The Annals of Applied Statistics
2022, Vol. 16, No. 4, 2554–2576
https://doi.org/10.1214/22-AOAS1603

SEMI-SUPERVISED NONPARAMETRIC BAYESIAN MODELLING OF SPATIAL PROTEOMICS

BY OLIVER M. CROOK[1,2,a], KATHRYN S. LILLEY[2,c], LAURENT GATTO[3,d] AND PAUL D. W. KIRK[1,b]

[1]*MRC Biostatistics Unit, School of Clinical Medicine, University of Cambridge,* [a]*oliver.crook@stats.ox.ac.uk,* [b]*paul.kirk@mrc-bsu.cam.ac.uk*

[2]*Cambridge Centre for Proteomics, Department of Biochemistry, University of Cambridge,* [c]*k.s.lilley@bioc.cam.ac.uk*

[3]*de Duve Institute, UCLouvain,* [d]*laurent.gatto@uclouvain.be*

Understanding subcellular 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 subcellular 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 nonparametric Bayesian framework, using K-component mixtures of Gaussian process regression models. The Gaussian process regression model accounts for correlation structure within a subcellular niche, with each mixture component capturing the distinct correlation structure observed within each niche. The availability of *marker proteins* (i.e., proteins with a priori known labelled locations) motivates a semi-supervised learning approach 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.

1. Introduction. Proteins are biomolecules that have a diverse set of functional roles within a cell enabling proliferation and survival. For a protein to be able to perform its function(s), it must interact with other binding partners and substrates which requires it to localise to the correct subcellular compartment (Gibson (2009)). There is mounting evidence implicating aberrant protein localisation in disease, including cancer and obesity (Olkkonen and Ikonen (2006), Laurila and Vihinen (2009), Luheshi, Crowther and Dobson (2008), De Matteis and Luini (2011), Cody, Iampietro and Lécuyer (2013), Kau, Way and Silver (2004), Rodriguez, Au and Henderson (2004), Latorre et al. (2005), Shin et al. (2013), Siljee et al. (2018)). Mapping the location of proteins within the cell using high-resolution spatial proteomic approaches is thus of high utility in the characterisation of therapeutic targets and in determining pathobiological mechanisms (Cook and Cristea (2019)). To interrogate the subcellular locations of thousands of proteins per experiment, recent advances in high-throughput spatial proteomics (Christoforou et al. (2016), Mulvey et al. (2017), Geladaki et al. (2019)), followed by rigorous data analysis (Gatto et al. (2010)) can be applied. As we elaborate in our exposition below, the methodology relies on the observation that each organelle (or, more generally, each subcellular niche) can be characterised by a subcellular fractionation profile that is shared by the proteins that localise to that organelle

Received July 2020; revised October 2021.
Key words and phrases. Proteomics, Bayesian mixture models, semi-supervised learning.

(De Duve and Beaufay (1981)). Applications of spatial proteomics experiments and analyses have enabled organelle-specific localisation to be determined for many proteins in many systems (Dunkley et al. (2006), Tan et al. (2009), Hall et al. (2009), Breckels et al. (2013)), including mouse pluripotent stem cells (mESCs) (Christoforou et al. (2016)) and cancer cell lines (Thul et al. (2017)). Mass spectrometry (MS) based spatial proteomics, which is what we consider here, has gained in popularity in recent years with several recent applications across many different organisms (Christoforou et al. (2016), Beltran, Mathias and Cristea (2016), Jadot et al. (2017), Itzhak et al. (2017), Mendes et al. (2017), Hirst et al. (2018), Davies et al. (2018), Orre et al. (2019), Nightingale et al. (2019), Shin et al. (2019), Barylyuk et al. (2020)).

An overview of a typical spatial proteomics experiment is provided in Figure 1(A). First, cells are 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 (Itzhak et al. (2016), Geladaki et al. (2019), Orre et al. (2019)) or equilibrium density centrifugation (Christoforou et al. (2016), Dunkley et al. (2004), Dunkley et al. (2006)), among others (Parsons, Fernández-Niño and Heazlewood (2014), Heard et al. (2015)). After centrifugation the proteins present in the fractions generated by this process are then extracted. The protein abundance of each protein in each fraction is then determined experimentally using high accuracy mass-spectrometry. This gives, for each protein, an abundance profile across the subcellular fractions.

In the localisation of organelle proteins by isotope tagging (LOPIT; Dunkley et al. (2004), Dunkley et al. (2006), Sadowski et al. (2006)) and *hyper*-LOPIT (Christoforou et al. (2016), Mulvey et al. (2017)) approaches, cell lysis is proceeded by the separation of subcellular 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) (Thompson et al. (2003)) and protein distributions revealing organelle specific correlation profiles within the fractions are achieved using synchronous precursor selection mass-spectrometry (SPS-MS^3). The resultant data are annotated using *marker proteins*; that is, proteins with unambiguous single localisations from the literature or appropriate databases such as the Human Protein Atlas (HPA) (Thul et al. (2017)) or Gene Ontology (Ashburner et al. (2000)); see Gatto et al. (2014a) for discussion. We, therefore, know a priori the unique subcellular niche to which each marker protein localises, and hence these proteins define a *labelled training dataset* comprising proteins for which we know the corresponding subcellular niche localisations (class labels). We denote by K the number of distinct subcellular niches that appear in this training dataset; that is, K is the number of classes. Typical spatial proteomics experiments can now provide information on several thousands of proteins; for example, 5032 were quantified for the mESC application (Section 3.2). Modern experiments are expected to resolve all major subcellular niches, but the precise number depends on experimental design. Indeed, in the *Drosophila* application (Section 3.1) no cytosolic component is observed because the supernatant, enriched in cytosolic proteins, was discarded; that is, all proteins belonging to the cytosol class were removed from the experiment. Furthermore, eukaryote cells with more complex subcellular organisation are likely to have more subcellular niches observed, if the data are sufficiently resolved (Barylyuk et al. (2020)). The experimental design (and thus the organelle separation) may be validated prior to quantitative analysis using western blotting (Mulvey et al. (2017)).

In work that contributed to the discovery of previously unknown organelles and the award of a Nobel prize, de Duve and colleagues (De Duve (1969), De Duve and Beaufay (1981), Blobel (2013)) observed that proteins belonging to the same organelle possessed very similar abundance profiles (Figure 1(B)). This motivates the following data analysis problem: given the abundance profiles of the marker proteins that are already known to localise to a particular

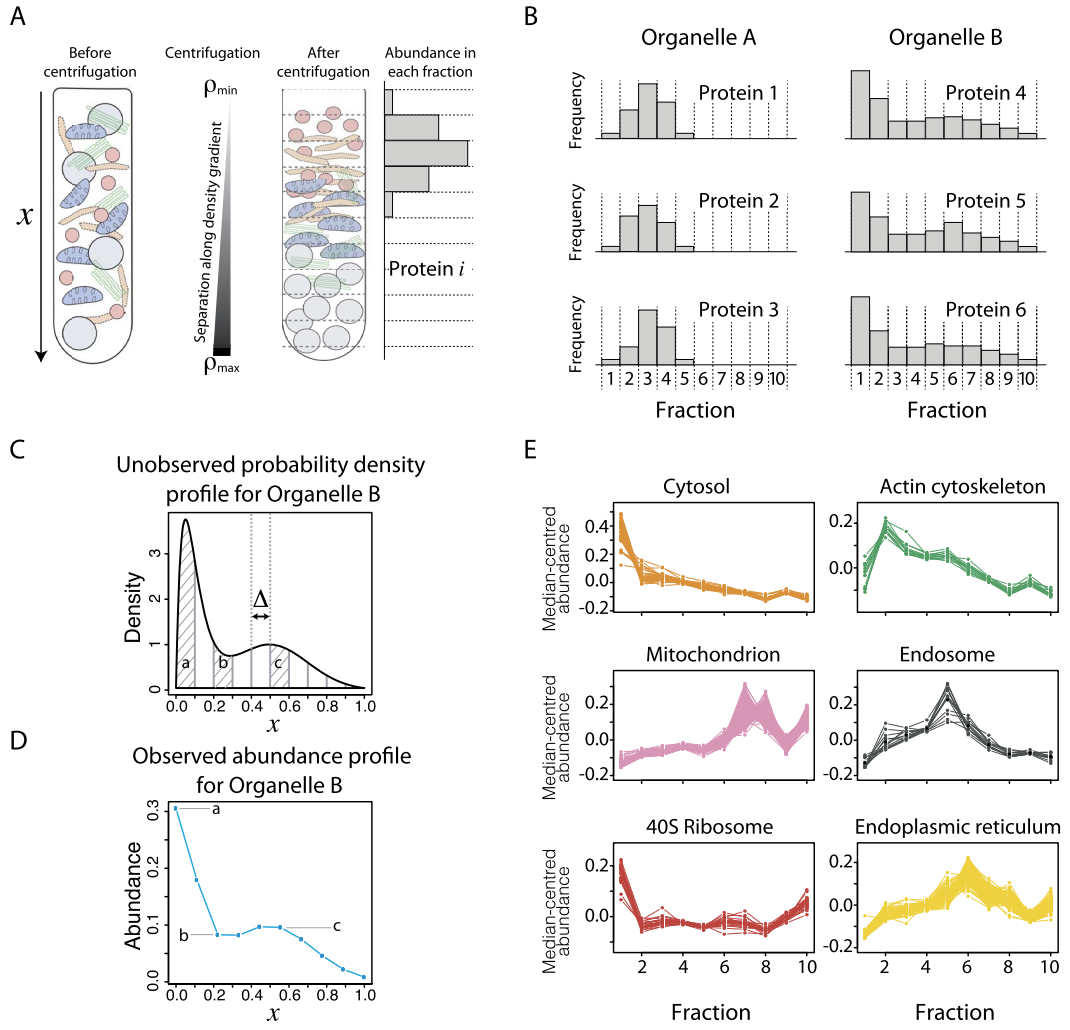

FIG. 1. *An overview of the experimental design of a spatial proteomics experiment using density-gradient centrifugation*: (*A*) *Cellular content is loaded onto a preformed iodixanol density gradient. The tube is then subject to centrifugation, typically at* $10^6 g$ *for eight hours. After centrifugation 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 subcellular niche share the same* (*median-centered*) *abundance profiles.*

subcellular niche (e.g., organelle), can we determine which other proteins might also localise to that niche? In many previous analyses this problem has been addressed as a black-box classification problem, with partial least squares discriminant analysis (Dunkley et al. (2006), Tan et al. (2009)) and the support vector machine (SVM) (Christoforou et al. (2016), Itzhak et al. (2016), Orre et al. (2019)) being the most popular approaches. However, other approaches are also used, such as nearest neighbour classifiers (Groen et al. (2014)), random forests (Ohta et al. (2010)), naive Bayes (Nikolovski et al. (2012)) and neural networks (Tardif et al. (2012), Beltran, Mathias and Cristea (2016)). We refer to Gatto et al. (2014a) for a review. Other advances include the use of transfer learning to incorporate additional sources of localisation information (Breckels et al. (2016)) and the development and application of outlier detection

techniques (Breckels et al. (2013)). A recent review of the improvements in resolution of spatial proteomics experiments over the last decade is provided by Gatto, Breckels and Lilley (2019).

The classification approaches listed above have a number of major limitations. For example, they 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. Posttranslation modifications and the presence of different 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 challenge of making protein-niche associations.

Crook et al. (2018) demonstrated the importance of uncertainty quantification in spatial proteomics analysis. This study developed a generative mixture model of MS spatial proteomics data and, using this model, computed posterior distributions of protein localisation probabilities. However, this model made a number of assumptions that simplified the analysis but which do not accurately reflect the data generating process. In the present manuscript we develop a generative model for the data that is more clearly motivated by the data generating process.

1.1. *Model development.* Let x be the spatial axis along which density gradient separation occurs (see Figure 1(A)), and let $x_1 < x_2$ be two distinct points along x. We assume that the kth organelle may be characterised by a smooth latent probability density function, $p_k(x)$ (Figure 1(C)), such that, for any protein i that uniquely localises to the kth organelle, the (unobserved) absolute quantity of protein i in the region $[x_1, x_2]$ after separation is given by

$$q_k(x_1, x_2) = \int_{x=x_1}^{x_2} p_k(x)\,\mathrm{d}x. \tag{1}$$

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 $q_k(x_j, x_j + \Delta)$, that is, for any protein that uniquely localises to the kth 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 1(D). 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 1(E)) and 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 2.4.4.2). The number of tandem mass tags available limits the number, D, of discrete observations we can observe and hence the resolution of the experiment. For example, in the case studies that we consider later, $D = 4$ for the *Drosophila* example (Tan et al. (2009)), whereas for the mouse embryonic stem cell (mESCs) example $D = 10$ (Christoforou et al. (2016)). As TMT chemistry improves, it is expected that more complex designs will become available.

Functional data analysis concerns itself with the analysis of data, where the sampled data for each subject is a function (Ramsay (2004)). Wang, Chiou and Müller (2016) recently reviewed the current major approaches in functional data analysis, including functional principal component analysis (Jones and Rice (1992)), functional linear regression (Morris (2015)), functional clustering (James and Sugar (2003)) and functional classification (Preda, Saporta and Lévéder (2007)). For classification the linear discriminant analysis method was extended to the functional setting using splines (James and Hastie (2001)). Mixture discriminant analysis in the functional setting applied to model bike sharing data was considered by Bouveyron, Côme and Jacques (2015), using a functional EM algorithm. Bayesian approaches to functional classification have also been considered, such as, the wavelet based functional mixed model approach (Zhu, Brown and Morris (2012)); Bayesian variable selection has also been extended to the functional setting (Zhu, Vannucci and Cox (2010)). Rodríguez, Dunson and Gelfand (2009) use dependant Dirichlet processes in the nonparametric Bayesian setting to cluster functional data. The Gaussian process approach to analysing functional data in biomedical applications is extensive (Honkela et al. (2010), Liu et al. (2010), Stegle et al. (2010), Kalaitzis and Lawrence (2011a), Heinonen et al. (2014), Topa et al. (2015)).

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 subcellular niche is described by distinct density-gradient profiles which display a nonlinear 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 Rasmussen (2004) and Rasmussen and Williams (2006) 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 (Kirk and Stumpf (2009), Cooke et al. (2011), Kalaitzis and Lawrence (2011b), Kirk et al. (2012), Hensman, Lawrence and Rattray (2013)). We remark that many of these applications consider unsupervised clustering problems. In contrast, here we have (partially) labelled data (since the localisations of the marker proteins are known prior to our experiments), and so we may consider semi-supervised approaches. We explore inference of GP hyperparameters in two ways: first, an empirical Bayes approach in which the hyperparameters are optimised by maximising a marginal likelihood; second, 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 (Duane et al. (1987)), avoiding random walk nature evident in traditional symmetric random walk proposals (Metropolis et al. (1953), Beskos et al. (2013)). Hamiltonian Monte Carlo (HMC) has been explored previously for hyperparameter inference in GP regression (Williams and Rasmussen (1996)), 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 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 (Zhang, Leithead and Leith (2005)).

2. Methods. We provide an overview of the key modelling choices and considerations below. A comprehensive mathematical summary of the model and inference procedure is provided in the Supplementary Material (Crook et al. (2022))

2.1. *Modelling protein abundances along the density gradient.* In our experiment we make discrete observations along a continuous density gradient $\mathbf{y}_i = [y_i(x_1), \ldots, y_i(x_D)]$, 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:

$$y_i(x) = \mu_i(x) + \epsilon_i, \tag{2}$$

where μ_i is an unknown deterministic function and ϵ_i a noise variable. We assume that $\epsilon_i \sim_{\text{iid}} \mathcal{N}(0, \sigma_i^2)$, 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 subcellular localisation, with all proteins associated with subcellular niche $k = 1, \ldots, K$ sharing the same regression model; that is, $\mu_i = \mu_k$ and $\sigma_i = \sigma_k$ for all proteins in the kth subcellular niche. For clarity we refer to subcellular structures, whether that be organelles, vesicles or large multiprotein 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 $\boldsymbol{\mu}_k = [\mu_k(x_1), \ldots, \mu_k(x_D)]$ and covariance matrix $\sigma_k^2 I_D$. To perform inference for the unknown function μ_k, as is typical for spatial correlated data (Gelfand, Kottas and MacEachern (2005), Steel and Fuentes (2010)), we specify a *Gaussian process* (GP) prior for each μ_k,

$$\mu_k(x) \sim \text{GP}\big(m_k(x), C_k\big(x, x'\big)\big), \tag{3}$$

where $m_k(x)$ is the *mean function* and $C_k(x, x')$ is the *covariance function* (sometimes also known as the kernel function) of the GP prior; see Rasmussen and Williams (2006). Each component is thus captured by a Gaussian process regression model. The full complement of proteins is then modelled as a K-component mixture of Gaussian process regression models, plus an "outlier component" to model proteins that are not captured well by any of the K known subcellular components. We provide a brief overview of Bayesian K-compinent mixtures in the next section, describe the modelling of outliers in Section 2.3 and further discuss the specification of the GP prior, including hyperparameter inference, in Section 2.4.

2.2. *Finite mixture models.* This section provides a brief review of finite mixture models (see, e.g., (Fraley and Raftery (2007), Lavine and West (1992)) for more details). Finite mixture models are of the form,

$$p(\mathbf{y}|\boldsymbol{\pi}, \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k f(\mathbf{y}|\boldsymbol{\theta}_k), \tag{4}$$

where K is the number of mixture components, π_k are the mixture proportions and $f(\mathbf{y}|\boldsymbol{\theta}_k)$ are the component densities. In our application, each mixture component corresponds to a distinct subcellular niche, and $\boldsymbol{\theta}_k$ is shorthand for μ_k and σ_k, as described in Section 2.1 above. As described in the Introduction, our data includes a set of marker proteins whose subcellular niche localisations are known a priori. Thus, K is known from the outset, since we assume that we have, at least, one marker protein localising to each subcellular niche; that is, we assume that all classes are represented among our labelled data (see Crook et al. (2020), for a relaxation of this assumption).

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, $Y = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$, that we seek to model using equation (4). We associate with each of these data points a component

indicator variable, $z_i \in \{1, \ldots, K\}$ which indicates which component generated observation $\mathbf{y}_i$. In our initial exposition we consider the unsupervised case, where all z_i are unknown, and then describe how we take into account the marker proteins which are (labelled) proteins for which the z_i are known. As is common for mixture models, we perform Gibbs sampling for the z_i and π_k by sampling from the conditionals described below.

Conditional for $\boldsymbol{\pi}$: If we assign the mixture proportions a symmetric Dirichlet prior with concentration parameter α/K, then we may marginalise the π_k (Murphy (2012)) or sample them. Although sampling these parameters can lead to increased posterior variance (Gelfand and Smith (1990), Casella and Robert (1996)), it can be computationally advantageous. Conjugacy of the Dirichlet prior and multinomial likelihood means that the conditional posterior distribution of the mixing proportions, given the component indicator variables is also Dirichlet,

$$\boldsymbol{\pi}|z_1, \ldots, z_n, \alpha \sim \mathrm{Dir}(\alpha/K + n_1, \ldots, \alpha/K + n_K), \tag{5}$$

where n_k is the number of data points $\mathbf{y}_i$ for which $z_i = k$. Selecting an appropriate value for α can be challenging. A sensitivity analysis for the choice of α is provided in the Supplementary Material (Crook et al. (2022)), and we show that $\alpha = 1$ is a good default choice in practice.

Conditional for z_i: Given the mixing proportions, the prior distribution of z_i is categorical with parameter vector $\boldsymbol{\pi} = [\pi_1, \ldots, \pi_K]$,

$$P(z_i = k|\boldsymbol{\pi}) = \pi_k, \tag{6}$$

and the conditional posterior for z_i is

$$P(z_i = k|\boldsymbol{\pi}, \mathbf{y}_i, \boldsymbol{\theta}_k) \propto \pi_k f(\mathbf{y}_i|\boldsymbol{\theta}_k). \tag{7}$$

In the present application we have a number of labelled marker proteins for which the component labels are known. If protein j is a marker protein, it is unnecessary to perform inference for z_j by sampling from equation (7) and, instead, z_j is fixed from the outset (see the Supplementary Material for details, Crook et al. (2022)).

2.3. *Modelling outliers.* Crook et al. (2018) demonstrated that many proteins are not captured well by any known subcellular component. This could be because of yet undiscovered biological novelty, technical variation or a manifestation of some proteins residing in multiple localisations. Modelling outliers in mixture models can be challenging (Cooke et al. (2011), Coretto and Hennig (2016), Hennig (2004), Murphy and Murphy (2019)). Here, we take the approach of Crook et al. (2018). Briefly, we introduce a binary latent variable ϕ so that, for each protein $\mathbf{y}_i$, we have a $\phi_i \in \{0, 1\}$ indicating whether $\mathbf{y}_i$ is modelled by one of the known subcellular components or an outlier component. The augmented model becomes the following:

$$\begin{aligned} p(\mathbf{y}_i|\boldsymbol{\pi}, \boldsymbol{\theta}, \phi) &= \sum_{k=1}^{K} \pi_k f(\mathbf{y}_i|\boldsymbol{\theta}_k)^{\phi_i} g(\mathbf{y}_i|\Phi)^{1-\phi_i} \\ &= \sum_{k=1}^{K} \pi_k \big(\phi_i f\big(\mathbf{y}_i|\boldsymbol{\theta}_k\big) + (1-\phi_i) g\big(\mathbf{y}_i|\Phi\big)\big), \end{aligned} \tag{8}$$

where g is the density of the outlier component. In our case we specify g as the density of a multivariate T distribution with degrees of freedom $\kappa = 4$, mean $\mathbf{M}$ and scale matrix V. $\mathbf{M}$ is taken as the empirical global mean of the data and the scale matrix V as half the empirical covariance of the data. These choices are motivated by considering a Gaussian component with the same mean and covariance but with heavier tails to better capture dispersed proteins.

We remark that other choices of g and parameters may be suitable and can be tailored to the application at hand. In typical Bayesian fashion we specify a prior for ϕ as $p_0(\phi_i = 0) = \epsilon$, where $\epsilon \sim \mathcal{B}(u, v)$. Marginalising ϕ in equation (8) leads to the following mixture of mixtures (Malsiner-Walli, Frühwirth-Schnatter and Grün (2017)):

$$p(\mathbf{y}_i|\boldsymbol{\pi}, \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k\big((1-\epsilon) f\big(\mathbf{y}_i|\boldsymbol{\theta}_k\big) + \epsilon g\big(\mathbf{y}_i|\Phi\big)\big). \tag{9}$$

We can also rewrite the above equation in the following way:

$$p(\mathbf{y}_i|\boldsymbol{\pi}, \boldsymbol{\theta}) = \sum_{k=1}^{K} \tilde{\pi}_k f(\mathbf{y}_i|\boldsymbol{\theta}_k) + \tilde{\pi}_0 g(\mathbf{y}_i|\Phi), \tag{10}$$

where $\tilde{\pi}_k = \pi_k(1-\epsilon)$ for $k = 1, \ldots, K$ and $\tilde{\pi}_0 = \pi_k\epsilon$ and, evidently, $\sum_{k=0}^{K} \tilde{\pi}_k = 1$. Thus, ϵ can be interpreted as the prior proportion of outliers. The Jeffreys prior in this scenario would set the parameters u and v of the $\mathcal{B}(u, v)$ prior to be $u = v = \frac{1}{2}$ (Jeffreys (1946)), while a uniform prior corresponds to $u = v = 1$. We prefer to specify a weakly informative prior based on prior data. From independent microscopy data, up to 50% of proteins do not have robust single localisations (Thul et al. (2017)), and it is unlikely that there are extremely few outliers (Christoforou et al. (2016)). These considerations lead us to placing small probability mass on the upper tail $\epsilon > 0.5$ and small probability around the lower tail close to 0. Since our prior information comes from different experiments, we only consider a weakly informative Beta prior: $\epsilon \sim B(2, 10)$. A sensitivity analysis is performed in the Supplementary Material (Crook et al. (2022)). All hyperparameter choices are stated in the appendix.

2.4. *Gaussian process prior specification.* A Gaussian process (GP) is a continuous stochastic process such that any finite collection of these random variables is jointly Gaussian. A Gaussian process prior is uniquely specified by a mean function m and covariance function C which determine the mean vectors and covariance matrices of the associated multivariate Gaussian distributions. To elaborate, assuming a GP prior for the function $\mu_k(x)$ means that, at spatial coordinates $x_1, \ldots, x_D$, the joint prior of $\boldsymbol{\mu}_k = [\mu_k(x_1), \ldots, \mu_k(x_D)]^T$ is multivariate Gaussian with mean vector $\boldsymbol{m}_k = [m_k(x_1), \ldots, m_k(x_D)]$ 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

$$C_k(x_i, x_j) = a_k^2 \exp\left(-\frac{\|x_i - x_j\|_2^2}{l_k}\right). \tag{11}$$

2.4.1. *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 $Y_k = \{\mathbf{y}_{i_1}, \ldots, \mathbf{y}_{i_{n_k}}\}$, where $i_1, \ldots, i_{n_k} \in \{1, \ldots, n\}$ are the indices for which $z_{i_1} = \cdots = z_{i_{n_k}} = k$. Our model tells us that

$$Y_k|\boldsymbol{\mu}_k, \sigma_k \sim \mathcal{N}(\boldsymbol{\mu}_k, \sigma_k^2 I_D). \tag{12}$$

Then, we can write this as

$$\begin{aligned} &Y_k(x_1), \ldots, Y_k(x_D)|\mu_k, \sigma_k \\ &\quad \sim \mathcal{N}\big(\mu_k(x_1), \ldots, \mu_k(x_D), \ldots, \mu_k(x_1), \ldots, \mu_k(x_D), \sigma_k^2 I_{n_k D}\big), \end{aligned} \tag{13}$$

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

$$\mu_k(x_1), \ldots, \mu_k(x_D), \ldots, \mu_k(x_1), \ldots, \mu_k(x_D)|a_k, l_k \sim \mathcal{N}(0, C_k), \tag{14}$$

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 kth component evaluated at $\tau = \{x_1, \ldots, x_D\}$,

$$C_k = \begin{bmatrix} A_k & A_k & \ldots & A_k \\ A_k & A_k & \ldots & A_k \\ \vdots & \vdots & \ddots & \vdots \\ A_k & A_k & \ldots & A_k \end{bmatrix}. \tag{15}$$

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

$$Y_k(x_1), \ldots, Y_k(x_D) | \boldsymbol{\rho}_k, \sigma_k^2 \sim \mathcal{N}\big(0, C_k + \sigma_k^2 I_{n_k D}\big), \tag{16}$$

thus avoiding inference of μ_k. Let $Y_k(\tau)$ denote the vector of length $n_k \times D$ equal to $[y_1(x_1), \ldots, y_1(x_D), \ldots, y_{n_k}(x_1), \ldots, y_{n_k}(x_D)]$. Then, we may rewrite equation (7) by marginalising μ_k to obtain

$$P(z_i = k | z_{-i}) \propto \pi_k \int p(\mathbf{y}_i | \mu_k) p\big(\mu_k | \boldsymbol{\rho_k}, Y_{-i,k}(\tau)\big) \, \mathrm{d}\mu_k, \tag{17}$$

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

2.4.2. *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 (Zhang, Leithead and Leith (2005)). Full derivations and step-by-step algorithms for computing this inverse and determinant can be found in the Supplementary Material (Crook et al. (2022)).

2.4.3. *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 $k = 1, \ldots, K$.

2.4.4. *Gaussian process hyperparameter inference*. To complete the specification of the GP prior, we need either to fix the hyperparameters a_k^2, l_k and σ_k^2 at the outset or to perform inference for these quantities. We consider two strategies for dealing with the hyperparameters: supervised optimisation and semi-supervised inference.

2.4.4.1. *Supervised approach*: *Optimising the hyperparameters*. Our first strategy is to fix the hyperparameters at the outset, via a maximum marginal likelihood, using only the labelled data. The marginal likelihood can be obtained quickly by recalling that

$$Y_k(x_1), \ldots, Y_k(x_D) | \boldsymbol{\rho_k}, \sigma_k^2 \sim N\big(0, C_k + \sigma_k^2 I_{n_k D}\big). \tag{18}$$

Thus, the log marginal likelihood is given by

$$\begin{aligned} &\log p\big(Y_k | \tau, \boldsymbol{\rho_k}, \sigma_k^2\big) \\ &\quad = -\frac{1}{2} Y_k(\tau) \big(C_k + \sigma_k^2 I_{n_k D}\big)^{-1} Y_k(\tau)^T - \frac{1}{2} \log \big| C_k + \sigma_k^2 I_{n_k D} \big| - \frac{n_k D}{2} \log 2\pi. \end{aligned} \tag{19}$$

For convenience of notation set $\hat{C}_k = C_k + \sigma_k^2 I_{n_k D}$. To maximise the marginal likelihood, given equation (19), we find the partial derivatives with respect to the parameters (Rasmussen (2004)). Hence, we can use a gradient based optimisation procedure. Positivity constraints on a_k^2, l_k, σ_k^2 are dealt with by reparametrisation, and so, dropping the dependence on k for notational convenience and abusing notation, we set $l = \exp(\nu_1)$, $a^2 = \exp(2\nu_2)$ and $\sigma^2 = \exp(2\nu_3)$. Application of the quasi-Newton L-BFGS algorithm (Liu and Nocedal (1989)) 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++ (Eddelbuettel and Francois (2011), Eddelbuettel and Sanderson (2014)).

2.4.4.2. *Semi-supervised approach*: *Bayesian inference of the hyperparameters*. 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 (Rasmussen (2004)). Furthermore, we consider a semi-supervised approach to hyperparameter inference. By a semi-supervised approach we mean that the hyperparameters are 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 kth component Y_k. We can partition this data into the unlabelled (U) and labelled data (L); in particular, $Y_k = [Y_k^{(L)}, Y_k^{(U)}]$. To clarify, the indicators z_i are known for $Y_k^{(L)}$ prior to inference, whilst allocations z_i for $Y_k^{(U)}$ are sampled at each iteration of our MCMC algorithm. In our semi-supervised approach to hyperparameter inference, we use the set Y_k of all data (labelled and unlabelled) currently associated with the kth component. We consider a Hamiltonian Monte Carlo (HMC) sampler for performing inference for these hyperparameters, as described in the Supplementary Material (Crook et al. (2022)), where we also compare to a Metropolis–Hastings sampler.

2.5. *MCMC algorithm for posterior Bayesian computation*. Full details of the procedure(s) for performing inference in our model are provided in the Supplementary Material (Crook et al. (2022)).

2.6. *Summarising uncertainty in posterior localisation probabilities*. Summarising uncertainty quantified by Bayesian analysis in an interpretable way can be challenging. As always, we can summarise uncertainty using credible intervals or regions (Gelman et al. (1995)). One particularly challenging quantity of interest to summarise is the uncertainty in posterior allocations. Whilst each individual allocation of a protein to a subcellular niche can be summarised by a credible interval, it is not clear what is the best way to summarise the posterior over all possible localisations for each individual protein. As in previous work (Crook et al. (2018)), we propose to summarise this uncertainty in an information-theoretic approach by computing the Shannon entropy of the localisation probabilities (Shannon (1948)) at each iteration of the MCMC algorithm

$$\left\{ H_{ik}^{(t)} = -\sum_{k=1}^{K} p_{ik}^{(t)} \log p_{ik}^{(t)} \right\}_{t=1}^{T}, \tag{20}$$

where $p_{ik}^{(t)}$ is the probability that protein i belong to component k at iteration t. We can then summarise this by a Monte Carlo average,

$$H_{ik} \approx \frac{1}{T} \sum_{t=1}^{T} H_{ik}^{(t)}. \tag{21}$$

We note that larger values of a Shannon entropy correspond to greater uncertainty in allocations.

2.7. *Proper scoring rules*. The primary goal of spatial proteomics is to assign proteins with unknown localisations to subcellular niches based on their quantitative functional measurements. Secondary goals include inference of organelle specific parameters and uncertainty quantification, because organelles have overlapping biochemical properties. To measure the ability of methodologies to correctly assign proteins to organelles, we desire a strictly proper and symmetric scoring rule (Gneiting and Raftery (2007)). The symmetry is a requirement because ruling out protein localisations are as important as confident assignments. The quadratic (Brier) loss, spherical loss and logarithmic loss are usually appropriate candidates (Gneiting and Raftery (2007)). We put equal value on whether probabilities are over or under estimated, and so the quadratic loss is appropriate, since the spherical loss puts more weight on lower entropy predictions (penalises underconfident predictions) and the log loss higher entropy predictions (rewards erring on the side of caution) (Gneiting and Raftery (2007), Machete (2013)). The unboundedness of the log loss is also problematic, since assigning potentially infinite penalty to an incorrect prediction is not useful in practice. We define the quadratic loss for a set of probabilistic forecasts $\mathbf{p}$ as

$$B(\mathbf{p}) = \sum_{i=1}^{n}\sum_{j=1}^{K} \|\delta_{ij} - p_{ij}\|_2^2, \tag{22}$$

where $\delta_{ij} = 1$ if protein i localises to component j and is 0 otherwise. It is useful to note a penalty of size 2 is incurred for completely incorrect predictions, that is, forecasting probability 1 for the wrong component. A smaller penalty is incurred for agnostic prediction amongst several classes. For example, suppose protein i localises to organelle 1, but we predict it belongs to organelle 2, 3, 4 and 5 with equal probability; the penalty incurred is 1.25. This is important in practice, because we favour methodologies that avoid us performing erroneous validation experiments.

3. Results.

3.1. *Case study I*: Drosophila melanogaster *embryos*.

3.1.1. *Application*. The first case study is the *Drosophila melanogaster* (common fruit fly) embryos (Tan et al. (2009)) 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 subcellular 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 2 demonstrates several phenomena. Reassuringly, the estimates of the noise parameters σ_k^2 for $k = 1, \ldots, K$, 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.

First, in many cases the posterior, using both the labelled and unlabelled data, is shifted right toward 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

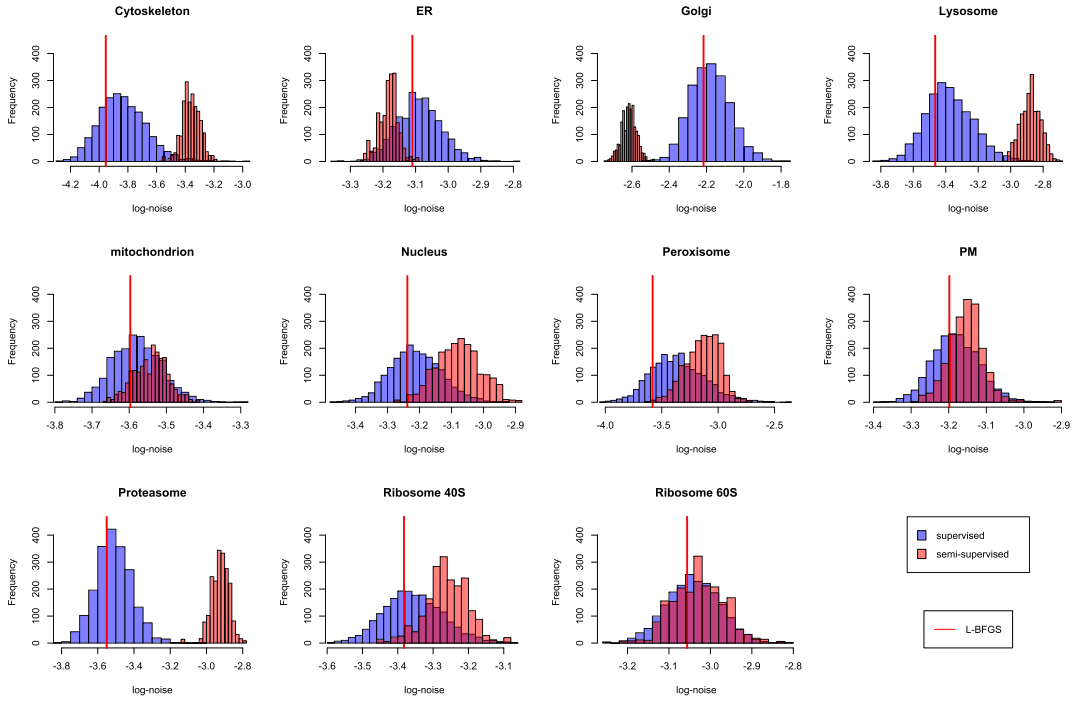

FIG. 2. *Posterior distributions for the log noise parameter σ^2 on the Drosophila data. In general, we observe a shift toward* 0, *indicating that the labelled data underestimates the value of the noise term for each component. We also observe increased posterior shrinkage for many components with the variance of the noise parameters reduced in the semi-supervised setting.*

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 $k = 1, \ldots, K$. This variance reduction is observed, in most cases, even when there 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 subcellular 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 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 ith protein to organelle j (see the Supplementary Material, Crook et al. (2022)).

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

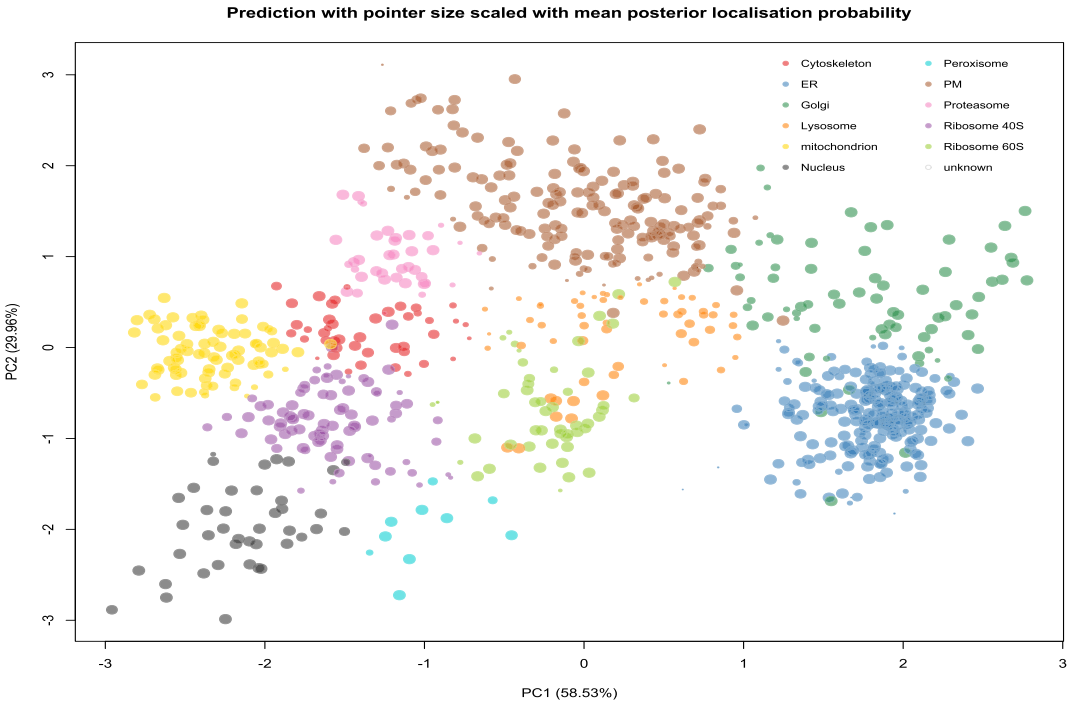

FIG. 3. *A PCA plot for the Drosophila data where points, representing proteins, are shaded by the component of greatest probability. The pointer for each protein is scaled according to membership probability with larger/smaller points indicating greater/lower allocation probabilities.*

nucleus, in Figure 4. We plot the labelled proteins, which were assigned to each component before our analysis, as well as 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 cofractionate 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.

3.1.2. *Sensitivity analysis for hyperprior specification.* We use the *Drosophila melanogaster* dataset to test for sensitivity of the hyperprior specification. To test for sensitivity, we see if predictive performance is affected by changes in the choice of hyperprior. The following cross-validation schema assesses whether predictive performance is affected by choice of

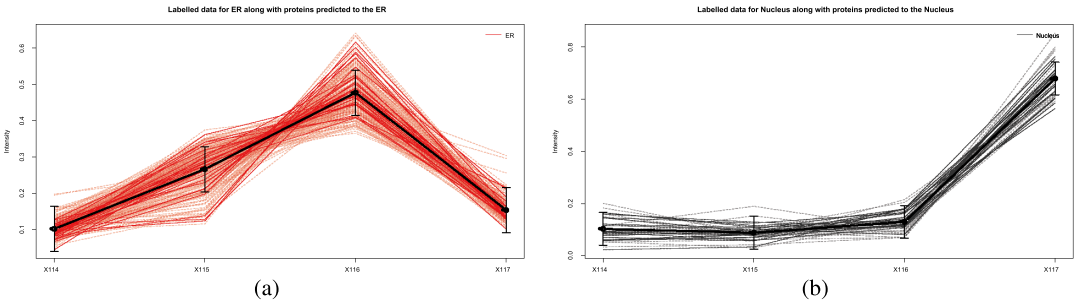

FIG. 4. *A plot of the gradient-density profiles for the ER and Nucleus with labelled proteins (solid lines) and proteins probabilistically assigned to those components (dashed lines). The profiles of the assigned proteins closely match the profiles of the components. The predictive posterior of these components is also overlaid.*

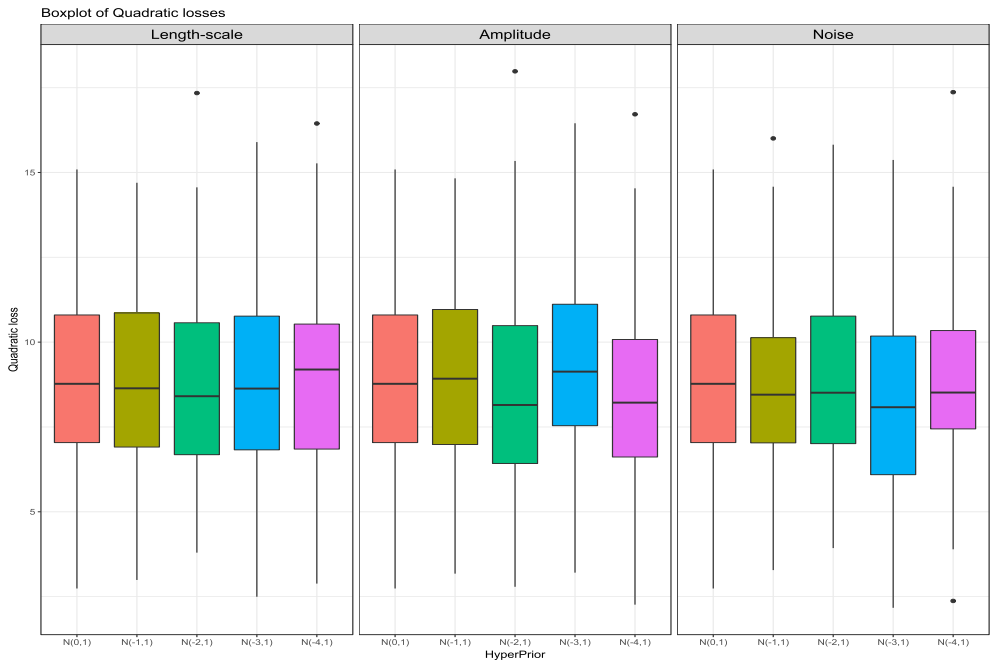

FIG. 5. *Boxplots of quadratic losses to assess the sensitivity of semi-supervised hyperparameter inference to hyperprior choices.*

hyperprior. 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 (Gneiting and Raftery (2007)). 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 hyperprior for each hyperparameter in turn for a grid of values $\tilde{m} = (0, -1, -2, -3, -4)$, keeping the hyperprior for the other variable held the same as a standard normal distribution. The results are displayed in Figure 5.

We observe only minor sensitivity to the choice of hyperprior, 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 (Rasmussen (2004)). 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 (Gelman (2006), Gelman et al. (1995), Lunn et al. (2000), Wang and Dunson (2011), Schuurman, Grasman and Hamaker ()).

3.2. *Case study II*: *Mouse pluripotent embryonic stems cells.*

3.2.1. *Application.* Our main case study is the mouse pluripotent E14TG2a stem cell dataset of Christoforou et al. (2016). This dataset contains 5032 quantitative protein profiles and resolves 14 subcellular niches. We first plot the density-gradient profiles of the marker proteins for each subcellular niche in Figure 6. We fit a Gaussian process prior regression model for each subcellular niche with the hyperparameters found by maximising

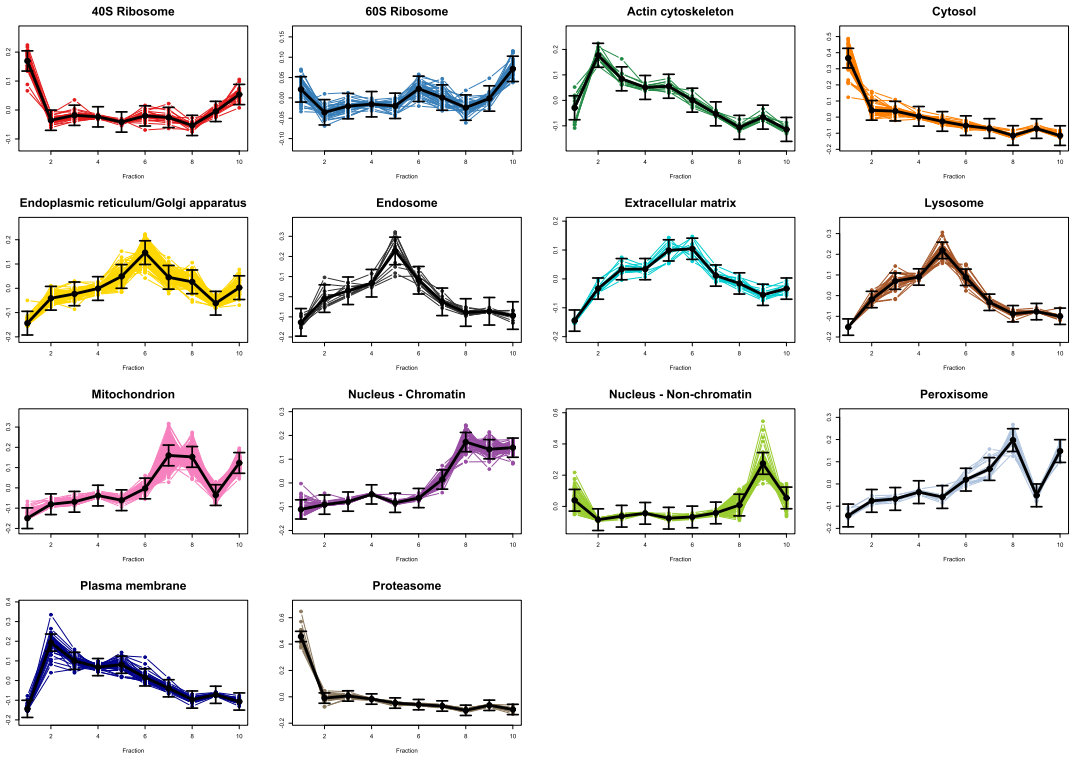

FIG. 6. *Quantitative profiles of protein markers for each subcellular niche. A GP prior regression model is fitted to these data, and the predictive distribution is displayed. We observe distinct distributions for each subcellular niche generated by the unique density-gradient properties of each subcellular niche.*

the marginal likelihood. A table of unconstrained log hyperparameter values found by maximising the marginal likelihood is found in the Supplementary Material (Crook et al. (2022)). 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 subcellular 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% equitailed credible interval which can also be found in the Supplementary Material (Crook et al. (2022)).

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 hyperpriors on the log of the hyperparameters. We run our MCMC algorithm for 20,000 iterations with half taken as burn-in and thin by 5 as well as using HMC to update the hyperparameters. The PCA plot in Figure 7 visualises our results. Each pointer represent a single protein and is scaled either to the probability of membership to the most probable component (left) or scaled with the Shannon entropy (right). As before, we also visualise the allocation probabilities for proteins across organelles in a heatmap (see the Supplementary Material, Crook et al. (2022)). 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.

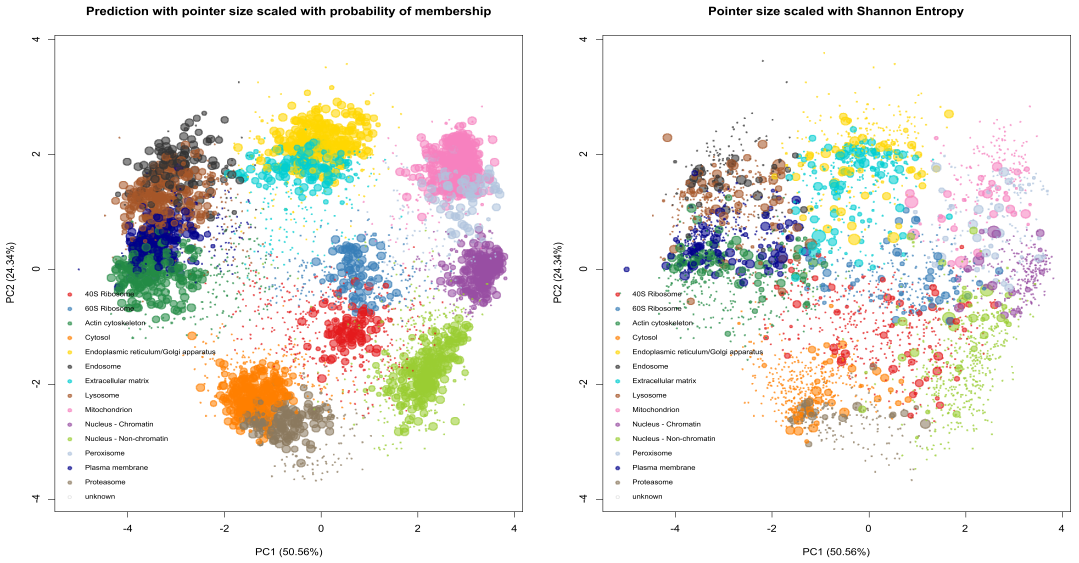

FIG. 7. *A PCA plot for the mouse pluripotent embryonic stem cell data where points, representing proteins, are shaded by the component of greatest probability. The pointer for each protein is scaled with membership probability* (*left*). (*right*) *The pointer for each protein is scaled with the Monte Carlo averaged Shannon Entropy.*

3.3. *Assessing predictive performance.* We compare the predictive performance of the methods proposed here as well as against the fully Bayesian TAGM model of Crook et al. (2018), where subcellular niches are described by multivariate Gaussian distributions rather than GPs. The following cross-validation schema is used to 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 (Gneiting and Raftery (2007)). 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. Prior specifications are stated in the Supplementary Material Crook et al. (2022).

We compare across *five* different spatial proteomics datasets across three different organisms. The datasets we compare our methods on are *Drosophila melanogaster* embryos from Tan et al. (2009), the mouse pluripotent embroyonic stem cell dataset of Christoforou et al. (2016), the HeLa cell line dataset of Itzhak et al. (2016), the mouse primary neuron dataset of Itzhak et al. (2017) and, finally, a CRISPR-CAS9 knock-out coupled to spatial proteomics analysis dataset (AP5Z1-KO1) of Hirst et al. (2018). The results are found in Figure 8.

We see that our in four out 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 Itzhak et al. (2016) and the HeLA AP5Z1 knock-out dataset of Hirst et al. (2018). However, the size of these difference is small, and there is, at most, a *six* point difference. This corresponds to better assignments for, at most, *three* proteins, which we do not believe to be worth the loss in uncertainty quantification in the GP hyperparameters and the lost ability

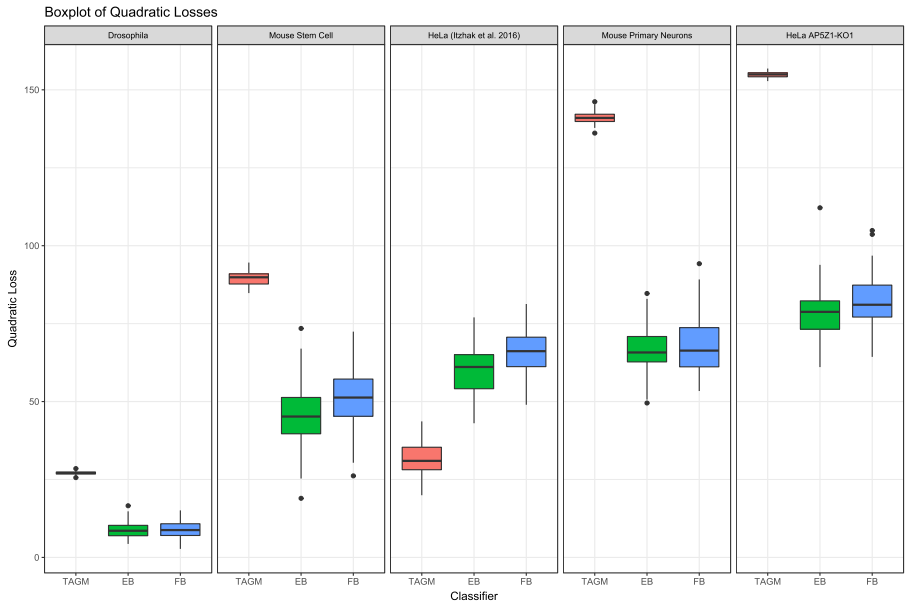

FIG. 8. *Boxplots of quadratic losses comparing predictive performance of the TAGM against the two semi–supervised Gaussian process models described here, where either an empirical Bayes (EB) approach or fully Bayesian (FB) approach is used for hyperparameter inference. That is, (EB) denotes the model where hyperparameters are fixed and learned 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.*

to provide expert prior information on the GP hyperparameters, both of which are provided by the fully Bayesian approach. Meanwhile, the improvement of the GP methods over the TAGM model is marked in the *four* datasets where we see improvement. Improvements range from score differences of roughly 16 to almost 80 which corresponds to *eight* to 40 proteins with better allocations. Moreover, we note that the GP methods have only *three* parameters for the structured covariance to be inferred, whilst the TAGM model requires inference of full unstructured covariance matrices.

We observe that the TAGM model outperforms the GP methods in the Itzhak et al. (2016) 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 subcellular structures, such as ribosomes, as well as cytosolic and nuclear proteins. In any case, our modelling assumptions are violated in both models, and this issue is exacerbated by parameterising 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 subcellular 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 would have to be carefully modelled, perhaps using reversible jump methods (Richardson and Green (1997)) or Dirichlet process approaches (Escobar and West (1995)).

4. Discussion. This article presents semi-supervised nonparametric Bayesian methods to model spatial proteomics data. Subcellular niches display unique signatures along subcellular fractions, and we exploit this information to construct GP regression models for each niche. The full complement of subcellular proteins is then described as 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 nonparametric framework that more closely reflects the biochemical process used to generate

the data. This greatly increases model interpretation and allows us to make more biological sound inferences from our model.

We compared the proposed semi-supervised models to the state-of-the-art model on *five* 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 overinterpreted, 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 decompositions 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 in the Supplementary Material (Crook et al. (2022)) and at the following GitHub repository: https://github.com/ococrook/toeplitz. These algorithms and associated formulae are useful to those outside the spatial proteomics community and 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 approach 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 hyperprior 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 hyperparameters is reflected in the uncertainty of the localisation of proteins to components. Quantifying uncertainty provide cell biologists with a wealth of information to make quantifiable inference about protein subcellular localisation.

We plan to disseminate our method via the Bioconductor project (Gentleman et al. (2004), Huber et al. (2015)) and to include our code in pRoloc package (Gatto et al. (2014b)). The pRoloc package includes methods for visualisation, processing data and disseminating code in a unified framework. All spatial proteomics data used here is freely available within the Bioconductor package pRolocdata (Gatto, Crook and Breckels (2018)).

One potential source of uncertainty in protein localisation is that they can be residents of multiple subcellular compartments. We believe that, by proposing a model which more closely reflects the underlying biochemical rationale for the experiment, we can facilitate models which can infer proteins with multiple locations with greater confidence. This is the subject of further work.

Funding. While completing this work, OMC was a Wellcome Trust Mathematical Genomics and Medicine student supported financially by the School of Clinical Medicine, University of Cambridge. KSL and LG were supported by Wellcome Trust Award 110170/Z/15/Z. PDWK is supported by MRC project reference MC_UU_00002/13, and the National Institute for Health Research (Cambridge Biomedical Research Centre at the Cambridge University Hospitals NHS Foundation Trust).

The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.

SUPPLEMENTARY MATERIAL

Supplement to "Semi-supervised nonparametric Bayesian modelling of spatial proteomics" (DOI: 10.1214/22-AOAS1603SUPP; .zip). In this supplement, we provide the additional derivations, results, and figures referenced in the main text, as well as our code for fast matrix inversion of Toeplitz matrices.

REFERENCES

ASHBURNER, M., BALL, C. A., BLAKE, J. A., BOTSTEIN, D., BUTLER, H., CHERRY, J. M., DAVIS, A. P., DOLINSKI, K., DWIGHT, S. S. et al. (2000). Gene ontology: Tool for the unification of biology. *Nat. Genet.* **25** 25–29.

BARYLYUK, K., KORENY, L., KE, H., BUTTERWORTH, S., CROOK, O. M., LASSADI, I., GUPTA, V., TROMER, E. C., MOURIER, T. et al. (2020). A subcellular atlas of toxoplasma reveals the functional context of the proteome. *BioRxiv*.

BELTRAN, P. M. J., MATHIAS, R. A. and CRISTEA, I. M. (2016). A portrait of the human organelle proteome in space and time during cytomegalovirus infection. *Cell Syst.* **3** 361–373.

BESKOS, A., PILLAI, N., ROBERTS, G., SANZ-SERNA, J.-M. and STUART, A. (2013). Optimal tuning of the hybrid Monte Carlo algorithm. *Bernoulli* **19** 1501–1534. MR3129023 https://doi.org/10.3150/12-BEJ414

BLOBEL, G. (2013). Christian de Duve (1917–2013). *Nature* **498** 300. https://doi.org/10.1038/498300a

BOUVEYRON, C., CÔME, E. and JACQUES, J. (2015). The discriminative functional mixture model for a comparative analysis of bike sharing systems. *Ann. Appl. Stat.* **9** 1726–1760. MR3456352 https://doi.org/10.1214/15-AOAS861

BRECKELS, L. M., GATTO, L., CHRISTOFOROU, A., GROEN, A. J., LILLEY, K. S. and TROTTER, M. W. (2013). The effect of organelle discovery upon sub-cellular protein localisation. *J. Proteomics* **88** 129–140.

BRECKELS, L. M., HOLDEN, S. B., WOJNAR, D., MULVEY, C. M., CHRISTOFOROU, A., GROEN, A., TROTTER, M. W., KOHLBACHER, O., LILLEY, K. S. et al. (2016). Learning from heterogeneous data sources: An application in spatial proteomics. *PLoS Comput. Biol.* **12** e1004920.

CASELLA, G. and ROBERT, C. P. (1996). Rao-Blackwellisation of sampling schemes. *Biometrika* **83** 81–94. MR1399157 https://doi.org/10.1093/biomet/83.1.81

CHRISTOFOROU, A., MULVEY, C. M., BRECKELS, L. M., GELADAKI, A., HURRELL, T., HAYWARD, P. C., NAAKE, T., GATTO, L., VINER, R. et al. (2016). A draft map of the mouse pluripotent stem cell spatial proteome. *Nat. Commun.* **7** 9992.

CODY, N. A., IAMPIETRO, C. and LÉCUYER, E. (2013). The many functions of mRNA localization during normal development and disease: From pillar to post. *Wiley Interdiscip. Rev.: Dev. Biol.* **2** 781–796.

COOK, K. C. and CRISTEA, I. M. (2019). Location is everything: Protein translocations as a viral infection strategy. *Curr Opin Chem Biol* **48** 34–43. https://doi.org/10.1016/j.cbpa.2018.09.021

COOKE, E. J., SAVAGE, R. S., KIRK, P. D. W., DARKINS, R. and WILD, D. L. (2011). Bayesian hierarchical clustering for microarray time series data with replicates and outlier measurements. *BMC Bioinform.* **12** 399. https://doi.org/10.1186/1471-2105-12-399

CORETTO, P. and HENNIG, C. (2016). Robust improper maximum likelihood: Tuning, computation, and a comparison with other methods for robust Gaussian clustering. *J. Amer. Statist. Assoc.* **111** 1648–1659. MR3601724 https://doi.org/10.1080/01621459.2015.1100996

CROOK, O. M., MULVEY, C. M., KIRK, P. D. W., LILLEY, K. S. and GATTO, L. (2018). A Bayesian mixture modelling approach for spatial proteomics. *PLoS Comput. Biol.* **14** 1–29.

CROOK, O. M., GELADAKI, A., NIGHTINGALE, D. J. H., VENNARD, O. L., LILLEY, K. S., GATTO, L. and KIRK, P. D. W. (2020). A semi-supervised Bayesian approach for simultaneous protein sub-cellular localisation assignment and novelty detection. *PLoS Comput. Biol.* **16** e1008288. https://doi.org/10.1371/journal.pcbi.1008288

CROOK, O. M., LILLEY, K. S, GATTO, L. and KIRK, P. D (2022). Supplement to "Semi-Supervised Nonparametric Bayesian Modelling of Spatial Proteomics." https://doi.org/10.1214/22-AOAS1603SUPP

DAVIES, A. K., ITZHAK, D. N., EDGAR, J. R., ARCHULETA, T. L., HIRST, J., JACKSON, L. P., ROBINSON, M. S. and BORNER, G. H. H. (2018). AP-4 vesicles contribute to spatial control of autophagy via RUSC-dependent peripheral delivery of ATG9A. *Nat. Commun.* **9** 3958. https://doi.org/10.1038/s41467-018-06172-7

DE DUVE, C. (1969). The peroxisome: A new cytoplasmic organelle. *Proc. R. Soc. Lond., B Biol. Sci.* **173** 71–83.

DE DUVE, C. and BEAUFAY, H. (1981). A short history of tissue fractionation. *J. Cell Biol.* **91** 293.

DE MATTEIS, M. A. and LUINI, A. (2011). Mendelian disorders of membrane trafficking. *N. Engl. J. Med.* **365** 927–938.

DUANE, S., KENNEDY, A. D., PENDLETON, B. J. and ROWETH, D. (1987). Hybrid Monte Carlo. *Phys. Lett. B* **195** 216–222. MR3960671 https://doi.org/10.1016/0370-2693(87)91197-x

DUNKLEY, T. P., WATSON, R., GRIFFIN, J. L., DUPREE, P. and LILLEY, K. S. (2004). Localization of organelle proteins by isotope tagging (LOPIT). *Mol. Cell. Proteomics* **3** 1128–1134.

DUNKLEY, T. P., HESTER, S., SHADFORTH, I. P., RUNIONS, J., WEIMAR, T., HANTON, S. L., GRIFFIN, J. L., BESSANT, C., BRANDIZZI, F. et al. (2006). Mapping the arabidopsis organelle proteome. *Proc. Natl. Acad. Sci. USA* **103** 6518–6523.

EDDELBUETTEL, D. and FRANCOIS, R. (2011). Rcpp: Seamless R and C++ integration. *J. Stat. Softw.* **40** 1–18.

EDDELBUETTEL, D. and SANDERSON, C. (2014). RcppArmadillo: Accelerating R with high-performance C++ linear algebra. *Comput. Statist. Data Anal.* **71** 1054–1063. MR3132026 https://doi.org/10.1016/j.csda.2013.02.005

ESCOBAR, M. D. and WEST, M. (1995). Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc.* **90** 577–588. MR1340510

FRALEY, C. and RAFTERY, A. E. (2007). Bayesian regularization for normal mixture estimation and model-based clustering. *J. Classification* **24** 155–181. MR2415725 https://doi.org/10.1007/s00357-007-0004-5

GATTO, L., BRECKELS, L. M. and LILLEY, K. S. (2019). Assessing sub-cellular resolution in spatial proteomics experiments. *Curr. Opin. Chem. Biol.* **48** 123–149. https://doi.org/10.1016/j.cbpa.2018.11.015

GATTO, L., CROOK, O. M. and BRECKELS, L. M. (2018). pRolocdata: Data accompanying the pRoloc package. R package version 1.19.1.

GATTO, L., VIZCAÍNO, J. A., HERMJAKOB, H., HUBER, W. and LILLEY, K. S. (2010). Organelle proteomics experimental designs and analysis. *Proteomics* **10** 3957–3969.

GATTO, L., BRECKELS, L. M., BURGER, T., NIGHTINGALE, D. J., GROEN, A. J., CAMPBELL, C., MULVEY, C. M., CHRISTOFOROU, A., FERRO, M. et al. (2014a). A foundation for reliable spatial proteomics data analysis. *Mol. Cell. Proteomics* mcp–M113.

GATTO, L., BRECKELS, L. M., WIECZOREK, S., BURGER, T. and LILLEY, K. S. (2014b). Mass-spectrometry-based spatial proteomics data analysis using pRoloc and pRolocdata. *Bioinformatics* **30** 1322–1324. https://doi.org/10.1093/bioinformatics/btu013

GELADAKI, A., BRITOVSEK, N. K., BRECKELS, L. M., SMITH, T. S. O. L. V., MULVEY, C. M., CROOK, O. M., GATTO, L. and LILLEY, K. S. (2019). Combining LOPIT with differential ultracentrifugation for high-resolution spatial proteomics. *Nat. Commun.* **10** 331.

GELFAND, A. E., KOTTAS, A. and MACEACHERN, S. N. (2005). Bayesian nonparametric spatial modeling with Dirichlet process mixing. *J. Amer. Statist. Assoc.* **100** 1021–1035. MR2201028 https://doi.org/10.1198/016214504000002078

GELFAND, A. E. and SMITH, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* **85** 398–409. MR1141740

GELMAN, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Anal.* **1** 515–533. MR2221284 https://doi.org/10.1214/06-BA117A

GELMAN, A., CARLIN, J. B., STERN, H. S. and RUBIN, D. B. (1995). *Bayesian Data Analysis. Texts in Statistical Science Series*. CRC Press, London. MR1385925

GENTLEMAN, R. C., CAREY, V. J., BATES, D. M., BOLSTAD, B., DETTLING, M., DUDOIT, S., ELLIS, B., GAUTIER, L., GE, Y. et al. (2004). Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biol.* **5** R80.

GIBSON, T. J. (2009). Cell regulation: Determined to signal discrete cooperation. *Trends Biochem. Sci.* **34** 471–482.

GNEITING, T. and RAFTERY, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *J. Amer. Statist. Assoc.* **102** 359–378. MR2345548 https://doi.org/10.1198/016214506000001437

GROEN, A. J., SANCHO-ANDRÉ, G., BRECKELS, L. M., GATTO, L., ANIENTO, F. and LILLEY, K. S. (2014). Identification of trans-Golgi network proteins in Arabidopsis thaliana root tissue. *J. Proteome Res.* **13** 763–776.

HALL, S. L., HESTER, S., GRIFFIN, J. L., LILLEY, K. S. and JACKSON, A. P. (2009). The organelle proteome of the DT40 lymphocyte cell line. *Mol. Cell. Proteomics* **8** 1295–1305.

HEARD, W., SKLENÁŘ, J., TOME, D. F., ROBATZEK, S. and JONES, A. M. (2015). Identification of regulatory and cargo proteins of endosomal and secretory pathways in arabidopsis thaliana by proteomic dissection. *Mol. Cell. Proteomics* **14** 1796–1813.

HEINONEN, M., GUIPAUD, O., MILLIAT, F., BUARD, V., MICHEAU, B., TARLET, G., BENDERITTER, M., ZEHRAOUI, F. and D'ALCHÉ BUC, F. (2014). Detecting time periods of differential gene expression using Gaussian processes: An application to endothelial cells exposed to radiotherapy dose fraction. *Bioinformatics* **31** 728–735.

HENNIG, C. (2004). Breakdown points for maximum likelihood estimators of location-scale mixtures. *Ann. Statist.* **32** 1313–1340. MR2089126 https://doi.org/10.1214/009053604000000571

HENSMAN, J., LAWRENCE, N. D. and RATTRAY, M. (2013). Hierarchical Bayesian modelling of gene expression time series across irregularly sampled replicates and clusters. *BMC Bioinform.* **14** 252. https://doi.org/10.1186/1471-2105-14-252

HIRST, J., ITZHAK, D. N., ANTROBUS, R., BORNER, G. H. H. and ROBINSON, M. S. (2018). Role of the AP-5 adaptor protein complex in late endosome-to-Golgi retrieval. *PLoS Biol.* **16** e2004411. https://doi.org/10.1371/journal.pbio.2004411

HONKELA, A., GIRARDOT, C., GUSTAFSON, E. H., LIU, Y.-H., FURLONG, E. E., LAWRENCE, N. D. and RATTRAY, M. (2010). Model-based method for transcription factor target identification with limited data. *Proc. Natl. Acad. Sci. USA* **107** 7793–7798.

HUBER, W., CAREY, V. J., GENTLEMAN, R., ANDERS, S., CARLSON, M., CARVALHO, B. S., BRAVO, H. C., DAVIS, S., GATTO, L. et al. (2015). Orchestrating high-throughput genomic analysis with bioconductor. *Nat. Methods* **12** 115.

ITZHAK, D. N., TYANOVA, S., COX, J. and BORNER, G. H. (2016). Global, quantitative and dynamic mapping of protein subcellular localization. *eLife* **5**. https://doi.org/10.7554/eLife.16950

ITZHAK, D. N., DAVIES, C., TYANOVA, S., MISHRA, A., WILLIAMSON, J., ANTROBUS, R., COX, J., WEEKES, M. P. and BORNER, G. H. H. (2017). A mass spectrometry-based approach for mapping protein subcellular localization reveals the spatial proteome of mouse primary neurons. *Cell Rep.* **20** 2706–2718. https://doi.org/10.1016/j.celrep.2017.08.063

JADOT, M., BOONEN, M., THIRION, J., WANG, N., XING, J., ZHAO, C., TANNOUS, A., QIAN, M., ZHENG, H. et al. (2017). Accounting for protein subcellular localization: A compartmental map of the rat liver proteome. *Mol. Cell. Proteomics* **16** 194–212.

JAMES, G. M. and HASTIE, T. J. (2001). Functional linear discriminant analysis for irregularly sampled curves. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **63** 533–550. MR1858401 https://doi.org/10.1111/1467-9868.00297

JAMES, G. M. and SUGAR, C. A. (2003). Clustering for sparsely sampled functional data. *J. Amer. Statist. Assoc.* **98** 397–408. MR1995716 https://doi.org/10.1198/016214503000189

JEFFREYS, H. (1946). An invariant form for the prior probability in estimation problems. *Proc. R. Soc. Lond. Ser. A* **186** 453–461. MR0017504 https://doi.org/10.1098/rspa.1946.0056

JONES, M. and RICE, J. A. (1992). Displaying the important features of large collections of similar curves. *Amer. Statist.* **46** 140–145.

KALAITZIS, A. A. and LAWRENCE, N. D. (2011a). A simple approach to ranking differentially expressed gene expression time courses through Gaussian process regression. *BMC Bioinform.* **12** 180. https://doi.org/10.1186/1471-2105-12-180

KALAITZIS, A. A. and LAWRENCE, N. D. (2011b). A simple approach to ranking differentially expressed gene expression time courses through Gaussian process regression. *BMC Bioinform.* **12** 180. https://doi.org/10.1186/1471-2105-12-180

KAU, T. R., WAY, J. C. and SILVER, P. A. (2004). Nuclear transport and cancer: From mechanism to intervention. *Nat. Rev. Cancer* **4** 106–117.

KIRK, P. D. and STUMPF, M. P. (2009). Gaussian process regression bootstrapping: Exploring the effects of uncertainty in time course data. *Bioinformatics* **25** 1300–1306.

KIRK, P., GRIFFIN, J. E., SAVAGE, R. S., GHAHRAMANI, Z. and WILD, D. L. (2012). Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics* **28** 3290–3297.

LATORRE, I. J., ROH, M. H., FRESE, K. K., WEISS, R. S., MARGOLIS, B. and JAVIER, R. T. (2005). Viral oncoprotein-induced mislocalization of select PDZ proteins disrupts tight junctions and causes polarity defects in epithelial cells. *J. Cell Sci.* **118** 4283–4293.

LAURILA, K. and VIHINEN, M. (2009). Prediction of disease-related mutations affecting protein localization. *BMC Genomics* **10** 122. https://doi.org/10.1186/1471-2164-10-122

LAVINE, M. and WEST, M. (1992). A Bayesian method for classification and discrimination. *Canad. J. Statist.* **20** 451–461. MR1208356 https://doi.org/10.2307/3315614

LIU, D. C. and NOCEDAL, J. (1989). On the limited memory BFGS method for large scale optimization. *Math. Program.* **45** 503–528. MR1038245 https://doi.org/10.1007/BF01589116

LIU, Q., LIN, K. K., ANDERSEN, B., SMYTH, P. and IHLER, A. (2010). Estimating replicate time shifts using Gaussian process regression. *Bioinformatics* **26** 770–776.

LUHESHI, L. M., CROWTHER, D. C. and DOBSON, C. M. (2008). Protein misfolding and disease: From the test tube to the organism. *Curr. Opin. Chem. Biol.* **12** 25–31.

LUNN, D. J., THOMAS, A., BEST, N. and SPIEGELHALTER, D. (2000). WinBUGS-a Bayesian modelling framework: Concepts, structure, and extensibility. *Stat. Comput.* **10** 325–337.

MACHETE, R. L. (2013). Contrasting probabilistic scoring rules. *J. Statist. Plann. Inference* **143** 1781–1790. MR3082233 https://doi.org/10.1016/j.jspi.2013.05.012

MALSINER-WALLI, G., FRÜHWIRTH-SCHNATTER, S. and GRÜN, B. (2017). Identifying mixtures of mixtures using Bayesian estimation. *J. Comput. Graph. Statist.* **26** 285–295. MR3640186 https://doi.org/10.1080/10618600.2016.1200472

MENDES, M., PELÁEZ-GARCÍA, A., LÓPEZ-LUCENDO, M., BARTOLOMÉ, R. A., CALVIÑO, E., BARDERAS, R. and CASAL, J. I. (2017). Mapping the spatial proteome of metastatic cells in colorectal cancer. *Proteomics* **17**. https://doi.org/10.1002/pmic.201700094

METROPOLIS, N., ROSENBLUTH, A. W., ROSENBLUTH, M. N., TELLER, A. H. and TELLER, E. (1953). Equation of state calculations by fast computing machines. *J. Chem. Phys.* **21** 1087–1092.

MORRIS, J. S. (2015). Functional regression. *Annu. Rev. Stat. Appl.* **2** 321–359.

MULVEY, C. M., BRECKELS, L. M., GELADAKI, A., BRITOVŠEK, N. K., NIGHTINGALE, D. J. H., CHRISTOFOROU, A., ELZEK, M., DEERY, M. J., GATTO, L. et al. (2017). Using hyperLOPIT to perform high-resolution mapping of the spatial proteome. *Nat. Protoc.* **12** 1110–1135. https://doi.org/10.1038/nprot.2017.026

MURPHY, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. MIT Press, Cambridge.

MURPHY, K. and MURPHY, T. B. (2019). Parsimonious model-based clustering with covariates. *Adv. Data Anal. Classif.*.

NIGHTINGALE, D. J., GELADAKI, A., BRECKELS, L. M., OLIVER, S. G. and LILLEY, K. S. (2019). The subcellular organisation of saccharomyces cerevisiae. *Curr. Opin. Chem. Biol.* **48** 86–95. https://doi.org/10.1016/j.cbpa.2018.10.026

NIKOLOVSKI, N., RUBTSOV, D., SEGURA, M. P., MILES, G. P., STEVENS, T. J., DUNKLEY, T. P., MUNRO, S., LILLEY, K. S. and DUPREE, P. (2012). Putative glycosyltransferases and other plant Golgi apparatus proteins are revealed by LOPIT proteomics. *Plant Physiol.* **160** 1037–1051.

OHTA, S., BUKOWSKI-WILLS, J.-C., SANCHEZ-PULIDO, L., DE LIMA ALVES, F., WOOD, L., CHEN, Z. A., PLATANI, M., FISCHER, L., HUDSON, D. F. et al. (2010). The protein composition of mitotic chromosomes determined using multiclassifier combinatorial proteomics. *Cell* **142** 810–821.

OLKKONEN, V. M. and IKONEN, E. (2006). When intracellular logistics fails-genetic defects in membrane trafficking. *J. Cell Sci.* **119** 5031–5045.

ORRE, L. M., VESTERLUND, M., PAN, Y., ARSLAN, T., ZHU, Y., WOODBRIDGE, A. F., FRINGS, O., FREDLUND, E. and LEHTIÖ, J. (2019). SubCellBarCode: Proteome-wide mapping of protein localization and relocalization. *Mol. Cell* **73** 166–182.e7.

PARSONS, H., FERNÁNDEZ-NIÑO, S. and HEAZLEWOOD, J. (2014). Separation of the plant Golgi apparatus and endoplasmic reticulum by free-flow electrophoresis. *Methods Mol. Biol.* (*Clifton N.J.*) **1072** 527.

PREDA, C., SAPORTA, G. and LÉVÉDER, C. (2007). PLS classification of functional data. *Comput. Statist.* **22** 223–235. MR2318457 https://doi.org/10.1007/s00180-007-0041-4

RAMSAY, J. O. (2004). Functional data analysis. *Encyc. Stat. Sci.* **4**.

RASMUSSEN, C. E. (2004). Gaussian processes in machine learning. In *Advanced Lectures on Machine Learning* 63–71. Springer, Berlin.

RASMUSSEN, C. E. and WILLIAMS, C. K. I. (2006). *Gaussian Processes for Machine Learning*. *Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. MR2514435

RICHARDSON, S. and GREEN, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components. *J. Roy. Statist. Soc. Ser. B* **59** 731–792. MR1483213 https://doi.org/10.1111/1467-9868.00095

RODRIGUEZ, J. A., AU, W. W. Y. and HENDERSON, B. R. (2004). Cytoplasmic mislocalization of BRCA1 caused by cancer-associated mutations in the BRCT domain. *Exp. Cell Res.* **293** 14–21. https://doi.org/10.1016/j.yexcr.2003.09.027

RODRÍGUEZ, A., DUNSON, D. B. and GELFAND, A. E. (2009). Bayesian nonparametric functional data analysis through density estimation. *Biometrika* **96** 149–162. MR2482141 https://doi.org/10.1093/biomet/asn054

SADOWSKI, P. G., DUNKLEY, T. P., SHADFORTH, I. P., DUPREE, P., BESSANT, C., GRIFFIN, J. L. and LILLEY, K. S. (2006). Quantitative proteomic approach to study subcellular localization of membrane proteins. *Nat. Protoc.* **1** 1778–1789.

SCHUURMAN, N. K., GRASMAN, R. P. P. P. and HAMAKER, E. L. A comparison of inverse-Wishart prior specifications for covariance matrices in multilevel autoregressive models. *Multivar. Behav. Res.* **51** 185–206. https://doi.org/10.1080/00273171.2015.1065398

SHANNON, C. E. (1948). A mathematical theory of communication. *Bell Syst. Tech. J.* **27** 379–423. MR0026286 https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

SHIN, S. J., SMITH, J. A., REZNICZEK, G. A., PAN, S., CHEN, R., BRENTNALL, T. A., WICHE, G. and KELLY, K. A. (2013). Unexpected gain of function for the scaffolding protein plectin due to mislocalization in pancreatic cancer. *Proc. Natl. Acad. Sci. USA* **110** 19414–19419.

SHIN, J. J., CROOK, O. M., BORGEAUD, A., CATTIN-ORTOLÁ, J., PEAK-CHEW, S.-Y., CHADWICK, J., LILLEY, K. S. and MUNRO, S. (2019). Determining the content of vesicles captured by golgin tethers using LOPIT-DC. *BioRxiv* 841965.

SILJEE, J. E., WANG, Y., BERNARD, A. A., ERSOY, B. A., ZHANG, S., MARLEY, A., VON ZASTROW, M., REITER, J. F. and VAISSE, C. (2018). Subcellular localization of MC4R with ADCY3 at neuronal primary cilia underlies a common pathway for genetic predisposition to obesity. *Nat. Genet.*.

STEEL, M. F. and FUENTES, M. (2010). *Non-Gaussian and Nonparametric Models for Continuous Spatial Data*. CRC Press, Boca Raton, FL.

STEGLE, O., DENBY, K. J., COOKE, E. J., WILD, D. L., GHAHRAMANI, Z. and BORGWARDT, K. M. (2010). A robust Bayesian two-sample test for detecting intervals of differential gene expression in microarray time series. *J. Comput. Biol.* **17** 355–367. MR2609139 https://doi.org/10.1089/cmb.2009.0175

TAN, D. J., DVINGE, H., CHRISTOFOROU, A., BERTONE, P., MARTINEZ ARIAS, A. and LILLEY, K. S. (2009). Mapping organelle proteins and protein complexes in drosophila melanogaster. *J. Proteome Res.* **8** 2667–2678.

TARDIF, M., ATTEIA, A., SPECHT, M., COGNE, G., ROLLAND, N., BRUGIÈRE, S., HIPPLER, M., FERRO, M., BRULEY, C. et al. (2012). PredAlgo: A new subcellular localization prediction tool dedicated to green algae. *Mol. Biol. Evol.* **29** 3625–3639.

THOMPSON, A., SCHÄFER, J., KUHN, K., KIENLE, S., SCHWARZ, J., SCHMIDT, G., NEUMANN, T. and HAMON, C. (2003). Tandem mass tags: A novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal. Chem.* **75** 1895–1904.

THUL, P. J., ÅKESSON, L., WIKING, M., MAHDESSIAN, D., GELADAKI, A., AIT BLAL, H., ALM, T., ASPLUND, A., BJÖRK, L. et al. (2017). A subcellular map of the human proteome. *Science*.

TOPA, H., JÓNÁS, Á., KOFLER, R., KOSIOL, C. and HONKELA, A. (2015). Gaussian process test for high-throughput sequencing time series: Application to experimental evolution. *Bioinformatics* **31** 1762–1770.

WANG, J.-L., CHIOU, J.-M. and MÜLLER, H.-G. (2016). Functional data analysis. *Annu. Rev. Stat. Appl.* **3** 257–295.

WANG, L. and DUNSON, D. B. (2011). Fast Bayesian inference in Dirichlet process mixture models. *J. Comput. Graph. Statist.* **20** 196–216. Supplementary material available online. MR2816545 https://doi.org/10.1198/jcgs.2010.07081

WILLIAMS, C. K. and RASMUSSEN, C. E. (1996). Gaussian processes for regression. In *Advances in Neural Information Processing Systems* 514–520.

ZHANG, Y., LEITHEAD, W. E. and LEITH, D. J. (2005). 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. 44*th IEEE Conference on* 3711–3716. IEEE, Los Alamitos.

ZHU, H., BROWN, P. J. and MORRIS, J. S. (2012). Robust classification of functional and quantitative image data using functional mixed models. *Biometrics* **68** 1260–1268. MR3040032 https://doi.org/10.1111/j.1541-0420.2012.01765.x

ZHU, H., VANNUCCI, M. and COX, D. D. (2010). A Bayesian hierarchical model for classification with selection of functional predictors. *Biometrics* **66** 463–473. MR2758826 https://doi.org/10.1111/j.1541-0420.2009.01283.x