimes A)(\sigma^{-2} I_{nD})(e_n \otimes I_D) \\ &= I_D + \sigma^{-2}(e_n^T e_n) \otimes (AI_D) \\ &= I_D + \sigma^{-2} n \otimes A \\ &= I_D + \sigma^{-2} nA \end{aligned} \tag{6.29}$$

Recall that A is a $D \times D$ Toeplitz matrix and so it is easy to see from the above that $R^{-1} + VM^{-1}U$ is also Toeplitz and hence its inverse may be computed efficiently. Denote this inverse by Z, then it follows from equation 6.28 that we have:

$$\begin{aligned} (M + URV)^{-1} &= \sigma^{-2} I_{nD} - \sigma^{-4}(e_n \otimes I_D)(Z)(e_n^T \otimes A) \\ &= \sigma^{-2} I_{nD} - \sigma^{-4}(e_n \otimes I_D)(e_n^T \otimes ZA) \\ &= \sigma^{-2} I_{nD} - \sigma^{-4}(e_n e_n^T) \otimes (ZA) \\ &= \sigma^{-2} I_{nD} - \sigma^{-4} J_n \otimes (ZA) \\ &= \sigma^{-2} I_{nD} - \frac{1}{n\sigma^2} J_n \otimes (I - Z), \end{aligned} \tag{6.30}$$

where the last line follows from the following computations, denoting $Z^{-1} = Q$:

$$\begin{aligned} & Q = I_D + \sigma^{-2} nA \\ \implies \quad & Q - \sigma^{-2} nA = I_D \\ \implies \quad & Q^{-1}Q - \sigma^{-2} nQ^{-1}A = Q^{-1} \\ \implies \quad & I_D - \sigma^{-2} nQ^{-1}A = Q^{-1} \\ \implies \quad & I_D - Z = \sigma^{-2} nZA \\ \implies \quad & ZA = \frac{(I_D - Z)\sigma^2}{n}. \end{aligned} \tag{6.31}$$

Thus the inversion of C requires only the inversion of a $D \times D$ matrix that can be performed in $O(D^2)$ computations. This should be compared with a naïve inversion of C requiring $O((nD)^3)$ computations, which represents significant savings. We also need the determinant of C and the

calculation is straightforward using an elementary determinant lemma.

$$\begin{aligned}\det(C) &= \det(M + URV) \\ &= \det(R^{-1} + VM^{-1}U)\det(R)\det(M) \\ &= \det(I_D + ((e_n^T \otimes A)(\sigma^{-2} I_{nD})(e_n \otimes I_D)))\det(M) \\ &= (\sigma^2)^{nD}\det(I_D + \sigma^{-2} nA).\end{aligned} \tag{6.32}$$

As before the term in the determinant is Toeplitz and hence the determinant can be computed efficiently. The extended Trench and Durbin algorithms are stated in the appendix.

6.3.5 Sampling the underlying function

Whilst it is often mathematically convenient to marginalise the unknown function μ_k from a computational perspective it is not always advantageous to do so. To be precise, marginalising μ_k induces dependencies among the observations; that is, we cannot exploit the conditional independence structure given the underlying function μ_k. After marginalising, Gibbs moves must be made sequentially for each protein in turn and this can slow down computation.

The alternative approach is to sample the underlying function and exploit conditional independence. Once a sample is obtained from the GP posterior on μ_k, conditional independence allows us to compute the likelihood for all proteins at once, exploiting vectorisation. If there are a particularly large number of observation in each component it is also possible to parallelize computation over the components $k = 1, ..., K$.

6.3.6 Gaussian process hyperparameter inference

Supervised approach: optimising the hyperparameters

Inference of the hyperparameters $\boldsymbol{\theta_k}$ can be dealt with in several ways. The first is to learn them using only the labelled data (i.e. data that pertains to proteins with well documented sub-cellular locations). Using the labelled data for each component constitutes maximise the marginal likelihood of the hyperparameters with respect to the data. These hyperparameters are then fixed throughout the inference of the unlabelled data. The marginal likelihood can be obtained quickly by recalling that

$$Y_k(x_1), ..., Y_k(x_D)|\boldsymbol{\theta_k} \sim N(0, C_k + \sigma_k^2 I_{n_k D}). \tag{6.33}$$

Thus the log marginal likelihood is given by

$$\begin{aligned}&\log p(Y_k|\tau,\boldsymbol{\theta_k}) \\ &\qquad = -\frac{1}{2}Y_k(\tau)\left(C_k + \sigma_k^2 I_{n_k D}\right)^{-1} Y_k(\tau)^T - \frac{1}{2}\log|C_k + \sigma_k^2 I_{n_k D}| - \frac{n_k D}{2}\log 2\pi.\end{aligned} \tag{6.34}$$

For convenience of notation set $\hat{C}_k = C_k + \sigma_k^2 I_{n_k D}$. To maximise the marginal likelihood given equation 6.34, we find the partial derivatives with respect to the parameters [375]. Hence, we can use a gradient based optimisation procedure. Positivity constraints on a_k^2, l_k, σ_k^2 are dealt with by re-parametrisation and so, dropping the dependence on k for notational convenience, and abusing notation, we set $l = \exp(\theta_1)$, $a^2 = \exp(2\theta_2)$ and $\sigma^2 = \exp(2\theta_3)$.

Application of the quasi-Newton L-BFGS algorithm [273] for numerical optimisation of the marginal likelihood with respect to the hyperparameters is now straightforward. The L-BFGS can only find a local optimum and so we initialise over a grid of values. We terminate the algorithm when successive iterations of the gradient are less than 10^{-8}. We make extensive use of high performance R packages to interface with C++ [122, 123].

Semi-supervised hyperparameter inference

The advantage of adopting a Bayesian approach to hyperparameter inference is that we can quantify uncertainty in these hyperparameters. Uncertainty quantification in GP hyperparameter inference is important, since different hyperparameters can have a strong effect on the GP posterior [375]. Furthermore, we consider a semi-supervised approach to hyperparameter inference. By a semi-supervised approach we mean that a posterior distribution for the hyperparameters can be inferred using both the labelled and unlabelled data, rather than just the labelled data.

Consider at some iteration of our MCMC algorithm the data associated to the k^{th} component Y_k. We can partition this data into the unlabelled (U) and labelled data (L); in particular, $Y_k = \left[Y_k^{(L)}, Y_k^{(U)}\right]$. To clarify, the indicators z_i are known for $Y_k^{(L)}$ prior to any inference, whilst allocations z_i for $Y_k^{(U)}$ are sampled at each iteration of our MCMC algorithm. If we believe our labelled data $Y_k^{(L)}$ are true representatives of the distribution of that component, it is computationally advantageous just to consider the labelled data when performing hyperparameter inference. However, there could be a sampling bias in the labelled data and so the labelled data alone is insufficient to explain the variability in the data. A semi-supervised approach allows the posterior distribution of the hyperparamters to reflect the uncertainty in the component allocations z_i and therefore improve our abilities to predict allocations and quantify uncertainty in allocations.

Semi-Supervised approach: hyperparameter inference using MH

In a Bayesian framework, we treat the hyperparameters as random variables and place hyperpriors overs them. Positivity constraints motivate working with the log of the hyperparameters and using, for example, standard normal priors [333]. Unfortunately loss of conjugacy between the prior on the hyperparameters $g_0(\boldsymbol{\theta})$ and the likelihood $f(\mathbf{y}|\boldsymbol{\theta})$ is unavoidable, and hence we use a Metropolis-Hastings step or Hamiltonian Monte-Carlo step for inference. The Metropolis-Hastings sampler can be summarised as follows:

Metropolis-Hastings algorithm with random walk proposals: Suppose θ_t is the most recently sampled value. Sample a value $\xi \sim N(0,1)$, setting $\theta_{t+1} = \theta_t + \xi$ and compute the Metropolis ratio

$$\Lambda = \frac{p(\theta_{t+1}|Y_k(\tau))}{p(\theta_t|Y_k(\tau))} = \frac{p(Y_k(\tau)|\theta_{t+1})p_0(\theta_{t+1})}{p(Y_k(\tau)|\theta_t)p_0(\theta_t)}. \tag{6.35}$$

This ratio can be computed in log form using equation 6.34. Then sample a uniform random number $u \sim U[0,1]$ if $\log(\Lambda) \geq \log(u)$ set $\theta_{t+1} = \theta_t + \xi$, otherwise $\theta_{t+1} = \theta_t$.

Semi-Supervised approach: hyperparameter inference using HMC

To avoid the random walk nature of the MH sampler, we also consider a Hamiltonian Monte-Carlo approach, which exploits the geometry of the space to provide more efficient proposals [115, 206, 336, 174]. In short, Hamiltonian Monte-Carlo allows us to construct Hamiltonian evolutions $H(\boldsymbol{y}, \boldsymbol{p})$ such that the resulting dynamics efficiently explore a target distribution $p(\boldsymbol{y})$. We augment our probability distribution with an auxiliary momentum component $\boldsymbol{p}$.

The Hamiltonian can be decomposed into potential and kinetic energies $H(\boldsymbol{y}, \boldsymbol{p}) = U(\boldsymbol{y}) + K(\boldsymbol{p})$. The canonical distribution is then given by:

$$p(\boldsymbol{y}, \boldsymbol{p}) \propto \exp(-H(\boldsymbol{y}, \boldsymbol{p})) \propto p(\boldsymbol{y})p(\boldsymbol{p}). \tag{6.36}$$

The distribution of momentum component is chosen as a Gaussian distribution with diagonal covariance matrix $M = diag(m_1, ..., m_r)$ and thus the distribution and kinetic energies are given by

$$\begin{aligned} p(\boldsymbol{p}) &= N(0, M) \\ K(\boldsymbol{p}) &= \frac{\boldsymbol{p}M^{-1}\boldsymbol{p}^T}{2} \\ \nabla K &= M^{-1}\boldsymbol{p}. \end{aligned} \tag{6.37}$$

It is easy to see from the canonical distribution that $U(\boldsymbol{y}) = -\log(p(\boldsymbol{y}))$ is the required choice for the potential. In practice, we need to simulate from Hamiltonian dynamics. Hamilton's

equations are given by a coupled system:

$$\begin{aligned} \frac{d\boldsymbol{p}}{dt} &= -\nabla_{\boldsymbol{y}} H(\boldsymbol{y}, \boldsymbol{p}) \\ \frac{d\boldsymbol{y}}{dt} &= \nabla_{\boldsymbol{p}} H(\boldsymbol{y}, \boldsymbol{p}). \end{aligned} \tag{6.38}$$

Such a system is called symplectic and thus a numerical schema which is a symplectic integrator is required to simulate the required dynamics [336]. The leapfrog algorithm is the standard choice [288]. This algorithm does not exactly conserve energy and so a Metropolis accept/reject step is required is remove the induced bias [27]. An MCMC algorithm can then be constructed to sample from the required distribution, where proposals are made using Hamiltonian evolutions. Recall, we are required to simulate the Hamiltonian evolutions. To simulate an evolution over time T, take L steps of size δ such that $L\delta \geq T$. One step of the leapfrog algorithm of size δ for Hamilton's dynamics starting at time t is given by the following

$$\begin{aligned} \boldsymbol{p}(t + \delta/2) &= \boldsymbol{p}(t) - \frac{\delta}{2} \nabla U \boldsymbol{y}(t) \\ \boldsymbol{y}(t + \delta) &= \boldsymbol{y}(t) + \delta \nabla K \boldsymbol{p}(t + \delta/2) \\ \boldsymbol{p}(t + \delta) &= \boldsymbol{p}(t + \delta/2) - \frac{\delta}{2} \nabla U \boldsymbol{y}(t + \delta). \end{aligned} \tag{6.39}$$

We can now summarise the HMC algorithm to sample n samples from a target distribution $p(\boldsymbol{y})$.

1. Set $t = 0$.
2. Sample a position value from the prior $\boldsymbol{y}_0 \sim p_0$.
3. Do until $t = n$
   (a) Set $t = t + 1$.
   (b) Sample an initial momentum variable $\boldsymbol{p}_0 \sim p(\boldsymbol{p})$.
   (c) Set $\boldsymbol{y}_0 = \boldsymbol{y}_{t-1}$.
   (d) Run algorithm 6.39 for L step of size δ and obtain proposal states $\boldsymbol{y}_*$ and $\boldsymbol{p}_*$.
   (e) Compute the Metropolis ratio:

$$\Lambda = \exp(-(U(\boldsymbol{y}_*) + K(\boldsymbol{p}_*)) + (U(\boldsymbol{y}_0) + K(\boldsymbol{p}_0))). \tag{6.40}$$

   (f) Sample $u \sim U[0, 1]$ if $\Lambda > u$ set $\boldsymbol{y}_t = \boldsymbol{y}_*$, else $\boldsymbol{y}_t = \boldsymbol{y}_{t-1}$.

We can now specify the details for sampling the hyperparameters of a Gaussian Process with standard normal hyperpriors. Using a squared exponential covariance function and

re-parametrising, as before, we first specify our target distribution $p(\boldsymbol{y}) = p(\boldsymbol{\theta}|X(\tau)) \propto p(X(\tau)|\boldsymbol{\theta})p_0(\boldsymbol{\theta})$. Now considering

$$U(\boldsymbol{y}) = -\log(p(\boldsymbol{y})) = -\log(p(X(\tau)|\boldsymbol{\theta})) - \log(p_0(\boldsymbol{\theta})) + constant, \tag{6.41}$$

the first term can be computed by marginalising and is recognised as the marginal likelihood. Recalling that we have a standard normal prior the negative log prior and its gradient is given by:

$$\begin{aligned} -\log(p_0(\boldsymbol{\theta})) &= \frac{3}{2}\log((2\pi)) + \frac{\boldsymbol{\theta}\boldsymbol{\theta}^T}{2} \\ \nabla(-\log(p_0(\boldsymbol{\theta}))) &= \boldsymbol{\theta}, \end{aligned} \tag{6.42}$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)$. Hence, we can write down the gradient of the potential energy. We further reintroduce the dependence on k,

$$\begin{aligned} \nabla U(\boldsymbol{y}) = \nabla(-\log(p(\boldsymbol{y}))) =& \frac{1}{2} tr\left(\left(\hat{C}_k^{-1} - \alpha\alpha^T\right)\nabla\hat{C}_k\right) + \boldsymbol{y} \\ & \alpha = \hat{C}_k^{-1} X_k(\tau). \end{aligned} \tag{6.43}$$

Thus we have everything we need to simulate Hamiltonian dynamics to explore our target distribution. In practice, we make a few standard adaptations to the above algorithm as detailed in [336]. We sample δ from a uniform distribution on $\mathcal{U}[a, b]$, as well as using a partial momentum refreshment with parameter α. More specifically, given $\boldsymbol{p}$ from the previous iteration of the HMC algorithm and a sample $n \sim p_0(\boldsymbol{p})$ set $\boldsymbol{p}'$ as

$$\boldsymbol{p}' = \alpha\boldsymbol{p} + (1 - \alpha^2)^{1/2} n. \tag{6.44}$$

In previous sections, we saw we can exploit a tensor decomposition to accelerate computation of the likelihood and similar formulae are available to accelerate computation of the gradient for use in L-BFGS and Hamiltonian Monte Carlo. These formulae can be found in appendix.

Hyperpriors for Gaussian process hyperparameters

The hyperpriors for the Gaussian process hyperparameters a^2, l, σ^2 need careful attention and the challenge of selecting them is well documented [26, 357, 103, 463, 143]. The values of the hyperparameters have a strong effect of the resultant sample paths of the Gaussian process and, in particular, a ridge in the marginal likelihood means different hyperparameters lead to unconditional prior simulations with the same spatial pattern but different scales [376, 143]. A number of priors are possible and we opt for log-normal priors, since they satisfy positivity constraints, are flexible, and allow for the encoding of expert knowledge [333, 376, 243]. We

choose log-normal priors with mean 0 and variance 1 in our analysis, since they provide the correct scale, whilst remaining somewhat vague (simulations from the prior predictive distribution led to infrequent extreme expression values: $p_0(|y| < 4) \approx 0.97$). The sharp left tail of these priors penalises small values, and the right tails allow large length-scales and variances - if supported by the data. If we desire a more informative prior the hyperprior mean could be selected using the labelled data. Later, our sensitivity analysis shows that changing the hyperprior mean has little effect on predictive performance. However, overly precise choices for the variance on unmotivated locations can lead to poor results.

An overview of the MCMC algorithm for posterior Bayesian computation

In our model $g_0(\boldsymbol{\theta})$ and $f(\mathbf{y}|\boldsymbol{\theta})$ are non-conjugate, which means the integral in equation 6.21 cannot be obtained analytically. A Gibbs sampling scheme with either an additional Metropolis-Hastings or Hamiltonian Monte Carlo update is used. Each iteration of the MCMC algorithm includes a sampled value for the component indicators, outlier components and current values of the hyperparameters. We also keep track of associated posterior probabilities and marginal likelihoods as appropriate. Furthermore, we can sample the hyperparameters every T iterations of the MCMC algorithm to accelerate computations.

6.3.7 Summary of Bayesian non-parametric model

We can reuse much of the machinery we have developed in previous chapters. Indeed, our non-parametric likelihood model can be integrated into our Robust mixture modelling framework seamlessly. Hence, the outlier components and summarisation of the model remains the same as earlier chapters. Furthermore, the quadratic loss (Brier score) is still used to compare model predictions.

6.4 Results

6.4.1 Case Study I: *Drosophila melanogaster* embryos

Application

The first case study is the *Drosophila melanogaster* (common fruit fly) embryos [448], in which we compare the supervised and semi-supervised approaches for updating the model hyperparameters. In particular, we explore the effect on the component specific noise term σ^2, by adopting different inference approaches. For each sub-cellular niche, we learn the hyperparameters by either maximising their marginal likelihood or sampling from their posterior using MCMC. The posterior distribution for the hyperparameters can either be found solely using the labelled data for each component or by making use of labelled and unlabelled data.

Figure 6.2 demonstrates several phenomena. Reassuringly, the estimates of the noise parameters σ_k^2 for $k = 1, ..., K$ obtained by using the L-BFGS algorithm to maximise the marginal likelihood coincide with the posterior distributions of the noise parameters, inferred using only the labelled data for each component. However, when we perform inference in a semi-supervised way, by using both the labelled and unlabelled data to make inferences, we make several important observations.

Firstly, in many cases, the posterior using both the labelled and unlabelled data is shifted right towards 0. Recalling that we are working with the log of the hyperparameters, this indicates that the noise parameters is smaller when solely using the labelled data. This is likely a manifestation of experimental bias, since it is reasonable to believe that proteins with known prior locations are those which have less variable localisations and are therefore easier to experimentally validate. A semi-supervised approach is able to overcome these issues, by adapting to proteins in a dense region of space. In some cases the shift is pronounced, with posteriors of the parameters using labelled and unlabelled data found in the tails of the posterior only using the labelled distribution. Furthermore, we notice shrinkage in the posterior distribution of the noise parameter in the semi-supervised setting. The reduction in variance reduces our uncertainty about the underlying true value of σ_k^2 for $k = 1, ..., K$. This variance reduction is observed in most cases even when these is little difference in the mean of the posteriors.

The primary goal of spatial proteomics is to predict the localisation of unknown proteins from data. Our modelling approach allows the allocation probability of each protein to each component to be used to predict the localisation of unknown proteins. Proteins may reside in multiple locations and some sub-cellular niches are challenging to separate because of confounding biochemical properties, leading to uncertainty in a proteins localisation. Thus adopting a Bayesian approach and quantifying this uncertainty is of great importance. Our methods allow point-estimates as well as interval estimates to be obtained for the posterior localisation probabilities. Figure 6.3 demonstrates the results of applying our method. Each protein in this PCA plot is scaled according to mean of the Monte-Carlo samples from the posterior localisation probability. To visualise the allocation probabilities for proteins across organelles, we produce a heatmap, M, where the $(i, j)^{th}$ entry of M is the Monte-Carlo estimate of the allocation probability of the i^{th} protein to organelle j (figure 6.4).

Further visualisation of the model and data are possible. We plot two representative example of gradient-density profiles for two components the endoplasmic reticulum (ER) and the nucleus, in figure 6.5. We plot both the labelled proteins, in colour, which were assigned to each component before our analysis. In grey, for both components, we plot the unlabelled proteins which have been allocated to these components probabilistically. We observe that they have the same gradient-density shape as the labelled proteins - in line with our beliefs about

the underlying biology: that proteins from the same components should co-fractionate and therefore have similar density gradient profiles. In addition, we overlay the posterior predictive distribution for these components and observe they represent the data well.

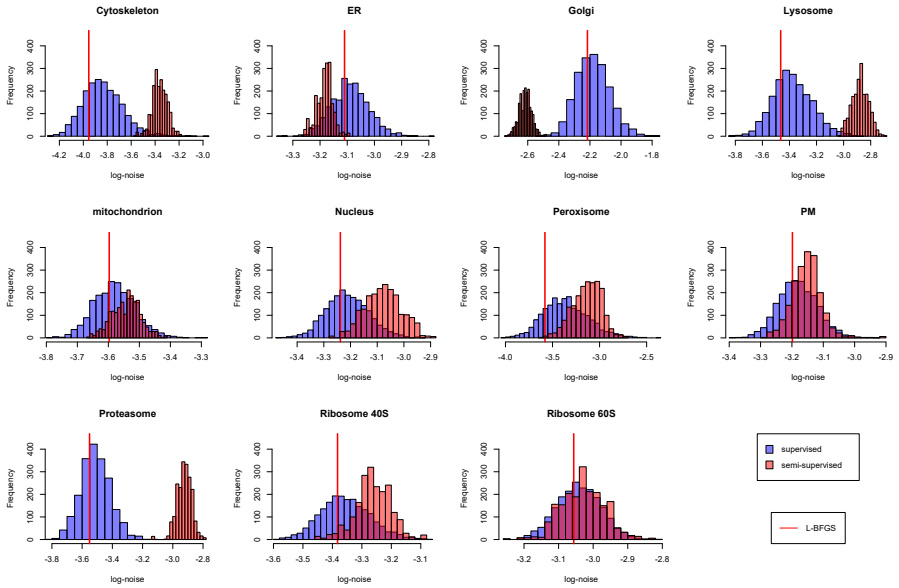


Fig. 6.2 Posterior distributions for the log noise parameter σ^2 on the *Drosophila* data. In general, we observe a shift towards 0, indicating that the labelled data underestimates the value of the noise term for each component. We also observe increased posterior shrinkage for many components with the variance of the noise parameters reduced in the semi-supervised setting.

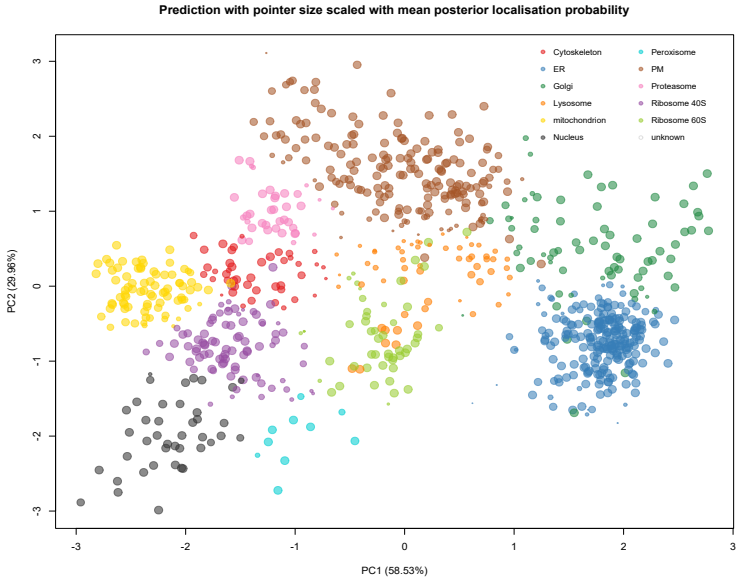


Fig. 6.3 A pca plot for the *Drosophila* data where points, representing proteins, are coloured by the component of greatest probability. The pointer for each protein is scaled according to membership probability with larger/smaller points indicating greater/lower allocation probabilities.

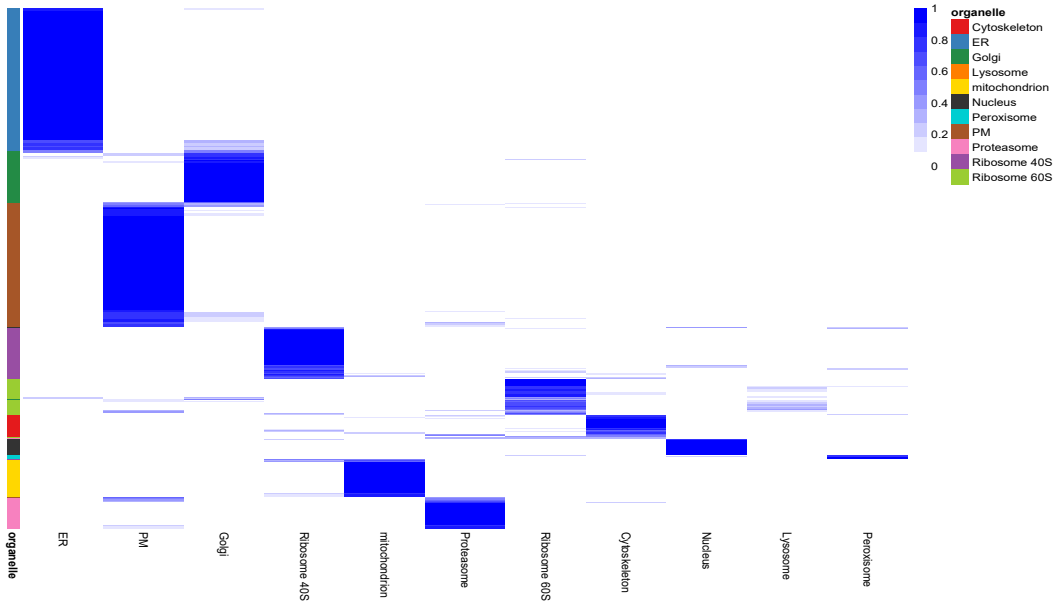
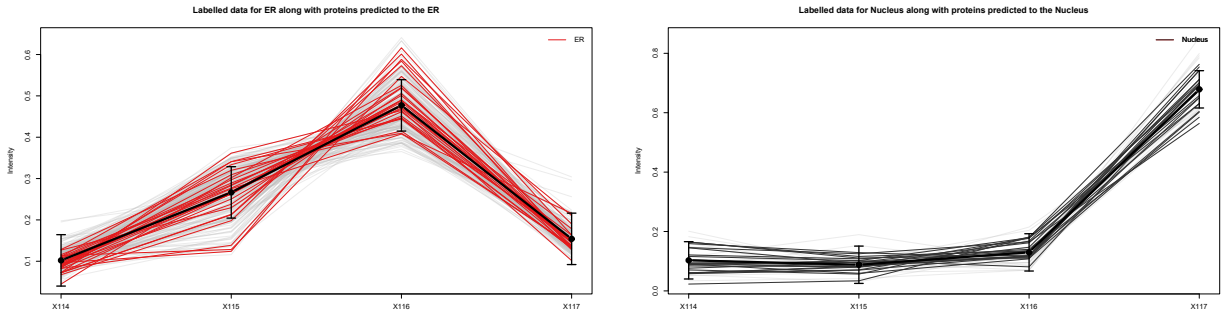


Fig. 6.4 A heatmap of organelles by proteins, where the $(i, j)^{th}$ entry is the Monte-Carlo estimate of the probability that a protein i belongs to organelle j. Allowing us to visualise the range of probabilities for each protein. Proteins are allocated to their most probable class and these allocations are shown in the colour bar on the left.



(a) (b)

Fig. 6.5 A plot of the gradient-density profiles for the ER and Nucleus with labelled proteins in colour and protein probabilistically assigned to those components in grey. The profiles of the assigned proteins closely match the profiles of the components. The predictive posterior of these components is also overlayed.

Sensitivity analysis for hyper-prior specification

We use the *Drosophila melanogaster* dataset to test for sensitivity of the hyper-prior specification. To test for sensitivity, we see if predictive performance is affected by changes in the choice of hyper-prior. The following cross-validation schema assesses whether predictive performance is affected by choice of hyper-prior. We split the labelled data for each experiment into class-stratified training (80%) and test (20%) partitions, with the separation formed at random. The true classes of the test profiles are withheld from the classifier, whilst MCMC is performed. This 80/20 data stratification is performed 100 times in order produce a distribution of scores. We compare the ability of the methods to probabilistically infer the true classes using the quadratic loss, also referred to as the Brier score [176]. Thus a distribution of quadratic losses is obtained for each method, with the preferred method minimising the quadratic loss. Each method is run for $10,000$ MCMC iterations with 1000 iterations for burn-in. We vary the mean of the standard normal hyper-prior for each hyperparameter in turn for a grid of values $\tilde{m} = (0, -1, -2, -3, -4)$, keeping the hyper-prior for the other variables held the same as a standard normal distribution. The results are displayed in figure 6.6.

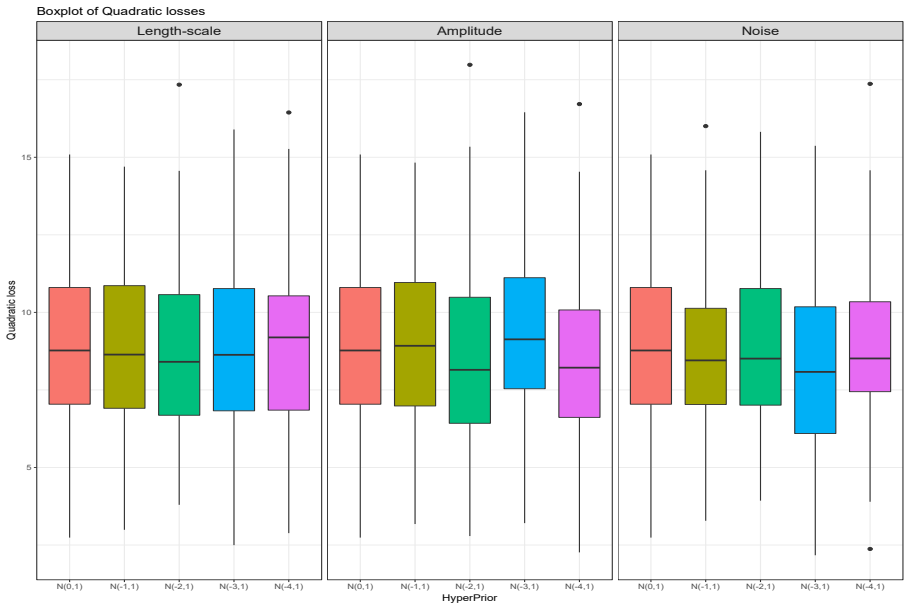


Fig. 6.6 Boxplots of quadratic losses to assess the sensitivity of semi-supervised hyperparameter inference to hyper-prior choices.

We observe only minor sensitivity to the choice of hyper-prior, with no significant difference in performance noted (KS test, threshold = 0.01). Sensitivity analysis for hyperparameters of GPs is vital, since these hyperparameters have a strong effect on the posterior of the GP [375]. The observed lack of sensitivity in our case is advantageous, since prior information can be included without fear of over fitting. However, practitioners should always take care when

specifying priors, especially for variance/covariance parameters as many authors have noted sensitivity of Bayesian models to these parameters [162, 287, 165, 471, 410].

6.4.2 Case Study II: mouse pluripotent embryonic stems cells

Application

Our main case study is the mouse pluripotent E14TG2a stem cell dataset of [68]. This dataset contains 5032 quantitative protein profiles, and resolves 14 sub-cellular niches. We first plot the density-gradient profiles of the marker proteins for each sub-cellular niche in figure 6.7. We fit a Gaussian process prior regression model for each sub-cellular niche with the hyperparameters found by maximising the marginal likelihood. A table of unconstrained log hyperparameter

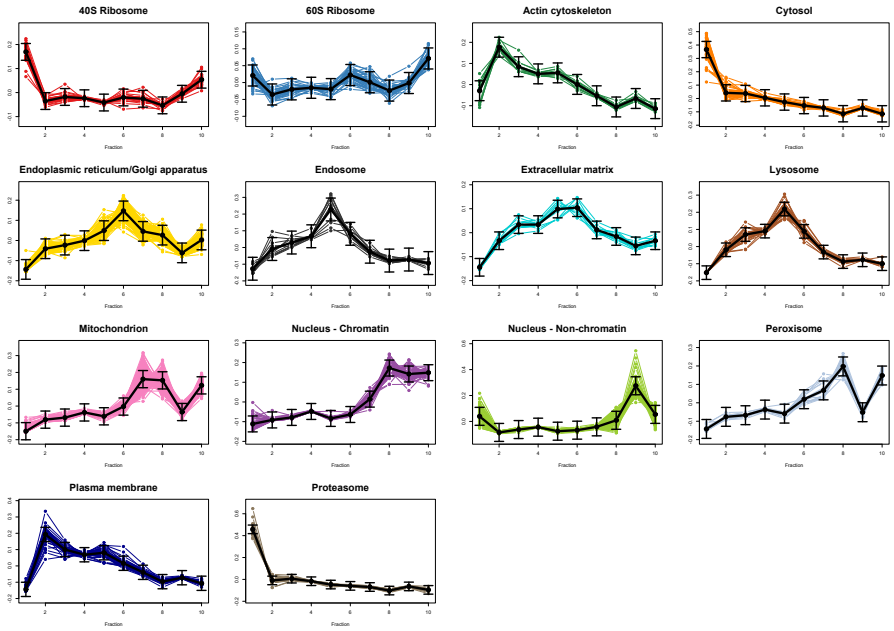


Fig. 6.7 Quantitative profiles of protein markers for each sub-cellular niche. A GP prior regression model is fitted to these data and the predictive distribution is displayed. We observe distinct distributions for each sub-cellular niche generated by the unique density-gradient properties of each sub-cellular niche.

values found by maximising the marginal likelihood in the appendix. Alternatively, placing standard normal priors on each of the log hyperparameters and using a Metropolis-Hastings update we can infer the distributions over these hyperparameters. We perform $20,000$ iterations for each sub cellular niche and discard $15,000$ iterations for burn-in and proceed to thin the remaining samples by 20. We summarise the Monte-Carlo sample by the expected value as well as the 95% equi-tailed credible interval, which can also be found in the appendix.

We go further to predict proteins with unknown localisation to annotated components using our proposed mixture of GP regression models. As before, we adopt a semi-supervised approach to hyperparameter inference. Again we place standard normal hyper-priors on the log of the hyperparameters. We run our MCMC algorithm for $20,000$ iterations with half taken as burnin and thin by 5, as well as using HMC to update the hyperparameters. The PCA plot in figure 6.8 visualises our results. Each pointer represent a single protein and is scaled either to the probability of membership to the coloured component (figure 6.8) or scaled with the Shannon entropy (figure 6.9) . As before, we also visualise the allocation probabilities for proteins across organelles in a heatmap (figure 6.10). In these plots we observe regions of high-probability and confidence to each organelle, as well as obtaining a global view of uncertainty. In this example, we observe regions of uncertainty, as measured by the Shannon entropy, concentrating where components overlap. We also observe uncertainty in regions where there is no dominant component. This Bayesian analysis provides a wealth of information on the global patterns of protein localisation in mouse pluripotent embryonic stem cells.

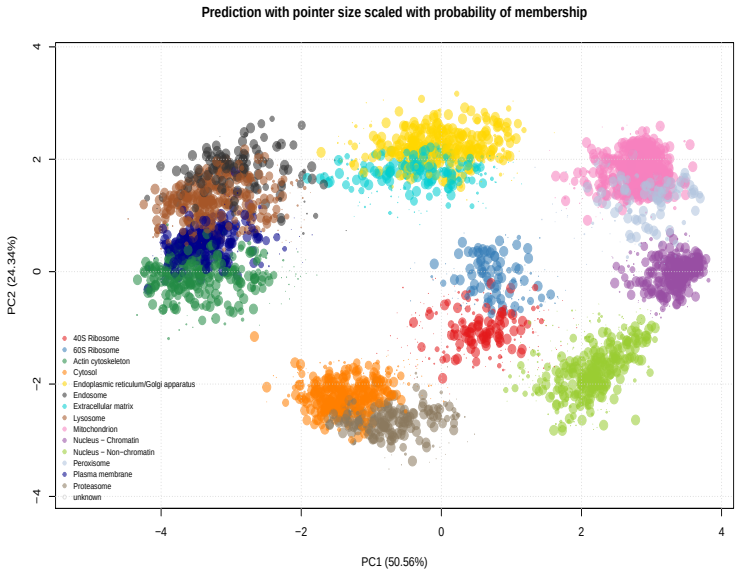


Fig. 6.8 A pca plot for the mouse pluripotent embryonic stem cell data where points, representing proteins, are coloured by the component of greatest probability. The pointer for each protein is scaled with membership probability.

6.4.3 Assessing predictive performance

We compare the predictive performance of the methods proposed here, as well as against the fully Bayesian TAGM model of [83], where sub-cellular niches are described by multivariate Gaussian distributions rather than GPs. The following cross-validation schema is used to

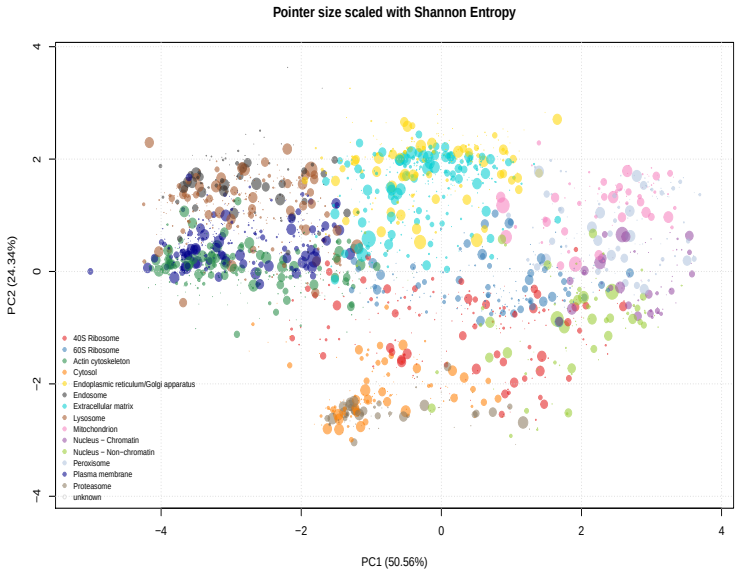


Fig. 6.9 A pca plot for the mouse pluripotent embryonic stem cell data where points, representing proteins, are coloured by the component of greatest probability. The pointer for each protein is scaled with the Monte-Carlo averaged Shannon Entropy.

compare the classifiers. We split the labelled data for each experiment into class-stratified training (80%) and test (20%) partitions, with the separation formed at random. The true classes of the test profiles are withheld from the classifier, whilst MCMC is performed. This 80/20 data stratification is performed 100 times in order produce a distribution of scores. We compare the ability of the methods to probabilistically infer the true classes using the quadratic loss, also referred to as the Brier score [176]. Thus a distribution of quadratic losses is obtained for each method, with the preferred method minimising the quadratic loss. Each method is run for $10,000$ MCMC iterations with 1000 iterations for burn-in. For fair comparison we held priors the same across all datasets.

We compare across 5 different spatial proteomics datasets across three different organisms. The datasets we compare our methods on are *Drosophila melanogaster* embryos from [448], the mouse pluripotent embryonic stem cell dataset of [68], the HeLa cell line dataset of [220], the mouse primary neuron dataset of [221] and finally a CRISPR-CAS9 knock-out coupled to spatial proteomics analysis dataset (AP5Z1-KO1) of [202]. The results are found in figure 6.11. We see that our in four out of five datasets there is an improvement of the GP models over the TAGM model (Kolmogorov-Smirnov (KS) two-sample test $p < 0.0001$), because the GP model is provided with more explicit correlation structure of the data. The empirical Bayes slightly method outperforms the fully Bayesian approach in three of the data sets ((KS) two-sample test $p < 0.01$). These are the mouse pluripotent embryonic stem cell dataset, the HeLa data set of [220] and the HeLA AP5Z1 knock-out dataset of [202]. However, the

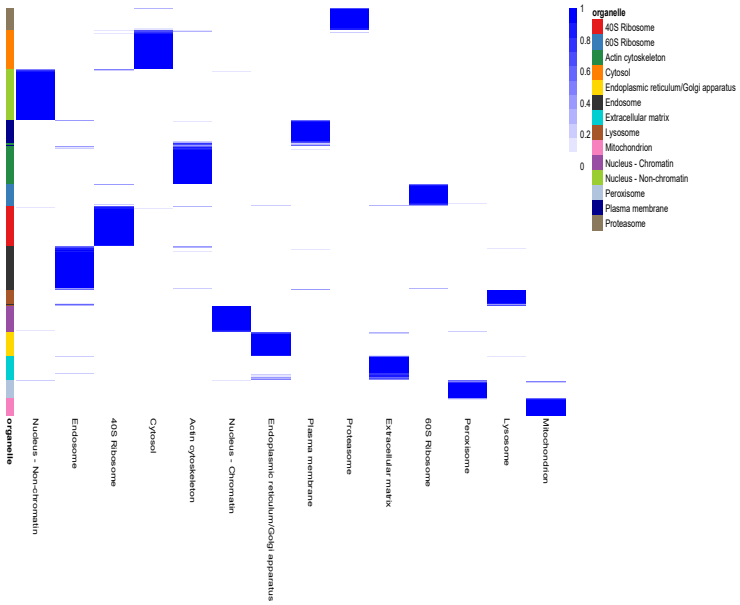


Fig. 6.10 A heatmap of organelles by proteins, where the $(i, j)^{th}$ entry is the Monte-Carlo estimate of the probability that a protein i belongs to organelle j. Allowing us to visualise the range of probabilities for each protein. Proteins are allocated to their most probable class and these allocations are shown in the colour bar on the left.

size of these difference is small, and there is at most a 6 point difference. This corresponds to better assignments for at most 3 proteins. This is hardly worth the loss in uncertainty quantification in the GP hyperparameters when using empirical Bayes over the fully Bayesian approach and the lost ability to provide expert prior information on the GP hyperparameters. Meanwhile, the improvement of the GP methods over the TAGM model is marked in the 4 datasets where we see improvement. Improvements range from score differences of roughly 16 to almost 80, this corresponds to 8 to 40 proteins with better allocations. The GP methods have only 3 parameters for the structured covariance to be inferred, whilst the TAGM model requires inference of full unstructured covariance matrices, which is potentially hundreds of parameters. Improved predictive performance in a lower parameter model is highly desirable.

We observe that the TAGM model outperforms the GP methods in the Itzhak et al. [220] dataset. The authors of this study used differential centrifugation to separate cellular content and curated a "large protein complex" class. This class could contain multiple sub-cellular structures such as ribosomes, as well as cytosolic and nuclear proteins - as observed in chapter 5. In any case, our modelling assumptions are violated in both models and this issue is exacerbated by parametrising the covariance structure. One solution to this would be to model this mixture of large protein complexes as its own class. However, as this class contains a quite diverse set of sub-cellular compartments, it is difficult to predict behaviour. This class could be itself a mixture of GPs, however the number of components of the class would be unknown and this

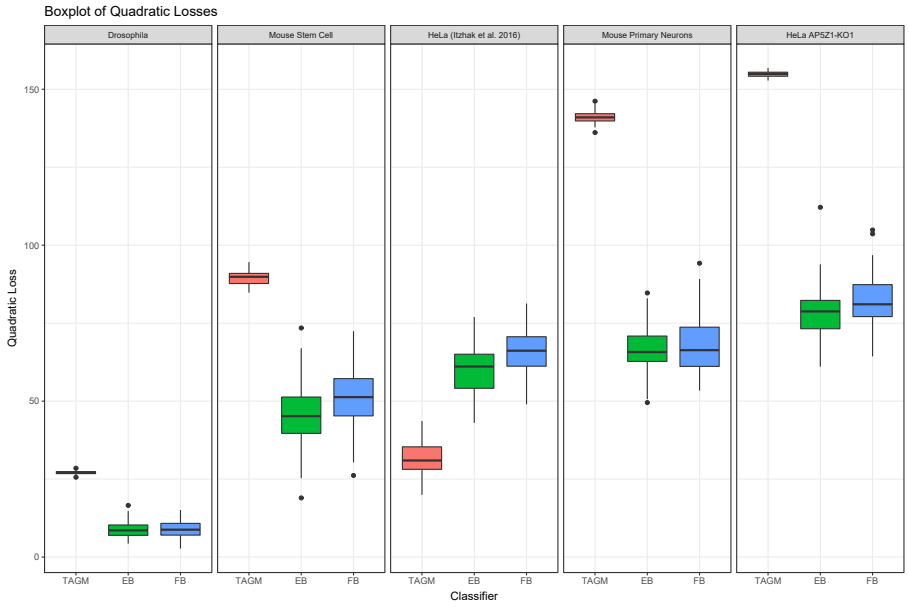


Fig. 6.11 Boxplots of quadratic losses comparing predictive performance of the TAGM against the two semi-supervised Gaussian process models described here, where either an empirical Bayes (EB) approach or fully Bayesian (FB) approach is used for hyperparameter inference. That is (EB) denotes the model where hyperparameters are fixed and learnt for the labelled data only, using L-BFGS to optimise the hyperparameters with respect to the marginal likelihood. (FB) denotes the semi-supervised model where hyperparameters are given priors and the unlabelled data are allowed in the inference of the hyperparameters.

would have to be carefully modelled, perhaps using reversible jump methods [380] or Dirichlet process approaches [125].

6.5 Discussion and limitations

This chapter presents semi-supervised non-parametric Bayesian methods to model spatial proteomics data. Sub-cellular niches display unique signatures along subcellular fractions and we exploit this information to construct GP regression models for each niche. The full complement of sub-cellular proteins is then described as a mixture of GP regression models, with outliers captured by an additional component in our mixture. This provides cell biologists with a fully Bayesian method to analyse spatial proteomics data in the non-parametric framework that more closely reflects the biochemical process used to generate the data. This greatly increases model interpretation and allows us to make more biologically sound inferences from our model.

We compared the proposed semi-supervised models to the state-of-the-art model on 5 different spatial proteomics datasets. Modelling the correlation structure along the subcellular fractions leads to competitive predictive performance over state-of-the-art models. Empirical

Bayes procedures perform either equally well or better than the fully Bayesian approach, at the loss of uncertainty quantification in the hyperparameters. Though this performance improvement should not be over interpreted, since cross-validation assessment is only performed on the labelled data and will not reflect any biased sampling mechanisms that could be at play.

To accelerate computation in our model, we note that the structure of our covariance matrix admits a tensor decomposition, which can be exploited so that fast algorithms for matrix inversion of Toeplitz matrices can be employed. These decomposition can then be used to derive formulae for fast computation of the likelihood and gradient of a GP. A stand-alone R-package implementing these methods using high-performance C++ libraries is available at https://github.com/ococrook/toeplitz. These algorithms and associate formulae are useful to those outside the spatial proteomics community to anyone using GPs with equally spaced observations, even in the unsupervised case.

We demonstrated that in the presence of labelled data there are two approaches to hyperparameter inference. This first, is to use empirical-Bayes to optimise the hyperparameters; the other a fully-Bayesian approach, taking into account the uncertainty in these hyperparameters. We propose to use HMC to update these hyperparameters, since highly correlated hyperparameters can induce high autocorrelation and exacerbate issues with random-walk MH updates. We demonstrate that, in the situation presented here, HMC updates can be up to an order of magnitude more efficient than MH updates. We further explored the sensitivity of our model to hyper-prior specification, which gives practitioners good default choices.

In two case-studies, we highlighted the value of taking a semi-supervised approach to hyperparameter inference, allowing us to explore the uncertainty in our hyperparameters. In a fully Bayesian approach the uncertainty in the hyperparamters is reflected in the uncertainty of the localisation of proteins to components. Quantifying uncertainty provides cell biologists with a wealth of information to make quantifiable inference about protein sub-cellular localisation.

Thus far in our method development, we have tackled several questions related to classification of proteins to organelles. These methods have allowed uncertainty quantification in the assignment of proteins to organelles, as well as the detection of unannotated niches. We also quantified uncertainty in the discovery of these niches. In this chapter, we framed spatial proteomics in the semi-supervised non-parametric Bayesian framework. These models can of course be extended, for example describing some niches themselves as mixtures of GPs or using hierarchical GPs to model the correlation between replicates (rather than assuming independence between biological replicates). However, most of these modelling tasks would be incremental improvements beyond the methods discussed here. A number of more interesting questions are still open in spatial proteomics; such as, which proteins change localisation upon external stimulus and what is the role of post-translation modifications? In the following chapter, we develop a Bayesian model for differential localisation.

Chapter 7

Inferring differential subcellular localisation in comparative spatial proteomics using BANDLE

7.1 Motivation

In previous chapters we have focused on the allocation problem in spatial proteomics. We have developed parametric and non-parametric Bayesian methods and we have reduced the reliance on marker proteins. We now turn to focus on the relatively new dynamic question: which proteins change their resident organelle upon subcellular perturbation? To answer this question we need to carefully consider the experimental design and define a new concept: *differential localisation*. Building on the work of previous chapters we develop a Bayesian model to answer the differential localisation problem. Motivation for this chapter and some of the concepts are published in [82]; however, the majority of this chapter is in preparation for submission.

7.1.1 Abstract

The steady-state localisation of proteins provides vital insight into their function. These localisations are context specific with proteins translocating between different sub-cellular niches upon perturbation of the subcellular environment. *Differential localisation* provides a step towards mechanistic insight of subcellular protein dynamics. Aberrant localisation has been implicated in a number of pathologies, thus *differential localisation* may help characterise disease states and facilitate rational drug discovery by suggesting novel targets. High-accuracy high-throughput mass spectrometry-based methods now exist to map the steady-state localisation and re-localisation of proteins. Here, we propose a principled Bayesian approach, BANDLE (Bayesian ANalysis of Differential Localisation Experiments), that uses these data to compute the

probability that a protein differentially localises upon cellular perturbation, as well quantifying the uncertainty in these estimates. Furthermore, BANDLE allows information to be shared across spatial proteomics datasets to improve statistical power. Extensive simulation studies demonstrate that BANDLE reduces the number of both type I and type II errors compared to existing approaches. Application of BANDLE to datasets studying EGF stimulation and AP-4 dependent localisation recovers well studied translocations, using only two-thirds of the provided data. Moreover, we implicate TMEM199 with AP-4 dependent localisation. In an application to cytomegalovirus infection, we obtain novel insights into the rewiring of the host proteome. Integration of high-throughput transcriptomic and proteomic data, along with degradation assays, acetylation experiments and a cytomegalovirus interactome allows us to provide the functional context of these data.

7.2 Introduction and literature review

Throughout this thesis, we have explored Bayesian methods for spatial proteomics in the static setting. Determining a protein's steady-state localisation can be a first step in determining its function. Furthermore, many biological processes are regulated by re-localisation of proteins, such as transcription factors shuttling from the cytoplasm to the nucleus, which are difficult to map using imaging methods at scale [366]. To simultaneously study the steady-state localisation and re-localisation of proteins, one approach is to couple gentle cell lysis and cell fractionation with high-accuracy mass spectrometry (MS) [68, 324, 159, 351]. Dynamic experiments have given us unprecedented insight into HCMV infection [24], EGF stimulation [220], EGFR inhibition [351]. In addition, CRISPR-Cas9 knockouts coupled with spatial proteomics has given insights into AP-4 vesicles [92], as well as AP-5 cargo [202]. In a study by Shin et al. [418], the golgin long coiled-coil proteins that selectively capture vesicles destined for the Golgi were re-located to the mitochondria by replacing their Golgi targeting domains with a mitochondrial transmembrane domain [418]. This allowed the authors to readily observe the vesicle cargo and regulatory proteins that are redirected to the mitochondria, whilst avoiding technical issues that arise because of the redundancy of the golgins and their transient interaction with vesicles. Together, these collections of experiments suggest spatial proteomics can provide unprecedented insight into biological function.

In dynamic and comparative experiments; that is, those where we expect re-localisation upon some stimulus to sub-cellular environment, the data analysis is more challenging. The task can no longer be phrased as a supervised learning problem, but the question under consideration is clear: which proteins have different sub-cellular niches after cellular perturbation? Procedures to answer this question have been presented by authors [24, 220, 221, 241] and reviewed in Crook et al. [81]. The approach of Itzhak et al. [220, 221] relies on coupling a multivariate

outlier test and a reproducibility score - termed the (movement-reproducibility) MR method. A threshold is then applied to these scores to obtain a list of proteins that re-locate; "moving" proteins. However, these scores can be challenging to interpret, since their ranges differ from one experiment to another and require additional replicates to calibrate the scores. Furthermore, the test ignores the spatial context of each protein, rendering the approach inefficient with some applications allowing false discovery rates of up to 23% [202]. Finally, the approach does not quantify uncertainty which is of clear importance when absolute purification of sub-cellular niches is impossible and multi-localising proteins are present. Recently, Kennedy et al. [241] introduced a computational pipeline for analysing dynamic spatial proteomics experiments by reframing it as a classification task. However, this formulation ignores that some changes in localisation might be shifts in multi-localisation patterns or only partial changes. Furthermore, their approach cannot be applied to replicated experiments and so its applicability is limited. In addition, the authors found that they needed to combine several of the organelle classes together to obtain good results. Finally, the framing of the problem as a classification task only allows a descriptive analysis of the data. These considerations motivate the development of a more sophisticated and reasoned methodology.

In this chapter, we present Bayesian ANalysis of Differential Localisation Experiments (BANDLE) - an integrative semi-supervised functional mixture model, to obtain the probability of a protein being differentially localised between two conditions. Posterior Bayesian computations are performed using Markov-chain Monte-Carlo and so uncertainty estimates are also available [173]. We associate the term *differentially localised* to those proteins which are assigned different sub-cellular localisations between two conditions. Then, we refer precisely to this phenomenon as *differential localisation*, throughout the text. Hence, our main quantity of interest is the probability that a protein is differentially localised between two conditions.

BANDLE models the quantitative protein profiles of each sub-cellular niche in each replicate of each experiment non-parametrically [84]. A first layer of integration combines replicate information in each experiment to obtain the localisation of proteins within a single experimental condition. Then a joint prior distribution on protein allocations across experimental conditions allows information to be shared across experiments and a differential localisation probability to be obtained. Two prior distributions are proposed: one using a matrix extension of the Dirichlet Distribution and another based on Pólya-Gamma augmentation [368, 65, 270].

A number of integrative mixture models have been proposed including Multiple Dataset Integration [243], infinite tensor factorisation approaches [18], Bayesian Consensus Clustering [278] and Clusternomics [146]. The methods have been developed mostly in the context of cancer sub-typing or transcriptional module discovery. Our approach is most similar to Clusternomics, which places a prior on the tensor product between the mixing proportions; but instead our model defines mixing proportions across datasets - upon which we introduce a

prior. Importantly, our approach demands that there is an explicit link between components in each dataset, which can be difficult to assume outside the semi-supervised setting because of a statistical issue known as label-switching [380].

In this chapter, we first review previous methods to tackle the differential localisation problem, as well as integrative mixture models. We then demonstrate the utility of BANDLE, by first performing extensive simulations and compare to the MR approach. We show that our approach reduces the number of Type I and Type II errors, and, as a result, can report an increased number of differentially localised proteins. These simulations also highlight the robustness of our approach to a number of experimental scenarios including batch effects. Our simulation studies also highlight that BANDLE provides interpretative improvements and clearer visualisations, and makes less restrictive statistical assumptions. We then apply our method to a number of datasets with well studied examples of differential localisation, including EGF stimulation and AP-4 dependent localisation. We recover known biology and provide additional cases of differential localisation, and demonstrate that TMEM199 localisation is AP-4 dependent. Finally, we apply BANDLE to a human cytomegalovirus (HCMV) dataset - a case where the MR approach is not applicable because the MR approach requires multiple replicates. Integration of high-throughput transcriptomic and proteomic data, along with degradation assays, acetylation experiments and a cytomegalovirus interactome allows us to provide the functional context of these data. In particular, we provide the spatial context of the interactome data.

7.3 Methods

7.3.1 Previous methods

The Movement-Reproducibility method

The movement-reproducibility (MR) method was proposed by Itzhak et al. [220, 221] and this is our interpretation of their method. We suppose that we are given two spatial proteomics experiments under a single contrast/perturbation/treatment, and denote unperturbed by $(d = 1)$ and $(d = 2)$ for the perturbed condition. Furthermore, assume we measure each condition with $r = 1, ..., R$ biological replicates. Let $X_1 = [X_1^{(1)}, ..., X_1^{(R)}]$ denote the concatenation of replicates for condition 1 and likewise for condition 2 denotes $X_2 = [X_2^{(1)}, ..., X_2^{(R)}]$. We first compute delta matrices as follows

$$\Delta = X_1 - X_2, \tag{7.1}$$

where $\Delta = [\Delta^{(1)}, ..., \Delta^{(R)}]$. This assumes that both features and replicates are comparable in some way; that is, a feature in the r^{th} replicate is directly comparable to the same feature in another replicate. Then for each Δ_r, $r = 1, .., R$, the squared Mahalanobis distance D_M from each protein to the empirical mean is computed using a robust estimate of the covariance matrix - the minimum covariance determination method [214]. Under a Gaussian assumption on Δ_r, $D_M(p_i)$ follows a chi-squared distribution with degrees of freedom equal to the dimension of the data G. Then, for each protein and each replicate a p-value is computed, such that there are R such p-values for each protein. These p-values are combined into a score by taking the cube of the largest p-value for each protein, correcting for multiple hypothesis testing using the Benjamini-Höchberg procedure and computing the $-\log_{10}$ of the resultant value. The final score is called the M score.

This process means that the computed value can no longer be interpreted as truly derived from a p-value. To maintain this interpretation one could instead combine p-values using Fisher's method [322]. Furthermore, the authors are, implicitly, concerned with finding *any* false positives and as such control over the FWER is desired rather than the FDR. Since FWER $\geq$ FDR, control of the FDR does not lead to control over the FWER.

A so-called reproducibility (R) score is obtained by first computing the Pearson correlation pairwise between matrices $\Delta_i, \Delta_j, i \neq j$ for each protein. A final R score, for each protein, is obtain by taking the minimum value for each protein. Again this score could have been interpreted in a formal testing procedure using a permutation test [124] and furthermore includes an assumption of bivariate normality. Moreover, Pearson's correlation is unresponsive to many non-linear relationships which might be present.

Finally, each protein has an associated pair of scores, referred to as the MR-score. To determine thresholds for these scores the authors take a desired FDR = 0.01. Thus they repeat

a control experiment 6 times to determine thresholds $M = 2, R = 0.9$ a region with no false discoveries.

Repeating the control experiment 6 times is a costly process and likely to be prohibitive for most experiments, particularly for cells that are expensive to culture. Furthermore, since the thresholds are empirically derived, this process needs to be repeated for every new experiment to determine optimal thresholds.

Integrative mixture models

For completeness, we include the background material on other Bayesian integrative mixture models. We introduce the different models and then compare them in detail. The first example we consider is the multiple dataset integration (MDI) method of Kirk et al. [243], where the joint prior for allocations (in the two dataset scenario) is given by:

$$\begin{aligned} \phi &\sim \mathcal{G}(a, b) \\ \pi_1 &\sim Dir(\frac{\alpha_1}{K_1}, ..., \frac{\alpha_1}{K_1}) \\ \pi_2 &\sim Dir(\frac{\alpha_2}{K_2}, ..., \frac{\alpha_2}{K_2}) \\ p(z_{i1}, z_{i2}|\phi) &\sim \pi_{z_i 1}\pi_{z_i 2}(1 + \phi \mathbb{1}(z_{i1} = z_{i2})). \end{aligned} \tag{7.2}$$

Meanwhile for clusternomics [146] the prior is

$$\begin{aligned} \rho &\sim Dir(\gamma \text{vec}(\pi_1 \otimes \pi_2)) \\ \pi_1 &\sim Dir(\frac{\alpha_1}{K_1}, ..., \frac{\alpha_1}{K_1}) \\ \pi_2 &\sim Dir(\frac{\alpha_2}{K_2}, ..., \frac{\alpha_2}{K_2}) \\ p(z_{i1} = k, z_{i2} = j) &= \rho_{kj}. \end{aligned} \tag{7.3}$$

The model for Bayesian Consensus Clustering (BCC) is the following [278]. First, define a global latent allocation $C = \{c_1, ..., c_n\}$, to one of K possible clusters. Then, for the d^{th} dataset define the local latent allocation z_{id}, where the conditional probability is given by

$$p(z_{id} = k|c_i) = \alpha_d \mathbb{1}(z_{id} = c_i) + \frac{1 - \alpha_d}{K - 1}(1 - \mathbb{1}(z_{id} = c_i)). \tag{7.4}$$

The key idea of BCC is that first a global latent allocation (or clustering) is defined and then local clusterings are defined conditional on the global clustering. The concentration parameter α_d controls the level of association between the global and local allocation. Importantly, note that if the k^{th} global component is empty, then corresponding local component probability is 0 if and only if $\alpha_d = 1$. Hence, in general there will be more local components than global

components. The approach is most similar to MDI, which allows the clusters to vary arbitrarily between the datasets. In the language of BCC and clusternomics, each dataset is allowed its own set of local clusters. Then the parameter ϕ up weights the prior probability that observations are allocated to the corresponding local components in each dataset. Note if $\phi = 0$ then we are in the independent case and so, in general, there is some up weighting of the joint allocation probabilities as $p_0(\phi > 0) > 0$ and more up weighting if the datasets are more similar. Clusternomics is, somewhat, the reverse of BCC. In contrast, allocations are defined first at the local level and then information is shared via a global allocation. As the number of local clusters increases so does the number of global clusters. Furthermore, the global concentration parameter γ controls the level of cluster sharing across the datasets.

A model for differential localisation

In the following, we layout our model for BANDLE, along with methods for inference, and approaches for summarising and visualising the output. Firstly, suppose we have two spatial proteomics experiments with unperturbed ($d = 1$) and perturbed conditions ($d = 2$). Furthermore, assume we measure each condition with $r = 1, ..., R$ biological replicates. Let $X_1 = [X_1^{(1)}, ..., X_1^{(R)}]$ denote the concatenation of replicates for condition 1 and likewise for condition 2 denotes $X_2 = [X_2^{(1)}, ..., X_2^{(R)}]$. We introduce the following latent allocation variable $z_{i,d}$, denoting the localisation of protein i in condition d. Thus, if $z_{i,d} = k$ this means that protein i localises to organelle k in dataset d. Given this latent allocation variable, we assume that the data from replicate $r = 1, ..., R$ arises from some component density $F(\cdot|\theta_k^{(r)})$. Hence, denoting by θ the set of all component parameters, we can write

$$x_{i,d}^{(r)}|z_{i,d}, \theta \sim F(x_{i,d}^{(r)}|\theta_{z_{i,d}}^{(r)}). \tag{7.5}$$

We assume that biological replicates are independent and so we factorise as follows

$$p(x_{i,d}|z_{i,d}, \theta) = \prod_{r=1}^{R} p(x_{i,d}^{(r)}|z_{i,d}, \theta_{z_{i,d}}^{(r)}). \tag{7.6}$$

To couple the two conditions together we assume a joint prior structure for the latent allocation variable in each dataset. To be more precise, we construct a prior for the pair $(z_{i,1}, z_{i,2})$. We fix the possible number of subcellular niches to which a protein may localise to be K. Now, we introduce the matrix Dirichlet distribution, which we denote as $\mathcal{M}\text{Dir}(\alpha, K)$. The concentration parameter α is a $K \times K$ matrix, such that for a matrix $\boldsymbol{\pi}$, the pdf of the matrix Dirichlet distribution is

$$f(\boldsymbol{\pi}|\alpha) = \prod_{k=1}^{K} \frac{1}{\mathcal{B}(\alpha_k)} \prod_{j=1}^{K} \pi_{jk}^{\alpha_{jk}-1}, \tag{7.7}$$

where $\mathcal{B}$ denotes the beta function, α_k denotes the k^{th} row of α, and $\sum_{j,k} \pi_{jk} = 1$. Thus, we propose the following hierarchical structure

$$\boldsymbol{\pi}|\alpha \sim \mathcal{M}\text{Dir}(\alpha, K) \tag{7.8}$$

$$(z_{i,1}, z_{i,2}) \sim cat(\boldsymbol{\pi}), \tag{7.9}$$

where $(z_{i,1}, z_{i,2}) \sim cat(\boldsymbol{\pi})$ means that the prior allocation probabilities are given by

$$p(z_{i,1} = k, z_{i,2} = k'|\boldsymbol{\pi}) = \pi_{kk'}. \tag{7.10}$$

The above model is conjugate, and so if $n_{j,k} = |\{(z_{i,1}, z_{i,2}) = (j,k)\}|$, it follows that the conditional posterior of $\boldsymbol{\pi}$ is

$$\boldsymbol{\pi}|(Z_1, Z_2), \alpha \sim \mathcal{M}\text{Dir}(\gamma, K) \tag{7.11}$$

where $\gamma_{j,k} = \alpha_{jk} + n_{j,k}$. The likelihood models for the data are Gaussian Random Fields, which we elaborate on in the following section. Hence, the conditional posterior of the allocation probabilities are

$$p(z_{i,1} = j, z_{i,2} = k|\boldsymbol{\pi}) \propto \pi_{jk} \prod_{r=1}^{R} p(x_{i,1}^{(r)}|z_{i,1} = j) p(x_{i,2}^{(r)}|z_{i,2} = k). \tag{7.12}$$

Likelihood Model

The model described in the previous section is presented in a general form, so it could be applied to many different modes of data. We describe the model for a single spatial proteomics experiment, since the same model is assumed across all spatial proteomics experiments, that are then subsequently joined together using the approach in the previous section. Though the model is the same across experiments, the parameters are experiment-specific.

The likelihood model is the same as the previous chapter but we repeat here to set notation, as well as language choices. We assume that the protein intensity x_i at each fraction s_j can be described by some regression model with unknown regression function:

$$x_i(s_j) = \mu_i(s_j) + \varepsilon_{ij}, \tag{7.13}$$

where μ_i is some unknown deterministic function of space and ε_{ij} is a noise variable, which we assume is $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma_i^2)$. Proteins are grouped together according to their subcellular localisation; such that, all proteins associated to subcellular niche $k = 1, ..., K$ share the same regression model. Hence, we write $\mu_i = \mu_k$ and $\sigma_i = \sigma_k$. Throughout, for clarity, we refer to sub-cellular structures, whether they are organelles, vesicles or large protein complexes, as *components*. The regression functions μ_k are unknown and thus we place priors over these

functions to represent our prior uncertainty. Protein intensities are spatially correlated and thus we place *Gaussian Random Field* (GRF) priors over these regression functions. We pedantically refer to these as GRF priors rather than *Gaussian Process* (GP) priors to make the distinction between the 1D spatial process that separates sub-cellular niches and the experimental cellular perturbations, which are potentially temporal in nature. Hence, we write the following

$$\mu_k \sim GRF(m_k(\boldsymbol{s}), C_k(\boldsymbol{s}, \boldsymbol{s'})), \tag{7.14}$$

which is defined as:

Definition 5. *Gaussian Random Field*
If $\mu(\boldsymbol{s}) \sim GRF(m_k(\boldsymbol{s}), C_k(\boldsymbol{s}, \boldsymbol{s'}))$ *then for any finite dimensional collection of indices* $s_1, ..., s_n$, $[\mu(s_1), ..., \mu(s_n)]$ *is multivariate Gaussian with mean* $[m(s_1), ..., m(s_n)]$ *and covariance matrix such that* $C_{ij} = C(s_i, s_j)$.

Each component is thus captured by a Gaussian Random Field model and the full complement of proteins as a finite mixture of GRF models. The protein intensity for each experiment maybe measured in replicate. For a sufficiently flexible model, we allow different regression models across different replicates. To be more precise, consider the protein intensity $x_i^{(r)}$ for the i^{th} protein measured in replicate r at fraction $s_j^{(r)}$, then we can write the following

$$x_i^{(r)}\left(s_j^{(r)}\right) = \mu_k^{(r)}\left(s_j^{(r)}\right) + \varepsilon_{ij}^{(r)}, \tag{7.15}$$

having assumed that the i^{th} protein is associated to the k^{th} component. The (hyper)parameters for the Gaussian Random Field priors for the r^{th} replicate in experiment d are denoted by $\theta_{k,d}^{(r)}$. We denote by $\boldsymbol{\theta}$ the collection of all hyperparameters and the collection of priors for these by $G_0(\boldsymbol{\theta})$. The loss of conjugacy between the prior on the hyperparameters and likelihood is unavoidable.

The GRF is used to model the uncertainty in the underlying regression functions; however, we have yet to consider the uncertainty that a protein belongs to each of these components. To capture these uncertainties, we can use the model in the previous section, allowing information to be shared across each condition. Following from the previous section, the conditional posterior of the allocation probabilities is

$$p(z_{i,1} = j, z_{i,2} = k|\boldsymbol{\pi}) \propto \pi_{jk} \prod_{r=1}^{R} p(x_{i,1}^{(r)}|z_{i,1} = j) p(x_{i,2}^{(r)}|z_{i,2} = k), \tag{7.16}$$

where, in the specific case of our likelihood model, the probabilities in the terms of the product can be computed using the appropriate GRF.

We assume that our Gaussian random fields are centred and that the covariance is from the Matérn class [434], as described in the previous chapter. The Matérn covariance is specified as follows

$$C_v(d) = a^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{8\nu} \frac{d}{\rho} \right)^{\nu} \mathcal{K}_v \left(\sqrt{8\nu} \frac{d}{\rho} \right), \tag{7.17}$$

Recall the parameter ν controls the differentiability of the resulting sample paths; such that, $\lceil v \rceil$ is the number of mean-square derivatives. For typical applications, ν is poorly identifiable and fixed. $\nu = 1/2$ recovers the exponential covariance, whereas taking the limit $\nu \to \infty$ one obtains the squared exponential (Gaussian) covariance. We fix $\nu = 2$.

A ridge in the marginal likelihood for the marginal variance and range parameters of the Matérn covariance makes inference challenging. Indeed, different hyperparameters lead to unconditional prior simulations with the same spatial pattern but different scales [376, 143]. Furthermore, when the intrinsic dimension of the Gaussian random field is less than four, there is no consistent estimator under in-fill asymptotics for ρ and a. A principled prior, which allows domain expertise to be expressed, is thus desired to enable stable inferences. A number of works considered reference priors for GRFs [26, 357, 103, 463]. Here, we employ a recently introduced collection of weakly-informative priors, which we introduce in the next section.

Penalised Complexity Priors

The penalised-complexity (PC) prior framework introduced by Simpson et al. [426] and Fuglstad et al. [143] allows priors to be specified in modular fashion. In brief, the PC prior framework considers model components, such as GRFs, a flexible extension of some *base model*. Priors are then constructed such that they shrink the more flexible model towards the base model. The first consideration is an appropriate distance from the base model P_0 to the flexible model P. Simpson et al. [426] choose the distance as $\sqrt{2\mathrm{KL}(P||P_0)}$, where $\mathrm{KL}(P||P_0)$ denotes the *Kullback-Leibler* divergence from P_0 to P. The square root and the factor 2 puts the distance on the appropriate scale [426]. A constant-rate penalisation principle is then used to derive the following condition

$$\frac{\pi(t+\delta)}{\pi(t)} = r^{\delta} \; t, \delta > 0, \tag{7.18}$$

where $0 < r < 1$ is the constant rate decay. This construction means that the strength of the penalty on the model increases as we depart from the base model. The only continuous distribution that satisfies this property is the exponential distribution $\pi(t) = \lambda \exp(-\lambda t)$ for $t > 0$. The hyperparameter λ allows expert information to be expressed, once the prior has been transformed onto the quantity of interest.

Penalised Complexity Priors for GRFs

To derive PC priors for GRF with a Matérn covariance requires some complex considerations. Firstly, it is required to parametrise the Matérn covariance to clarify which parameters are identifiable under in-fill asymptotics [143]. This parametrisation loses physical interpretation but is more amenable to theoretical considerations:

$$\kappa = \frac{\sqrt{8\nu}}{\rho} \text{ and } \tau = a\kappa^{\nu}\sqrt{\frac{\Gamma(\nu+1/2)(4\pi)^{1/2}}{\Gamma(\nu)}}. \tag{7.19}$$

Thus, since τ can be inferred under in-fill asymptotics and κ cannot, a joint PC prior $\pi(\kappa,\tau) = \pi(\tau|\kappa)\pi(\kappa)$ is constructed in two stages. Fuglstad et al. [143] argue that the appropriate base model in the this scenario are GRFs with infinite length-scales and zero marginal variance. Using spectral representations of the GRFs Fuglstad et al. [143] derive the required joint PC prior for κ and τ. It is then simple to parametrise the prior for the parameters ρ and a. The joint PC prior is stated below [143]:

Theorem 3. *Joint PC prior for GRFs*
Let u be a GRF defined on $\mathbb{R}$, with Matérn covariance with parameters a, ρ and ν. Then the joint PC prior $\pi(a,\rho)$ corresponding to a base model with infinite range and zero variance is

$$\pi(a,\rho) = \frac{\lambda_1\lambda_2}{2}\rho^{-3/2}\exp(-\lambda_1\rho^{-1/2} - \lambda_2 a), \tag{7.20}$$

where $P(\rho < \rho_0) = \alpha_1$ and $P(a > a_0) = \alpha_2$ are achieved by

$$\lambda_1 = -\log(\alpha_1)\rho^{1/2} \text{ and } \lambda_2 = \frac{-\log(\alpha_2)}{a_0}. \tag{7.21}$$

Penalised complexity prior for the noise model

The noise effect is distributed according to $\varepsilon_{ij} \sim \mathcal{N}(0,\sigma_k^2)$ for $k = 1,...,K$. We additionally choose a PC prior in this scenario, first we reparametrize in terms of a precision $\tau_k = 1/\sigma_k^2$ for $k = 1,...,K$. Then appealing to Simpson et al. [426] the PC prior is a type-2 Gumbel distribution

$$\pi(\tau) = \frac{\lambda_3}{2}\tau^{-3/2}\exp(-\lambda_3\tau^{-1/2}). \tag{7.22}$$

The penalised complexity prior in this case shrinks towards zero variance. The hyperparameter λ_3 can be set using the following tail probability $p(\sigma_k > U) = \alpha$ results in $\lambda_3 = \frac{-\log(\alpha)}{U}$.

Modelling outliers and hyperparameter inference

Outlier modelling is performed as described in previous chapter using an additional heavy tailed student's t-component. Hyperparameters are inferred using optimisation rather than sampling

from the posterior distribution using MCMC and we refer to chapter 6 for details. This is to reduce the computational burden of this more complex model.

Calibration of Dirichlet prior

The following section describes how to calibrate the Dirichlet prior based on expert information and prior predictive checks. Recall the prior on the allocation probabilities is the following

$$p(z_{i,1} = k, z_{i,2} = k' | \boldsymbol{\pi}) = \pi_{kk'}. \tag{7.23}$$

The matrix $\boldsymbol{\pi}$ has π_{jk} as its $(j,k)^{th}$ entry and π_{jk} is the prior probability that a protein belongs to organelle j in dataset 1 (control) and k in dataset 2 (contrast). The diagonal terms represent the probability that the protein was allocated to the same organelle in each dataset. The non-diagonal terms are the prior probability that the protein was not allocated to the same organelle. Since the number of non-diagonal terms greatly exceeds to the number of diagonal entries it is important to specify this prior carefully. Recall that the prior is given a matrix Dirichlet distribution with concentration parameter α.

Firstly, we are interested in the prior expectation of the number of proteins that are *differentially localised*; that is, proteins not allocated to the same organelle in both conditions. Let ρ be the prior probability that a protein is not allocated to the same organelle. Then it follows that

$$p(z_{i,1} \neq z_{i,2} | \boldsymbol{\pi}) =: \rho = \sum_{j,k;j\neq k} \pi_{jk}. \tag{7.24}$$

By properties of the Dirichlet distribution we have that

$$\pi_{jk} \sim \mathcal{B}(\alpha_{jk}, \alpha_0 - \alpha_{jk}). \tag{7.25}$$

Thus, the expected value of ρ is computed as follows

$$\begin{aligned} \mathbb{E}[\rho] &= \sum_{j,k;j\neq k} \mathbb{E}[\pi_{jk}] \\ &= \sum_{j,k;j\neq k} \frac{\alpha_{jk}}{\alpha_0}. \end{aligned} \tag{7.26}$$

We are further interested in the probability that a certain number of proteins, say q, are differentially localised. Letting N_U be the number of unlabelled proteins in the experiment, then the distribution of the prior number of differential localised proteins is

$$p(N_U \rho > q) = p\left(N_U \sum_{j,k;j\neq k} \pi_{jk} > q \right) = \delta. \tag{7.27}$$

Computing δ is not simple; however, it is straightforward to estimate δ using Monte-Carlo by simply sampling from Beta distributions:

$$p\left(N_U \sum_{j,k;j\neq k} \pi_{jk} > q\right) \approx \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}\left(N_U \sum_{j,k;j\neq k} \pi_{jk}^{(t)} > q\right). \tag{7.28}$$

Thus, we recommend calibrating the Dirichlet prior using the above expectation and quantile. It many be important to calibrate several quantiles to ensure sufficient mass is placed on desired regions of the probability space. For example, let $q_1 < q_2$, then we may desire that δ_1, below, is not so small to rule out reasonable inferences and that $\delta_2 < \delta_1$ is sufficiently large. These can be computed from the equations below.

$$p\left(N_U \sum_{j,k;j\neq k} \pi_{jk} > q\right) \approx \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}\left(N_U \sum_{j,k;j\neq k} \pi_{jk}^{(t)} > q_1\right) = \delta_1, \tag{7.29}$$

$$p\left(N_U \sum_{j,k;j\neq k} \pi_{jk} > q\right) \approx \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}\left(N_U \sum_{j,k;j\neq k} \pi_{jk}^{(t)} > q_2\right) = \delta_2. \tag{7.30}$$

More precise and informative prior biological knowledge can be specified; for example, should we suspect that some relocalisation events between particular organelles are more likely than others due to the stimuli, these can be encoded into the prior. If we expect more relocalisation events between organelle j and k_1 than organelle j and k_2, this can be encoded by ensuring

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{1}\left(\pi_{jk_1}^{(t)} > \pi_{jk_2}^{(t)}\right) > \delta_3 > 0. \tag{7.31}$$

Alternatively, if an objective Bayesian analysis is preferred, the Jeffery's prior sets $\alpha_{jk} = 0.5$ for every $j, k = 1, .., K$. We do not generally recommend this approach, because the diagonal terms of $\boldsymbol{\pi}$ have a different interpretation to the off-diagonal terms.

Differential localisation probability

The main posterior quantity of interest is the probability that a protein is differentially localised. This can be approximated from the T Monte-Carlo samples as follows, suppressing notational dependence on all data and parameters for clarity

$$\chi_i = p(z_{i,1} \neq z_{i,2}) \approx \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}(z_{i,1}^{(t)} \neq z_{i,2}^{(t)}), \tag{7.32}$$

where t denotes the t^{th} sample of the MCMC algorithm. It is important to note that this quantity is agnostic to the assigned subcellular niche.

To perform uncertainty quantification on the the differential localisation probability, we use the non-parametric bootstrap on the Monte-Carlo samples. More precisely, first sample *uniformly with replacement* from $\{z_{i,1}^{(t)}\}_{t=1}^{T}$ and $\{z_{i,2}^{(t)}\}_{t=1}^{T}$ to a total of T samples. This produces a bootstrap sample indexed by B_1. Then we compute our statistic of interest

$$\chi_{i,B_1}^{*} \approx \frac{1}{|B_1|} \sum_{t \in B_1} \mathbb{1}(z_{i,1}^{(t)} \neq z_{i,2}^{(t)}). \tag{7.33}$$

This process is then repeated to obtain a set of bootstrap samples $\mathbb{B} = \{B_1, ..., B_b\}$, for some large b, say 1000. For each $B_r \in \mathbb{B}$, we compute χ_{i,B_r}^{*} for $r = 1, ..., b$, obtaining a sampling distribution for χ_r from which we can compute functionals of interest.

Posterior localisation probabilities

A further quantity of interest is the posterior probability that a protein belongs to each of the K sub-cellular niches present in the data. For the control this is given by the following Monte-Carlo average

$$p(z_{i,1} = k|\Theta) \approx \frac{1}{T} \sum_{t=1}^{T} p(z_{i,1}^{(t)} = k|\Theta), \tag{7.34}$$

where Θ denotes all other quantities in the model. A corresponding formula also holds for the second dataset

$$p(z_{i,2} = k|\Theta) \approx \frac{1}{T} \sum_{t=1}^{T} p(z_{i,2}^{(t)} = k|\Theta). \tag{7.35}$$

The posterior distribution of these quantities and uncertainty estimates can be computed and visualised in standard ways described in the previous chapters.

A non-conjugate prior

Thus far, we have been using a conjugate Dirichlet prior for the a priori mixing proportions $\boldsymbol{\pi}$. Our model assumes no correlation across $\boldsymbol{\pi}$ due to the use of the Dirichlet distribution. However, we can extend the model to include correlation efficiently using Polya-Gamma augmentation. In the interest of brevity these models are relegated to the appendix, but the results of the model are shown in the chapter.

7.4 Results

7.4.1 The BANDLE workflow

For clarity, we visualise the BANDLE workflow in figure 7.1. The workflow begins with a well defined mass-spectrometry based spatial proteomics experiment. A cellular perturbation of

interest is performed alongside control experiments in wild-type cells. The usual principles of experimental design for proteomics apply, to avoid confounding [154]. Additional quality control steps are undertaken specifically for spatial proteomics experiments [154, 155, 157]. To apply the Bayesian model, we first calibrate the prior based on *prior predictive checks* [162]. In all scenarios, we check the prior expected number of differentially localised proteins and the probability that more than l proteins are differentially localised. These are reported in the appendix. We then proceed with Bayesian parameter inference using Markov-chain Monte-Carlo (MCMC) [173] and the checking of convergence. We then visualise our results principally using rank plots, where proteins are ranked from those most likely to be differentially localised to those least likely. Results are then interpreted using other functional screens, assays and databases.

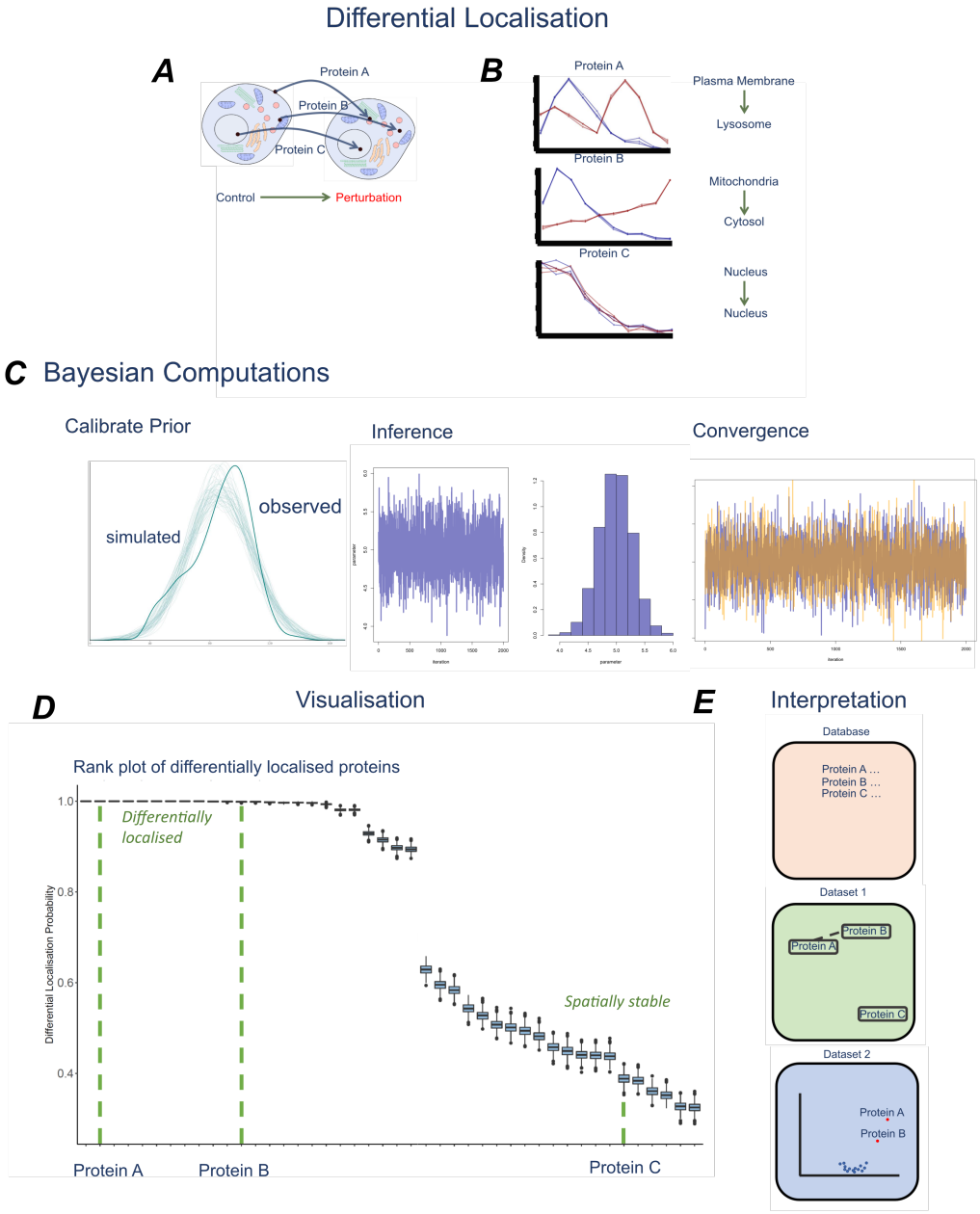


Fig. 7.1 An overview of the BANDLE workflow. (A) A motivated differential localisation experiment is set-up with a perturbation of interest (B) Mass-spectrometry based spatial proteomics methods are applied to generate the data. (C) BANDLE is applied by first calibrating the prior, then performing inference using MCMC, as well as algorithmic assessing convergence. (D) The major results of BANDLE are represented in a rank plot. (E) Results are interpreted using auxiliary data or additional experiments.

7.4.2 Simulations demonstrate superior performance of BANDLE

BANDLE reduces false positives and increases power

To assess the performance of BANDLE and the MR approach, we run a number of simulations allowing us to ascertain the difference between each method in scenarios where we know the ground truth. We first start with a real dataset on Drosophila embryos and simulate replicates, as well as 20 protein re-localisations [448]. To simulate these datasets a bootstrapping approach is used, coupled with additional noise effects. The first simulation uses a simple bootstrapping approach, where a niche-specific noise component is included (see appendix). The subsequent simulations start with the basic bootstrapping approach and add additional effects. The second and third simulations add batch effects: random and systematic respectively (see appendix). The fourth simulation generates misaligned features across datasets by permuting them (fraction swapping) - this models misaligned fractions between replicates (see appendix). The final simulation includes both batch effects and feature permutations. The simulations are repeated 10 times, where each time we simulate entirely new datasets and re-localisations - this is repeated for each simulation task. We assess the methods on two metrics - the area under the curve (AUC) of the true positive rate and false positive rate for the detection of differentially localised proteins. Furthermore, we determine the number of correctly differentially localised proteins at fixed thresholds (see appendix).

Our proposed method, BANDLE, significantly outperforms the MR method with respect to AUC in all scenarios (t-test $p < 0.01$). Furthermore, it demonstrates that BANDLE is robust to a variety of situations, including batch effects. The performance of BANDLE based on the Dirichlet prior is already very good and thus it is unsurprising that we do not observe any significant improvements in AUC by including prior information on correlations captured by the Pólya-Gamma prior. Additional comparisons are made in the appendix where we make similar observations

The improved AUC, which demonstrates improved control of false positives and increased power, translates into increased discovery of differentially localised proteins. Indeed, BANDLE with the Dirichlet prior discovers around twice as many such re-localising proteins. Allowing prior correlations through the Pólya-Gamma prior demonstrates that additional differentially localised proteins are discovered. This is an important reality of those performing comparative and dynamic spatial proteomics experiments, since the experiments become more worthwhile with additional biological discoveries. In practice, the authors of the MR approach advocate additional replicates to calibrate which thresholds are used to declare a protein differentially localised. This assumes that the perturbation of interest does not have a strong effect on the properties of the sub-cellular niches, which restricts applicability. In contrast, BANDLE

does not need additional mass-spectrometry experiments to calibrate its probabilistic ranking meaning more discoveries are made at lower cost.

In the following section, we examine the differences between the approaches in a simulated example. There we focus on the output, interpretation and statistical qualities of each approach, rather than the predictive performance of the methods.

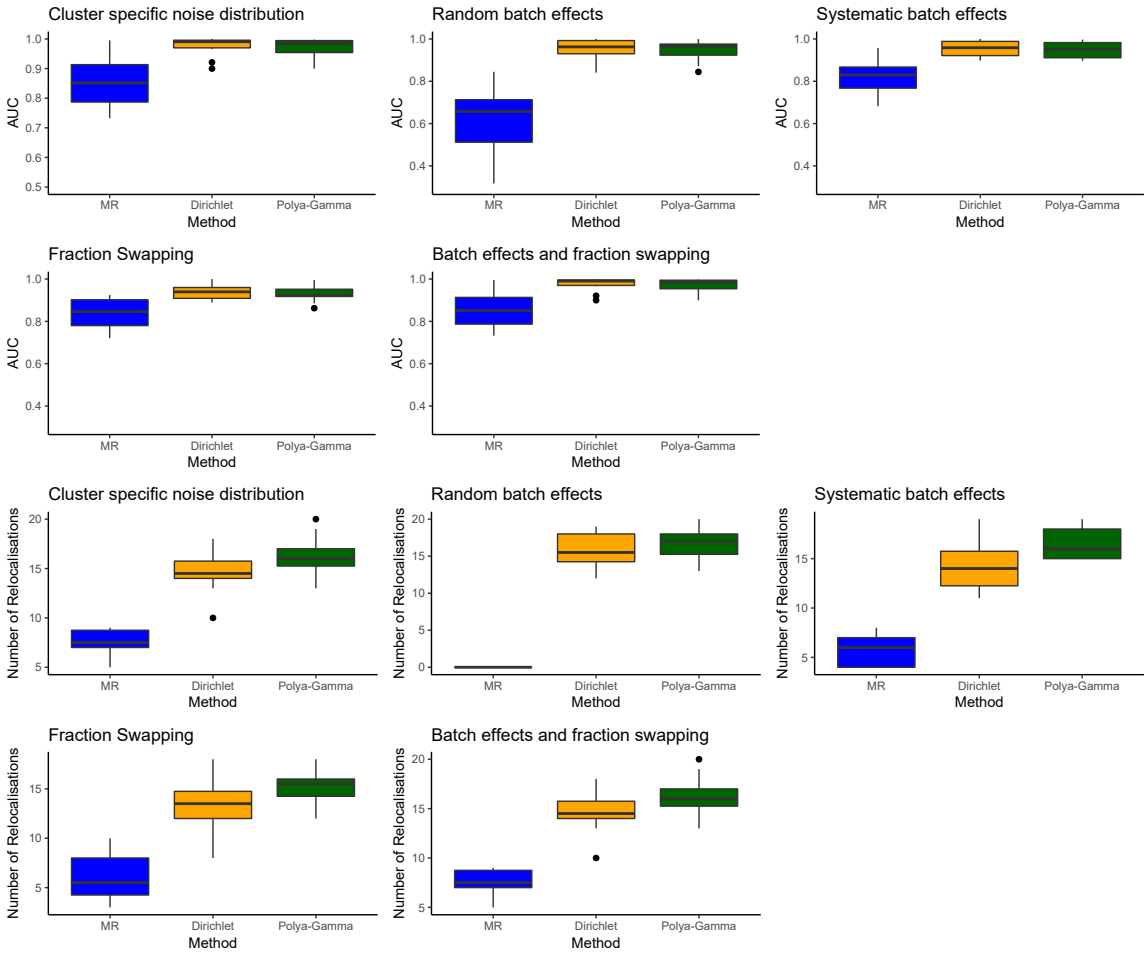


Fig. 7.2 Boxplots comparing the performance of the MR approach and our proposed method BANDLE. BANDLE is separated into whether a Dirichlet-based prior was used or if the Polya-Gamma augmentation was applied. Each boxplot correspond to a different simulation scenario. The first 5 boxplots show BANDLE has significantly improved AUC in all scenarios. These AUCs are translated into the correct number of re-localisations and we can see that our method clearly outperforms the MR approach.

BANDLE quantifies uncertainty and is straightforward to interpret

In this section, we further explore the application of BANDLE with a Dirichlet prior and the MR approach, focusing on the interpretation and statistical properties of the two methods. Again, we simulate dynamic spatial proteomics data, starting from the Drosophila experiment in the scenario in which the MR method performed best. This is where there are cluster specific noise distributions but no other effects, such as batch effects, were included. Sample PCA plots of the data are presented in figure 7.3 A. There is a clear pattern of localisations across the data where proteins with known sub-cellular localisations are closer to each other. However, the organelle distributions clearly overlap and in some cases are highly dispersed - a representation of the challenges faced in real data. These data are annotated with 11 sub-cellular niches and 888 proteins are measured across 3 replicates of control and 3 of treatment (totalling 6 experiments). Re-localisations are simulated for 20 proteins.

We first apply the MR method according to the methods in [220, 221]. We provide a brief description of the approach with full details in the methods. To begin, the difference profiles are computed by subtracting the quantitative values for each treatment from each control. Then the squared Mahalanobis distance is computed to the centre of the data and under a Gaussian assumption the null hypothesis is that these distances follow a Chi-squared distribution, ergo a p-value is obtained. This process is repeated across the 3 replicates and the largest p-value was then cubed and then corrected from multiple hypothesis testing using the Benjamini-Höchberg procedure [25]. A negative $\log_{10}$ transform is then performed to obtain the M-score. To produce the R-score, Pearson correlations are computed between each difference profile for all pairwise combination of difference profiles. The lowest of the three R-scores is reported. The M-score and R-score are plotted against each other (see figure 7.3 B) and the proteins with high M-score and high R-score are considers “hits”.

There are a number of assumptions underlying the MR methodology. Firstly, comparing difference profiles pairwise assumes that the features in both datasets exactly correspond. However, this precludes any stimuli that changes the biochemical properties of the organelles, since changing these properties may result in differing buoyant densities or pelleting of niches at different centrifugation speeds. Thus, whether density-gradient or differential centrifugation is used for organelle separation this assumption must be carefully assessed. Secondly, the Gaussian assumption ignores the natural clustering structure of the data because of the different organelle properties. Indeed, examination of the p-value distributions in a histogram (figure 7.3 C) shows that it clearly deviates from the mixture of distributions expected (p-values are uniformly distributed under the null). The peaking of p-value towards 1 suggests poor distributional assumptions [204]. Thus perhaps the Chi-squared distribution is a poor fit for the statistic of interested. Exploring this further, we fit a Chi-squared and Gamma distribution empirically to the statistics using maximum likelihood estimation (MLE) of the parameters. Figure 7.3

D show that the Gamma distribution is a better distributional fit - successfully capturing the tail behaviour of the statistic. The Chi-squared family is nested in the Gamma family of distributions, so if the theoretical Chi-squared distribution was a good fit the distributions would overlap. For a quantitative assessment of model fits we compute the negative log-likelihood of the data given the optimal distributions - the Gamma distribution has a markedly lower negative log-likelihood (Figure 7.3 E). This provides strong evidence that the underlying Gaussian assumptions are likely violated. Thirdly, it is inappropriate to cube p-values: to combine p-value across experiments one could use Fisher's method [322, 49, 250] or the Harmonic mean p-value (HMP) [178, 481] depending on the context. Indeed, the cube of the p-value is no longer a p-value. To elaborate, if $\mathcal{P}$ are a set of p-values, then under the assumption of the null hypothesis $\mathcal{P}$ is uniformly distributed; however, the cube is clearly not uniformly distributed. Since we no longer work with p-values, Benjamini-Höchberg correction becomes meaningless in this context. Transforming these values to a "Movement score", conflates significance with effect size which confounds data interpretation. Finally, summarising to a single pair of scores ignores their variability across experimental replicates.

BANDLE first models each sub-cellular niche non-parametrically (since the underlying functional forms are unknown [84]). Visualisation of the posterior predictive distributions from these fits for selected sub-cellular niches is given in figure 7.3 H - we observe a good correspondence between the model and the data. We can see that the different sub-cellular niches have contrasting correlation structures and thus niche specific distributions are required. These distributions are specific for each replicate of the experiment and also the two experimental conditions. The information from the replicates, and the control and treatment are combined using an integrative mixture model. Briefly, mixing proportions are defined across datasets allowing information to be shared between the control and treatment (see methods for more details). This formulation allows us to compute the probability that a protein is assigned to a different sub-cellular niche between the two experiments - the *differential localisation probability*. The proteins can then be ranked from most probably differentially localised to least (figure 7.3 H). The figure is simple to interpret: the proteins with highest rank are the most likely to have differentially localised during the experiment, having been confidently assigned to different sub-cellular niches in the control versus treatment. The proteins with lowest rank are highly unlikely to have moved during the experiment - the localisations are stable. This is important information in itself, especially when combined with other information such as changes in abundance or post translational modification. Figure 7.3 G shows the 30 proteins with highest rank visualising the uncertainty in the differential localisation probability (see methods). This ranking allows us to prioritise which proteins to follow up in validation experiments. The ranking can also be mapped onto other experimental data, such as expression or protein-protein interaction data. The probabilistic ranking produced by BANDLE is more closely aligned with

the phenomenon of interest. Indeed, we divide the data into the proteins that were differentially localised and those that were not. Then from plotting the distribution of the statistics from the respective methods, it is clear that output from BANDLE is most closely associated with re-localisation events (figure 7.3 I).

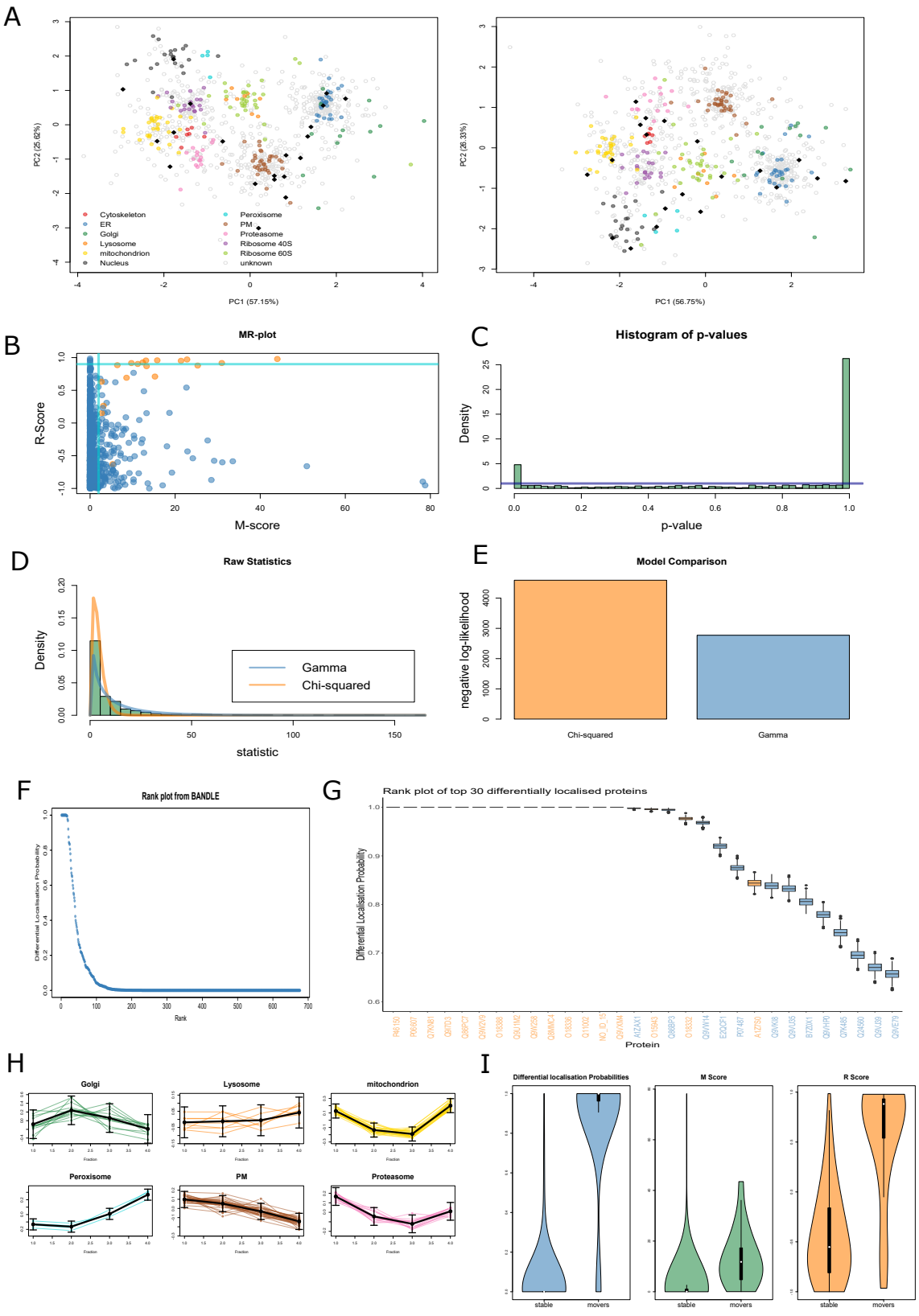

A
PC2 (25.62%)
PC1 (57.15%)
Cytoskeleton
ER
Golgi
Lysosome
mitochondrion
Nucleus
Peroxisome
PM
Proteasome
Ribosome 40S
Ribosome 60S
unknown
PC2 (26.33%)
PC1 (56.75%)
B
MR-plot
R-Score
M-score
C
Histogram of p-values
Density
p-value
D
Raw Statistics
Density
statistic
Gamma
Chi-squared
E
Model Comparison
negative log-likelihood
Chi-squared
Gamma
F
Rank plot from BANDLE
Differential Localisation Probability
Rank
G
Rank plot of top 30 differentially localised proteins
Differential Localisation Probability
Protein
H
Golgi
Lysosome
mitochondrion
Peroxisome
PM
Proteasome
I
Differential localisation Probabilities
M Score
R Score
stable
movers

Fig. 7.3

Fig. 7.3 (A) example PCA plots where pointers correspond to proteins. Marker proteins are coloured according to their subcellular niche, whilst proteins with unknown localisation are in grey. Simulated translocations are highlighted in black, where the left corresponds to control and right to the perturbed dataset. (B) An MR-plot showing movement score against reproducibility score. Each pointer correspond to a protein and orange pointers correspond to simulated translocations and blue otherwise. Teal lines are drawn at suggested thresholds with proteins in the top right corner considered hits. (C) A p-value histogram from the statistic underlying the MR-method. A purple line indicates uniformity. This histogram clearly deviates from uniform behaviour. (D) A histogram of the raw statistics underlying the MR method. A Chi-square (orange) and Gamma (blue) fit are overlaid (obtained using maximum likelihood estimation). The Gamma distribution clearly captures the tail behaviour. (E) model selection on the raw statistic using the Chi-squared and Gamma models. The Gamma model has lower negative log-likelihood and is thus a better model fit. (F) A BANDLE rank plot where proteins are ranked from most to least likely to differentially localised. The differentially localisation probability is recorded on the y-axis. (G) A BANDLE rank plot of the top 30 differentially localised proteins with uncertainty estimates for the differential localisation probability. Proteins marked in orange were simulated translocations. (H) Posterior predictive distributions (black) overlayed on the marker profiles for different subcellular niches showing the quality of the non-parametric BANDLE fits. (I) Violin plots for the differential localisation probabilities, the M score and R score. The distribution are split between differentially localised (movers) and spatially stable proteins. Clearly, the differential localisation probabilities correlate most closely with the phenomena of interest.

7.4.3 Applications to differential localisation experiments

Characterising differential localisation upon EGF Stimulation

Having carefully assessed the statistical properties of our approach, BANDLE, and the MR method, we apply these approaches to a number of datasets. First, we consider the *Dynamics Organeller Maps* (DOMs) dataset of [220], exploring the effects EGF stimulation in HeLa cells. In this experiment, SILAC labelled HeLa cell were cultured and recombinant EGF was added to the culture at a concentration of 20 ng ml^{-1} (see [220]). A total of 2237 complete protein profiles were measured across 3 replicates of control and 3 replicates of EGF treated HeLa cells. Principal Component Analysis (PCA) projections of the data can be visualised in the appendix. A quality control assessment was performed using the approach of [157]. As a result, nuclear pore complex, peroxisome and Golgi annotations were removed, since the marker proteins of these classes were highly dispersed.

The MR method was applied as described in the methods and the results can be visualised in figure 7.4 C. 7 proteins are predicted to be differentially localised with the MR method with the thresholds suggested by [220]. These include 3 core proteins of the EGF signalling pathway SHC1, GRB2 and EGFR [346] and other, potentially related, proteins TMEM214, ACOT2, AHNAK, PKN2. Since the MR approach does not provide information about how the functional residency of the proteins change, it is challenging to interpret these results without further analytical approaches.

To quantify uncertainty and gain deeper insight into the perturbation of HeLa cell after EGF stimulation we applied our BANDLE pipeline. Firstly, the rank plots display a characteristic shape suggesting that most proteins are unlikely to be differentially localisation upon EGF stimulation (figure 7.4 D). Furthermore, we provide uncertainty estimates in the probability that a protein is differentially localised for select top proteins (figure 7.4 E). We also visualise the change in localisation for the proteins known to re-localise upon EGF stimulation: SHC1, GRB2 and EGFR (figure 7.4 G). This is displayed by projecting the posterior localisation probabilities on to the corresponding PCA coordinates. These probabilities are then smoothed using a Nadaraya-Watson kernel estimator [329, 474] and visualised as contours.

Given the well-documented interplay between phosphorylation and sub-cellular localisation [265, 67, 371, 17], we hypothesised that proteins with the greatest differential phosphorylation would correlate with proteins that were more likely to be differentially localised. To this end, we integrated our analysis with a time-resolved phosphoproteomic dataset of EGF stimulation using MS-based quantitation [249]. In their study, EGF stimulated cells were cultured to 8 different time points: $0, 2, 4, 8, 16, 32, 64, 128$ mins. For MS-based quantitation trypsin digested peptides are laballed using iTRAQ 8-plex and pooled. Immunoprecipitation was used to enrich for phosphorylated tyrosine residues [369] and the enrichment of phosphosites on serine and

threonine residues was performed via immobilized metal affinity chromatography (IMAC) [133, 321].

For each phosphopeptide corresponding to a unique protein, we computed the largest $\log_2$ fold change observed across the time course. Given that the changes in localisation occur within 20 minutes, we restricted ourselves to the first 6 time points [220]. We then took the top 10 proteins ranked by each of the MR method and BANDLE. These rankings are then correlated with rankings obtained from the changes in phosphorylation. The Spearman rank correlations were recomputed for $5,000$ bootstrap resamples to obtain bootstrap distributions of correlations (see figure 7.4). We report the mean correlation and the 95% boostrap confidence intervals. The correlation between the ranks of the MR method and the phosphoproteomic dataset was $\rho_S = 0.40\,(-0.49, 0.85)$, whilst the the correlation when using the ranking of BANDLE was $\rho_S = 0.68\,(0.02, 0.98)$. That is to say the proteins which are more likely to be declared as differentially localised according to BANDLE are more likely to differentially phosphorylated than those declared as differentially localised according to the MR method. Alongside the statistical and interpretable benefits of BANDLE, it is clear the approach has the utility to provide insight into localisation dynamics.

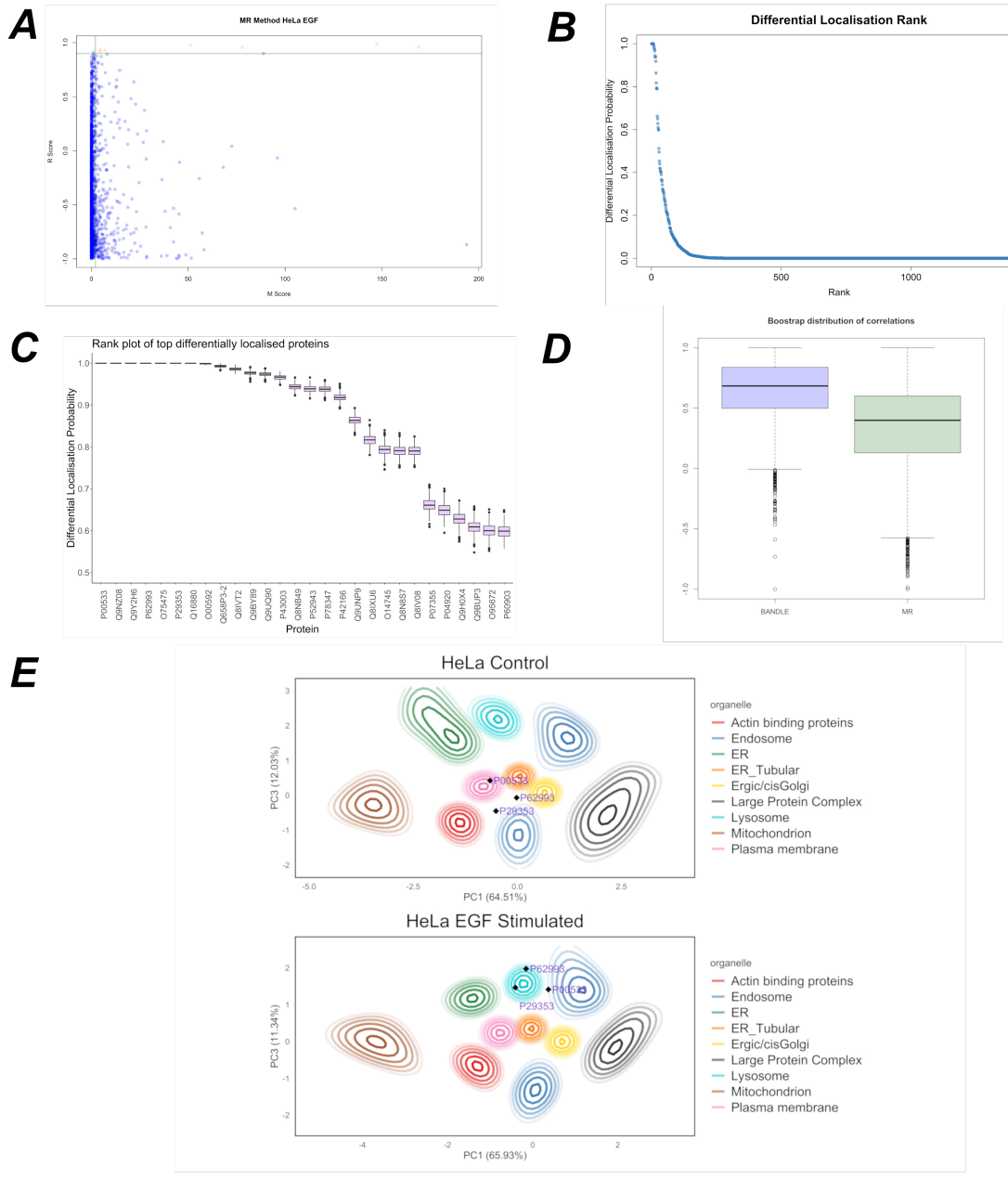

A
MR Method HeLa EGF
R Score
M Score
B
Differential Localisation Rank
Differential Localisation Probability
Rank
C
Rank plot of top differentially localised proteins
Differential Localisation Probability
Protein
D
Boostrap distribution of correlations
BANDLE
MR
E
HeLa Control
PC3 (12.03%)
PC1 (64.51%)
P00533
P62993
P29353
organelle
Actin binding proteins
Endosome
ER
ER_Tubular
Ergic/cisGolgi
Large Protein Complex
Lysosome
Mitochondrion
Plasma membrane
HeLa EGF Stimulated
PC3 (11.34%)
PC1 (65.93%)
P62993
P29353
organelle
Actin binding proteins
Endosome
ER
ER_Tubular
Ergic/cisGolgi
Large Protein Complex
Lysosome
Mitochondrion
Plasma membrane

Fig. 7.4

Fig. 7.4 (A) An MR-plot where dark green lines are drawn at suggested threshold and hits are highlighted in orange. (B) BANDLE rank plot showing the distribution of differentially localised proteins. (C) The top differentially localised proteins from BANDLE plotted with uncertainty estimates. (D) Boostrap distributions of correlations with a phosphoprotemomic time-course experiment. The BANDLE confidence intervals differ significantly from 0, whilst the MR method do not. (E) PCA plots with (smoothed) localisation probabilities project onto them. Each colour represent an organelle and ellipses represent lines of isoprobability. The inner ellipse corresponds to 0.99 and the proceed line 0.95 with further lines decreasing by 0.05 each time. The protein are highlight demonstrating example relocalisations. EGFR (P005330) clearly relocalises from the PM to endosome, whilst SHC-1 (P29353) and GRB2 (P62993) relocalise from unknown localisation to the Lysosome.

BANDLE obtains deeper insights into AP-4 dependent localisation

The adaptor protein (AP) complexes are a set of heterotetrameric complexes, which transport transmembrane cargo protein vesicles [390]. The AP1-3 complexes are well characterised: AP-1 mediates the transport of lysosomal hydrolases from the trans-Golgi to the endsomes [239, 199]; AP-2 has a significant role in the regulation of endocytosis [323]; AP-3 is involved in the sorting of trans-Golgi proteins targeted to the lysosome [107]. However, the role of the AP-4 complex is still poorly understood [200, 201], despite loss-of-function mutations resulting in early-onset progressive spastic paraplegia [319].

AP-4 consists of four subunits ($\beta 4$, ε, $\mu 4$ and $\sigma 4$) forming an obligate complex [107]. Davies et al. [92] study the functional role of AP-4 using spatial proteomics; in particular, the DOM workflow mentioned previously. As part of their study, they use AP-4 CRISPR knockout cells to interrogate the effect on the spatial proteome when AP-4 function has been ablated.

The DOM experiment we re-analyse from [92] provides full quantitative measurements for 3926 proteins across two replicates of wild-type cells and two replicates where the $\beta 4$ subunit has been knocked-out. The data are visualised as PCA plots (see appendix) . As in the previous analysis, we run a quality control step removing the Actin binding protein and Nuclear pore complex annotations [157]. This dataset is particular challenging to analyse because there are only two replicates for control and treatment. The value of Bayesian analysis is the ability to provide prior information to regularise, as well as the quantification of uncertainty which is more critical in data sparse scenarios.

Previous application of the MR methods led to authors to find that SERINC 1 (Q9NRX5) and SERINC 3 (Q13530) were differentially localised [92]. Their results suggest that SERINC 1 and 3 are cargo proteins of the AP-4 complex that are packaged into vesicles at the trans-Golgi before being transported to the cell periphery. All together their results suggest AP-4 provides spatial regulation of autophagy and that AP-4 neurological pathology is linked to disturbances in membrane trafficking in neurons [297, 92].

We apply our method BANDLE in order to gain further insights into AP-4 dependent localisation. We compute the differential localisation probability and rank proteins according to this statistic (see figures 7.5 A and B). Characteristic S shape plots are observed with most proteins not differentially localisation upon knock-out of AP-4 $\beta 4$. The results of both SERINC 1 and 3 are validated, as we compute a differentially localisation probability greater than 0.95 for these proteins. Furthermore, 16 of the top 20 proteins are membrane-bound or membrane-associated proteins (FDR < 0.01 hyper-geometric test). To demonstrate the benefit of our probabilistic ranking, we perform two-sided KS rank test against the functional annotations provided in the STRING database (corrected for multiple testing within each functional framework) [443]. We find that processes such as ER to Golgi transport and lipid metabolism are more highly ranked that would be expected at random (FDR < 0.01), as well as

endosome and Golgi localisations (FDR < 0.01). Whilst processes associated with translation, ribosome localisation and function appear significantly lower in the ranking (FDR < 0.01). As expected, this provides a high level overview and evidence for the functional nature of AP-4 in the secretary pathway.

Taking a more precise view on our results, we examine the top 20 differentially localised proteins in more detail. We compute the Spearman correlation matrix between these proteins and observe clustering, suggesting the proteins act in a coordinated way (see figure 7.5 C). Visualising the data in a heatmap (figure 7.5 D), after mean and variance normalisation, we observe a highly concordant pattern: most proteins are enriched in fractions 4 and 5. These fractions are obtained from the highest centrifugation speeds and so differentially pellet light membrane organelles, such as endosomes and lysosomes [220, 159]. Again, further evidence for the role of AP-4 dependent localisation dynamics within the secretary pathway.

In figure 7.5 C, we observe a large cluster of 9 proteins, which includes SERINC 1 and 3. Amongst these 9 proteins is SLC38A2, a ubiquitously expressed amino-acid transporter that is widely express in the central nervous system and is recruited to the plasma membrane from a pool localised in the trans-Golgi [188, 30, 177, 306] Thus, its differential localisation here provides further evidence for the role of AP-4 as a membrane trafficker from the trans-Golgi. Another protein in this cluster is TMEM 199 (Q8N511) a protein of unknown function that is involved in lysosomal degradation [313]. Furthermore, it has been implicated in Golgi homoeostasis but the functional nature of this process is unknown [226]. Probing further, we observe that TMEM199 acts in a coordinated fashion with SERINC 1 and 3. Marked re-localisations are observed on PCA plots toward the endo/lysosomal regions (see figure 7.5 E) and we note that the quantitative profiles of SERINC 1, SERINC 3 and TMEM199 act in an analogous way upon AP-4 knockout (see figure 7.5 F). Our findings motivate additional studies to elucidate AP-4 dependent localisation.

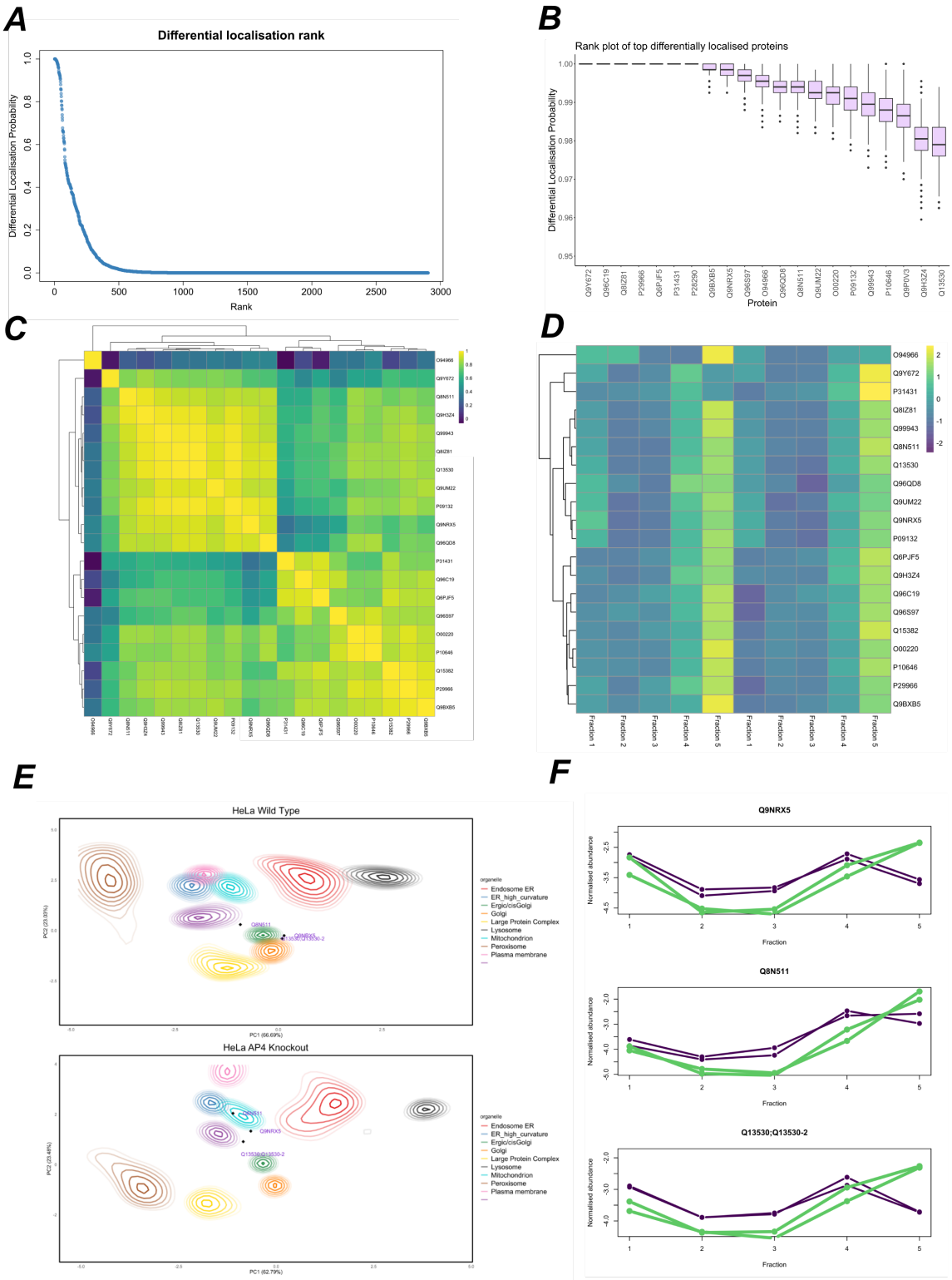

A
Differential localisation rank
Differential Localisation Probability
Rank
B
Rank plot of top differentially localised proteins
Differential Localisation Probability
Protein
C
D
E
HeLa Wild Type
HeLa AP4 Knockout
organelle
Endosome ER
ER_high_curvature
Ergic/cisGolgi
Golgi
Large Protein Complex
Lysosome
Mitochondrion
Peroxisome
Plasma membrane
F
Q9NRX5
Q8N511
Q13530;Q13530-2
Normalised abundance
Fraction

Fig. 7.5

Fig. 7.5 (A) BANDLE rank plot showing the distribution of differentially localised proteins. (B) The top differentially localised proteins from BANDLE plotted with uncertainty estimates. (C) A Spearman correlation heatmap showing strong correlations and coclustering behaviour of proteins that have AP-4 dependent localisation (D) Normalised mass-spectrometry profiles plotted as a heatmap from the AP-4 knockout data. Proteins are shown to have similar behaviour with greater intensity in fraction 5, where light membrane organelles are likely to pellet. (E) PCA plots with (smoothed) localisation probabilities project onto them. Each colour represent an organelle and ellipses represent lines of isoprobability. The inner ellipse corresponds to 0.99 and the proceed line 0.95 with further lines decreasing by 0.05 each time. The proteins SERINC 1 and 3, as well as TMEM199 are highlight demonstrating example relocalisations. (F) Normalised abundance profiles showing that SERINC 1, SERINC 3 and TMEM199 show similar behaviour upon knockout of AP-4.

7.4.4 Rewiring the proteome in response to Cytomegalovirus infection

The host spatial-temporal proteome

Human Cytomegalovirus (HCMV) infection is a ubiquitous herpesvirus that burdens the majority of the populous [56]. In healthy immune systems, HCMV establishes latent infection following initial viral communication [377] and reactivation can lead to serious pathology in certain imunno-compromised individuals [36]. HCMV has a highly expanded genome with vast capabilities to encode functional proteins [325, 437]. For the virus to succeed it carefully modulates cellular functions *en masse* [230].

Initial viral infection involves endocytosis of the virion into the cell [218], host machinery is then used to transport viral capsids into the nucleus [348]. Within the host nucleus viral transcription and genome replication occurs [312, 172, 236]. Meanwhile, other viral proteins are targeted to the secretory pathway to inhibit immune response and regulate the expression of viral genes [431, 126, 216, 316, 79, 266], rewire signalling pathways [491] and modulate metabolism [490]. These processes perform part of the early phases on the infection cycle. In later phases, the cellular trafficking pathways and the secretory organelles are hijacked for the formation of the viral assembly complex (vAC) [50, 318, 8, 89, 88]. Thus, HCMV biology is a paradigm to analyse complex viral processes [475].

There has been a recent flurry in applying system-wide proteomic approaches to the HCMV infection model. Weekes et al. [475] developed *quantitative temporal viromics* a multiplexed proteomic approach to understand the temporal response of thousands of cellular host and viral proteins. More recently, to discover proteins with innate immune function a multiplexed proteasome-lysosome degradation assay found that more than 100 proteins are degraded shortly after infection [340]. Meanwhile, a comprehensive mass spectrometry interactome analysis has identified thousands of host-virus interactions [343]. Furthermore, high-throughput temporal proteomic analysis has revealed acetylation, a lysine posttransational modification, as an integral component of HCMV infection [328].

Beltran et al. [24] use spatial and temporal proteomics to investigate the response of the human host proteome to HCMV infection. The authors perform subcellular fractionation on uninfected (control) and HCMV infected (treated) cells at 5 different time point $(24, 48, 72, 96, 120)$ hours post infection (hpi). The authors then used neural networks to classify proteins to subcellular niches at each time point in the control and treated cells, allowing a descriptive initial analysis of the data. Proteins with differential classification at each time point are those that are believed to be differentially localised. However, the challenge of this study is that only a single replicate is produced in each situation. This renders the MR method of Itzhak et al. [220] inapplicable.

Differential classification is a reasonable approach to probe differential localisation though it neglects information shared across both experiments and it is not quantitative. In the case of single replicates, by sharing information and providing prior information we are able to improve inference and obtain deeper insights. We apply BANDLE to control and HCMV-treated cells at 24 hpi, in the interest of brevity, to explore further the host spatial-temporal proteome. Our analysis reflects extensive rewiring of the proteome with hundreds of proteins differentially localised on HCMV infection. We highlight an example of differential localisation with SCARB1 (see figure 7.6 A), with a localisation in the secretory pathway shifting toward a PM/cytosolic localisation, similar to what has previously been observed [24].

To obtain global insights into the functional behaviour of the differentially localised proteins, we performed a Gene Ontology (GO) enrichment analysis. An extensive list of terms are enriched and these can be divided broadly into subcategories such as translation and transcription; transport; viral processes; and immune process (see appendix). These results reflect closely the early phase of HCMV infection [230]. Pathway enrichment analysis highlights terms related to a viral infection (Viral mRNA Translation, Influenza Life Cycle, Infectious disease, Innate Immune System, Immune System, MHC class II antigen presentation, Antigen processing-Cross presentation, Host Interactions of HIV factors, HIV Infection) (see figure 7.6 B). Pathway analysis also reveals known processes that are modulated on HCMV infection, such as membrane trafficking [40, 338, 492], Extracellular matrix organization [378] and rab regulation of trafficking [282].

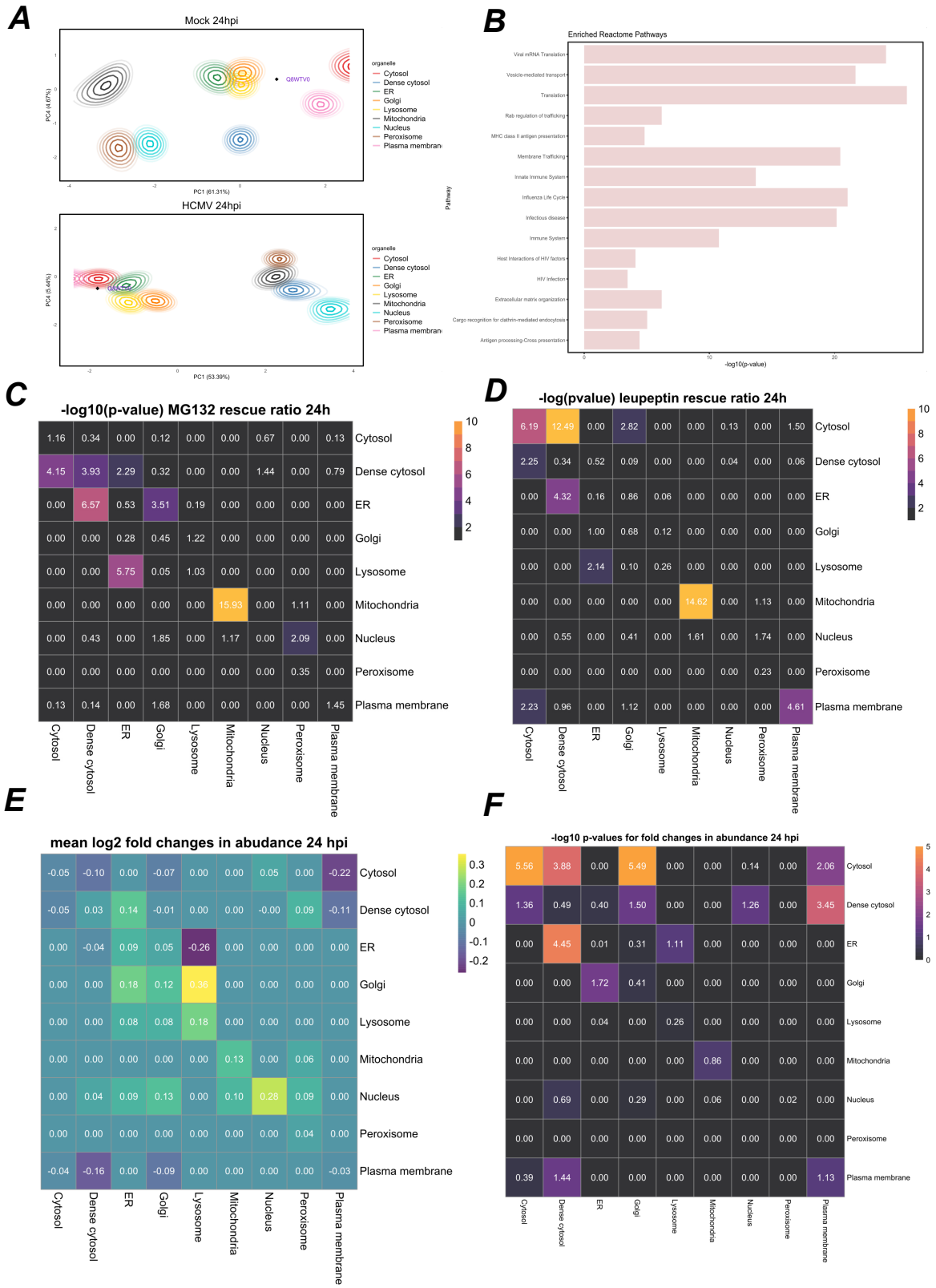

A
Mock 24hpi
HCMV 24hpi
organelle
Cytosol
Dense cytosol
ER
Golgi
Lysosome
Mitochondria
Nucleus
Peroxisome
Plasma membrane
PC1 (61.31%)
PC4 (4.67%)
PC1 (53.39%)
PC4 (5.44%)
B
Enriched Reactome Pathways
Viral mRNA Translation
Vesicle-mediated transport
Translation
Rab regulation of trafficking
MHC class II antigen presentation
Membrane Trafficking
Innate Immune System
Influenza Life Cycle
Infectious disease
Immune System
Host Interactions of HIV factors
HIV Infection
Extracellular matrix organization
Cargo recognition for clathrin-mediated endocytosis
Antigen processing-Cross presentation
Pathway
-log10(p-value)
C
-log10(p-value) MG132 rescue ratio 24h
D
-log(pvalue) leupeptin rescue ratio 24h
E
mean log2 fold changes in abudance 24 hpi
F
-log10 p-values for fold changes in abundance 24 hpi

Fig. 7.6

Fig. 7.6 (A) PCA plots with (smoothed) localisation probabilities project onto them. Each colour represent an organelle and ellipses represent lines of isoprobability. The inner ellipse corresponds to 0.99 and the proceed line 0.95 with further lines decreasing by 0.05 each time. The relocalisation of SCARB1 is highlighted on the plot (B) Reactome pathway enrichment results. (C) A heatmap representation of the MG132 inhibitor degradation data at 24 hpi. $\log_{10} p$-values are overlaid onto the spatial patterns across MOCK and HCMV infected cell 24hpi. The y-axis corresponds to localisation in the MOCK dataset whilst the x-axis corresponds to HCMV infected cells. (D) as for C but for the leupeptin inhibitor. (E) mean $\log_2$ abundance fold changes are overlaid on a heatmap according to their spatial pattern (F) the p-values corresponding to the fold changes observed in E.

Integrating HCMV proteomic datasets to add functional relevance to spatial proteomics data

The spatial information obtained here allows us to perform careful integration with other high-resolution proteomic datasets. The degradation screens by [340] identified proteins that were actively degraded during HCMV infection but gave no information regarding the spatial location of the targets. To determine the location of host proteins targeted by HCMV for degradation, the BANDLE revised spatial data at 24 hpi was overlapped with proteins that were degraded by the proteasome or lysosome. The subcellular location of the host proteins is displayed for the 24 h timepoint. To determine the spatial granularity of the degradation data we tested whether the proteins assigned to each spatial pattern had a significantly different degradation distribution that the distribution of all proteins in the experiment (t-test). We note that proteins that are differentially localised are no more likely to be targeted for degradation than those that are not (see appendix).

Degradation data from [340] are overlaid as a heatmap, showing a $-\log_{10}(p\text{-value})$ for each inhibitor (figure 7.6 C and D). For proteasomal targeted proteins (MG132), the data highlight a high number of proteins degraded from the mitochondria. The mitochondria act as a signalling platform for apoptosis and innate immunity and it is already well established that HCMV can subvert these processes to its advantage [87]. Furthermore, there is a high degree of protein degradation as one might expect in proteasome fractions (dense cytosol), with an enrichment of proteins recruited from the ER and cytosol (see appendix). For lysosomal targeted proteins (leupeptin) there was a high degree of proteins degraded from the mitochondria, cytosol and plasma membrane. There were also several proteins degraded that moved from the cytosol to the dense cytosol.

Many host proteins are up or down regulated upon HCMV infection [475]. We examine more recent abundance data from [328] at 24 hpi and first we note that differentially localised proteins are not more abundant than spatially stable proteins (see appendix). However, we see a strong spatial pattern when we overlay the abundance pattern on a heatmap. In figure 7.6 E , we report the mean $\log_2$ fold change for proteins stratified according to predicted subcellular localisation. It is important to combine spatial and abundance data, since a differentially localised protein may not undergo a true translocation event but rather a new pool of proteins is synthesised. The significance of these abundance changes is highlighted in figure 7.6 F. For example, there is a significant decrease in the abundance of the protein recruited to the dense cytosol from the ER (see appendix). Some of the larger changes are not significant because there are too few proteins with the same spatial pattern. We note that FAM3C, a protein involved in platelet degranulation, is upregulated at 24 hpi. Furthermore, FAM3C relocalises from the Golgi to the Lysosome, its Golgi localisation is in concordance with the Human Protein

Atlas (HPA) [455] and its Lysosome relocalisation suggests that it is trafficked through the secretory pathway before undergoing degranulation.

Upon integration of the acetylation data of [328], the spatial patterns are much more nuanced (see figures 7.7 A and B). Perhaps surprisingly, we do not observe increased acetylation levels amongst differentially localised proteins (see appendix). The only significant pattern is for proteins relocalising from the dense cytosol to the cytosol; however, we observe this is driven by a single protein Skp1 (see appendix), which shows a 2.5-fold increase in acetylation at 24 hpi for Skp1 and there is an increase in its RNA transcript at 24 hpi [340]. The Skp1 protein is part of an E3 ubiquitin ligase complex that targets proteins for degradation. E3 ligases are often manipulated by viruses in order to control cellular processes to create a cell states that benefit viral replication and survival [289]. It is therefore possible that HCMV is controlling Skp1 activity through acetylation at its C-terminus, leading to its translocation and likely change in function.

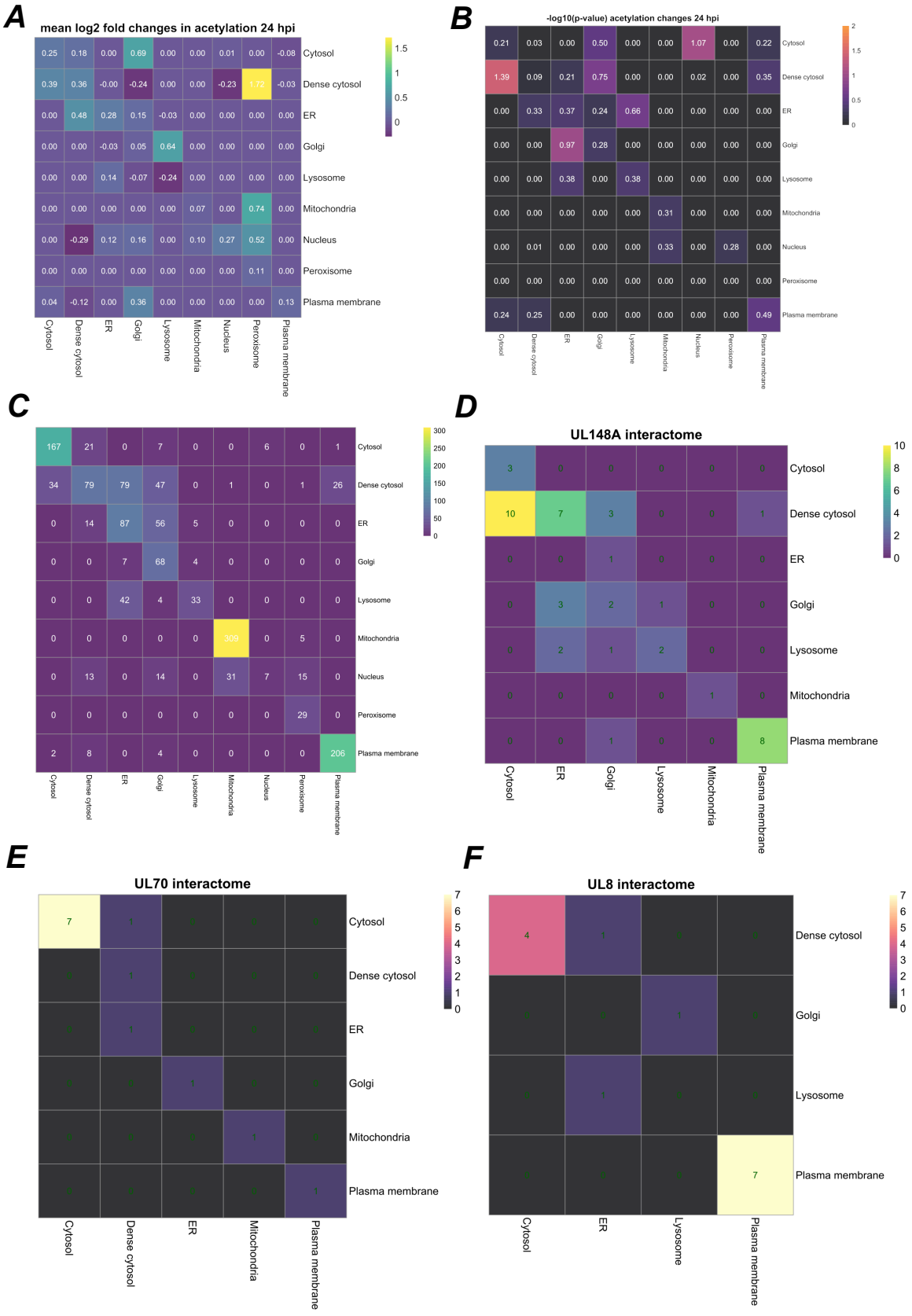

A
mean log2 fold changes in acetylation 24 hpi
B
-log10(p-value) acetylation changes 24 hpi
C
D
UL148A interactome
E
UL70 interactome
F
UL8 interactome
Cytosol
Dense cytosol
ER
Golgi
Lysosome
Mitochondria
Nucleus
Peroxisome
Plasma membrane

Fig. 7.7

Fig. 7.7 (A) A heatmap representation of the mean $\log_2$ fold changes in acetylation overlaid on spatial pattern of HCMV infection 24 hpi. (B) p-values for the changes shown in figure A. (C) The spatial allocation derived from BANDLE where each entry of the heatmap is the number of proteins. The y-axis represents localisation in the mock dataset and the x-axis localisation in the HCMV infected cells 24 hpi. (D) UL148A interactome mapped onto the BANDLE determined spatial patterns. (E) UL70 interactome mapped onto the BANDLE determined spatial patterns (F) UL8 interactome mapped onto the BANDLE determine spatial patterns.

The recent publication of the HCMV interactome has provided a wealth of data that gives insights into the function of the 170 canonical and 2 non-canonical viral protein-coding genes [343]. However, a common difficulty with analysing large interactome projects is the ability to reduce the number of false-positive interactions, leading to poor agreement between experimental and computational datasets. This can be controlled through replicates, supervised machine learning and increased statistical stringency; however, background contamination can never be eliminated. If a protein is located in a single location, it would be expected that true positive interactors to be located in the same subcellular compartment. Therefore, to narrow the list of viral-protein interactors, we overlapped spatial information from [24] with the viral interactors from [343] (figure 7.7 D,E,F).

We plot heatmaps to indicate the spatial distribution of the host proteins (figure 7.7). The overall distribution is plotted in the heatmap of figure 7.7 C. Firstly, we are interested in scenarios where the interacting host proteins were more likely to retain their localisation upon HCMV infection (than the computed posterior distribution would have predicted). Thus, for each viral bait, we simulated from a binomial $A \sim Bin(n, p)$ where p is the posterior probability that a random protein was assigned to the same localisation and n is the number of interactors of that viral bait. We then simulated from this distribution $5,000$ times to obtain a histogram (see appendix). Viral baits of interest are those were the observed statistic in the tails of these histograms.

Examples of such cases are shown for viral proteins UL8 and UL70 (see figure 7.7 E and F). The majority of UL8 interactors were located in the plasma membrane and cytosol. UL8 is a transmembrane protein that is transiently localised at the cell surface, with a small cytoplasmic pool [360], perfectly mimicking the location of the majority of UL8 interactors. Practically all UL70 interactors were located in the cytosol. Viral UL70 is a primase known to locate to both the nucleus and cytoplasmic compartments during HCMV infection [417]. As the nucleus was removed prior to fractionation then one expects only to be able to interrogate cytosolic interactors. An example were the host proteins were spatially diffuse was UL148A an elusive viral protein of unknown function, believed to be involved with modulating the innate immune response [90]. UL148A appears to interact with host proteins distributed throughout the cell suggesting it is highly promiscuous (figure 7.7 D). Perhaps UL148A is a moonlighting protein [231] making its function hard to pinpoint and such an observation would not be uncommon for viral proteins because of limited genomic size [70, 73]. These results illustrate the strength in overlapping spatial proteomics with interactome studies to decrease the number of false positives and focus research on higher confidence protein-protein interactions. The entire list of spatially resolved viral protein interactions is shown in the appendix.

7.5 Discussion and limitations

We have presented a Bayesian model for comparative and dynamic spatial proteomic experiments. Unlike current approaches, our flexible integrative mixture model allows any number of replicate experiments to be included. Furthermore, subcellular profiles are modelled separately for each condition and each replicate, allowing cases where the correlation profiles differ between experiments. Crucially, our model facilitates the computation of differential localisation probability, which cannot be performed by other methods in the literature. Furthermore, BANDLE probabilistically assigns proteins to organelles and can model outliers meaning that further supervised machine learning after application of BANDLE is not required. The probabilistic ranking obtained from BANDLE can be used for downstream pathway or GO enrichment analysis, likewise it can be mapped onto other high-throughput datasets.

We compared BANDLE to the MR approach of [220, 221]. The MR method is not as broadly applicable as BANDLE, and BANDLE does not require additional experiments to interpret the thresholds. In our careful simulation study, we demonstrate reduced Type 1 error and increased power when using our approach. In a further simulation, we demonstrated that BANDLE has more desirable statistical properties than the MR approach, the results are easier to interpret and more information is available. Since we are in a Bayesian framework, our approach also quantifies uncertainty.

Application of our approach to 3 dynamic and comparative mass-spectrometry based spatial proteomic experiments demonstrates the broad applicability of our approach. We validate many previously known findings in the literature, placing confidence in these results. When BANDLE was applied to EGF stimulation dataset, we saw increased correlation between our differential localisation results and a phosphoproteomic timecourse than when compared to the results of the MR approach.

We applied BANDLE to an AP-4 knockout dataset to investigate AP-4 dependant localisation and, as with other studies, we observe SERINC 1 and SERINC 3 are AP-4 Cargo. Furthermore, we implicate TMEM199 as potentially overlooked AP-4 cargo. We apply BANDLE to datasets where the MR approach is not applicable - an HCMV infection spatial proteomic dataset. Pathway and GO enrichment results implicate differentially localised protein in well-studied processes of early viral infection; such as, membrane trafficking and immune response.

We then carefully integrated several HCMV proteomic datasets and place a spatial perspective on these data, including proteins targeted for degradation, as well as abundance and acetylation dataset. In addition, we augment a recent HCMV interactome by placing it in its spatial context and note that most host protein interactomes are in the same localisation as their viral bait. This provides an excellent resource for the community and highlights the benefit of

integrating spatial proteomics and interactomics datasets. This analysis also reveals potential moonlighting proteins.

Our analysis here highlights the potential role for post-translational modifications (PTMs) and their influence on localisation. The current datasets are limited because the spatial information is averaged over different PTMs. Thus, it is vital to develop methods to obtain spatial PTM information and develop corresponding computational tools to analyse these data. In this scenario, a testing approaching might be more realistic and there is a fear that our computational methods will not scale to peptide-centric studies. Furthermore, our approach here can only look at a single condition at a time. In the future, more complex spatial proteomics designs will be available that will study multiple perturbations simultaneously. A clear limitation of our work is not being able to analysis dynamic experiments that depend on additional covariates.

Overall, differential localisation experiments seek to add an orthogonal perspective to other assays, such as classical high-throughput differential abundance testing. Currently, differential localisation has not been extensively explored in high-throughput. We hope rigorous statistical methods will spur extensive and illuminating applications.

Another limitation of our analysis is that integration with other datasets happens in a multiple step approach, feeding our results into the output of our methods. From a statistical point of view sharing information across datasets in an integrative approach is clearly desirable. Modelling covariate based designs and integration with other data sources also, again, raises the question of scalability of our computational methods. We have also assumed independence between biological replicates, but note that statistical dependence structure could be modelled using hierarchical Gaussian processes [198]. However, this will also reduce the scalability of our approach.

Chapter 8

Conclusion

Christian De Duve's principle of fractionating the cell and associating proteins within subcellular niches by their shared profiles across subcellular fractions, has been turned into a powerful profiling spatial proteomic technique. Coupling either density-gradient or differential centrifugation to high-accuracy mass spectrometry allows for high-throughput interrogation of the spatial proteome. The modern inceptions of De Duve's principle not only provide localisation information on thousands of proteins but they have also uncovered significant levels of multi-localisation and hence potentially multifunctional proteins [159]. The methodology has also been refashioned to study protein dynamic in the form of trans-locations or, more precisely, differential localisation.

The data analysis challenges posed by mass spectrometry based spatial proteomics has been well documented [154, 155, 157]. Machine learning algorithms trained on marker proteins have a number of limitations and often cannot be easily manipulated to answer more challenging questions. In this thesis, we explored Bayesian approaches to the spatial proteomics problem both in the parametric and non-parametric framework. The Bayesian treatment allows for quantification of uncertainty through the posterior distribution of the parameters and of latent variables. As mass spectrometry based spatial proteomics becomes more prominent and it is extended to more complex designs and questions, we believe the use of Bayesian modelling will form a key part of the process.

8.1 Main findings and contributions

In chapter 2, we developed a class of semi-supervised robust Bayesian mixture models. We showed that the predictive output of these models matches that of the state-of-the-art machine learning algorithms currently applied to spatial proteomics data. Application to many datasets showed that our model was flexible enough to perform well in diverse experimental designs. Moreover, our model provides more information. Using MCMC sampling we can sample from

the posterior distribution of localisation probabilities. Through a detailed case study on mouse pluripotent embryonic stem cells, we showed that the uncertainty quantification is biologically meaningful, allowing us to report on potential cases of multi-localisation. We developed ways to summarise the uncertainty, as well as clear visualisations allowing a thorough treatment of the spatial proteomics data. We also highlight a more general interpretation of our model, which could useful in applications to other datasets. We continued along this line in chapter 3, in which we walked through the analysis of a typical spatial proteomics dataset and provided an easy to use software implementation for our approach. We developed further visualisations and provided a primer on Bayesian analysis with an unversed practitioner in mind. This allowed a more extensive discussion of the model priors.

Chapter 4 presented a substantial application of our Bayesian model to *Toxoplasma gondii*. Since *T. gondii* is not a model organism most of its functional annotation arises from various indirect sources, such as homology with other organisms [21]. Many proteins are still classed as "hypothetical" from the genome. Many knowledge gaps are overcome in our analysis, where we can confidently allocate thousands of proteins to 26 distinct subcellular niches. This allows us to integrate other datasets with the data. One example is a genome-wide CRISPR-CAS9 knockout screen, in which we identify, for example, that the apicoplast is enriched for indispensable proteins. Uncertainty quantification also allows us to explore some aspects of localisation dynamics. We also highlight some major limitations of our approach in light on this application.

Chapters 5 and 6 explore extensions to our original model. In chapter 5, we first discuss several approaches for selecting or inferring the number of clusters in mixture models. We opt to apply the method of *over fitted* mixtures. This can be considered an additional semi-supervised extension of our model, which allows us to perform novelty detection. We show how we can perform uncertainty quantification in this model using the discovery probability. Furthermore, application to 10 spatial proteomics datasets, covering a broad range of biological systems, provides new putative annotations in every dataset. Moving forward this will allow spatial proteomics to be applied to poorly annotated organisms and reduce the reliance on marker proteins. On the other hand, chapter 6 sets out to develop a model that more closely captures the data generating mechanisms. The increased computation of this Bayesian non-parametric approach is offset by the development of several matrix algorithms that extend classical methods. Extensive simulations demonstrate that our model is largely robust to prior choices and can make more accurate predictions than previous models.

In chapter 7, we rigorously defined the concept of differential localisation - a fundamental biological phenomenon. We review current data analysis methods and show that they have poor statistical properties and restrictive assumptions. This motivates a Bayesian model for uncertainty quantification in differential localisation experiments. We show that our method reduces false positives and increases power over current methods. Furthermore, it has clearly

motivated statistical properties and is straightforward to interpret. We developed several visualisations to help interrogate the data. Application to three case studies shows the benefit of our approach and we uncover new differential localisations in several high profile experiments.

8.2 Limitations and future work

In this final section, we summarise the limitations of this thesis and suggest some directions for future research.

8.2.1 Theoretical and empirical properties of mixed mixtures

If we defined a *mixed mixture* model as a mixture model with parametric components with an additional parametric term from a family with heavier tails than the other components, we can ask questions about its theoretical properties and empirical behaviour. In particular, one would wish to obtain results as in Coretto and Hennig [75] beyond frequentist estimators in Gaussian mixture models. Finite sample robustness properties in the Bayesian setting for these models would a valuable contribution and is currently missing from the literature. Extensive empirical comparisons, especially in the context of likelihood misspecification would also be valuable. We have not established these results for our models and this is a clear limitation of our work.

8.2.2 Missing values

Our models cannot currently handle missing values. The main motivation for TMT multiplexing is that it reduced missing values; however, this is only partially true, since in different replicates one still observes missingness between different batches. Imputation is a general strategy for handling missing values but the optimal strategy will depend on experimental design [263]. Missing values in the Bayesian framework can be handled in two ways. The first is to introduce latent indicators of missingness and treat them as values to be inferred [170]. The second is to restrict to the values of interest for the required computations so that missing data does not contribute to the likelihood [170]. Either method could be implemented within our model.

8.2.3 Hierarchical models

Throughout this thesis, we have assumed that biological replicates are statistically independent. Of course, they are independent experiments; however, they share useful information. A hierarchical model could be developed to share information across replicated experiments. The challenge here is that the density gradients are not exactly the same and so the integration might involve some registration or careful assumptions at that level of the hierarchy.

8.2.4 Sub-niche resolution

Some subcellular niches show sub-clustering behaviour. For example the ER might be split into lumen or membrane components. One strategy to investigate sub-clustering would be to take the output of our model and consider each cluster in turn. A mixture model could then be fitted to each component again in turn and the number of components inferred. If more than one sub-cluster is observed then that might be evidence for sub-organeller resolution. However, multiple components could also arise purely as an effect of model mis-specification. One might wish to compare the marginal likelihood with a scale-skew t-distribution as a strategy to diagnose potential mis-specification. One could also approach this within the model itself by building a mixture of mixture models, though there are severe identifiability issues with such models [292].

8.2.5 Protein-protein interaction and protein complexes

There is visual evidence that interacting proteins or protein complexes exhibit co-behaviour in spatial proteomics data [324]. Computational strategies could be developed to allow the probability of two proteins interacting given spatial proteomics data to be computed. The challenge here is demonstrating that the results are valid and would require extensive external validation. Some work in this direction is present in the literature [113, 403] but no methods have been put proposed to deduce protein complexes from LOPIT data.

8.2.6 Computation

For increasingly large datasets our approaches can require excessive computation. Reimplementing in a low-level language could alleviate this problem, as well as approximate Bayesian methods, such as variational inference [34], could be employed.

8.2.7 Data integration

None of the models we have presented can integrate datasets of different modalities that provide complementary information on protein subcellular localisation. However, there is clear evidence that there is utility in such approaches [44]. One strategy could be to use the multiple dataset integration framework of Kirk et al. [243], extended to the semi-supervised setting.

8.2.8 Summarisation of raw data

A mass-spectrometry based spatial proteomics dataset actually measures peptide spectrum matches (PSMs), which are quickly summarised to proteins. However, in summarising, one might average over two protein isoforms that have different sub-cellular localisation or similarly

average over different modified version of the protein, each of which with localisations. There are several ways to tackle this problem. One way would be to use a model based summarisation approach and propagate the uncertainty to the protein level or qualitatively explore the behaviour of proteins with uncertainty in their quantitation. Another approach would be to apply the models at the PSM level - though this could place a strain on computation. This thesis has not explored issues that arise because of uncertainty in the protein quantitation. Furthermore, this is a general problem in all of proteomics and warrants substantial attention. Let y_{ijkl} be log PSM quantitation for protein i in sample j for peptide k for PSM l. Then consider the following linear Bayesian model

$$y_{ijkl} = \beta_{ij}^{\text{protein}} + \beta_{ik}^{\text{peptide}} + \beta_{ikl}^{\text{PSM}} + \varepsilon_{ijkl} \tag{8.1}$$

$$\beta_{ij}^{\text{protein}} \sim \mathcal{N}(0, \sigma_{ij}^{\text{Protein}}) \tag{8.2}$$

$$\beta_{ik}^{\text{peptide}} \sim \mathcal{N}(0, \sigma_{ik}^{\text{Peptide}}) \tag{8.3}$$

$$\beta_{ikjl}^{\text{PSM}} \sim \mathcal{N}(0, \sigma_{ikl}^{\text{PSM}}) \tag{8.4}$$

$$\varepsilon_{ijkl} \sim \mathcal{T}(0, \sigma_{ijkl}), \tag{8.5}$$

$$\tag{8.6}$$

where suitable priors, such as folded Normal distributions, are placed on the variances of the noise terms. One could then use MCMC to sample from this model. The Monte-Carlo estimator for $\beta_{ij}^{\text{protein}}$ quantifies the amount of protein i in sample j. This could then be used for downstream analysis. However, one could also obtain samples from the posterior distribution $q \sim p(\beta_{ij}^{\text{protein}}|Y)$. The downstream analysis could then be performed for all values of this posterior and we could observe the variation in the final quantities of interest. Though this would be a computationally intensive process, the results could be illuminating.

8.2.9 Subcellular localisation of post-translational modifications

An under explored area is the effect of post-translational modification on subcellular localisation, despite the well studied case of phosphorylation. Using a variety of different enrichment approaches including titanium dioxide metal cation chelates, it is possible to extend spatial proteomics methods to allow the quantitation of phosphopetides and other antibody enrichments could be used for other modifications [252]. The goal in this scenario would be to test whether the profiles were different for the modified peptide and non-modified form. A testing approach is permissible because the same gradient is used for organelle separation and subsequent enrichment. A Bayesian semi-parametric two-sample test could be developed to tackle this question. More precisely, a shared model, where non-modified and modified peptides share the same functional profile would be compared to an independent model, where non-modified

and modified peptides have different models. These models would be nested and the standard Bayesian approach to hypothesis testing could be used [164]. Extension to multiple simultaneous modifications would also be of interest.

8.2.10 Differential localisation with multiple perturbations

Linear models are frequently used when one is interested in multiple contrasts, that is comparing multiple perturbations. For example, we might be interested in differential localisation at three different stages of the cell cycle. Currently, our differential localisation model cannot handle this scenario and can only perform pairwise comparisons. Now, let us denote $z_{i,j}$ the localisation of protein i in experiment j. We would be interested in the following quantities of interest $p(z_{i,1} \neq z_{i,2})$, $p(z_{i,2} \neq z_{i,3})$, $p(z_{i,1} \neq z_{i,3})$. Naively, we could generate a mixture model over the three experiments; however if there were K subcellular niches this would require K^3 parameters. With careful prior choices this would be possible but a new, bespoke approach might be better starting point.

8.2.11 Differential localisation with temporal perturbations

Time course experiments are performed frequently in systems biology and additive models are usually employed to handle these situations. Again, our differential localisation model does not model these scenarios. To formalise the setting, let t_1, t_2 and t_3 be three time points at which we have interest. For example, times after infection or cellular heat shock. One possibility would be to assume a Markov structure on the allocations at these times. For example, we could assume $p(z_{i,t_1}|X_{t_1}, X_{t_2}, X_{t_3}) = p(z_{i,t_1}|X_{t_1})$ and $p(z_{i,t_2}|X_{t_1}, X_{t_2}, X_{t_3}) = p(z_{i,t_2}|X_{t_2}, z_{i,1})$, $p(z_{i,t_3}|X_{t_1}, X_{t_2}, X_{t_3}) = p(z_{i,t_3}|X_{t_3}, z_{i,2})$. Other independence structures might also be possible.

8.2.12 Differential localisation with covariates

The previous two scenarios can be summarised more compactly as differential localisation with covariates. Let β be some covariate either continuous or discrete, for example space, time, temperature, life cycle stage and many more. The independence structure of the localisation with respect to the covariate is the modelling aspect that is challenging in this setting. That is, what is the independence structure of the following probability $p(z_{i,\beta(s)}|X_{\beta(S/s)}, z_{i,\beta(S/s)})$, where S is some indexing set and S/s indicates the set excluding s.

Future experiments might also combine several functional proteomics techniques into a single experiment to obtain, for example, subcellular structural information. These and the limitations we have mentioned are the subject of future work.

References

[1] Enzyme-linked immunosorbent assay, ELISA: III. Quantitation of specific antibodies by enzyme-labeled anti-immunoglobulin in antigen-coated tubes.

[2] Ruedi Aebersold and Matthias Mann. Mass spectrometry-based proteomics. *Nature*, 422 (6928):198–207, 2003.

[3] Ruedi Aebersold, Jeffrey N Agar, I Jonathan Amster, Mark S Baker, Carolyn R Bertozzi, Emily S Boja, Catherine E Costello, Benjamin F Cravatt, Catherine Fenselau, Benjamin A Garcia, et al. How many human proteoforms are there? *Nature chemical biology*, 14(3): 206, 2018.

[4] J Aitchison and Sheng M Shen. Logistic-normal distributions: Some properties and uses. *Biometrika*, 67(2):261–272, 1980.

[5] John Aitchison. The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2):139–160, 1982.

[6] Hirotugu Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974.

[7] Bruce Alberts, Dennis Bray, Karen Hopkin, Alexander D Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. *Essential cell biology*. Garland Science, 2013.

[8] James C Alwine. The human cytomegalovirus assembly compartment: a masterpiece of viral manipulation of cellular processes that facilitates assembly and egress. *PLoS pathogens*, 8(9), 2012.

[9] Giovanna FL Ames and Kishiko Nikaido. Two-dimensional gel electrophoresis of membrane proteins. *Biochemistry*, 15(3):616–623, 1976.

[10] Christophe Andrieu, Nando De Freitas, Arnaud Doucet, and Michael I Jordan. An introduction to MCMC for machine learning. *Machine learning*, 50(1-2):5–43, 2003.

[11] Charles E. Antoniak. Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems. *Ann. Statist.*, 2(6):1152–1174, 11 1974. doi: 10.1214/aos/1176342871. URL http://dx.doi.org/10.1214/aos/1176342871.

[12] Toshiyuki Araki and Jeffrey Milbrandt. ZNRF proteins constitute a family of presynaptic E3 ubiquitin ligases. *Journal of Neuroscience*, 23(28):9385–9394, 2003.

[13] Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404, 1950.

[14] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

[15] Kriti Bahl, Shuwei Xie, Gaelle Spagnol, Paul Sorgen, Naava Naslavsky, and Steve Caplan. EHD3 protein is required for tubular recycling endosome stabilization, and an asparagine-glutamic acid residue pair within its Eps15 homology (EH) domain dictates its selective binding to NPF peptides. *Journal of Biological Chemistry*, 291(26):13465–13478, 2016.

[16] Monya Baker. Blame it on the antibodies. *Nature*, 521(7552):274, 2015.

[17] Elli-Anna Balta, Marie-Theres Wittmann, Matthias Jung, Elisabeth Sock, Benjamin Martin Haeberle, Birgit Heim, Felix von Zweydorf, Jana Heppt, Julia von Wittgenstein, Christian Johannes Gloeckner, et al. Phosphorylation modulates the subcellular localization of SOX11. *Frontiers in molecular neuroscience*, 11:211, 2018.

[18] Anjishnu Banerjee, Jared Murray, and David Dunson. Bayesian learning of joint distributions of objects. In *Artificial Intelligence and Statistics*, pages 1–9, 2013.

[19] Jeffrey D Banfield and Adrian E Raftery. Model-based gaussian and non-gaussian clustering. *Biometrics*, pages 803–821, 1993.

[20] Marcus Bantscheff, Markus Schirle, Gavain Sweetman, Jens Rick, and Bernhard Kuster. Quantitative mass spectrometry in proteomics: a critical review. *Analytical and bioanalytical chemistry*, 389(4):1017–1031, 2007.

[21] Konstantin Barylyuk, Ludek Koreny, Huiling Ke, Simon Butterworth, Oliver M. Crook, Imen Lassadi, Vipul Gupta, Eelco Tromer, Tobias Mourier, Tim J. Stevens, Lisa M. Breckels, Arnab Pain, Kathryn S. Lilley, and Ross F. Waller. A subcellular atlas of Toxoplasma reveals the functional context of the proteome. *bioRxiv*, 2020. doi: 10.1101/2020.04.23.057125.

[22] John Robert Baxter and Jeffrey S Rosenthal. Rates of convergence for everywhere-positive Markov chains. *Statistics & probability letters*, 22(4):333–338, 1995.

[23] Archana Belle, Amos Tanay, Ledion Bitincka, Ron Shamir, and Erin K O'Shea. Quantification of protein half-lives in the budding yeast proteome. *Proceedings of the National Academy of Sciences*, 103(35):13004–13009, 2006.

[24] Pierre M Jean Beltran, Rommel A Mathias, and Ileana M Cristea. A portrait of the human organelle proteome in space and time during cytomegalovirus infection. *Cell systems*, 3(4):361–373, 2016.

[25] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300, 1995.

[26] James O Berger, Victor De Oliveira, and Bruno Sansó. Objective Bayesian analysis of spatially correlated data. *Journal of the American Statistical Association*, 96(456): 1361–1374, 2001.

[27] Alexandros Beskos, Natesh Pillai, Gareth Roberts, Jesus-Maria Sanz-Serna, and Andrew Stuart. Optimal tuning of the hybrid Monte Carlo algorithm. *Bernoulli*, 19(5A):1501–1534, 11 2013. doi: 10.3150/12-BEJ414. URL https://doi.org/10.3150/12-BEJ414.

[28] Michael Betancourt. A conceptual introduction to Hamiltonian Monte Carlo. *arXiv preprint arXiv:1701.02434*, 2017.

[29] Michael Betancourt. Calibrating model-based inferences and decisions. *arXiv preprint arXiv:1803.08393*, 2018.

[30] Elena Bevilacqua, Ovidio Bussolati, Valeria Dall Asta, Francesca Gaccioli, Roberto Sala, Gian C Gazzola, and Renata Franchi-Gazzola. SNAT2 silencing prevents the osmotic induction of transport system A and hinders cell recovery from hypertonic stress. *FEBS letters*, 579(16):3376–3380, 2005.

[31] Harish S Bhat and Nitesh Kumar. On the derivation of the Bayesian Information Criterion. *technical report*, 2010.

[32] Christophe Biernacki, Gilles Celeux, and Gérard Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE transactions on pattern analysis and machine intelligence*, 22(7):719–725, 2000.

[33] David Blackwell, James B MacQueen, et al. Ferguson distributions via Pólya urn schemes. *The annals of statistics*, 1(2):353–355, 1973.

[34] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

[35] Günter Blobel. Christian de duve (1917–2013), 2013.

[36] Michael Boeckh and W Garrett Nichols. The impact of cytomegalovirus serostatus of donor and recipient before hematopoietic stem cell transplantation in the era of antiviral prophylaxis and preemptive therapy. *Blood*, 103(6):2003–2008, 2004.

[37] Ingwer Borg and Patrick Groenen. Modern multidimensional scaling: Theory and applications. *Journal of Educational Measurement*, 40(3):277–280, 2003.

[38] Charles Bouveyron, Etienne Côme, Julien Jacques, et al. The discriminative functional mixture model for a comparative analysis of bike sharing systems. *The Annals of Applied Statistics*, 9(4):1726–1760, 2015.

[39] Elizabeth I Boyle, Shuai Weng, Jeremy Gollub, Heng Jin, David Botstein, J Michael Cherry, and Gavin Sherlock. GO::TermFinder – open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics*, 20(18):3710–3715, 2004.

[40] Petros Bozidis, Chad D Williamson, Daniel S Wong, and Anamaris M Colberg-Poley. Trafficking of UL37 proteins into mitochondrion-associated membranes during permissive human cytomegalovirus infection. *Journal of virology*, 84(15):7898–7903, 2010.

[41] Allan Bradley, Martin Evans, Matthew H Kaufman, and Elizabeth Robertson. Formation of germ-line chimaeras from embryo-derived teratocarcinoma cell lines. *Nature*, 309(5965): 255–256, 1984.

[42] Tess C Branon, Justin A Bosch, Ariana D Sanchez, Namrata D Udeshi, Tanya Svinkina, Steven A Carr, Jessica L Feldman, Norbert Perrimon, and Alice Y Ting. Efficient proximity labeling in living cells and organisms with TurboID. *Nature biotechnology*, 36 (9):880–887, 2018.

[43] Lisa M Breckels, Laurent Gatto, Andy Christoforou, Arnoud J Groen, Kathryn S Lilley, and Matthew WB Trotter. The effect of organelle discovery upon sub-cellular protein localisation. *Journal of proteomics*, 88:129–140, 2013.

[44] Lisa M Breckels, Sean B Holden, David Wojnar, Claire M Mulvey, Andy Christoforou, Arnoud Groen, Matthew WB Trotter, Oliver Kohlbacher, Kathryn S Lilley, and Laurent Gatto. Learning from heterogeneous data sources: an application in spatial proteomics. *PLoS computational biology*, 12(5):e1004920, 2016.

[45] Lisa M Breckels, Claire M Mulvey, Kathryn S Lilley, and Laurent Gatto. A Bioconductor workflow for processing and analysing spatial proteomics data. *F1000Research*, 5, 2016.

[46] Stephen P Brooks and Andrew Gelman. General methods for monitoring convergence of iterative simulations. *Journal of computational and graphical statistics*, 7(4):434–455, 1998.

[47] Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng. *Handbook of markov chain monte carlo*. CRC press, 2011.

[48] Markus Brosch, Lu Yu, Tim Hubbard, and Jyoti Choudhary. Accurate and sensitive peptide identification with Mascot Percolator. *Journal of proteome research*, 8(6):3176–3181, 2009.

[49] Morton B Brown. A method for combining non-independent, one-sided tests of significance. *Biometrics*, pages 987–992, 1975.

[50] Nicholas J Buchkovich, Tobi G Maguire, and James C Alwine. Role of the endoplasmic reticulum chaperone BiP, SUN domain proteins, and dynein in altering nuclear morphology during human cytomegalovirus infection. *Journal of virology*, 84(14):7005–7017, 2010.

[51] Shannon M Buckley, Beatriz Aranda-Orgilles, Alexandros Strikoudis, Effie Apostolou, Evangelia Loizou, Kelly Moran-Crusio, Charles L Farnsworth, Antonius A Koller, Ramanuj Dasgupta, Jeffrey C Silva, et al. Regulation of pluripotency and cellular reprogramming by the ubiquitin-proteasome system. *Cell stem cell*, 11(6):783–798, 2012.

[52] Catherine A Bue, Christine M Bentivoglio, and Charles Barlowe. Erv26p directs pro-alkaline phosphatase into endoplasmic reticulum–derived coat protein complex ii transport vesicles. *Molecular biology of the cell*, 17(11):4780–4789, 2006.

[53] Ellen Bushell, Ana Rita Gomes, Theo Sanderson, Burcu Anar, Gareth Girling, Colin Herd, Tom Metcalf, Katarzyna Modrzynska, Frank Schwach, Rowena E Martin, et al. Functional profiling of a Plasmodium genome reveals an abundance of essential genes. *Cell*, 170(2):260–272, 2017.

[54] Or Cabasso, Olga Pekar, and Mia Horowitz. SUMOylation of EHD3 modulates tubulation of the endocytic recycling compartment. *PloS one*, 10(7):e0134053, 2015.

[55] Ben J Callahan, Kris Sankaran, Julia A Fukuyama, Paul J McMurdie, and Susan P Holmes. Bioconductor workflow for microbiome data analysis: from raw reads to community analyses. *F1000Research*, 5, 2016.

[56] Michael J Cannon, D Scott Schmid, and Terri B Hyde. Review of cytomegalovirus seroprevalence and demographic characteristics associated with infection. *Reviews in medical virology*, 20(4):202–213, 2010.

[57] Corinna Cappellaro, Vladimir Mrsa, and Widmar Tanner. New Potential Cell Wall Glucanases ofSaccharomyces cerevisiae and Their Involvement in Mating. *Journal of bacteriology*, 180(19):5030–5037, 1998.

[58] Annalisa Carlucci, Monia Porpora, Corrado Garbi, Mario Galgani, Margherita Santoriello, Massimo Mascolo, Domenico di Lorenzo, Vincenzo Altieri, Maria Quarto, Luigi Terracciano, et al. PTPD1 supports receptor stability and mitogenic signaling in bladder cancer cells. *Journal of biological chemistry*, 285(50):39260–39270, 2010.

[59] Vern B Carruthers and LD Sibley. Sequential protein secretion from three distinct organelles of toxoplasma gondii accompanies invasion of human fibroblasts. *European journal of cell biology*, 73(2):114–123, 1997.

[60] George Casella and Christian P Robert. Rao-Blackwellisation of sampling schemes. *Biometrika*, 83(1):81–94, 1996.

[61] Antoine Chambaz, Judith Rousseau, et al. Bounds for Bayesian order identification with application to mixtures. *The Annals of Statistics*, 36(2):938–962, 2008.

[62] Ian Chambers and Simon R Tomlinson. The transcriptional foundation of pluripotency. *Development*, 136(14):2311–2322, 2009.

[63] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):1–27, 2011.

[64] Pin-I Chen, Kristine Schauer, Chen Kong, Andrew R Harding, Bruno Goud, and Philip D Stahl. Rab5 isoforms orchestrate a "division of labor" in the endocytic network; Rab5C modulates Rac-mediated cell motility. *PloS one*, 9(2):e90384, 2014.

[65] Hee Min Choi, James P Hobert, et al. The Polya-Gamma Gibbs sampler for Bayesian logistic regression is uniformly ergodic. *Electronic Journal of Statistics*, 7:2054–2064, 2013.

[66] Hyungwon Choi, Sinae Kim, Anne-Claude Gingras, and Alexey I Nesvizhskii. Analysis of protein complexes through model-based biclustering of label-free quantitative ap-ms data. *Molecular Systems Biology*, 6(1):385, 2010. doi: 10.1038/msb.2010.41. URL https://onlinelibrary.wiley.com/doi/abs/10.1038/msb.2010.41.

[67] Frank Christian, Emma L Smith, and Ruaidhrí J Carmody. The regulation of NF-κB subunits by phosphorylation. *Cells*, 5(1):12, 2016.

[68] Andy Christoforou, Claire M Mulvey, Lisa M Breckels, Aikaterini Geladaki, Tracey Hurrell, Penelope C Hayward, Thomas Naake, Laurent Gatto, Rosa Viner, Alfonso Martinez Arias, et al. A draft map of the mouse pluripotent stem cell spatial proteome. *Nature communications*, 7:9992, 2016.

[69] Neal AL Cody, Carole Iampietro, and Eric Lécuyer. The many functions of mRNA localization during normal development and disease: from pillar to post. *Wiley Interdisciplinary Reviews: Developmental Biology*, 2(6):781–796, 2013.

[70] Jonathan D Cook and Jeffrey E Lee. The secret life of viral entry glycoproteins: moonlighting in immune evasion. *PLoS pathogens*, 9(5), 2013.

[71] Emma J Cooke, Richard S Savage, Paul DW Kirk, Robert Darkins, and David L Wild. Bayesian hierarchical clustering for microarray time series data with replicates and outlier measurements. *BMC bioinformatics*, 12(1):399, 2011.

[72] Albert H Coons, Hugh J Creech, R Norman Jones, and Ernst Berliner. The demonstration of pneumococcal antigen in tissues by the use of fluorescent antibody. *The Journal of Immunology*, 45(3):159–170, 1942.

[73] Shelley D Copley. An evolutionary perspective on protein moonlighting, 2014.

[74] Adrian Corduneanu and Christopher M Bishop. Variational Bayesian model selection for mixture distributions.

[75] Pietro Coretto and Christian Hennig. Robust improper maximum likelihood: tuning, computation, and a comparison with other methods for robust gaussian clustering. *Journal of the American Statistical Association*, 111(516):1648–1659, 2016.

[76] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3): 273–297, 1995.

[77] Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27, 1967.

[78] Jurgen Cox, Nadin Neuhauser, Annette Michalski, Richard A Scheltema, Jesper V Olsen, and Matthias Mann. Andromeda: a peptide search engine integrated into the MaxQuant environment. *Journal of proteome research*, 10(4):1794–1805, 2011.

[79] Ileana M Cristea, Nathaniel J Moorman, Scott S Terhune, Christian D Cuevas, Erin S O'Keefe, Michael P Rout, Brian T Chait, and Thomas Shenk. Human cytomegalovirus pUL83 stimulates activity of the viral immediate-early promoter through its interaction with the cellular IFI16 protein. *Journal of virology*, 84(15):7803–7814, 2010.

[80] Nello Cristianini, John Shawe-Taylor, et al. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.

[81] Oliver Crook, Aikaterini Geladaki, Daniel JH Nightingale, Owen Vennard, Kathryn Susan Lilley, Laurent Gatto, and Paul DW Kirk. A semi-supervised bayesian approach for simultaneous protein sub-cellular localisation assignment and novelty detection. *bioRxiv*, 2020.

[82] Oliver M Crook, Tom Smith, Mohamed Elzek, and Kathryn S Lilley. Moving Profiling Spatial Proteomics Beyond Discrete Classification. *Proteomics*, page 1900392.

[83] Oliver M Crook, Claire M Mulvey, Paul DW Kirk, Kathryn S Lilley, and Laurent Gatto. A Bayesian mixture modelling approach for spatial proteomics. *PLoS computational biology*, 14(11), 2018.

[84] Oliver M Crook, Lisa M Breckels, Kathryn S Lilley, Paul DW Kirk, and Laurent Gatto. A Bioconductor workflow for the Bayesian analysis of spatial proteomics. *F1000Research*, 8, 2019.

[85] Oliver M Crook, Laurent Gatto, and Paul DW Kirk. Fast approximate inference for variable selection in Dirichlet process mixtures, with an application to pan-cancer proteomics. *Statistical Applications in Genetics and Molecular Biology*, 18(6), 2019.

[86] Oliver M Crook, Kathryn S Lilley, Laurent Gatto, and Paul DW Kirk. Semi-Supervised Non-Parametric Bayesian Modelling of Spatial Proteomics. *arXiv preprint arXiv:1903.02909*, 2019.

[87] Marni S Crow, Krystal K Lum, Xinlei Sheng, Bokai Song, and Ileana M Cristea. Diverse mechanisms evolved by DNA viruses to inhibit early host defenses. *Critical reviews in biochemistry and molecular biology*, 51(6):452–481, 2016.

[88] Subhendu Das and Philip E Pellett. Spatial relationships between markers for secretory and endosomal machinery in human cytomegalovirus-infected cells versus those in uninfected cells. *Journal of virology*, 85(12):5864–5879, 2011.

[89] Subhendu Das, Amit Vasanji, and Philip E Pellett. Three-dimensional structure of the human cytomegalovirus cytoplasmic virion assembly complex includes a reoriented secretory apparatus. *Journal of virology*, 81(21):11861–11869, 2007.

[90] Liat Dassa, Einat Seidel, Esther Oiknine-Djian, Rachel Yamin, Dana G Wolf, Vu Thuy Khanh Le-Trilling, and Ofer Mandelboim. The human cytomegalovirus protein UL148A downregulates the NK cell-activating ligand MICA to avoid NK cell attack. *Journal of virology*, 92(17):e00162–18, 2018.

[91] Alexandra K Davies, Daniel N Itzhak, James R Edgar, Tara L Archuleta, Jennifer Hirst, Lauren P Jackson, Margaret S Robinson, and Georg HH Borner. AP-4 vesicles contribute to spatial control of autophagy via RUSC-dependent peripheral delivery of ATG9A. *Nature Communications*, 9:3958, 2018.

[92] Alexandra K Davies, Daniel N Itzhak, James R Edgar, Tara L Archuleta, Jennifer Hirst, Lauren P Jackson, Margaret S Robinson, and Georg HH Borner. AP-4 vesicles contribute to spatial control of autophagy via RUSC-dependent peripheral delivery of ATG9A. *Nature communications*, 9(1):3958, 2018.

[93] Anthony C Davison, David V Hinkley, and G Alastair Young. Recent developments in bootstrap methodology. *Statistical Science*, pages 141–157, 2003.

[94] Christian De Duve. The lysosome. *Scientific American*, 208(5):64–73, 1963.

[95] Christian de Duve. Principles of tissue fractionation. *Journal of Theoretical Biology*, 6(1): 33–59, 1964.

[96] Christian De Duve. Functions of microbodies (peroxisomes). *J. Cell Biol.*, 27:25A–26A, 1965.

[97] Christian de Duve. Tissue fraction-past and present. *The Journal of Cell Biology*, 50(1): 20, 1971.

[98] Christian De Duve and Henri Beaufay. A short history of tissue fractionation. *The Journal of cell biology*, 91(3):293, 1981.

[99] Christian De Duve and Robert Wattiaux. Functions of lysosomes. *Annual review of physiology*, 28(1):435–492, 1966.

[100] Christian De Duve, BvC Pressman, R Gianetto, R Wattiaux, and Françoise Appelmans. Tissue fractionation studies. 6. Intracellular distribution patterns of enzymes in rat-liver tissue. *Biochemical Journal*, 60(4):604–617, 1955.

[101] Christian De Duve, Thierry De Barsy, Brian Poole, Paul Tulkens, et al. Lysosomotropic agents. *Biochemical pharmacology*, 23(18):2495–2531, 1974.

[102] Maria Antonietta De Matteis and Alberto Luini. Mendelian disorders of membrane trafficking. *New England Journal of Medicine*, 365(10):927–938, 2011.

[103] Victor De Oliveira. Objective Bayesian analysis of spatial data with measurement error. *Canadian Journal of Statistics*, 35(2):283–301, 2007.

[104] Gillian Barbara Dealtry and David Rickwood. *Cell biology labfax*. Distributed in the United States and Canada by Academic Press, 1992.

[105] Pierre Del Moral, Arnaud Doucet, and Ajay Jasra. Sequential monte carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):411–436, 2006.

[106] Marizela Delic, Minoska Valli, Alexandra B Graf, Martin Pfeffer, Diethard Mattanovich, and Brigitte Gasser. The secretory pathway: exploring yeast diversity. *FEMS microbiology reviews*, 37(6):872–914, 2013.

[107] Esteban C Dell'Angelica, Judith Klumperman, Willem Stoorvogel, and Juan S Bonifacino. Association of the AP-3 adaptor complex with clathrin. *Science*, 280(5362):431–434, 1998.

[108] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

[109] Silvia Munoz Descalzo, Pau Rué, Fernando Faunes, Penelope Hayward, Lars Martin Jakt, Tina Balayo, Jordi Garcia-Ojalvo, and Alfonso Martinez Arias. A competitive protein interaction network buffers Oct4-mediated differentiation to promote pluripotency in embryonic stem cells. *Molecular systems biology*, 9(1), 2013.

[110] Jean Diebolt and Christian P Robert. Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(2):363–375, 1994.

[111] Meik Dilcher, Beate Köhler, and Gabriele Fischer von Mollard. Genetic Interactions with the Yeast Q-SNARE VTI1Reveal Novel Functions for the R-SNARE YKT6. *Journal of Biological Chemistry*, 276(37):34537–34544, 2001.

[112] Arnaud Doucet, Nando De Freitas, and Neil Gordon. An introduction to sequential Monte Carlo methods. In *Sequential Monte Carlo methods in practice*, pages 3–14. Springer, 2001.

[113] Kevin Drew, Chanjae Lee, Ryan L Huizar, Fan Tu, Blake Borgeson, Claire D McWhite, Yun Ma, John B Wallingford, and Edward M Marcotte. Integration of over 9,000 mass spectrometry experiments builds a global map of human protein complexes. *Molecular systems biology*, 13(6):932, 2017.

[114] Mathias Drton and Martyn Plummer. A Bayesian information criterion for singular models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79 (2):323–380, 2017.

[115] Simon Duane, Anthony D Kennedy, Brian J Pendleton, and Duncan Roweth. Hybrid monte carlo. *Physics letters B*, 195(2):216–222, 1987.

[116] JP Dubey, DS Lindsay, and CA Speer. Structures of Toxoplasma gondiitachyzoites, bradyzoites, and sporozoites and biology and development of tissue cysts. *Clinical microbiology reviews*, 11(2):267–299, 1998.

[117] JF Dubremetz, Nathalie Garcia-Réguet, Valérie Conseil, and Marie Noëlle Fourmaux. Apical organelles and host-cell invasion by Apicomplexa. *International journal for parasitology*, 28(7):1007–1013, 1998.

[118] Tom PJ Dunkley, Rod Watson, Julian L Griffin, Paul Dupree, and Kathryn S Lilley. Localization of organelle proteins by isotope tagging (LOPIT). *Molecular & Cellular Proteomics*, 3(11):1128–1134, 2004.

[119] Tom PJ Dunkley, Svenja Hester, Ian P Shadforth, John Runions, Thilo Weimar, Sally L Hanton, Julian L Griffin, Conrad Bessant, Federica Brandizzi, Chris Hawes, et al. Mapping the arabidopsis organelle proteome. *Proceedings of the National Academy of Sciences*, 103(17):6518–6523, 2006.

[120] C de Duve. The peroxisome: a new cytoplasmic organelle. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 173(1030):71–83, 1969.

[121] Florence Dzierszinski, Manami Nishi, Lillian Ouko, and David S Roos. Dynamics of Toxoplasma gondii differentiation. *Eukaryotic cell*, 3(4):992–1003, 2004.

[122] Dirk Eddelbuettel and Romain Francois. Rcpp: Seamless R and C++ Integration. *Journal of Statistical Software, Articles*, 40(8):1–18, 2011. ISSN 1548-7660. doi: 10.18637/jss.v040.i08. URL https://www.jstatsoft.org/v040/i08.

[123] Dirk Eddelbuettel and Conrad Sanderson. RcppArmadillo: Accelerating R with High-performance C++ Linear Algebra. *Comput. Stat. Data Anal.*, 71:1054–1063, 2014. ISSN 0167-9473. doi: 10.1016/j.csda.2013.02.005. URL http://dx.doi.org/10.1016/j.csda.2013.02.005.

[124] Bradley Efron. *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, volume 1. Cambridge University Press, 2012.

[125] Michael D Escobar and Mike West. Bayesian density estimation and inference using mixtures. *Journal of the american statistical association*, 90(430):577–588, 1995.

[126] Xuyan Feng, Jörg Schröer, Dong Yu, and Thomas Shenk. Human cytomegalovirus pUS24 is a virion protein that functions very early in the replication cycle. *Journal of virology*, 80(17):8371–8378, 2006.

[127] John B Fenn, Matthias Mann, Chin Kai Meng, Shek Fu Wong, and Craig M Whitehouse. Electrospray ionization for mass spectrometry of large biomolecules. *Science*, 246(4926):64–71, 1989.

[128] John B Fenn, Matthias Mann, Chin Kai Meng, Shek Fu Wong, and Craige M Whitehouse. Electrospray ionization–principles and practice. *Mass Spectrometry Reviews*, 9(1):37–70, 1990.

[129] DJP Ferguson and WM Hutchison. An ultrastructural study of the early development and tissue cyst formation of Toxoplasma gondii in the brains of mice. *Parasitology research*, 73(6):483–491, 1987.

[130] DJP Ferguson, A Birch-Andersen, J Chr Siim, and WM Hutchison. Observations on the ultrastructure of the sporocyst and the initiation of sporozoite formation in Toxoplasma gondii. *Acta Pathologica Microbiologica Scandinavica Section B Microbiology*, 86(1-6):165–168, 1978.

[131] Thomas S. Ferguson. Prior Distributions on Spaces of Probability Measures. *Ann. Statist.*, 2(4):615–629, 07 1974. doi: 10.1214/aos/1176342752. URL http://dx.doi.org/10.1214/aos/1176342752.

[132] Nicola Festuccia, Rodrigo Osorno, Valerie Wilson, and Ian Chambers. The role of pluripotency gene regulatory network components in mediating transitions between pluripotent cell states. *Current opinion in genetics & development*, 23(5):504–511, 2013.

[133] Scott B Ficarro, Mark L McCleland, P Todd Stukenberg, Daniel J Burke, Mark M Ross, Jeffrey Shabanowitz, Donald F Hunt, and Forest M White. Phosphoproteome analysis by mass spectrometry and its application to Saccharomyces cerevisiae. *Nature biotechnology*, 20(3):301–305, 2002.

[134] Evelyn Fix. *Discriminatory analysis: nonparametric discrimination, consistency properties*. USAF school of Aviation Medicine, 1951.

[135] Leonard J Foster, Carmen L de Hoog, Yanling Zhang, Yong Zhang, Xiaohui Xie, Vamsi K Mootha, and Matthias Mann. A mammalian organelle map by protein correlation profiling. *Cell*, 125(1):187–199, 2006.

[136] Bernardo J Foth and Geoffrey I McFadden. The apicoplast: a plastid in Plasmodium falciparum and other Apicomplexan parasites. *International review of cytology*, 224: 57–110, 2003.

[137] Chris Fraley and Adrian E Raftery. How many clusters? which clustering method? answers via model-based cluster analysis. *The computer journal*, 41(8):578–588, 1998.

[138] Chris Fraley and Adrian E Raftery. Bayesian regularization for normal mixture estimation and model-based clustering. Technical report, Washington Univ Seattle Dept of Statistics, 2005.

[139] Chris Fraley and Adrian E Raftery. Bayesian regularization for normal mixture estimation and model-based clustering. *Journal of classification*, 24(2):155–181, 2007.

[140] Chris Fraley, Adrian E Raftery, T Brendan Murphy, and Luca Scrucca. mclust version 4 for r: Normal mixture modeling for model-based clustering, classification, and density estimation. 2012.

[141] Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2):95–110, 1956.

[142] Arno Fritsch, Katja Ickstadt, et al. Improved criteria for clustering based on the posterior similarity matrix. *Bayesian analysis*, 4(2):367–391, 2009.

[143] Geir-Arne Fuglstad, Daniel Simpson, Finn Lindgren, and Håvard Rue. Constructing priors that penalize the complexity of Gaussian random fields. *Journal of the American Statistical Association*, 114(525):445–452, 2019.

[144] Masayoshi Fukasawa, Oleg Varlamov, William S Eng, Thomas H Söllner, and James E Rothman. Localization and activity of the SNARE Ykt6 determined by its regulatory domain and palmitoylation. *Proceedings of the National Academy of Sciences*, 101(14): 4815–4820, 2004.

[145] Jairo Fúquene, Mark Steel, and David Rossell. On choosing mixture components via non-local priors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(5):809–837, 2019.

[146] Evelina Gabasova, John Reid, and Lorenz Wernisch. Clusternomics: Integrative context-dependent clustering for heterogeneous datasets. *PLoS computational biology*, 13(10): e1005781, 2017.

[147] Jonah Gabry, Daniel Simpson, Aki Vehtari, Michael Betancourt, and Andrew Gelman. Visualization in Bayesian workflow. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 182(2):389–402, 2019.

[148] Michael D Gallagher and Alice S Chen-Plotkin. The post-GWAS era: from association to function. *The American Journal of Human Genetics*, 102(5):717–730, 2018.

[149] Marıa Teresa Gallegos. Clustering in the presence of outliers. In *Exploratory Data Analysis in Empirical Research*, pages 58–66. Springer, 2003.

[150] María Teresa Gallegos, Gunter Ritter, et al. A robust method for cluster analysis. *The Annals of Statistics*, 33(1):347–380, 2005.

[151] Dani Gamerman and Hedibert F Lopes. *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. Chapman and Hall/CRC, 2006.

[152] Luis Ángel García-Escudero and Alfonso Gordaliza. Robustness properties of k means and trimmed k means. *Journal of the American Statistical Association*, 94(447):956–969, 1999.

[153] Laurent Gatto and Kathryn Lilley. MSnbase - an R/Bioconductor package for isobaric tagged mass spectrometry data visualization, processing and quantitation. *Bioinformatics*, 28:288–289, 2012.

[154] Laurent Gatto, Juan Antonio Vizcaíno, Henning Hermjakob, Wolfgang Huber, and Kathryn S Lilley. Organelle proteomics experimental designs and analysis. *Proteomics*, 10(22):3957–3969, 2010.

[155] Laurent Gatto, Lisa M Breckels, Thomas Burger, Daniel JH Nightingale, Arnoud J Groen, Callum Campbell, Claire M Mulvey, Andy Christoforou, Myriam Ferro, and Kathryn S Lilley. A foundation for reliable spatial proteomics data analysis. *Molecular & Cellular Proteomics*, pages mcp–M113, 2014.

[156] Laurent Gatto, Lisa M. Breckels, Samuel Wieczorek, Thomas Burger, and Kathryn S. Lilley. Mass-spectrometry based spatial proteomics data analysis using pRoloc and pRolocdata. *Bioinformatics*, 2014.

[157] Laurent Gatto, Lisa M Breckels, and Kathryn S Lilley. Assessing sub-cellular resolution in spatial proteomics experiments. *Current Opinion in Chemical Biology*, 48:123–149, 2019.

[158] Alexis Gautreau, Ksenia Oguievetskaia, and Christian Ungermann. Function and regulation of the endosomal fusion and fission machineries. *Cold Spring Harbor perspectives in biology*, 6(3):a016832, 2014.

[159] Aikaterini Geladaki, Nina Kocevar Britovsek, Lisa M Breckels, Tom Sand Owen L Vennard Smith, Claire M Mulvey, Oliver M Crook, Laurent Gatto, and Kathryn S Lilley. Combining LOPIT with differential ultracentrifugation for high-resolution spatial proteomics. *Nature Communications*, 10:331, 2019.

[160] Alan E Gelfand and Adrian FM Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, 85(410):398–409, 1990.

[161] Alan E Gelfand, Athanasios Kottas, and Steven N MacEachern. Bayesian nonparametric spatial modeling with Dirichlet process mixing. *Journal of the American Statistical Association*, 100(471):1021–1035, 2005.

[162] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman & Hall, London, 1995.

[163] Andrew Gelman and Donald B Rubin. Inference from iterative simulation using multiple sequences. *Statistical science*, pages 457–472, 1992.

[164] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 2013.

[165] Andrew Gelman et al. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian analysis*, 1(3):515–534, 2006.

[166] Robert C Gentleman, Vincent J Carey, Douglas M Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, Laurent Gautier, Yongchao Ge, Jeff Gentry, et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, 5(10):R80, 2004.

[167] Manju George, GuoGuang Ying, Mark A Rainey, Aharon Solomon, Pankit T Parikh, Qingshen Gao, Vimla Band, and Hamid Band. Shared as well as distinct roles of EHD proteins revealed by biochemical and functional comparisons in mammalian cells and C. elegans. *BMC cell biology*, 8(1):3, 2007.

[168] John Geweke. Bayesian inference in econometric models using Monte Carlo integration. *Econometrica: Journal of the Econometric Society*, pages 1317–1339, 1989.

[169] John Geweke. Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments. *BAYESIAN STATISTICS*, 1992.

[170] Zoubin Ghahramani and Michael I Jordan. Learning from incomplete data. 1995.

[171] Toby J Gibson. Cell regulation: determined to signal discrete cooperation. *Trends in biochemical sciences*, 34(10):471–482, 2009.

[172] W Gibson. Structure and formation of the cytomegalovirus virion. In *Human cytomegalovirus*, pages 187–204. Springer, 2008.

[173] Walter R Gilks, Sylvia Richardson, and David Spiegelhalter. *Markov chain Monte Carlo in practice*. Chapman and Hall/CRC, 1995.

[174] Mark Girolami and Ben Calderhead. Riemann manifold langevin and hamiltonian monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214, 2011.

[175] Tilmann Gneiting. Nonseparable, stationary covariance functions for space–time data. *Journal of the American Statistical Association*, 97(458):590–600, 2002.

[176] Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

[177] IM Gonzalez-Gonzalez, B Cubelos, C Gimenez, and F Zafra. Immunohistochemical localization of the amino acid transporter SNAT2 in the rat brain. *Neuroscience*, 130(1): 61–73, 2005.

[178] IJ Good. Significance Tests in Parallel and in Series. *Journal of the American Statistical Association*, 53(284):799–813, 1958.

[179] Dov Greenbaum, Christopher Colangelo, Kenneth Williams, and Mark Gerstein. Comparing protein abundance and mRNA expression levels on a genomic scale. *Genome biology*, 4(9):1–8, 2003.

[180] Geoffrey Grimmett, Geoffrey R Grimmett, David Stirzaker, et al. *Probability and random processes*. Oxford university press, 2001.

[181] Arnoud J Groen, Gloria Sancho-Andreés, Lisa M Breckels, Laurent Gatto, Fernando Aniento, and Kathryn S Lilley. Identification of trans-Golgi network proteins in Arabidopsis thaliana root tissue. *Journal of proteome research*, 13(2):763–776, 2014.

[182] Kristin M Hager, Boris Striepen, Lewis G Tilney, and David S Roos. The nuclear envelope serves as an intermediary between the ER and Golgi complex in the intracellular parasite Toxoplasma gondii. *Journal of cell science*, 112(16):2631–2638, 1999.

[183] Mohamed-Ali Hakimi, Philipp Olias, and L David Sibley. Toxoplasma effectors targeting host signaling and transcription. *Clinical microbiology reviews*, 30(3):615–645, 2017.

[184] Peter Hall, Byeong U Park, Richard J Samworth, et al. Choice of neighbor order in nearest-neighbor classification. *The Annals of Statistics*, 36(5):2135–2152, 2008.

[185] Stephanie L Hall, Svenja Hester, Julian L Griffin, Kathryn S Lilley, and Antony P Jackson. The organelle proteome of the dt40 lymphocyte cell line. *Molecular & Cellular Proteomics*, 8(6):1295–1305, 2009.

[186] Jan Hartmann, Ke Hu, Cynthia Y He, Laurence Pelletier, David S Roos, and Graham Warren. Golgi and centrosome cycles in Toxoplasma gondii. *Molecular & Biochemical Parasitology*, 1(145):125–127, 2006.

[187] W Keith Hastings. Monte Carlo sampling methods using Markov chains and their applications. 1970.

[188] Takahiro Hatanaka, Wei Huang, Haiping Wang, Mitsuru Sugawara, Puttur D Prasad, Frederick H Leibach, and Vadivel Ganapathy. Primary structure, functional characteristics and tissue expression pattern of human ATA2, a subtype of amino acid transport system A. *Biochimica et Biophysica Acta (BBA)-Biomembranes*, 1467(1):1–6, 2000.

[189] Arie H Havelaar, Martyn D Kirk, Paul R Torgerson, Herman J Gibb, Tine Hald, Robin J Lake, Nicolas Praet, David C Bellinger, Nilanthi R De Silva, Neyla Gargouri, et al. World Health Organization global estimates and regional comparisons of the burden of foodborne disease in 2010. *PLoS medicine*, 12(12):e1001923, 2015.

[190] Hussein Hazimeh and ChengXiang Zhai. Axiomatic analysis of smoothing methods in language models for pseudo-relevance feedback. *Proceedings of the 2015 International Conference on The Theory of Information Retrieval*, pages 141–150, 2015.

[191] Cynthia Y He, Boris Striepen, Charles H Pletcher, John M Murray, and David S Roos. Targeting and processing of nuclear-encoded apicoplast proteins in plastid segregation mutants of Toxoplasma gondii. *Journal of Biological Chemistry*, 276(30):28436–28442, 2001.

[192] Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009.

[193] Xin He, Chris K Fuller, Yi Song, Qingying Meng, Bin Zhang, Xia Yang, and Hao Li. Sherlock: detecting gene-disease associations by matching patterns of expression QTL and GWAS. *The American Journal of Human Genetics*, 92(5):667–680, 2013.

[194] William Heard, Jan Sklenář, Daniel FA Tome, Silke Robatzek, and Alexandra ME Jones. Identification of regulatory and cargo proteins of endosomal and secretory pathways in arabidopsis thaliana by proteomic dissection. *Molecular & Cellular Proteomics*, 14(7): 1796–1813, 2015.

[195] Markus Heinonen, Olivier Guipaud, Fabien Milliat, Valérie Buard, Béatrice Micheau, Georges Tarlet, Marc Benderitter, Farida Zehraoui, and Florence d'Alché Buc. Detecting time periods of differential gene expression using Gaussian processes: an application to endothelial cells exposed to radiotherapy dose fraction. *Bioinformatics*, 31(5):728–735, 2014.

[196] Yuji Henmi, Natsuko Oe, Nozomu Kono, Tomohiko Taguchi, Kohji Takei, and Kenji Tanabe. Phosphatidic acid induces EHD3-containing membrane tubulation and is required for receptor recycling. *Experimental cell research*, 342(1):1–10, 2016.

[197] Christian Hennig et al. Breakdown points for maximum likelihood estimators of location–scale mixtures. *The Annals of Statistics*, 32(4):1313–1340, 2004.

[198] James Hensman, Neil D Lawrence, and Magnus Rattray. Hierarchical bayesian modelling of gene expression time series across irregularly sampled replicates and clusters. *BMC bioinformatics*, 14(1):252, 2013.

[199] Jochen Hess, Peter Angel, and Marina Schorpp-Kistner. AP-1 subunits: quarrel and harmony among siblings. *Journal of cell science*, 117(25):5965–5973, 2004.

[200] Jennifer Hirst, Nicholas A Bright, Brian Rous, and Margaret S Robinson. Characterization of a fourth adaptor-related protein complex. *Molecular biology of the cell*, 10(8):2787–2802, 1999.

[201] Jennifer Hirst, Carol Irving, and Georg HH Borner. Adaptor protein complexes AP-4 and AP-5: new players in endosomal trafficking and progressive spastic paraplegia. *Traffic*, 14 (2):153–164, 2013.

[202] Jennifer Hirst, Daniel N Itzhak, Robin Antrobus, Georg HH Borner, and Margaret S Robinson. Role of the AP-5 adaptor protein complex in late endosome-to-Golgi retrieval. *PLoS biology*, 16(1):e2004411, 2018.

[203] Sebastian Hoepfner, Fedor Severin, Alicia Cabezas, Bianca Habermann, Anja Runge, David Gillooly, Harald Stenmark, and Marino Zerial. Modulation of receptor recycling and degradation by the endosomal kinesin KIF16B. *Cell*, 121(3):437–450, 2005.

[204] Susan Holmes and Wolfgang Huber. *Modern statistics for modern biology*. Cambridge University Press, 2018.

[205] Antti Honkela, Charles Girardot, E Hilary Gustafson, Ya-Hsin Liu, Eileen EM Furlong, Neil D Lawrence, and Magnus Rattray. Model-based method for transcription factor target identification with limited data. *Proceedings of the National Academy of Sciences*, 107(17):7793–7798, 2010.

[206] Alan M Horowitz. A generalized guided Monte Carlo algorithm. *Physics Letters B*, 268 (2):247–252, 1991.

[207] Gerta Hoxhaj, Edward Caddye, Ayaz Najafov, Vanessa P Houde, Catherine Johnson, Kumara Dissanayake, Rachel Toth, David G Campbell, Alan R Prescott, and Carol MacKintosh. The E3 ubiquitin ligase ZNRF2 is a substrate of mTORC1 and regulates its activation by amino acids. *elife*, 5:e12278, 2016.

[208] Chih-wei Hsu, Chih-chung Chang, and Chih-jen Lin. A practical guide to support vector classification, 2010.

[209] Ke Hu, Tara Mann, Boris Striepen, Con JM Beckers, David S Roos, and John M Murray. Daughter cell assembly in the protozoan parasite Toxoplasma gondii. *Molecular biology of the cell*, 13(2):593–606, 2002.

[210] Ke Hu, Jeff Johnson, Laurence Florens, Martin Fraunholz, Sapna Suravajjala, Camille DiLullo, John Yates, David S Roos, and John M Murray. Cytoskeletal components of an invasion machine—the apical complex of Toxoplasma gondii. *PLoS pathogens*, 2(2), 2006.

[211] Qizhi Hu, Robert J Noll, Hongyan Li, Alexander Makarov, Mark Hardman, and R Graham Cooks. The Orbitrap: a new mass spectrometer. *Journal of mass spectrometry*, 40(4):430–443, 2005.

[212] Wolfgang Huber, Vincent J Carey, Robert Gentleman, Simon Anders, Marc Carlson, Benilton S Carvalho, Hector Corrada Bravo, Sean Davis, Laurent Gatto, Thomas Girke, et al. Orchestrating high-throughput genomic analysis with bioconductor. *Nature methods*, 12(2):115–121, 2015.

[213] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2 (1):193–218, 1985.

[214] Mia Hubert and Michiel Debruyne. Minimum covariance determinant. *Wiley interdisciplinary reviews: Computational statistics*, 2(1):36–43, 2010.

[215] Won-Ki Huh, James V Falvo, Luke C Gerke, Adam S Carroll, Russell W Howson, Jonathan S Weissman, and Erin K O'shea. Global analysis of protein localization in budding yeast. *Nature*, 425(6959):686, 2003.

[216] Jiwon Hwang and Robert F Kalejta. Proteasome-dependent, ubiquitin-independent degradation of Daxx by the viral pp71 protein in human cytomegalovirus-infected cells. *Virology*, 367(2):334–338, 2007.

[217] Hironori Inadome, Yoichi Noda, Hiroyuki Adachi, and Koji Yoda. Immunoisolaton of the yeast Golgi subcompartments and characterization of a novel membrane protein, Svp26, discovered in the Sed5-containing compartments. *Molecular and cellular biology*, 25(17): 7696–7710, 2005.

[218] MK Isaacson, LK Juckem, and T Compton. Virus entry and innate immune activation. In *Human Cytomegalovirus*, pages 85–100. Springer, 2008.

[219] Hemant Ishwaran and Lancelot F James. Generalized weighted Chinese restaurant processes for species sampling mixture models. *Statistica Sinica*, pages 1211–1235, 2003.

[220] Daniel N Itzhak, Stefka Tyanova, Jürgen Cox, and Georg HH Borner. Global, quantitative and dynamic mapping of protein subcellular localization. *Elife*, 5:e16950, 2016.

[221] Daniel N Itzhak, Colin Davies, Stefka Tyanova, Archana Mishra, James Williamson, Robin Antrobus, Jürgen Cox, Michael P Weekes, and Georg HH Borner. A Mass Spectrometry-Based Approach for Mapping Protein Subcellular Localization Reveals the Spatial Proteome of Mouse Primary Neurons. *Cell reports*, 20(11):2706–2718, 2017.

[222] Michel Jadot, Marielle Boonen, Jaqueline Thirion, Nan Wang, Jinchuan Xing, Caifeng Zhao, Abla Tannous, Meiqian Qian, Haiyan Zheng, John K Everett, et al. Accounting for protein subcellular localization: A compartmental map of the rat liver proteome. *Molecular & Cellular Proteomics*, 16(2):194–212, 2017.

[223] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer.

[224] Gareth M James and Trevor J Hastie. Functional linear discriminant analysis for irregularly sampled curves. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3):533–550, 2001.

[225] Gareth M James and Catherine A Sugar. Clustering for sparsely sampled functional data. *Journal of the American Statistical Association*, 98(462):397–408, 2003.

[226] Jos C Jansen, Sharita Timal, Monique Van Scherpenzeel, Helen Michelakakis, Dorothée Vicogne, Angel Ashikov, Marina Moraitou, Alexander Hoischen, Karin Huijben, Gerry Steenbergen, et al. TMEM199 deficiency is a disorder of Golgi homeostasis characterized by elevated aminotransferases, alkaline phosphatase, and cholesterol and abnormal glycosylation. *The American Journal of Human Genetics*, 98(2):322–330, 2016.

[227] Søren F Jarner and Gareth O Roberts. Polynomial convergence rates of Markov chains. *The Annals of Applied Probability*, 12(1):224–247, 2002.

[228] Søren Fiig Jarner and Ernst Hansen. Geometric ergodicity of Metropolis algorithms. *Stochastic processes and their applications*, 85(2):341–361, 2000.

[229] Ajay Jasra, Chris C Holmes, and David A Stephens. Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statistical Science*, pages 50–67, 2005.

[230] Pierre M Jean Beltran and Ileana M Cristea. The life cycle and pathogenesis of human cytomegalovirus infection: lessons from proteomics. *Expert review of proteomics*, 11(6): 697–711, 2014.

[231] Constance J Jeffery. Moonlighting proteins - an update. *Molecular BioSystems*, 5(4): 345–350, 2009.

[232] Harold Jeffreys. *The theory of probability*. OUP Oxford, 1998.

[233] Keith A Joiner and David S Roos. Secretory traffic in the eukaryotic parasite Toxoplasma gondii: less is more. *The Journal of cell biology*, 157(4):557–563, 2002.

[234] MC Jones and John A Rice. Displaying the important features of large collections of similar curves. *The American Statistician*, 46(2):140–145, 1992.

[235] Alfredo A Kalaitzis and Neil D Lawrence. A simple approach to ranking differentially expressed gene expression time courses through gaussian process regression. *BMC bioinformatics*, 12(1):180, 2011.

[236] RF Kalejta. Functions of human cytomegalovirus tegument proteins prior to immediate early gene expression. In *Human cytomegalovirus*, pages 101–115. Springer, 2008.

[237] Kaniav Kamary, Kerrie Mengersen, Christian P Robert, and Judith Rousseau. Testing hypotheses via a mixture estimation model. *arXiv preprint arXiv:1412.2044*, 2014.

[238] Alexandros Karatzoglou, Alex Smola, Kurt Hornik, and Achim Zeileis. kernlab-an S4 package for kernel methods in R. *Journal of statistical software*, 11(9):1–20, 2004.

[239] Michael Karin, Zheng-gang Liu, and Ebrahim Zandi. AP-1 function and regulation. *Current opinion in cell biology*, 9(2):240–246, 1997.

[240] Tweeny R Kau, Jeffrey C Way, and Pamela A Silver. Nuclear transport and cancer: from mechanism to intervention. *Nature Reviews Cancer*, 4(2):106–117, 2004.

[241] Michelle A. Kennedy, William A. Hofstadter, and Ileana M. Cristea. TRANSPIRE: A Computational Pipeline to Elucidate Intracellular Protein Movements from Spatial Proteomics Data Sets. *Journal of the American Society for Mass Spectrometry*, 0(0):null, 2020. doi: 10.1021/jasms.0c00033. URL https://doi.org/10.1021/jasms.0c00033. PMID: 32401031.

[242] George S Kimeldorf and Grace Wahba. A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics*, 41(2):495–502, 1970.

[243] Paul Kirk, Jim E Griffin, Richard S Savage, Zoubin Ghahramani, and David L Wild. Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics*, 28(24): 3290–3297, 2012.

[244] Paul DW Kirk and Michael PH Stumpf. Gaussian process regression bootstrapping: exploring the effects of uncertainty in time course data. *Bioinformatics*, 25(10):1300–1306, 2009.

[245] Paul DW Kirk, Maxime Huvet, Anat Melamed, Goedele N Maertens, and Charles RM Bangham. Retroviruses integrate into a shared, non-palindromic DNA motif. *Nature microbiology*, 2:16212, 2016.

[246] PDW Kirk, AC Babtie, and MPH Stumpf. Systems biology (un) certainties. *Science*, 350 (6259):386–388, 2015.

[247] Jessica C Kissinger, Bindu Gajria, Li Li, Ian T Paulsen, and David S Roos. ToxoDB: accessing the Toxoplasma gondii genome. *Nucleic acids research*, 31(1):234–236, 2003.

[248] Sabine Köhler, Charles F Delwiche, Paul W Denny, Lewis G Tilney, Paul Webster, RJM Wilson, Jeffrey D Palmer, and David S Roos. A plastid of probable green algal origin in Apicomplexan parasites. *Science*, 275(5305):1485–1489, 1997.

[249] Ali Sinan Köksal, Kirsten Beck, Dylan R Cronin, Aaron McKenna, Nathan D Camp, Saurabh Srivastava, Matthew E MacGilvray, Rastislav Bodík, Alejandro Wolf-Yadlin, Ernest Fraenkel, et al. Synthesizing signaling pathways from temporal phosphoproteomic data. *Cell reports*, 24(13):3607–3618, 2018.

[250] James T Kost and Michael P McDermott. Combining dependent P-values. *Statistics & Probability Letters*, 60(2):183–190, 2002.

[251] Karen L Kotloff, James P Nataro, William C Blackwelder, Dilruba Nasrin, Tamer H Farag, Sandra Panchalingam, Yukun Wu, Samba O Sow, Dipika Sur, Robert F Breiman, et al. Burden and aetiology of diarrhoeal disease in infants and young children in developing countries (the Global Enteric Multicenter Study, GEMS): a prospective, case-control study. *The Lancet*, 382(9888):209–222, 2013.

[252] Natalie Krahmer, Bahar Najafi, Florian Schueder, Fabiana Quagliarini, Martin Steger, Susanne Seitz, Robert Kasper, Favio Salinas, Jürgen Cox, Nina Henriette Uhlenhaut, et al. Organellar proteomics and phospho-proteomics reveal subcellular reorganization in diet-induced hepatic steatosis. *Developmental cell*, 47(2):205–221, 2018.

[253] Anders R Kristensen and Leonard J Foster. Protein correlation profiling-SILAC to study protein-protein interactions. In *Stable Isotope Labeling by Amino Acids in Cell Culture (SILAC)*, pages 263–270. Springer, 2014.

[254] Anders R Kristensen, Joerg Gsponer, and Leonard J Foster. A high-throughput approach for measuring temporal changes in the interactome. *Nature methods*, 9(9):907, 2012.

[255] Eric S Lander, Lauren M Linton, Bruce Birren, Chad Nusbaum, Michael C Zody, Jennifer Baldwin, Keri Devon, Ken Dewar, Michael Doyle, William FitzHugh, et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–922, 2001.

[256] Letizia Lanzetti, Vladimir Rybin, Maria Grazia Malabarba, Savvas Christoforidis, Giorgio Scita, Marino Zerial, and Pier Paolo Di Fiore. The Eps8 protein coordinates EGF receptor signalling through Rac and trafficking through Rab5. *Nature*, 408(6810):374, 2000.

[257] Isabel J Latorre, Michael H Roh, Kristopher K Frese, Robert S Weiss, Ben Margolis, and Ronald T Javier. Viral oncoprotein-induced mislocalization of select PDZ proteins disrupts tight junctions and causes polarity defects in epithelial cells. *Journal of cell science*, 118(18):4283–4293, 2005.

[258] Yee-Ling Lau, Wenn-Chyau Lee, Ranganath Gudimella, GuiPing Zhang, Xiao-Teng Ching, Rozaimi Razali, Farhanah Aziz, Arif Anwar, and Mun-Yik Fong. Deciphering the draft genome of Toxoplasma gondii RH strain. *PLoS One*, 11(6), 2016.

[259] Kirsti Laurila and Mauno Vihinen. Prediction of disease-related mutations affecting protein localization. *BMC genomics*, 10(1):122, 2009.

[260] Richard A Laursen. Solid-Phase Edman Degradation: An Automatic Peptide Sequencer. *European journal of biochemistry*, 20(1):89–102, 1971.

[261] Michael Lavine and Mike West. A Bayesian method for classification and discrimination. *Canadian Journal of Statistics*, 20(4):451–461, 1992.

[262] Fiona Law, Jung Hwa Seo, Ziqing Wang, Jennifer L DeLeon, Yousstina Bolis, Ashley Brown, Wei-Xing Zong, Guangwei Du, and Christian E Rocheleau. The VPS34 PI3K negatively regulates RAB-5 during endosome maturation. *J Cell Sci*, 130(12):2007–2017, 2017.

[263] Cosmin Lazar, Laurent Gatto, Myriam Ferro, Christophe Bruley, and Thomas Burger. Accounting for the multiple natures of missing values in label-free quantitative proteomics data sets to compare imputation strategies. *Journal of proteome research*, 15(4):1116–1125, 2016.

[264] Elias Lazarides and Klaus Weber. Actin antibody: the specific visualization of actin filaments in non-muscle cells. *Proceedings of the National Academy of Sciences*, 71(6): 2268–2272, 1974.

[265] A-Young Lee, Wei Chen, Steve Stippec, Jon Self, Fan Yang, Xiaojun Ding, She Chen, Yu-Chi Juang, and Melanie H Cobb. Protein kinase WNK3 regulates the neuronal splicing factor Fox-1. *Proceedings of the National Academy of Sciences*, 109(42):16841–16846, 2012.

[266] Tuo Li, Jin Chen, and Ileana M Cristea. Human cytomegalovirus tegument protein pUL83 inhibits IFI16-mediated DNA sensing for immune evasion. *Cell host & microbe*, 14(5): 591–599, 2013.

[267] James Liley, John A Todd, and Chris Wallace. A method for identifying genetic heterogeneity within phenotypically defined disease subgroups. *Nature genetics*, 49 (2):310, 2017.

[268] Dayin Lin, David L Tabb, and John R Yates III. Large-scale protein identification using mass spectrometry. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, 1646 (1-2):1–10, 2003.

[269] Hsuan-Tien Lin, Chih-Jen Lin, and Ruby C Weng. A note on platt's probabilistic outputs for support vector machines. *Machine learning*, 68(3):267–276, 2007.

[270] Scott Linderman, Matthew J Johnson, and Ryan P Adams. Dependent multinomial models made easy: Stick-breaking with the polya-gamma augmentation. In *Advances in Neural Information Processing Systems*, pages 3456–3464, 2015.

[271] Finn Lindgren, Håvard Rue, and Johan Lindström. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73 (4):423–498, 2011.

[272] Karen Linnemannstöns, Leonie Witte, Jeanette Clarissa Kittel, Adi Danieli, Denise Müller, Lena Nitsch, Mona Honemann-Capito, Ferdi Grawe, Andreas Wodarz, Julia Christina Gross, et al. Ykt6 membrane-to-cytosol cycling regulates exosomal Wnt secretion. *bioRxiv*, page 485565, 2018.

[273] Dong C Liu and Jorge Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical programming*, 45(1-3):503–528, 1989.

[274] Jun S Liu. *Monte Carlo strategies in scientific computing*. Springer Science & Business Media, 2008.

[275] Jun S Liu and Rong Chen. Sequential Monte Carlo methods for dynamic systems. *Journal of the American statistical association*, 93(443):1032–1044, 1998.

[276] Qiang Liu, Kevin K Lin, Bogi Andersen, Padhraic Smyth, and Alexander Ihler. Estimating replicate time shifts using Gaussian process regression. *Bioinformatics*, 26(6):770–776, 2010.

[277] Yansheng Liu, Andreas Beyer, and Ruedi Aebersold. On the dependency of cellular protein levels on mRNA abundance. *Cell*, 165(3):535–550, 2016.

[278] Eric F Lock and David B Dunson. Bayesian consensus clustering. *Bioinformatics*, 29(20): 2610–2616, 2013.

[279] FA Londry and James W Hager. Mass selective axial ion ejection from a linear quadrupole ion trap. *Journal of the American Society for Mass Spectrometry*, 14(10):1130–1147, 2003.

[280] Tapio Lönnberg, Valentine Svensson, Kylie R James, Daniel Fernandez-Ruiz, Ismail Sebina, Ruddy Montandon, Megan SF Soon, Lily G Fogg, Arya Sheela Nair, Urijah N Liligeto, et al. Single-cell RNA-seq and computational analysis using temporal mixture modeling resolves TH1/TFH fate bifurcation in malaria. *Science Immunology*, 2(9), 2017.

[281] Hernan Lorenzi, Asis Khan, Michael S Behnke, Sivaranjani Namasivayam, Lakshmipuram S Swapna, Michalis Hadjithomas, Svetlana Karamycheva, Deborah Pinney, Brian P Brunk, James W Ajioka, et al. Local admixture of amplified and diversified secreted pathogenesis determinants shapes mosaic Toxoplasma gondii genomes. *Nature communications*, 7(1):1–13, 2016.

[282] Pero Lučin, Ljerka Kareluša, Gordana Blagojević Zagorac, Hana Mahmutefendić Lučin, Valentino Pavišić, Natalia Jug Vučko, Silvija Lukanović Jurić, Marina Marcelić, Berislav Lisnić, and Stipan Jonjić. Cytomegaloviruses Exploit Recycling Rab Proteins in the Sequential Establishment of the Assembly Compartment. *Frontiers in Cell and Developmental Biology*, 6, 2018.

[283] Leila M Luheshi, Damian C Crowther, and Christopher M Dobson. Protein misfolding and disease: from the test tube to the organism. *Current opinion in chemical biology*, 12 (1):25–31, 2008.

[284] Aaron TL Lun and Gordon K Smyth. From reads to regions: a Bioconductor workflow to detect differential binding in ChIP-seq data. *F1000Research*, 4, 2015.

[285] Aaron TL Lun, Davis J McCarthy, and John C Marioni. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Research*, 5, 2016.

[286] Emma Lundberg and Georg HH Borner. Spatial proteomics: a powerful discovery tool for cell biology. *Nature Reviews Molecular Cell Biology*, 20(5):285–302, 2019.

[287] David J Lunn, Andrew Thomas, Nicky Best, and David Spiegelhalter. WinBUGS-a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and computing*, 10(4):325–337, 2000.

[288] David JC MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[289] Cathal Mahon, Nevan J Krogan, Charles S Craik, and Elah Pick. Cullin E3 ligases and their rewiring by viral factors. *Biomolecules*, 4(4):897–930, 2014.

[290] Alexander Makarov. Electrostatic axially harmonic orbital trapping: a high-performance technique of mass analysis. *Analytical chemistry*, 72(6):1156–1162, 2000.

[291] Jovana Maksimovic, Belinda Phipson, and Alicia Oshlack. A cross-package Bioconductor workflow for analysing methylation array data. *F1000Research*, 5, 2016.

[292] Gertraud Malsiner-Walli, Sylvia Frühwirth-Schnatter, and Bettina Grün. Identifying mixtures of mixtures using Bayesian estimation. *Journal of Computational and Graphical Statistics*, 26(2):285–295, 2017.

[293] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008. ISBN 0521865719, 9780521865715.

[294] Jean-Michel Marin, Kerrie Mengersen, and Christian P Robert. Bayesian modelling and inference on mixtures of distributions. *Handbook of statistics*, 25:459–507, 2005.

[295] Alan G Marshall, Christopher L Hendrickson, and George S Jackson. Fourier transform ion cyclotron resonance mass spectrometry: a primer. *Mass spectrometry reviews*, 17(1): 1–35, 1998.

[296] Takahide Matsui, Peidu Jiang, Saori Nakano, Yuriko Sakamaki, Hayashi Yamamoto, and Noboru Mizushima. Autophagosomal YKT6 is required for fusion with lysosomes independently of syntaxin 17. *J Cell Biol*, 217(8):2633–2645, 2018.

[297] Rafael Mattera, Sang Yoon Park, Raffaella De Pace, Carlos M Guardia, and Juan S Bonifacino. AP-4 mediates export of ATG9A from the trans-Golgi network to promote autophagosome formation. *Proceedings of the National Academy of Sciences*, 114(50): E10697–E10706, 2017.

[298] Graeme C McAlister, David P Nusinow, Mark P Jedrychowski, Martin Wühr, Edward L Huttlin, Brian K Erickson, Ramin Rad, Wilhelm Haas, and Steven P Gygi. MultiNotch MS3 enables accurate, sensitive, and multiplexed detection of differential expression across cancer cell line proteomes. *Analytical chemistry*, 86(14):7150–7158, 2014.

[299] Richard McElreath. *Statistical rethinking: A Bayesian course with examples in R and Stan*. CRC press, 2020.

[300] Geoffrey J McLachlan. On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 36(3):318–324, 1987.

[301] Geoffrey J McLachlan and Kaye E Basford. *Mixture models: Inference and applications to clustering*, volume 38. M. Dekker New York, 1988.

[302] Geoffrey J McLachlan and Thriyambakam Krishnan. *The EM algorithm and extensions*, volume 382. John Wiley & Sons, 2007.

[303] Geoffrey J McLachlan and David Peel. *Finite mixture models*. John Wiley & Sons, 2004.

[304] Christoph TA Meiringer, Kathrin Auffarth, Haitong Hou, and Christian Ungermann. Depalmitoylation of Ykt6 prevents its entry into the multivesicular body pathway. *Traffic*, 9(9):1510–1521, 2008.

[305] EJL Melo, M Attias, and W De Souza. The single mitochondrion of tachyzoites of Toxoplasma gondii. *Journal of structural biology*, 130(1):27–33, 2000.

[306] M Melone, H Varoqui, JD Erickson, and F Conti. Localization of the Na+-coupled neutral amino acid transporter 2 in the cerebral cortex. *Neuroscience*, 140(1):281–292, 2006.

[307] Marta Mendes, Alberto Peláez-García, María López-Lucendo, Rubén A. Bartolomé, Eva Calviño, Rodrigo Barderas, and J. Ignacio Casal. Mapping the Spatial Proteome of Metastatic Cells in Colorectal Cancer. *proteomics*, 17(19), 2017. ISSN 1615-9861. doi: 10.1002/pmic.201700094. URL http://dx.doi.org/10.1002/pmic.201700094. 1700094.

[308] Pablo Mendoza, Rina Ortiz, Jorge Díaz, Andrew FG Quest, Lisette Leyton, Dwayne Stupack, and Vicente A Torres. Rab5 activation promotes focal adhesion disassembly, migration and invasiveness in tumor cells. *J Cell Sci*, 126(17):3835–3847, 2013.

[309] Nicholas Metropolis and Stanislaw Ulam. The monte carlo method. *Journal of the American statistical association*, 44(247):335–341, 1949.

[310] Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.

[311] Sean P Meyn, Robert L Tweedie, et al. Computable bounds for geometric convergence rates of Markov chains. *The Annals of Applied Probability*, 4(4):981–1011, 1994.

[312] Jens Milbradt, Sabrina Auerochs, and Manfred Marschall. Cytomegaloviral proteins pUL50 and pUL53 are associated with the nuclear lamina and interact with cellular protein kinase C. *Journal of general virology*, 88(10):2642–2650, 2007.

[313] Anna L Miles, Stephen P Burr, Guinevere L Grice, and James A Nathan. The vacuolar-ATPase complex and assembly factors, TMEM199 and CCDC115, control HIF1α prolyl hydroxylation by regulating cellular iron levels. *Elife*, 6:e22693, 2017.

[314] Thomas Minka and Zoubin Ghahramani. Expectation propagation for infinite mixtures. 2003.

[315] Thomas P Minka. Expectation propagation for approximate Bayesian inference. *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, 2001.

[316] Dora P Mitchell, John P Savaryn, Nathaniel J Moorman, Thomas Shenk, and Scott S Terhune. Human cytomegalovirus UL28 and UL29 open reading frames encode a spliced mRNA and stimulate accumulation of immediate-early RNAs. *Journal of virology*, 83 (19):10187–10197, 2009.

[317] John C Mitchell and Krzysztof Apt. *Concepts in programming languages*. Cambridge University Press, 2003.

[318] Nathaniel J Moorman, Ronit Sharon-Friling, Thomas Shenk, and Ileana M Cristea. A targeted spatial-temporal proteomics approach implicates multiple cellular trafficking pathways in human cytomegalovirus virion maturation. *Molecular & Cellular Proteomics*, 9(5):851–860, 2010.

[319] Andres Moreno-De-Luca, Sandra L Helmers, Hui Mao, Thomas G Burns, Amanda MA Melton, Karen R Schmidt, Paul M Fernhoff, David H Ledbetter, and Christa L Martin. Adaptor protein complex-4 (AP-4) deficiency causes a novel autosomal recessive cerebral palsy syndrome with microcephaly and intellectual disability. *Journal of medical genetics*, 48(2):141–144, 2011.

[320] Jeffrey S Morris. Functional regression. *Annual Review of Statistics and Its Application*, 2:321–359, 2015.

[321] Katrin Moser and Forest M White. Phosphoproteomic analysis of rat liver by high capacity IMAC and LC- MS/MS. *Journal of proteome research*, 5(1):98–104, 2006.

[322] Frederick Mosteller and R. A. Fisher. Questions and answers. *The American Statistician*, 2(5):30–31, 1948. ISSN 00031305. URL http://www.jstor.org/stable/2681650.

[323] Alison Motley, Nicholas A Bright, Matthew NJ Seaman, and Margaret S Robinson. Clathrin-mediated endocytosis in AP-2–depleted cells. *The Journal of cell biology*, 162 (5):909–918, 2003.

[324] Claire M Mulvey, Lisa M Breckels, Aikaterini Geladaki, Nina Kočevar Britovšek, Daniel JH Nightingale, Andy Christoforou, Mohamed Elzek, Michael J Deery, Laurent Gatto, and Kathryn S Lilley. Using hyperLOPIT to perform high-resolution mapping of the spatial proteome. *Nature Protocols*, 12(6):1110–1135, 2017.

[325] Eain Murphy, Isidore Rigoutsos, Tetsuo Shibuya, and Thomas E Shenk. Reevaluation of human cytomegalovirus coding potential. *Proceedings of the National Academy of Sciences*, 100(23):13585–13590, 2003.

[326] Kevin P Murphy. Conjugate Bayesian analysis of the Gaussian distribution. *Techincal Report*, 1:16, 2007.

[327] Kevin P Murphy. *Machine learning: a probabilistic perspective*. 2012.

[328] LA Murray, X Sheng, and Ileana M Cristea. Orchestration of protein acetylation as a toggle for cellular defense and virus replication. *Nature communications*, 9(1):1–17, 2018.

[329] Elizbar A Nadaraya. On estimating regression. *Theory of Probability & Its Applications*, 9(1):141–142, 1964.

[330] Nikolaos Nasios and Adrian G Bors. Variational learning for Gaussian mixture models. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 36(4): 849–862, 2006.

[331] Naava Naslavsky, Juliati Rahajeng, Mahak Sharma, Marko Jovic, and Steve Caplan. Interactions between EHD proteins and Rab11-FIP2: a role for EHD3 in early endosomal transport. *Molecular biology of the cell*, 17(1):163–177, 2006.

[332] Naava Naslavsky, Jenna McKenzie, Nihal Altan-Bonnet, David Sheff, and Steve Caplan. EHD3 regulates early-endosome-to-Golgi transport and preserves Golgi morphology. *Journal of cell science*, 122(3):389–400, 2009.

[333] Radford M Neal. Monte Carlo implementation of Gaussian process models for Bayesian regression and classification. *arXiv preprint physics/9701026*, 1997.

[334] Radford M Neal. Slice sampling. *Annals of statistics*, pages 705–741, 2003.

[335] Radford M Neal. MCMC Using Hamiltonian Dynamics. *Handbook of Markov Chain Monte Carlo*, page 113, 2011.

[336] Radford M Neal et al. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2(11), 2011.

[337] Lan Huong Nguyen and Susan Holmes. Ten quick tips for effective dimensionality reduction. *PLOS Computational Biology*, 15(6):e1006907, 2019.

[338] Ina Niemann, Anna Reichel, and Thomas Stamminger. Intracellular trafficking of the human cytomegalovirus-encoded 7-trans-membrane protein homologs pus27 and pul78 during viral infection: a comparative analysis. *Viruses*, 6(2):661–682, 2014.

[339] Daniel JH Nightingale, Aikaterini Geladaki, Lisa M Breckels, Stephen G Oliver, and Kathryn S Lilley. The subcellular organisation of Saccharomyces cerevisiae. *Current opinion in chemical biology*, 48:86–95, 2019.

[340] Katie Nightingale, Kai-Min Lin, Benjamin J Ravenhill, Colin Davies, Luis Nobre, Ceri A Fielding, Eva Ruckova, Alice Fletcher-Etherington, Lior Soday, Hester Nichols, et al. High-definition analysis of host protein stability during human cytomegalovirus infection reveals antiviral factors and viral evasion mechanisms. *Cell host & microbe*, 24(3):447–460, 2018.

[341] Nino Nikolovski, Denis Rubtsov, Marcelo P Segura, Godfrey P Miles, Tim J Stevens, Tom PJ Dunkley, Sean Munro, Kathryn S Lilley, and Paul Dupree. Putative glycosyltransferases and other plant Golgi apparatus proteins are revealed by LOPIT proteomics. *Plant physiology*, 160(2):1037–1051, 2012.

[342] Manami Nishi, Ke Hu, John M Murray, and David S Roos. Organellar dynamics during the cell cycle of Toxoplasma gondii. *Journal of cell science*, 121(9):1559–1568, 2008.

[343] Luis V Nobre, Katie Nightingale, Benjamin J Ravenhill, Robin Antrobus, Lior Soday, Jenna Nichols, James A Davies, Sepehr Seirafian, Eddie CY Wang, Andrew J Davison, et al. Human cytomegalovirus interactome analysis identifies degradation hubs, domain associations and viral protein functions. *eLife*, 8:e49894, 2019.

[344] Malgorzata Nowicka, Carsten Krieg, Lukas M Weber, Felix J Hartmann, Silvia Guglietta, Burkhard Becher, Mitchell P Levesque, and Mark D Robinson. CyTOF workflow: differential discovery in high-throughput high-dimensional cytometry datasets. *F1000Research*, 6, 2017.

[345] Esa Nummelin. *General irreducible Markov chains and non-negative operators*, volume 83. Cambridge University Press, 2004.

[346] Kanae Oda, Yukiko Matsuoka, Akira Funahashi, and Hiroaki Kitano. A comprehensive pathway map of epidermal growth factor receptor signaling. *Molecular systems biology*, 1 (1), 2005.

[347] Patrick H O'Farrell. High resolution two-dimensional electrophoresis of proteins. *Journal of biological chemistry*, 250(10):4007–4021, 1975.

[348] K Ogawa-Goto, K Tanaka, W Gibson, E Moriishi, Y Miura, T Kurata, S Irie, and T Sata. Microtubule network facilitates nuclear targeting of human cytomegalovirus capsid. *Journal of virology*, 77(15):8541–8547, 2003.

[349] Shinya Ohta, Jimi-Carlo Bukowski-Wills, Luis Sanchez-Pulido, Flavia de Lima Alves, Laura Wood, Zhuo A Chen, Melpi Platani, Lutz Fischer, Damien F Hudson, Chris P Ponting, et al. The protein composition of mitotic chromosomes determined using multiclassifier combinatorial proteomics. *Cell*, 142(5):810–821, 2010.

[350] Vesa M Olkkonen and Elina Ikonen. When intracellular logistics fails-genetic defects in membrane trafficking. *Journal of cell science*, 119(24):5031–5045, 2006.

[351] Lukas Minus Orre, Mattias Vesterlund, Yanbo Pan, Taner Arslan, Yafeng Zhu, Alejandro Fernandez Woodbridge, Oliver Frings, Erik Fredlund, and Janne Lehtiö. SubCellBarCode: Proteome-wide Mapping of Protein Localization and Relocalization. *Molecular Cell*, 73(1):166 – 182, 2019. ISSN 1097-2765. doi: https://doi.org/10.1016/j.molcel.2018.11.035. URL http://www.sciencedirect.com/science/article/pii/S1097276518310050.

[352] Stefan Otte, William J Belden, Matthew Heidtman, Jay Liu, Ole N Jensen, and Charles Barlowe. Erv41p and Erv46p: new components of COPII vesicles involved in transport between the ER and Golgi complex. *The Journal of cell biology*, 152(3):503–518, 2001.

[353] Bernadett Papp and Kathrin Plath. Epigenetics of reprogramming to induced pluripotency. *Cell*, 152(6):1324–1343, 2013.

[354] Nishant Pappireddi, Lance Martin, and Martin Wühr. A review on quantitative multiplexed proteomics. *ChemBioChem*, 20(10):1210–1224, 2019.

[355] Mercedes Pardo, Lucía Monteoliva, Paloma Vazquez, Raquel Martínez, Gloria Molero, Cesar Nombela, and Concha Gil. PST1 and ECM33 encode two yeast cell surface GPI proteins important for cell wall integrity. *Microbiology*, 150(12):4157–4170, 2004.

[356] HT Parsons, SM Fernández-Niño, and JL Heazlewood. Separation of the plant Golgi apparatus and endoplasmic reticulum by free-flow electrophoresis. *Methods in molecular biology (Clifton, NJ)*, 1072:527, 2014.

[357] Rui Paulo et al. Default priors for Gaussian processes. *The Annals of Statistics*, 33(2): 556–582, 2005.

[358] David Peel and Geoffrey J McLachlan. Robust mixture modelling using the t distribution. *Statistics and computing*, 10(4):339–348, 2000.

[359] Laurence Pelletier, Charlene A Stern, Marc Pypaert, David Sheff, Huân M Ngô, Nitin Roper, Cynthia Y He, Ke Hu, Derek Toomre, Isabelle Coppens, et al. Golgi biogenesis in Toxoplasma gondii. *Nature*, 418(6897):548–552, 2002.

[360] Natàlia Pérez-Carmona, Pablo Martínez-Vicente, Domènec Farré, Ildar Gabaev, Martin Messerle, Pablo Engel, and Ana Angulo. A prominent role of the human cytomegalovirus UL8 glycoprotein in restraining proinflammatory cytokine production by myeloid cells at late times during infection. *Journal of virology*, 92(9):e02229–17, 2018.

[361] Lena Pernas, Yaw Adomako-Ankomah, Anjali J Shastri, Sarah E Ewald, Moritz Treeck, Jon P Boyle, and John C Boothroyd. Toxoplasma effector MAF1 mediates recruitment of host mitochondria and impacts the host response. *PLoS biology*, 12(4), 2014.

[362] Fanny Perraudeau, Davide Risso, Kelly Street, Elizabeth Purdom, and Sandrine Dudoit. Bioconductor workflow for single-cell RNA sequencing: Normalization, dimensionality reduction, clustering, and lineage inference. *F1000Research*, 6, 2017.

[363] Belinda Phipson and Gordon K Smyth. Permutation P-values should never be zero: calculating exact P-values when permutations are randomly drawn. *Statistical applications in genetics and molecular biology*, 9(1).

[364] Jim Pitman and Marc Yor. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, pages 855–900, 1997.

[365] John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. 1999.

[366] Alexander Plotnikov, Eldar Zehorai, Shiri Procaccia, and Rony Seger. The MAPK cascades: signaling components, nuclear roles and mechanisms of nuclear translocation. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, 1813(9):1619–1633, 2011.

[367] Martyn Plummer, Nicky Best, Kate Cowles, and Karen Vines. CODA: Convergence Diagnosis and Output Analysis for MCMC. *R News*, 6(1):7–11, 2006. URL https://journal.r-project.org/archive/.

[368] Nicholas G Polson, James G Scott, and Jesse Windle. Bayesian inference for logistic models using Pólya–Gamma latent variables. *Journal of the American statistical Association*, 108(504):1339–1349, 2013.

[369] Anthony P Possemato, Joao A Paulo, Daniel Mulhern, Ailan Guo, Steven P Gygi, and Sean A Beausoleil. Multiplexed phosphoproteomic profiling using titanium dioxide and immunoaffinity enrichments reveals complementary phosphorylation events. *Journal of proteome research*, 16(4):1506–1514, 2017.

[370] Cristian Preda, Gilbert Saporta, and Caroline Lévéder. PLS classification of functional data. *Computational Statistics*, 22(2):223–235, 2007.

[371] Rosa Puertollano, Shawn M Ferguson, James Brugarolas, and Andrea Ballabio. The complex relationship between TFEB transcription factor phosphorylation and subcellular localization. *The EMBO journal*, 37(11), 2018.

[372] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017. URL https://www.R-project.org/.

[373] James O Ramsay. Functional data analysis. *Encyclopedia of Statistical Sciences*, 4, 2004.

[374] William M Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971.

[375] Carl Edward Rasmussen. Gaussian processes in machine learning. In *Advanced lectures on machine learning*, pages 63–71. Springer, 2004.

[376] Carl Edward Rasmussen and Christopher KI Williams. *Gaussian processes for machine learning*. MIT Press, 2006.

[377] MB Reeves, PA MacAry, PJ Lehner, JGP Sissons, and JH Sinclair. Latency, chromatin remodeling, and reactivation of human cytomegalovirus in the dendritic cells of healthy carriers. *Proceedings of the National Academy of Sciences*, 102(11):4140–4145, 2005.

[378] Barbara Reinhardt, Michael Winkler, Peter Schaarschmidt, Robert Pretsch, Shaoxia Zhou, Bianca Vaida, Alexandra Schmid-Kotsas, Detlef Michel, Paul Walther, Max Bachem, et al. Human cytomegalovirus-induced reduction of extracellular matrix proteins in vascular smooth muscle cell cultures: a pathomechanism in vasculopathies? *Journal of general virology*, 87(10):2849–2858, 2006.

[379] Boyu Ren, Sergio Bacallado, Stefano Favaro, Susan Holmes, and Lorenzo Trippa. Bayesian nonparametric ordination for the analysis of microbial communities. *Journal of the American Statistical Association*, 112(520):1430–1442, 2017.

[380] Sylvia Richardson and Peter J Green. On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: series B (statistical methodology)*, 59(4):731–792, 1997.

[381] Jochen Rink, Eric Ghigo, Yannis Kalaidzidis, and Marino Zerial. Rab conversion as a mechanism of progression from early to late endosomes. *Cell*, 122(5):735–749, 2005.

[382] Christian P Robert and George Casella. The Metropolis—Hastings Algorithm. In *Monte Carlo Statistical Methods*, pages 231–283. Springer, 1999.

[383] Gareth O Roberts and Nicholas G Polson. On the geometric convergence of the Gibbs sampler. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(2): 377–384, 1994.

[384] Gareth O Roberts and Jeffrey S Rosenthal. Markov-chain monte carlo: Some practical implications of theoretical results. *Canadian Journal of Statistics*, 26(1):5–20, 1998.

[385] Gareth O Roberts and Jeffrey S Rosenthal. Optimal scaling of discrete approximations to Langevin diffusions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1):255–268, 1998.

[386] Gareth O Roberts and Sujit K Sahu. Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(2):291–317, 1997.

[387] Gareth O Roberts and Adrian FM Smith. Simple conditions for the convergence of the Gibbs sampler and Metropolis-Hastings algorithms. *Stochastic processes and their applications*, 49(2):207–216, 1994.

[388] Gareth O Roberts, Jeffrey S Rosenthal, et al. Two convergence properties of hybrid samplers. *The Annals of Applied Probability*, 8(2):397–407, 1998.

[389] Gareth O Roberts, Jeffrey S Rosenthal, et al. Optimal scaling for various Metropolis-Hastings algorithms. *Statistical science*, 16(4):351–367, 2001.

[390] Margaret S Robinson. Forty years of clathrin-coated vesicles. *Traffic*, 16(12):1210–1238, 2015.

[391] Abel Rodríguez, David B Dunson, and Alan E Gelfand. Bayesian nonparametric functional data analysis through density estimation. *Biometrika*, 96(1):149–162, 2009.

[392] Jesse Rodriguez, Nitin Gupta, Richard D Smith, and Pavel A Pevzner. Does trypsin cut before proline? *Journal of proteome research*, 7(01):300–305, 2008.

[393] José Antonio Rodriguez, Wendy WY Au, and Beric R Henderson. Cytoplasmic mislocalization of BRCA1 caused by cancer-associated mutations in the BRCT domain. *Experimental cell research*, 293(1):14–21, 2004.

[394] David S Roos, Michael J Crawford, Robert GK Donald, Jessica C Kissinger, Leszek J Klimczak, and Boris Striepen. Origin, targeting, and function of the apicomplexan plastid. *Current opinion in microbiology*, 2(4):426–432, 1999.

[395] Jeffrey S Rosenthal. Minorization conditions and convergence rates for Markov chain Monte Carlo. *Journal of the American Statistical Association*, 90(430):558–566, 1995.

[396] Jeffrey S Rosenthal. A review of asymptotic convergence for general state space Markov chains. *Far East J. Theor. Stat*, 5(1):37–50, 2001.

[397] Judith Rousseau and Kerrie Mengersen. Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(5):689–710, 2011.

[398] Kyle J Roux, Dae In Kim, Manfred Raida, and Brian Burke. A promiscuous biotin ligase fusion protein identifies proximal and interacting proteins in mammalian cells. *The Journal of cell biology*, 196(6):801–810, 2012.

[399] Reuven Y Rubinstein and Dirk P Kroese. *Simulation and the Monte Carlo method*, volume 10. John Wiley & Sons, 1981.

[400] Pawel G Sadowski, Tom PJ Dunkley, Ian P Shadforth, Paul Dupree, Conrad Bessant, Julian L Griffin, and Kathryn S Lilley. Quantitative proteomic approach to study subcellular localization of membrane proteins. *Nature protocols*, 1(4):1778–1789, 2006.

[401] JPJ Saeij, JP Boyle, S Coller, S Taylor, LD Sibley, ET Brooke-Powell, JW Ajioka, and JC Boothroyd. Polymorphic secreted kinases are key virulence factors in toxoplasmosis. *Science*, 314(5806):1780–1783, 2006.

[402] JPJ Saeij, S Coller, JP Boyle, ME Jerome, MW White, and JC Boothroyd. Toxoplasma co-opts host gene expression by injection of a polymorphic kinase homologue. *Nature*, 445(7125):324–327, 2007.

[403] Daniela Salas, R Greg Stacey, Mopelola Akinlaja, and Leonard J Foster. Next-generation interactomics: considerations for the use of co-elution to measure protein interaction networks. *Molecular & Cellular Proteomics*, 19(1):1–10, 2020.

[404] Prabha Sampath, David K Pritchard, Lil Pabon, Hans Reinecke, Stephen M Schwartz, David R Morris, and Charles E Murry. A hierarchical network controls protein translation during murine embryonic stem cell self-renewal and differentiation. *Cell stem cell*, 2(5): 448–460, 2008.

[405] Daniel J Schad, Michael Betancourt, and Shravan Vasishth. Toward a principled Bayesian workflow in cognitive science. *arXiv preprint arXiv:1904.12765*, 2019.

[406] Mark J Schervish and Bradley P Carlin. On the convergence of successive substitution sampling. *Journal of Computational and Graphical statistics*, 1(2):111–127, 1992.

[407] Ulrike Schnell, Freark Dijk, Klaas A Sjollema, and Ben NG Giepmans. Immunolabeling artifacts and the need for live-cell imaging. *Nature methods*, 9(2):152, 2012.

[408] Bernhard Schölkopf, Ralf Herbrich, and Alex J Smola. A generalized representer theorem. In *International conference on computational learning theory*, pages 416–426. Springer, 2001.

[409] Bernhard Schölkopf, Alexander J Smola, Francis Bach, et al. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[410] NK Schuurman, RPPP Grasman, and EL Hamaker. A comparison of inverse-wishart prior specifications for covariance matrices in multilevel autoregressive models. *Multivariate Behavioral Research*, 51(2-3):185–206, 2016.

[411] Gideon Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6 (2):461–464, 1978.

[412] Luca Scrucca, Michael Fop, T Brendan Murphy, and Adrian E Raftery. mclust 5: clustering, classification and density estimation using gaussian finite mixture models. *The R journal*, 8(1):289, 2016.

[413] Frank Seeber, David JP Ferguson, and Uwe Gross. Toxoplasma gondii: a paraformaldehyde-insensitive diaphorase activity acts as a specific histochemical marker for the single mitochondrion. *Experimental parasitology*, 89(1):137–139, 1998.

[414] Ophir Shalem, Neville E Sanjana, Ella Hartenian, Xi Shi, David A Scott, Tarjei S Mikkelsen, Dirk Heckl, Benjamin L Ebert, David E Root, John G Doench, et al. Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science*, 343(6166):84–87, 2014.

[415] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, July 1948. ISSN 0005-8580. doi: 10.1002/j.1538-7305.1948.tb01338.x.

[416] Harley G Sheffield and Marjorie L Melton. The fine structure and reproduction of Toxoplasma gondii. *The Journal of parasitology*, pages 209–226, 1968.

[417] Ao Shen, Ji Lei, Edward Yang, Yonggang Pei, Yuan-Chuan Chen, Hao Gong, Gengfu Xiao, and Fenyong Liu. Human cytomegalovirus primase UL70 specifically interacts with cellular factor Snapin. *Journal of virology*, 85(22):11732–11741, 2011.

[418] John J.H. Shin, Oliver M. Crook, Alicia Borgeaud, Jérôme Cattin-Ortolá, Sew-Yeu Peak-Chew, Jessica Chadwick, Kathryn S. Lilley, and Sean Munro. Determining the content of vesicles captured by golgin tethers using LOPIT-DC. *bioRxiv*, 2019. doi: 10.1101/841965. URL https://www.biorxiv.org/content/early/2019/11/14/841965.

[419] Soo J Shin, Jeffrey A Smith, Günther A Rezniczek, Sheng Pan, Ru Chen, Teresa A Brentnall, Gerhard Wiche, and Kimberly A Kelly. Unexpected gain of function for the scaffolding protein plectin due to mislocalization in pancreatic cancer. *Proceedings of the National Academy of Sciences*, 110(48):19414–19419, 2013.

[420] Ng Shyh-Chang and George Q Daley. Lin28: primal regulator of growth and metabolism in stem cells. *Cell stem cell*, 12(4):395–406, 2013.

[421] Saima M Sidik, Diego Huet, Suresh M Ganesan, My-Hang Huynh, Tim Wang, Armiyaw S Nasamu, Prathapan Thiru, Jeroen PJ Saeij, Vern B Carruthers, Jacquin C Niles, et al. A genome-wide CRISPR screen in Toxoplasma identifies essential apicomplexan genes. *Cell*, 166(6):1423–1435, 2016.

[422] Saima M Sidik, Diego Huet, and Sebastian Lourido. CRISPR-Cas9-based genome-wide screening of Toxoplasma gondii. *Nature protocols*, 13(2):307, 2018.

[423] J E Siljee, Y Wang, A A Bernard, B A Ersoy, S Zhang, A Marley, M Von Zastrow, J F Reiter, and C Vaisse. Subcellular localization of MC4R with ADCY3 at neuronal primary cilia underlies a common pathway for genetic predisposition to obesity. *Nat Genet*, Jan 2018. doi: 10.1038/s41588-017-0020-9.

[424] Bernhard W Silverman. Algorithm AS 176: Kernel density estimation using the fast Fourier transform. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 31(1):93–99, 1982.

[425] Anne Simonsen, Roger Lippe, Savvas Christoforidis, Jean-Michel Gaullier, Andreas Brech, Judy Callaghan, Ban-Hock Toh, Carol Murphy, Marino Zerial, and Harald Stenmark. EEA1 links PI (3) K function to Rab5 regulation of endosome fusion. *Nature*, 394(6692): 494, 1998.

[426] Daniel Simpson, Håvard Rue, Andrea Riebler, Thiago G Martins, Sigrunn H Sørbye, et al. Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical science*, 32(1):1–28, 2017.

[427] Adrian FM Smith and Gareth O Roberts. Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 3–23, 1993.

[428] Austin G Smith. Embryo-derived stem cells: of mice and men. *Annual review of cell and developmental biology*, 17(1):435–462, 2001.

[429] Alex J Smola and Bernhard Schölkopf. A tutorial on support vector regression. *Statistics and computing*, 14(3):199–222, 2004.

[430] Charlotte Stadler, Elton Rexhepaj, Vasanth R Singan, Robert F Murphy, Rainer Pepperkok, Mathias Uhlén, Jeremy C Simpson, and Emma Lundberg. Immunofluorescence and fluorescent-protein tagging show high correlation for protein localization in mammalian cells. *Nature methods*, 10(4):315–323, 2013.

[431] Thomas Stamminger, Matthias Gstaiger, Konstanze Weinzierl, Kerstin Lorz, Michael Winkler, and Walter Schaffner. Open reading frame UL26 of human cytomegalovirus encodes a novel tegument protein that contains a strong transcriptional activation domain. *Journal of virology*, 76(10):4836–4847, 2002.

[432] Mark FJ Steel and Montserrat Fuentes. *Non-gaussian and nonparametric models for continuous spatial data*. CRC press, 2010.

[433] Oliver Stegle, Katherine J Denby, Emma J Cooke, David L Wild, Zoubin Ghahramani, and Karsten M Borgwardt. A robust bayesian two-sample test for detecting intervals of differential gene expression in microarray time series. *Journal of Computational Biology*, 17(3):355–367, 2010.

[434] Michael L Stein. *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media, 1999.

[435] Ingo Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of machine learning research*, 2(Nov):67–93, 2001.

[436] Matthew Stephens. Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):795–809, 2000.

[437] Noam Stern-Ginossar, Ben Weisburd, Annette Michalski, Vu Thuy Khanh Le, Marco Y Hein, Sheng-Xiong Huang, Ming Ma, Ben Shen, Shu-Bing Qian, Hartmut Hengel, et al. Decoding human cytomegalovirus. *Science*, 338(6110):1088–1093, 2012.

[438] Boris Striepen. Parasitic infections: time to tackle cryptosporidiosis. *Nature News*, 503 (7475):189, 2013.

[439] Boris Striepen, Cynthia Yingxin He, Mariana Matrajt, Dominique Soldati, and David S Roos. Expression, selection, and organellar targeting of the green fluorescent protein in Toxoplasma gondii. *Molecular and biochemical parasitology*, 92(2):325–338, 1998.

[440] Boris Striepen, Michael J Crawford, Michael K Shaw, Lewis G Tilney, Frank Seeber, and David S Roos. The plastid of Toxoplasma gondii is divided by association with the centrosomes. *The Journal of cell biology*, 151(7):1423–1434, 2000.

[441] Devin P Sullivan, Casper F Winsnes, Lovisa Åkesson, Martin Hjelmare, Mikaela Wiking, Rutger Schutten, Linzi Campbell, Hjalti Leifsson, Scott Rhodes, Andie Nordgren, et al. Deep learning is combined with massive-scale citizen science to improve large-scale image classification. *Nature biotechnology*, 36(9):820, 2018.

[442] Lakshmipuram Seshadri Swapna and John Parkinson. Genomics of apicomplexan parasites. *Critical reviews in biochemistry and molecular biology*, 52(3):254–273, 2017.

[443] Damian Szklarczyk, Annika L Gable, David Lyon, Alexander Junge, Stefan Wyder, Jaime Huerta-Cepas, Milan Simonovic, Nadezhda T Doncheva, John H Morris, Peer Bork, et al. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic acids research*, 47 (D1):D607–D613, 2019.

[444] David L Tabb, Jimmy K Eng, and John R Yates. Protein identification by SEQUEST. In *Proteome Research: Mass Spectrometry*, pages 125–142. Springer, 2001.

[445] Guihua Tai, Lei Lu, Tuan Lao Wang, Bor Luen Tang, Bruno Goud, Ludger Johannes, and Wanjin Hong. Participation of the syntaxin 5/Ykt6/GS28/GS15 SNARE complex in transport from the early/recycling endosome to the trans-Golgi network. *Molecular biology of the cell*, 15(9):4011–4022, 2004.

[446] Szabolcs Takáts, Gábor Glatz, Győző Szenci, Attila Boda, Gábor V Horváth, Krisztina Hegedűs, Attila L Kovács, and Gábor Juhász. Non-canonical role of the SNARE protein Ykt6 in autophagosome-lysosome fusion. *PLoS genetics*, 14(4):e1007359, 2018.

[447] Vivian Tam, Nikunj Patel, Michelle Turcotte, Yohan Bossé, Guillaume Paré, and David Meyre. Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics*, 20(8):467–484, 2019.

[448] Denise JL Tan, Heidi Dvinge, Andrew Christoforou, Paul Bertone, Alfonso Martinez Arias, and Kathryn S Lilley. Mapping organelle proteins and protein complexes in drosophila melanogaster. *Journal of proteome research*, 8(6):2667–2678, 2009.

[449] Martin A Tanner and Wing Hung Wong. The calculation of posterior distributions by data augmentation. *Journal of the American statistical Association*, 82(398):528–540, 1987.

[450] Marianne Tardif, Ariane Atteia, Michael Specht, Guillaume Cogne, Norbert Rolland, Sabine Brugière, Michael Hippler, Myriam Ferro, Christophe Bruley, Gilles Peltier, et al. PredAlgo: a new subcellular localization prediction tool dedicated to green algae. *Molecular biology and evolution*, 29(12):3625–3639, 2012.

[451] S Taylor, A Barragan, C Su, B Fux, SJ Fentress, K Tang, WL Beatty, H El Hajj, M Jerome, MS Behnke, et al. A secreted serine-threonine kinase determines virulence in the eukaryotic pathogen Toxoplasma gondii. *Science*, 314(5806):1776–1780, 2006.

[452] Yee W Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Sharing clusters among related groups: Hierarchical Dirichlet processes. In *Advances in neural information processing systems*, pages 1385–1392, 2005.

[453] Andrew Thompson, Jürgen Schäfer, Karsten Kuhn, Stefan Kienle, Josef Schwarz, Günter Schmidt, Thomas Neumann, and Christian Hamon. Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by ms/ms. *Analytical chemistry*, 75(8):1895–1904, 2003.

[454] Robert L Thorndike. Who belongs in the family. In *Psychometrika*. Citeseer, 1953.

[455] Peter J. Thul, Lovisa Åkesson, Mikaela Wiking, Diana Mahdessian, Aikaterini Geladaki, Hammou Ait Blal, Tove Alm, Anna Asplund, Lars Björk, Lisa M. Breckels, Anna Bäckström, Frida Danielsson, Linn Fagerberg, Jenny Fall, Laurent Gatto, Christian Gnann, Sophia Hober, Martin Hjelmare, Fredric Johansson, Sunjae Lee, Cecilia Lindskog, Jan Mulder, Claire M. Mulvey, Peter Nilsson, Per Oksvold, Johan Rockberg, Rutger Schutten, Jochen M. Schwenk, Åsa Sivertsson, Evelina Sjöstedt, Marie Skogs, Charlotte Stadler, Devin P. Sullivan, Hanna Tegel, Casper Winsnes, Cheng Zhang, Martin Zwahlen, Adil Mardinoglu, Fredrik Pontén, Kalle von Feilitzen, Kathryn S. Lilley, Mathias Uhlén, and Emma Lundberg. A subcellular map of the human proteome. *Science*, 2017. ISSN 0036-8075. doi: 10.1126/science.aal3321. URL http://science.sciencemag.org/content/early/2017/05/10/science.aal3321.

[456] Luke Tierney. Markov chains for exploring posterior distributions. *the Annals of Statistics*, pages 1701–1728, 1994.

[457] Lily Ting, Ramin Rad, Steven P Gygi, and Wilhelm Haas. MS3 eliminates ratio distortion in isobaric multiplexed quantitative proteomics. *Nature methods*, 8(11):937, 2011.

[458] Hande Topa, Ágnes Jónás, Robert Kofler, Carolin Kosiol, and Antti Honkela. Gaussian process test for high-throughput sequencing time series: application to experimental evolution. *Bioinformatics*, 31(11):1762–1770, 2015.

[459] Catherine Toursel, Florence Dzierszinski, Annie Bernigaud, Marlène Mortuaire, and Stanislas Tomavo. Molecular cloning, organellar targeting and developmental expression of mitochondrial chaperone HSP60 in Toxoplasma gondii. *Molecular and biochemical parasitology*, 111(2):319–332, 2000.

[460] Laura Trinkle-Mulcahy. Recent advances in proximity-based labeling methods for interactome mapping. *F1000Research*, 8, 2019.

[461] Shipra Vaishnava, David P Morrison, Rajshekhar Y Gaji, John M Murray, Rolf Entzeroth, Daniel K Howe, and Boris Striepen. Plastid segregation and cell division in the apicomplexan parasite Sarcocystis neurona. *Journal of cell science*, 118(15):3397–3407, 2005.

[462] Daniel Valcarce, Javier Parapar, and Álvaro Barreiro. Additive Smoothing for Relevance-Based Language Modelling of Recommender Systems. *Proceedings of the 4th Spanish Conference on Information Retrieval*, pages 1–8, 2016. doi: 10.1145/2934732.2934737. URL http://doi.acm.org/10.1145/2934732.2934737.

[463] Aad W van der Vaart, J Harry van Zanten, et al. Adaptive Bayesian estimation using a Gaussian random field with inverse gamma bandwidth. *The Annals of Statistics*, 37(5B): 2655–2675, 2009.

[464] Giel G van Dooren, Matthias Marti, Christopher J Tonkin, Luciana M Stimmler, Alan F Cowman, and Geoffrey I McFadden. Development of the endoplasmic reticulum, mitochondrion and apicoplast during the asexual life cycle of Plasmodium falciparum. *Molecular microbiology*, 57(2):405–419, 2005.

[465] Zoé Van Havre, Nicole White, Judith Rousseau, and Kerrie Mengersen. Overfitting Bayesian mixture models with an unknown number of components. *PloS one*, 10(7), 2015.

[466] Dootika Vats and Christina Knudson. Revisiting the Gelman-Rubin Diagnostic. *arXiv preprint arXiv:1812.09384*, 2018.

[467] Aki Vehtari, Andrew Gelman, Daniel Simpson, Bob Carpenter, and Paul-Christian Bürkner. Rank-normalization, folding, and localization: An improved $\hat{R}$ for assessing convergence of MCMC. *arXiv preprint arXiv:1903.08008*, 2019.

[468] J Votỳpka, D Modrỳ, M Oborník, J Šlapeta, J Lukeš, et al. Apicomplexa. *Cham: springer*, 2017.

[469] Benjamin S Waldman, Dominic Schwarz, Marc H Wadsworth II, Jeroen P Saeij, Alex K Shalek, and Sebastian Lourido. Identification of a master regulator of differentiation in Toxoplasma. *Cell*, 2020.

[470] Jane-Ling Wang, Jeng-Min Chiou, and Hans-Georg Müller. Functional data analysis. *Annual Review of Statistics and Its Application*, 3:257–295, 2016.

[471] Lianming Wang and David B Dunson. Fast Bayesian inference in Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 20(1):196–216, 2011.

[472] Yuchong Wang, Kathryn S Lilley, and Stephen G Oliver. A protocol for the subcellular fractionation of Saccharomyces cerevisiae using nitrogen cavitation and density gradient centrifugation. *Yeast*, 31(4):127–135, 2014.

[473] Joe H Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244, 1963.

[474] Geoffrey S Watson. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 359–372, 1964.

[475] Michael P Weekes, Peter Tomasec, Edward L Huttlin, Ceri A Fielding, David Nusinow, Richard J Stanton, Eddie CY Wang, Rebecca Aicheler, Isa Murrell, Gavin WG Wilkinson, et al. Quantitative temporal viromics: an approach to investigate host-pathogen interaction. *Cell*, 157(6):1460–1472, 2014.

[476] Holger Wendland. Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree. *Advances in computational Mathematics*, 4(1):389–396, 1995.

[477] Franz Wendler, Alison K Gillingham, Rita Sinka, Cláudia Rosa-Ferreira, David E Gordon, Xavier Franch-Marro, Andrew A Peden, Jean-Paul Vincent, and Sean Munro. A genome-wide RNA interference screen identifies two novel components of the metazoan secretory pathway. *The EMBO journal*, 29(2):304–314, 2010.

[478] Hadley Wickham. *Advanced r*. CRC press, 2019.

[479] Christopher KI Williams and Carl Edward Rasmussen. Gaussian processes for regression. In *Advances in neural information processing systems*, pages 514–520, 1996.

[480] Matthias Wilm. Principles of electrospray ionization. *Molecular & cellular proteomics*, 10 (7), 2011.

[481] Daniel J Wilson. The harmonic mean p-value for combining dependent tests. *Proceedings of the National Academy of Sciences*, 116(4):1195–1200, 2019.

[482] Yong H Woo, Hifzur Ansari, Thomas D Otto, Christen M Klinger, Martin Kolisko, Jan Michálek, Alka Saxena, Dhanasekaran Shanmugam, Annageldi Tayyrov, Alaguraj Veluchamy, et al. Chromerid genomes reveal the evolutionary path from photosynthetic algae to obligate intracellular parasites. *elife*, 4:e06974, 2015.

[483] Ben J Woodcroft, Paul J McMillan, Chaitali Dekiwadia, Leann Tilley, and Stuart A Ralph. Determination of protein subcellular localization in apicomplexan parasites. *Trends in parasitology*, 28(12):546–554, 2012.

[484] Philip G Woodman. Biogenesis of the sorting endosome: the role of Rab5. *Traffic*, 1(9): 695–701, 2000.

[485] Ting-Fan Wu, Chih-Jen Lin, and Ruby C Weng. Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research*, 5(Aug): 975–1005, 2004.

[486] Qing Yuan Yin, Piet WJ de Groot, Henk L Dekker, Luitzen de Jong, Frans M Klis, and Chris G de Koster. Comprehensive proteomic analysis of Saccharomyces cerevisiae cell walls identification of proteins covalently attached via glycosylphosphatidylinositol remnants or mild alkali-sensitive linkages. *Journal of Biological Chemistry*, 280(21): 20894–20901, 2005.

[487] Ido Yofe, Uri Weill, Matthias Meurer, Silvia Chuartzman, Einat Zalckvar, Omer Goldman, Shifra Ben-Dor, Conny Schütze, Nils Wiedemann, Michael Knop, et al. One library to make them all: streamlining the creation of yeast libraries via a SWAp-Tag strategy. *Nature methods*, 13(4):371, 2016.

[488] Cheryl Qian Ying Yong and Bor Luen Tang. Another longin SNARE for autophagosome-lysosome fusion-how does Ykt6 work? *Autophagy*, 15(2):352–357, 2019.

[489] Guangchuang Yu, Li-Gen Wang, Yanyan Han, and Qing-Yu He. clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics: a journal of integrative biology*, 16(5):284–287, 2012.

[490] Yongjun Yu, Amy J Clippinger, and James C Alwine. Viral effects on metabolism: changes in glucose and glutamine utilization during human cytomegalovirus infection. *Trends in microbiology*, 19(7):360–367, 2011.

[491] AD Yurochko. Human cytomegalovirus modulation of signal transduction. In *Human Cytomegalovirus*, pages 205–220. Springer, 2008.

[492] Sebastian Zeltzer, Carol A Zeltzer, Suzu Igarashi, Jean Wilson, Julie G Donaldson, and Felicia Goodrum. Virus control of trafficking from sorting endosomes. *MBio*, 9(4): e00683–18, 2018.

[493] Marino Zerial and Heidi McBride. Rab proteins as membrane organizers. *Nature reviews Molecular cell biology*, 2(2):107, 2001.

[494] Yunong Zhang, William E Leithead, and Douglas J Leith. Time-series Gaussian process regression based on Toeplitz computation of O (N 2) operations and O (N)-level storage. In *Decision and Control, 2005 and 2005 European Control Conference. CDC-ECC'05. 44th IEEE Conference on*, pages 3711–3716. IEEE, 2005.

[495] Hongxiao Zhu, Marina Vannucci, and Dennis D Cox. A Bayesian hierarchical model for classification with selection of functional predictors. *Biometrics*, 66(2):463–473, 2010.

[496] Hongxiao Zhu, Philip J Brown, and Jeffrey S Morris. Robust classification of functional and quantitative image data using functional mixed models. *Biometrics*, 68(4):1260–1268, 2012.

Appendix A

Appendix to chapter 2

A.1 Appendix 1: Derivation of EM algorithm for TAGM model

This appendix give a formal derivation of the EM algorithm used for our model. Computations are standard but useful and similar technical summaries can be found (for example see [138, 326]) We let $H = \{\boldsymbol{\mu}_0, \lambda_0, \nu_0, S_0\}$ denote the parameters of the normal-inverse-Wishart prior. More precisely:

$$\boldsymbol{\mu}_k, \Sigma_k \sim \mathcal{N}\left(\boldsymbol{\mu}_k | \boldsymbol{\mu}_0, \frac{\Sigma_k}{\lambda_0}\right) IW\left(\Sigma_k | \nu_0, S_0\right). \tag{A.1}$$

Furthermore, let $\boldsymbol{\theta}_k = \{\boldsymbol{\mu}_k, \Sigma_k\}$, and let $\Theta = \{\kappa, \mathbf{M}, V\}$ be the parameters of the global $\mathcal{T}$ distribution. We specify the following hierarchical Bayesian model.

$$\begin{aligned}
\pi|\beta &\sim Dir(\beta), \\
\theta_k|H &\sim \mathcal{NIW}(H), \\
z_i|\pi &\sim cat(\pi), \\
\epsilon|u,v &\sim \mathcal{B}(u,v) \\
\phi_i|\epsilon &\sim Ber(1-\epsilon) \\
\mathbf{x}_i|z_i = k, \theta, \Phi, \Theta &\sim \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_k, \Sigma_k)^{\mathbb{1}(\phi_i=1)} \mathcal{T}(\mathbf{x}_i|\kappa, \mathbf{M}, V)^{\mathbb{1}(\phi_i=0)}
\end{aligned} \tag{A.2}$$

Since $p(\phi_i = 1) = 1 - \epsilon$, we can rewrite the last line of the model (A.2) as the following:

$$p(\mathbf{x}_i|z_i = k, \theta, \Phi, \Theta) = (1-\epsilon)\mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_k, \Sigma_k) + \epsilon\mathcal{T}(\mathbf{x}_i|\kappa, \mathbf{M}, V).$$

The total joint probability is

$$
\begin{aligned}
p(\theta, \Theta, X, Z, \Phi) =& p(X, Z, \Phi|\theta, \pi, \epsilon)p(\epsilon|u, v)p(\theta|H)p(\pi|\beta) \\
=& \prod_{i=1}^{n}\prod_{k=1}^{K}\left(\pi_k((1-\epsilon)\mathcal{N}(x_i|\boldsymbol{\mu}_k, \Sigma_k))^{\mathbb{1}(\phi_i=1)}(\epsilon\mathcal{T}(x_i|\kappa, \mathbf{M}, V))^{\mathbb{1}(\phi_i=0)}\right)^{\mathbb{1}(z_i=k)} \\
& \cdot \left(\prod_{k=1}^{K}\mathcal{NIW}(H)\right) \cdot Dir(\beta) \cdot \mathcal{B}(u, v).
\end{aligned} \tag{A.3}
$$

Before we formally derive an EM algorithm for this model, we derive a few useful quantities. Let $f(\mathbf{x}|\boldsymbol{\mu}, \Sigma)$ denote the density of the multivariate normal with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ evaluated at $\mathbf{x}$ and further let $g(\mathbf{x}|\kappa, \mathbf{M}, V)$ denote the density of the multivariate T-distribution. We compute that

$$
\begin{aligned}
p(\phi_i = 1|z_i = k, \mathbf{x}_i) =& \frac{p(\phi_i = 1, \mathbf{x}_i|z_i = k)}{p(\mathbf{x}_i|z_i = k)} \\
=& \frac{p(\mathbf{x}_i|z_i = k, \phi_i = 1)P(\phi_i = 1|z_i = k)}{p(\mathbf{x}_i|z_i = k)} \\
=& \frac{(1-\epsilon)f(\mathbf{x}_i|\boldsymbol{\mu}_k, \Sigma_k)}{(1-\epsilon)f(\mathbf{x}_i|\boldsymbol{\mu}_k, \Sigma_k) + \epsilon g(\mathbf{x}_i|\kappa, \mathbf{M}, V)}.
\end{aligned} \tag{A.4}
$$

Likewise we see that,

$$
p(\phi_i = 0|z_i = k, \mathbf{x}_i) = \frac{\epsilon f(\mathbf{x}_i|M, V)}{(1-\epsilon)f(\mathbf{x}_i|\boldsymbol{\mu}_k, \Sigma_k) + \epsilon g(\mathbf{x}_i|\kappa, \mathbf{M}, V)}. \tag{A.5}
$$

Thus

$$
\begin{aligned}
& p(\phi_i = 1, z_i = k|\mathbf{x}_i) \\
& = p(\phi_i = 1|z_i = k, \mathbf{x}_i)p(z_i = k|\mathbf{x}_i) \\
& = p(\phi_i = 1|z_i = k, \mathbf{x}_i)\frac{p(\mathbf{x}_i|z_i = k)p(z_i = k)}{p(\mathbf{x}_i)} \\
& = p(\phi_i = 1|z_i = k, \mathbf{x}_i)\frac{(p(\mathbf{x}_i|z_i = k, \phi_i = 0)p(\phi_i = 0) + p(\mathbf{x}_i|z_i = k, \phi_i = 1)p(\phi_i = 1))\, p(z_i = k)}{p(\mathbf{x}_i)}
\end{aligned} \tag{A.6}
$$

and then substituting values leads to

$$\frac{(1-\epsilon)f(\mathbf{x}_i|\boldsymbol{\mu}_k,\Sigma_k)}{(1-\epsilon)f(\mathbf{x}_i|\boldsymbol{\mu}_k,\Sigma_k)+\epsilon g(\mathbf{x}_i|\kappa,\mathbf{M},V)}\frac{\pi_k\left((1-\epsilon)f(\mathbf{x}_i|\boldsymbol{\mu}_k,\Sigma_k)+\epsilon g(\mathbf{x}_i|\kappa,\mathbf{M},V)\right)}{\sum_{k=1}^{K}\pi_k\left((1-\epsilon)f(\mathbf{x}_i|\boldsymbol{\mu}_k,\Sigma_k)+\epsilon g(\mathbf{x}_i|\kappa,\mathbf{M},V)\right)}= \\ \frac{\pi_k(1-\epsilon)f(\mathbf{x}_i|\boldsymbol{\mu}_k,\Sigma_k)}{\sum_{k=1}^{K}\pi_k\left((1-\epsilon)f(\mathbf{x}_i|\boldsymbol{\mu}_k,\Sigma_k)+\epsilon g(\mathbf{x}_i|\kappa,\mathbf{M},V)\right)}. \tag{A.7}$$

We also see that

$$p(\phi_i=0,z_i=k|\mathbf{x}_i)=\frac{\pi_k\epsilon g(\mathbf{x}_i|\kappa,\mathbf{M},V)}{\sum_{k=1}^{K}\pi_k\left((1-\epsilon)f(\mathbf{x}_i|\boldsymbol{\mu}_k,\Sigma_k)+\epsilon g(\mathbf{x}_i|\kappa,\mathbf{M},V)\right)}. \tag{A.8}$$

We can now formally derive the EM algorithm for this model. First, we compute the expected value of the log-posterior function with respect to the conditional distribution of the latent variable given the observations (under the current estimate of the parameters). For notational convenience we suppress the dependence on the parameters.

$$\begin{aligned}
&Q(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}) \\
&\quad=E_{Z,\Phi|X,\hat{\boldsymbol{\theta}}}[\log p(\boldsymbol{\theta};X,Z,\Phi)] \\
&\quad=\sum_{i=1}^{n}E_{Z,\Phi|X,\hat{\boldsymbol{\theta}}}[\log p(\boldsymbol{\theta};\mathbf{x}_i,z_i,\phi_i)] \\
&\quad=\sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{r=0}^{1}p(z_i=k,\phi_i=r|\mathbf{x}_i)\log(L(\boldsymbol{\theta}_k|\mathbf{x}_i,z_i=k,\phi_i))+\log(p(\pi)+\sum_{k=1}^{K}\log(p(\boldsymbol{\theta}_k)) \\
&\quad=\sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{r=0}^{1}p(z_i=k,\phi_i=r|\mathbf{x}_i)\log(p(\mathbf{x}_i,z_i=k,\phi_i|\boldsymbol{\theta}_k))+\log(p(\pi)+\sum_{k=1}^{K}\log(p(\boldsymbol{\theta}_k)) \\
&\quad=Q'(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}})+D(\boldsymbol{\pi},\boldsymbol{\theta})
\end{aligned} \tag{A.9}$$

We note that the equation splits up into a likelihood term Q' plus the log prior D. The coefficient of the first term in the equation above has already been derived and the other term is given by:

$$\begin{aligned}
&p(\mathbf{x}_i,z_i=k,\phi_i)|\boldsymbol{\theta}_k) \\
&\qquad=p(\mathbf{x}_i,\phi_i|\boldsymbol{\theta}_k,z_i=k)p(z_i=k|\boldsymbol{\theta}_k) \\
&\qquad=\pi_k p(\mathbf{x}_i,\phi_i|\boldsymbol{\theta}_k,z_i=k) \\
&\qquad=\pi_k\left(p(\mathbf{x}_i|\boldsymbol{\theta}_k,z_i=k,\phi_i)p(\phi_i|\boldsymbol{\theta}_k,z_i=k)\right) \\
&\qquad=\pi_k\left(((1-\epsilon)f(\mathbf{x}_i|\boldsymbol{\mu}_k,\Sigma_k))^{\phi_i}(\epsilon g(\mathbf{x}_i|\kappa,\mathbf{M},V))^{1-\phi_i}\right),
\end{aligned} \tag{A.10}$$

where we used that ϕ_i was a binary random variable. Thus we see that

$$\begin{aligned}
&Q'(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}) \\
=&\sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{\Phi} p(z_i=k,\phi_i|\mathbf{x}_i)\log(p(\mathbf{x}_i,z_i=k,\phi_i|\boldsymbol{\theta}_k)) \\
=&\sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{\Phi} p(z_i=k,\phi_i|\mathbf{x}_i)\log(\pi_k((1-\epsilon)f(\mathbf{x}_i|\boldsymbol{\mu}_k,\Sigma_k))^{\phi_i}(\epsilon g(\mathbf{x}_i|\kappa,\mathbf{M},V))^{1-\phi_i}) \\
=&\sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{\Phi} p(z_i=k,\phi_i|\mathbf{x}_i)\left(\log(\pi_k)+\phi_i\log((1-\epsilon)f(\mathbf{x}_i|\boldsymbol{\mu}_k,\Sigma_k))+(1-\phi_i)\log(\epsilon g(\mathbf{x}_i|\kappa,\mathbf{M},V))\right) \\
=&(A)+(B)+(C)+(D)
\end{aligned} \tag{A.11}$$

where

$$\begin{aligned}
(A)&=\sum_{i=1}^{n}\sum_{k=1}^{K} p(z_i=k|\mathbf{x}_i)\log(\pi_k) \\
(B)&=\sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{\Phi} p(z_i=k,\phi_i|\mathbf{x}_i)(\phi_i\log(1-\epsilon)+(1-\phi_i)\log(\epsilon)) \\
(C)&=\sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{\Phi} p(z_i=k,\phi_i|\mathbf{x}_i)\phi_i\log(f(\mathbf{x}_i|\boldsymbol{\mu}_k,\Sigma_k)) \\
(D)&=\sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{\Phi} p(z_i=k,\phi_i|\mathbf{x}_i)(1-\phi_i)\log(g(\mathbf{x}_i|\kappa,\mathbf{M},V)).
\end{aligned} \tag{A.12}$$

Then again using that ϕ_i is binary we can make the following simplifications.

$$\begin{aligned}
(B)&=\sum_{i=1}^{n}\sum_{k=1}^{K} p(z_i=k,\phi_i=1|\mathbf{x}_i)\log(1-\epsilon)+p(z_i=k,\phi_i=0|x_i)\log(\epsilon) \\
(C)&=\sum_{i=1}^{n}\sum_{k=1}^{K} p(z_i=k,\phi_i=1|\mathbf{x}_i)\log(f(\mathbf{x}_i|\boldsymbol{\mu}_k,\Sigma_k)) \\
(D)&=\sum_{i=1}^{n}\sum_{k=1}^{K} p(z_i=k,\phi_i=0|\mathbf{x}_i)\log(g(\mathbf{x}_i|\kappa,\mathbf{M},V)).
\end{aligned} \tag{A.13}$$

Terms can now be maximised by considering terms independently because of linearity. Note that the equations 2.54 and 2.55 are computed with respect to the current estimated values of the parameters. For convenience set the following notation

$$\begin{aligned} a_{ik} =& p(z_i = k, \phi_i = 1|\mathbf{x}_i) \\ b_{ik} =& p(z_i = k, \phi_i = 0|\mathbf{x}_i) \\ w_{ik} =& p(z_i = k|\mathbf{x}_i) = a_{ik} + b_{ik} \\ a_k =& \sum_{i=1}^{n} a_{ik}, a = \sum_{k=1}^{K} a_k \\ b_k =& \sum_{i=1}^{n} b_{ik}, b = \sum_{k=1}^{K} b_k \\ r_k =& \sum_{i=1}^{n} w_{ik} \end{aligned} \tag{A.14}$$

The maximisation step requires finding $argmax_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}})$, this can be found for parameter separately for each linear term. To find $\hat{\epsilon}$, we need only consider computing the maximisation step from equation (B). First set $\epsilon_1 = 1 - \epsilon$ and $\epsilon_2 = \epsilon$ and add the log prior term to equation (B). Thus, the required Lagrangian is

$$\mathcal{L}_\epsilon = a\log(\epsilon_1) + b\log(\epsilon_2) + (u-1)\log(\epsilon_2) + (v-1)\log((\epsilon_1) + \lambda(\epsilon_1 + \epsilon_2 - 1) + constant. \tag{A.15}$$

Solving this system leads to

$$\epsilon = \frac{u + b - 1}{(a + b) + (u + v) - 2}. \tag{A.16}$$

To find the MAP estimate for $\boldsymbol{\pi}$, we examine equation (A) and add the log prior. Furthermore we must maximise $\boldsymbol{\pi}$ under the constraint that $\sum_{k=1}^{K} \pi_k = 1$. The Lagrangian for this constrained optimisation problem is the following,

$$\mathcal{L} = \sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\log(\pi_k) - \log(B(\beta)) + \sum_{k=1}^{K}(\beta_k - 1)\log(\pi_k) + \lambda\left(\sum_{k=1}^{K}\pi_k - 1\right). \tag{A.17}$$

The fixed point of this Lagrangian solves the required constrained optimisation problem and $B(\beta)$ denotes the Beta function with parameter β.

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \pi_k} &= \frac{r_k}{\pi_k} + \frac{\beta_k - 1}{\pi_k} + \lambda = 0 \\ \frac{\partial \mathcal{L}}{\partial \lambda} &= \sum_{k=1}^{K} \pi_k - 1 = 0\end{aligned} \tag{A.18}$$

Solving this pair of equations yields

$$\pi_k = \frac{r_k + \beta_k - 1}{N + \sum \beta_k - K}. \tag{A.19}$$

To find the posterior mode of the remaining parameters requires some work. First we recall that the normal inverse-Wishart prior is proportional to:

$$\prod_{k=1}^{K} |\Sigma_k|^{\frac{\nu_0 + D + 2}{2}} \exp\left(-\frac{1}{2} tr(\Sigma_k^{-1} S_0^{-1})\right) \exp\left(-\frac{\lambda_0}{2} tr(\Sigma_k^{-1}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_0)^T(\boldsymbol{\mu}_k - \boldsymbol{\mu}_0))\right). \tag{A.20}$$

The required equation we are interested in is (C).

$$\begin{aligned}&\sum_{i=1}^{n}\sum_{k=1}^{K} a_{ik} \log(f(\mathbf{x}_i|\boldsymbol{\mu}_k, \Sigma_k)) \\ &\quad= \sum_{k=1}^{K}\left\{-\sum_{i=1}^{n} a_{ik}\frac{D\log(2\pi)}{2} - \frac{1}{2}\sum_{k=1}^{n} a_{ik}\log|\Sigma_k| - \frac{1}{2}\sum_{i=1}^{n} a_{ik} tr\left(\Sigma_k^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_k)^T(\mathbf{x}_i - \boldsymbol{\mu}_k)\right)\right\} \\ &\quad= \sum_{k=1}^{K}\left\{-a_k\frac{D\log(2\pi)}{2} - \frac{1}{2}a_k\log|\Sigma_k| - \frac{1}{2} tr\left(\Sigma_k^{-1}\sum_{i=1}^{n} a_{ik}(\mathbf{x}_i - \boldsymbol{\mu}_k)^T(\mathbf{x}_i - \boldsymbol{\mu}_k)\right)\right\}.\end{aligned} \tag{A.21}$$

Now to derive the M-step objective we remove the constant terms and add on the log prior. This leads to

$$\begin{aligned}&\sum_{k=1}^{K}\left\{\frac{\nu_0 + D + 2}{2}\log|\Sigma_k| - \frac{1}{2} tr\left(\Sigma_k^{-1} S_0^{-1}\right) - \frac{\lambda_0}{2} tr\left(\Sigma_k^{-1}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_0)^T(\boldsymbol{\mu}_k - \boldsymbol{\mu}_0)\right)\right\} \\ &+ \sum_{k=1}^{K}\left\{-\frac{1}{2}a_k\log|\Sigma_k| - \frac{1}{2} tr\left(\Sigma_k^{-1}\sum_{i=1}^{n} a_{ik}(\mathbf{x}_i - \boldsymbol{\mu}_k)^T(\mathbf{x}_i - \boldsymbol{\mu}_k)\right)\right\}.\end{aligned} \tag{A.22}$$

This can be rewritten as

$$\sum_{k=1}^{K}\left\{\frac{\nu_0+D+2+a_k}{2}\log|\Sigma_k|-\frac{1}{2}tr\left(\Sigma_k^{-1}S_0^{-1}\right)-\frac{\lambda_0}{2}tr\left(\Sigma_k^{-1}(\boldsymbol{\mu}_k-\boldsymbol{\mu}_0)^T(\boldsymbol{\mu}_k-\boldsymbol{\mu}_0)\right)\right\} + \sum_{k=1}^{K}\left\{-\frac{1}{2}tr\left(\Sigma_k^{-1}\sum_{i=1}^{n}a_{ik}(\mathbf{x}_i-\boldsymbol{\mu}_k)^T(\mathbf{x}_i-\boldsymbol{\mu}_k)\right)\right\}. \tag{A.23}$$

Now define $\bar{\mathbf{x}}_k=(\sum_{i=i}^{n}a_{ik}\mathbf{x}_i)/a_k$ and note the following algebraic rearrangements.

$$\begin{aligned}
&\sum_{i=1}^{n}a_{ik}(\mathbf{x}_i-\boldsymbol{\mu}_k)^T(\mathbf{x}_i-\boldsymbol{\mu}_k)\\
&=\sum_{i=1}^{n}a_{ik}\mathbf{x}_i^T\mathbf{x}_i-\boldsymbol{\mu}_k^T\mathbf{x}_i-\mathbf{x}_i^T\boldsymbol{\mu}_k+\boldsymbol{\mu}_k^T\boldsymbol{\mu}_k\\
&=\sum_{i=1}^{n}a_{ik}\mathbf{x}_i^T\mathbf{x}_i-\boldsymbol{\mu}_k^T\sum_{i=1}^{n}a_{ik}\mathbf{x}_i-\left(\sum_{i=1}^{n}a_{ik}\mathbf{x}_i^T\right)\boldsymbol{\mu}_k+a_k\boldsymbol{\mu}_k^T\boldsymbol{\mu}_k\\
&=\sum_{i=1}^{n}a_{ik}\mathbf{x}_i^T\mathbf{x}_i-a_k\boldsymbol{\mu}_k^T\bar{\mathbf{x}}_k-a_k\bar{\mathbf{x}}_k^T\boldsymbol{\mu}_k+a_k\boldsymbol{\mu}_k^T\boldsymbol{\mu}_k\\
&=\sum_{i=1}^{n}a_{ik}\mathbf{x}_i^T\mathbf{x}_i-a_k\bar{\mathbf{x}}_k^T\bar{\mathbf{x}}_k+a_k(\bar{\mathbf{x}}_k-\boldsymbol{\mu}_k)^T(\bar{\mathbf{x}}_k-\boldsymbol{\mu}_k)\\
&=\sum_{i=1}^{n}a_{ik}(\mathbf{x}_i-\bar{\mathbf{x}}_k)^T(\mathbf{x}_i-\bar{\mathbf{x}}_k)+a_k(\bar{\mathbf{x}}_k-\boldsymbol{\mu}_k)^T(\bar{\mathbf{x}}_k-\boldsymbol{\mu}_k)
\end{aligned} \tag{A.24}$$

This allows us to rewrite equation A.23 as

$$\sum_{k=1}^{K}\left\{\frac{\nu_0+D+2+a_k}{2}\log|\Sigma_k|-\frac{1}{2}tr\left(\Sigma_k^{-1}\left(S_0^{-1}+\sum_{i=1}^{n}a_{ik}(\mathbf{x}_i-\bar{\mathbf{x}}_k)^T(\mathbf{x}_i-\bar{\mathbf{x}}_k)\right)\right)\right\} + \sum_{k=1}^{K}\left\{-\frac{1}{2}tr\left(\Sigma_k^{-1}\left(\lambda_0(\boldsymbol{\mu}_k-\boldsymbol{\mu}_0)^T(\boldsymbol{\mu}_k-\boldsymbol{\mu}_0)\right)+a_k(\bar{\mathbf{x}}_k-\boldsymbol{\mu}_k)^T(\bar{\mathbf{x}}_k-\boldsymbol{\mu}_k)\right)\right\} \tag{A.25}$$

This can be written as:

$$\sum_{k=1}^{K}\left\{\frac{\nu_k+D+2}{2}\log|\Sigma_k|-\frac{1}{2}tr\left(\Sigma_k^{-1}S_k^{-1}\right)-\frac{1}{2}tr\left(\Sigma_k^{-1}\left(\lambda_k(\boldsymbol{\mu}_k-\boldsymbol{m}_k)^T(\boldsymbol{\mu}_k-\boldsymbol{m}_k)\right)\right)\right\} \tag{A.26}$$

where,

$$\begin{aligned}
\lambda_k =& \lambda_0 + a_k \\
\nu_k =& \nu_0 + a_k \\
\boldsymbol{m}_k =& \frac{a_k \bar{\mathbf{x}}_k + \lambda_0 \boldsymbol{\mu}_0}{\lambda_k} \\
S_k^{-1} =& S_0^{-1} + \frac{\lambda_0 a_k}{\lambda_k} (\bar{\mathbf{x}}_k - \boldsymbol{\mu}_0)^T (\bar{\mathbf{x}}_k - \boldsymbol{\mu}_0) + \sum_{i=1}^{n} a_{ik} (\mathbf{x}_i - \bar{\mathbf{x}}_k)^T (\mathbf{x}_i - \bar{\mathbf{x}}_k)
\end{aligned} \tag{A.27}$$

Thus the parameters of the posterior mode are:

$$\begin{aligned}
\hat{\boldsymbol{\mu}}_k =& \boldsymbol{m}_k \\
\hat{\Sigma}_k =& \frac{1}{\nu_k + D + 2} S_k^{-1}
\end{aligned} \tag{A.28}$$

To summarise the EM algorithm, we iterate between the two steps:

E-Step: Given the current parameters compute the values given by equations (A.14), with formulas provided in equations (2.54) and (2.55).

M-Step: Compute

$$\epsilon = \frac{u + b - 1}{(a + b) + (u + v) - 2},$$

and

$$\pi_k = \frac{r_k + \beta_k - 1}{N + \sum \beta_k - K},$$

as well as

$$\bar{\mathbf{x}}_k = \frac{1}{a_k} \left(\sum_{i=i}^{n} a_{ik} \mathbf{x}_i \right)$$

Compute the MAP estimates given by equations (A.28). These estimates are then used in the following iteration of the E-step. Iterate until $|Q(\boldsymbol{\theta}|\boldsymbol{\theta}_t) - Q(\boldsymbol{\theta}|\boldsymbol{\theta_{t-1}})| < \delta$ for some pre-specified $\delta > 0$.

A.2 Appendix 2: Derivation of collapsed Gibbs sampler for TAGM model

To derive the Gibbs sampler, we write down all the conditional probabilities. Then, exploiting conjugacy, we can marginalise parameters in the model. Recall the total joint probability is the following:

$$\begin{aligned} p(\boldsymbol{\theta}, \Theta, X, Z, \Phi) =& p(X, Z, \Phi|\theta, \boldsymbol{\pi}, \epsilon)p(\epsilon|u, v)p(\theta|H)p(\boldsymbol{\pi}|\beta) \\ =& \prod_{i=1}^{n}\prod_{k=1}^{K}\left(\pi_k((1-\epsilon)\mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_k, \Sigma_k))^{\mathbb{1}(\phi_i=1)}(\epsilon\mathcal{T}(\mathbf{x}_i|\kappa, \mathbf{M}, V))^{\mathbb{1}(\phi_i=0)}\right)^{\mathbb{1}(z_i=k)} \\ & \cdot\left(\prod_{k=1}^{K}\mathcal{NIW}(H)\right)\cdot Dir(\beta)\cdot\mathcal{B}(u, v). \end{aligned} \tag{A.29}$$

Suppose we know the hidden latent component allocations z_i and outlier allocations ϕ_i. Then we could sample from the a required normal distribution. The conditional probability of the parameters given the allocations is given by:

$$p(\theta_k|X, Z, \Phi, \theta_{-k}, \beta, u, v, H) \propto p_0(\theta_k)\prod_{i=1}^{n}N(\mathbf{x}_i|\boldsymbol{\mu}_k, \Sigma_k)^{\mathbb{1}(\phi_i=1)}. \tag{A.30}$$

The prior is conjugate and so the posterior belongs to the same parametric family as the prior, a NIW distribution, and so the parameters can be updated as follows:

$$\begin{aligned} m_k =& \frac{n_k\bar{\mathbf{x}}_k + \lambda_0\boldsymbol{\mu}_0}{\lambda_k} \\ \lambda_k =& \lambda_0 + n_k \\ \nu_k =& \nu_0 + n_k \\ S_k =& S_0 + \sum_{i:z_i=k,\phi_i=1}(\mathbf{x}_i - \bar{\mathbf{x}})^T(\mathbf{x}_i - \bar{\mathbf{x}}) + \frac{\lambda_0 n_k}{\lambda_k}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)^T(\bar{\mathbf{x}} - \boldsymbol{\mu}_0), \end{aligned} \tag{A.31}$$

where $n_k = |\{\mathbf{x}_i|z_i = k, \phi_i = 1\}|$. Now we write down the conditional of the component allocations

$$p(z_i = k|X, z_{-i}, \Phi, \theta, \beta, u, v, H) \propto p_0(z_i = k|z_{-i}, \beta)p(\mathbf{x}_i|\mathbf{x}_{-i}, z_{-i}, z_i = k, \Phi, H). \tag{A.32}$$

The first term in this equation is

$$p_0(z_i = k|z_{-i}, \beta) = \frac{p(z_i = k, z_{-i}|\beta)}{p(z_{-i}|\beta)} = \frac{p(Z|\beta)}{p(z_{-i}|\beta)}. \tag{A.33}$$

To calculate the numerator we proceed by marginalising over $\boldsymbol{\pi}$ as follows

$$p(Z|\beta) = \int p(z|\boldsymbol{\pi})p(\boldsymbol{\pi}|\beta)d\boldsymbol{\pi} = \frac{\Gamma(\beta)}{\Gamma(n+\beta)} \prod_{k=1}^{K} \frac{\Gamma(n_k + \beta_k)}{\Gamma(\beta_k)}. \tag{A.34}$$

Hence, we arrive at the following probability:

$$p_0(z_i = k|z_{-i}, \beta) = \frac{n_{k\setminus i} + \beta_k}{n + \sum \beta_k - 1}. \tag{A.35}$$

The conditional for the second term of A.32 is more tricky. First note the following conditional distributions

$$\begin{aligned}
&\mathbf{x}_i|z_i = k, X_{k\setminus i}, \phi_i = 1, \Phi, z_{-i} \sim \mathcal{N}(\mathbf{x}_i|\theta_k) \\
&\mathbf{x}_i|z_i = k, X_{k\setminus i}, \phi_i = 0, \Phi, z_{-i} \sim \mathcal{T}(\mathbf{x}_i|\kappa, \mathbf{M}, V), \\
&\mathbf{x}_i|z_i = k, X_{k\setminus i}, \phi_i, \Phi, z_{-i} \sim N(\mathbf{x}_i|\theta_k)^{\mathbb{1}(\phi_i=1)}\mathcal{T}(\mathbf{x}_i|, \kappa, \mathbf{M}, V)^{\mathbb{1}(\phi_i=0)},
\end{aligned} \tag{A.36}$$

where we denote $X_{k\setminus i}$ as the observations associated with class k, besides x_i. Now, we first note that:

$$p(\mathbf{x}_i|z_i = k, X_{k\setminus i}, \phi_i, \Phi, H, z_{-i}) = p(\mathbf{x}_i|X_{k\setminus i}, \phi_i, \Phi, H) = \frac{p(\mathbf{x}_i, X_{k\setminus i}|\phi_i, \Phi, H)}{p(X_{k\setminus i}|\phi_i, \Phi, H)}. \tag{A.37}$$

Thus, we find an equation for the numerator, using the fact that terms associated with $\phi_i = 0$ do not depend on k and thus can be absorbed into the normalising constant.

$$p(X_k|\phi_i, \Phi, H) \propto \prod_{i:\phi_i=1} \int p(\mathbf{x}_i|z_i = k, \Phi, H, \theta_k)p(\theta_k|H)d\theta_k. \tag{A.38}$$

This is the marginal likelihood of the data. Thus the ratio in A.37 is the posterior predictive which is given by the non-centred T-distribution with formula given by:

$$\mathcal{T}\left(v_k - d + 1, m_k, \frac{(1+\lambda_k)S_k}{\lambda_k(v_k - d + 1)}\right).$$

Thus, we can compute the following:

$$
\begin{aligned}
p(z_i = k|X, z_{-i}, \Phi, \theta, \beta, u, v, H) &\propto p_0(z_i = k|z_{-i}, \beta)p(\mathbf{x}_i|\mathbf{x}_{-i}, z_{-i}, \Phi, z_i = k, H) \\
&= \frac{n_{k\setminus i} + \beta_k}{n + \sum \beta_k - 1}\mathcal{T}\left(\mathbf{x}_i|v_k - d + 1, m_k, \frac{(1+\lambda_k)S_k}{\lambda_k(v_k - d + 1)}\right).
\end{aligned} \tag{A.39}
$$

It remains to compute the conditional for the ϕ_i. By first recalling that ϕ_i is binary we see that

$$
p(\phi_i|X, Z, \theta, \beta, u, v, H) \propto p_0(\phi_i)\prod_{i=1}^{n} N(\mathbf{x}_i|\theta_{z_i})^{\mathbb{1}(\phi_i=1)}T(\mathbf{x}_i|\kappa, M, V)^{\mathbb{1}(\phi_i=0)} \tag{A.40}
$$

can be written as

$$
\begin{aligned}
p(\phi_i = 1|X, Z, \theta, \phi_{-i}, \beta, u, v, H) &\propto p_0(\phi_i = 1|\phi_{-i}, u, v)p(\mathbf{x}_i|\mathbf{x}_{-i}, \phi_i = 1, Z, \theta, \Phi, \beta, u, v, H), \\
p(\phi_i = 0|X, Z, \theta, \phi_{-i}, \beta, u, v, H) &\propto p_0(\phi_i = 0|\phi_{-i}, u, v)p(\mathbf{x}_i|\mathbf{x}_{-i}, \phi_i = 0, Z, \theta, \Phi, \beta, u, v, H).
\end{aligned} \tag{A.41}
$$

First we need to compute a formula for $p_0(\phi_i|\phi_{-i}, u, v)$. First we see that

$$
p_0(\phi_i|\phi_{-i}, u, v) = \frac{p(\Phi|u, v)}{p(\phi_{-i}|u, v)}. \tag{A.42}
$$

The numerator can be computed by marginalising over ϵ:

$$
p(\Phi|u, v) = \int p(\Phi|\epsilon)p(\epsilon|u, v)d\epsilon. \tag{A.43}
$$

We denote $\sum \mathbb{1}(\phi_i = 1) = \tau_1$ and $\sum \mathbb{1}(\phi_i = 0) = \tau_0 = 1 - \tau_1$. Then it is easy to see that

$$
\begin{aligned}
p(\Phi|u, v) &= \int p(\Phi|\epsilon)p(\epsilon|u, v)d\epsilon \\
&= \frac{1}{B(u, v)}\int (1-\epsilon)^{\tau_1+v-1}\epsilon^{\tau_0+u-1}d\epsilon \\
&= \frac{B(\tau_0 + u, \tau_1 + v)}{B(u, v)}.
\end{aligned} \tag{A.44}
$$

Hence,

$$\begin{aligned} p(\phi_i = 1|\phi_{-i}, u, v) = &\frac{B(\tau_0 + u, \tau_1 + v)}{B(u, v)} \cdot \frac{B(u, v)}{B(\tau_0 + u, \tau_1 + v - 1)} \\ = &\frac{\tau_1 + v - 1}{n + u + v - 1}, \end{aligned} \tag{A.45}$$

where $n = \tau_1 + \tau_2$. In general,

$$p(\phi_i = s|\phi_{-i}, u, v) = \frac{\tau_{s\setminus i} + v^s u^{1-s}}{n + u + v - 1}. \tag{A.46}$$

Now we return to computing $p(\mathbf{x}_i|\mathbf{x}_{-i}, Z, \theta, \phi_i = 1, \Phi, \beta, u, v, H)$. First we see that

$$p(\mathbf{x}_i|\mathbf{x}_{-i}, Z, \theta, \phi_i = 1, \Phi, \beta, u, v, H) = \frac{p(X|Z, \theta, \phi_i = 1, \Phi, \beta, u, v, H)}{p(\mathbf{x}_{-i}|Z, \theta, \phi_i = 1, \Phi, \beta, u, v, H)}. \tag{A.47}$$

Thus if we integrate over the parameters, we would have a ratio of marginal likelihoods giving the posterior predictive which is a non-centred T-distribution:

$$p(\mathbf{x}_i|\mathbf{x}_{-i}, Z, \theta, \phi_i = 1, \Phi, \beta, u, v, H) = \mathcal{T}\left(v_k - d + 1, m_k, \frac{(1 + \lambda_k)S_k}{\lambda_k(v_k - d + 1)}\right). \tag{A.48}$$

In the other case that $\phi = 0$, we have that

$$p(x_i|x_{-i}, Z, \theta, \phi_i = 0, \Phi, \beta, u, v, H) = \mathcal{T}(x_i|\kappa, \mathbf{M}, V). \tag{A.49}$$

Thus we can compute:

$$p(\phi_i|X, Z, \theta, \phi_{-i}, \beta, u, v, H) \tag{A.50}$$

and sample from the required distribution. Thus, we can summarise the collapsed Gibbs sampler as follows:

1. Update the priors with the labelled data
2. For the unlabelled observations, in turn, compute the probability of assigning to each component
3. Sample a label according to this probability
4. Compute the probability of belonging to this class or the outlier component
5. Sample an indicator to a class specific component or the outlier component

6. If we assign to the class specific component update the class specific posterior distribution with the statistics of this observation
7. Update other posteriors as appropriate.
8. Once all unlabelled observations have a been assigned, consider the observations sequentially, removing the statistics from the posteriors and then performing steps 2-7. We repeat this process for all unlabelled observations.
9. repeat 7-8 until convergence of the Markov-chain.

The computational bottleneck in the algorithm is computing the posterior updates for the parameters

$$\begin{aligned} m_k =& \frac{n_k \bar{\mathbf{x}}_k + \lambda_0 \boldsymbol{\mu}_0}{\lambda_k} \\ \lambda_k =& \lambda_0 + n_k \\ \nu_k =& \nu_0 + n_k \\ S_k =& S_0 + \sum_{i:z_i=k,\phi_i=1} (\mathbf{x}_i - \bar{\mathbf{x}})^T (\mathbf{x}_i - \bar{\mathbf{x}}) + \frac{\lambda_0 n_k}{\lambda_k} (\bar{\mathbf{x}} - \boldsymbol{\mu}_0)^T (\bar{\mathbf{x}} - \boldsymbol{\mu}_0), \end{aligned} \tag{A.51}$$

We first note that

$$S_k = S_0 + \sum_{i:z_i=k,\phi_i=1} \mathbf{x}_i^T \mathbf{x}_i + \lambda_0 \boldsymbol{\mu}_0^T \boldsymbol{\mu}_0 - \lambda_k \boldsymbol{\mu}_k^T \boldsymbol{\mu}_k \tag{A.52}$$

Let us denote $T = \sum_{i:z_i=k,\phi_i=1} \mathbf{x}_i^T \mathbf{x}_i$. Thus we can derive a set of iterative updates to speed up computation when adding/removing statistics from clusters. More precisely, indicating updated posterior parameters by a prime, if we remove statistics of observation i from cluster k, we see that

$$\begin{aligned} m'_k =& \frac{\lambda_k m_k - \mathbf{x}_i}{\lambda_k - 1} \\ \lambda'_k =& \lambda_k - 1 \\ \nu'_k =& \nu_k - 1 \\ T' =& T - \mathbf{x}_i^T \mathbf{x}_i \\ S'_k =& S_0 + T' + \lambda_0 \boldsymbol{\mu}_0^T \boldsymbol{\mu}_0 - \lambda_k m_k'^T m'_k. \end{aligned} \tag{A.53}$$

Likewise if we add the statistics of observation i to cluster k, we see that

$$
\begin{aligned}
m'_k =& \frac{\lambda_k m_k + \mathbf{x}_i}{\lambda_k + 1} \\
\lambda'_k =& \lambda_k + 1 \\
\nu'_k =& \nu_k + 1 \\
T' =& T + \mathbf{x}_i^T \mathbf{x}_i \\
S'_k =& S_0 + T' + \lambda_0 \boldsymbol{\mu}_0^T \boldsymbol{\mu}_0 - \lambda_k m_k'^T m'_k.
\end{aligned}
\tag{A.54}
$$

A.3 Appendix 3: Convergence diagnostics of EM algorithm

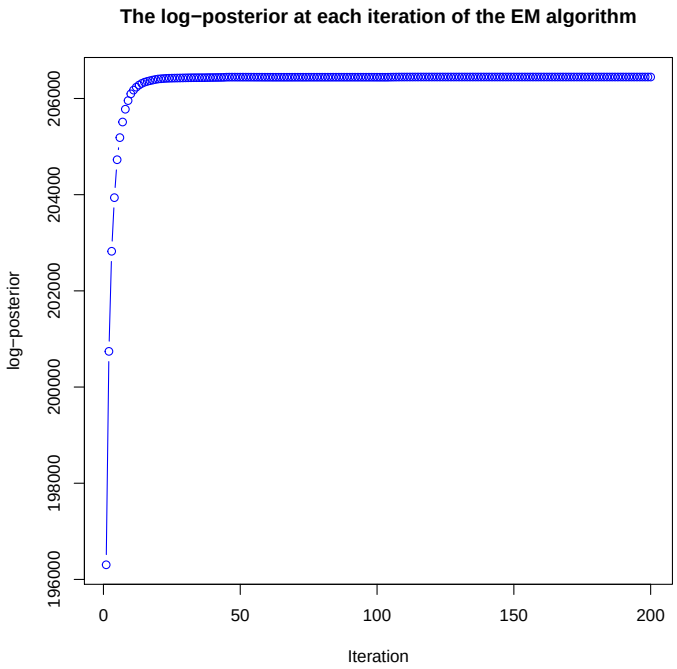


Fig. A.1 Plot of the log-posterior at each iteration of the EM algorithm to demonstrate monotonicity and convergence

A.4 Appendix 4: Trace plots for assessing MCMC convergence

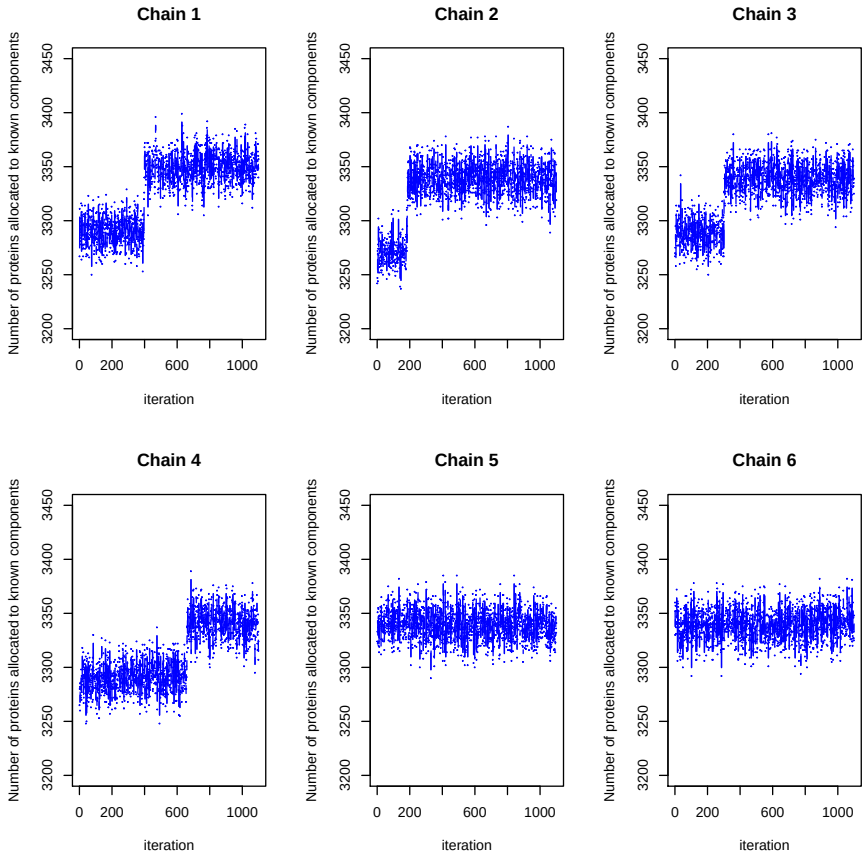


Fig. A.2 Trace plots of the number of proteins allocated to the known components in each of 6 parallel MCMC runs. Chain 4 is discarded because of lack of convergence. 600 samples are retained from remaining chains and pooled.

A.5 Appendix 5: F1 t-tests

	SVM	KNN	MAP
KNN	2.7E-03		
MAP	3.3E-02	3.4E-01	
MCMC	3.4E-01	3.3E-02	2.3E-01

Table A.1 Adjusted P-values for pairwise T-tests for Macro F-1 score classifier evaluation on the Drosophila dataset

	SVM	KNN	MAP
KNN	1.2E-02		
MAP	2.7E-01	1.5E-01	
MCMC	4.9E-01	1.9E-03	1.1E-01

Table A.2 Adjusted P-values for pairwise T-tests for Macro F-1 score classifier evaluation on the Chicken DT40 dataset

	SVM	KNN	MAP
KNN	1.0E+00		
MAP	1.0E+00	1.0E+00	
MCMC	3.3E-01	6.0E-02	1.1E-05

Table A.3 Adjusted P-values for pairwise T-tests for Macro F-1 score classifier evaluation on the mouse dataset

	SVM	KNN	MAP
KNN	1.4E-35		
MAP	3.3E-06	6.7E-21	
MCMC	8.0E-59	3.2E-91	2.4E-70

Table A.4 Adjusted P-values for pairwise T-tests for Macro F-1 score classifier evaluation on the HeLa dataset

	SVM	KNN	MAP
KNN	1.3E-02		
MAP	4.3E-04	3.3E-09	
MCMC	5.8E-01	3.5E-03	3.1E-03

Table A.5 Adjusted P-values for pairwise T-tests for Macro F-1 score classifier evaluation on the U2-OS dataset

	SVM	KNN	MAP
KNN	2.2E-08		
MAP	1.0E-34	6.8E-14	
MCMC	7.4E-05	5.3E-02	1.0E-20

Table A.6 Adjusted P-values for pairwise T-tests for Macro F-1 score classifier evaluation on the HeLa wild (Hirst et al.) dataset

	SVM	KNN	MAP
KNN	5.3E-02		
MAP	1.7E-23	7.9E-27	
MCMC	9.1E-02	5.8E-04	1.8E-19

Table A.7 Adjusted P-values for pairwise T-tests for Macro F-1 score classifier evaluation on the HeLa KO1 (Hirst et al.) dataset

	SVM	KNN	MAP
KNN	1.3E-01		
MAP	1.1E-55	1.1E-55	
MCMC	1.0E-18	6.3E-22	2.0E-26

Table A.8 Adjusted P-values for pairwise T-tests for Macro F-1 score classifier evaluation on the HeLa KO2 (Hirst et al.) dataset

	SVM	KNN	MAP
KNN	9.6E-02		
MAP	4.1E-07	1.1E-09	
MCMC	2.8E-27	1.0E-28	6.3E-10

Table A.9 Adjusted P-values for pairwise T-tests for Macro F-1 score classifier evaluation on the Primary Fibroblasts Mock 24hpi dataset

	SVM	KNN	MAP
KNN	6.6E-07		
MAP	1.3E-10	2.0E-01	
MCMC	1.6E-05	2.0E-01	6.2E-03

Table A.10 Adjusted P-values for pairwise T-tests for Macro F-1 score classifier evaluation on the Primary Fibroblasts Mock 48hpi dataset

	SVM	KNN	MAP
KNN	3.9E-03		
MAP	9.5E-01	8.6E-03	
MCMC	6.4E-02	3.0E-01	8.6E-02

Table A.11 Adjusted P-values for pairwise T-tests for Macro F-1 score classifier evaluation on the Primary Fibroblasts Mock 72hpi dataset

	SVM	KNN	MAP
KNN	8.6E-03		
MAP	1.1E-02	8.6E-01	
MCMC	3.7E-06	1.6E-02	3.3E-02

Table A.12 Adjusted P-values for pairwise T-tests for Macro F-1 score classifier evaluation on the Primary Fibroblasts Mock 96hpi dataset

	SVM	KNN	MAP
KNN	1.9E-23		
MAP	1.4E-02	2.3E-34	
MCMC	3.8E-07	1.6E-81	2.0E-02

Table A.13 Adjusted P-values for pairwise T-tests for Macro F-1 score classifier evaluation on the Primary Fibroblasts Mock 120hpi dataset

	SVM	KNN	MAP
KNN	4.6E-01		
MAP	2.6E-05	1.7E-04	
MCMC	1.7E-04	1.3E-03	5.5E-01

Table A.14 Adjusted P-values for pairwise T-tests for Macro F-1 score classifier evaluation on the Primary Fibroblasts HCMV 24hpi dataset

	SVM	KNN	MAP
KNN	1.0E-02		
MAP	4.6E-01	1.5E-03	
MCMC	1.2E-02	7.3E-01	1.5E-03

Table A.15 Adjusted P-values for pairwise T-tests for Macro F-1 score classifier evaluation on the Primary Fibroblasts HCMV 48hpi dataset

	SVM	KNN	MAP
KNN	5.5E-02		
MAP	9.5E-06	3.4E-02	
MCMC	1.1E-01	6.2E-01	6.4E-03

Table A.16 Adjusted P-values for pairwise T-tests for Macro F-1 score classifier evaluation on the Primary Fibroblasts HCMV 72hpi dataset

	SVM	KNN	MAP
KNN	2.8E-01		
MAP	2.6E-09	7.2E-08	
MCMC	4.2E-10	5.6E-09	5.7E-01

Table A.17 Adjusted P-values for pairwise T-tests for Macro F-1 score classifier evaluation on the Primary Fibroblasts HCMV 96hpi dataset

	SVM	KNN	MAP
KNN	2.3E-04		
MAP	7.1E-04	3.8E-10	
MCMC	1.4E-01	5.7E-02	6.0E-05

Table A.18 Adjusted P-values for pairwise T-tests for Macro F-1 score classifier evaluation on the Primary Fibroblasts HCMV 120hpi dataset

	SVM	KNN	MAP
KNN	6.7E-06		
MAP	6.3E-05	4.4E-01	
MCMC	4.4E-01	6.7E-06	8.3E-05

Table A.19 Adjusted P-values for pairwise T-tests for Macro F-1 score classifier evaluation on the E14TG2a dataset

A.6 Appendix 6: Quadratic loss t-tests

	SVM	KNN	MAP
KNN	5.9E-13		
MAP	1.1E-04	9.6E-124	
MCMC	2.2E-23	3.3E-58	5.9E-171

Table A.20 Adjusted P-values for pairwise T-tests for Quadratic Loss classifier evaluation on the Drosphila dataset

	SVM	KNN	MAP
KNN	3.2E-08		
MAP	1.7E-26	1.3E-128	
MCMC	4.2E-13	8.8E-37	7.0E-135

Table A.21 Adjusted P-values for pairwise T-tests for Quadratic Loss classifier evaluation on the Chicken DT40 dataset

	SVM	KNN	MAP
KNN	5.5E-14		
MAP	3.0E-25	6.3E-128	
MCMC	7.4E-26	1.7E-129	1.6E-14

Table A.22 Adjusted P-values for pairwise T-tests for Quadratic Loss classifier evaluation on the mouse dataset

	SVM	KNN	MAP
KNN	1.2E-02		
MAP	9.4E-07	7.4E-86	
MCMC	5.5E-08	2.7E-89	2.4E-12

Table A.23 Adjusted P-values for pairwise T-tests for Quadratic Loss classifier evaluation on the HeLa dataset

	SVM	KNN	MAP
KNN	6.8E-02		
MAP	7.4E-17	1.1E-73	
MCMC	1.4E-20	6.7E-81	8.3E-41

Table A.24 Adjusted P-values for pairwise T-tests for Quadratic Loss classifier evaluation on the U2-OS dataset

	SVM	KNN	MAP
KNN	2.3E-92		
MAP	9.0E-13	2.4E-83	
MCMC	6.6E-19	3.0E-81	1.1E-01

Table A.25 Adjusted P-values for pairwise T-tests for Quadratic Loss classifier evaluation on the HeLa wild (Hirst et al.) dataset

	SVM	KNN	MAP
KNN	5.2E-97		
MAP	1.4E-02	1.2E-90	
MCMC	2.3E-09	7.0E-95	2.2E-02

Table A.26 Adjusted P-values for pairwise T-tests for Quadratic Loss classifier evaluation on the HeLa KO1 (Hirst et al.) dataset

	SVM	KNN	MAP
KNN	8.9E-93		
MAP	3.1E-01	8.1E-91	
MCMC	9.0E-06	1.5E-83	8.9E-05

Table A.27 Adjusted P-values for pairwise T-tests for Quadratic Loss classifier evaluation on the HeLa KO2 (Hirst et al.) dataset

	SVM	KNN	MAP
KNN	6.1E-13		
MAP	1.4E-18	4.4E-81	
MCMC	3.2E-18	7.2E-77	5.9E-03

Table A.28 Adjusted P-values for pairwise T-tests for Quadratic Loss classifier evaluation on the Primary Fibroblasts Mock 24hpi dataset

	SVM	KNN	MAP
KNN	6.1E-18		
MAP	3.6E-24	2.2E-57	
MCMC	1.4E-24	3.6E-61	3.6E-04

Table A.29 Adjusted P-values for pairwise T-tests for Quadratic Loss classifier evaluation on the Primary Fibroblasts Mock 48hpi dataset

	SVM	KNN	MAP
KNN	1.2E-15		
MAP	4.5E-23	2.5E-89	
MCMC	4.2E-23	5.1E-91	4.4E-01

Table A.30 Adjusted P-values for pairwise T-tests for Quadratic Loss classifier evaluation on the Primary Fibroblasts Mock 72hpi dataset

	SVM	KNN	MAP
KNN	1.8E-13		
MAP	1.4E-20	3.6E-126	
MCMC	5.0E-20	1.5E-109	5.3E-07

Table A.31 Adjusted P-values for pairwise T-tests for Quadratic Loss classifier evaluation on the Primary Fibroblasts Mock 96hpi dataset

	SVM	KNN	MAP
KNN	6.7E-14		
MAP	1.0E-19	2.6E-45	
MCMC	8.0E-20	2.4E-45	2.5E-02

Table A.32 Adjusted P-values for pairwise T-tests for Quadratic Loss classifier evaluation on the Primary Fibroblasts Mock 120hpi dataset

	SVM	KNN	MAP
KNN	6.0E-22		
MAP	2.8E-27	6.4E-53	
MCMC	1.4E-27	1.5E-56	3.0E-03

Table A.33 Adjusted P-values for pairwise T-tests for Quadratic Loss classifier evaluation on the Primary Fibroblasts HCMV 24hpi dataset

	SVM	KNN	MAP
KNN	1.9E-26		
MAP	1.3E-33	2.7E-84	
MCMC	1.3E-33	2.7E-84	6.0E-01

Table A.34 Adjusted P-values for pairwise T-tests for Quadratic Loss classifier evaluation on the Primary Fibroblasts HCMV 48hpi dataset

	SVM	KNN	MAP
KNN	6.3E-20		
MAP	1.9E-25	2.7E-57	
MCMC	1.2E-25	3.4E-58	1.5E-02

Table A.35 Adjusted P-values for pairwise T-tests for Quadratic Loss classifier evaluation on the Primary Fibroblasts HCMV 72hpi dataset

	SVM	KNN	MAP
KNN	1.7E-25		
MAP	9.3E-32	1.9E-56	
MCMC	9.3E-32	1.2E-54	7.1E-01

Table A.36 Adjusted P-values for pairwise T-tests for Quadratic Loss classifier evaluation on the Primary Fibroblasts HCMV 96hpi dataset

	SVM	KNN	MAP
KNN	6.5E-25		
MAP	5.3E-32	1.1E-71	
MCMC	7.1E-32	8.4E-71	5.7E-02

Table A.37 Adjusted P-values for pairwise T-tests for Quadratic Loss classifier evaluation on the Primary Fibroblasts HCMV 120hpi dataset

	SVM	KNN	MAP
KNN	4.7E-04		
MAP	4.7E-21	1.5E-103	
MCMC	3.3E-12	1.8E-57	1.3E-137

Table A.38 Adjusted P-values for pairwise T-tests for Quadratic Loss classifier evaluation on the E14TG2a dataset

A.7 Appendix 7: GO enrichment analysis figures

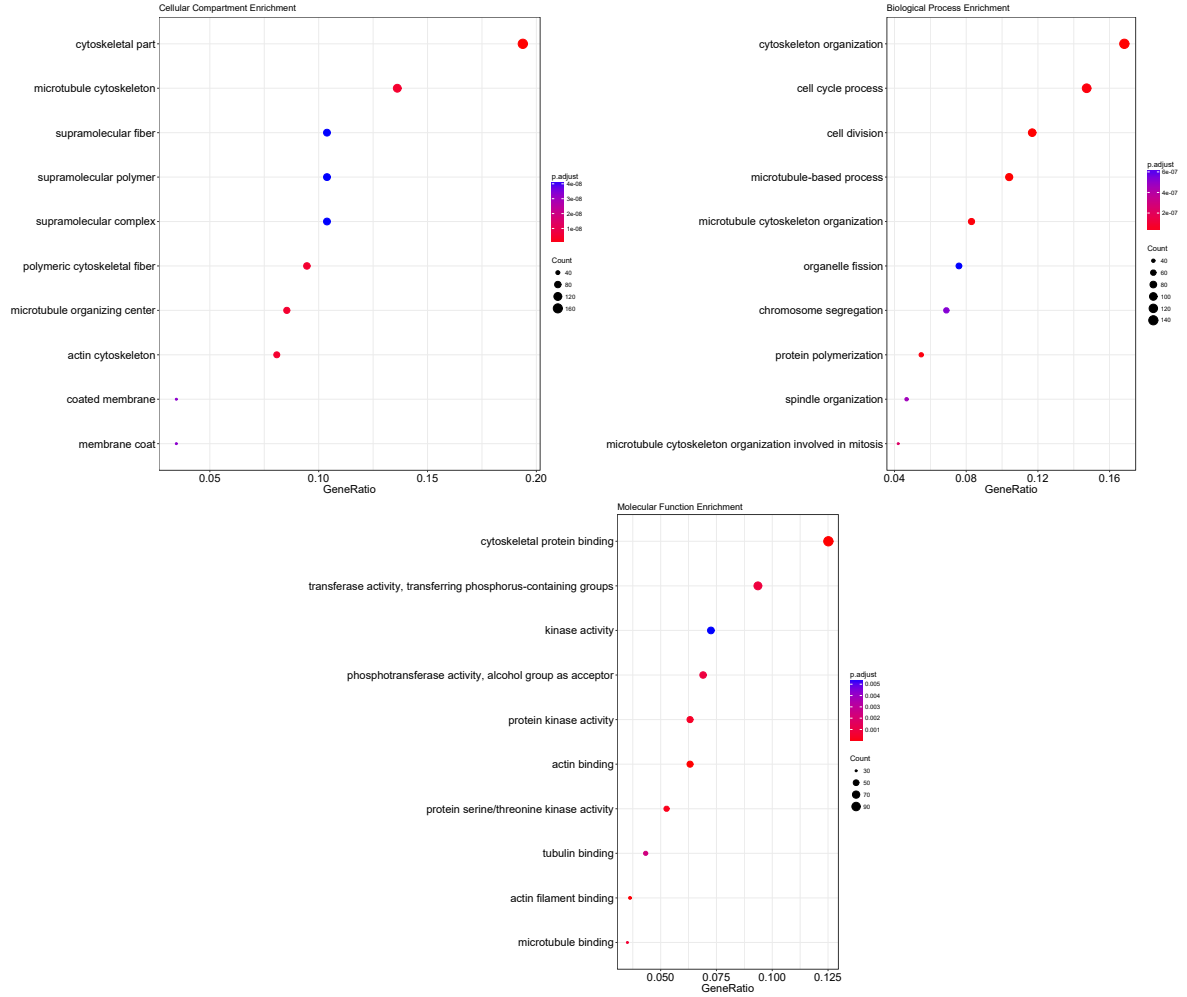


Fig. A.3 Gene Ontology over representation analysis on outlier proteins - that is proteins allocated with less than probability 0.95. We analyse the enrichment of terms in the cellular compartment, biological process, and molecular function ontologies. We display the top 10 significant results in the dotplots.

Appendix B

Appendix to chapter 6

B.1 Appendix 1: Matrix algorithms

We state here the require algorithm to invert the covariance matrix $C = \sigma^2 I_{nD} + J_n \otimes A$ for a Toeplitz matrix A. The algorithms are a minor modification of the algorithms found in [494] to handle the tensor product.

Algorithm 1 Tensor extended Trench algorithm

procedure TRENCH($C = \sigma^2 I_{nD} + J_n \otimes A$) ▷ C^{-1} and $\log\det C$, for Toeplitz A
  $Q \leftarrow I_D + \sigma^{-2} n A$
  $q \leftarrow Q^T_{1,:}$
  Input q to algorithm 2, returning $v \in \mathbb{R}^D$ and $l \in \mathbb{R}^D$
  $\bar{Q}(1, 1:D) \leftarrow v(D:1)$
  $\bar{Q}(1:D, 1) \leftarrow v(D:1)$
  $\bar{Q}(D, 1:D) \leftarrow v(1:D)$
  $\bar{Q}(1:D, D) \leftarrow v(1:D)$
  for $i = 2 : \lfloor (D-1)/2 \rfloor + 1$ **do**
    for $j = i : N - i + 1$ **do**
      $\bar{Q}(i,j) \leftarrow \bar{Q}(i-1, j-1) + \frac{v(D+1-j)v(D+1-i) - v(i-1)v(j-1)}{v(D)}$
      $\bar{Q}(j,i) \leftarrow \bar{Q}(i,j)$
      $\bar{Q}(N-i+1, N-j+1) \leftarrow \bar{Q}(i,j)$
      $\bar{Q}(N-j+1, N-i+1) \leftarrow \bar{Q}(i,j)$
    end for
  end for
  $Z \leftarrow \bar{Q}$
  $C^{-1} = \sigma^{-2} I_{nD} - \frac{1}{n\sigma^2} J_n^T \otimes (I - Z)$
  $\log\det C \leftarrow nD \log(\sigma^2) + l$
end procedure

Algorithm 2 Vector-Inverse and log-determinant algorithm

procedure VECTOR-INVERSE(q) ▷ v and l as required by algorithm 1
  $\xi \leftarrow \frac{q(2:D)}{q(1)}$
  Input $D-1$ and ξ to algorithm 3, returning $z \in \mathbb{R}^{D-1}$ and $l \in \mathbb{R}^D$
  $l \leftarrow l + D \log q(1)$
  $v(D) \leftarrow \frac{1}{(1+\xi^T z)q(1)}$
  $v(1:D-1) \leftarrow v(D)z(D-1:1)$
end procedure

Algorithm 3 extended Durbin's algorithm

procedure DURBIN(m, ξ) ▷ z and l as required by algorithm 1
  $z(1) \leftarrow -\xi(1)$
  $\beta \leftarrow \alpha \leftarrow 1$
  $l \leftarrow 0$
  for $i = 1 : m-1$ **do**
    $\beta \leftarrow (1-\alpha^2)\beta$
    $l = l + \log \beta$
    $\alpha \leftarrow \frac{\xi(i+1)+\xi(i:1)^T z(1:i)}{\beta}$
    $z(1:i) \leftarrow z(1:i) + \alpha z(i:1)$
    $z(i+1) \leftarrow \alpha$
  end for
  $\beta \leftarrow (1-\alpha^2)\beta$
  $l \leftarrow l + \log \beta$
end procedure

B.2 Appendix 2: Derivative of the marginal likelihood

The derivatives of the marginal likelihood given in the main text are given by [375]

$$\frac{\partial}{\partial\theta_j}\log\left\{p(Y_k|\tau,\boldsymbol{\theta_k})\right\} = \frac{1}{2}Y_k(\tau)^T\hat{C}_k^{-1}\left(\frac{\partial\hat{C}_k}{\partial\theta_j}\right)\hat{C}_k^{-1}Y_k(\tau) - \frac{1}{2}tr\left\{\hat{C}_k^{-1}\left(\frac{\partial\hat{C}_k}{\partial\theta_j}\right)\right\}. \tag{B.1}$$

The partial derivatives of the covariance functions can obtained in a straightforward manner and once evaluated at observations can be structured into blocks as was performed in the main text. Letting $\hat{A}_k$ be the diagonal blocks of the covariance matrix. The corresponding diagonal blocks of the derivative are given in equation B.2. Blocks not on the diagonal are similar and do not include the derivative with respect to θ_3.

$$\begin{aligned}\left[\frac{\partial\hat{A}_k}{\partial\theta_1}\right]_{rs} =& a\exp\left\{\left(-\frac{(x_r-x_s)^2}{e^{\theta_1}}\right)\right\}\left(\frac{(x_r-x_s)^2}{e^{\theta_1}}\right)\\ \left[\frac{\partial\hat{A}_k}{\partial\theta_2}\right]_{rs} =& 2e^{2\theta_2}\exp\left(-\frac{(x_r-x_s)^2}{l}\right)\\ \left[\frac{\partial\hat{A}_k}{\partial\theta_3}\right]_{rs} =& 2e^{2\theta_3}\delta_{rs}.\end{aligned} \tag{B.2}$$

B.3 Appendix 3: Tensor decompositions for derivatives of the marginal likelihood

In this appendix we derive formulae for the derivative of the marginal likelihood exploiting the block structure of our matrices. We first make some preliminary manipulations. We set the following notation $\partial_{\theta_j} = \frac{\partial}{\partial_{\theta_j}}$. First we note that

$$\hat{C}_k^{-1}(\partial_{\theta_j}\hat{C}_k)\hat{C}_k^{-1} = -\partial_{\theta_j}\hat{C}_k^{-1}. \tag{B.3}$$

We recall the following

$$\hat{C}_k^{-1} = \sigma^{-2}I_{nD} - \sigma^{-4}J_n\otimes(ZA), \tag{B.4}$$

and hence the following is true

$$\partial_{\theta_j}\hat{C}_k^{-1} = \partial_{\theta_j}(\sigma^{-2}I_{nD}) - \partial_{\theta_j}\left\{\sigma^{-4}J_n\otimes(ZA)\right\}. \tag{B.5}$$

We then note that $\partial_{\theta_j}J_n = 0$ and so the following algebraic manipulations hold

$$\begin{aligned}\partial_{\theta_j}\left\{J_n\otimes(ZA)\right\} =& \partial_{\theta_j}J_n\otimes(ZA) + J_n\otimes\partial_{\theta_j}(ZA)\\ =& J_n\otimes(\partial_{\theta_j}Z\cdot A + Z\cdot\partial_{\theta_j}A).\end{aligned} \tag{B.6}$$

We recall that

$$Z = (I_D + \sigma^{-2} nA)^{-1} = Q^{-1} \tag{B.7}$$

and so

$$\partial_{\theta_j} Z = -Q^{-1}(\partial_{\theta_j} Q)Q^{-1}. \tag{B.8}$$

It is obvious that

$$\partial_{\theta_j} Q = \partial_{\theta_j}(\sigma^{-2} nA), \tag{B.9}$$

and so

$$\partial_{\theta_j} Z = -Z\partial_{\theta_j}(\sigma^{-2} nA)Z. \tag{B.10}$$

Whence it follows that

$$\partial_{\theta_j} \hat{C}_k^{-1} = \partial_{\theta_j}(\sigma^{-2} I_{nD}) - \partial_{\theta_j}(\sigma^{-4}) J_n \otimes (ZA) - \sigma^{-4} \left\{ -Z\partial_{\theta_j}(\sigma^{-2} nA)ZA + Z\partial_{\theta_j} A \right\}. \tag{B.11}$$

Recall that

$$\begin{aligned} \partial_{\theta_1} A_{rs} &= A_{rs} S_{rs} \\ \partial_{\theta_2} A_{rs} &= 2A_{rs} \end{aligned} \tag{B.12}$$

where $S_{rs} = \frac{(t_r - t_s)^2}{l}$. We now derive formulae for the derivatives of the marginal likelihood in which we denote by $A \odot B$ the Hadamard (element-wise) product of matrices A and B.

Proposition 1. *The derivative of the marginal likelihood in equation B.1 with respect to θ_1 is given by*

$$\partial\theta_1 \log\{p(X|\tau, \boldsymbol{\theta})\} = \frac{1}{2} X(\tau)^T \sigma^{-4} J_n \otimes (ZASZ) X(\tau) - \frac{1}{2} tr\left(\hat{C}_k^{-1} \partial_{\theta_1} \hat{C}_k \right), \tag{B.13}$$

where

$$tr\left(\hat{C}_k^{-1} \partial_{\theta_1} \hat{C}_k \right) = \sigma^{-2} n \sum_i (AS)_{i,i} - \sigma^{-2} n \sum_{i,j} \{(I_D - Z) \odot (AS)\}_{ij}. \tag{B.14}$$

Proof. We note the following equalities, which follow from our preliminary manipulations

$$\begin{aligned} \partial_{\theta_1} \hat{C}_k^{-1} &= \sigma^{-4}(-Z(\sigma^{-2} n\partial_{\theta_1} A)ZA + Z\partial_{\theta_1} A) \\ &= -\sigma^{-4} J_n \otimes \left\{ (Z\partial_{\theta_1} A)(-\sigma^{-2} nZA + I_D) \right\} \\ &= -\sigma^{-4} J_n \otimes \{Z(\partial_{\theta_1} A)Z\} \\ &= -\sigma^{-4} J_n \otimes (ZASZ), \end{aligned} \tag{B.15}$$

where the third line follows from the second because

$$\begin{aligned} & Q = I_D + \sigma^{-2} n A \\ \implies \quad & Q - \sigma^{-2} n A = I_D \\ \implies \quad & Q^{-1}Q - \sigma^{-2} n Q^{-1} A = Q^{-1} \\ \implies \quad & I_D - \sigma^{-2} n Q^{-1} A = Q^{-1} \end{aligned} \tag{B.16}$$

For the trace term, recall that the trace of a product of two matrices is the sum of the Hadamard product of those two matrices. That is

$$tr\left(\hat{C}_k^{-1}\partial_{\theta_j}\hat{C}_k\right) = \sum_{i,j}\left(\hat{C}_k^{-1} \odot \partial_{\theta_j}\hat{C}_k\right)_{i,j}. \tag{B.17}$$

Applying the mixed product property, we see that the following equalities hold

$$\begin{aligned} \hat{C}_k^{-1} \odot \partial_{\theta_1}\hat{C}_k &= \left\{\sigma^{-2} I_{nD} - \sigma^{-4} J_n \otimes (ZA)\right\} \odot \{J_n \otimes (AS)\} \\ &= \sigma^{-2} I_{nD} \odot \{J_n \otimes (AS)\} - \sigma^{-4}\{J_n \otimes (ZA)\} \odot \{J_n \otimes (AS)\} \\ &= \sigma^{-2} I_{nD} diag(AS, AS, \ldots, AS) - \sigma^{-4}\left[J_n \otimes \{(ZA) \odot (AS)\}\right]. \end{aligned} \tag{B.18}$$

Hence,

$$tr\left(\hat{C}_k^{-1}\partial_{\theta_1}\hat{C}_k\right) = \sigma^{-2} n \sum_i (AS)_{i,i} - \sigma^{-4} n^2 \sum_{i,j}\{(ZA) \odot (AS)\}_{ij}. \tag{B.19}$$

Thus the derivative of the log marginal likelihood is

$$\partial\theta_1 \log\{p(X|\tau, \boldsymbol{\theta})\} = \frac{1}{2} X(\tau)^T \sigma^{-4} J_n \otimes (ZASZ) X(\tau) - \frac{1}{2} tr\left(\hat{C}_k^{-1}\partial_{\theta_1}\hat{C}_k\right) \tag{B.20}$$

Then we can substitute $ZA = (I - Z)\frac{\sigma^2}{n}$ to obtain the required result. □

Proposition 2. *The derivative of the marginal likelihood in equation B.1 with respect to θ_2 is given by*

$$\partial\theta_2 \log\{p(X|\tau, \boldsymbol{\theta})\} = \frac{1}{2} X(\tau)^T \sigma^{-4} J_n \otimes (2ZAZ) X(\tau) - \frac{1}{2} tr\left(\hat{C}_k^{-1}\partial_{\theta_2}\hat{C}_k\right) \tag{B.21}$$

where

$$tr\left(\hat{C}_k^{-1}\partial_{\theta_2}\hat{C}_k\right) = 2\sigma^{-2} n \sum_i (A)_{i,i} - \sigma^{-2} n \sum_{i,j}\{(I - Z) \odot (2A)\}_{ij}. \tag{B.22}$$

Proof. As in the previous proposition we observe:

$$\begin{aligned}\partial_{\theta_2}\hat{C}_k^{-1} &= \sigma^{-4}(-Z(\sigma^{-2}n\partial_{\theta_1}A)ZA + Z\partial_{\theta_2}A)\\ &= -\sigma^{-4}J_n \otimes \left\{(Z\partial_{\theta_2}A)(-\sigma^{-2}nZA + I)\right\}\\ &= -\sigma^{-4}J_n \otimes \{Z(\partial_{\theta_2}A)Z\}\\ &= -\sigma^{-4}J_n \otimes (2ZAZ).\end{aligned} \tag{B.23}$$

For the trace term, as for θ_1 we proceed as follows

$$\begin{aligned}\hat{C}_k^{-1} \odot \partial_{\theta_2}\hat{C}_k &= \left\{\sigma^{-2}I_{nD} - \sigma^{-4}J_n \otimes (ZA)\right\} \odot \{J_n \otimes (2A)\}\\ &= \sigma^{-2}I_{nD} \odot \{J_n \otimes (2A)\} - \sigma^{-4}\{J_n \otimes (ZA)\} \odot \{J_n \otimes (2A)\}\\ &= 2\sigma^{-2}I_{nD}diag(A, A, \ldots, A) - \sigma^{-4}\left[J_n \otimes \{(ZA) \odot (2A)\}\right].\end{aligned} \tag{B.24}$$

Hence,

$$tr\left(\hat{C}_k^{-1}\partial_{\theta_2}\hat{C}_k\right) = 2\sigma^{-2}n\sum_i (A)_{i,i} - \sigma^{-4}n^2\sum_{i,j}\{(ZA) \odot (2A)\}_{ij}. \tag{B.25}$$

Thus the derivative of the log marginal likelihood is

$$\partial\theta_2 \log\{p(X|\tau, \boldsymbol{\theta})\} = \frac{1}{2}X(\tau)^T\sigma^{-4}J_n \otimes (2ZAZ)X(\tau) - \frac{1}{2}tr\left(\hat{C}_k^{-1}\partial_{\theta_2}\hat{C}_k\right) \tag{B.26}$$

Then we can substitute $ZA = (I - Z)\frac{\sigma^2}{n}$ to obtain the required result. □

Proposition 3. *The derivative of the marginal likelihood in equation B.1 with respect to θ_1 is given by*

$$\partial\theta_3 \log\{p(X|\tau, \boldsymbol{\theta})\} = \sigma^{-2}\|X(\tau)\|_2^2 + X(\tau)^T J_n \otimes \left\{\frac{(Z^2 - I)}{\sigma^2 n}\right\}X(\tau) - \frac{1}{2}tr\left(\hat{C}_k^{-1}\partial_{\theta_3}\hat{C}_k\right), \tag{B.27}$$

where

$$tr(\hat{C}_k^{-1}\partial_{\theta_3}\hat{C}_k) = 2nD - 2\sum_i (I - Z)_{ii}. \tag{B.28}$$

Proof. We note that $\partial_{\theta_3}\hat{C}_k = 2\sigma^2 I_{nD}$ is a scalar multiple of the identity matrix and thus commutes. Hence, we need only compute $\hat{C}_k^{-1}\hat{C}_k^{-1}$ and the trace term. Note the following

algebraic manipulations:

$$\begin{aligned}
\hat{C}_k^{-1}\hat{C}_k^{-1} &= \left\{\sigma^{-2}I_{nD} - \sigma^{-4}J_n \otimes (ZA)\right\}\left\{\sigma^{-2}I_{nD} - \sigma^{-4}J_n \otimes (ZA)\right\} \\
&= \sigma^{-4}I_{nD} - 2\sigma^{-6}J_n \otimes (ZA) + \sigma^{-8}(J_nJ_n) \otimes (ZAZA) \\
&= \sigma^{-4}I_{nD} - 2\sigma^{-6}J_n \otimes (ZA) + n\sigma^{-8}(J_n) \otimes (ZAZA) \\
&= \sigma^{-4}I_{nD} + J_n \otimes (-2\sigma^{-6}ZA + n\sigma^{-8}ZAZA) \\
&= \sigma^{-4}I_{nD} + J_n \otimes \left\{\sigma^{-6}(-2I_D + n\sigma^{-2}ZA)ZA\right\} \\
&= \sigma^{-4}I_{nD} + J_n \otimes \left\{\sigma^{-6}(-2I_D + I_D - Z)ZA\right\} \\
&= \sigma^{-4}I_{nD} + J_n \otimes \left\{-\sigma^{-6}(I_D + Z)ZA\right\} \\
&= \sigma^{-4}I_{nD} + J_n \otimes \left\{-\sigma^{-4}(I_D - Z^2)/n\right\}.
\end{aligned} \tag{B.29}$$

To compute the trace we note that the following follows directly from the tensor decomposition of $\hat{C}_k^{-1}$:

$$tr(\hat{C}_k^{-1}) = nD\sigma^{-2} - \sigma^{-4}n\sum_i (ZA)_{ii} = nD\sigma^{-2} - \sigma^{-2}\sum_i (I - Z)_{ii}. \tag{B.30}$$

Substituting the formulae provides the desired result. □

In practice, we never need to compute or even store the full $nD \times nD$ inverse matrix C^{-1}, since we can only need to keep track of summaries of the data matrix rather than the full data matrix itself. This is demonstrated in the following proposition.

Proposition 4. *Let*

$$X = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1n} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{D1} & x_{D2} & x_{D3} & \dots & x_{Dn} \end{bmatrix},$$

be a $D \times n$ matrix. Let $Y_i = \sum_j X_{i,j}$ be the sum of the i^{th} row of X and written concisely $Y = Xe_n$, where e_n is a $n \times 1$ vector of ones. We write J_n to be the $n \times n$ matrix of ones. Let R be any $D \times D$ matrix. Then the following holds

$$vec(X)^T(J_n \otimes R)vec(X) = YRY, \tag{B.31}$$

where $vec(X)$ denotes the vectorisation of X; that is, the $Dn \times 1$ vector formed by stacking columns of X.

Proof. Firstly, observe the following standard algebraic manipulations

$$\begin{aligned}(J_n \otimes R)vec(X) &= vec(RXJ_n) \\ &= vec(RXe_ne_n^T) \\ &= vec(RYe_n^T) \\ &= (e_n \otimes R)vec(Y) \\ &= (e_n \otimes R)Y.\end{aligned} \tag{B.32}$$

Thus, using the above, it follows that

$$\begin{aligned}vec(X)^T(J_n \otimes R)vec(X) &= vec(X)^T(e_n \otimes R)Y \\ &= vec(R^TXe_n)^TY \\ &= vec(R^TY)^TY \\ &= (R^TY)^TY \\ &= Y^TRY,\end{aligned} \tag{B.33}$$

as required. □

B.4 Appendix 4: Further sensitivity analysis

In this appendix, we assess the effects of prior choices on our applied analysis. Our investigation is two-fold: we establish how the posterior distributions change under different prior choices, as well as how this impacts the partitioning of the data. The partitioning of the data is produced by assigning proteins to their most probable organelle. To visualise the uncertainty in the potential partitions of the data, we use the posterior similarity matrix [142]. The posterior similarity matrix is the matrix S such that entry (i, j) is given by

$$S_{ij} \approx \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}(z_i^{(t)} = z_j^{(t)}). \tag{B.34}$$

In words, this is the proportion of times that protein i and protein j are allocated to the same component, during the MCMC algorithm. The similarity of two partitions is calculated using the adjusted Rand index (ARI) [213], which is 0 under partitions from a random model. The ARI is 1 for identical partitions and can be negative if two partitions are less similar than one would be expected by random.

We begin by assessing the sensitivity with respect to the prior on the outlier proportions, which is given a beta prior: $\phi_i \sim B(u, v)$. The default choices for this parameters are $(u, v) = (2, 10)$,

we assess how $(u, v) = (2, 4)$ and $(2, 2)$ impacts our results. All other prior choices are held at their defaults. First, we plot the inferred posterior distribution of these parameters and we also include the samples from the prior for reference (figure B.1). We observe posterior shrinkage and it is clear that the priors are strongly informed by the data. The similarity of the posterior distributions reflects the insensitivity of the results to the prior choice.

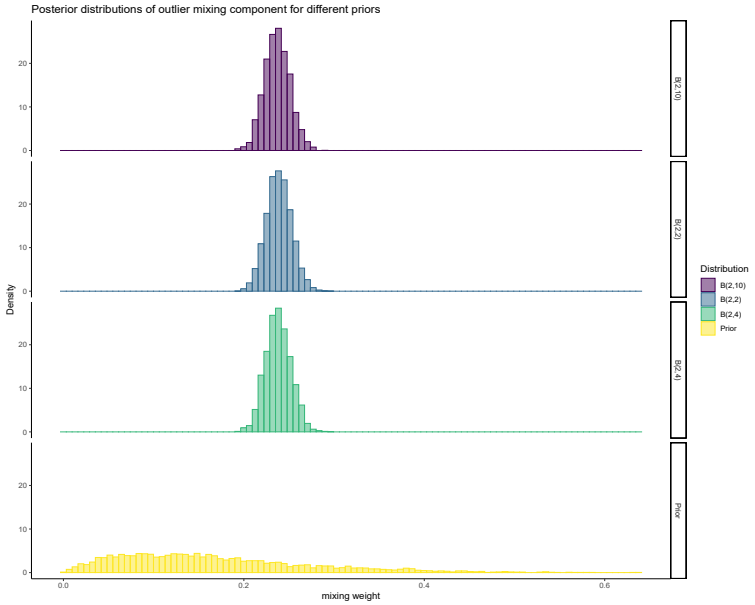
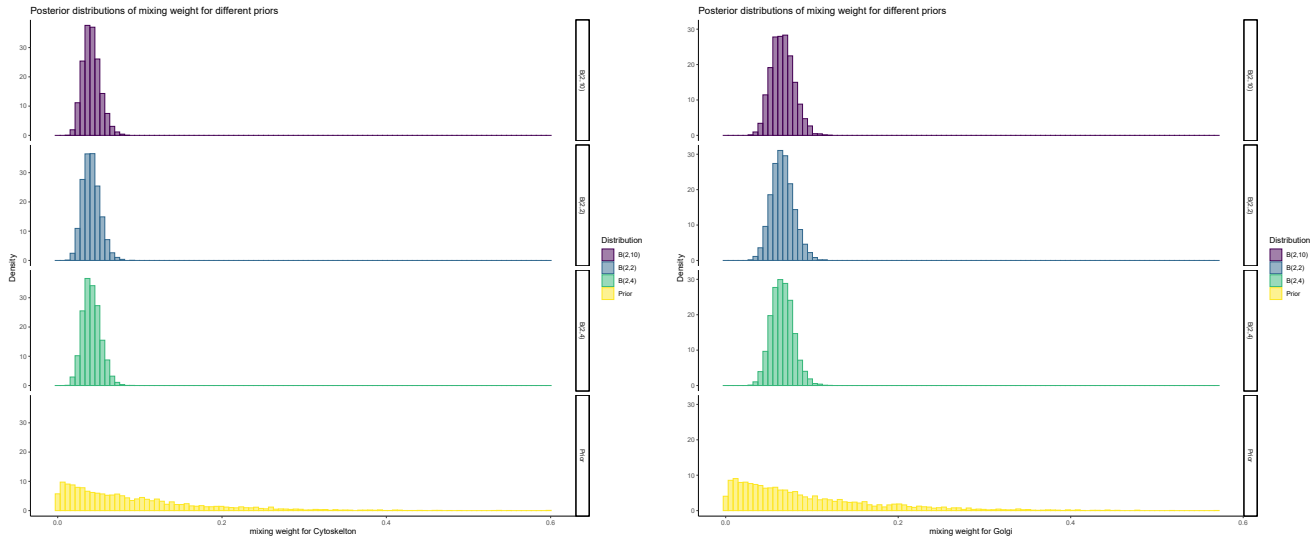


Fig. B.1 The first three facets display histograms of the posterior distribution for the mixture weight of the outlier component under different prior choices. The prior choices are indicated on the right hand panel, as well as in the legend. We note that the posterior distribution are similar across these different choices. The lowest facet is a histogram of samples from the default prior distribution and is shown for reference.

Furthermore, we assess whether component mixture proportions are affected by the choice of prior. Again, we observe insensitivity to the prior choice. Two examples are plotted in figure B.2.

Now, that we have established that the posterior distributions are insensitive to the choice of prior, we analyse whether different prior choices generate substantially different partitions of the data. We visualise these partitions and associated uncertainty in the following PSMs (figure B.3). It is clear that the PSMs are similar across the different prior choice. For further quantitative analysis, we compute the adjusted Rand Index across the different partitions. The partitions are near identical across the different choices (table B.1).

We now turn to the (hyper)-prior on the GP hyperparameters. The default choice for the amplitude a^2, length-scale l and noise σ^2 are log-normal priors with zero mean and standard deviation 1. Holding the mean constant at zero, we assess how the posterior distributions and data partitions are impacted on setting the standard deviation to 0.5 and 0.1 (whilst other



(a) (b)

Fig. B.2 (a,b) In each plot the first three facets display histograms of the posterior distribution for mixture weights for (a) the Cytoskelton component and (b) the Golgi component. The prior choice are indicated in the right hand panel, as well as the legend. The posterior distributions are similar across the different choices. The lowest facet in both figures is a histogram of samples from the default prior (a marginal Beta distribution)

Table B.1 adjusted Rand index for partitions generated from different prior choices

Prior	B(2,10)	B(2,4)	B(2,2)
B(2,10)	1	0.983	0.983
B(2,4)	-	1	1
B(2,2)	-	-	1

choices are held at defaults). We plot representative and illustrative examples of posterior distributions for the hyperparameters and the default prior is plotted for reference (figure B.4). We observe posterior shrinkage in all situations; however, to differing degrees. σ^2 is the most strongly informed parameter by the data. Furthermore, smaller prior standard deviations causes concentration of the posterior around the prior locations. In these scenarios, we may conclude there is insufficient data to inform such a strong prior assertion on these hyperparameters. Thus, our diffuse default choice appears a good choice in practice. If we wish to specify strongly informative priors, we could estimate the mean parameter of the hyperprior by examining the fitted values from maximum marginal likelihood estimation (Type II ML) for the labelled data only.

Again, we explore how these differing prior settings induce different posterior partitions of the data. The posterior similarity matrices are robust to alteration of the prior in most of the

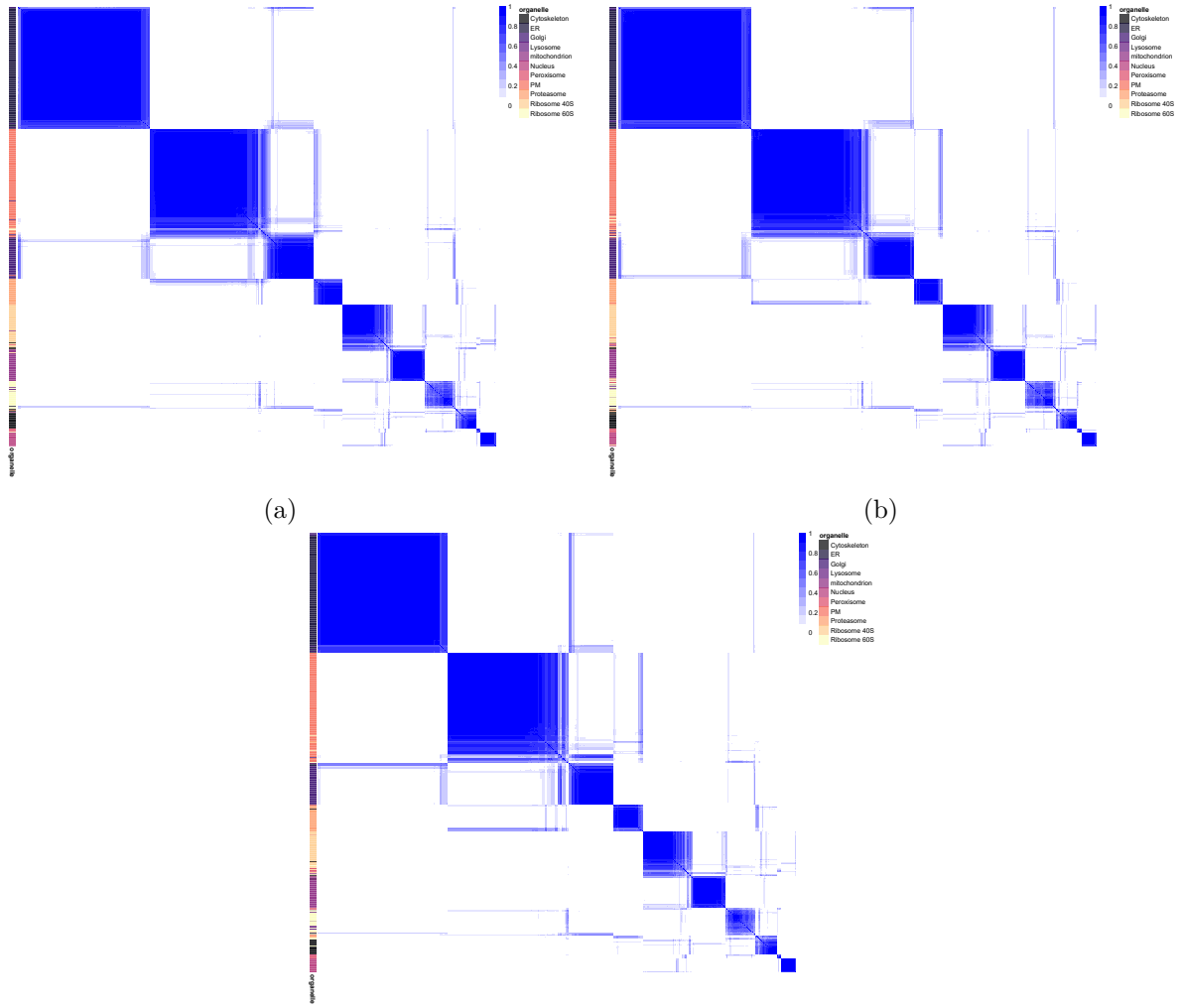


Fig. B.3 Posterior similarity matrices generated from different choices of prior for the outlier component weight. The colour bar on the left indicates the assigned organelle and we note that the PSMs are similar across each of the choices. (a) $B(2, 10)$, (b) $B(2, 4)$, (c) $B(2, 2)$.

scenarios, showing only slight differences (figures B.5, B.6 and B.7). This is corroborated by extremely similar ARIs across the different settings (tables B.2, B.3 and B.4). However, when we specify the following prior for the noise: $\sigma^2 \sim \mathcal{LN}(0, 0.1)$, we no longer observe reasonable results. This is a manifestation of a strong prior on an inappropriate configuration of the model. This is not a limitation of the analysis, because it demonstrates that if expert opinion is available it can be encoded, with effect, into the data. However, we wishes to demonstrate that without care to the choice of the hyperprior, one should expect poor results. We note that such sensitivity is not observed for the other hyperparameters.

Finally, we perform sensitivity analysis for the prior for the mixing proportions. The default choice is the symmetric uniform prior $\boldsymbol{\pi} \sim Dir(1)$ and we test sensitivity with respect to the

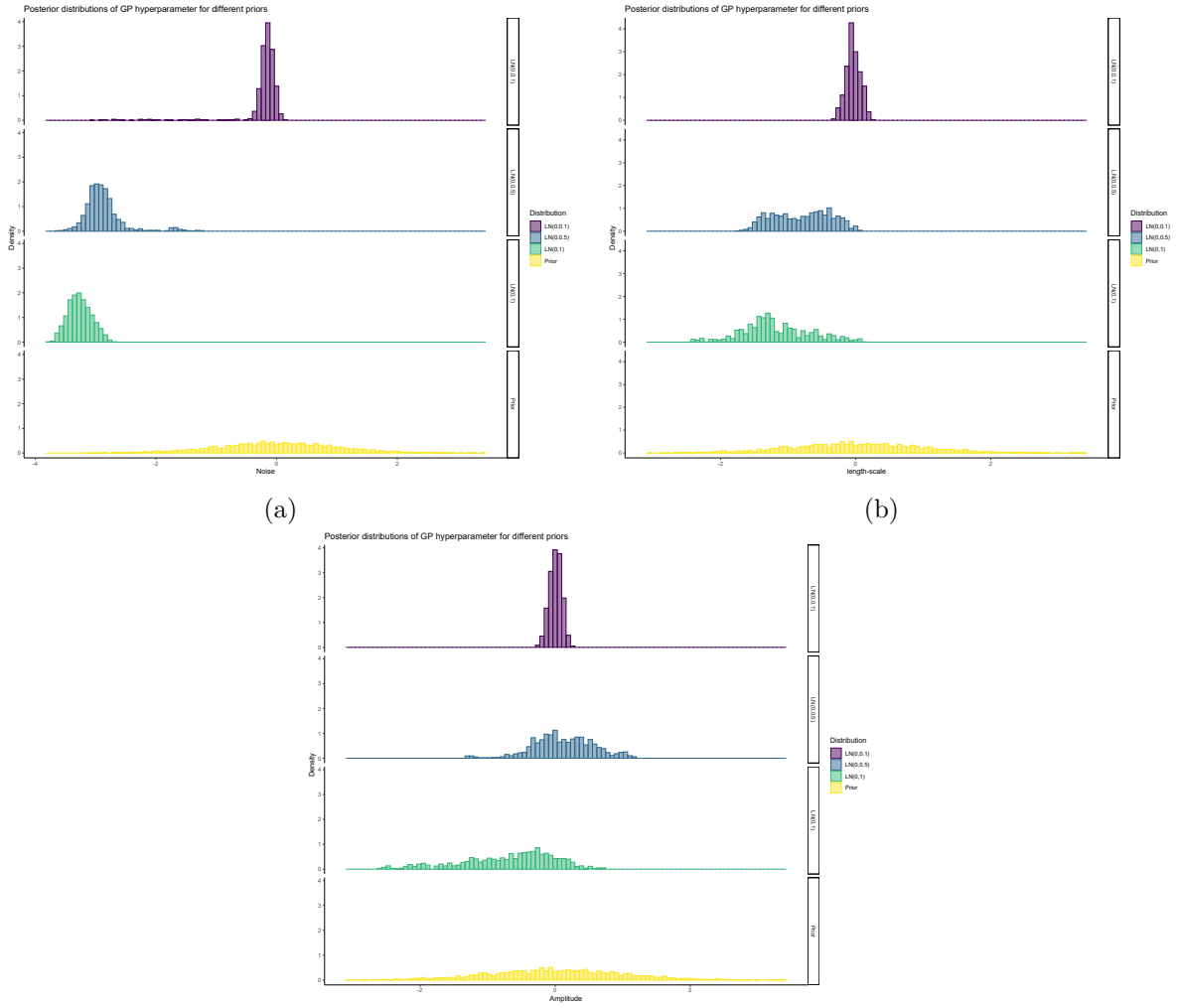


Fig. B.4 The first three histograms in each figure are posterior distribution for the GP hyperparameters, (a) noise, (b) length-scale, (c) amplitude, for different choices of hyperprior. These choices are shown in the right hand bar, as well as in the legend. For reference, the lowest facet in each figure displays a histogram of samples from the prior distribution. The distributions are plotted on the log scale to aid visualisation.

Table B.2 adjusted Rand index for partitions generated from different prior choices on the amplitude

Prior	LN(0, 1)	LN(0, 0.5)	LN(0, 0.1)
LN(0, 1)	1	0.970	0.986
LN(0, 0.5)	-	1	0.976
LN(0, 0.1)	-	-	1

Jeffrey's prior $\boldsymbol{\pi} \sim Dir(0.5)$ and $\boldsymbol{\pi} \sim Dir(0.1)$. All other priors are set to default. We plot

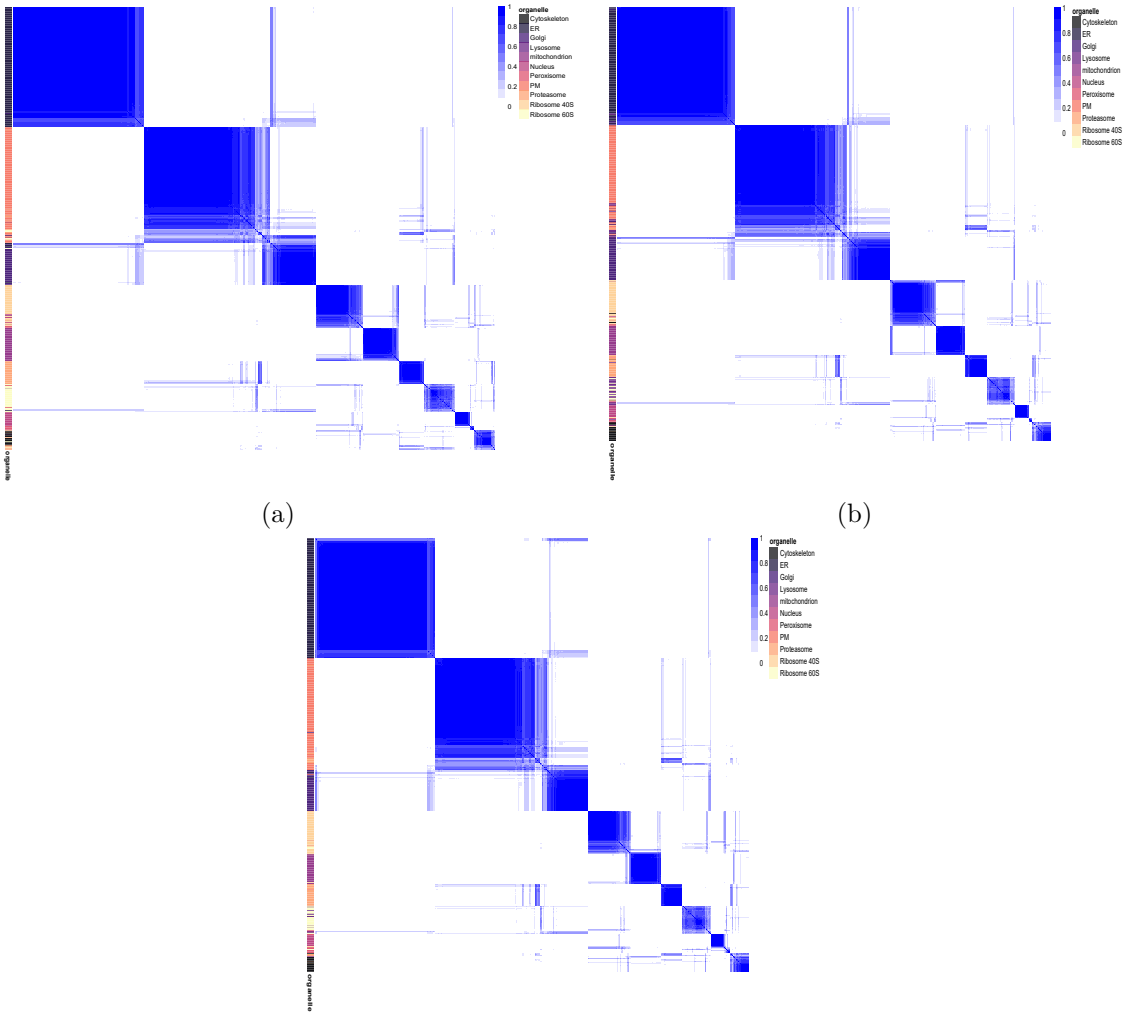


Fig. B.5 Posterior similarity matrices generated from different choices of hyperprior for the amplitude hyperparameter. The colour bar on the left indicates the assigned organelle and we note that the PSMs are similar across each of the choices. (a) $\mathcal{LN}(0,1)$, (b) $\mathcal{LN}(0,0.5)$, (c) $\mathcal{LN}(0,0.1)$.

Table B.3 adjusted Rand index for partitions generated from different prior choices on the length-scale

Prior	LN(0, 1)	LN(0, 0.5)	LN(0, 0.1)
LN(0, 1)	1	0.969	0.965
LN(0, 0.5)	-	1	0.990
LN(0, 0.1)	-	-	1

two illustrative examples in figure B.8, showing insensitivity of our analysis to the prior choice. As with previous example, we explore the how the different prior induce different posterior

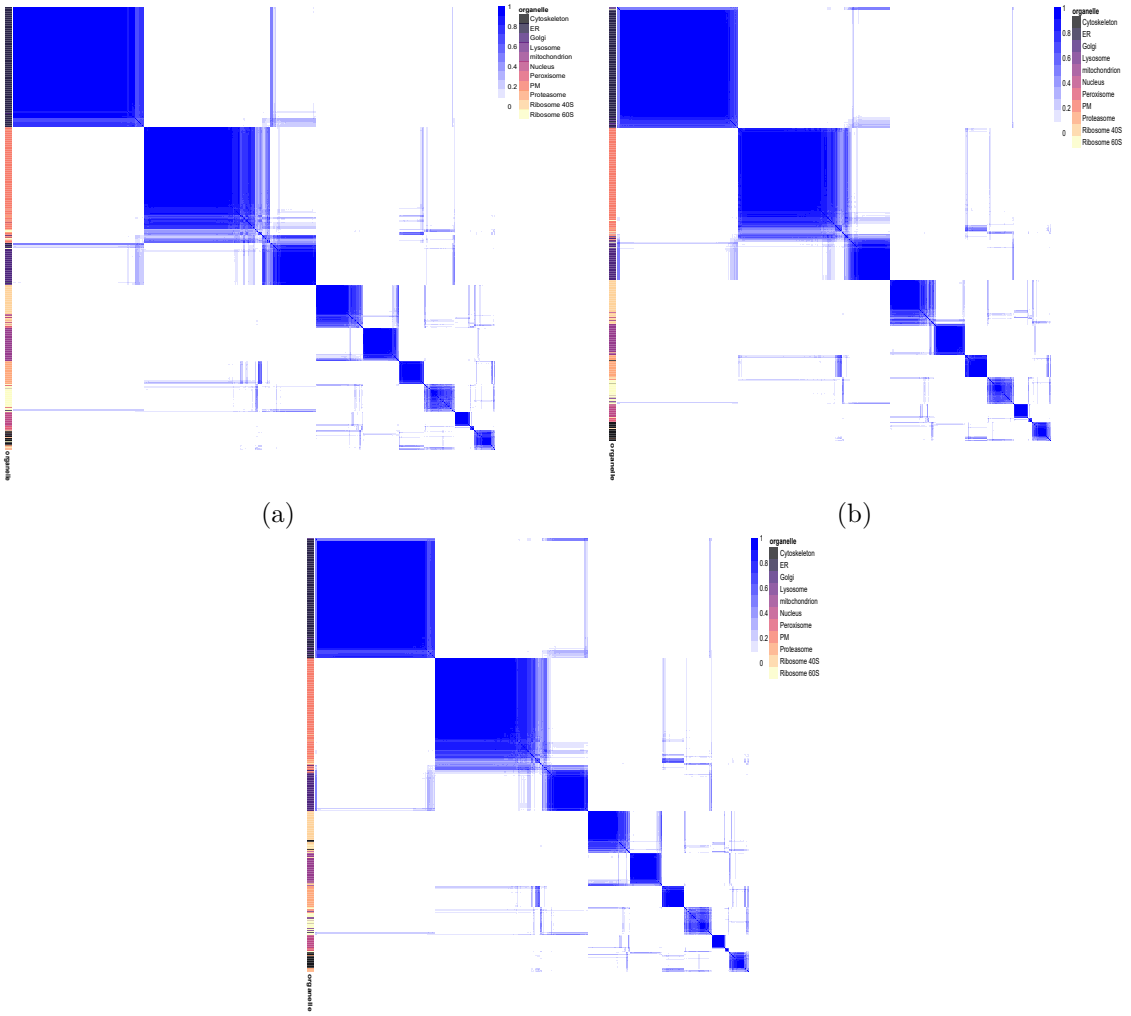


Fig. B.6 Posterior similarity matrices generated from different choices of hyperprior for the length-scale hyperparameter. The colour bar on the left indicates the assigned organelle and we note that the PSMs are similar across each of the choices. (a) $\mathcal{LN}(0,1)$, (b) $\mathcal{LN}(0,0.5)$, (c) $\mathcal{LN}(0,0.1)$.

Table B.4 adjusted Rand index for partitions generated from different prior choices on the noise

Prior	LN(0, 1)	LN(0, 0.5)	LN(0, 0.1)
LN(0, 1)	1	0.966	0.196
LN(0, 0.5)	-	1	0.200
LN(0, 0.1)	-	-	1

partitions of the data and accordingly visualise the PSMs, as well reporting the ARIs across these partitions (figure B.9 and table B.5). The reported PSMs are very similar and the ARIs are all close to 1.

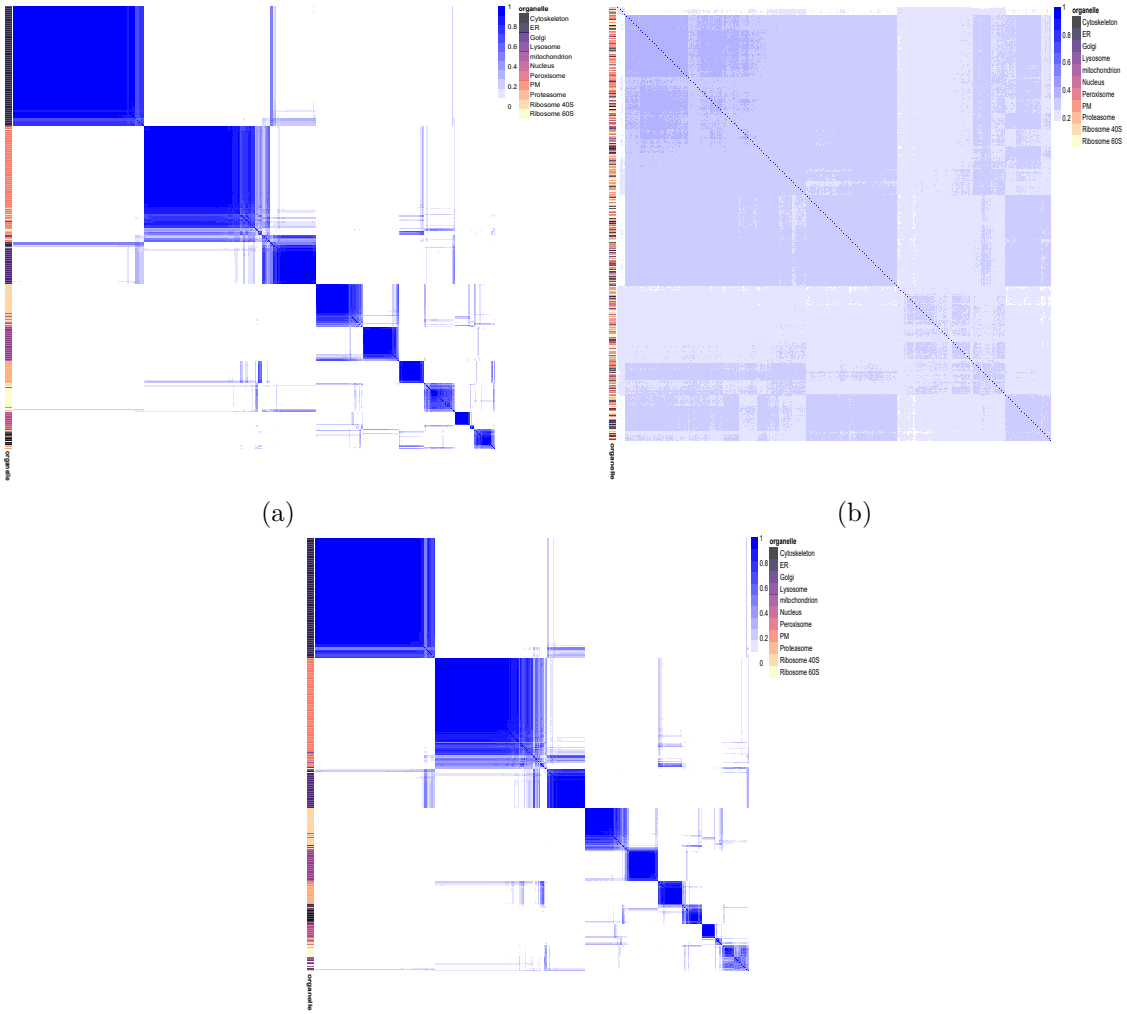
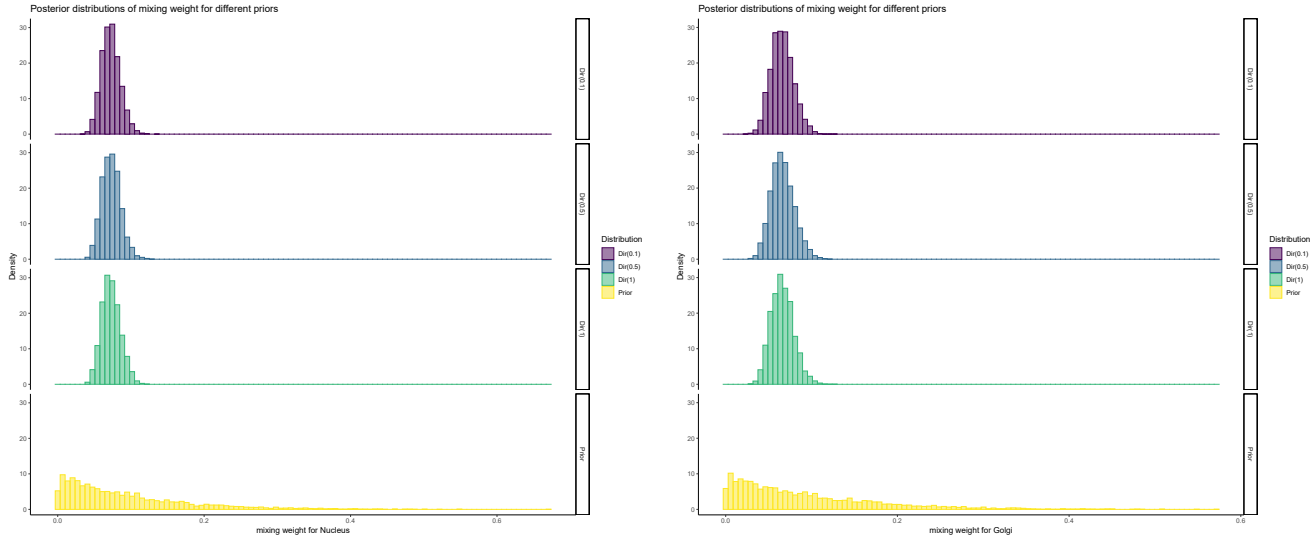


Fig. B.7 Posterior similarity matrices generated from different choices of hyperprior for the noise hyperparameter. The colour bar on the left indicates the assigned organelle and we note that the PSMs are similar across each of the choices. (a) $\mathcal{LN}(0,1)$, (b) $\mathcal{LN}(0,0.5)$, (c) $\mathcal{LN}(0,0.1)$.

Table B.5 adjusted Rand Index for partitions generated from different prior choices on the mixing proportions

Prior	Dir(1)	Dir(0.5)	Dir(0.1)
Dir(1)	1	0.984	0.986
Dir(0.5)	-	1	0.999
Dir(0.1)	-	-	1



(a) (b)

Fig. B.8 In both figures the first three facets are histograms of the posterior distribution for the mixing weight of (a) the Nucleus component (b) the Golgi component as representative examples. Each facet is for a different choice of prior distribution stated in the right hand bar and also the legend. The lowest facet in both figures is a histogram of samples from the prior distribution as a reference.

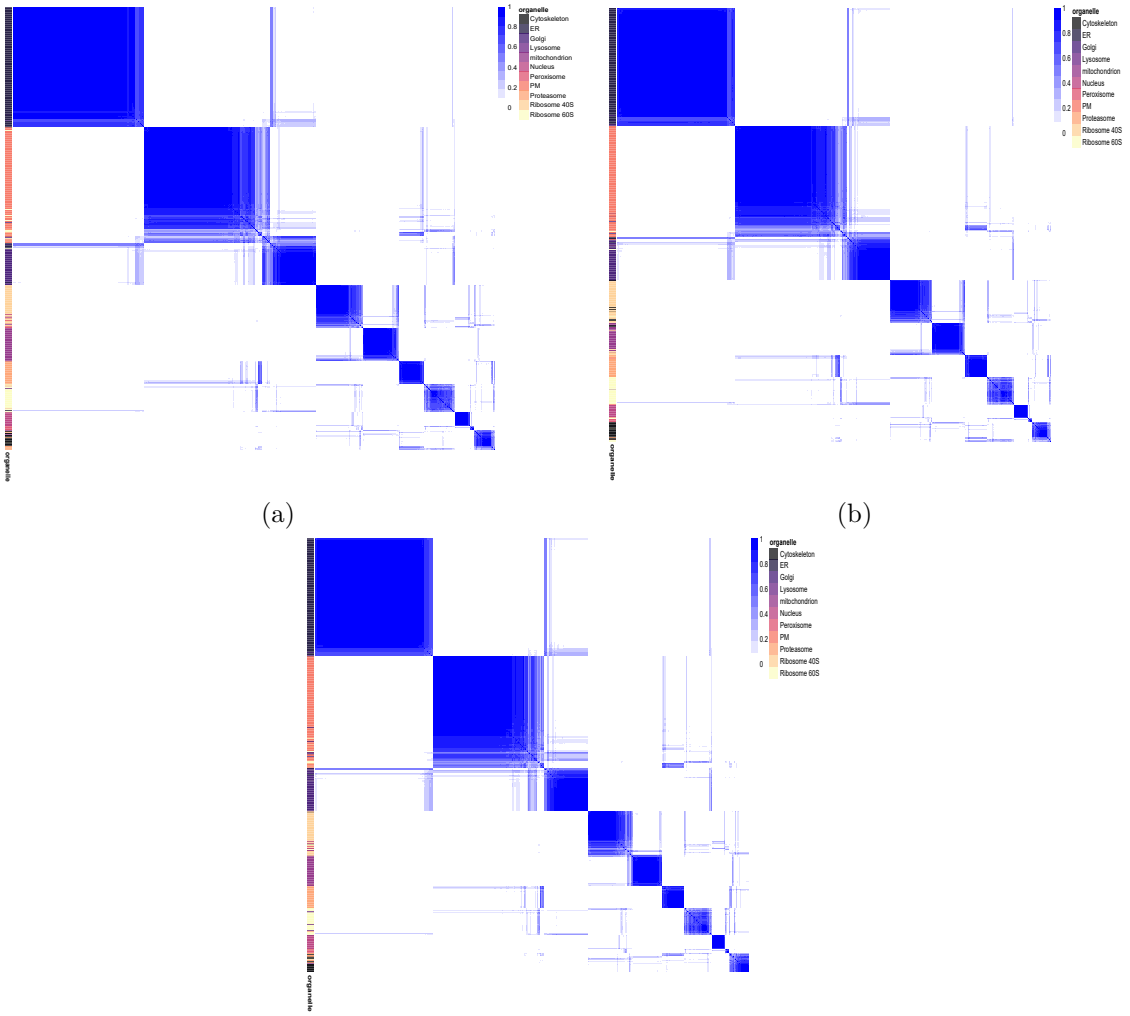


Fig. B.9 Posterior similarity matrices generated from different choices of prior for the mixture proportions. The colour bar on the left indicates the assigned organelle and we note that the PSMs are similar across each of the choices. (a) Dir (1), (b) Dir (0.5), (c) Dir (0.1).

B.5 Appendix 5: Simulation study

We perform a simulation study to assess the robustness of our model to different settings of the priors, assumptions about the distribution of the unlabelled data, and misspecification of the covariance function and of the outlier distribution. To simulate new spatial proteomics data, we simulate from the posterior predictive distribution using only the labelled data for model fitting. The hyperparameters are found by optimising the marginal likelihood using L-BFGS. To visualise the variability that these simulated datasets capture, we take each dataset in turn and compute the mean location of each organelle (as the mean of all proteins arising from that organelle). We then align each dataset onto the same PCA coordinates and visualise the projected means of the organelles with an overlay of density contours (see figure B.10). In each of the simulation scenarios we consider, we simulate 10 datasets and report the distribution of scores across these simulated datasets.

Given that we know the class from which protein comes from, we focus on analysing predictive performance (as measured using the quadratic loss) and whether proteins were correctly or incorrectly identified as outliers. Our first set of simulations assess the effect of predictive performance on changing the prior hyperparameters, whilst keeping all other values are their defaults. The following scenarios are assessed:

- Default Settings
- $B(2, 4)$ prior on the outlier component
- $B(2, 2)$ prior on the outlier component
- $\mathcal{LN}(0, 0.5)$ prior on the Amplitude
- $\mathcal{LN}(0, 0.5)$ prior on the length-scale
- $\mathcal{LN}(0, 0.5)$ prior on the Noise

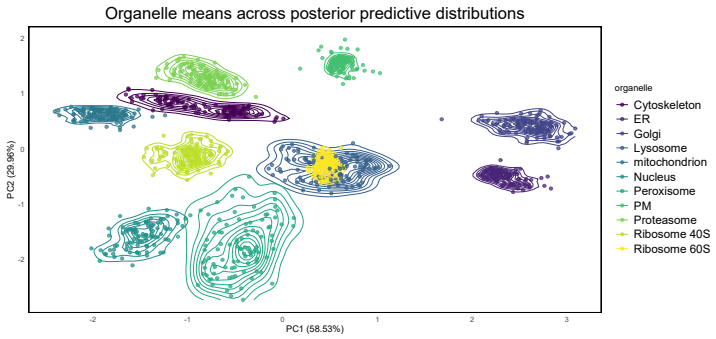
Figure B.11 demonstrate that there is no considerable effect on changing the prior to the predictive performance of our method. Next, we explore the effect of changing the value of the noise parameter when simulating new data. More precisely, once σ_k^2 for $k = 1, ..., K$ has been computed for the labelled data only, we simulate unlabelled data according to the following scenarios:

- The variance of the unlabelled data is twice that of the labelled data
- The variance of the unlabelled data is five times that of the labelled data
- The variance of the unlabelled data is ten times that of the labelled data

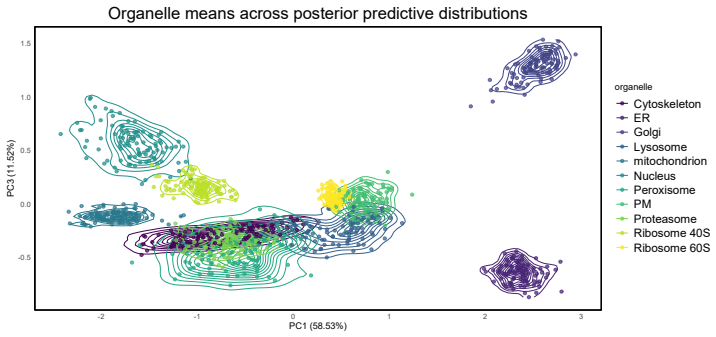
We note that the observed variance of the unlabelled data is always less than twice that of the labelled data in case studies and thus the simulations are extreme scenarios. Unsurprisingly, as the variance of the unlabelled data departs from that of the labelled data the predictive performance decreases (see figure B.11). We now turn to the covariance function and the effect of misspecification by using the Matérn covariance. The marginal variance and range parameters are found by optimisation with respect to the marginal likelihood. Our original model with squared exponential covariance is then fitted to the model and predictive performance is assessed. We assess the following scenarios:

- Matérn covariance with $\nu = 2$
- Matérn covariance with $\nu = 3$.

Figure B.11 demonstrates that there is little sensitivity to misspecification with slightly improved performance when simulations were draw from a smoother Matérn covariance. Finally, we consider a simulation scenario where we consider a different distribution for the outlier component. In robust mixture modelling it is typical to use spatial poisson process to model outliers [19, 140]. Thus, we simulate outlier data according to a spatial poisson process on the D-disk (disk of dimension D) with radius $r = 2 \times \max(|X|)$. The predictive performance in this scenario is comparable to that of our simulation settings (see figure B.11). Though for this example of misspecification of the outlier distribution it is more interesting to examine whether outliers are correctly identified. In table B.6 we report the proportion of proteins that are incorrect identified as outliers (amongst those that are not outliers) and those that are correctly identified as outliers. We also report the 95% confidence interval for these results. We also report these results for the other simulation settings and note that results are stable across the different scenarios.

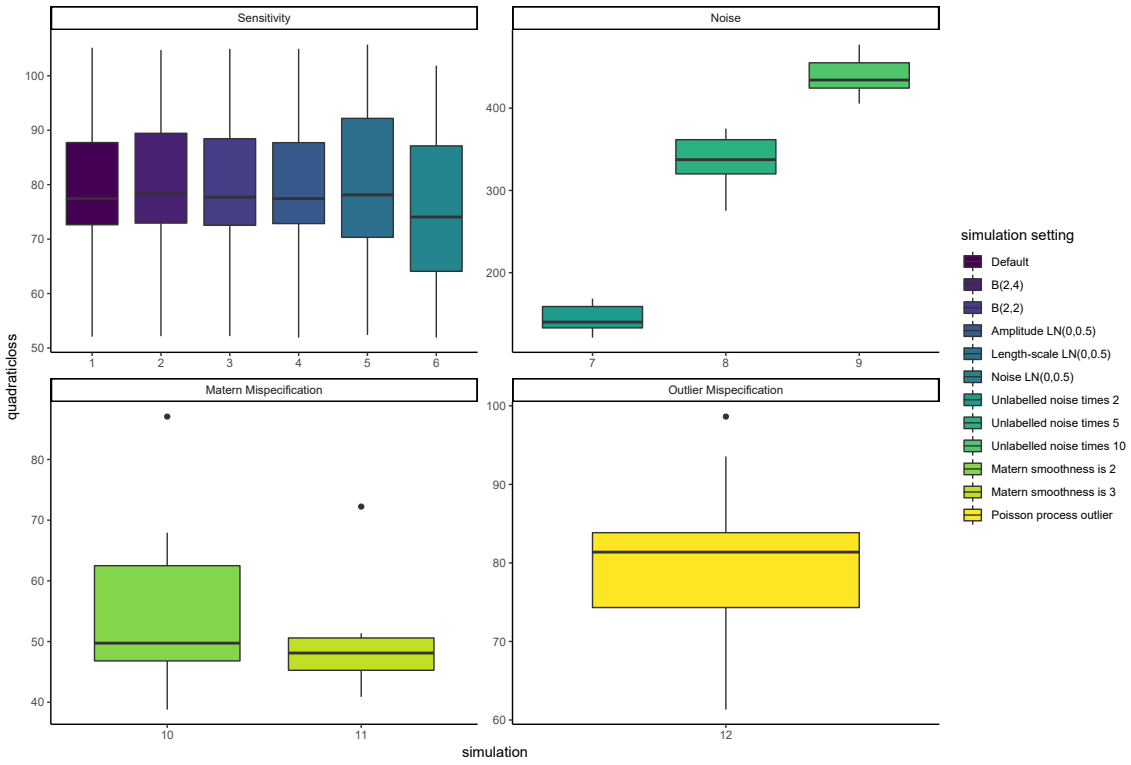


(a)



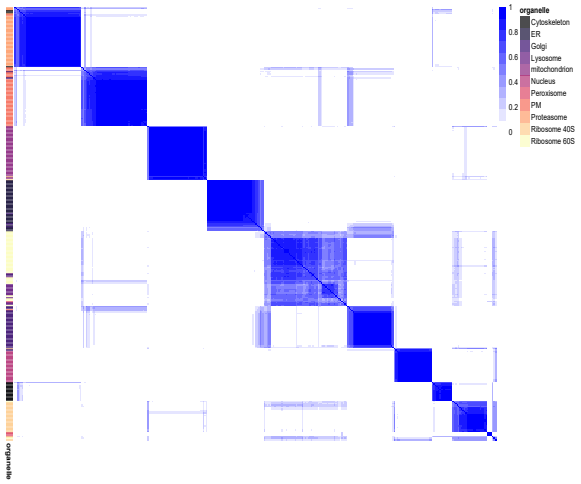
(b)

Fig. B.10 PCA projection of organelle means generated from posterior predictive distributions. We simulate 100 datasets from the posterior predictive using labelled data only. For each dataset, we compute the mean of each organelle as the mean of all proteins associated with that organelle. We then align each of these datasets onto the same PCA coordinates and visualise. Contours are overlaid to visualise uncertainty in the location in the PCA plot across different datasets. (a) Principal components 1 and 2 are shown; (b) visualises principal components 1 and 3.

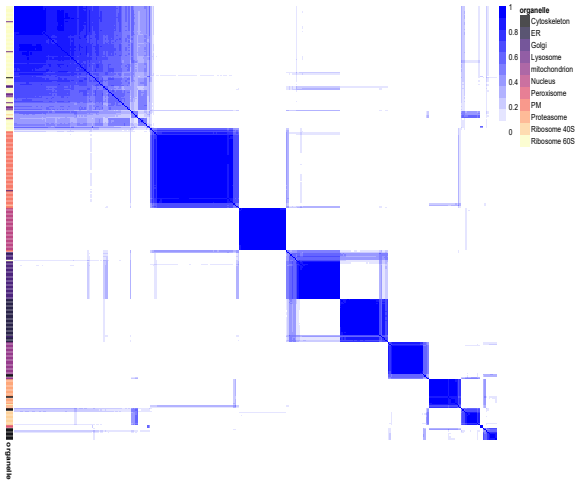


(a)

Fig. B.11 The quadratic loss (Brier Score) across different simulation scenarios. The simulations are in order of those presented in the text. Furthermore, the legend indicates the different simulation settings that have been achieved. Results are similar across most of the scenarios, whilst we see decreased performance as the noise of the unlabelled data increases.



(a)



(b)

Fig. B.12 Example PSMs of the induced posterior partition when the data are simulated from the Matern model and inference is perform using the squared exponential covariance; that is, the case of covariance function misspecification. (a) Example PSM when the smoothness is 2; (b) Example PSM when the smoothness is 3.

Table B.6 A table reporting the proportion of outliers that were incorrectly allocated as outliers along with 95% confidence intervals. Those correctly allocated as outliers are also report along with with 95% confidence intervals. The left hand column indicates the corresponding simulation scenario.

Simulation setting	Incorrectly allocated as outlier	Correctly allocated as outlier
Defaults	0 [0,0.02]	0.82[0.44,0.89]
B(2,4)	0 [0,0.02]	0.82[0.40,0.89]
B(2,2)	0 [0,0.02]	0.82[0.41,0.92]
Amplitude LN(0,0.5)	0 [0,0.02]	0.82[0.42,0.92]
Length Scale LN(0,0.5)	0 [0,0.02]	0.82[0.40,0.89]
Noise LN(0,0.5)	0 [0,0.01]	0.82[0.38,0.87]
Unlabelled Noise times 2	0 [0,0.01]	0.96[0.72,1.00]
Unlabelled Noise times 5	0 [0,0.02]	0.87[0.42,0.97]
Unlabelled Noise times 10	0.02[0,0.04]	0.88[0.50,0.99]
Matern smoothness is 2	0 [0,0.01]	0.85[0.68,1.00]
Matern smoothness is 3	0 [0,0.00]	0.85[0.58,0.99]
Poisson process outlier	0 [0,0.01]	0.83[0.74,0.99]

B.6 Appendix 6: Efficiency of HMC versus MH for hyparameter updates

Table B.7 A table summarising the difference in performance between Metropolis-Hastings and Hamiltonian Monte Carlo at sampling the hyperparameters of a GP for several different organelles. For each organelle and for each method we report the acceptance rate and the time-normalised effective sample size. It is clear that HMC outperforms MH according to this metric.

Component	Method	Iterations	Acceptance rate	Length-scale	Amplitude	Noise
Cytosol	MH	50,000	0.240	523	659	9375
	HMC	500	0.716	35348	54730	134485
Ribosome 40S	MH	50,000	0.297	259	582	10756
	HMC	500	0.742	14114	44662	27758
Lysosome	MH	50,000	0.273	403	821	10385
	HMC	500	0.710	28558	40955	543828
Proteosome	MH	50,000	0.267	408	712	10410
	HMC	500	0.800	16243	27186	55923
Actin	MH	50,000	0.409	436	1129	10841
	HMC	500	0.598	5750	479	6342

B.7 Appendix 7: Tables of hyperparameters

Tables of hyperparameters and hyperparameter distributions for the mouse pluripotent stem cell data.

Table B.8 A table of log hyperparameters for a GP found by optimising the marginal likelihood using L-BFGS

Sub-cellular niche	Length-scale	Amplitude	Noise
40S Ribosome	0.81	-2.45	-4.23
60S Ribosome	0.61	-2.90	-4.28
Actin cytoskeleton	0.44	-2.67	-3.77
Cytosol	0.80	-2.17	-3.66
ER/Golgi apparatus	0.96	-2.60	-3.82
Endosome	0.48	-2.48	-3.49
Extracellular matrix	0.53	-2.74	-4.06
Lysosome	0.64	-2.43	-4.03
Mitochondrion	0.55	-2.26	-3.77
Nucleus - Chromatin	0.46	-2.23	-3.71
Nucleus - Non-chromatin	0.23	-2.25	-3.47
Peroxisome	0.78	-2.40	-3.78
Plasma membrane	0.28	-2.41	-3.92
Proteasome	0.70	-2.01	-4.16

Table B.9 A table of log GP hyperparameters with 95% equi-tailed credible intervals summarised from samples produced using HMC

	Length-scale	Amplitude	Noise
40S Ribosome	$0.54\,[-0.64, 1.08]$	$-2.39\,[-2.74, -2.01]$	$-4.23\,[-4.29, -4.17]$
60S Ribosome	$0.51\,[-0.20, 0.93]$	$-2.77\,[-3.18, -2.31]$	$-4.28\,[-4.31, -4.23]$
Actin cytoskeleton	$0.33\,[-0.52, 0.81]$	$-2.55\,[-2.89, -2.20]$	$-3.76\,[-3.84, -3.68]$
Cytosol	$0.69\,[-0.01, 1.11]$	$-2.04\,[-2.43, -1.60]$	$-3.66\,[-3.70, -3.61]$
ER/Golgi apparatus	$0.89\,[0.29, 1.37]$	$-2.53\,[-2.90, -1.89]$	$-3.82\,[-3.85, -3.79]$
Endosome	$0.39\,[-0.24, 0.84]$	$-2.37\,[-2.68, -1.92]$	$-3.48\,[-3.58, -3.39]$
Extracellular matrix	$0.37[-0.32, 0.92]$	$-2.65[-2.97, -2.24]$	$-4.05\,[-4.14, -3.96]$
Lysosome	$0.54\,[-0.31, 0.94]$	$-2.36\,[-2.69, -2.00]$	$-4.03\,[-4.09, -3.98]$
Mitochondrion	$0.53\,[0.12, 0.95]$	$-2.12\,[-2.38, -1.80]$	$-3.77\,[-3.78, -3.75]$
Nucleus - Chromatin	$0.46\,[0.05, 0.86]$	$-2.14\,[-2.45, -1.81]$	$-3.71\,[-3.75, -3.68]$
Nucleus - Non-chromatin	$0.05\,[-1.19, 0.69]$	$-2.09\,[-2.48, -1.71]$	$-3.47\,[-3.50, -3.44]$
Peroxisome	$0.75\,[0.28, 1.17]$	$-2.31\,[-2.62, -1.92]$	$-3.78\,[-3.85, -.3.69]$
Plasma membrane	$0.02\,[-1.03, 0.67]$	$-2.32\,[-2.65, -1.91]$	$-3.91\,[-3.95, -3.86]$
Proteasome	$0.59\,[0.16.0.97]$	$-1.94\,[-2.26, -1.52]$	$-4.15\,[-4.21, -4.10]$

Appendix C

Appendix to chapter 7

C.1 Appendix 1: Additional simulations

We perform additional simulations comparing the MR approach to BANDLE. The simulation scenarios are the same as performed in the main text. However, we start from the LOPIT-DC dataset of [159] instead. The conclusion are as for the main text that BANDLE significantly outperforms the MR method.

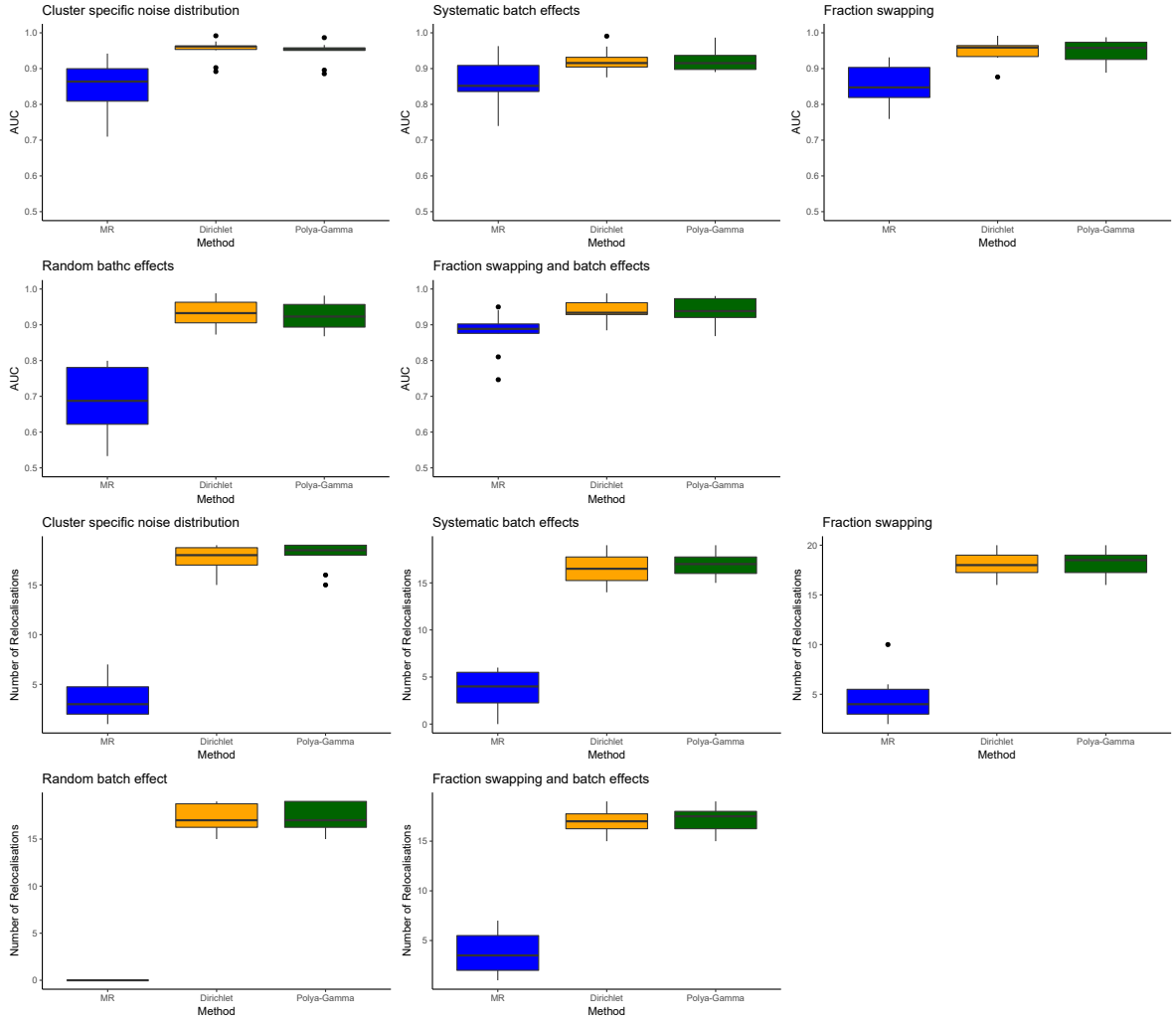


Fig. C.1 The first 5 boxplots compare MR to BANDLE with two different prior settings, using the area under curve (AUC). Distributions are over new simulated datasets. The second set of boxplots demonstrate how these AUCs translate into confident differential localisation events.

C.2 Appendix 2: Convergence analysis EGF stimulation

We ran our MCMC sampler for $20,000$ iterations, where we discarded $10,000$ iterations for burn-in and retained every 10^{th} iteration for thinning to reduce autocorrelation. 8 chains were run in parallel and two were discarded for lack convergence by visual inspection. Example trace plots are plotted below. We further assessed convergence by computing $\hat{R}$ for parallel chains of the mixing weights and confirmed that they were less than 1.01 indicating that are chains are well-mixed. Finally, we concatenated the 6 remaining chains and computed the rank of each sample. These ranks are the plotted in separate histograms for each chain separately. Departures from uniformity of these histograms indicates non-convergence and we observe well behaved rank plots.

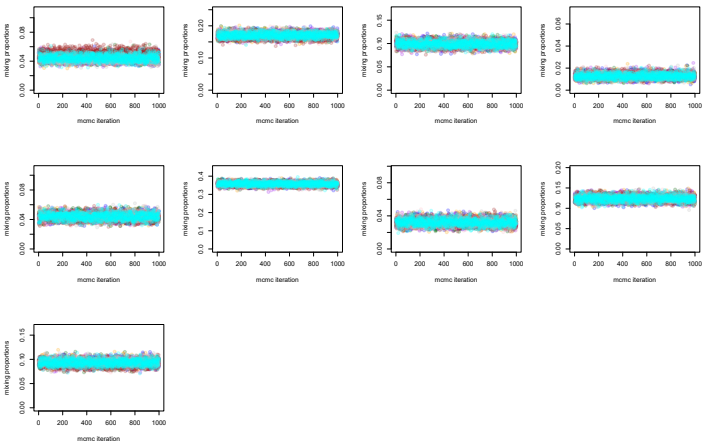


Fig. C.2 MCMC traceplot for EGF data

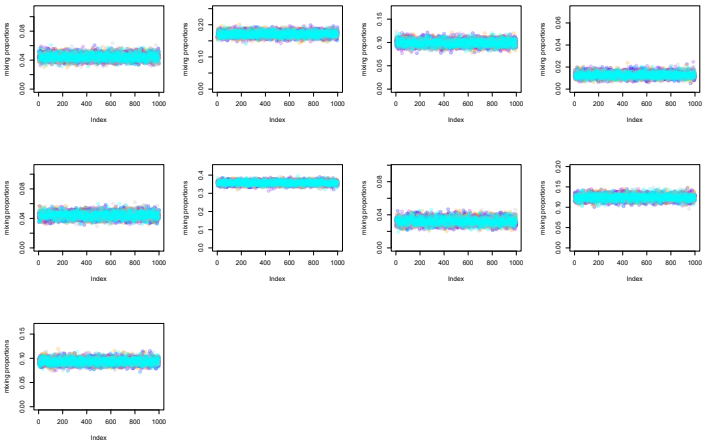


Fig. C.3 MCMC traceplot for EGF data

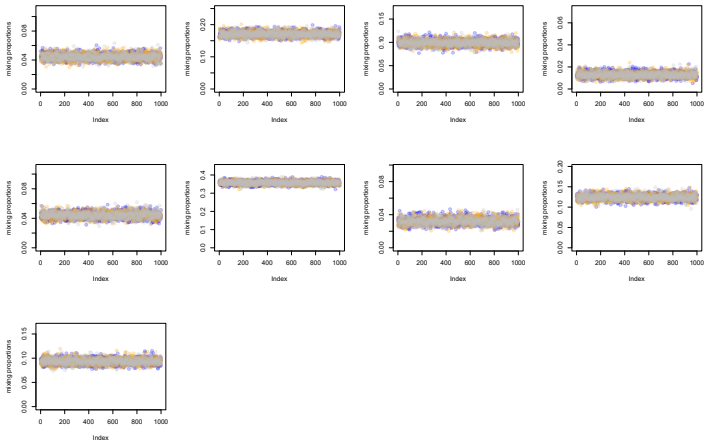


Fig. C.4 MCMC traceplot for EGF data

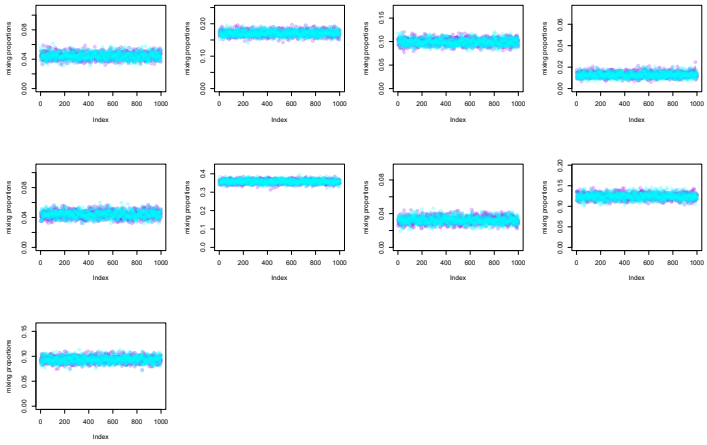


Fig. C.5 MCMC traceplot for EGF data

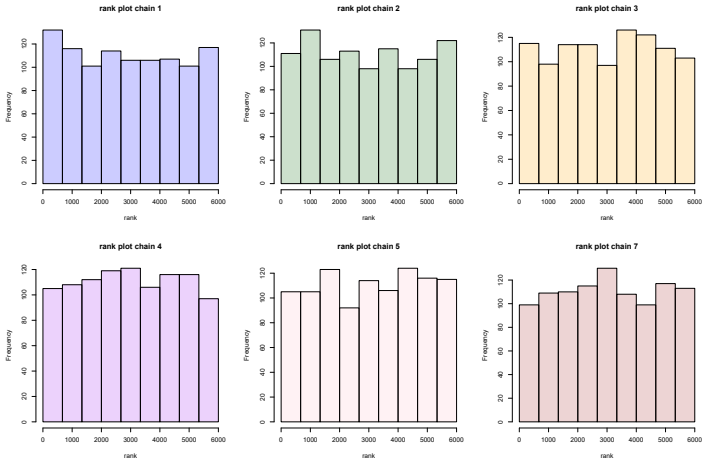


Fig. C.6 MCMC rank plot for EGF data

C.3 Appendix 3: EGF stimulation phosphoproteomics time course

Example abundance changes for the phosphoproteomic time course experiment.

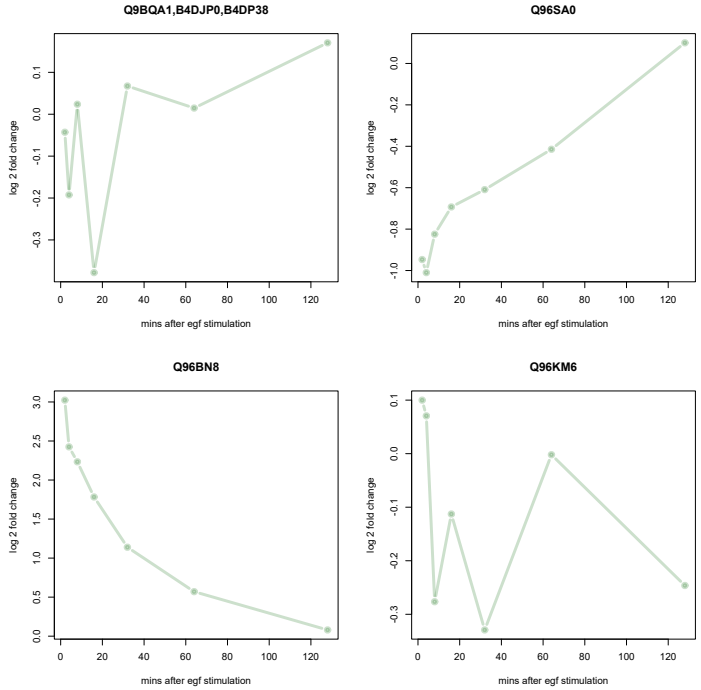


Fig. C.7 Example trajectories from the timecourse phosphoproteomics experiment

C.4 Appendix 4: Convergence analysis AP-4 knockout

We ran our MCMC sampler for $20,000$ iterations, where we discarded $10,000$ iterations for burn-in and retained every 50^{th} iteration for thinning to reduce autocorrelation. 6 chains were run in parallel and one was discarded for lack convergence by visual inspection. Example trace plots are plotted below. We further assessed convergence by computing $\hat{R}$ for parallel chains of the mixing weights and confirmed that they were less than 1.01 indicating that are chains are well-mixed. Finally, we concatenated the 5 remaining chains and computed the rank of each sample. These ranks are the plotted in separate histograms for each chain separately. Departures from uniformity of these histograms indicates non-convergence and we observe well behaved rank plots.

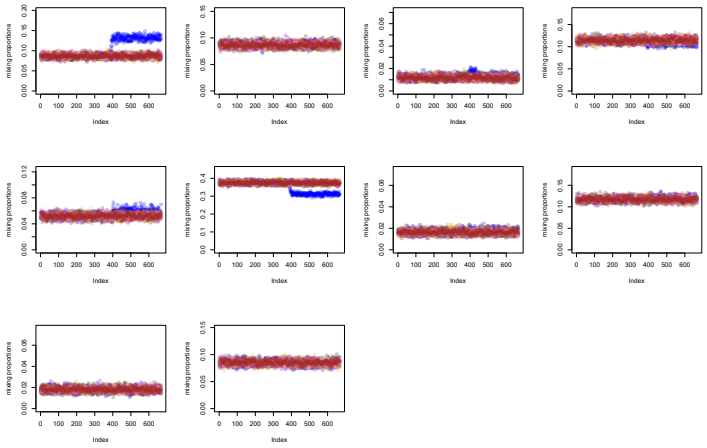


Fig. C.8 MCMC trace plot for AP-4 dataset

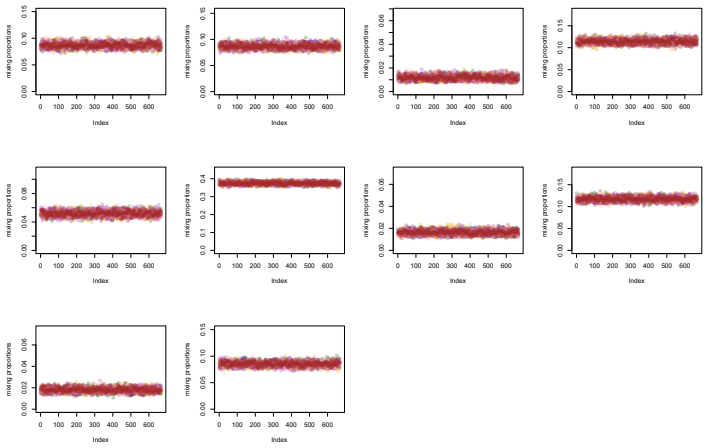


Fig. C.9 MCMC traceplot for AP-4 dataset

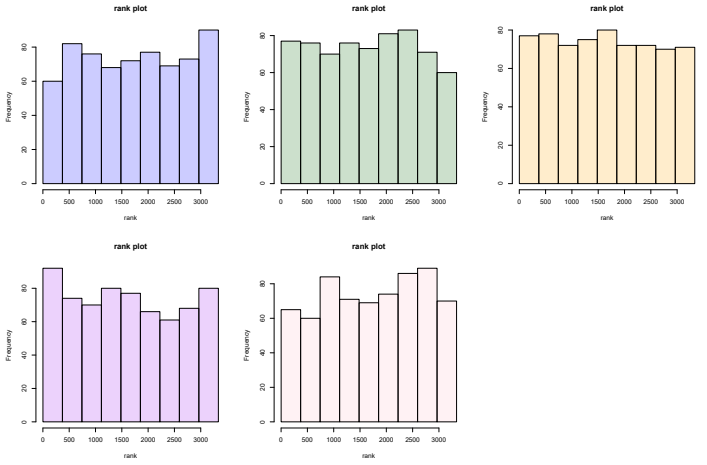


Fig. C.10 MCMC rank plot for AP-4 dataset

C.5 Appendix 5: Convergence analysis for HCMV datasets

We ran our MCMC sampler for $20,000$ iterations, where we discarded $10,000$ iterations for burn-in and retained every 50^{th} iteration for thinning to reduce autocorrelation. 6 chains were run in parallel and convergence was analysed by visual inspection, and one chains was discarded. Example trace plots are plotted below. We further assessed convergence by computing $\hat{R}$ for parallel chains of the mixing weights and confirmed that they were less than 1.01 indicating that are chains are well-mixed. Finally, we concatenated the 5 remaining chains and computed the rank of each sample. These ranks are the plotted in separate histograms for each chain separately. Departures from uniformity of these histograms indicates non-convergence and we observe well behaved rank plots.

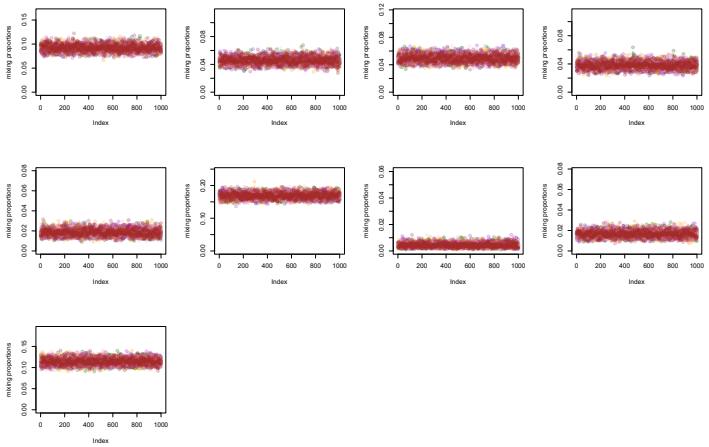


Fig. C.11 MCMC trace plot for HCMV dataset 24 hpi

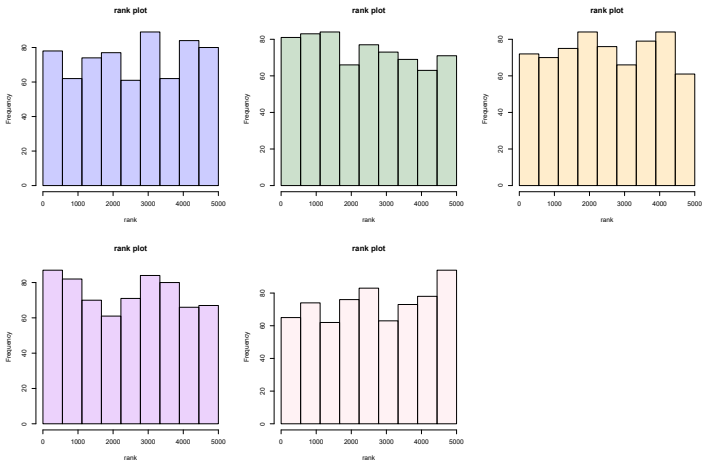


Fig. C.12 MCMC rank plot for HCMV dataset 24 hpi

C.6 Appendix 6: Prior settings and sensitivity analysis

The hyperparameter α is the prior on the mixing weights $\boldsymbol{\pi}$, where π_{ij} is the prior probability that a protein belongs to the i^{th} niche in the control dataset and niche j in the treatment dataset. The entries of α can be interpreted as the prior relative proportions of protein allocations. Let J be the matrix of all ones, it is typical in Bayesian mixture modelling to set $\alpha = 0.5J$ or $\alpha = J$, corresponding to the Jeffreys' prior and the symmetric prior respectively. However, in our scenario the diagonal and off-diagonal terms have different meanings. The diagonal terms correspond to proteins allocated to the same niche in both datasets and the off-diagonal terms correspond to differential localised proteins. However, there are far more off diagonal terms than diagonal terms. Hence, the Jeffreys' and symmetric priors implicitly assume that the there are more differentially localised proteins that spatial stable. Of course, this is at odds with our expectations and thus we opt for a more sensible weakly informative prior as a default. We set $\alpha_{jj} = 1$ and $\alpha_{jk} = 0.01$ for $k \neq j$. This assume that there are roughly an order of magnitude fewer differentially localised proteins that spatially stable ones. This default is used in all simulations and application except the EGF simulation dataset. In that case, we have prior knowledge of a differentially localisation between the Plasma membrane and the Endosome and so we set the corresponding entry of α to 1.

In general, we do not find that our analysis is very sensitive to the prior choice. To demonstrate, we perform a sensitivity analysis for the results using the Jeffreys' prior, the symmetric prior and our weakly informative prior. In the context of the simulation example in section 7.4.2, we apply the different prior choice an examine the results. Since the primary quantity of interest is the prediction of differentially localised proteins we examine this quantity. The following ROC curve demonstrate that the results are almost identical across the different prior choices.

Prior information is carefully encoded using domain knowledge and previous analysis. In brief, for the EGF application, we encode that the most likely transition is from plasma membrane to Lysosome. To evaluate the coherence of our prior we perform a *prior predictive check* [162]. The summary statistic of interest is the number of differential localisation and the expected number of differential localisation, given our prior, is roughly 4.6. Furthermore, the prior probability that there are more than 15 differential localisation is less than 0.01.

For the AP-4 Application, we first performed a prior predictive check and find that our prior configuration leads to a 5.3 proteins *a priori* differential localisation in expectation and the probability that more than 15 proteins are differential localised is ≈ 0.03.

For the HCMV application, priors are set such that the expected prior number of differential localisation is roughly 3.

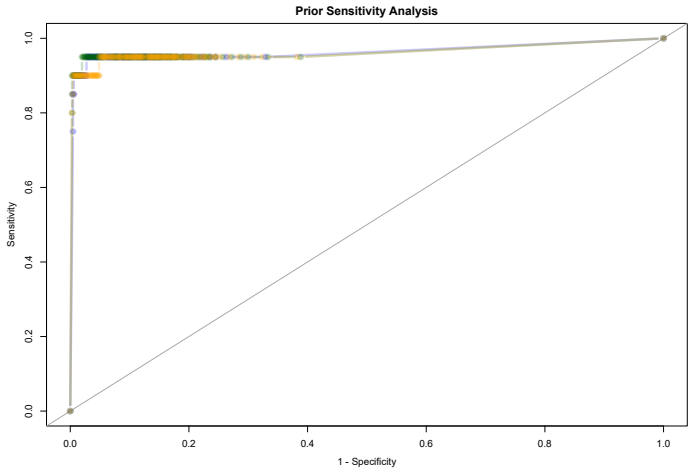


Fig. C.13 ROC curve for examining prior sensitivity. Blue corresponds to default, dark green to the Jeffreys' prior and orange to the symmetric prior. The curves are essentially indistinguishable.

C.7 Appendix 7: Selecting τ

One hyperparameter that we haven not yet discussed in the choice of τ when using the empirical strategy to select the prior for the Pólya-Gamma based prior (see supplementary methods). One possible way to select τ is to first perform a prior predictive check. However, this can be arduous if a reasonable value is not known in advance. We suggest on strategy for generating appropriate values of τ. The first is to select a weakly informative Dirichlet prior, for example, the default we have suggested in the previous section. We then compute the standardised KL divergence between this weakly informative Dirichlet prior and the a range of possible Pólya-Gamma based priors. If we believe that our Dirichlet prior is sensible then a sensible Pólya-Gamma prior will have low KL divergence. In figure C.14, we vary the value of τ (on the log scale) for different values of mean for the Pólya-Gamma prior. There is a clear elbow in this plot. We do not advise purely selecting the value of τ which minimises this KL divergence, rather choose τ roughly in the that region and perform a prior predictive check to ensure that it leads to sensible prior inferences. A default value of $\tau = 0.3$ appears to work well in practice.

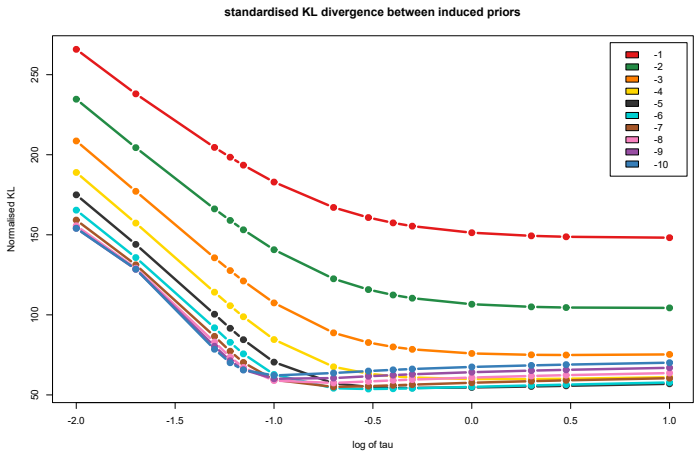


Fig. C.14 KL divergence plot show KL divergence between the Polya-Gamma prior and the weakly informative default Dirichlet prior for vary values of τ. Each colour indicates a different choice of mean for the Polya-Gamma based prior.

C.8 Appendix 8: EGF stimulation figures

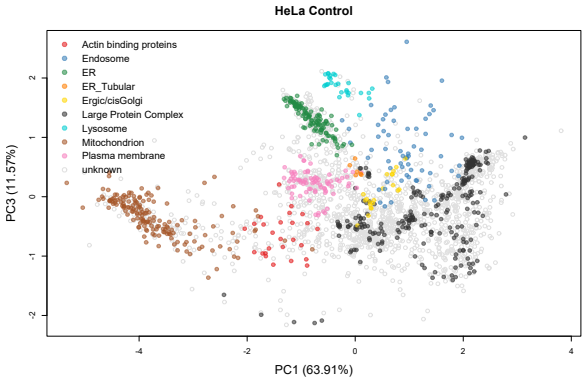


Fig. C.15 A PCA plot of the control HeLA dataset from [220]. Each pointer corresponds to a protein and marker proteins are highlighted according to their subcellular niche.

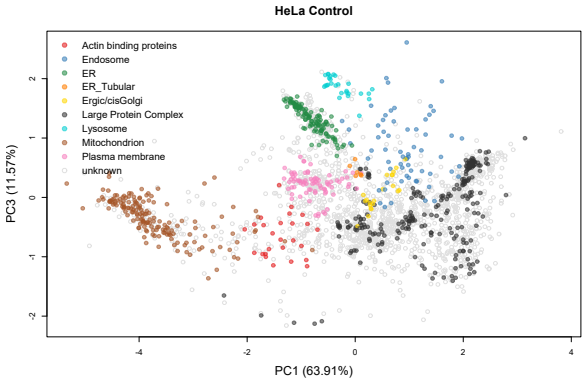


Fig. C.16 A PCA plot of the EGF stimulated HeLA dataset from [220]. Each pointer corresponds to a protein and marker proteins are highlighted according to their subcellular niche.

C.9 Appendix 9: AP-4 knockout figures

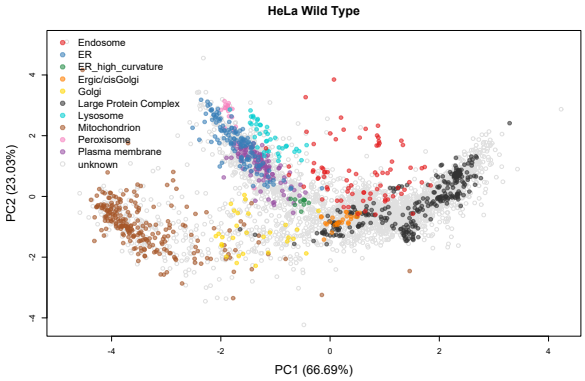


Fig. C.17 A PCA plot of the control HeLA dataset from [92]. Each pointer corresponds to a protein and marker proteins are highlighted according to their subcellular niche.

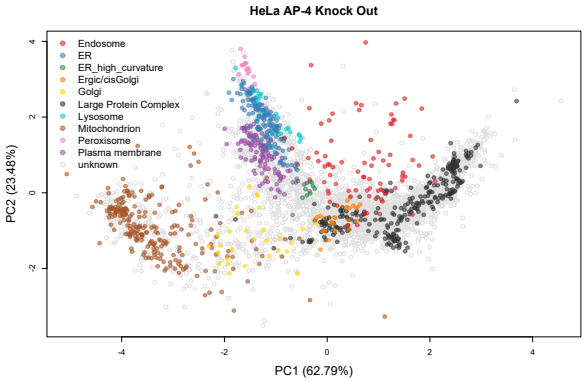


Fig. C.18 A PCA plot of the AP-4 knockout HeLA dataset from [92]. Each pointer corresponds to a protein and marker proteins are highlighted according to their subcellular niche.

C.10 Appendix 10: HCMV PCA plots

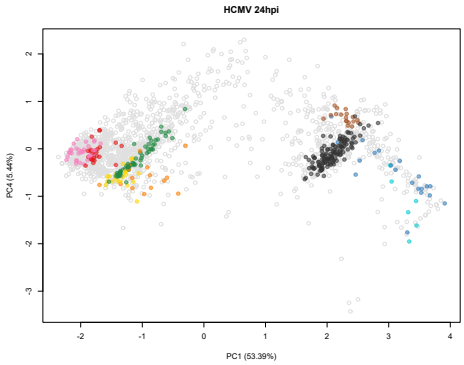


Fig. C.19 A PCA plot of control fibroblast cells dataset from [24]. Each pointer corresponds to a protein and marker proteins are highlighted according to their subcellular niche.

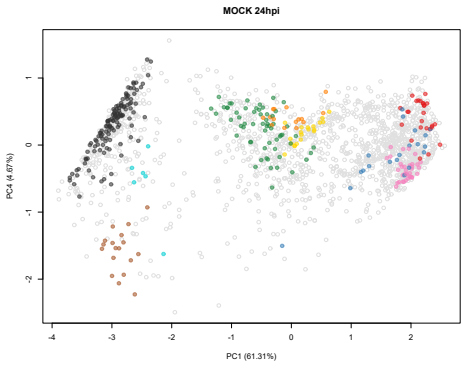


Fig. C.20 A PCA plot of HCMV infected fibroblast cells dataset from [24]. Each pointer corresponds to a protein and marker proteins are highlighted according to their subcellular niche.

C.11 Appendix 11: GO enrichment analysis HCMV dataset

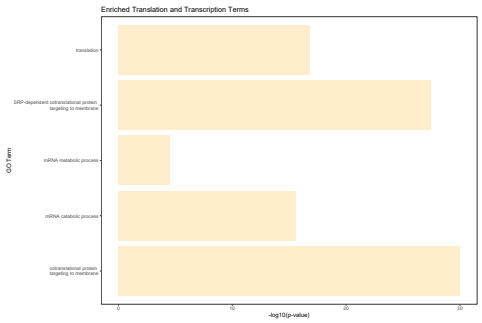


Fig. C.21 GO enrichment results (Translation and Transcription terms)

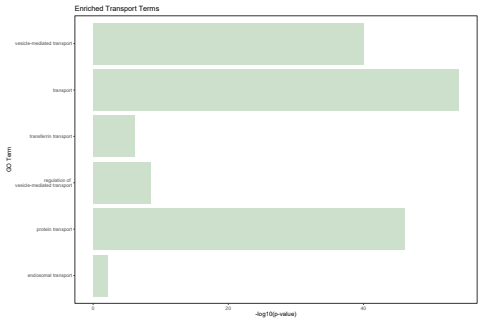


Fig. C.22 GO enrichment results (Transport terms)

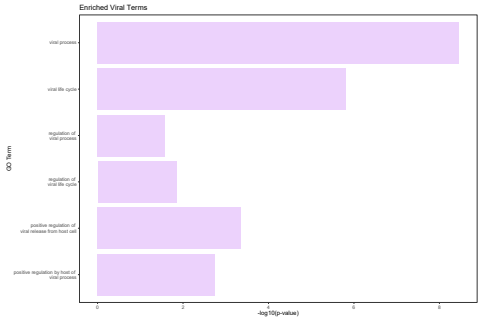


Fig. C.23 GO enrichment results (Viral processes)

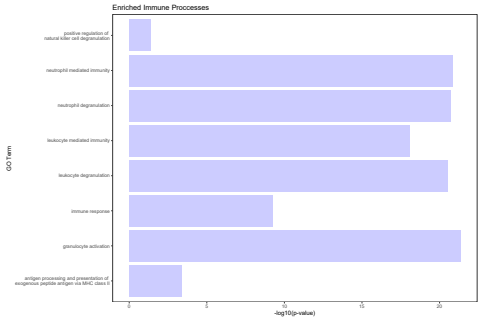


Fig. C.24 GO enrichment results (Immune processes)

C.12 Appendix 12: HCMV additional figures abundance and degradation assays

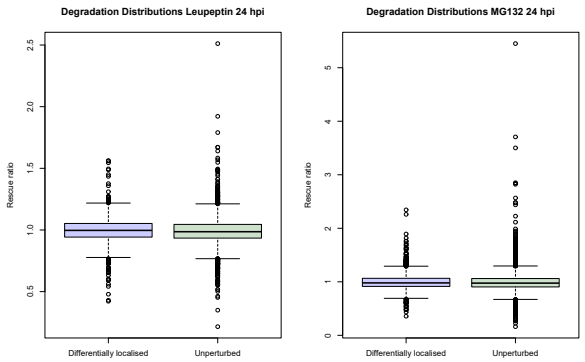


Fig. C.25 Boxplots of the global degradation distributions for MG132 and leupeptin. Separate distributions are plotted for differentially localised proteins are those are not. No difference is observed.

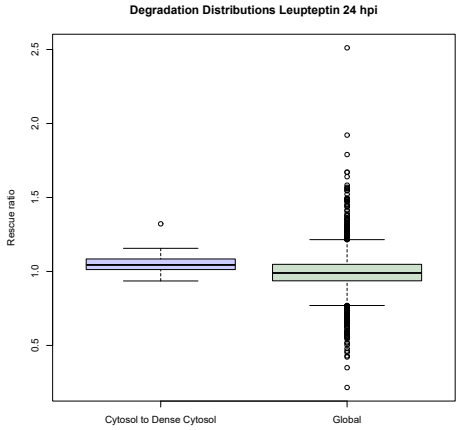


Fig. C.26 Leupeptin distributions of protein recruited from the cytosol to the dense cytosol, showing increased proteins targeted for degradation.

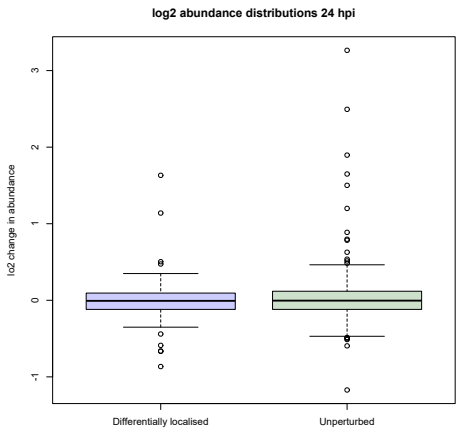


Fig. C.27 Global abundance distributions for proteins 24 hpi separated into differentially localised or not. There is no difference between differentially proteins and those that are not.

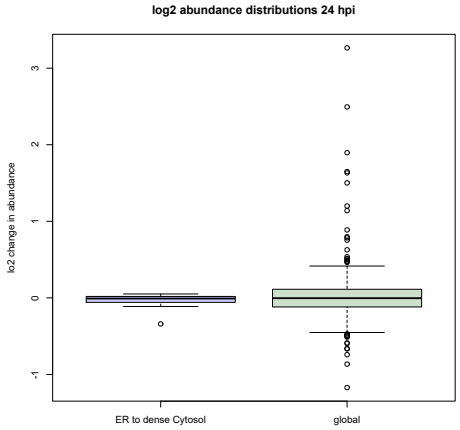


Fig. C.28 Boxplots for $\log_2$ normalised abundance distributions. Proteins recruited from the ER to dense cytosol show a decrease in abundance when compared to the global distribution.

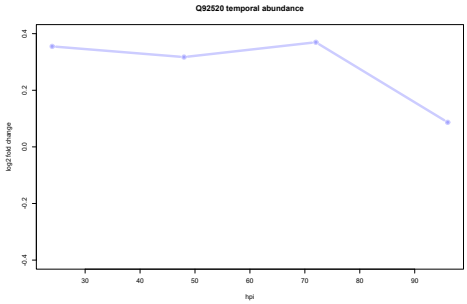


Fig. C.29 The temporal abundance of Q92520, clearly Q92520 is upregulated until 92 hpi.

C.13 Appendix 13: HCMV additional figures acetylation data

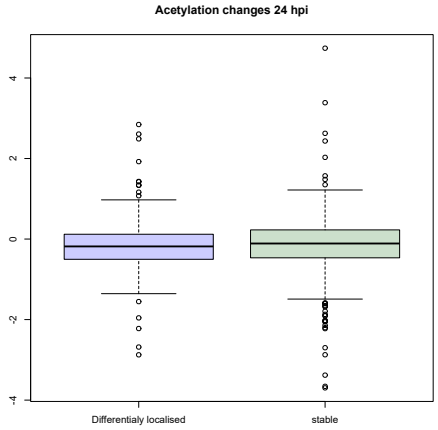


Fig. C.30 Global distributions for acetylation changes for HCMV 24 hpi compare to MOCK. We do not observe any correlations between differential localisation and acetylation changes.

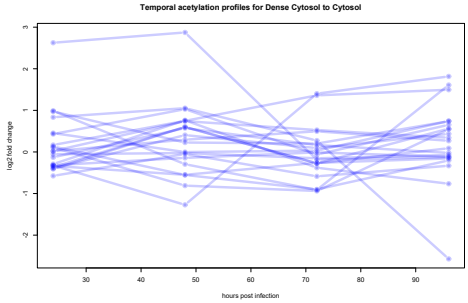


Fig. C.31 Temporal acetylation profiles for HCMV infected cells which relocalise from dense Cytosol to the Cytosol. Skp1 has a 2.5 fold increase in acetylation at 24 hpi.

C.14 Appendix 14: HCMV interactome figures

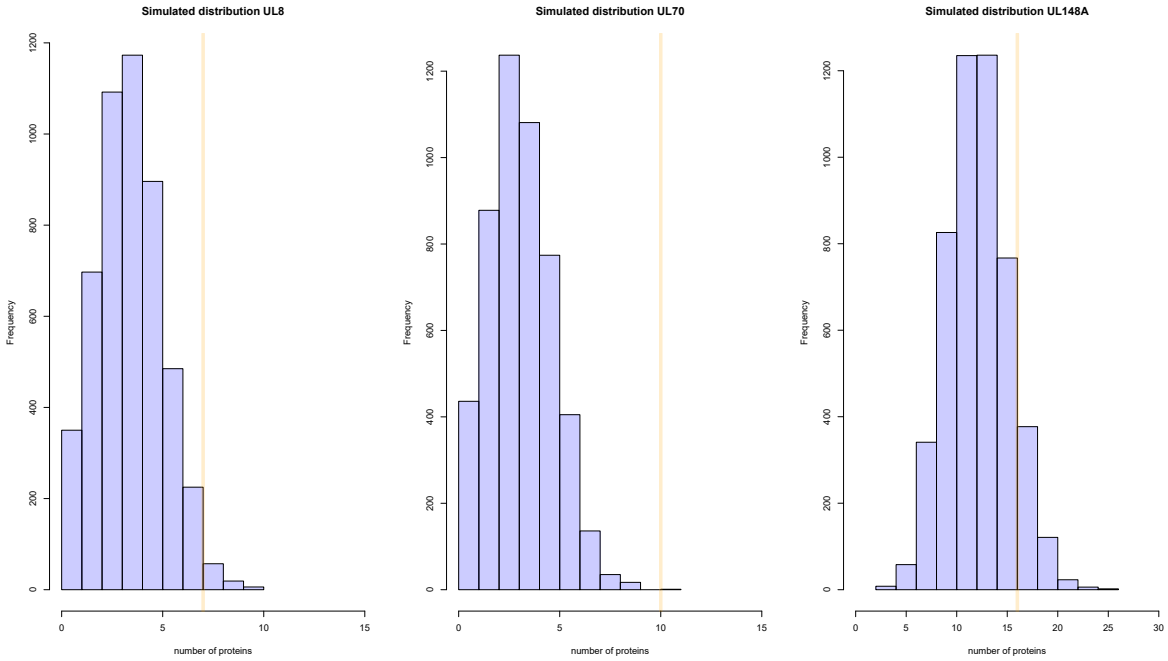


Fig. C.32 Distributions for predicted number of proteins to be in the same localisation given the spatial pattern observed in A for each of the viral interactomes. The observed statistic is marked in orange, UL8 and UL70 have more proteins in the same location than would be expected at random. This is not true for UL148A.

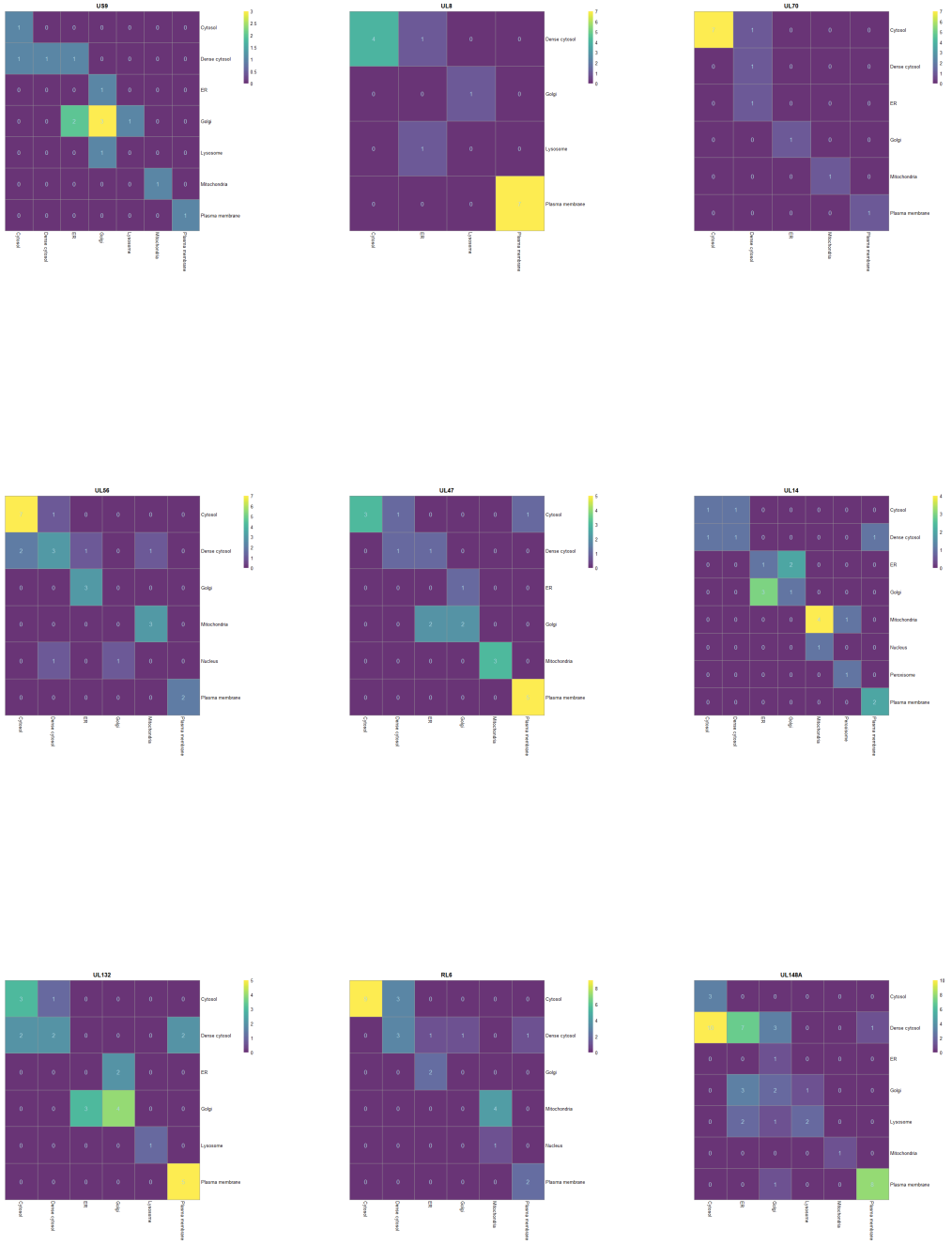


Fig. C.33 Protein localisation distribution and relocalisation for viral interactomes

C.15 Appendix 15: Supplementary methods

C.15.1 A non-conjugate prior

Thus far we have been using a conjugate Dirichlet prior for the a priori mixing proportions $\boldsymbol{\pi}$. Our model assumes no correlation across $\boldsymbol{\pi}$ due to the use of the Dirichlet distribution. However, we describe how we can extend the model to include correlations. Firstly, the joint prior of the allocation probabilities is

$$(z_{i,1}, z_{i,2}) \sim cat(f(\boldsymbol{\pi})), \tag{C.1}$$

where $f(\boldsymbol{\pi}) = \frac{\exp(\boldsymbol{\pi})}{\sum_{j,k}\exp(\pi_{jk})}$. Thus the prior correlations between organelles can be included using a multivariate Gaussian

$$vec(\boldsymbol{\pi})|\mu, \Sigma \sim \mathcal{N}(\mu, \Sigma). \tag{C.2}$$

Given $\boldsymbol{\pi}$ the underlying conditional posterior allocation probabilities are the same as before. The conditional posterior of $\boldsymbol{\pi}$ now changes and because of loss of conjugacy a metropolis-hastings step is required. In the next section, we develop a prior using stick-breaking Pólya-Gamma augmentation to facilitate Gibbs sampling.

C.15.2 Pólya-Gamma augmentation

A random variable X has a Pólya-Gamma distribution with parameters $b > 0$ and $c \in \mathbb{R}$, denoted $X \sim \mathcal{PG}(b, c)$ if [368]

$$X =_d \frac{1}{2\pi^2}\sum_{k=1}^{\infty}\frac{g_k}{(k-\frac{1}{2})^2 + \frac{c^2}{4\pi^2}}, \tag{C.3}$$

where $g_k \sim_{iid} \mathcal{G}(b, 1)$. The fundamental equation that renders Pólya-Gamma augmentation useful is the following

$$\frac{(e^{\phi})^a}{(1+e^{\phi})^b} = 2^{-b}e^{\kappa\phi}\int_0^{\infty} e^{-\omega\phi^2/2}p(\omega)\,d\omega, \tag{C.4}$$

where $\kappa = a - b/2$ and $w \sim \mathcal{PG}(b, 0)$. This is advantageous because of the following construction. Consider the binomial regression problem

$$y_i = \text{Binom}\left(n_i, \frac{1}{1+e^{-x_i^T\beta}}\right), \tag{C.5}$$

where β has the Gaussian prior $\mathcal{N}(\mu, \Sigma)$. To sample from the posterior using Pólya-Gamma augmentation, we first introduce strategic variables w and then sample according to

$$w_i|\beta \sim \mathcal{PG}(n_i, x_i^T\beta) \tag{C.6}$$

$$\beta|y, w \sim \mathcal{N}(\tilde{\mu}, \tilde{\Sigma}), \tag{C.7}$$

where

$$\tilde{\Sigma} = (X^T\Omega X + \Sigma^{-1})^{-1} \tag{C.8}$$

$$\tilde{\mu} = \tilde{\Sigma}(X^T\kappa + \Sigma^{-1}\mu), \tag{C.9}$$

and

$$\kappa = (y_1 - n_1/2, ..., y_N - n_N/2) \tag{C.10}$$

$$\Omega = diag(\omega_1, ..., \omega_N). \tag{C.11}$$

Thus, a simple tuning free two step auxiliary variable sampler is required rather than Metropolis-Hastings move. To adapt this method to our situation we consider a slightly different construction in the following section.

C.15.3 Stick-breaking Pólya-Gamma augmentation

In this section, we extended the Pólya-Gamma augmentation of the previous section to multinomial variables using a stick-breaking approach [270]. Consider a likelihood of the form

$$p(x|\phi) = c(x)\frac{(e^{\phi})^{a(x)}}{(1+e^{\phi})^{b(x)}}. \tag{C.12}$$

The joint probability distribution can then be written as

$$p(\phi, x) = p(\phi)c(x)\frac{(e^{\phi})^{a(x)}}{(1+e^{\phi})^{b(x)}} = p(\phi)c(x)2^{-b(x)}e^{\kappa(x)\phi}\int_0^{\infty} e^{-\omega\phi^2/2}p(\omega)\,d\omega \tag{C.13}$$

Thus, the conditional distribution can be written as

$$p(\phi|x, \omega) \propto p(\phi)e^{\kappa(x)\phi}e^{-\omega\phi^2/2}, \tag{C.14}$$

which is Gaussian if $p(\phi)$ is Gaussian. Furthermore, by the exponential tilting property of the Pólya-Gamma distribution [368] it follows that

$$\omega|\phi, x \sim \mathcal{PG}(b(x), \phi). \tag{C.15}$$

Now consider a multinomial model on K categories with N trials with probability vector π. It can be written as a stick-breaking construction of binomials as follows

$$\text{Multi}(x|N,\pi) = \prod_{k=1}^{K-1} \text{Binom}(x_k|N_k,\tilde{\pi}_k), \tag{C.16}$$

where

$$N_k = N - \sum_{j<k} x_j \tag{C.17}$$

$$\tilde{\pi}_k = \frac{\pi_k}{1-\sum_{j<k}\pi_j} \tag{C.18}$$

Let $\sigma(\phi_k) = \exp(\phi_k)/(1+\exp(\phi_k))$ and $\tilde{\pi}_k = \sigma(\phi_k)$. Now Substituting into the stick-breaking model

$$\text{Multi}(x|N,\pi) = \prod_{k=1}^{K-1} \text{Binom}(x_k|N_k,\sigma(\phi_k)) \tag{C.19}$$

$$= \prod_{k=1}^{K-1} \binom{N_k}{x_k} \sigma(\phi_k)^{x_k}(1-\sigma(\phi_k))^{N_k-x_k} \tag{C.20}$$

$$= \prod_{k=1}^{K-1} \binom{N_k}{x_k} \frac{(e^{\phi_k})^{x_k}}{(1+e^{\phi_k})^{N_k}}. \tag{C.21}$$

Thus, we can set $a_k(x) = x_k$ and $b_k(x) = N_k$ and introduce Pólya-Gamma variables w_k. Then

$$p(x,w|\phi) \propto \prod_{k=1}^{K-1} \exp\left[\left(x_k - \frac{N_k}{2}\right)\phi_k - \frac{w_k\phi_k^2}{2}\right] \propto N(\phi|\Omega^{-1}\kappa(x),\Omega^{-1}), \tag{C.22}$$

where $\Omega = diag(\omega_1,...,\omega_K)$ and $\kappa(x_k) = x_k - N_k/2$.

C.15.4 A correlated model for differential localisation

The above schema allows us to construct a correlated differential localisation model, using stick-breaking Pólya-Gamma augmentation. Suppose that there are K organelles to which a protein could localises. Then we specify a joint model on the allocation probabilities

$$vec(\boldsymbol{\pi})|\mu,\Sigma \sim \mathcal{N}(\mu,\Sigma) \tag{C.23}$$

$$(z_{i,1},z_{i,2}) \sim cat(f(\boldsymbol{\pi})) \tag{C.24}$$

$$\omega \sim \mathcal{PG}(1,0), \tag{C.25}$$

For easy of notation let $\psi = vec(\boldsymbol{\pi})$ and f is the stick-breaking map. Then it follows from the previous sections

$$p(\psi|Z_1, Z_2, \omega) \propto N(\psi|\Omega^{-1}\kappa, \Omega^{-1})N(\psi|\mu, \Sigma) \propto N(\psi|\tilde{\mu}, \tilde{\Sigma}), \tag{C.26}$$

where

$$\tilde{\mu} = \tilde{\Sigma}\left(\kappa + \Sigma^{-1}\mu\right) \tag{C.27}$$

$$\tilde{\Sigma} = \left(\Omega + \Sigma^{-1}\right)^{-1}. \tag{C.28}$$

To compute κ, first let $n_{j,k} = \sum_i \mathbb{1}(z_{i1} = j, z_{i2} = k)$ and let $\boldsymbol{n} = vec(n)$. Then $\kappa_l = \boldsymbol{n}_l - \frac{1}{2}$ for $l = 1, ..., K^2$. Finally, we can sample the conditional posterior of the Pólya-Gamma variables

$$\omega_l|Z_1, Z_2, \psi \sim \mathcal{PG}(1, \psi_l). \tag{C.29}$$

C.15.5 Calibration of Polya-Gamma prior

The Polya-Gamma augmentation method was used to take advantage of the knowledge that some classes were known to be correlated *a priori*. The Polya-Gamma prior admits are more flexible prior to be placed on the prior allocation probabilities. Recall that the following prior on the allocation probabiltiies

$$p(z_{i,1} = k, z_{i,2} = k'|\boldsymbol{\pi}) = f(\pi_{kk'}). \tag{C.30}$$

This prior is then expanded hierarchically in the following fasion:

$$vec(\boldsymbol{\pi})|\mu, \Sigma \sim \mathcal{N}(\mu, \Sigma) \tag{C.31}$$

$$(z_{i,1}, z_{i,2}) \sim cat(f(\boldsymbol{\pi})) \tag{C.32}$$

$$\omega \sim \mathcal{PG}(1, 0), \tag{C.33}$$

where,

$$f(\pi_{kk'}) = \sigma(\pi_{kk'})(1 - \sum_{j<k, j'<k'} f(\pi_{jj'})) \tag{C.34}$$

There are no analytic formula for the moments of the logit-normal distribution and thus analysing the behaviour of the above prior above is challenging. The implied distribution on $f(\boldsymbol{\pi})$ can be computed by standard transformations:

$$p(f(\boldsymbol{\pi})|\mu, \Sigma) = \mathcal{N}(vec(\boldsymbol{\pi})|\mu, \Sigma) \cdot \prod_{k,k'} \left[\frac{1 - \sum_{j<k, j'<k'} f(\pi_{jj'})}{f(\pi_{kk'})\left(1 - \sum_{j\leq k, j'\leq k'} f(\pi_{jj'}\right)} \right]. \tag{C.35}$$

This equation clearly demonstrate the complexity of the prior. Recall we are interested in the quantity

$$p(z_{i,1} \neq z_{i,2}|\boldsymbol{\pi}) =: \rho_{pg} = \sum_{j,k;j\neq k} f(\pi_{jk}). \tag{C.36}$$

The prior expectation of the above and the following prior quantile can be used to calibrate the prior:

$$p(N_U \rho_{pg} > q) = p\left(N_U \sum_{j,k;j\neq k} f(\pi_{jk}) > q\right) = \delta. \tag{C.37}$$

This computation can be performed via Monte-Carlo simulation and corresponding quantiles as for the Dirichlet prior can be calibrated.

$$p\left(N_U \sum_{j,k;j\neq k} f(\pi_{jk}) > q\right) \approx \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}\left(N_U \sum_{j,k;j\neq k} f(\pi_{jk}^{(t)}) > q\right) \tag{C.38}$$

However, this is impractical in general for user to specify such a complex prior, since it requires the specification of a full covariance matrix. To alleviate this we suggest using prior data to set this prior. We suggest computing Σ_1 ,the covariance between the classes using the marker data from the first dataset, and likewise Σ_2, the covariance between the classes from the second dataset. We then set the prior covariance

$$\Sigma = \tau^{-1} \cdot (\Sigma_1 + \lambda I_K \otimes \Sigma_2 + \lambda I_K) \tag{C.39}$$

or the precision

$$\Sigma^{-1} = \tau \cdot \left((\Sigma_1 + \lambda I_K)^{-1} \otimes (\Sigma_2 + \lambda I_K)^{-1}\right), \tag{C.40}$$

where τ is a tuning parameter that is user specified and λI_K is constant multiple of the identity to provide stability.

C.15.6 Prior Coherence Analysis

The previous sections have constructed two different priors, that capture prior beliefs in different ways. The Dirichlet prior is considerably easier to specify and illicit for domain expertise; however, the stick-breaking Pólya-Gamma prior is much more flexible and can encode more complex prior beliefs - with the task that the prior is more challenging to specify. This section elaborates on the differences between the priors.

Suppose we are given a Dirichlet prior on $\boldsymbol{\pi}$, we can compute the corresponding distribution on $\psi = g_{SB}^{-1}(\boldsymbol{\pi})$, where g_{SB} denotes the stick-breaking map. Recall the matrix Dirichlet prior:

$$q(\boldsymbol{\pi}|\alpha) = \prod_{k=1}^{K} \frac{1}{\mathcal{B}(\alpha_k)} \prod_{j=1}^{K} \pi_{jk}^{\alpha_{jk}-1}. \tag{C.41}$$

The induced prior on ψ, computed from a change of variables, is the following

$$q(\psi|\alpha) = \frac{1}{\mathcal{B}(\alpha)} \prod_{k,k'} \sigma(\psi_{k,k'})^{\alpha_{kk'}} \sigma(-\psi_{k,k'})^{\sum_{j>k,j>k'} \alpha_{jj'}}. \tag{C.42}$$

As well as looking at the induced priors on the corresponding parameter spaces, we can compute how far about these priors are from each other. Aitchison demonstrated that the the Dirichlet distribution and logit-Normal are never equal for any choice of parameters; however, there are parameters choices that minimise the *Kullback-Leibler* (KL) divergence between them [4, 5]. The stick-breaking Polya-Gamma prior in less straightforward to work with than the Logit-Normal, but facilitates Gibbs sampling. Furthermore, the Logit-Normal transform preserves permutation symmetry in the density; while the stick-breaking transform does not preserve symmetry.

In light of similar analysis, we compute the KL divergence, defined below, between the two priors: the Gaussian Prior and the prior induced on this space by the inverse stick-breaking map from the Dirichlet prior. The KL divergence is

$$KL(P||Q) = \int_{\mathcal{X}} \log\left(\frac{dP}{dQ}\right) dP, \tag{C.43}$$

for probability measures P and Q defined on measurable space $\mathcal{X}$ and $\frac{dP}{dQ}$ the Radon-Nikodym derivative of P with respect to Q. Thus, we compute as follows, where, for ease of notation, we re-label the indexes, such that $vec(\alpha) = [\alpha_1, ..., \alpha_D]$ and likewise for ψ (with abuse of notation).

$$\begin{aligned} KL(p(\psi|\mu,\Sigma)||q(\psi|\alpha)) &= \int p(\psi|\mu,\Sigma) \log \frac{p(\psi|\mu,\Sigma)}{q(\psi|\alpha)} d\psi \\ &= \mathbb{E}[\log \mathcal{N}(\psi|\mu,\Sigma)] - \mathbb{E}[\log q(\psi|\alpha)] \\ &= (A) - (B), \end{aligned} \tag{C.44}$$

where the expectations are computed with respect to p. Continuing the computation

$$\begin{aligned}(A) &= \mathbb{E}\left[\log\left((2\pi)^{-D/2}|\Sigma|^{-1/2}\exp\left(\frac{1}{2}(\psi-\mu)^T\Sigma^{-1}(\psi-\mu)\right)\right)\right] \\ &= -\frac{D}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma| - \frac{1}{2}\mathbb{E}\left[tr\left((\psi-\mu)^T\Sigma^{-1}(\psi-\mu)\right)\right] \\ &= -\frac{D}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma| - \frac{1}{2}\Sigma^{-1}tr\left(\mathbb{E}\left[(\psi-\mu)(\psi-\mu)^T\right]\right) \\ &= -\frac{D}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma| - \frac{1}{2}D \\ &= -\frac{1}{2}\log((2\pi e)^D|\Sigma|),\end{aligned} \tag{C.45}$$

where in the second line we employed the trace trick and in the third line the linearity of the expectation. For part (B), we write

$$\begin{aligned}(B) &= \mathbb{E}\left[\log q(\psi|\alpha)\right] \\ &= \mathbb{E}\left[\log\left(\frac{1}{\mathcal{B}(\alpha)}\prod_{k=1}^{D-1}\sigma(\psi_k)^{\alpha_k}\sigma(-\psi_k)^{\sum_{j=k+1}^{D}\alpha_j}\right)\right] \\ &= -\log\mathcal{B}(\alpha) + \sum_{k=1}^{D-1}\mathbb{E}\left[\alpha_k\log(\sigma(\psi_k)) + \sum_{j=k+1}^{D}\alpha_j\log(\sigma(-\psi_k))\right] \\ &= -\log\mathcal{B}(\alpha) + \sum_{k=1}^{D-1}\alpha_k\mathbb{E}\left[\log(\sigma(\psi_k))\right] + \sum_{k=1}^{D-1}\sum_{j=k+1}^{D}\alpha_j\mathbb{E}\left[\log(\sigma(-\psi_k))\right] \\ &= -\log\mathcal{B}(\alpha) + \sum_{k=1}^{D-1}\alpha_k\mathbb{E}\left[\log(\sigma(\psi_k))\right] + \sum_{k=2}^{D}(k-1)\alpha_k\mathbb{E}\left[\log(\sigma(-\psi_k))\right]\end{aligned} \tag{C.46}$$

To compute the first summand, we expand the logistic function and then make a second order Taylor approximation about $x_0 = \mathbb{E}[x]$.

$$\begin{aligned}\sum_{k=1}^{D-1}\alpha_k\mathbb{E}\left[\log(\sigma(\psi_k))\right] &= -\sum_{k=1}^{D-1}\alpha_k\mathbb{E}\left[\log(1+e^{-\psi_k})\right] \\ &\approx -\sum_{k=1}^{D-1}\alpha_k\left(\log(1+e^{-\mathbb{E}[\psi_k]}) + \frac{e^{\mathbb{E}[\psi_k]}}{(1+e^{\mathbb{E}[\psi_k]})^2}\cdot\mathbb{V}(\psi_k)\right) \\ &= -\sum_{k=1}^{D-1}\alpha_k\left(\log(1+e^{-\mu_k}) + \frac{e^{\mu_k}}{(1+e^{\mu_k})^2}\cdot\Sigma_{kk}\right)\end{aligned} \tag{C.47}$$

Then, likewise for the second summand

$$\begin{aligned}\sum_{k=2}^{D}(k-1)\alpha_k\mathbb{E}\left[\log(\sigma(-\psi_k))\right] &= -\sum_{k=2}^{D}(k-1)\alpha_k\mathbb{E}\left[\log(1+e^{\psi_k})\right] \\ &\approx -\sum_{k=2}^{D}(k-1)\alpha_k\left(\log(1+e^{\mathbb{E}[\psi_k]}) - \frac{e^{\mathbb{E}[\psi_k]}}{(1+e^{\mathbb{E}[\psi_k]})^2}\cdot\mathbb{V}(\psi_k)\right) \\ &= -\sum_{k=2}^{D}(k-1)\alpha_k\left(\log(1+e^{\mu_k}) - \frac{e^{\mu_k}}{(1+e^{\mu_k})^2}\cdot\Sigma_{kk}\right)\end{aligned} \tag{C.48}$$

Hence,

$$\begin{aligned}KL(p(\psi|\mu,\Sigma)||q(\psi|\alpha)) \approx &-\frac{1}{2}\log((2\pi e)^D|\Sigma|) + \log\mathcal{B}(\alpha) \\ &+ \sum_{k=1}^{D-1}\alpha_k\left(\log(1+e^{-\mu_k}) + \frac{e^{\mu_k}}{(1+e^{\mu_k})^2}\cdot\Sigma_{kk}\right) \\ &+ \sum_{k=2}^{D}(k-1)\alpha_k\left(\log(1+e^{\mu_k}) - \frac{e^{\mu_k}}{(1+e^{\mu_k})^2}\cdot\Sigma_{kk}\right)\end{aligned} \tag{C.49}$$

To obtain a reasonable scale for the above result, we state the KL divergence between two Dirichlet distributions and two Gaussian distributions. Let us note that the KL divergence between two Dirichlet distribution is the following

$$\begin{aligned}KL(Dir(\pi|\alpha)||Dir(\pi|\alpha')) = &\log\Gamma(\alpha_0) - \sum_{k=1}^{K}\log\Gamma(\alpha_k) - \log\Gamma(\alpha'_0) \\ &+ \sum_{k=1}^{K}\log\Gamma(\alpha'_k) + \sum_{k=1}^{K}(\alpha_k-\alpha'_k)(\psi(\alpha_k)-\psi(\alpha_0)),\end{aligned} \tag{C.50}$$

where ψ denotes the digamma function. Likewise the KL divergence between two Gaussian distributions is the following

$$KL(p(x|\mu,\Sigma)||q(x|\mu',\Sigma')) = \frac{1}{2}\left(tr(\Sigma'^{-1}\Sigma) + (\mu'-\mu)\Sigma'^{-1}(\mu'-\mu) - K + \log\frac{|\Sigma'|}{|\Sigma|}\right) \tag{C.51}$$

C.15.7 Simulating dynamic spatial proteomics experiments

We describe the ways in which we produce produce synthetic dynamic spatial proteomics experiments from real dataset. The expression value for each protein can be written as follows:

$$y_i = f_k + \varepsilon_i \tag{C.52}$$

for some value $k = 1, ..., K$ which index the K possible subcellular niches. The value of f_k is unknown so we estimate it from the data. We use K-NN classification, with the number of nearest neighbours $\widehat{K} = 10$, to assign every protein to an organelle. That is, the probability the i^{th} protein belongs to the j^{th} organelle is approximated by:

$$P(z_i = j|Y = y_i) \approx \frac{1}{\widehat{K}} \sum_{l \in \mathcal{N}_i} \mathbb{1}(y_l = j), \tag{C.53}$$

where $\mathcal{N}_i$ is the set of $\widehat{K}$ closest labelled points to y_i. We then assign proteins to their most probable subcellular niche. We proceed to estimate f_k for $k = 1, ..., K$ by the mean of expression values of all the proteins allocated to that niche:

$$\hat{f}_k \approx \frac{1}{|n_k|} \sum_{i \in n_k} y_i, \tag{C.54}$$

where n_k indexes the proteins assigned to the k^{th} subcellular niche. We then use the residual bootstrap to generate synthetic data. To be precise, we first compute the residuals

$$\hat{\varepsilon}_i = y_i - \hat{f}_k \; i = 1, ..., N, \tag{C.55}$$

where k is the organelle to which protein i was assigned by K-NN classification. We then obtain $\mathcal{E} = \{\hat{\varepsilon}_{i,g}\}_{g=1}^{G}$, where G is length of the vector y_i. The we use a nonparamateric bootstrap (uniform sampling with replacement) to obtain $\mathcal{E}_B = \{\hat{\varepsilon}^*_{i,g}\}_{g=1}^{G}$. Replicates of the data are then obtained as follows

$$y_i^{rep} = \hat{f}_k + \hat{\varepsilon}^*_i \; i = 1, ..., N. \tag{C.56}$$

We further propose to use

$$y_i^{rep} = \hat{f}_k + \nu\hat{\varepsilon}^*_i \; i = 1, ..., N, \tag{C.57}$$

where ν is some deterministic or random value. In addition, we consider organelle specific multiplicative noise:

$$y_i^{rep} = \hat{f}_k + \nu_k\hat{\varepsilon}^*_i \; i = 1, ..., N, \tag{C.58}$$

where ν_k are different random values for $k = 1, ..., K$.

The above process produce replicates without any translocation events. To simulate translocation events we randomly select, with equal probability, L proteins. Then for each of these l proteins we randomly select, with equal probability, one of the K possible organelles to which we translocate the protein. Then we replace the quantitative value for l^{th} protein with a sample from the following distribution

$$y_l \sim \mathcal{N}(\hat{f}_k, \hat{\sigma}^2_{f_k}), \tag{C.59}$$

where k is the newly assigned organelle and $\hat{\sigma}^2_{f_k}$ is an unbiased estimator of the population variance of $\hat{f}_k$:

$$\hat{\sigma}^2_{f_k} = \frac{1}{|n_k| - 1} \sum_{i \in n_k} (y_i - \hat{f}_k). \tag{C.60}$$

Different spatial proteomics experiments are usually run on different mass-spectrometry runs and thus both random and systematic batch effects can occur. Furthermore, differences in the labelling efficiency of each tag, as well as slight differences in the amount of protein labelled and how well ions fly in the mass-spectrometer can lead to systematic difference between experiments. Furthermore, there is inherent technical variability in the apparatus and sample handling; for example, density-gradients or differential centrifugation speeds are never precisely the same. We propose three approaches to test the robustness of the available methods to these effects.

Random batch effects. After the replicates have been produced and translocation events simulated. We propose to generate random batch effects through the following process. For each replicate in turn, we sample a fraction with equal probability from $S_G = \{1, ..., G\}$. For the sampled fraction, say g, we add a random biased effect, μ_{batch}, to that fraction; such that,

$$y^{rep,batch}_{i,g} = y^{rep}_{i,g} + \mu_{batch} \tag{C.61}$$

Systematic batch effects. Systematic batch effects are produce in identical manner to random batch effects, but instead the fraction is sampled first and the effect is added to same fraction across the experiments. The magnitude of the effect is allow to differ across experiments.

Fraction permutations We permute the fractions in different experiments, which is designed to reflect the inherent technical variabilities of the procedure. Let $\sigma : S_G \to S_G$ be a permutation such that $\sigma(S_G) = \{\sigma(1), ..., \sigma(G)\}$. We then replace each fraction with its permuted value, as follows:

$$y^{rep,perm}_{i,g} = y^{rep}_{i,\sigma(g)}. \tag{C.62}$$

In the five possible simulation scenarios, which are all repeated 10 times, the following setting are used.

- $\nu_k \sim U[1, 2]$

For the systematic and random batch effects we take

- $\mu_{batch} = 0.3$