Challenges and opportunities for Bayesian statistics in proteomics

Oliver M. Crook * [1], Chun-wa Chung[2], and Charlotte M. Deane[1]

[1] *Department of Statistics, University of Oxford, Oxford, UK*
[2] *Structural and Biophysical Sciences, GlaxoSmithKline R&D, Stevenage, UK*

September 27, 2021

Abstract

Proteomics is a data-rich science with complex experimental designs and an intricate measurement process. To obtain insights from large datasets, statistical methodology and machine learning is routinely applied. For a quantity of interest, many of these approaches only produce a point estimate, such as a mean, leaving little room for bespoke interpretations. In contrast, Bayesian statistics quantifies uncertainty using probability distributions. These probability distributions allow scientists to ask complex questions of their proteomics data which would otherwise be challenging using alternative approaches. Bayesian statistics also offers a modular framework for specifying complex hierarchies of parameter dependencies. This allows us to use statistical methodology which equals, rather than neglects, the sophistication of experimental design and instrumentation present in proteomics. Here, we review Bayesian methods applied to proteomics and argue for a broader uptake, whilst also highlighting the challenges posed by adopting a new statistical framework. To illustrate our review, we present a walk-through of the development of a Bayesian model for dynamic organic orthogonal phase-separation (OOPS) data.

1 Introduction

Decision making spans the entire research process. Ultimately, it is a choice to believe an explanation for a phenomena given the current evidence. For some theories, the evidence is overwhelming: careful mechanistic experiments and verifiable model predictions have never contradicted that theory[106]. This scenario is, however, rare. In practice, we make decisions under uncertainty and the evidence is not clear-cut[101]. Bayesian statistics allows us to

*oliver.crook@stats.ox.ac.uk

make inferences from that evidence to enable decision making in those cases[35]. In contrast to *frequentist* methods, Bayesian inference allows us to use probability to model degrees of belief, rather than just frequencies. Consequently, models that are consistent with the available evidence are more probable and incompatible models are less probable[37,87]. By using probability theory in this manner, there is a recipe for taking *prior beliefs* (i.e. information encoded by domain expertise) and updating them to *posterior beliefs* using observed data [35]. As a result, this posterior probability distribution quantifies the models compatible with domain expertise and our experimental data[37]. This recipe is known more formally as *Bayes' theorem*.

Mass-spectrometry-based proteomics is a complex scientific field[2]. The techniques versatility allows us to explore differential abundance[2], protein turnover[66], interactions[46], thermal stability[65], structure[64,89], spatial information[3,33,38] and more[48,74,99]. In each case, data are manipulated, thresholded and filtered so that a statistical test or machine learning algorithm can be applied. The results are then frequently concluded with a single value, which we have granted the role of arbiter of truth. These decisions are often made without consideration of what we might be happening at each step. Bayesian statistics could propagate or quantify the uncertainty in these steps, replace implicit or ad-hoc approaches with explicit models and summarise the output with a probability distribution consistent with our data[37]. This paradigm progression not only provides us an ability to ask new questions of our data but a consistent way to perform inference and criticize our models[35,37].

Bayesian statistics offers proteomics considerable possibilities; despite that, it has not been readily adopted in the community. This may stem from a lack of familiarity, a lack of awareness of available tools, complex language, impenetrable literature, inability to communicate results from an analysis and computational difficulties. Here, we review the contribution Bayesian statistics has already made to proteomics, clarify the Bayesian workflow and how it can be applied to proteomics, highlight a number of modelling strategies and outline current challenges for the community. Throughout, we illustrate our analysis with examples from the proteomics literature, focusing on building a model for dynamic organic orthogonal phase separation data[80].

2 Main

2.1 Bayes, in brief

Before we review the contributions Bayesian statistics has already made to proteomics, we introduce the fundamental technical background and notation. We use $P(E)$ to denote the probability of the event E. E can be anything from "it rains tomorrow" to "my parameter falls between the values a and b". We let D be notation for the observed data, for example from a shotgun proteomics experiment. Let x be a data point from D, such as a measurement for a particular protein. We assume that x arises from some probability distribution p

and we write $x \sim p(x|\theta)$, for example this could be a log normal distribution. Let α be hyperparameters of the parameter distribution, such that θ themselves are drawn from a probability distribution $\theta \sim p(\theta|\alpha)$. For example, the mean of the log normal distribution could be drawn from a normal distribution.

The *prior distribution* captures our domain expertise and is the distribution of the parameters before any data is observed: $p(\theta|\alpha)$. The *prior* could capture, say, that abundance values are positive and are unlikely to exceed the number of grains of sand on Earth. The sampling distribution is the distribution of the data given the parameters $p(D|\theta)$, we can write this as a function $L(\theta|D)$ called the *likelihood*. If we average (or *marginalise*) the distribution of the data over the parameters, we obtain the so-called marginal likelihood:

$$p(D|\alpha) = \int L(\theta|D)p(\theta|\alpha)\,\mathrm{d}\theta. \tag{1}$$

The posterior distribution of the parameters is determined by Bayes' theorem, as the following:

$$p(\theta|D,\alpha) = \frac{p(D|\theta,\alpha)p(\theta|\alpha)}{p(D|\alpha)}. \tag{2}$$

Bayes' theorem tells us the mathematical way to update beliefs in light of evidence: simply multiply our prior and likelihood and renormalise by the marginal likelihood. Bayes' theorem implies a self-consistency property: the posterior averaged over the data returns the prior[95]:

$$p(\theta) = \int\int p(\theta|\tilde{y})p(\tilde{y}|\tilde{\theta})p(\tilde{\theta})\,\mathrm{d}\tilde{y}\mathrm{d}\tilde{\theta}, \tag{3}$$

where $\tilde{y} \sim p(D|\tilde{\theta})$. When Bayesian's predict, instead of simply taking a single parameter value forward, they make predictions by averaging:

$$p(\tilde{x}|D,\alpha) = \int p(\tilde{x}|\theta)p(\theta|X,\alpha)\,\mathrm{d}\theta. \tag{4}$$

In summary, Bayesian statistics provides us with a distribution of plausible parameter values from the *posterior* and a distribution of hypothetical predicted values from the *posterior predictive distribution*. We can then ask bespoke question of these probability distributions; for example, $P(\theta > 2|D,\alpha) = \int_2^\infty p(\theta|D\,\alpha)\,\mathrm{d}\theta$ is the probability that a parameter is greater than 2. For proteomics, these could be the probability that a fold-change exceeded a certain value or a the probability that a spectra belongs to a particular peptide. One possible interpretation of the prior is as a penalty that regularises the parameters, many applications exploit this and so can apply bespoke regularisation to their model.

2.2 Bayesian contributions to proteomics

2.2.1 Bottom-up proteomics and differential abundance

Here, we highlight Bayesian approaches already applied to proteomics. We focus on proteomics data generated via mass-spectrometry but refer to contributions for the reverse-

phase-protein-array (RPPA) literature[21,61,73] and 2D gel electrophoresis[71]. A number of approaches have been aimed at quantification and differential abundance analysis of bottom-up proteomics data[13,68,75,78,79,86,90,91,97,98]. Carvalho *et al.* [13] appreciate that Bayesian statistics can be used to improve analysis of spectral counting data using a Poisson likelihood and calculating the probability of detecting a protein in a particular sample. However, they do not exploit the full Bayesian toolkit by working with probability distributions and simply end-up interpreting their probabilities as p-values. Thus, it is not clear whether their approach is any better than simply using a likelihood-based approach. A number of approaches[78,90,97,98] argue for propagating and quantifying the uncertainty in the analysis rather than simply the underlying quantitation values, either through relaxing the parsimony assumption[90], including ion statistics via a two-level Beta-Binomial model[78], or jointly modelling identification and quantitation statistics[97,98]. Millikin *et al.* [68] argue to use an analogue of the t-test and exploit the posterior distribution by using an interval thresholding approach. O'Brien *et al.* [75] note the inherent compositional nature of labelled proteomics approaches and include a modelling parameter to model ratio compression allowing better estimation of the true fold changes. Importantly this parameter is shared across all proteins, which allowed them to estimate the parameter accurately even for proteins with few observed peptides. Jow *et al.* [49] also model isobaric laballed mass-spectrometry data but do not model ratio compression. Meanwhile for label-free experiments, O'Brien *et al.* [76] explicitly model missingness showing that jointly modelling missingness and abundance leads to improved performance. All of these approaches demonstrate some benefit over previously applied approaches suggests that combining these method would provide further improvements. It also suggests translating these methods to other proteomics techniques would be a fruitful endeavour. However, many of the methods differ in a number of key modelling choices and it is not always clear how these choices were made.

2.2.2 Protein and peptide identification

One of the fundamental problems in mass-spectrometry based proteomics is identifying a peptide from a spectra. A spectra can be very noisy and $b-, y-$ ions can be missing which results in a complex observation process. Furthermore, we have prior knowledge of observing particular amino acid sequences and knowledge of the cleavage process. This appears an ideal scenario for application of Bayesian methods. Indeed, a number of approaches have been applied[15,19,42,55]. Chen *et al.* [15] use a fairly simple framework to calculate peptide identification probabilities based on peptide coordinates. Halloran *et al.* [42] employ a dynamic Bayesian network in a method called DRIP, allowing for insertions and deletions to the spectra. By modelling possible alignments between theoretical and observed spectra they can calculate the most probable peptide match. The authors find their approach improves over available methods particularly for low resolution MS2 data. Lewis *et al.* [55] take a different approach to the same problem, incorporating a scoring function into a likelihood model. Their model also allows deletions directly via indicator functions. Insertions are

characterised by excessive deviations from the spectra of the candidate peptide, where excessive is characterised probabilistically via laplace noise. The authors also include prior information about possible cleavage pairs, as well as prior information about the probability of observing a particular peptide sequence in the dataset. Finally, in constrast to Halloran *et al.* [42] they make full use of Bayesian methods and provide a posterior distribution over possible peptides and parameters. This could allow multiple peptides to be associated to a spectrum with differing certainty which could be used in downstream analysis. Claassen *et al.* [19] tackle a slightly different problem and use a non-parametric Bayesian model to predict the coverage in sequential LC-MS/MS experiments but suggest their approach could also be adapted to database searching and de novo sequencing. Again, these approaches all have benefits over previously applied methodology. The clearest is the inclusion of more information and the ability to provide a flexible, and well rationalised, model to the underlying data. The ability to exploit uncertainty captured by the posterior distribution for downstream analysis is far more insightful that simply point estimates from a Bayesian analysis. However, these approaches have not been adopted by the community. This could be because the methods are difficult to apply, the benefits are not compelling or computation is excessive.

2.2.3 Proteoforms and post-translation modifications

A number of approaches are interested in applications to proteoform analysis (splice isoforms) or post-translational modifications[18,57,62,93,105]. Chung *et al.* [18] employ a non-parametric mixture model to jointly model the modification mass for each PTM group and the true (unobserved) location of the modified amino acid. Their approach outperforms other approaches, is fully automated and provides modification confidence scores. However, the approach does not model the underlying spectrum, which could result in unnecessary false positives. Webb-Robertson *et al.* [105] tackle the proteoform problem by deconvolving peptides into signatures even if they are associated with the same protein. However, they only use a Bayesian point estimate rather than exploiting the full posterior distribution. Lim *et al.* [57] use a Bayesian model to estimate the phosphorylation stoichiometry using Bayesian statistics. By incorporating a physically plausible model they remove problems with previous models that could allow negative stoichiometry. Their joint model allows them to borrow power across replicates and they report downstream uncertainty. Shteynberg *et al.* [93] use a Bayesian mixture model to compute compute probabilities for modification sites. This allows them to combine precomputed scores in a rational way but again they do not examine the full posterior distributions. Mallikarjun *et al.* [62] employ a Bayesian linear regression modelling strategy to analyse differential PTM data, suggesting their approach outperformed other methods and could allow uncertainty in missing values. The main benefit here appears to be the regularisation of the parameters using priors rather than specifically the uncertainty quantification in the analysis.

2.2.4 Biomarkers and clinical proteomics

Protein biomarkers, molecular indicators of aberrant processes or disease, and clinical proteomics are a key component of proteomics research. Hernández *et al.* [44] provide a review of Bayesian method development in biomarker development and we refer to them for additional details. Morris *et al.* [69,70] develop Bayesian wavelet-based functional mixed models for mass-spectrometry-based proteomics data. Their advanced framework, allows the simultaneous use of nonparametric fixed and random effects, which facilitates adjustment for clinical and experimental covariates that could affect the intensity and location of a spectra. Working with posterior distributions they are able to compute important quantities such as the probability of intensity changes for fixed fold levels and are able to control a Bayesian false discovery rate. Liao *et al.* [56] combine the above framework with image analysis methods to enable biomarker discovery from LC-MS data. Hwang *et al.* [47] develop a pipeline, MS-BID, for biomarker analysis which uses a Bayesian anova. Harris *et al.* [43] apply a Bayesian hierarchical linear probit regression model to determine discriminative biomarkers from mass spectrometry data. They find their approach improves over a K-nearest neighbour method. Furthermore, by using posterior probabilities they are able to determine which samples will be the most promising for prognostics. Kuschner *et al.* [52] propose a Bayesian network to perform feature selection from mass-spectrometry data. The selected features then provided excellent predictive power. Though again this approach still uses a Bayesian point estimate and could have obtained more information by computing full posterior distributions. Deng *et al.* [27] develop a Bayesian network which allows them to integrate mass-spectrometry and microarray data, allowing them to borrow power between mRNA and protein levels. Here Bayesian statistics is sufficiently flexible to incorporate different modalities and weigh up the uncertainty between different datasets. More recently Liu *et al.* [58] developed Bayesian Function-on-Scalar quantile regression for mass-spectrometry data. This approach notes that biomarker difference may not be apparent at mean regression but rather at a particular quantile (such as the 0.95 quantile). This method simultaneously accounts for the functional nature of MALDI-TOF data, incorporates prior knowledge for adaptive regularization and a basis representation which allows borrowing of power. They find their method identifies biomarkers overlooked by mean regression. Bayesian methods for biomarker and clinical proteomics are more developed than other examined proteomics sub-fields with several exemplary methods that make full use of the flexibility of Bayesian modelling and the rich output of the posterior distribution.

2.2.5 Chromatography

To facilitate identification in mass-spectrometry a liquid-chromatography step is usually applied. The time at which a peptide elutes from the liquid chromatography, called the retention time, can be used as additional information to help identify peptides. However, there is uncertainty in this retention time and they can vary from one run to another. Chen *et al.* [14] develop a Bayesian model called DART-ID, which models a latent (unobserved) global

retention time alignment. This alignment allows them to combine the outputted posterior error probability of MaxQuant with the inferred RT density in each experiment. Hence, by using this result they can update their confidences and improve coverage in experiments by 50%. Whilst their approach appears powerful, they only use a point estimate and obtain uncertainty through bootstrapping. Though the author reference computational challenges it would have been useful to examine the posterior distribution so we could be explicit about the computational trade-offs. Maboudi Afkham *et al.* [60] are interested in the uncertainty in peptide retention time methods. Using a Gaussian process regression method they were able to accurately predict retention times and obtain uncertainty estimates. They then use the posterior distribution from the regression analysis as a variable retention time window to identify potentially incorrect peptides. This improves over fixed windowing strategies. One strategy they overlooked, in a similar vain to DART-ID, would be to update the identification probabilities based the deviation probability from the predicted retention time. This approach naturally fits within a Bayesian framework.

2.2.6 Intact, top-down and structural proteomics

Proteoform analysis is one of the key challenges in top-down proteomics. LeDuc *et al.* [54] introduce a C-score, not be confused with a C-statistic, to facilitate automated identification and characeterisation of proteoforms from top-down proteomics data. Ultimately, their approach allows them to rank probable proteoforms having observed their data. Performing an analysis in a Bayesian framework allows them to specify a generative model, provide expert prior information and carefully model the underlying noise distribution. Their proposed C-score is essentially a transformed posterior error probability. However, despite their Bayesian framework, they opt for a point estimate of their model, which could have been greatly enhanced by examining the full posterior of their model. Marty *et al.* [63] proposed a Bayesian deconvolution algorithm for Ion Mobility spectra, and extended in Kostelic *et al.* [51]. Their approach allows the convolution of the charge distribution with the peak shape to obtain a flexible deconvolution approach. The extent of their applications is extensive, demonstrating a clear benefit of their method. However, their approach also uses a point estimate from their analysis. Indeed, the idea of a posterior distribution is not mentioned at all within the paper. Hence, apart from the use of prior information, it is not clear what particular benefit a Bayesian analysis has for their approach. Saltzberg *et al.* [84] propose a Bayesian model to resolve residue level information from hydrogen-deuterium exchange mass-spectrometry. They choose uninformative priors, and though they perform inference using Monte-Carlo methods they do not use the posterior distribution. Furthermore, they do not justify why there model allows for negative deuterium incorporation, which maybe arises from a misunderstanding of the positivity constraint induced by their exponential likelihood model.

2.2.7 Functional proteomics

Functional proteomics methods aim to decipher protein-function on a system-wide scale. One approach is spatial subcellular proteomics[17,33] where protein are localised to their subcellular niche using mass-spectrometry data. Bayesian approaches have been developed for biochemical fractionation-based subcellular proteomics[20,21,22,23,24]. Crook *et al.* [20, 21, 22] demonstrate Bayesian modelling can quantify uncertainty in protein subcellular localisation and identify cases where this may correspond to multi-localising proteins. Crook *et al.* [20] show that a even a Bayesian point estimate may overlook these cases and more information is obtained by examining the full posterior distribution. Crook *et al.* [23] allow the uncertainty in the number of subcellular niches to be accounted for and show that allowing additional niches can be uncovered. However, the model appears sensitive the prior choices and should be chosen carefully. Crook *et al.* [24] build on these experiments to analyse differential localisation experiments showing that modelling uncertainty improves power and interpretation compared with other methods. This fully Bayesian analysis; however, is computationally intensive as it attempts to model many datasets at once. Another functional approach is AP-MS, which allows us to determine protein interactions and complexes[17]. Choi *et al.* [16] develop a non-parametric Bayesian model to bi-cluster AP-MS data. They sample from the posterior distribution and are hence able to report the uncertainty in the clustering. However, their nested model assumes that the conditional on the Bait cluster the Prey clusters are independent and their model assumes exchangability (permutation leads to the same probability distribution) of the rows and columns. Fang *et al.* [28] propose a semi-parametric model for thermal protein profiling after identifying proteins that deviate from classic sigmoid behaviour. Semi-parametric models combine interpretable parametric models with more flexible non-parametric models. Using Bayesian analysis they can critically assess the semi-parametric and parametric model fits and demonstrate those that are better modelled by the semi-parametric model share functional enrichments. Again this fully Bayesian approach has demanding computational requirements.

2.3 The Bayesian workflow

2.3.1 Motivating example

To illustrate the Bayesian workflow, we examine some recently introduced proteomics data generated using the orthogonal organic phase separation (OOPS) method of Queiroz *et al.* [80]. This method is able to efficiently enrich for RNA-binding proteins and hence, by adapting to the dynamic setting, is able to examine differential RNA binding. This is where the proportion of a particular protein bound to RNA changes depending on the condition. Here, we examine such an experiment where thymidine-nocodazole was used to induce cellular arrest. Total and oops-enriched protein abundances where obtained at $0, 6$ and 23 hours post treatment. Each experiment was performed in triplicate, except for at 6h when four replicates were taken. The 10 total and 10 oops sample where labelled using 2 separate TMT

10-plex kits and quantitative mass spectrometry was performed in two runs using SPS-MS3 on an Orbitrap Fusion Lumos. Here, we attempt to use the Bayesian toolkit to model this data and answer question about change in RNA-binding. A heatmap of the data is shown in figure 3. A protein was chosen at random for the modelling illustration, NCAPD2, a regulatory subunit of the condensin complex, a protein known to have differential subcellular localisation throughout the cell-cycle[88].

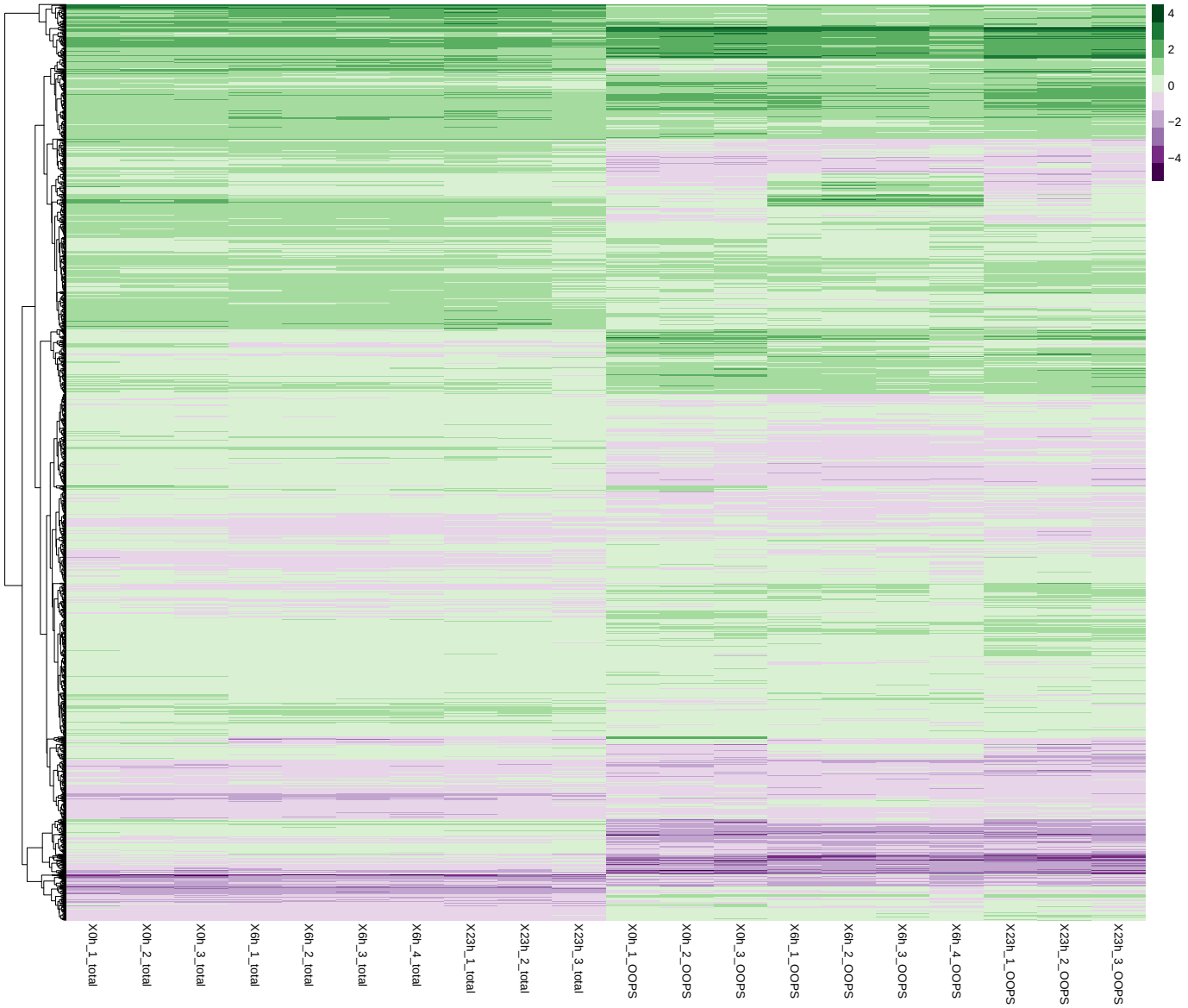

Figure 1: **Exploratory data analysis of OOPS.** A heatmap of the mass-spectrometry data generated by the oops experiment

2.3.2 Generative modelling

Having highlighted the successes and limitations of some of the contributions of Bayesian methods to mass spectrometry-based proteomics, we clarify the Bayesian workflow to facilitate it for proteomics. The first tension of Bayesian analysis is the pairing of the likelihood and the prior[7,36,37]. On one hand, the word *prior* suggests it must be chosen first; however, without knowledge of the likelihood it makes little sense to start selecting priors - we may not even know the parameters of the model. Thinking of the likelihood and prior as a pair reduces this conceptual tension. It also leads to an explicit way to check our modelling assumptions via generative and predictive modelling[7]. A generative model generates data consistent with the data. The prior has good predictive properties if the *posterior predictive distribution* can predict new data generated from similar experiments. To be explicit, given a likelihood and prior, we can simulate data y. First, sample the parameters of the likelihood from the prior and then given these parameters sample data from the model:

$$\begin{aligned} \tilde{\theta} &\sim p(\theta|\alpha) \\ \tilde{y} &\sim p(y|\tilde{\theta}). \end{aligned} \tag{5}$$

This leads us to define the *prior predictive distribution*:

$$p(\tilde{y}|\alpha) = \int_{\theta} p(\tilde{y}|\theta)p(\theta|\alpha)\,\mathrm{d}\theta. \tag{6}$$

There are a number of key observations. Firstly, the prior predictive distribution has no knowledge of the data, aside from the modelling assumptions of the domain expert. Secondly, the likelihood and prior are now explicitly coupled and so poor modelling choices in either the likelihood or prior will be apparent via the prior predictive. Thirdly, the failure of uniform or uninformative priors as a default is clear, they will generate unrealistic data.

In our oops example, we model log protein abundance as a linear model of sample type (whether total or oops) and time $(0, 6, 23\text{h})$. Since we are interested in changes in the proportion of protein bound to RNA, we include a interaction effect between time and sample type. We then use Gaussian priors on the coefficients of the effects and an exponential prior on the standard deviation of the Gaussian noise. Formally, the model can be written as

$$\begin{aligned} \log y &= \beta_{type} + \beta_{time} + \beta_{time:type} + \epsilon \\ \epsilon &\sim \mathcal{N}(0, \sigma^2) \\ \beta &\sim \mathcal{N}(0, 1) \\ \sigma &\sim \mathrm{Exp}(1). \end{aligned} \tag{7}$$

The priors were chosen fairly arbitrarily and we can use a prior predictive check to see whether this is a sensible generative model. Figure 2 show a variety of prior predictive checks using different summaries of our observed and simulated data. We see that the our generative model is too diffuse compare with the observed data and produces excessive large deviations

beyond what we would expect from a typical proteomics dataset. In the accompanying vignette, we show that are inferences can be better calibrated using an exponential prior with rate 4, which corresponds to 1 in 5000 proteins having a standard deviation in their log abundance above 2.5.

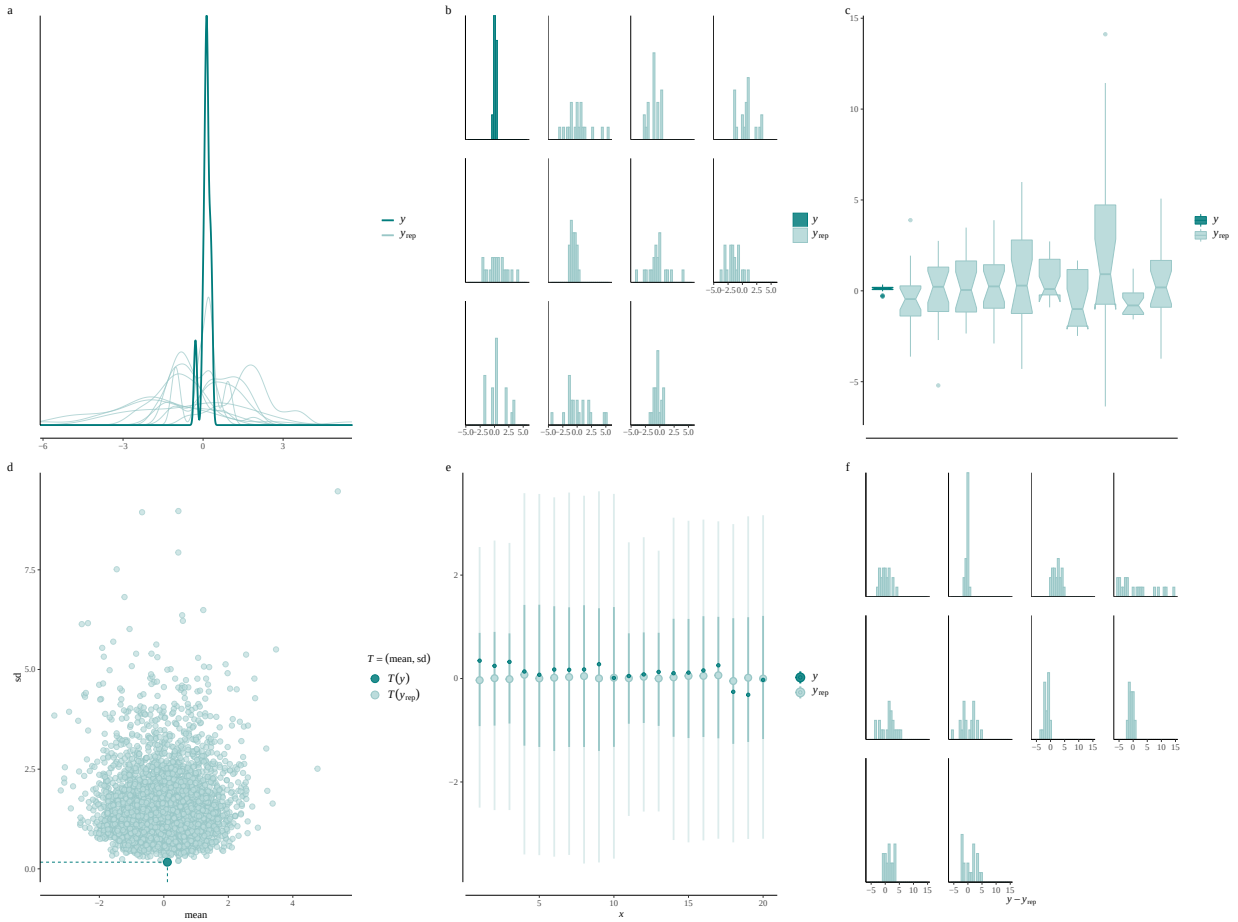

Figure 2: **Prior predictive checks.** Prior predictive checks applied to oops data. y denotes the observed data, whilst y_{rep} denotes the simulated data from the prior predictive distribution. (a) kernel density estimation based checks (b) Histogram based checks (c) boxplot based checks (d) summary statistics checks (e) interval plot based checks (f) error histogram based checks

2.3.3 Predictive modelling

Once our prior and likelihood have seen the data, D, they are updated into the posterior distribution. We can then sample new data by first sampling parameters from the posterior distribution and then again sampling from the likelihood:

$$\begin{aligned} \tilde{\theta} &\sim p(\theta|D, \alpha) \\ \tilde{y} &\sim p(y|\tilde{\theta}). \end{aligned} \tag{8}$$

This leads to the definition of the posterior predictive distribution:

$$p(\tilde{y}|D,\alpha) = \int_\theta p(\tilde{y}|\theta)p(\theta|D,\alpha)\,\mathrm{d}\theta = \int_\theta p(\tilde{y}|\theta)\frac{p(D|\theta,\alpha)p(\theta|\alpha)}{p(D|\alpha)}\mathrm{d}\theta. \tag{9}$$

We have expanded integrand using Bayes theorem to make a key point explicit: the posterior predictive distribution depends on the likelihood, the prior and the data. This coupling allows us to make a number of observations. A good choice of prior and likelihood leads to good predictive performance and over-fitting can be examined via the posterior predictive distribution.

Having fitted the model to the data, we can perform a posterior predictive check on our inferences. Figure 3 shows a number of posterior predictive checks and that clearly the model has learnt from the data. Visualisation show that samples from the posterior predictive distribution look similar to the observed data. Contrast this with prior predictive checks where the samples form the distribution were very diffuse.

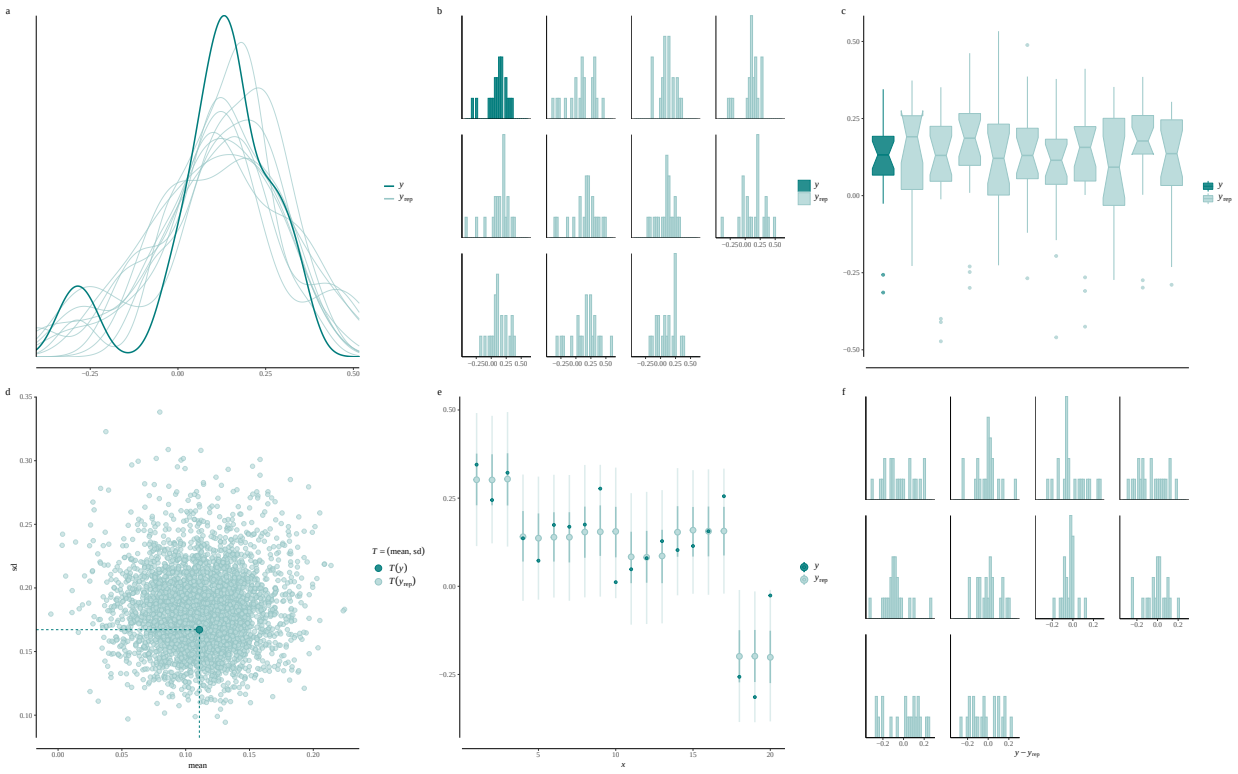

Figure 3: **Posterior predictive checks.** Posterior predictive checks applied to oops data. y denotes the observed data, whilst y_{rep} denotes the simulated data from the prior predictive distribution. (a) kernel density estimation based checks (b) Histogram based checks (c) boxplot based checks (d) summary statistics checks (e) interval plot based checks (f) error histogram based checks

2.3.4 Fitting a model: Bayesian computation

In practice, the integrals and probability distribution required for sufficiently flexible modelling are intractable. We can perform inference in a wide array of models using Markov-chain Monte Carlo (MCMC) methods[10,39], including Gibbs sampling[34,94], Metropolis sampling[82], and Hamiltonian Monte Carlo[45]. Bayesian inference an also be performed using sequential Monte Carlo[26] or variational inference[9]. Although the latter can provide a fast approximation of the posterior distribution, it can be arbitrarily inaccurate. Here, we focus on Hamiltonian Monte Carlo, as it forms the basis of modern probabilistic programming languages[12]. Initially, when an MCMC algorithm begins it will "move" towards the posterior distribution producing a "sample" at each iteration. An initial warm-up or burn-in section is require to remove bias due to dependence of the algorithms starting values and to adapt some of the algorithms tuning parameters to provide efficient inference. Once the warm-up section is complete, there is a sampling period which is run until multiple chains have mixed. One measure of mixing chains is $\hat{R}$, which is essentially a measure of between and within chain variance[104]. Current standard practise is that $\hat{R}$ should be close to 1. It is also recommend to visualise trace plots and rank histograms for samples from an MCMC algorithm[31]. Some tools included further diagnostic checks such as divergences but this is beyond the scope of this review[6]. Table 1 highlights some probablistic programming languages that can be used to fit general purpose Bayesian models. For our oops example Bayesian computations are reliable, see accompanying vignette and the supplement.

Packages for Bayesian computation			
Computational tool	language	Inference method	Reference
stan	c++	HMC variant	12
brms	R	HMC variant	11
MCMCglmm	R	Metropolis/Slice sampling	41
PyMC3	Python	HMC variant	85
BUGS	BUGS/R	Gibbs sampling	59
Edward	Python	Various including variational inference	100
Pyro	Python	HMC variant	8
Turing.jl	Julia	Various including HMC	32

Table 1: **General purpose probabilistic programming languages.** A variety of probabilistic programming languages are available in several languages using modern and efficient inference methods. Amongst these languages, one can fit the vast majority of models used in practice.

2.3.5 Posterior z-scores and contraction

It is often desirable to evaluate the behaviour of a model, and if any model assumptions are preventing us from making sensible inferences. The *posterior z-score* and *posterior contraction* are useful metrics to identify several problems with a model[7]. Let's assume, we have access to a parameter,θ^*, of the true data generating process. The *posterior z-score* for a parameter is defined as:

$$z_{\text{post}}(\theta|\tilde{y}, \theta^*) = \frac{E_{\text{post}}[\theta|\tilde{y}] - \theta^*}{s_{\text{post}}(\theta|\tilde{y})}, \tag{10}$$

whilst the *posterior contraction* is defined as

$$c(\theta|\tilde{y}) = 1 - \frac{V_{\text{post}}(\theta|\tilde{y})}{V_{\text{prior}}(\theta|\tilde{y})}. \tag{11}$$

Together these quantities tell us about how the posterior is learning from the data. If the posterior z-score is large and the posterior contraction is small, then the prior modelling conflicts with the true process - we are unable to learn the true parameter well. If the posterior z-score is large and the posterior contraction is close to 1, this suggest we are concentrating on an incorrect part of the probability space and so the model is over-fitting. If the posterior z-score is small and the posterior contraction is also small then the model is poorly informed by the data. The ideal scenario is that posterior contractions are close to 1, whilst having posterior z-scores that are to 0. This tells us that the data is highly informative and the prior was not biased away from the data generating mechanism. Examples of posterior contractions are shown in the accompanying vignette.

2.3.6 Model selection and averaging

Using probability allows us to select between competing models that may generate the data. Given two models, $\mathcal{M}_1$ and $\mathcal{M}_2$, we can ask for the $P(D|\mathcal{M}_i)$ for $i = 1, 2$. The relative plausibility of two models is referred to as the *Bayes factor*[50],

$$\text{BF}_{12} = \frac{P(D|\mathcal{M}_1)}{P(D|\mathcal{M}_2)} = \frac{P(\mathcal{M}_1|D)}{P(\mathcal{M}_2|D)} \frac{P(\mathcal{M}_2)}{P(\mathcal{M}_1)}. \tag{12}$$

The Bayes factor allows an interpretable and quantitative way to evaluate the relative plausibility of two models, by examining the ratio of the probabilities of that model generating the data. From a brief calculation, we can see that:

$$P(D|\mathcal{M}_1) = \int p(D|\theta_1, \mathcal{M}_1) p(\theta_1|\mathcal{M}_1) \, \mathrm{d}\theta_1, \tag{13}$$

where θ_1 are parameters that parametrise model $\mathcal{M}_1$. Here, we see the dependence of the Bayes Factor on the prior and the implicit assumption that we are evaluating models on their prior predictive performance becomes explicit. Thus, using improper/uninformative priors with Bayes factor would be inappropriate. However, there a further complexities, the most

alarming perhaps is that one can inflate the Bayes factor by simply choosing a prior that places probability on unrealistic parts of the parameter space. Typically a uniform prior would have such an effect. If you are not willing to rigorously defend your prior choices, model evaluation may be better using functions of the posterior predictive distributions[7].

We have already seen that one of the key mechanics of Bayesian statistics is the ability to average over quantities, rather than simply taking the best parameters forward. This can also be performed with models using so-called Bayesian model averaging[81]. Let ϕ be a quantity of interest and given models, $\mathcal{M}_i$ $i = 1, ..., n$, we may average them:

$$p(\phi|D) = \sum_{i=1}^{n} p(\phi|D, \mathcal{M}_i)p(\mathcal{M}_i|D). \tag{14}$$

This is simply the average of the posterior predictive distribution for ϕ under the models considered, weighted by their posterior model probability. If we are simply interested in the Bayesian model average estimate of a particular parameter, we can compute

$$\hat{\theta} = E_{\text{BMA}}[\theta] = \sum_{i=1}^{n} E_{\mathcal{M}_i}[\theta_i]p(\mathcal{M}_i|D). \tag{15}$$

Given the sensitivity of the Bayes factor to the prior, it sometimes useful to consider model selection based on the posterior predictive distribution. One example is the log pointwise predictive density (lpd)[103]:

$$\text{lpd} = \sum_{i=1}^{n} \log p(y_i|D) = \sum_{i=1}^{n} \log \int p(y_i|\theta)p(\theta|D)\,\mathrm{d}\theta. \tag{16}$$

It is frequently useful to consider an out-of-sample predictive fit via leave-one-out (LOO) cross-validation[103]:

$$\text{lpd}_{\text{LOO}} = \sum_{i=1}^{n} \log p(y_i|D_{-i}), \tag{17}$$

where D_{-i} is data without data point i. This quantity can be efficiently approximated using the LOO package[103]. We note that the above definitions can be adapted to any utility or loss function so that the metric of interest can be characterised.

Returning to our oops example, in our second vignette we develop more complex models of the data. These models include group-level (random) effects for the replicate number and TMT tag used (see model strategies section). However, it is not clear that these models are actually more plausible. To clarify the three competing models are
Model 1:

$$\begin{aligned}
\log y &= \beta_{type} + \beta_{time} + \beta_{time:type} + \epsilon \\
\epsilon &\sim \mathcal{N}(0, \sigma^2) \\
\beta &\sim \mathcal{N}(0, 1) \\
\log \sigma &= \beta_{\sigma,time} + \beta_{\sigma,type} \\
\beta_\sigma &\sim \mathcal{T}(3, 0, 1).
\end{aligned} \tag{18}$$

Model 2:

$$
\begin{aligned}
\log y &= \beta_{type} + \beta_{time} + \beta_{time:type} + u_{replicate} + \epsilon \\
\epsilon &\sim \mathcal{N}(0, \sigma^2) \\
\beta &\sim \mathcal{N}(0, 1) \\
\log \sigma &= \beta_{\sigma,time} + \beta_{\sigma,type} \\
\beta_\sigma &\sim \mathcal{T}(3, 0, 1) \\
u_{replicate} &\sim \mathcal{N}(0, \sigma^2_{replicate}) \\
\log \sigma_{replicate} &\sim \mathcal{T}(3, 0, 0.1)
\end{aligned} \tag{19}
$$

Model 3:

$$
\begin{aligned}
\log y &= \beta_{type} + \beta_{time} + \beta_{time:type} + u_{replicate} + u_{TMT} + \epsilon \\
\epsilon &\sim \mathcal{N}(0, \sigma^2) \\
\beta &\sim \mathcal{N}(0, 1) \\
\log \sigma &= \beta_{\sigma,time} + \beta_{\sigma,type} \\
\beta_\sigma &\sim \mathcal{T}(3, 0, 1) \\
u_{replicate} &\sim \mathcal{N}(0, \sigma^2_{replicate}) \\
u_{TMT} &\sim \mathcal{N}(0, \sigma^2_{TMT}) \\
\log \sigma_{replicate} &\sim \mathcal{T}(3, 0, 0.1) \\
\log \sigma_{TMT} &\sim \mathcal{T}(3, 0, 0.1)
\end{aligned} \tag{20}
$$

Each of the models progresses with more complexity. We compute the posterior model probabilities for each of these examples and find that $P(\mathcal{M}_1|D) = 0.35$, $P(\mathcal{M}_2|D) = 0.48$ and $P(\mathcal{M}_3|D) = 0.17$ (see vignette for more details). This suggest that a group-level effect for replicate is warranted but there is less support for the more complex model 3. Note that because computing the posterior model probabilities includes integration against the prior, these probabilities are automatically penalised for model complexity. See accompanying vignette for further exploration.

2.3.7 Using uncertainty from a Bayesian analysis

Bayesian's quantify uncertainty using probability distributions. Perhaps the most commonly used representation of uncertainty is the credible interval[35]. A credible interval is an interval (a, b) such that a parameter lies within this interval with some probability. For example, the we could ask for interval such that the probability that a protein's log abundance falls between a and b with probability 0.95. In notation used earlier $P(a < \log x < b) = 0.95$. We can see that the interval (a, b) is not unique.

The analogous quantity in frequentist statistics is the confidence interval; however, it is an entirely different concept. This is seen most clearly by asking which part of the constructions are random. For credible intervals, it is the quantity of interest θ that is random and the

interval that is a fixed quantity. Whilst, for a confidence interval the parameter is fixed and the interval is random, since it depends on the randomly observe sample.

However, Bayesian's can report any quantity that can be derived from the posterior distribution or posterior predictive distribution, which in practice can very complex representation of uncertainty. Since summarisation can distort the representation of uncertainty, we recommend reporting the full posterior distribution whenever that is practical.

For our oops example, we are interested in the interaction effects, since these allow us to determine whether the proportion of protein bound to RNA is changing between conditions. In figure 4, we plot the joint distribution of the two interaction effects. We can then ask a number of question of this joint distribution. Some examples include the probability of being positive or negative, the probability of having the opposite signs, the probability that the absolute effects are exceed 0.1, and many more (see accompanying vignette).

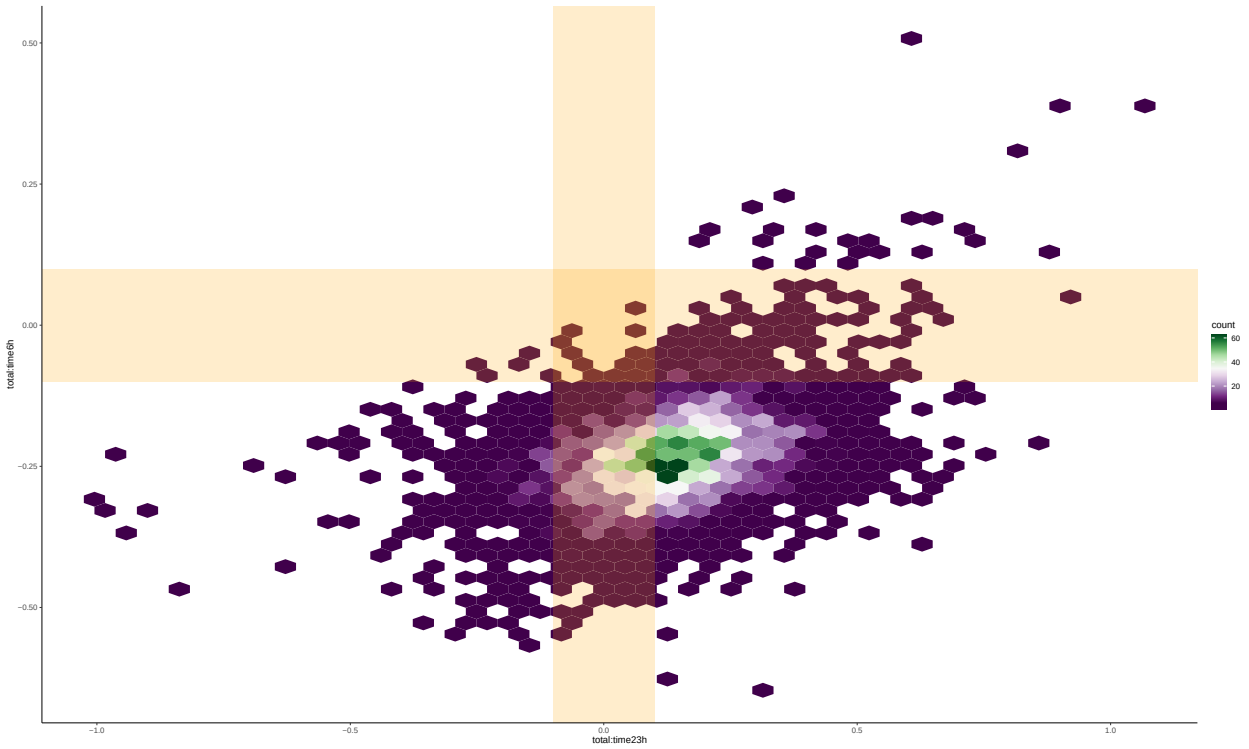

Figure 4: **Joint posterior distribution of interaction effects** Joint posterior distribution of the interaction effect of type with time at six hours and twenty-three hours. The distribution is shown as a 2d histogram with hexagon based density estimation. Orange regions highlight absolute effect sizes less than 0.1 and we see the density is concentrated outside this region, though is more overlapping at twenty-three hours.

2.4 Modelling strategies

2.4.1 Parametric models

Here, we outline some commonly used modelling strategies and related them to the proteomics literature. This is by no means meant to be exhaustive, since there are infinitely many possible models one could specify. One of the most commonly used model is the linear model, where we wish to link a set of predictor to outcomes:

$$y = \beta X + \epsilon. \tag{21}$$

If we choose ϵ to be Gaussian noise, we can write down the model as follows:

$$y \sim \mathcal{N}(\beta X, \sigma^2). \tag{22}$$

There is nothing Bayesian about this model until we specify priors. Remember, the choice of prior should be motivated by generative and predictive modelling and of course the priors should respect the domain of the parameters. Typically, one may start with a Gaussian or Student-t prior on β. The prior on σ could be a number of probability distributions that respect positivity. Usually recommendations include half-normal, exponential, half-student-t and half-Cauchy depending how confident we are about the scale of the noise[36]. Since protein abundances are positive quantities, it is has been typical to model them as a log normal distribution

$$\log y \sim \mathcal{N}(\beta X, \sigma^2). \tag{23}$$

If our observed data were counts then it maybe sensible to use Poisson or Negative binomial regression[53]:

$$\begin{aligned} y &\sim \text{Pois}(\lambda) \\ \log(\lambda) &= \beta X \\ \beta &\sim \mathcal{N}(0, \sigma^2), \end{aligned} \tag{24}$$

and

$$\begin{aligned} y &\sim \text{NB}(r, p) \\ \text{logit}(p) &= \beta X \\ \beta &\sim \mathcal{N}(0, \sigma^2). \end{aligned} \tag{25}$$

In each of the above cases, we would have to choose appropriate priors on the model parameters. Again, using the evaluation strategies previously discussed. Many data have an excess of zero's which are not captured by the usual statistical models. Many distribution can be extended to hurdle or zero-inflated models to account for these observations. The distribution of the noise process can be, again, as exotic as needed for the task at hand. Consistent outliers might call for student-t distribution or perhaps the noise itself depends

on some covariates, such as time or spatial location. See Goeminne *et al.* [40] for an example, albeit non-Bayesian, of a hurdle model applied to proteomics.

Another useful model strategy is to allow parameters at the population-level and group-level, note that these are sometimes referred to as fixed and random effects. For example, a paired t-test is a linear model with grouping specified by the subject or replicate. More complex groupings are allowed, including interactions between groupings and groups that are nested within each other. If β and u are population-level and group-level coefficients with design matrices X and Z, then a log linear (mixed) model would be[4]:

$$\log y = \beta X + uZ + \epsilon. \tag{26}$$

As before the flexibility of the Bayesian analysis, allows you to build any sensible probability distribution on top of this initial model. See Morris *et al.* [70, 71] for application of mixed-models to proteomics data, as well as the accompanying vignette.

Another useful modelling strategy is mixture models, which occurs frequently in the context of clustering and classification[67]. The mixture model assume that data arises from different components each with the same parametric density with different parameters:

$$\begin{aligned} y_i|z_i, \theta &\sim F(\theta_{z_i}) \\ z_i|\pi &\sim \text{cat}(\pi) \\ \pi|\alpha &\sim \text{Dir}(\alpha) \\ \theta &\sim p(\theta). \end{aligned} \tag{27}$$

The priors and the likelihood can be chosen based on the specific application at hand and the workflow recommendations can be applied. It is often insightful to write, using the law of total probability, the mixture model as

$$p(y_i) = \sum_{k=1}^{K} \pi_k p(y_i|\theta_k). \tag{28}$$

Note that because a Dirichlet prior is placed on π, the entries must all be non-negative and sum to unity. Hence, the entries of π can be interpreted as weights. The data cluster by being associated to the component density which fits those observations through the variables z_i. Examples of mixture models applied to proteomics include Chung *et al.* [18], Crook *et al.* [20, 22]

2.4.2 Non-parametric models

In contrast to parametric models, non-parametric models allow more parameters as more data is observed. Phrased another way, in a parametric model there are finitely many parameters, whilst in a non-parametric model there are infinitely many such parameters. This makes non-parametric models more flexible; however, to avoid the over-fitting concerns raised in earlier section we ought to be prudent with our choice of priors. One of the most

popular non-parametric model is the Gaussian process (GP), which can be used to model functions f. Suppose we observe data $\{(x_i, y_i)_{i=1,\ldots,n}\}$, we wish to find a function f such that $f(x_i)$ models y_i. Let us assume a Gaussian regression set-up, using a *Gaussian process prior* to model f:

$$\begin{aligned} y &\sim \mathcal{N}(f, \sigma^2) \\ f &\sim \mathcal{GP}(m, C). \end{aligned} \tag{29}$$

The Gaussian process is a distribution over *functions* that is uniquely characterised by its mean and covariance functions. The choice of mean and covariance functions are modelling choices to be made by the domain expert. Typically, the covariance function is parametrised by some parameters $C = C(\theta)$ and we can also place priors on these parameters so that $\theta \sim p(\theta)$. Again, these modelling choices can be evaluated using prior/posterior predictive checks. We refer to several discussion on choose priors for Gaussian processes[5,25,30,77,102]. For applications of Gaussian process to proteomics data see Crook *et al.* [22], Fang *et al.* [28], Maboudi Afkham *et al.* [60], Shin *et al.* [92].

The other non-parametric model that is frequently used is the Dirichlet process[1,29]. Dirichlet processes are a popular tool for modelling data with parameter repetitions. For example, when we cluster data, all observation associated with cluster 1 share the same parameter θ_1. The Dirichlet process is defined using a base distribution G and a concentration parameter α and is written $\mathrm{DP}(G, \alpha)$. For example, suppose that $G = \mathcal{N}(0, 1)$, then we can simulate from the Dirichlet process as follows. For any $i \geq 1$, with probability $\frac{\alpha}{\alpha+i-1}$ sample $x_i \sim \mathcal{N}(0, 1)$ and with probability $\frac{n_x}{\alpha+i-1}$ let $x_i = x$, where n_x is the number of previous observations of x. As we can see, if we have already observed a value, then we are increasingly likely to observe it in the future. This property is sometime referred to as the "rich get richer property".

The Dirichlet process allows us to work with mixture models with infinitely many components, which is useful for characterising the uncertainty in the number of components. Once we have a sensible parametric likelihood for the observations $F(\theta_i)$, the Dirichlet process can be used as a prior to construct the Dirichlet process mixture model:

$$\begin{aligned} y_i|\theta_i &\sim F(\theta_i) \\ \theta_i|P &\sim P \\ P|\alpha, G &\sim \mathrm{DP}(G, \alpha). \end{aligned} \tag{30}$$

Since P will be discrete, the set$\{\theta_i\}_{i=1,\ldots,n}$ will contain repetitions. This allows us to think of this model as a mixture model, where the groups of parameters define the components. Extensions are available[83,96] and MCMC algorithms for fitting these models can be found in Neal [72]. For applications of Dirichlet processes to proteomics data see Claassen *et al.* [19] and Choi *et al.* [16].

3 Discussion

Despite Bayesian statistics offering a powerful and flexible framework for performing proteomics data analysis, we have seen that few problems have been tackled using this methodology. Even when Bayesian statistics has been applied, the methodology has not made complete use of the information available form such an analysis. Many analysis have simply resorted to proxies from frequentist based approaches. One of the key advantages of the Bayesian approach is to be able to jointly model several quantities and provide uncertainty estimates in any parameters. Another advantage of Bayesian statistics is that it makes modelling assumption explicit; hence, it becomes clear how the models can be improved and what is the extent of their limitations.

We summarised key modelling ideas in Bayesian statistics starting with the workflow. We highlighted that the Bayesian workflow has a consistent approach to model building, model criticism and evaluation grounded in probability theory. Using a case study, we provided a workflow for developing a Bayesian model for Organic Orthogonal Phase Separation (OOPS) data. We then proceeded with common modelling strategies to allow proteomics researchers to understand key models in the literature and link them to current methods used in the literature.

We conclude that mass spectrometry-based proteomics has had resisted uptake on Bayesian methods for various reasons. These include, but are not limited to, lack of familiarity with the workflow and tools available, lack of compelling examples in literature, lack of desire to invest in bespoke model development. We hope that this review goes some way in removing some of the barriers to applying and understand Bayesian methods. As recommendations, we actively encourage proteomics researchers to collaborate with applied Bayesian researchers, for those developing Bayesian tools to provide workflows and open software so their approach can be used by the community and finally for the community to be clear about their modelling choices. We believe that computation is a secondary issue given the rapid uptake of deep learning methods, which are also computationally intensive.

References

[1] Antoniak, C. E. (1974). Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *The annals of statistics*, pages 1152–1174.

[2] Bantscheff, M. et al. (2007). Quantitative mass spectrometry in proteomics: a critical review. *Analytical and bioanalytical chemistry*, **389**(4), 1017–1031.

[3] Barylyuk, K. et al. (2020). A comprehensive subcellular atlas of the toxoplasma proteome via hyperlopit provides spatial context for protein functions. *Cell host & microbe*, **28**(5), 752–766.

[4] Bates, D. et al. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.

[5] Berger, J. O. et al. (2001). Objective bayesian analysis of spatially correlated data. *Journal of the American Statistical Association*, **96**(456), 1361–1374.

[6] Betancourt, M. (2017). A conceptual introduction to hamiltonian monte carlo. *arXiv preprint arXiv:1701.02434*.

[7] Betancourt, M. (2021). Towards a principled bayesian workflow. *https://github.com/betanalpha/knitr_case_studies/tree/master/principled_bayesian_workflow*.

[8] Bingham, E. et al. (2019). Pyro: Deep universal probabilistic programming. *The Journal of Machine Learning Research*, **20**(1), 973–978.

[9] Blei, D. M. et al. (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association*, **112**(518), 859–877.

[10] Brooks, S. et al. (2011). Handbook of markov chain monte carlo.

[11] Bürkner, P.-C. (2017). brms: An r package for bayesian multilevel models using stan. *Journal of statistical software*, **80**(1), 1–28.

[12] Carpenter, B. et al. (2017). Stan: A probabilistic programming language. *Journal of statistical software*, **76**(1), 1–32.

[13] Carvalho, P. C. et al. (2011). Analyzing marginal cases in differential shotgun proteomics. *Bioinformatics*, **27**(2), 275–276.

[14] Chen, A. T. et al. (2019). Dart-id increases single-cell proteome coverage. *PLoS computational biology*, **15**(7), e1007082.

[15] Chen, S. S. et al. (2005). Improving mass and liquid chromatography based identification of proteins using bayesian scoring. *Journal of proteome research*, **4**(6), 2174–2184.

[16] Choi, H. et al. (2010). Analysis of protein complexes through model-based biclustering of label-free quantitative ap-ms data. *Molecular systems biology*, **6**(1), 385.

[17] Christopher, J. A. et al. (2021). Subcellular proteomics. *Nature Reviews Methods Primers*, **1**(1), 1–24.

[18] Chung, C. et al. (2013). Non-parametric bayesian approach to post-translational modification refinement of predictions from tandem mass spectrometry. *Bioinformatics*, **29**(7), 821–829.

[19] Claassen, M. et al. (2009). Proteome coverage prediction with infinite markov models. *Bioinformatics*, **25**(12), i154–i160.

[20] Crook, O. M. et al. (2018). A bayesian mixture modelling approach for spatial proteomics. *PLoS computational biology*, **14**(11), e1006516.

[21] Crook, O. M. et al. (2019a). A bioconductor workflow for the bayesian analysis of spatial proteomics. *F1000Research*, **8**.

[22] Crook, O. M. et al. (2019b). Semi-supervised non-parametric bayesian modelling of spatial proteomics. *arXiv preprint arXiv:1903.02909*.

[23] Crook, O. M. et al. (2020). A semi-supervised bayesian approach for simultaneous protein sub-cellular localisation assignment and novelty detection. *PLoS computational biology*, **16**(11), e1008288.

[24] Crook, O. M. et al. (2021). Inferring differential subcellular localisation in comparative spatial proteomics using bandle. *bioRxiv*.

[25] De Oliveira, V. (2007). Objective bayesian analysis of spatial data with measurement error. *Canadian Journal of Statistics*, **35**(2), 283–301.

[26] Del Moral, P. et al. (2006). Sequential monte carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **68**(3), 411–436.

[27] Deng, X. et al. (2007). Cross-platform analysis of cancer biomarkers: a bayesian network approach to incorporating mass spectrometry and microarray data. *Cancer informatics*, **3**, 117693510700300001.

[28] Fang, S. et al. (2021). A bayesian semi-parametric model for thermal proteome profiling. *Communications biology*, **4**(1), 1–15.

[29] Ferguson, T. S. (1973). A bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230.

[30] Fuglstad, G.-A. et al. (2019). Constructing priors that penalize the complexity of gaussian random fields. *Journal of the American Statistical Association*, **114**(525), 445–452.

[31] Gabry, J. et al. (2019). Visualization in bayesian workflow. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **182**(2), 389–402.

[32] Ge, H. et al. (2018). Turing: A language for flexible probabilistic inference. In *International conference on artificial intelligence and statistics*, pages 1682–1690. PMLR.

[33] Geladaki, A. et al. (2019). Combining lopit with differential ultracentrifugation for high-resolution spatial proteomics. *Nature communications*, **10**(1), 1–15.

[34] Gelfand, A. E. (2000). Gibbs sampling. *Journal of the American statistical Association*, **95**(452), 1300–1304.

[35] Gelman, A. et al. (1995). *Bayesian data analysis*. Chapman and Hall/CRC.

[36] Gelman, A. et al. (2017). The prior can often only be understood in the context of the likelihood. *Entropy*, **19**(10), 555.

[37] Gelman, A. et al. (2020). Bayesian workflow. *arXiv preprint arXiv:2011.01808*.

[38] Gessel, M. M. et al. (2014). Maldi imaging mass spectrometry: spatial molecular analysis to enable a new age of discovery. *Journal of proteomics*, **107**, 71–82.

[39] Gilks, W. R. et al. (1995). *Markov chain Monte Carlo in practice*. CRC press.

[40] Goeminne, L. J. et al. (2020). Msqrob takes the missing hurdle: uniting intensity-and count-based proteomics. *Analytical chemistry*, **92**(9), 6278–6287.

[41] Hadfield, J. D. (2010). Mcmc methods for multi-response generalized linear mixed models: the mcmcglmm r package. *Journal of statistical software*, **33**(1), 1–22.

[42] Halloran, J. T. et al. (2016). Dynamic bayesian network for accurate detection of peptides from tandem mass spectra. *Journal of proteome research*, **15**(8), 2749–2759.

[43] Harris, K. et al. (2009). Definition of valid proteomic biomarkers: a bayesian solution. pages 137–149.

[44] Hernández, B. et al. (2015). Bayesian methods for proteomic biomarker development. *EuPA Open Proteomics*, **9**, 54–64.

[45] Hoffman, M. D. et al. (2014). The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.*, **15**(1), 1593–1623.

[46] Huttlin, E. L. et al. (2015). The bioplex network: a systematic exploration of the human interactome. *Cell*, **162**(2), 425–440.

[47] Hwang, D. et al. (2008). Ms-bid: a java package for label-free lc-ms-based comparative proteomic analysis. *Bioinformatics*, **24**(22), 2641–2642.

[48] Johnson, D. T. et al. (2019). Fast photochemical oxidation of proteins (fpop): A powerful mass spectrometry–based structural proteomics tool. *Journal of Biological Chemistry*, **294**(32), 11969–11979.

[49] Jow, H. et al. (2014). Bayesian identification of protein differential expression in multi-group isobaric labelled mass spectrometry data. *Statistical applications in genetics and molecular biology*, **13**(5), 531–551.

[50] Kass, R. E. et al. (1995). Bayes factors. *Journal of the american statistical association*, **90**(430), 773–795.

[51] Kostelic, M. et al. (2021). Unideccd: Deconvolution of charge detection-mass spectrometry data.

[52] Kuschner, K. W. et al. (2010). A bayesian network approach to feature selection in mass spectrometry data. *BMC bioinformatics*, **11**(1), 1–10.

[53] Lawless, J. F. (1987). Negative binomial and mixed poisson regression. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique*, pages 209–225.

[54] LeDuc, R. D. et al. (2014). The c-score: a bayesian framework to sharply improve proteoform scoring in high-throughput top down proteomics. *Journal of proteome research*, **13**(7), 3231–3240.

[55] Lewis, N. H. et al. (2018). Peptide refinement by using a stochastic search. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **67**(5), 1207–1236.

[56] Liao, H. et al. (2014). A new paradigm for clinical biomarker discovery and screening with mass spectrometry through biomedical image analysis principles. pages 1332–1335.

[57] Lim, M. Y. et al. (2017). Improved method for determining absolute phosphorylation stoichiometry using bayesian statistics and isobaric labeling. *Journal of proteome research*, **16**(11), 4217–4226.

[58] Liu, Y. et al. (2020). Function-on-scalar quantile regression with application to mass spectrometry proteomics data. *The Annals of Applied Statistics*, **14**(2), 521–541.

[59] Lunn, D. et al. (2009). The bugs project: Evolution, critique and future directions. *Statistics in medicine*, **28**(25), 3049–3067.

[60] Maboudi Afkham, H. et al. (2017). Uncertainty estimation of predictions of peptides' chromatographic retention times in shotgun proteomics. *Bioinformatics*, **33**(4), 508–513.

[61] Maity, A. K. et al. (2020). Bayesian data integration and variable selection for pan-cancer survival prediction using protein expression data. *Biometrics*, **76**(1), 316–325.

[62] Mallikarjun, V. et al. (2020). Bayesenproteomics: Bayesian elastic nets for quantification of peptidoforms in complex samples. *Journal of proteome research*, **19**(6), 2167–2184.

[63] Marty, M. T. et al. (2015). Bayesian deconvolution of mass and ion mobility spectra: from binary interactions to polydisperse ensembles. *Analytical chemistry*, **87**(8), 4370–4376.

[64] Masson, G. R. et al. (2019). Recommendations for performing, interpreting and reporting hydrogen deuterium exchange mass spectrometry (hdx-ms) experiments. *Nature methods*, **16**(7), 595–602.

[65] Mateus, A. et al. (2020). Thermal proteome profiling for interrogating protein interactions. *Molecular systems biology*, **16**(3), e9232.

[66] Mathieson, T. et al. (2018). Systematic analysis of protein turnover in primary cells. *Nature communications*, **9**(1), 1–10.

[67] McLachlan, G. J. et al. (2019). Finite mixture models. *Annual review of statistics and its application*, **6**, 355–378.

[68] Millikin, R. J. et al. (2020). A bayesian null interval hypothesis test controls false discovery rates and improves sensitivity in label-free quantitative proteomics. *Journal of proteome research*, **19**(5), 1975–1981.

[69] Morris, J. S. et al. (2006). Analysis of mass spectrometry data using bayesian wavelet-based functional mixed models. bayesian inference for gene expression and proteomics.

[70] Morris, J. S. et al. (2008). Bayesian analysis of mass spectrometry proteomic data using wavelet-based functional mixed models. *Biometrics*, **64**(2), 479–489.

[71] Morris, J. S. et al. (2011). Automated analysis of quantitative image data using isomorphic functional mixed models, with application to proteomics data. *The annals of applied statistics*, **5**(2A), 894.

[72] Neal, R. M. (2000). Markov chain sampling methods for dirichlet process mixture models. *Journal of computational and graphical statistics*, **9**(2), 249–265.

[73] Ni, Y. et al. (2019). Bayesian hierarchical varying-sparsity regression models with application to cancer proteogenomics. *Journal of the American Statistical Association*, **114**(525), 48–60.

[74] Nightingale, K. et al. (2018). 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.

[75] O'Brien, J. J. et al. (2018). Compositional proteomics: Effects of spatial constraints on protein quantification utilizing isobaric tags. *Journal of proteome research*, **17**(1), 590–599.

[76] O'Brien, J. J. et al. (2018). The effects of nonignorable missing data on label-free mass spectrometry proteomics experiments. *The annals of applied statistics*, **12**(4), 2075.

[77] Paulo, R. et al. (2005). Default priors for gaussian processes. *The Annals of Statistics*, **33**(2), 556–582.

[78] Peshkin, L. et al. (2019). Bayesian confidence intervals for multiplexed proteomics integrate ion-statistics with peptide quantification concordance. *Molecular & Cellular Proteomics*, **18**(10), 2108–2120.

[79] Phillips, A. et al. (2021). Uncertainty aware protein-level quantification and differential expression analysis of proteomics data with seamass. *Statistical methods for proteomics*.

[80] Queiroz, R. M. et al. (2019). Comprehensive identification of rna–protein interactions in any organism using orthogonal organic phase separation (oops). *Nature biotechnology*, **37**(2), 169–178.

[81] Raftery, A. E. et al. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, **92**(437), 179–191.

[82] Robert, C. P. et al. (1999). The metropolis—hastings algorithm. In *Monte Carlo Statistical Methods*, pages 231–283. Springer.

[83] Rodriguez, A. et al. (2008). The nested dirichlet process. *Journal of the American statistical Association*, **103**(483), 1131–1154.

[84] Saltzberg, D. J. et al. (2017). A residue-resolved bayesian approach to quantitative interpretation of hydrogen–deuterium exchange from mass spectrometry: application to characterizing protein–ligand interactions. *The Journal of Physical Chemistry B*, **121**(15), 3493–3501.

[85] Salvatier, J. et al. (2016). Probabilistic programming in python using pymc3. *PeerJ Computer Science*, **2**, e55.

[86] Santra, T. et al. (2016). A bayesian algorithm for detecting differentially expressed proteins and its application in breast cancer research. *Scientific reports*, **6**(1), 1–10.

[87] Schad, D. J. et al. (2021). Toward a principled bayesian workflow in cognitive science. *Psychological methods*, **26**(1), 103.

[88] Schmiesing, J. A. et al. (2000). A human condensin complex containing hcap-c–hcap-e and cnap1, a homolog of xenopus xcap-d2, colocalizes with phosphorylated histone h3 during the early stage of mitotic chromosome condensation. *Molecular and cellular biology*, **20**(18), 6996–7006.

[89] Schopper, S. et al. (2017). Measuring protein structural changes on a proteome-wide scale using limited proteolysis-coupled mass spectrometry. *Nature protocols*, **12**(11), 2391–2410.

[90] Serang, O. et al. (2012). Recognizing uncertainty increases robustness and reproducibility of mass spectrometry-based protein inferences. *Journal of proteome research*, **11**(12), 5586–5591.

[91] Serang, O. et al. (2013). Nonparametric bayesian evaluation of differential protein quantification. *Journal of proteome research*, **12**(10), 4556–4565.

[92] Shin, J. J. et al. (2020). Spatial proteomics defines the content of trafficking vesicles captured by golgin tethers. *Nature communications*, **11**(1), 1–13.

[93] Shteynberg, D. D. et al. (2019). Ptmprophet: fast and accurate mass modification localization for the trans-proteomic pipeline. *Journal of proteome research*, **18**(12), 4262–4272.

[94] Smith, A. F. et al. (1993). Bayesian computation via the gibbs sampler and related markov chain monte carlo methods. *Journal of the Royal Statistical Society: Series B (Methodological)*, **55**(1), 3–23.

[95] Talts, S. et al. (2018). Validating bayesian inference algorithms with simulation-based calibration. *arXiv preprint arXiv:1804.06788*.

[96] Teh, Y. W. et al. (2006). Hierarchical dirichlet processes. *Journal of the american statistical association*, **101**(476), 1566–1581.

[97] The, M. et al. (2019). Integrated identification and quantification error probabilities for shotgun proteomics. *Molecular & cellular Proteomics*, **18**(3), 561–570.

[98] The, M. et al. (2021). Triqler for maxquant: Enhancing results from maxquant by bayesian error propagation and integration. *Journal of proteome research*, **20**(4), 2062–2068.

[99] Toby, T. K. et al. (2016). Progress in top-down proteomics and the analysis of proteoforms. *Annual review of analytical chemistry*, **9**, 499–519.

[100] Tran, D. et al. (2016). Edward: A library for probabilistic modeling, inference, and criticism. *arXiv preprint arXiv:1610.09787*.

[101] Tversky, A. et al. (1974). Judgment under uncertainty: Heuristics and biases. *science*, **185**(4157), 1124–1131.

[102] van der Vaart, A. W. et al. (2009). Adaptive bayesian estimation using a gaussian random field with inverse gamma bandwidth. *The Annals of Statistics*, **37**(5B), 2655–2675.

[103] Vehtari, A. et al. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and computing*, **27**(5), 1413–1432.

[104] Vehtari, A. et al. (2019). Rank-normalization, folding, and localization: An improved r-hat for assessing convergence of mcmc. *arXiv preprint arXiv:1903.08008*.

[105] Webb-Robertson, B.-J. M. et al. (2014). Bayesian proteoform modeling improves protein quantification of global proteomic measurements. *Molecular & cellular proteomics*, **13**(12), 3639–3646.

[106] Wilkie, D. (1974). Second law of thermodynamics. *Nature*, **251**(5476), 601–602.