Semi-supervised Bayesian integration of multiple spatial proteomics datasets

Stephen D. Coleman[1*], Lisa Breckels[2], Ross F. Waller[2], Kathryn S. Lilley[2], Chris Wallace[1,3,4], Oliver M. Crook[5†], Paul D.W. Kirk[1,3,4†]

1 MRC Biostatistics Unit, University of Cambridge, Cambridge, UK

2 Department of Biochemistry, University of Cambridge, Cambridge, UK

3 Department of Medicine, University of Cambridge, Cambridge, UK

4 Cambridge Institute of Therapeutic Immunology & Infectious Disease, Cambridge, UK

5 Department of Statistics, University of Oxford, Oxford, UK

† These authors contributed equally to this work.

* Correspondence: colemste@ohsu.edu

Abstract

The subcellular localisation of proteins is a key determinant of their function. High-throughput analyses of these localisations can be performed using mass spectrometry-based spatial proteomics, which enables us to examine the localisation and relocalisation of proteins. Furthermore, complementary data sources can provide additional sources of functional or localisation information. Examples include protein annotations and other high-throughput 'omic assays. Integrating these modalities can provide new insights as well as additional confidence in results, but existing approaches for integrative analyses of spatial proteomics datasets are limited in the types of data they can integrate and do not quantify uncertainty. Here we propose a semi-supervised Bayesian approach to integrate spatial proteomics datasets with other data sources, to improve the inference of protein sub-cellular localisation. We demonstrate our approach outperforms other transfer-learning methods and has greater flexibility in the data it can model. To demonstrate the flexibility of our approach, we apply our method to integrate spatial proteomics data generated for the parasite *Toxoplasma gondii* with time-course gene expression data generated over its cell cycle. Our findings suggest that proteins linked to invasion organelles are associated with expression programs that peak at the end of the first

cell-cycle. Furthermore, this integrative analysis divides the dense granule proteins into heterogeneous populations suggestive of potentially different functions. Our method is disseminated via the `mdir` R package available on the lead author's Github.

Author summary

Proteins are located in subcellular environments to ensure that they are near their interaction partners and occur in the correct biochemical environment to function. Where a protein is located can be determined from a number of data sources. To integrate diverse datasets together we develop an integrative Bayesian model to combine the information from several datasets in a principled manner. We learn how similar the dataset are as part of the modelling process and demonstrate the benefits of integrating mass-spectrometry based spatial proteomics data with timecourse gene-expression datasets.

Introduction

Proteins are key to all cellular processes including cell proliferation and survival. To function correctly, a protein must interact with other binding partners and substrates which requires them to localise to the correct subcellular compartment [1]. For example, signalling and metabolic pathways are dependent on the location of the constituent proteins [2]. Mislocation to the incorrect compartment is increasingly associated with disease, including cancer and obesity [3, 4, 5, 6, 7]. Thus, understanding the location of proteins within the cell is of use in the development of therapeutic targets and determining disease aetiology [8]. Mass-spectrometry (**MS**) based spatial proteomics can be used to study the localisation and relocalisation of proteins on a proteome- and system-wide scale. These experiments leverage so-called marker proteins with an unambiguous single localisation and then apply machine learning approaches, such as the support vector machine (**SVM**), to predict the localisation of annotated proteins.

In a typical MS-based spatial proteomics experiment cells are first gently lysed, to dissociate the cellular compartments while preserving organelle integrity. This cellular material is then separated into subcellular fractions using one of a range of methods such as differential centrifugation [9, 10] or equilibrium density centrifugation [11, 12]. Following subcellular fractionation, the proteins are digested to peptides and subject to high accuracy quantitative MS generating a signature of abundance across the subcellular fractions which correlates with subcellular localisation. A wide variety of methods exploit these central principles, including protein correlation profiling, SubCellBarCode, dynamic organeller maps amongst others [11, 13, 9, 14, 15]. In the localisation of organelle proteins by isotope tagging (**LOPIT**) [11, 16] and

hyperplexed LOPIT (**hyper-LOPIT**) [12, 17] approaches, cell lysis is preceded by the separation of subcellular components along a continuous density gradient. Discrete fractions along this gradient are then collected and multiplexed quantification is achieved using tandem mass tags (**TMT**) [18]. Protein distributions revealing organelle specific profiles within the fractions are acquired by using synchronous precursor selection mass-spectrometry (**SPS-MS3**). Proteins belonging to the same organelle possess shared quantitative abundance profiles [19, 20], it is these (normalised) abundance profiles that are the input to machine learning algorithms to predict the localisation of the unannotated proteins. This means the marker proteins form a labelled subset of the data which can be used as a training set for supervised methods (e.g., partial least squares discriminant analysis [16, 21]; the support vector machine [12, 9]). Using a supervised method assumes that both the marker proteins are fully representative of their compartment, that all compartments have known marker proteins and that all subcellular niches are observed in the labelled set. Supervised classification remains the most common type of Machine Learning approach to prediction tasks in proteomics [22]. If we do not believe these assumptions hold, we can employ semi-supervised methods to both infer the unobserved localisations and update the parameters describing the components simultaneously [23, 24]. These methods can typical uncover unannotated subcellular niches and have better accuracy at inferring the underlying model parameters.

'Omic technologies that infer, say, gene expression, report on different properties of proteins that can help us to understand the aetiology of diseases and conditions that cause functional changes in the system, e.g., cancers, neurodegenerative diseases and immune-mediated diseases, etc. [25, 26, 27, 28, 29, 30]. Deepening understanding of gene expression might elucidate the mechanisms by which viruses and some parasites hijack host organelles to replicate [31]. Additionally they can have implication for disease diagnosis and therapeutic success [32, 33]. More generally, 'omic measurements can give us an integrated view on the molecular mechanisms of biological processes [34]. However, integrating diverse datasets (sometimes referred to as modalities in the machine learning literature) is challenging because the different measurement processes generate heterogeneous data structures and each method has its own set of limitations. Simply concatenating the datasets assumes each dataset contributes equal information towards inference and ignores dataset specific variability.

Despite the successes of multi-omic data integration [35], most methods are unsupervised and tend to infer so-called latent factors that explain variation in the data; that is, a de-convolution of the dataset into a low-dimensional representation [36, 37] or to identify shared clusters [38, 39, 40]. Here, we perform integrative analysis of MS-based spatial proteomics data with other datasets (e.g. Gene Ontology (**GO**) annotations and Gene-expression datasets) as they contain powerful signals that may uncover subcellular regulation of

processes that are not apparent from considering another dataset in isolation. Furthermore, the signal in one MS-based spatial proteomics dataset may be boosted by considering multiple datasets simultaneously. More generally, we wish to jointly model structure in across datasets and uncover shared clusters, groups or regulation.

When considering pairs of datasets (though we can consider larger combinations), we can have one of the following three scenarios. Either both our modalities have some observed classes, neither do, or one modality does. MS-based spatial proteomic data are annotated with marker proteins and so at least one of our modalities contains some known structure which should be incorporated into the modelling. This motivates developing semi-supervised integrative methodology and account for dataset-specific models. Our approach is applicable beyond MS-based spatial proteomics to general semi-supervised integrative tasks and is implemented in an R-package: `mdir` available from `https://github.com/stcolema/MDIr`.

The paper is organised as follows. In the Materials and methods section the models used and data analysed in this paper are presented. This includes a description of the simulation study comparing some of the models, the datasets used to validate our approach on real data and our case study, two 'omics datasets for the model apicomplexan, *Toxoplasma gondii*. The Results section deals with the results from analysing these data showing the advantage of leveraging information from observed markers/labels and across datasets via integrative modelling. In the analysis of *T. Gondii*, we find evidence for dividing the dense granule proteins into heterogeneous populations with potentially different functions. Finally, in the Discussion we consider the results described in preceding sections.

Materials and methods

Methods

We introduce Gaussian mixture models as the foundation of the TAGM [23], GP mixture and (multiple dataset integration) MDI [40] models, which we then introduce in turn. Finally we introduce the KNN transfer learning algorithm, a supervised method for multi-modal data in spatial proteomics.

Mixture models

Let $X = (X_1, \ldots, X_N)$ be the observed data, with $X_n = [X_{n,1}, \ldots, X_{n,P}]^\top$ for each item being considered, where each observation has P variables. We wish to model the data using a mixture of densities:

$$p(X_n|\theta = \{\theta_1, \ldots, \theta_K\}, \pi) = \sum_{k=1}^{K} \pi_k f(X_n|\theta_k) \tag{1}$$

independently for each $n = 1, \ldots, N$. Here, $f(\cdot)$ is a probability density function such as the Gaussian density function or the categorical density function, and each component has its own weight, π_k, and set of parameters, θ_k. The component weights are restricted to the unit simplex, i.e., $\sum_{k=1}^{K} \pi_k = 1$. To capture the discrete structure in the data, we introduce an allocation variable, $c = [c_1, \ldots, c_N]^\top, c_n \in [1, K] \subset \mathbb{N}$, to indicate which component a sample was drawn from, introducing conditional independence between the components,

$$p(c_n = k) = \pi_k, \tag{2}$$

$$p(X_n|c_n = k, \theta_k) = f(X_n|\theta_k). \tag{3}$$

The joint model can then be written

$$p(X, c, K, \pi, \theta) = p(X|c, \pi, K, \theta)p(\theta|c, \pi, K)p(c|\pi, K)p(\pi|K)p(K). \tag{4}$$

We assume conditional independence between certain parameters such that the model reduces to

$$p(X, c, \theta, \pi, K) = p(\pi|K)p(\theta|K)p(K) \prod_{n=1}^{N} p(X_n|c_n, \theta_{c_n})p(c_n|\pi). \tag{5}$$

In terms of the hierarchical model this is:

$$X_n|c_n, \theta \sim F(\theta_{c_n}), \tag{6}$$

$$c_n|\pi \sim \text{Categorical}(\pi_1, \ldots, \pi_K), \tag{7}$$

$$\pi_1, \ldots, \pi_K \sim \text{Dirichlet}(\alpha/K, \ldots, \alpha/K), \tag{8}$$

$$\theta_k \sim G^{(0)}, \tag{9}$$

where F is the appropriate distribution (e.g., Gaussian, Categorical, etc.) and $G^{(0)}$ is some prior over the component parameters. We use the empirical priors suggested by Fraley and Raftery [41] for the Gaussian mixture models. For the Categorical model we use the average proportion of each class for each feature to define our prior density.

T-augmented Gaussian mixture model

Crook et al. [23] proposed capturing outliers within the Gaussian mixture model using a heavy-tailed multivariate t (**MVT**) distribution in their t-augmented Gaussian mixture (**TAGM**) model. They consider each component a mixture of a multivariate normal (**MVN**) density and this MVT which has common parameters across all components, i.e.,

$$p(X_n|c_n = k, \cdot) = \begin{cases} f_t(X_n|M, V, \eta), & \text{if the } n^{th} \text{ item is an outlier} \\ f_{\mathcal{N}}(X_n|\mu_k, \Sigma_k) & \text{otherwise,} \end{cases} \tag{10}$$

where μ_k and Σ_k are the mean vector and covariance matrix of the kth component, and M, V and η are the parameters of the MVT which are set empirically and not sampled. Being considered an outlier is an inferred quantity indicated through the $\psi = [\psi_1, \ldots, \psi_N]^\top$ binary variable. Note that the items of known label (in our applications, the marker proteins) are never considered outliers. This extends the hierarchical model described above to:

$$(X_n|c_n = k, \psi_n\mu_k, \Sigma_k) \sim N(\mu_k, \Sigma_k)^{\psi_n}\mathcal{T}(\eta, M, V)^{1-\psi_n}, \tag{11}$$

$$c_n|\pi \sim \text{Categorical}(\pi_1, \ldots, \pi_K), \tag{12}$$

$$\psi_n|\epsilon \sim \text{Bernoulli}(1 - \epsilon), \tag{13}$$

$$\pi_1, \ldots, \pi_K \sim \text{Dirichlet}(\alpha/K, \ldots, \alpha/K), \tag{14}$$

$$(\mu_k, \Sigma_k) \sim \mathcal{NIW}(\mu_0, \kappa, \nu, \Psi), \tag{15}$$

where we set η to 4 to ensure heavy-tails, and set M to empirical mean and V to half of the global variance of the data following Crook et al. [23].

Gaussian process

The Gaussian Process (**GP**) is a generalisation of the Gaussian probability distribution over functions or infinite vectors rather than finite vectors. The GP is immensely flexible and can describe measurements of arbitrary proximity, most relevant in modelling temporal or spatial data. Specifically, the GP is a continuous stochastic process with index set T such that its realisation at any finite collection of points $\{t_1, \ldots, t_P\} \subset T$ is jointly Gaussian. It is uniquely specified by a mean function $m(t)$ and a positive semi-definite covariance kernel $C(\cdot, \cdot)$, which determine the mean vectors and covariance matrices of the associated multivariate Gaussian distributions. That is, $f(x)$ is a GP with mean function $m(t)$ and covariance kernel $C(\cdot, \cdot)$ if for

any $\{t_1, \ldots, t_P\} \subset T$

$$f(t_1, \ldots, t_P) \sim \mathcal{N}((m(t_1), \ldots, m(t_P)), \Sigma), \tag{16}$$

$$\Sigma_{ij} = C(t_i, t_j). \tag{17}$$

GP models have been used in a range of problems such as systems biology analyses [42, 43, 44, 45, 46] environmental and health studies [47, 48], and monitoring patient health [49, 50] and are built upon rigorous theoretical foundations in terms of support and contraction rates [51, 52, 53].

In practice, a prior mean function of $m(x) = 0$ is commonly used and the covariance function is chosen from a small family of easy to implement functions such as the squared exponential function [54]. Lacking any strong belief about periodicity, symmetry or other constraints on the functions, we use the squared exponential function which is defined for two time points/locations t, t', as

$$c(t, t') = a^2 \exp\left\{-\frac{d(t, t')^2}{2l^2}\right\}, \tag{18}$$

where a and l are the amplitude and length-scale respectively. We use the most common choice for the metric $d(\cdot, \cdot)$, Euclidean distance or the l^2-norm.

Gaussian process mixture models

If we consider the mean vector of a Gaussian density as the finite realisation of an infinite function in our feature space, we can place a Gaussian process prior over the space of functions it is drawn from. Thus, for the kth component of this mixture, mean vector μ_k and observed data which has been transformed to a vector, $X_k = [X_{k_1}, \ldots, X_{k_{N_k}}]$,

$$X_k | \mu_k, \sigma_k \sim \mathcal{N}(\mu_k, \sigma_k^2 I_P), \tag{19}$$

$$\mu_k | a_k, l_k \sim GP(0, C_k). \tag{20}$$

As we assume exchangeable data for ease of notation we permute the original dataset such that we can denote X_k with contiguous indices, i.e., $X_k = [X_1, \ldots, X_{N_k}]$. The posterior distribution can be derived by using the properties of the multivariate Gaussian relating to conditional distribution of two jointly distributed

Gaussian variables [55],

$$\begin{bmatrix} y \\ f(X^*) \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} C(X,X) + \sigma^2\mathbf{I} & C(X,X^*) \\ C(X,X^*) & C(X^*,X^*) \end{bmatrix}\right), \tag{21}$$

$$\implies \begin{cases} \mathbb{E}(f(X^*)|y) & = C(X^*,X)(C(X,X)+\sigma^2\mathbf{I})^{-1}y, \\ \mathbb{C}ov(f(X^*)|y) & = C(X^*,X^*) - C(X^*,X)(C(X,X)+\sigma^2\mathbf{I})^{-1}C(X,X^*). \end{cases} \tag{22}$$

We can then say

$$(\mu_k|X_k,\sigma_k,a_k,l_k) \sim GP(m_k,\tilde{C}_k), \tag{23}$$

$$m_k = C_k[1:P](C_k+\sigma_k^2 I_{N_K P})^{-1}X_k, \tag{24}$$

$$\tilde{C}_k = C_k[1:P,1:P] - C_k[1:P,](C_k+\sigma_k^2 I_{N_K P})^{-1}C_k[,1:P]. \tag{25}$$

This involves an expensive calculation, taking the inverse of an $N_k P \times N_k P$ matrix. There are methods to scale GP models, e.g., sparse approximations [56] and low-dimensional approximations [57], see Liu et al. [58] for a review of the subject. However, in our application we have rich structure in the data as all items within a modality have the same time/spatial measurements with consistent distance between each measurement and we can use this to reduce the computational load without using any of these more complex methods [59], we can instead rewrite the posterior parameters as

$$m_k = \frac{N_k}{\sigma_k^2}(A_k - A_k \times (I_p - Q^{-1}))\bar{X}_k, \tag{26}$$

$$\tilde{C}_k = A_k - \frac{N_k}{\sigma_k^2}(A_k - A_k \times (I_p - Q^{-1}))A_k. \tag{27}$$

Please see the supplementary material section S.1.1 for this derivation.

We sample the hyper-parameters (a^2, l^2, σ^2) according to a Metropolis-Hastings scheme. Standard log-normal hyperpriors were used.

Multiple Dataset Integration

MDI is a Bayesian integrative or multi-modal clustering method. Bayesian statistics allows quantification of uncertainty through the use of probability distributions and offers a modular framework for data analysis by making dependencies between data and parameters explicit [60]. In MDI, signal sharing is defined by

the prior on the cluster label of the nth observation in the M modalities/datasets:

$$p(c_n^{(1)}, \ldots, c_n^{(M)}|\cdot) = \prod_{m=1}^{M} \gamma_{c_n^{(m)}}^{(m)} \prod_{m=1}^{M-1} \prod_{l=m+1}^{M} (1 + \phi_{(m,l)} \mathbb{I}(c_n^{(m)} = c_n^{(l)})), \tag{28}$$

where $c_n^{(m)}$ is the label of the nth observation in the mth modality, $\gamma_k^{(m)}$ is the weight of the kth cluster in the mth modality, $\phi_{(m,l)}$ is the similarity between the clusterings of the mth and l^{th} modalities and $\mathbb{I}(x)$ is the indicator function returning 1 if x is true and 0 otherwise. Attractively, $\phi_{(m,l)}$ is inferred from the data and if there is no shared signal it will tend towards 0 giving us no dependency between these modalities in which case each dataset will have independent clusters.

Within each dataset/modality MDI uses a specific density to model the probability of allocation to a specific cluster. For example, one might wish to model continuous data with Gaussian densities, categorical data with Categorical distributions and/or time-series data with Gaussian processes. Thus we have flexibility in how we model each dataset within the MDI framework.

We implement this in C++ with the novel functionality to allow for observed classes and labels in a subset of datasets/modes and make this available through the R package `MDIr`. In the Bayesian paradigm, moving from an unsupervised cluster method to a semi-supervised method can be phrased as a missing data problem. Let $I = \{i_1, \ldots, i_N\}$ be a vector indicating if the label c_n is observed or not. If $i_n = 0$ (i.e. the label is unobserved), then c_n is sampled as usual, but if $i_n = 1$ then c_n is known and held fixed. Observing labels means that we can no longer arbitrarily switch labels as the labels are inherently linked to a specific class (in our case, subcellular niches). The Gibbs sampler to perform inference on the model is implemented in C++ and includes a correction to the conditional distribution of the $\phi(m, l)$ parameters which was incorrect in previous implementations as observed by Stephen Johnson, Daniel Henderson, and Richard Boys in 2017 (see supplement section S.1.2).

As per Kirk et al. [40], we use an overfitted mixture model [61] to approximate a Dirichlet process mixture model [62, 63, 64, 65] in each unsupervised modality, allowing inference of the number of clusters/classes present. If a number of classes are observed but one believes that additional classes might be present in the data one may use a semi-supervised overfitted mixture to detect novel classes as per Crook et al. [66].

Markov chain Monte Carlo methods

We use Markov chain Monte Carlo (**MCMC**) methods [67, 68, 69, 70] to perform inference on our Bayesian models. MCMC is the gold standard for Computational Bayes [71, 72, 73, 74], constructing a Markov chain of sampled parameters to describe the conditional posterior density for model parameters, with asymptotic

guarantees that this description is exact. However, in practice MCMC is prone to becoming trapped in local modes of the target density. Designing algorithms that perform better in this scenario is an area of active research [75, 76, 77, 78, 79, 80, 81], but implementing these methods can be time-consuming and with no guarantee of overcoming the problem. We instead use Gibbs sampling [82] in our implementation of MDI and the consensus clustering approach to circumvent multi-modality [83]. We can use the collection of sampled values to infer a point estimate and provide a quantification of uncertainty about this. Common choices of point estimate for continuous parameters are the median or the mean, but to obtain a summary clustering we instead use that which minimises the variation 134 of information [84], which has been found to have desirable properties [85], though alternative approaches exist [86, 87]. To summarise the uncertainty about a partition, the posterior similarity matrix (**PSM**) is often used. This considers the average co-clustering probability of each pair of items based on the sampled clusterings. It is an $N \times N$ symmetric matrix. For a classification, we assign the samples to the class with the highest average probability of allocation across the retained MCMC samples.

Transfer learning using a k-nearest neighbours framework

The k-nearest neighbours (**KNN**) classifier is a supervised method in which predictions are made using the majority vote of an item's k nearest labelled neighbours. The KNN-transfer learning (**KNN-TL**) classifier extends this to transfer information from an auxiliary modality to a primary modality [88, 89]. To use the transfer learning algorithm, the best choice of k_1, k_2, the number of neighbours considered when classifying items with missing labels for each modality, must be estimated from the KNN classifier to be used as inputs into the KNN-TL classifier. This is performed via a grid search on a pre-specified grid, with the optimal value selected through cross-validation. Breckels et al. [89] adapted this method to predict protein localisation using LOPIT data as the primary dataset and GO terms as the auxiliary dataset (in contrast to MDI, the KNN-TL classifier does consider an ordering of datasets). In this context, this algorithm finds the k_1 nearest marker proteins to the nth protein in the LOPIT data and the k_2 nearest marker proteins in the GO data. The representation of each class among these neighbours is calculated, i.e., if Z_1, Z_2 are the sets of neighbours in the first and second modalities respectively,

$$p(c_n^{(v)} = j) = \frac{1}{|Z_v|} \sum_{i \in Z_v} c_i^{(v)} = j, \text{for } j = 1, \ldots, K^{(v)}, v = 1, 2, \tag{29}$$

from which a localisation is predicted

$$p(c_n = j) = \theta_j p(c_n^{(1)} = j) + (1 - \theta_j) p(c_n^{(2)} = j) \tag{30}$$

where $\theta \in [0, 1]$ is estimated using a grid search (again, the optimal value is selected through cross-validation on the marker proteins). The final estimate is $\hat{c}_n = \arg\max_j p(c_n = j)$.

Materials

Simulation study

Our simulation study investigates the benefit of having observed labels available in a modality (such as marker proteins in a hyperLOPIT experiment), and the advantage of performing an integrative analysis rather than independent analyses of each modality. To achieve this, we construct three scenarios. Each scenario is defined by the density used to generate the clusters (the generative model) and by the ϕ vector which determines the similarity of the clustering structure across modalities. We first generate the labels indicating the cluster an item is generated from in each modality and then sample measurements based on this cluster's parameters. In every simulation, the first dataset is generated from six clusters, the second from seven clusters and the third from eight clusters. We generate two hundred labels and data points in each datasets from this model:

$$p(c_n^{(v)} = k|\gamma, \phi) \propto \gamma_k \prod_{w=1;w\neq v}^{V} (1 + \phi_{v,w}), \tag{31}$$

$$(X_n^{(v)}|c_n^{(v)} = k) \sim f(\theta_k). \tag{32}$$

There are fifteen measurements / feature for each data point (i.e., $N = 200, P = 15$ in each dataset in all simulations). For the semi-supervised models, a random 30% of the labels are given as observed in the first dataset with no labels or classes observed in the second and third datasets in all models. Within a generating seed the labels in each modality are the same across all scenarios and each dataset is informative of the others ($\phi = (12, 8, 4)$). The Adjusted Rand Index (**ARI**) [90] between the true labels of each pair of datasets across simulations is shown in figure 1 A. The ARI is a score comparing two partitions, adjusted for chance. A value of 0 means two partitions are no more similar than a pair of random partitions would be expected to be, 1 means they are identical.

We consider three different generating models which give the scenarios their names. The first two are

$$\text{Scenario 1 (Gaussian): } (X_n^{(v)}|c_n^{(v)} = k) \sim \mathcal{N}(\mu_k, \Sigma_k), \tag{33}$$

$$\text{Scenario 2 (MVT): } (X_n^{(v)}|c_n^{(v)} = k) \sim t_{\eta_k}(\mu_k, \Sigma_k). \tag{34}$$

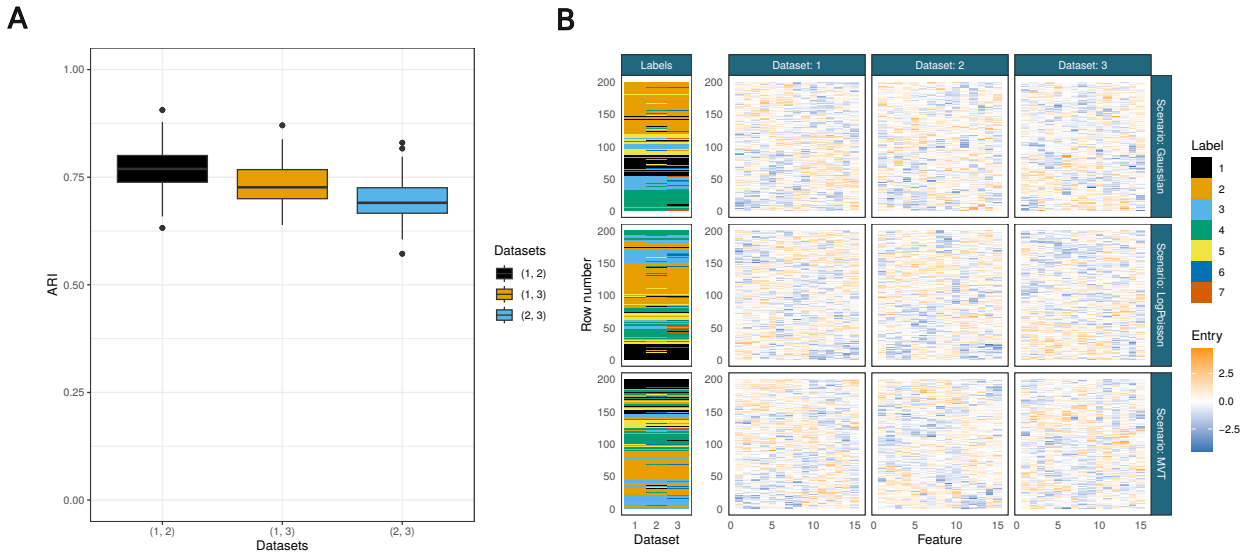

Fig 1. A. ARI between the generating labels for each combination of dataset pairings within a given simulation across all scenarios. For example, a point contributing to the (1, 2) boxplot is the ARI between the generating clusterings for the first and second datasets for a given random seed and scenario. B. Example of simulated data from each scenario and the generating labels all ordered by the data in the first modality for each scenario.

Scenario 3, the Log-Poisson case, is more complex.

$$(Y_{n,p}|c_n^{(v)} = k) \sim Poisson(\lambda_{k,p}), \tag{35}$$

$$\epsilon_{n,p} \sim \mathcal{N}(0, 1), \tag{36}$$

$$X_{n,p} = \log(1 + Y_{n,p}) + \epsilon_{n,p}. \tag{37}$$

This is inspired by the simulation study of Chandra et al. [91], but the Gaussian noise is added after the log-transform to ensure this transform is always possible. Each cluster strongly deviates from Gaussian in this case. An example of the simulated data is shown in figure 1 B. A full description of the generating mechanisms and choice of parameters to differentiate the datasets are given in S.2 of the supplementary material.

We fit five different models in each simulation. These are distinguished with the information available to them in the first dataset and whether they jointly model the datasets (i.e., MDI) or model each dataset separately using a mixture model. In all cases datasets 2 and 3 are unsupervised and the number of clusters present is inferred.

- Unsupervised MDI: MDI with no observed labels and the number of clusters unknown,
- Overfitted semi-supervised MDI: MDI with known labels, but the true number of clusters is hidden,

- Semi-supervised MDI: MDI with both observed labels and the true number of clusters known,
- Overfitted semi-supervised mixture model: mixture model with known labels, but the true number of clusters is hidden,
- Semi-supervised mixture model: mixture model with both observed labels and the true number of clusters known,

We run 10 chains of each method to each simulated dataset. Each chain is 15,000 iterations long with the first 7,500 discarded as burn-in. We thin to every 100th sample to reduce autocorrelation. The best chain for the MDI methods is determined by the chain with the highest mean ARI between the point estimate clustering and the truth across datasets whereas for the mixture model the chains are selected based on the highest ARI between the point estimate and the truth in each dataset (rather than using the same seed across datasets).

Validation study

We compare our approach to a state-of-the-art transfer learning method previously applied to LOPIT data, the KNN-TL classifier, using the marker proteins from four of the same datasets analysed by Breckels et al. [89]. This includes two datasets for *Arabidopsis thaliana*, the *Root Callus* [16] dataset (referred to hereafter as *Callus* to distinguish it from the second *A. thaliana* dataset) which contains 261 marker proteins and 16 fractions, and the *Root* dataset [92] which contains 185 marker proteins and six fractions. a dataset for human embryonic kidney fibroblast cells [93] which contains 404 marker proteins and 8 fractions, and a mouse pluripotent stem cell dataset [12] which contains 8 fractions for 387 marker proteins. These are available through the `pRolocdata` Bioconductor package. Figure 2 shows PCA plots of these proteins. These are paired with GO term data. GO terms record observed and inferred characteristics of gene and gene products in terms of three sub-ontologies. The cellular component (**CC**) the product localises to; its molecular function (**MF**), such as binding or catalysis; and the biological processes (**BP**) the product is involved, e.g., chromatin binding. The data we use are the same as are available through the `pRolocdata` Bioconductor package [94]. This data was prepared by Breckels et al. [89] by using all possible GO CC terms associated to the proteins in the experiment to construct a binary matrix representing the presence/absence of a given term for each protein, for each experiment.

We model the proteomic datasets with a TAGM model and the GO terms using a categorical mixture model (see the supplementary material of [40] for a description of this model) within the MDI framework. We run five chains of the TAGM and MDI models for 15,000 iterations, discard the first 5,000 as warm-up, thin to every 50th and then combine samples across chains (i.e. construct an ensemble). We follow the

vignette of Breckels et al. [93] for the KNN-TL algorithm. This involves identifying the inputs k_1, k_2 and θ in a grid search using cross-validation. We consider a range of $k_1, k_2 = (3, 5, 7, 9, 11, 13, 15, 17, 19)$ and $\theta_k = (0.0, 0.25, 0.50, 0.75, 1.0)$ for $k = 1, \ldots K$. As all the combinations of θ must be considered this becomes computationally intensive as K increases. We considered the smaller of either the full combination matrix or a random sample of 20,000 rows. We inspect performance across ten folds with approximately 30% of the marker proteins observed in each organelle and 70% of the localisations predicted.

We compare performance using the Brier loss, macro F1 score and accuracy (defined below). Accuracy is the number of predictions that match the ground truth divided by the total number of predictions; this is the simplest score we use as it does not account for either uncertainty or class size. The F1 score is a common measure of performance in binary classification examples that corrects for sensitivity and specificity. The variation considered here is an extension to the multi-class setting. Letting $\hat{c}_n$ denote the predicted class for the nth item and c_n denote the true class, then

$$\text{Accuracy} = \frac{1}{N}\sum_{n=1}^{N} \mathbb{I}(\hat{c}_n = c_n), \tag{38}$$

$$\text{true positives (k)} = \sum_{n=1}^{N} \mathbb{I}(\hat{c}_n = k)\mathbb{I}(c_n = k), \tag{39}$$

$$\text{false positives (k)} = \sum_{n=1}^{N} \mathbb{I}(\hat{c}_n = k)\mathbb{I}(c_n \neq k), \tag{40}$$

$$\text{false negatives (k)} = \sum_{n=1}^{N} \mathbb{I}(\hat{c}_n \neq k)\mathbb{I}(c_n = k), \tag{41}$$

$$\text{F1 (k)} = \frac{2 \times \text{ true positives}}{2 \times \text{ true positives} + \text{false negatives} + \text{false positives}}, \tag{42}$$

$$\text{Macro F1} = \frac{1}{K}\sum_{k=1}^{K} \text{F1 (k)}, \tag{43}$$

$$\text{Brier loss} = \frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K} (\hat{p}(c_n = k) - z_{n,k})^2, \tag{44}$$

where $z_{n,k}$ is the $(n, k)^{th}$ entry in the one-hot-encoding of the true allocation matrix, and $\hat{p}(c_n = k)$ is the inferred probability of the nth item being allocated to the kth class. The KNN-TL classifier does not provide a measure of uncertainty and thus we use $\hat{p}(c_n = k) = 1$ for $\hat{c}_n = k$ and 0 otherwise. The Brier score captures how well the Bayesian methods quantify the uncertainty about the classification.

To explore performance on more modern datasets and show that MDI has application beyond analysing pairs of datasets (in contrast to the KNN-TL classifier) we also investigate performance in predicting allocation

of the marker proteins in data from Beltran et al. [95] and Orre et al. [14]. The first is spatial proteomics data for human cells infected by Human Cytomegalovirus (**HCMV**) across 5 time points (24, 48, 72, 96 and 120 hours post infection) and contains 295 marker proteins with measurements in every time point. The second contains measurements for 3,330 marker proteins across five different human cancer cell lines (epidermoid carcinoma A431, glioblastoma U251, breast cancer MCF7, lung cancer NCI-H322, and lung cancer HCC-827). We use a TAGM model for each dataset within the MDI model and as in the previous datasets we consider performance over ten different splits of 30% localisations observed and the remaining 70% inferred.

Toxoplasma gondii

The Apicomplexa phylum of highly successful obligate parasites that can invade practically all warm-blooded animals [96, 97, 98, 99, 100]. As such, the apicomplexans are the source of much scientific interest, particularly understanding the mechanisms by which they infect a host and evade the innate and adaptive immune responses. However, this phylum is deeply divergent from the canonical models of molecular biology [31]. This means that results and knowledge do not always translate well from these well-studied organisms to the Apicomplexa. *T. gondii* is a natural candidate for the model apicomplexan as it can be cultured continuously in a lab and is readily amenable to genomic engineering for targeted experiments [101]. Beyond implications for other apicomplexans such as *Plasmodium falciparum*, the cause of malaria, *T. gondii* is also of medical interest in and of itself being the cause of Toxoplasmosis which, in immunocompromised individuals and as a congenital disease in infected infants, can be life-threatening [102]. From a proteomic perspective, the apicomplexans are a fascinating group; they have a myriad of unique cell structures and compartments that play a key role in their success. For example, they contain an apical structure central to penetration and invasion of animal cells; this contains secretory compartments including the micronemes, rhoptries, and dense granules for the staged release of molecules required for successful invasion and establishment within the host cell[103, 104, 105, 106]. Some of the de novo subcellular niches are formed during the invasion of the host cell by the parasite [107] and it is hoped that better understanding of these organelles and their formation during invasion might lead to novel, targeted treatments [108].

Motivated by the importance of these organisms and their complex cellular structure, we perform an multi-omic analysis of the model apicomplexan, *T. gondii*, combining hyperLOPIT data from Barylyuk et al. [31] and transcriptomic data for the cell-cycle over twelve hours of post-invasion parasite replication [109] (figure 2). We expect that this cell cycle data can reveal the transcriptional programmes of the genes involved in organelle formation and the coexpression patterns of the secreted proteins. The hyperLOPIT data consists

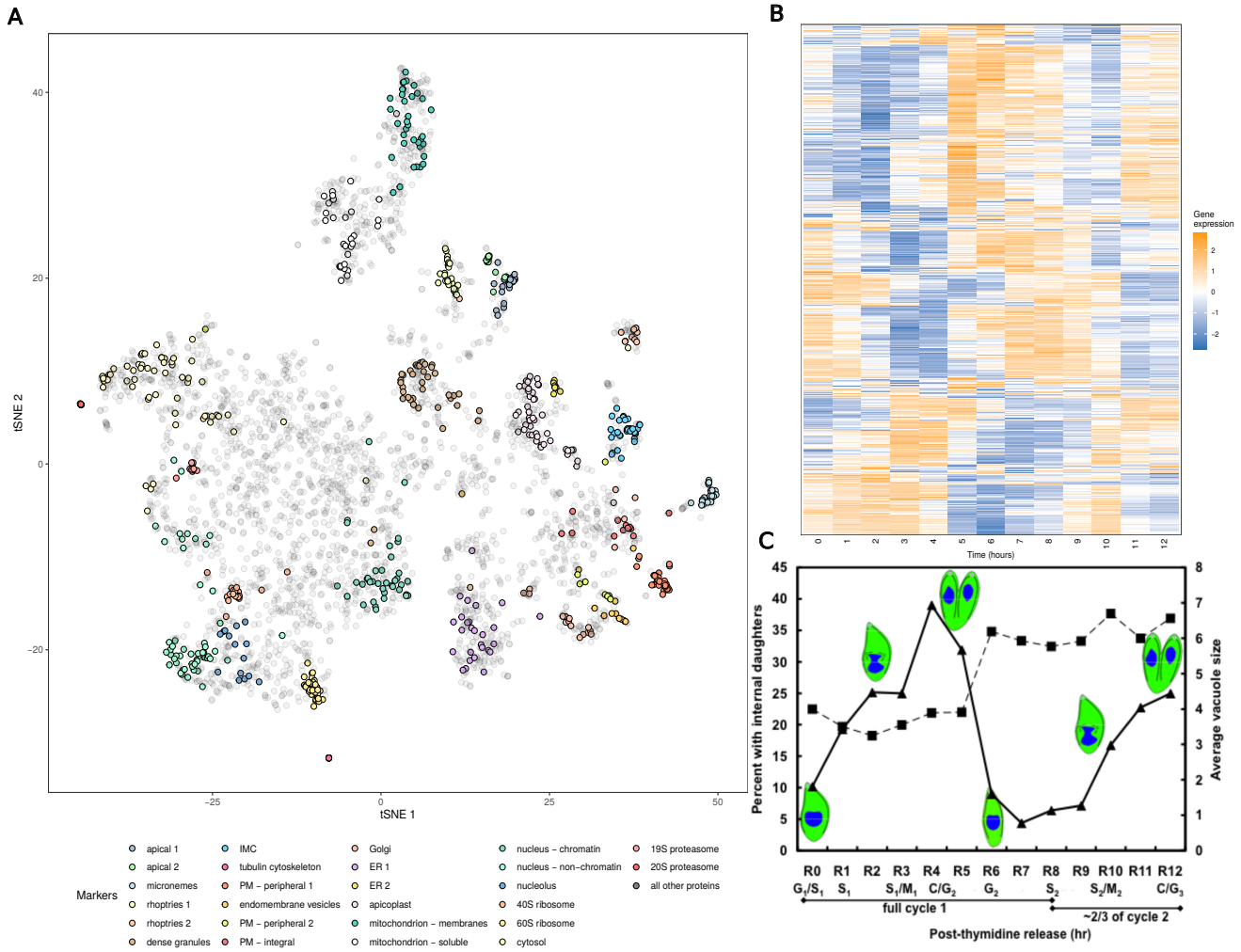

Fig 2. The *T. gondii* data. A) A tSNE of the hyperLOPIT data from Baryluk et al. with the marker proteins of each organelle highlighted. B) The gene expression data from Behnke et al. C) Figure 1 from Behnke et al. showing progression of cell number per vacuole (vacuole size, squares) and portion of cells undergoing division (triangles) at different timepoints in the gene expression data.

of 3 experimental iterations, each containing 10 fractions. By jointly modelling gene expression data with hyperLOPIT data annotated using marker proteins we hope to draw on the temporal programmes of the cell that are specific to spatial niches, and in doing so further resolve protein functional cohorts. However, jointly modelling the two datasets might also shed light on the localisation of proteins. As it stands, proteomes of most parasite compartments remain poorly characterised. Even the locations of proteins of predicted function based on conserved sequences in other organisms are largely untested in apicomplexans, though recent work from Barylyuk et al. [31] has improved our knowledge enormously. To generate both datasets the parasites were grown in and purified from human foreskin fibroblasts. For the gene-expression dataset parasites were synchronised using thymidine following the protocol developed by Radke and White [110]. We consider only the genes which have products in both datasets, reducing to 3,643 genes modelled.

Within the MDI model we choose data-specific probability densities to capture the structure in each dataset. We model the hyperLOPIT data using a TAGM model, setting the number of components to 26 which is the number of subcellular niches with associated marker proteins. In the cell cycle data, we are interested in co-expression patterns. We therefore mean-centre and standardise the measurements for each gene rather than considering absolute transcript abundance. To capture the temporal nature of the data, this modality is modelled using a mixture of MVNs with a GP prior on the mean function, as described in section Gaussian process mixture models. The proposal window for the Metropolis-Hastings sampling of the Gaussian process hyperparameters was set by achieving an acceptance rate between 0.2 and 0.8 in a short chain. For the gene expression data we use 125 components to infer the number of clusters present (approximating a Dirichlet process mixture model [61] in this modality). We also model this data with an unsupervised mixture model with the same choices and parameters as the corresponding submodel in MDI, but the number of components required to be overfitted is much larger and we use 300 components in the mixture model. Due to the complexity of the data and a finite computational budget we used the consensus clustering approach of Coleman et al. [83] to circumvent the issue of poor mixing in individual chains and used the 15,000th sample from 150 randomly initialised chains (see supplement section S.3 for details of assessing stability). We inferred a point estimate partition for the hyperLOPIT data by using the localisation with the highest average allocation probability across MCMC samples, in the transcriptomic data we used the estimate which minimised the lower bound to the posterior expected Variation of Information with a credible ball around the estimate as per the method of Wade and Ghahramani [84].

Results

Development of semi-supervised integrative models for spatial proteomics

To integrate a variety of datasets with spatial proteomics data we developed a semi-supervised Bayesian integrative model. This model allows the integration of a spatial proteomics specific Bayesian model [23, 66] with Bayesian mixture models of other diverse data types. For example, categorical data can be integrated using a Categorical model and time-course data integrated using a Gaussian process based model. Our approach is flexible in that it allows labels/markers to be observed in none, all or one of the data-sets being integrated. Our model is built on the multiple dataset integration (MDI) framework [40], which means that information is shared using a parameter (which is learnt from the data) which up-weights the prior likelihood of two observations clustering together in each dataset. If the latent structures of two datasets are inferred to be unrelated, the model fits separate models in each. Our approach is built in the Bayesian framework and

so quantifies uncertainty in parameters of interest. To enable others to apply our model to other datasets, it is implemented and disseminated as the R-package `mdir` (available from `https://github.com/stcolema/MDIr`).

Simulation study

The predictive performance of the methods is shown in figure 3. This figure demonstrates that, in the examples considered, knowing the true number of generating clusters in the semi-supervised methods is not better than estimating the number of clusters. However, we note that in these examples the observed labels are sampled completely at random from the class, and that all classes are observed, enabling the overfitted mixture models to estimate K accurately.

However, if the sampling of the observed labels was biased in some way or the choice of density used to model the classes was more severely misspecified then the overfitted model might need to use additional components to have a better fit. Investigating this further might be interesting, but it is reassuring that when the model is misspecified here, as in the MVT and Log-Poisson scenario, that it is providing a point estimate partition similar to the ground truth. In fact, it appears to have a mild advantage over the model with K known. We suspect that this is due to additional components capturing observations not well described by the Gaussian density describing their class of origin rather then attempting to fit all of the points into a smaller set of Gaussian densities. We can also see that semi-supervised MDI is the best performer across all datasets and scenarios, though it is matched by unsupervised MDI in the Gaussian and MVT scenarios, the scenarios where the distributional assumptions of the model are either correct or close to correct. More generally, MDI outperforms or matches the mixture models. Unsupervised MDI frequently matches the semi-supervised mixture models in the first modality, with similar median performance across scenarios. Unsurprisingly, semi-supervised MDI, which uses the the most information available to us by considering shared structure across datasets and the observed labels, performs the best in all datasets in the Log-Poisson scenario where the density choice is most strongly misspecified. The relative weight of the density likelihood in defining allocations is smaller in MDI than in the mixture model as the component weights are upweighted by the ϕ parameter. Furthermore, in the first modality some labels are identified and thus the model has less reliance on the distributional assumptions than an unsupervised model or independent mixture models. Thus we would expect semi-supervised MDI to be more robust to model misspecification than the other methods investigated here, and it is reassuring to see this borne out.

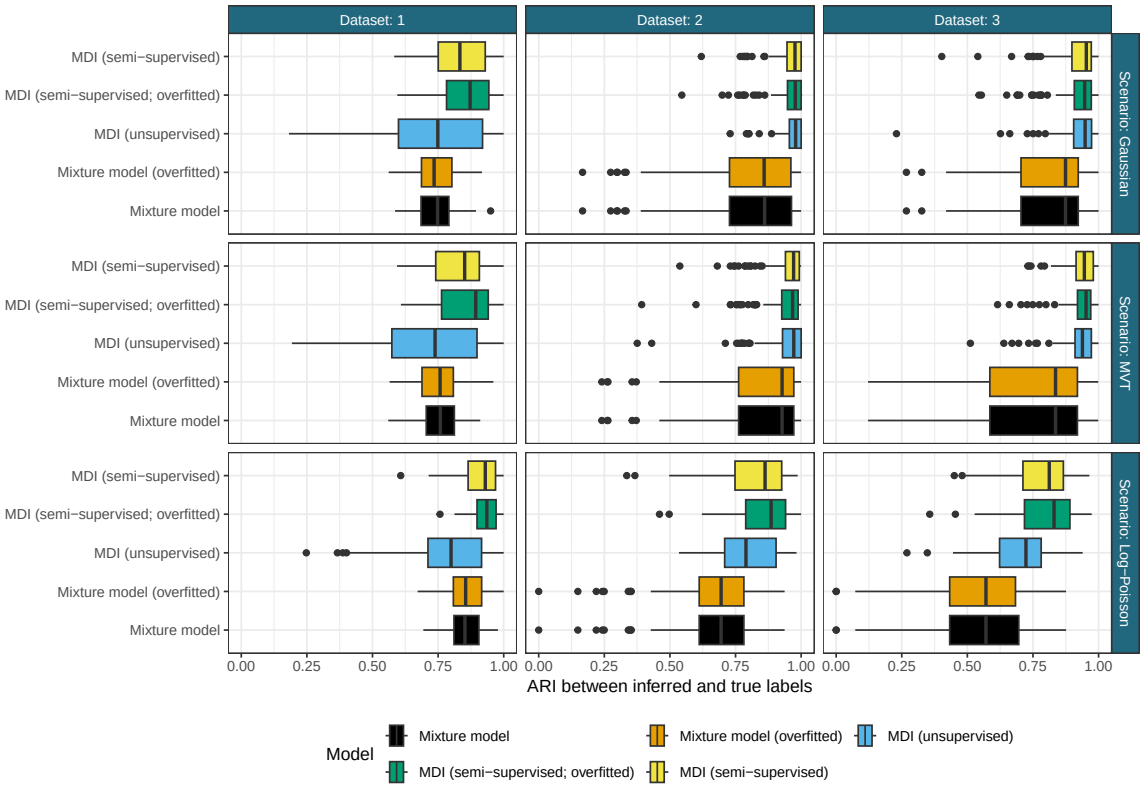

Fig 3. Predictive performance of methods in the simulation study. Horizontal facets are the different datasets, vertical facets are the different generative scenarios. The y-axis shows the different methods compared and the x-axis is the Adjusted Rand Index (ARI) between the inferred labels and the ground truth. In the first modality the ARI is calculated for the same set of unknown labels even in the case of unsupervised MDI where the known labels are not used in the inference to make the comparison more appropriate. Each boxplot is composed of 100 points, the ARI between the clustering point estimate for a given model in a given dataset; e.g., the yellow box plot in the first facet is composed of the ARI between the true and inferred partition for the test (or unobserved) labels in the first dataset from the 100 simulations for the Gaussian scenario.

Validation study

Having demonstrated that our semi-supervised model improves over other integrative mixture models, we wish to evaluate the performance of our method in spatial proteomics applications. We compare our integrative approach with TAGM - a non-integrative spatial proteomics model that can only see the spatial proteomics data. We also compare to the TL KNN method previously developed for integrative spatial proteomics. In this situation, we integrate the GO annotations, which are matrices of presence/absence of a particular annotation, with the MS-based spatial proteomics data. This constitutes a supervised task in both datasets.

In Fig 4 it can be seen that the integrative methods always match or outperform the TAGM model for all scores in all datasets for every test/train split, which is in keeping with the results from the simulation study. This is as expected for the informative second modality available to MDI and the KNN-TL algorithm in this study. MDI matches or outperforms the KNN-TL algorithm under all scores for the Human, Mouse and Callus datasets for all fractions of test/train splits. In the Root dataset the KNN-TL appears to perform slightly better than MDI under accuracy and the F1 scores, but they are more similar under the Brier loss, suggesting that the proteins that MDI mis-classifies have a high uncertainty for their localisation in this dataset. The uncertainty quantification and sampled distributions mean MDI and TAGM give much richer outputs than the KNN-TL algorithm. This demonstrates that our method is a significant advancement over previously developed approaches.

To further assess the advantages an integrative approach, we examined two MS-based spatial proteomics studies with multiple experiments. The first was the spatio-temporal study of [95] in which human fibroblasts where infected with human cytomegalovirus (HCMV). Five spatial proteomics datasets were generated at timepoints of hours post infection (24,48,72,96,120). The focus of their study was to examine the differences between infection and control. Here, we integrate all five datasets in the infected experiment to demonstrate that predictive performance improves compared to considering each dataset in isolation compared with TAGM (see Fig 5). This demonstrates that our approach can also scale to many datasets, however the transfer KNN algorithm could not scale because of its combinatorial approach. As an additional example, [14] generated subcellbarcode - spatial proteomics datasets over five different human cancer cell lines. There approach used extensive offline chromatography and so the number of measured protein approaches close to $10,000$ per cell line across 15 subcellular niches. We also integrated each of these cell lines and saw predictive performance increase in all of the datasets. This demonstrates that these atlases could be improved by borrowing information across related experiments to obtain more accurate predictions whilst allowing for

difference between the datasets. In the next section, we demonstrate the increased flexibility of our approach by integrating a dataset without any annotations at all.

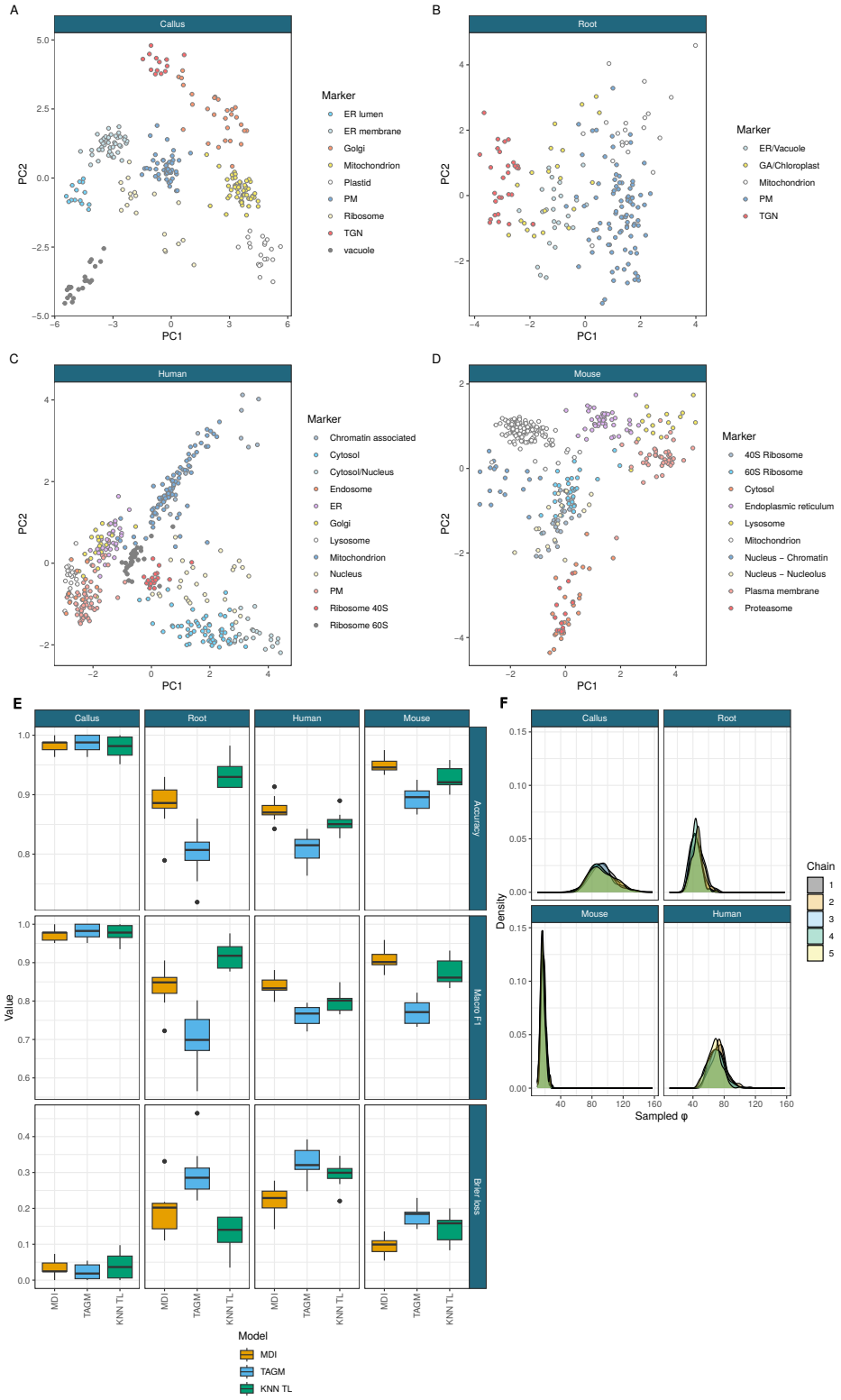

Fig 4. A - D) PCA plots of the marker proteins from the different LOPIT datasets. E) Performance of the MDI, TAGM and transfer learning models for the held out data across ten folds under accuracy, macro F1 and the Brier loss. F) The inferred integration parameter ϕ suggesting a high degree of information sharing.

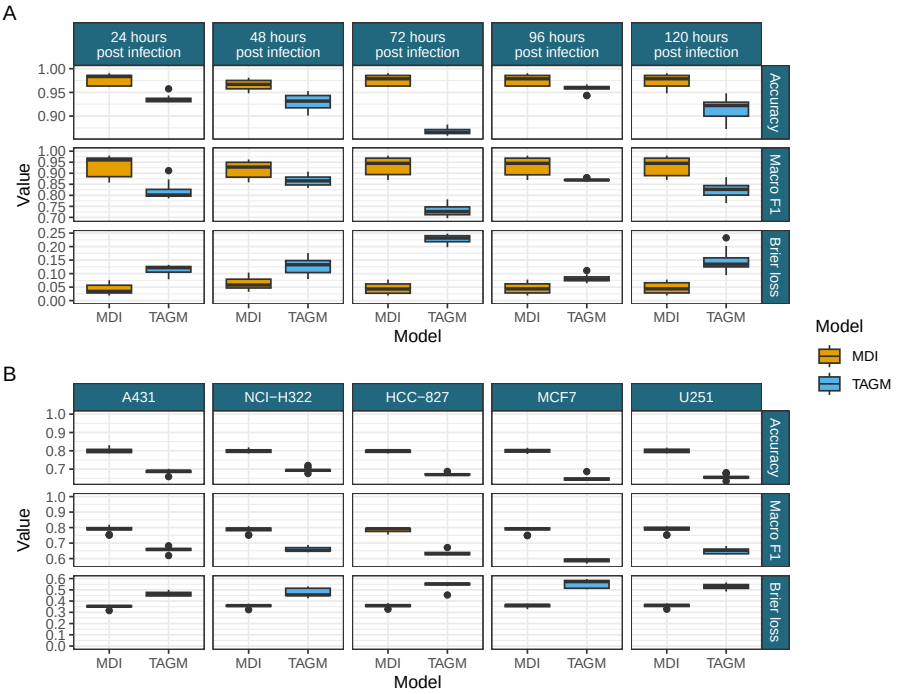

Fig 5. A) Comparison of the performance of the independent TAGM models and the MDI model in predicting a random test set of 70% of the marker proteins in each time point from Beltran et al. [95]. B) Comparison of the performance of the independent TAGM models and the MDI model in predicting a test set of 70% of the marker proteins in each cell line from Orre et al. [14].

Integrating spatial proteomics and gene expression time-course data from *Toxoplasma gondii*

To demonstrate the versatility of our approach, we applied our semi-supervised integrative framework to *Toxoplasma gondii* spatial proteomics data and cell-cycle transcriptomics data. Briefly, the spatial proteomics is modelled using the TAGM model whilst the transcriptomic data us modelled using a mixture of Gaussian proccesses, these two models are then joined using the MDI framework. For additional modelling details, we refer to the method section.

The localisation predictions from this joint analysis, which included both spatial proteomics and time-course gene expression, are plotted in tSNE coordinates of the hyperLOPIT data visualised in figure 6A. Broadly, the allocations continue to cluster in accordance with the tSNE coordinates demonstrating the strong localisation signal from the spatial proteomics data.

Subsequently, we focus on the behaviour of subset of proteins from secretory organelles and ER. For these sub-cellular niches the proteins that are allocated to them from our joint analysis are shown in Figure 6B. We would expect that the majority of the corresponding genes are “invasion-ready”, highly expressed from

beginning (invasion G/S-phase) of the cell-cycle and the invasion of the host's cell. However, the gene expression may not be high for some protein of the secretory organelles because they are already highly abundant. Thus, more specifically, gene-expression is high when the proteins need to be remade due to them being secreted or cellular division dilute their concentration and so more proteins are needed. This is the case for the proteins allocated to the rhoptries (a specialised organelle whose proteins are secreted during penetration)(Fig 6 C). The proteins allocated to the micronemes and the apical complex compartments lag slightly behind that of the rhoptries. The lag in expression of micronemes tells us that rhoptries are made first, then new micronemes, and this displacement in time is interpreted to be because they use similar sorting routes through the Golgi and therefore the cell temporarily separates their synthesis so that they do not get mislocalised.

To assess the benefit of including the temporal data, we plot the allocation probabilities from the TAGM model applied to the spatial proteomics alone against the allocation probabilities from the integrative analysis for secretory organelles (Fig 6 D). We observe a general trend where by the integrative analysis was more confident than the TAGM analysis with 73% of proteins having higher probabilities with the integrative approach. Only four proteins had different allocations between the two approaches but three of those where simply reallocated to the other apical cluster. TGME49_285150, a protein of unknown function, was reallocated to the apical rather than the dense granules albeit with low probability 0.64. A few proteins had much lower allocation probability with the integrative analysis including two proteins with a difference of more than 0.3. One of these was a protein of unknown function whilst the other was ERK7, which typically has an apical localisation but here allocated to the dense granules with low probability. ERK7 localisation appears to be AC9 dependent [111] and so the low probability here may be indicative of a diffuse subcellular localisation. These observations suggest that integrative analysis helps further calibrate our localisation analysis and inclusion of temporal datasets in future studies will produce a more refined and confident set of allocation.

We then explore the behaviour of the proteins predicted to localise to the dense granules, an important organelle in the apicomplexan post-invasion strategy for re-engineering the host cell environment to enable the parasites' safe growth and replication.(Fig 7). These secretory vesicles are key in the formation of the parasitophorous vacuole in which the parasite resides. It was recently discovered that the related parasite *Cryptosporidium* boasts a small granule as a subset of the dense granules [112] and the question remains of whether in *Toxoplasma* dense granules also form a heterogeneous population of secretory compartments [113]. We consider the jointly inferred clusters of the transcripts corresponding to these proteins in the transcriptomic data, and find that the majority of these proteins are highly expressed towards the G/S transition

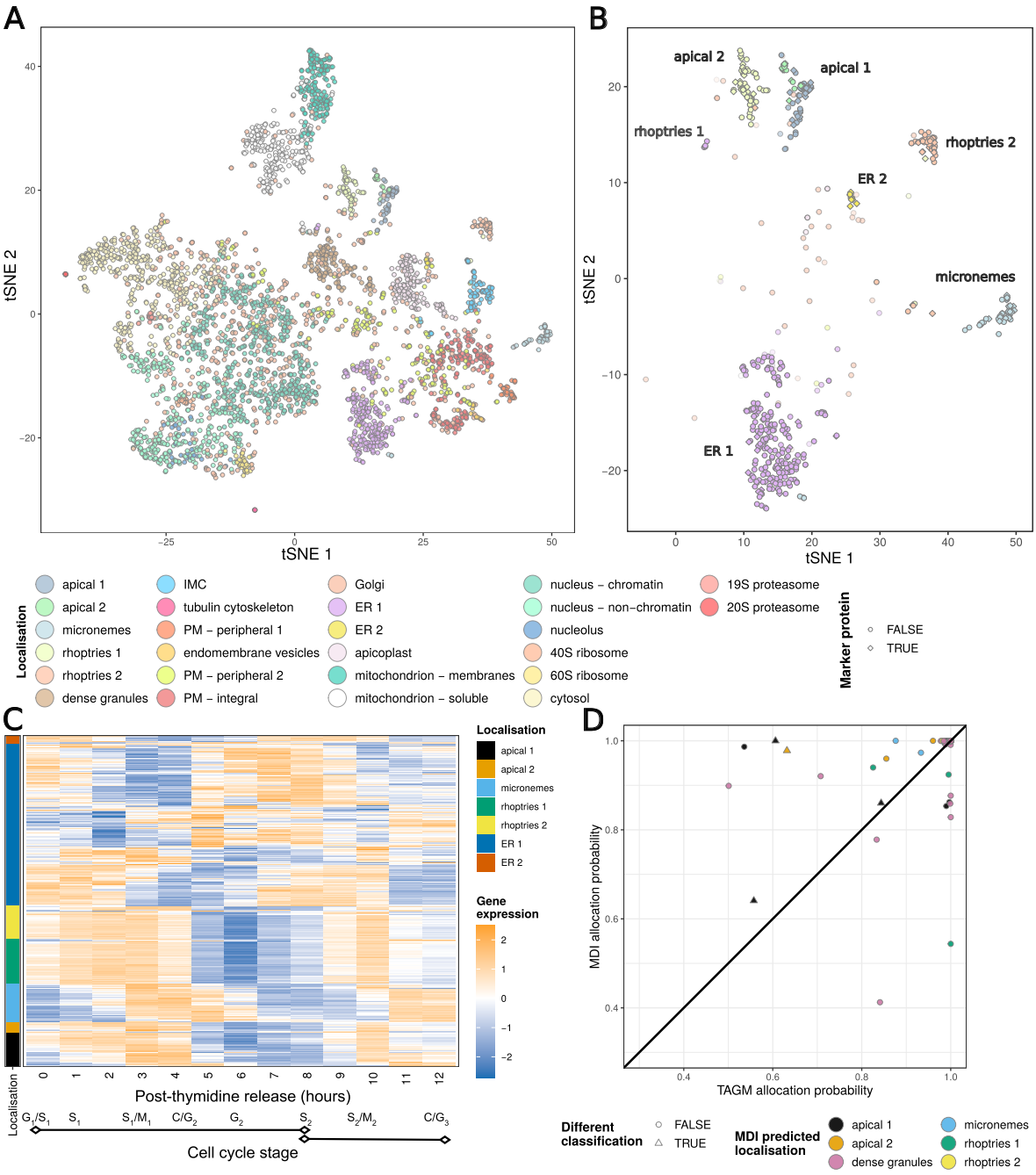

Fig 6. A) tSNE projection of the hyperLOPIT data from Baryluk et al coloured by the predicted localisations using both the hyperLOPIT and gene-expression data. B) The location in A) of the proteins predicted to localise to a subset of localisations (apical 1, apical 2, the micronemes, rhoptries 1, rhoptries 2, ER1 and ER2), opacity set by inferred probability of allocation. C) The gene co-expression data for these proteins annotated by the predicted localisation. D) Comparison of the predicted localisations and the associated probabilities from the MDI and TAGM models for these proteins.

stage of the cell cycle. However, there is heterogeneity in these dense granule gene expression programmes, for example a subset is more highly expressed closer to the S/M transition (Fig 7 B, cluster 8) — more in line with the timing of biogenesis of small granules within *Cryptosporidium*. Moreover, considering the

distributions of the protein-average ratio of nonsynonymous and synonymous mutation rates in the four largest clusters shows that cluster 8 has a significantly higher mean d_N/d_S ratio than the other clusters (Fig 7 C). This indicates that these proteins are under stronger selection pressures compared with the other dense granule proteins and, therefore, that there might be a distinctive role for this cohort of proteins. It is currently unknown if these differences might manifest only in their expression timing and levels, or if they might occupy distinctive spatial niches similar to the *Cryptosporidium* small granules from dense granules. But either way, our integrative analysis has uncovered a further layer of regulation, the first at the level of subcellular localisation and the second at the level of transcriptional regulation. These findings suggest that semi-supervised Bayesian integration can uncover previously unknown functional relationships in cell organisation.

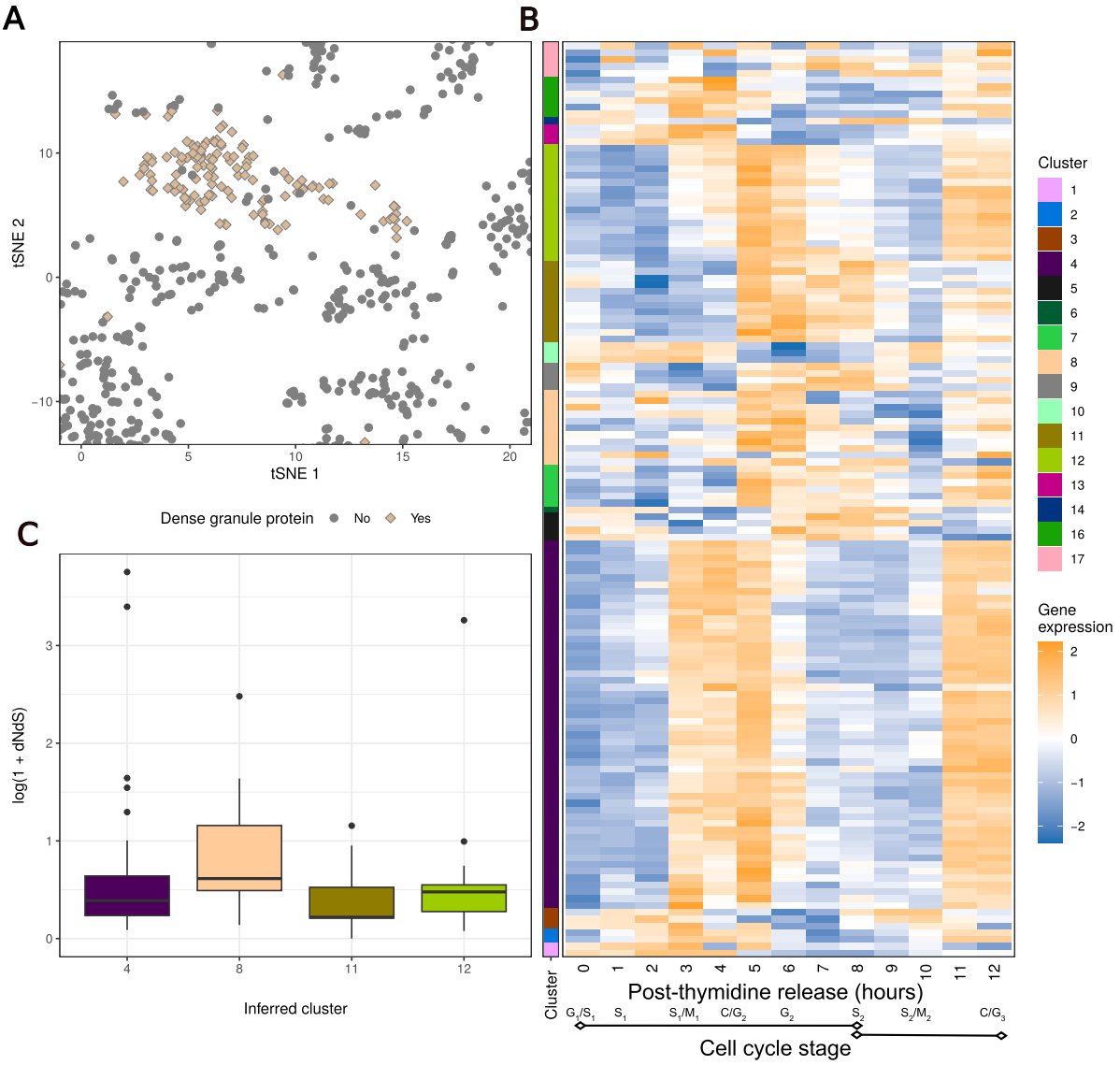

Fig 7. A) The subsection of the tSNE plot for the hyperLOPIT data focusing on the proteins predicted to localise to the dense granules using both dataset. B) The gene expression data for the genes whose members are predicted to localise to the Dense Granules, annotated by the point estimate clustering. The predicted localisation of the proteins in the hyperLOPIT data and the point estimate clustering in the gene expression data are jointly inferred with the same model run of the MDI model on these data. C) Comparison of the log ratio of dN to dS for the proteins from the 4 largest clusters from figure B).

Discussion

We have introduced a semi-supervised, multi-modal Bayesian integrative method for spatial proteomics applications. We have shown on real and simulated datasets that it matches or outperforms state-of-the-art methods and offers a rich and informative output, particularly a principled quantification of uncertainty about the predictions. We then applied our method to a pair of 'omics datasets for *T. gondii*. Our analysis goes beyond that of Barylyuk et al. [31] where expression data is considered independently. We validated our analysis by considering the expression data for the secretory organelles, demonstrating that they correlated with their functions. We also found evidence suggesting that the dense granules may form a non-homogeneous functional population that have varying expression patterns during the cell cycle that are identified by transcriptional timing rather than on subcellular localisation alone.

Our approach is broadly applicable and various likelihood functions are implemented to allow integration of diverse data types, including discrete, continuous and time-course data with spatial proteomics data. We have shown that semi-supervised, integrative analysis is a useful tool for spatial proteomics and provided an R package as an open-source dissemination of our method. We envisage a broad range of applications of our method.

Our method has some limitations. Firstly, Bayesian methods are computationally intensive, though users are provided with uncertainty quantification in return for this cost. Secondly, whilst we have shown some robustness to model misspecification, our analysis is likely to be affected by gross misspecification of the likelihood. In the integrative case, the misspecification of a model for a single dataset may have a negative impact on the modelling of all other datasets, as the misspecification "leaks" from one dataset into another where there is no misspecification, corrupting the analysis. In this case, one might apply "cut-models" or adapt our approach using an appropriate likelihood.

Acknowledgments

This research was funded by the MRC (MC_UU_000024), Wellcome Trust (WT220788). This research was funded in whole, or in part, by the Wellcome Trust [WT220788]. RFW was supported by the Wellcome Trust, United Kingdom, Investigator award 214298/Z/18/Z. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

Conflicts of Interest

CW receives funding from MSD and GSK and is a part-time employee of GSK. These companies had no involvement in the work presented here.

References

1. Gibson TJ. Cell regulation: determined to signal discrete cooperation. Trends in Biochemical Sciences. 2009;34(10):471–482. doi:https://doi.org/10.1016/j.tibs.2009.06.007.
2. Laurila K, Vihinen M. Prediction of disease-related mutations affecting protein localization. BMC Genomics. 2009;10(1):122. doi:10.1186/1471-2164-10-122.
3. Olkkonen VM, Ikonen E. When intracellular logistics fails - genetic defects in membrane trafficking. Journal of Cell Science. 2006;119(24):5031–5045. doi:10.1242/jcs.03303.
4. Luheshi LM, Crowther DC, Dobson CM. Protein misfolding and disease: from the test tube to the organism. Current Opinion in Chemical Biology. 2008;12(1):25–31. doi:https://doi.org/10.1016/j.cbpa.2008.02.011.
5. De Matteis MA, Luini A. Mendelian Disorders of Membrane Trafficking. New England Journal of Medicine. 2011;365(10):927–938. doi:10.1056/NEJMra0910494.
6. Shin SJ, Smith JA, Rezniczek GA, Pan S, Chen R, Brentnall TA, et al. Unexpected gain of function for the scaffolding protein plectin due to mislocalization in pancreatic cancer. Proceedings of the National Academy of Sciences. 2013;110(48):19414–19419. doi:10.1073/pnas.1309720110.
7. Siljee JE, Wang Y, Bernard AA, Ersoy BA, Zhang S, Marley A, et al. Subcellular localization of MC4R with ADCY3 at neuronal primary cilia underlies a common pathway for genetic predisposition to obesity. Nature Genetics. 2018;50(2):180–185. doi:10.1038/s41588-017-0020-9.
8. Cook KC, Cristea IM. Location is everything: protein translocations as a viral infection strategy. Current Opinion in Chemical Biology. 2019;48:34–43. doi:https://doi.org/10.1016/j.cbpa.2018.09.021.
9. Itzhak DN, Tyanova S, Cox J, Borner GH. Global, quantitative and dynamic mapping of protein subcellular localization. eLife. 2016;5:e16950. doi:10.7554/eLife.16950.
10. Geladaki A, Kočevar Britovšek N, Breckels LM, Smith TS, Vennard OL, Mulvey CM, et al. Combining LOPIT with differential ultracentrifugation for high-resolution spatial proteomics. Nature Communications. 2019;10(1):331. doi:10.1038/s41467-018-08191-w.

11. Dunkley TP, Watson R, Griffin JL, Dupree P, Lilley KS. Localization of organelle proteins by isotope tagging (LOPIT). Molecular & Cellular Proteomics. 2004;3(11):1128–1134.

12. Christoforou A, Mulvey CM, Breckels LM, Geladaki A, Hurrell T, Hayward PC, et al. A draft map of the mouse pluripotent stem cell spatial proteome. Nature Communications. 2016;7(1):9992. doi:10.1038/ncomms9992.

13. Foster LJ, de Hoog CL, Zhang Y, Zhang Y, Xie X, Mootha VK, et al. A mammalian organelle map by protein correlation profiling. Cell. 2006;125(1):187–199.

14. Orre LM, Vesterlund M, Pan Y, Arslan T, Zhu Y, Fernandez Woodbridge A, et al. SubCellBarCode: Proteome-wide Mapping of Protein Localization and Relocalization. Molecular Cell. 2019;73(1):166–182.e7. doi:10.1016/j.molcel.2018.11.035.

15. Shin JJ, Crook OM, Borgeaud AC, Cattin-Ortolá J, Peak-Chew SY, Breckels LM, et al. Spatial proteomics defines the content of trafficking vesicles captured by golgin tethers. Nature communications. 2020;11(1):5987.

16. Dunkley TPJ, Hester S, Shadforth IP, Runions J, Weimar T, Hanton SL, et al. Mapping the *Arabidopsis* organelle proteome. Proceedings of the National Academy of Sciences. 2006;103(17):6518–6523. doi:10.1073/pnas.0506958103.

17. Mulvey CM, Breckels LM, Geladaki A, Britovšek NK, Nightingale DJ, Christoforou A, et al. Using hyperLOPIT to perform high-resolution mapping of the spatial proteome. Nature protocols. 2017;12(6):1110–1135.

18. Thompson A, Schäfer J, Kuhn K, Kienle S, Schwarz J, Schmidt G, et al. Tandem Mass Tags: A Novel Quantification Strategy for Comparative Analysis of Complex Protein Mixtures by MS/MS. Analytical Chemistry. 2003;75(8):1895–1904. doi:10.1021/ac0262560.

19. De Duve C. The peroxisome: a new cytoplasmic organelle. Proceedings of the Royal Society of London Series B Biological Sciences. 1969;173(1030):71–83. doi:10.1098/rspb.1969.0039.

20. De Duve C, Beaufay H. A short history of tissue fractionation. J Cell Biol. 1981;91(3 Pt 2):293s–299s.

21. Tan DJL, Dvinge H, Christoforou A, Bertone P, Martinez Arias A, Lilley KS. Mapping Organelle Proteins and Protein Complexes in Drosophila melanogaster. Journal of Proteome Research. 2009;8(6):2667–2678. doi:10.1021/pr800866n.

22. Jiang Y, Rex DAB, Schuster D, Neely BA, Rosano GL, Volkmar N, et al.. Comprehensive Overview of Bottom-up Proteomics using Mass Spectrometry; 2023.

23. Crook OM, Mulvey CM, Kirk PDW, Lilley KS, Gatto L. A Bayesian Mixture Modelling Approach for Spatial Proteomics. PLOS Computational Biology. 2018;14(11):e1006516. doi:10.1371/journal.pcbi.1006516.

24. Crook OM, Lilley KS, Gatto L, Kirk PDW. Semi-supervised nonparametric Bayesian modelling of spatial proteomics. The Annals of Applied Statistics. 2022;16(4):2554 – 2576. doi:10.1214/22-AOAS1603.

25. Qattan AT, Mulvey C, Crawford M, Natale DA, Godovac-Zimmermann J. Quantitative Organelle Proteomics of MCF-7 Breast Cancer Cells Reveals Multiple Subcellular Locations for Proteins in Cellular Functional Processes. Journal of Proteome Research. 2010;9(1):495–508. doi:10.1021/pr9008332.

26. Allen M, Zou F, Chai HS, Younkin CS, Crook J, Pankratz VS, et al. Novel late-onset Alzheimer disease loci variants associate with brain gene expression. Neurology. 2012;79(3):221–228. doi:10.1212/WNL.0b013e3182605801.

27. Bali J, Gheinani AH, Zurbriggen S, Rajendran L. Role of genes linked to sporadic Alzheimer's disease risk in the production of β-amyloid peptides. Proceedings of the National Academy of Sciences. 2012;109(38):15307–15311. doi:10.1073/pnas.1201632109.

28. Krahmer N, Najafi B, Schueder F, Quagliarini F, Steger M, Seitz S, et al. Organellar Proteomics and Phospho-Proteomics Reveal Subcellular Reorganization in Diet-Induced Hepatic Steatosis. Developmental Cell. 2018;47(2):205–221.e7. doi:https://doi.org/10.1016/j.devcel.2018.09.017.

29. Allou L, Balzano S, Magg A, Quinodoz M, Royer-Bertrand B, Schöpflin R, et al. Non-coding deletions identify Maenli lncRNA as a limb-specific En1 regulator. Nature. 2021;592(7852):93–98. doi:10.1038/s41586-021-03208-9.

30. Akerman I, Maestro MA, De Franco E, Grau V, Flanagan S, García-Hurtado J, et al. Neonatal diabetes mutations disrupt a chromatin pioneering function that activates the human insulin gene. Cell reports. 2021;35(2):108981.

31. Barylyuk K, Koreny L, Ke H, Butterworth S, Crook OM, Lassadi I, et al. A Comprehensive Subcellular Atlas of the Toxoplasma Proteome via hyperLOPIT Provides Spatial Context for Protein Functions. Cell Host & Microbe. 2020;28(5):752–766.e9. doi:https://doi.org/10.1016/j.chom.2020.09.011.

32. Bai JP, Alekseyenko AV, Statnikov A, Wang IM, Wong PH. Strategic applications of gene expression: from drug discovery/development to bedside. The AAPS journal. 2013;15(2):427–437.

33. Emmert-Streib F, Dehmer M, Haibe-Kains B. Gene regulatory networks and their applications: understanding biological and medical problems in terms of networks. Frontiers in cell and developmental biology. 2014;2:38.

34. The Cancer Genome Atlas Network. Comprehensive Molecular Portraits of Human Breast Tumours. Nature. 2012;490(7418):61–70. doi:10.1038/nature11412.

35. Manzoni C, Kia DA, Vandrovcova J, Hardy J, Wood NW, Lewis PA, et al. Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. Briefings in Bioinformatics. 2016;19(2):286–302. doi:10.1093/bib/bbw114.

36. Lock EF, Hoadley KA, Marron JS, Nobel AB. Joint and Individual Variation Explained (JIVE) for Integrated Analysis of Multiple Data Types. The annals of applied statistics. 2013;7(1):523–542. doi:10.1214/12-AOAS597.

37. Argelaguet R, Velten B, Arnol D, Dietrich S, Zenz T, Marioni JC, et al. Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. Molecular systems biology. 2018;14(6):e8124.

38. Lock EF, Dunson DB. Bayesian consensus clustering. Bioinformatics. 2013;29(20):2610–2616. doi:10.1093/bioinformatics/btt425.

39. Gabasova E, Reid J, Wernisch L. Clusternomics: Integrative context-dependent clustering for heterogeneous datasets. PLoS computational biology. 2017;13(10):e1005781.

40. Kirk P, Griffin JE, Savage RS, Ghahramani Z, Wild DL. Bayesian correlated clustering to integrate multiple datasets. Bioinformatics. 2012;28(24):3290–3297.

41. Fraley C, Raftery AE. Model-based clustering, discriminant analysis, and density estimation. Journal of the American statistical Association. 2002;97(458):611–631.

42. Kirk PDW, Stumpf MPH. Gaussian process regression bootstrapping: exploring the effects of uncertainty in time course data. Bioinformatics. 2009;25(10):1300–1306. doi:10.1093/bioinformatics/btp139.

43. Kalaitzis AA, Lawrence ND. A Simple Approach to Ranking Differentially Expressed Gene Expression Time Courses through Gaussian Process Regression. BMC Bioinformatics. 2011;12(1):180. doi:10.1186/1471-2105-12-180.

44. Cooke EJ, Savage RS, Kirk PD, Darkins R, Wild DL. Bayesian hierarchical clustering for microarray time series data with replicates and outlier measurements. BMC Bioinformatics. 2011;12(1):399. doi:10.1186/1471-2105-12-399.

45. Hensman J, Lawrence ND, Rattray M. Hierarchical Bayesian modelling of gene expression time series across irregularly sampled replicates and clusters. BMC Bioinformatics. 2013;14(1):252. doi:10.1186/1471-2105-14-252.

46. Rouanet A, Johnson R, Strauss ME, Richardson S, Tom BD, White SR, et al.. Bayesian profile regression for clustering analysis involving a longitudinal response and explanatory variables; 2021. Available from: `https://arxiv.org/abs/2111.04518`.

47. Lee D, Mukhopadhyay S, Rushworth A, Sahu SK. A rigorous statistical framework for spatio-temporal pollution prediction and estimation of its long-term impact on health. Biostatistics. 2016;18(2):370–385. doi:10.1093/biostatistics/kxw048.

48. Ferrari F, Dunson DB. Identifying main effects and interactions among exposures using Gaussian Processes. Ann Appl Stat. 2020;14(4):1743–1758.

49. Colopy GW, Pimentel MAF, Roberts SJ, Clifton DA. Bayesian Gaussian processes for identifying the deteriorating patient. In: 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); 2016. p. 5311–5314.

50. Cheng LF, Dumitrascu B, Darnell G, Chivers C, Draugelis M, Li K, et al. Sparse multi-output Gaussian processes for online medical time series prediction. BMC Medical Informatics and Decision Making. 2020;20(1):152. doi:10.1186/s12911-020-1069-4.

51. van der Vaart AW, van Zanten JH. Rates of contraction of posterior distributions based on Gaussian process priors. The Annals of Statistics. 2008;36(3):1435 – 1463. doi:10.1214/009053607000000613.

52. van der Vaart A, van Zanten H. Information Rates of Nonparametric Gaussian Process Methods. Journal of Machine Learning Research. 2011;12(6).

53. Teckentrup AL. Convergence of Gaussian Process Regression with Estimated Hyper-Parameters and Applications in Bayesian Inverse Problems. SIAM/ASA Journal on Uncertainty Quantification. 2020;8(4):1310–1337. doi:10.1137/19M1284816.

54. Stephenson WT, Ghosh S, Nguyen TD, Yurochkin M, Deshpande S, Broderick T. Measuring the robustness of Gaussian processes to kernel choice. In: Camps-Valls G, Ruiz FJR, Valera I, editors. Proceedings of The 25th International Conference on Artificial Intelligence and Statistics. vol. 151 of Proceedings of Machine Learning Research. PMLR; 2022. p. 3308–3331. Available from: `https://proceedings.mlr.press/v151/stephenson22a.html`.

55. Williams CK, Rasmussen CE. Gaussian processes for machine learning. vol. 2. MIT press Cambridge, MA; 2006.

56. Quinonero-Candela J, Rasmussen CE. A unifying view of sparse approximate Gaussian process regression. The Journal of Machine Learning Research. 2005;6:1939–1959.

57. Banerjee A, Dunson DB, Tokdar ST. Efficient Gaussian process regression for large datasets. Biometrika. 2012;100(1):75–89. doi:10.1093/biomet/ass068.

58. Liu H, Ong YS, Shen X, Cai J. When Gaussian Process Meets Big Data: A Review of Scalable GPs. IEEE Transactions on Neural Networks and Learning Systems. 2020;31(11):4405–4423. doi:10.1109/TNNLS.2019.2957109.

59. Zhang Y, Leithead WE, Leith DJ. Time-series Gaussian Process Regression Based on Toeplitz Computation of O(N2) Operations and O(N)-level Storage. In: Proceedings of the 44th IEEE Conference on Decision and Control; 2005. p. 3711–3716.

60. Crook OM, Chung Cw, Deane CM. Challenges and Opportunities for Bayesian Statistics in Proteomics. Journal of Proteome Research. 2022;21(4):849–864. doi:10.1021/acs.jproteome.1c00859.

61. Rousseau J, Mengersen K. Asymptotic behaviour of the posterior distribution in overfitted mixture models. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2011;73(5):689–710.

62. Ferguson TS. A Bayesian analysis of some nonparametric problems. The annals of statistics. 1973; p. 209–230.

63. Antoniak CE. Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems. The Annals of Statistics. 1974;2(6):1152–1174. doi:10.1214/aos/1176342871.

64. Ferguson TS. Bayesian Density Estimation by Mixtures of Normal Distributions. In: Rizvi MH, Rustagi JS, Siegmund D, editors. Recent Advances in Statistics. Academic Press; 1983. p. 287–302.

65. Lo AY. On a Class of Bayesian Nonparametric Estimates: I. Density Estimates. The Annals of Statistics. 1984;12(1):351–357.

66. Crook OM, Geladaki A, Nightingale DJH, Vennard OL, Lilley KS, Gatto L, et al. A semi-supervised Bayesian approach for simultaneous protein sub-cellular localisation assignment and novelty detection. PLOS Computational Biology. 2020;16(11):1–21. doi:10.1371/journal.pcbi.1008288.

67. Metropolis N, Ulam S. The Monte Carlo Method. Journal of the American Statistical Association. 1949;44(247):335. doi:10.2307/2280232.

68. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. Equation of State Calculations by Fast Computing Machines. The Journal of Chemical Physics. 1953;21(6):1087–1092. doi:10.1063/1.1699114.

69. Hastings WK. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. Biometrika. 1970; p. 14.

70. Gelfand AE, Smith AFM. Sampling-Based Approaches to Calculating Marginal Densities. Journal of the American Statistical Association. 1990;85(410):398–409. doi:10.1080/01621459.1990.10476213.

71. Robert CP, Casella G. Monte Carlo statistical methods. vol. 2. Springer; 1999.

72. Berger JO. Bayesian Analysis: A Look at Today and Thoughts of Tomorrow. Journal of the American Statistical Association. 2000;95(452):1269–1276. doi:10.1080/01621459.2000.10474328.

73. Robert CP. The Bayesian choice: a decision-theoretic motivation. Springer-Verlag; 1994.

74. McGrayne SB. The Theory That Would Not Die: How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines, & Emerged Triumphant from Two Centuries of C. Yale University Press; 2011.

75. Dahl DB. An improved merge-split sampler for conjugate Dirichlet process mixture models. Technical Report. 2003;1:086.

76. Jain S, Neal RM. A Split-Merge Markov chain Monte Carlo Procedure for the Dirichlet Process Mixture Model. Journal of Computational and Graphical Statistics. 2004;13(1):158–182. doi:10.1198/1061860043001.

77. Bouchard-Côté A, Doucet A, Roth A. Particle Gibbs split-merge sampling for Bayesian inference in mixture models. Journal of Machine Learning Research. 2017;18(28).

78. Dunson DB, Johndrow JE. The Hastings algorithm at fifty. Biometrika. 2019;107(1):1–23. doi:10.1093/biomet/asz066.

79. Lu Y, Lu J, Nolen J. Accelerating Langevin Sampling with Birth-death; 2019. Available from: `https://arxiv.org/abs/1905.09863`.

80. Syed S, Romaniello V, Campbell T, Bouchard-Cote A. Parallel tempering on optimized paths. In: Meila M, Zhang T, editors. Proceedings of the 38th International Conference on Machine Learning. vol. 139 of Proceedings of Machine Learning Research. PMLR; 2021. p. 10033–10042. Available from: `https://proceedings.mlr.press/v139/syed21a.html`.

81. Syed S, Bouchard-Côté A, Deligiannidis G, Doucet A. Non-reversible parallel tempering: A scalable highly parallel MCMC scheme. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2022;84(2):321–350. doi:https://doi.org/10.1111/rssb.12464.

82. Geman S, Geman D. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1984;PAMI-6(6):721–741. doi:10.1109/TPAMI.1984.4767596.

83. Coleman S, Kirk PDW, Wallace C. Consensus clustering for Bayesian mixture models. BMC Bioinformatics. 2022;23(1):290. doi:10.1186/s12859-022-04830-8.

84. Wade S, Ghahramani Z. Bayesian Cluster Analysis: Point Estimation and Credible Balls (with Discussion). Bayesian Analysis. 2018;13(2):559–626. doi:10.1214/17-BA1073.

85. Chaumeny Y, Moris JvdM, Davison AC, Kirk PDW. Bayesian nonparametric mixture inconsistency for the number of components: How worried should we be in practice?; 2022. Available from: `https://arxiv.org/abs/2207.14717`.

86. Fritsch A, Ickstadt K. Improved criteria for clustering based on the posterior similarity matrix. Bayesian analysis. 2009;4(2):367–391.

87. Dahl DB, Johnson DJ, Mueller P. Search Algorithms and Loss Functions for Bayesian Clustering. arXiv:210504451 [stat]. 2021;.

88. Wu P, Dietterich TG. Improving SVM Accuracy by Training on Auxiliary Data Sources. In: Proceedings of the Twenty-First International Conference on Machine Learning. ICML '04. New York, NY, USA: Association for Computing Machinery; 2004. p. 110. Available from: `https://doi-org.ezp.lib.cam.ac.uk/10.1145/1015330.1015436`.

89. Breckels LM, Holden SB, Wojnar D, Mulvey CM, Christoforou A, Groen A, et al. Learning from Heterogeneous Data Sources: An Application in Spatial Proteomics. PLOS Computational Biology. 2016;12(5):1–26. doi:10.1371/journal.pcbi.1004920.

90. Hubert L, Arabie P. Comparing partitions. Journal of Classification. 1985;2(1):193–218. doi:10.1007/BF01908075.

91. Chandra NK, Canale A, Dunson DB. Escaping the curse of dimensionality in Bayesian model based clustering; 2020. Available from: `https://arxiv.org/abs/2006.02700`.

92. Groen AJ, Sancho-Andrés G, Breckels LM, Gatto L, Aniento F, Lilley KS. Identification of Trans-Golgi Network Proteins in Arabidopsis thaliana Root Tissue. Journal of Proteome Research. 2014;13(2):763–776. doi:10.1021/pr4008464.

93. Breckels LM, Gatto L, Christoforou A, Groen AJ, Lilley KS, Trotter MWB. The effect of organelle discovery upon sub-cellular protein localisation. Journal of Proteomics. 2013;88:129–140. doi:https://doi.org/10.1016/j.jprot.2013.02.019.

94. Gatto L, Crook O, Breckels L. pRolocdata: Data accompanying the pRoloc package; 2023. Available from: `https://bioconductor.org/packages/pRolocdata`.

95. Jean Beltran P, Mathias RA, Cristea IM. A Portrait of the Human Organelle Proteome In Space and Time during Cytomegalovirus Infection. Cell Systems. 2016;3(4):361–373.e6. doi:10.1016/j.cels.2016.08.012.

96. Morrissette NS, Sibley LD. Cytoskeleton of Apicomplexan Parasites. Microbiology and Molecular Biology Reviews. 2002;66(1):21–38. doi:10.1128/MMBR.66.1.21-38.2002.

97. Mercier C, Adjogble KDZ, Däubener W, Delauw MFC. Dense granules: Are they key organelles to help understand the parasitophorous vacuole of all apicomplexa parasites? International Journal for Parasitology. 2005;35(8):829–849. doi:https://doi.org/10.1016/j.ijpara.2005.03.011.

98. Rashid M, Rashid MI, Akbar H, Ahmad L, Hassan MA, Ashraf K, et al. A systematic review on modelling approaches for economic losses studies caused by parasites and their associated diseases in cattle. Parasitology. 2019;146(2):129–141. doi:10.1017/S0031182018001282.

99. Kotloff KL, Nataro JP, Blackwelder WC, Nasrin D, Farag TH, Panchalingam S, et al. Burden and aetiology of diarrhoeal disease in infants and young children in developing countries (the Global Enteric Multicenter Study, GEMS): a prospective, case-control study. The Lancet. 2013;382(9888):209–222. doi:https://doi.org/10.1016/S0140-6736(13)60844-2.

100. Havelaar AH, Kirk MD, Torgerson PR, Gibb HJ, Hald T, Lake RJ, et al. World Health Organization Global Estimates and Regional Comparisons of the Burden of Foodborne Disease in 2010. PLOS Medicine. 2015;12(12):1–23. doi:10.1371/journal.pmed.1001923.

101. Di Cristina M, Carruthers VB. New and emerging uses of CRISPR/Cas9 to genetically manipulate apicomplexan parasites. Parasitology. 2018;145(9):1119–1126. doi:10.1017/S003118201800001X.

102. Carruthers VB. Host cell invasion by the opportunistic pathogen Toxoplasma gondii. Acta Tropica. 2002;81(2):111–122. doi:https://doi.org/10.1016/S0001-706X(01)00201-7.

103. Kats LM, Cooke BM, Coppel RL, Black CG. Protein Trafficking to Apical Organelles of Malaria Parasites – Building an Invasion Machine. Traffic. 2008;9(2):176–186. doi:https://doi.org/10.1111/j.1600-0854.2007.00681.x.

104. Lebrun M, Carruthers VB, Cesbron-Delauw MF. Chapter 12 - Toxoplasma Secretory Proteins and Their Roles in Cell Invasion and Intracellular Survival. In: Weiss LM, Kim K, editors. Toxoplasma Gondii (Second Edition). second edition ed. Boston: Academic Press; 2014. p. 389–453. Available from: https://www.sciencedirect.com/science/article/pii/B978012396481600012X.

105. Yeoh S, O'Donnell RA, Koussis K, Dluzewski AR, Ansell KH, Osborne SA, et al. Subcellular Discharge of a Serine Protease Mediates Release of Invasive Malaria Parasites from Host Erythrocytes. Cell. 2007;131(6):1072–1083. doi:https://doi.org/10.1016/j.cell.2007.10.049.

106. Hakimi MA, Olias P, Sibley LD. *Toxoplasma* Effectors Targeting Host Signaling and Transcription. Clinical Microbiology Reviews. 2017;30(3):615–645. doi:10.1128/CMR.00005-17.

107. Dobrowolski JM, Sibley LD. Toxoplasma Invasion of Mammalian Cells Is Powered by the Actin Cytoskeleton of the Parasite. Cell. 1996;84(6):933–939. doi:https://doi.org/10.1016/S0092-8674(00)81071-5.

108. Maréchal E, Cesbron-Delauw MF. The apicoplast: a new member of the plastid family. Trends in Plant Science. 2001;6(5):200–205. doi:https://doi.org/10.1016/S1360-1385(01)01921-5.

109. Behnke MS, Wootton JC, Lehmann MM, Radke JB, Lucas O, Nawas J, et al. Coordinated Progression through Two Subtranscriptomes Underlies the Tachyzoite Cycle of Toxoplasma gondii. PLOS ONE. 2010;5(8):1–20. doi:10.1371/journal.pone.0012354.

110. Radke JR, White MW. A cell cycle model for the tachyzoite of Toxoplasma gondii using the Herpes simplex virus thymidine kinase. Molecular and Biochemical Parasitology. 1998;94(2):237–247. doi:https://doi.org/10.1016/S0166-6851(98)00074-7.

111. Back PS, O'Shaughnessy WJ, Moon AS, Dewangan PS, Hu X, Sha J, et al. Ancient MAPK ERK7 is regulated by an unusual inhibitory scaffold required for Toxoplasma apical complex biogenesis. Proceedings of the National Academy of Sciences. 2020;117(22):12164–12173.

112. Guérin A, Strelau KM, Barylyuk K, Wallbank BA, Berry L, Crook OM, et al. Cryptosporidium uses multiple distinct secretory organelles to interact with and modify its host cell. Cell Host & Microbe. 2023;31(4):650–664.

113. Mercier C, Cesbron-Delauw MF. Toxoplasma secretory granules: one population or more? Trends in parasitology. 2015;31(2):60–71.

Supplement: Semi-supervised Bayesian integration of multiple spatial proteomics datasets

Stephen Coleman[1*], Lisa Breckels[2], Ross F. Waller[2], Kathryn S. Lilley[2], Chris Wallace[1,3,4], Oliver M. Crook[5†], Paul D.W. Kirk[1,3,4†]

1 MRC Biostatistics Unit, University of Cambridge, Cambridge, UK
2 Department of Biochemistry, University of Cambridge, Cambridge, UK
3 Department of Medicine, University of Cambridge, Cambridge, UK
4 Cambridge Institute of Therapeutic Immunology & Infectious Disease, Cambridge, UK
5 Department of Statistics, University of Oxford, Oxford, UK

† These authors contributed equally to this work.
* colemste@ohsu.edu

S.1 Derivations

S.1.1 Gaussian process mixture models

If we consider the mean vector of a Gaussian density as the finite realisation of an infinite function in our feature space, we can place a Gaussian process prior on it. Thus, for the kth component of this mixture, mean vector μ_k and observed data which has been transformed to a vector, $X_k = [X_{k_1}, \ldots, X_{k_{N_k}}]$,

$$X_k|\mu_k, \sigma_k \sim \mathcal{N}(\mu_k, \sigma_k^2 I_P), \tag{1}$$

$$\mu_k|a_k, l_k \sim GP(0, C_k). \tag{2}$$

As we assume exchangeable data for ease of notation we permute the original dataset such that we can denote X_k with contiguous indices, i.e., $X_k = [X_1, \ldots, X_{N_k}]$. The posterior distribution can be derived by using the properties of the multivariate Gaussian relating to conditional distribution of two jointly distributed Gaussian variables (see Williams and Rasmussen, 2006),

$$\begin{bmatrix} y \\ f(X^*) \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} C(X,X) + \sigma^2\mathbf{I} & C(X,X^*) \\ C(X,X^*) & C(X^*,X^*) \end{bmatrix}\right), \tag{3}$$

$$\implies \begin{cases} \mathbb{E}(f(X^*)|y) & = C(X^*,X)(C(X,X) + \sigma^2\mathbf{I})^{-1}y, \\ \mathbb{C}ov(f(X^*)|y) & = C(X^*,X^*) - C(X^*,X)(C(X,X) + \sigma^2\mathbf{I})^{-1}C(X,X^*). \end{cases} \tag{4}$$

We can then say

$$(\mu_k|X_k, \sigma_k, a_k, l_k) \sim GP(m_k, \tilde{C}_k), \tag{5}$$

$$m_k = C_k[1:P](C_k + \sigma_k^2 I_{N_K P})^{-1}X_k, \tag{6}$$

$$\tilde{C}_k = C_k[1:P, 1:P] - C_k[1:P,](C_k + \sigma_k^2 I_{N_K P})^{-1}C_k[,1:P]. \tag{7}$$

Note that this involves taking the inverse of an $N_k P \times N_k P$ matrix, which is frequently a very costly calculation. There has been much exploration of scaling GP models such as sparse approximations (Quinonero-Candela and Rasmussen, 2005) and low-dimensional approximations (Banerjee et al., 2012), see Liu et al. (2020) for a recent review of the subject. However, in our application we have rich structure in the data as all items within a view have the same time/spatial measurements with consistent distance between each measurement and we can use this to reduce the computational load without using any of these more complex methods (Zhang et al., 2005).

Let A_k be the $P \times P$ matrix defined by the squared exponential kernel function for the P measurements of a single observation, i.e.,

$$(A_k)_{ij} = a_k^2 \exp\left\{-\frac{d(t_i,t_j)^2}{2l_k^2}\right\},\ i,j = \{1,\ldots,P\}, \tag{8}$$

then A_k is positive symmetric and Toeplitz (Golub and Van Loan, 2013) and

$$C_k + \sigma_k^2 I_{N_K P} = \begin{bmatrix} A_k + \sigma_k^2 I_P & A_k & \cdots & A_k \\ A_k & A_k + \sigma_k^2 I_P & \cdots & A_k \\ \vdots & \vdots & \ddots & \vdots \\ A_k & A_k & \cdots & A_k + \sigma_k^2 I_P \end{bmatrix}. \tag{9}$$

As Crook et al. (2022) have shown, we can write

$$C_k = J_{N_k} \otimes A_k \tag{10}$$

where J_{N_k} is $N_k \times N_k$ matrix of ones and $\otimes$ is the Kronecker product. Following the derivation from the supplement of Crook et al. (2022), by letting $Q_k = I_P + \sigma_k^{-2} N_k A_k$ we can write

$$(C_k + \sigma_k^2 I_{N_K P})^{-1} = \frac{1}{\sigma_k^2} I_{N_k P} - \frac{1}{N_k \sigma_k^2} J_{N_k} \otimes (I_P - Q^{-1}). \tag{11}$$

Thus only a $P \times P$ matrix has to be inverted. Writing this out explicitly:

$$(C_k + \sigma_k^2 I_{N_K P})^{-1} = \frac{1}{\sigma_k^2} \begin{bmatrix} I_P - \frac{1}{N_k}(I_P - Q^{-1}) & -\frac{1}{N_k}(I_P - Q^{-1}) & \cdots & -\frac{1}{N_k}(I_P - Q^{-1}) \\ -\frac{1}{N_k}(I_P - Q^{-1}) & I_P - \frac{1}{N_k}(I_P - Q^{-1}) & \cdots & -\frac{1}{N_k}(I_P - Q^{-1}) \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{1}{N_k}(I_P - Q^{-1}) & -\frac{1}{N_k}(I_P - Q^{-1}) & \cdots & I_P - \frac{1}{N_k}(I_P - Q^{-1}) \end{bmatrix} \tag{12}$$

By noticing that each block column consists of a single occurrence of the identity matrix and N_k occurrences of $-\frac{1}{N_k}(I_P - Q^{-1})$, and this object only occurs when multiplied by C_k, a matrix consisting of N_k repeated blocks of A_k, we can simplify the object of interest to

$$C_k[1:P,](C_k + \sigma_k^2 I_{N_K P})^{-1} = \frac{1}{\sigma_k^2} \begin{bmatrix} (A_k - A_k \times (I_p - Q^{-1})) & \cdots & (A_k - A_k \times (I_p - Q^{-1})) \end{bmatrix}. \tag{13}$$

This object occurs in both the posterior covariance function and the posterior mean function. In the latter,

$$C_k[1:P,](C_k + \sigma_k^2 I_{N_K P})^{-1} X_k = \frac{1}{\sigma_k^2} \begin{bmatrix} (A_k - A_k \times (I_p - Q^{-1})) & \cdots & (A_k - A_k \times (I_p - Q^{-1})) \end{bmatrix} X_k \tag{14}$$

$$= \frac{1}{\sigma_k^2}(A_k - A_k \times (I_p - Q^{-1}))(X_1 + X_2 + \cdots + X_{N_k}) \tag{15}$$

$$= \frac{N_k}{\sigma_k^2}(A_k - A_k \times (I_p - Q^{-1}))\bar{X}_k, \tag{16}$$

where $\bar{X}_k$ is the sample mean of the component data.

For the final product of $C_k[1:P,](C_k + \sigma_k^2 I_{N_K P})^{-1} C_k[,1:P]$, we note that we have reduced the multiplication of the first two objects to a simple repetitive block structure. Thus

$$C_k[1:P,](C_k + \sigma_k^2 I_{N_K P})^{-1} C_k[,1:P] = \frac{N_k}{\sigma_k^2}(A_k - A_k \times (I_p - Q^{-1}))A_k. \tag{17}$$

This means that only a single $P \times P$ matrix needs to be inverted and only two cases of multiplying a pair of $P \times P$ matrices must be performed rather than considering the inversion of a $N_k P \times N_k P$ matrix and its multiplication by a $P \times N_k P$ matrix and the result's multiplication by $N_k P \times P$ matrix.

We can now rewrite the posterior parameters as

$$m_k = \frac{N_k}{\sigma_k^2}(A_k - A_k \times (I_p - Q^{-1}))\bar{X}_k, \tag{18}$$

$$\tilde{C}_k = A_k - \frac{N_k}{\sigma_k^2}(A_k - A_k \times (I_p - Q^{-1}))A_k. \tag{19}$$

S.1.2 MDI: Derivation of $p(\phi_{(i,j)}|\cdot)$

A popular methods for performing inference on a Bayesian clustering method is to construct a Gibbs sampler. This uses the conditional densities of the model parameters. To derive these for MDI, we begin with the likelihood for the allocation of the nth individual in each of the V views, $c_n = (c_n^{(1)}, \ldots, c_n^{(V)})$ is

$$p(c_n^{(1)}, \ldots, c_n^{(V)}|\cdot) = \frac{S^{N-1}\exp(-SZ)}{(N-1)!} \times \prod_{n=1}^{N}\left(\prod_{v=1}^{V}\gamma_{c_n^{(v)}}^{(v)}\prod_{v=1}^{V-1}\prod_{w=v+1}^{V}(1+\phi_{(v,w)}\mathbb{I}(c_n^{(v)}=c_n^{(w)})\right), \tag{20}$$

for N items being clustered using K components in each view, normalising constant Z, unnormalised component weights γ, information sharing parameter vector ϕ, and a strategic latent variable S that is introduced to induce closed posterior distributions for many of the variables. The conditional posterior distribution for these objects can be found in the supplementary material of Kirk et al. (2012). In this supplement, the conditional distribution of the information sharing parameters, $\phi_{(i,j)}(i,j \in \{1,\ldots,V\} \subset \mathbb{N}, i<j)$ is stated as

$$\phi_{(i,j)} \sim Ga(\alpha,\beta), \tag{21}$$

where $Ga(a,b)$ is a Gamma distribution parameterised by a shape and a rate and

$$\alpha = 1 + \sum_{n=1}^{N}\mathbb{I}(c_n^{(i)}=c_n^{(j)}), \tag{22}$$

$$\beta = S\sum_{k_i=k_j=1}^{K}\sum_{k_1=1}^{K}\cdots\sum_{k_{i-1}=1}^{K}\sum_{k_{i+1}=1}^{K}\cdots\sum_{k_{j-1}=1}^{K}\sum_{k_{j+1}=1}^{K}\cdots\sum_{k_V=1}^{K}\prod_{v=1}^{V}\gamma_{k_v}^{(v)} \times \prod_{v=1}^{V-1}\prod_{w=v+1;w\neq j}^{V}(1+\phi_{(v,w)}\mathbb{I}(k_v=k_w))\prod_{v=1}^{j-1}(1+\phi_{(v,j)}\mathbb{I}(k_v=k_j)). \tag{23}$$

However, in 2017 Stephen Johnson, Daniel Henderson, and Richard Boys noticed that this requires a correction which we derive here. Note that we assume $K_1 = K_2 = \cdots = K_V = K$, but this is not necessary and if it does not hold any quantities any occurrence of K can be replaced with the appropriate K_v for all $v = 1,\ldots,V$. To derive $p(\phi_{(i,j)}|\cdot)$, for some $i,j \in \{1,\ldots,V\}, i<j$, we derive the marginal likelihood of $\phi_{(i,j)}$ and then use this to find the closed form of the posterior distribution.

$$p(\phi_{(i,j)}|\cdot) \propto \exp\{-SZ\}\prod_{n=1}^{N}\left(\prod_{v=1}^{V-1}\prod_{w=v+1}^{V}(1+\phi_{(v,w)}\mathbb{I}(c_n^{(v)}=c_n^{(w)})\right) \tag{24}$$

we consider this in two parts, first the normalising constant in the exponential function and then the product on the right. The normalising constant, Z can be found by considering all possible combinations of states of the labels:

$$Z = \sum_{k_1=1}^{K}\cdots\sum_{k_V=1}^{K}\left(\prod_{v=1}^{V}\gamma_{k_v}^{(v)}\prod_{v=1}^{V-1}\prod_{w=v+1}^{V}(1+\phi_{(v,w)}\mathbb{I}(c_n^{(v)}=c_n^{(w)})\right). \tag{25}$$

Considering only the dependency of Z on $\phi_{(i,j)}$:

$$Z \propto \phi_{(i,j)}\sum_{k_i=k_j=1}^{K}\sum_{k_1=1}^{K}\cdots\sum_{k_{i-1}=1}^{K}\sum_{k_{i+1}=1}^{K}\cdots\sum_{k_{j-1}=1}^{K}\sum_{k_{j+1}=1}^{K}\cdots\sum_{k_V=1}^{K}\prod_{v=1}^{V}\gamma_{k_v}^{(v)} \times \prod_{v=1}^{V-1}\prod_{w=v+1;w\neq j}^{V}(1+\phi_{(v,w)}\mathbb{I}(k_v=k_w))\prod_{v=1}^{j-1}(1+\phi_{(v,j)}\mathbb{I}(k_v=k_j)). \tag{26}$$

The product simplifies as a function of $\phi_{(i,j)}$ to

$$\begin{aligned}\prod_{n=1}^{N}\prod_{v=1}^{V-1}\prod_{w=v+1}^{V}(1+\phi_{(v,w)}\mathbb{I}(c_n^{(v)}=c_n^{(w)}) &\propto \prod_{n=1}^{N}(1+\phi_{(i,j)})\mathbb{I}(c_n^{(i)}=c_n^{(j)})\\ &= (1+\phi_{(i,j)})^{\sum_{n=1}^{N}\mathbb{I}(c_n^{(i)}=c_n^{(j)})}\\ &= \sum_{r=0}^{\sum_{n=1}^{N}\mathbb{I}(c_n^{(i)}=c_n^{(j)})}\binom{\sum_{n=1}^{N}\mathbb{I}(c_n^{(i)}=c_n^{(j)})}{r}\phi_{(i,j)}^{r}, \qquad (27)\end{aligned}$$

by the binomial theorem. $\sum_{n=1}^{N}\mathbb{I}(c_n^{(i)}=c_n^{(j)})$ is the number of items with the same label in both datasets. We denote this quantity by N_{ij} to reduce clutter.

Combining equations 26 and 27 yields:

$$\begin{aligned}p(\phi_{(i,j)}|\cdot) \propto \exp\Bigg(&-\phi_{(i,j)}S\sum_{k_i=k_j=1}^{K}\sum_{k_1=1}^{K}\cdots\sum_{k_{i-1}=1}^{K}\sum_{k_{i+1}=1}^{K}\cdots\sum_{k_{j-1}=1}^{K}\sum_{k_{j+1}=1}^{K}\cdots\sum_{k_V=1}^{K}\prod_{v=1}^{V}\gamma_{k_v}^{(v)}\\ &\times\prod_{v=1}^{V-1}\prod_{w=v+1;v\neq j}^{V}(1+\phi_{(v,w)}\mathbb{I}(k_v=k_w))\prod_{v=1}^{j-1}(1+\phi_{(v,j)}\mathbb{I}(k_v=k_j))\Bigg)\\ &\times\sum_{r=0}^{N_{ij}}\binom{N_{ij}}{r}\phi_{(i,j)}^{r}. \qquad (28)\end{aligned}$$

If we let $g(x;a,b) := \frac{b^a}{\Gamma(a)}x^{a-1}\exp(-bx)$ be the gamma density function, we can transform this to a mixture of Gamma distributions:

$$\begin{aligned}p(\phi_{(i,j)}|\cdot) \propto \Bigg(&S\sum_{k_i=k_j=1}^{K}\sum_{k_1=1}^{K}\cdots\sum_{k_{i-1}=1}^{K}\sum_{k_{i+1}=1}^{K}\cdots\sum_{k_{j-1}=1}^{K}\sum_{k_{j+1}=1}^{K}\cdots\sum_{k_V=1}^{K}\prod_{v=1}^{V}\gamma_{k_v}^{(v)}\\ &\times\prod_{v=1}^{V-1}\prod_{w=v+1;w\neq j}^{V}(1+\phi_{(v,w)}\mathbb{I}(k_v=k_w))\prod_{v=1}^{j-1}(1+\phi_{(v,j)}\mathbb{I}(k_v=k_j))\Bigg)^{r+1}\Gamma(r+1)^{-1}\\ &\times\Bigg(S\sum_{k_i=k_j=1}^{K}\sum_{k_1=1}^{K}\cdots\sum_{k_{i-1}=1}^{K}\sum_{k_{i+1}=1}^{K}\cdots\sum_{k_{j-1}=1}^{K}\sum_{k_{j+1}=1}^{K}\cdots\sum_{k_V=1}^{K}\prod_{v=1}^{V}\gamma_{k_v}^{(v)}\\ &\times\prod_{v=1}^{V-1}\prod_{w=v+1;w\neq j}^{V}(1+\phi_{(v,w)}\mathbb{I}(k_v=k_w))\prod_{v=1}^{j-1}(1+\phi_{(v,j)}\mathbb{I}(k_v=k_j))\Bigg)^{-(r+1)}\Gamma(r+1)\\ &\times\exp\Bigg(-\phi_{(i,j)}S\sum_{k_i=k_j=1}^{K}\sum_{k_1=1}^{K}\cdots\sum_{k_{i-1}=1}^{K}\sum_{k_{i+1}=1}^{K}\cdots\sum_{k_{j-1}=1}^{K}\sum_{k_{j+1}=1}^{K}\cdots\sum_{k_V=1}^{K}\prod_{v=1}^{V}\gamma_{k_v}^{(v)}\\ &\times\prod_{v=1}^{V-1}\prod_{w=v+1;w\neq j}^{V}(1+\phi_{(v,w)}\mathbb{I}(k_v=k_w))\prod_{v=1}^{j-1}(1+\phi_{(v,j)}\mathbb{I}(k_v=k_j))\Bigg)\\ &\times\sum_{r=0}^{N_{ij}}\binom{N_{ij}}{r}\phi_{(i,j)}^{r}, \qquad (29)\\ =&\sum_{r=0}^{N_{ij}}\binom{N_{ij}}{r}\frac{\Gamma(r+1)}{\beta^{r+1}}g(\phi_{(i,j)};r+1,\beta), \qquad (30)\end{aligned}$$

where

$$\beta = S \sum_{k_i=k_j=1}^{K} \sum_{k_1=1}^{K} \cdots \sum_{k_{i-1}=1}^{K} \sum_{k_{i+1}=1}^{K} \cdots \sum_{k_{j-1}=1}^{K} \sum_{k_{j+1}=1}^{K} \cdots \sum_{k_V=1}^{K} \prod_{v=1}^{V} \gamma_{k_v}^{(v)}$$
$$\times \prod_{v=1}^{V-1} \prod_{w=v+1;w\neq j}^{V} (1+\phi_{(v,w)}\mathbb{I}(k_v=k_w)) \prod_{v=1}^{j-1}(1+\phi_{(v,j)}\mathbb{I}(k_v=k_j)) \tag{31}$$

is the same as in the derivation for this posterior in the original paper.

If we then introduce a gamma prior distribution over $\phi_{(i,j)}$, say

$$\phi_{(i,j)} \sim Ga(\alpha_0, \beta_0), \tag{32}$$

then the posterior distribution becomes

$$p(\phi_{(i,j)}|\cdot) \propto \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \phi_{(i,j)}^{\alpha_0-1} \exp(-\beta_0\phi_{(i,j)}) \exp(-\phi_{(i,j)}\beta) \sum_{r=0}^{N_{ij}} \binom{N_{ij}}{r} \phi_{(i,j)}^{r} \tag{33}$$
$$\propto \sum_{r=0}^{N_{ij}} \binom{N_{ij}}{r} \phi_{(i,j)}^{r+\alpha_0-1} \exp(-\phi_{(i,j)}(\beta+\beta_0)) \tag{34}$$
$$= \sum_{r=0}^{N_{ij}} \binom{N_{ij}}{r} g(\phi_{(i,j)}|r+\alpha_0, \beta+\beta_0)\Gamma(r+\alpha_0)/(\beta+\beta_0)^{(r+\alpha_0)}. \tag{35}$$

which is clearly a mixture of Gamma densities, which can be rewritten as

$$p(\phi_{ij}|\cdot) = \sum_{r=0}^{N_{ij}} \pi_r g(\phi_{ij}|r+\alpha_0, \beta+\beta_0), \tag{36}$$

where

$$\pi_r = \frac{\binom{N_{ij}}{r}\Gamma(r+\alpha_0)/(\beta+\beta_0)^{(r+\alpha_0)}}{\sum_{r=0}^{N_{ij}} \binom{N_{ij}}{r}\Gamma(r+\alpha_0)/(\beta+\beta_0)^{(r+\alpha_0)}}. \tag{37}$$

The component of this mixture with the largest expected value is the same as the quantity in equation 21. This correction means that the sampled $\phi_{(i,j)}$'s will be smaller on average than in the original implementations, resulting in slightly less weight being placed by the model on common or overlapping partitions across views than in previous implementations.

S.2 Simulation study

Each scenario is defined by the density used to generate the clusters (the generative model) and by the ϕ vector ($\phi = (12, 8, 4)$ for our simulation study) which determines the similarity of the clustering structure across modalities. We first generate the labels indicating the cluster an item is generated from in each modality and then sample measurements based on this cluster's parameters. In every simulation, the first dataset is generated from six clusters, the second from seven clusters and the third from eight clusters. We

generate two hundred labels and data points in each datasets from this model:

$$p(c_n^{(v)} = k|\gamma, \phi) \propto \gamma_k \prod_{w=1;w\neq v}^{V} (1 + \phi_{v,w}), \tag{38}$$

$$(X_n^{(v)}|c_n^{(v)} = k) \sim f(\theta_k), \tag{39}$$

$$\phi = (12, 8, 4), \tag{40}$$

$$K = (6, 7, 8), \tag{41}$$

$$\gamma^{(1)} = [4.0, 7.0, 7.0, 5.5, 3.0, 5.5]^\top, \tag{42}$$

$$\gamma^{(2)} = [5.5, 4.0, 4.0, 4.0, 4.0, 4.0, 3.0]^\top, \tag{43}$$

$$\gamma^{(3)} = [6.0, 6.0, 5.0, 3.0, 5.5, 7.0, 4.0, 4.0]^\top. \tag{44}$$

There are fifteen measurements / feature for each data point (i.e., $N = 200, P = 15$ in each dataset in all simulations), and in the first dataset of each simulation we have 6 clusters, in the second 7 and the third 8. Within a generating seed the labels in each modality are the same across all scenarios and each dataset is informative of the others ($\phi = (12, 8, 4)$). We randomly sample an expected 30% of the labels in the first dataset to be observed (or considered as training data) in the semi-supervised model runs.

We consider three different generating models which give the scenarios their names. The first two are

$$\text{Scenario 1 (Gaussian): } (X_n^{(v)}|c_n^{(v)} = k) \sim \mathcal{N}(\mu_k, \Sigma_k), \tag{45}$$

$$\text{Scenario 2 (MVT): } (X_n^{(v)}|c_n^{(v)} = k) \sim t_{\eta_k}(\mu_k, \Sigma_k). \tag{46}$$

Scenario 3, the Log-Poisson case, is more complex.

$$(Y_{n,p}|c_n^{(v)} = k) \sim Poisson(\lambda_{k,p}), \tag{47}$$

$$\epsilon_{n,p} \sim \mathcal{N}(0, 1), \tag{48}$$

$$X_{n,p} = \log(1 + Y_{n,p}) + \epsilon_{n,p}. \tag{49}$$

We randomly permute the cluster parameters in each feature/measurement. In the case of Scenarios 1 and 2, the generating means and standard deviations are:

$$\mu^{(1)} = [-1.50, -1.25, 1.25, 0.50, -1.00, 1.00, 1.75]^\top \tag{50}$$

$$\mu^{(2)} = [-1.00, -0.50, -0.75, 0.00, 0.25, 0.50, 0.75, 0.50, 0.00]^\top, \tag{51}$$

$$\mu^{(3)} = [-1.00, -0.50, -1.25, -0.50, -0.25, -0.25, 0.25, 1.25, 1.25, 1.00, 0.50]^\top, \tag{52}$$

$$\sigma^{(1)} = [1.50, 1.25, 1.75, 2.25, 2.75, 1.50, 3.0]^\top, \tag{53}$$

$$\sigma^{(2)} = [0.5, 1.0, 0.5, 1.0, 0.5, 0.25, 1.25, 0.5, 1.00^\top \tag{54}$$

$$\sigma^{(3)} = [0.50, 0.75, 1.50, 1.00, 0.50, 0.75, 0.75, 0.50, 1.50, 1.00, 0.50]^\top. \tag{55}$$

For scenario 2 we have, in addition to these means and standard deviations, the following degrees of freedom:

$$\nu^{(1)} = [5, 10, 5, 5, 10, 5, 10]^\top, \tag{56}$$

$$\nu_k^{(2)} \sim \mathcal{U}\{5, 10\} \text{ independently for each } k = 1 \ldots, K_2, \tag{57}$$

$$\nu_k^{(3)} \sim \mathcal{U}\{5, 10\} \text{ independently for each } k = 1 \ldots, K_3. \tag{58}$$

In Scenario 3 we have:

$$\lambda^{(1)} = [5, 15, 25, 35, 30, 15, 40]^\top \quad , \tag{59}$$

$$\lambda^{(2)} = [4, 13, 21, 26, 55, 43, 31, 28, 15]^\top, \tag{60}$$

$$\lambda^{(3)} = [4, 8, 44, 32, 16, 28, 55, 24, 12, 18, 28]^\top. \tag{61}$$

S.3 *Toxoplasma gondii* model convergence

The initial model runs displayed poor mixing in the LOPIT data (see figure S.1 A). To circumvent this issue we used consensus clustering Coleman et al. (2022). We found that an ensemble composed of the 15,000th sample from 150 chains displayed stability (figure S.2). The non-monotonic behaviour seen in figure S.2 A for the wider ensembles is due to the emergence of the shared signal between the 9,000 and 12,000 iterations in a majority of chains (note the mass shifting up the scale in the figure S.2 C as the depth increases). The consensus matrices for this inference and a similar consensus clustering for an unsupervised mixture model run on the microarray data can be seen in figure S.1 B; this also shows that the clusters driving the disagreement across individual chains have their membership uncertainty well-captured by consensus clustering. We also performed consensus clustering of a mixture model on the microarray data to investigate the impact of the joint modelling of the views. We use the same consensus clustering algorithm of the 15,000th iteration from 150 different chains as in the integrative analysis except we had to increase the number of components modelled to 300 as at 125 all available components were consistently occupied in MCMC samples in the initial model runs using the smaller number of components. This inference resulted in a much sparser consensus matrix than the MDI inference (see figure S.1 B).

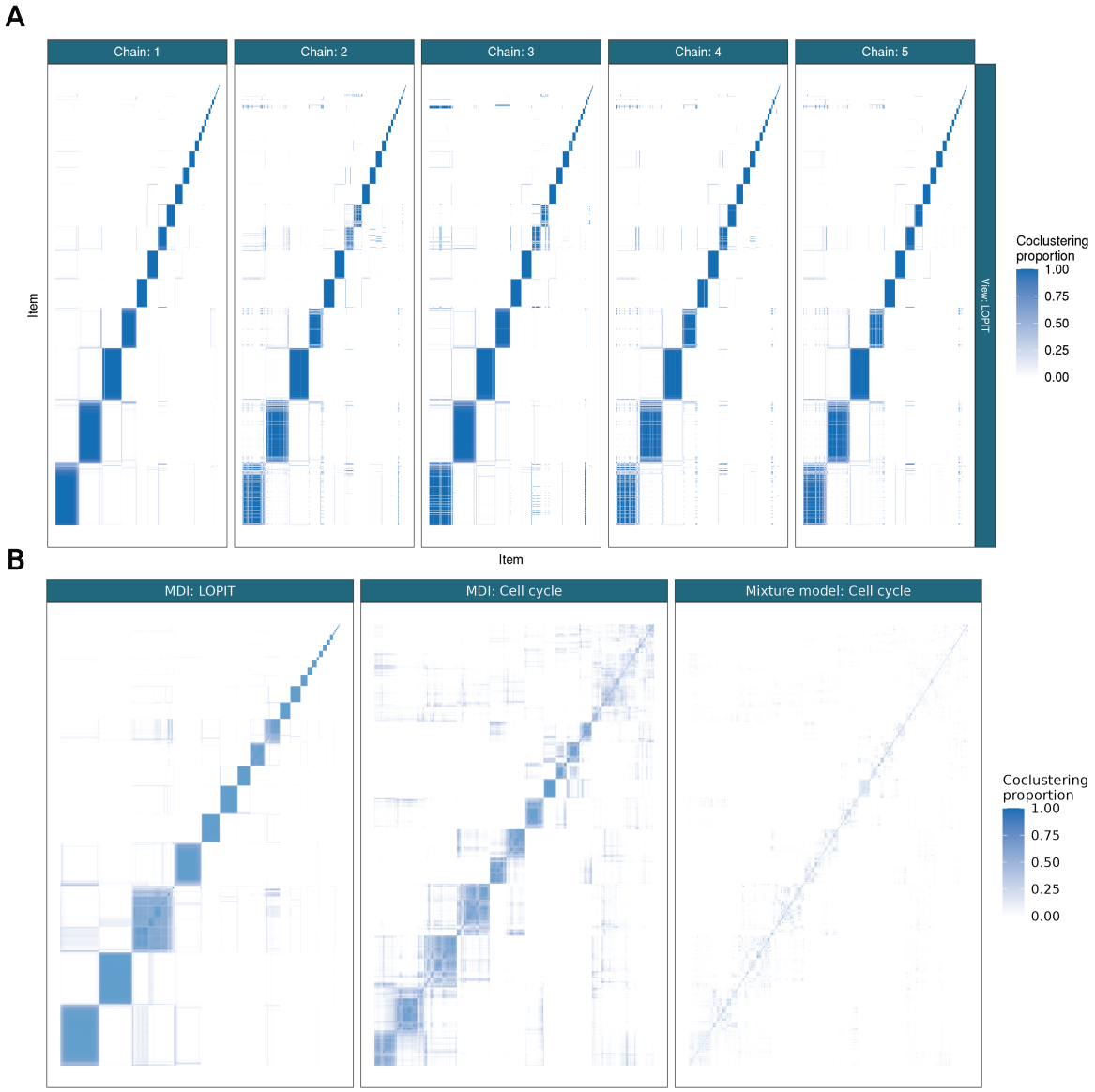

Figure S.1: A. PSMs for the LOPIT view of 5 chains of MDI run for 30,000 iterations with common row and column ordering (defined by chain 1). B. Consensus matrices for MDI in LOPIT and the cell-cycle microarray view and an unsupervised mixture model in the cell-cycle microarray data. The clusters that appear to be the cause of the poor mixing in the individual chains are highlighted with orange boxes in the LOPIT view.

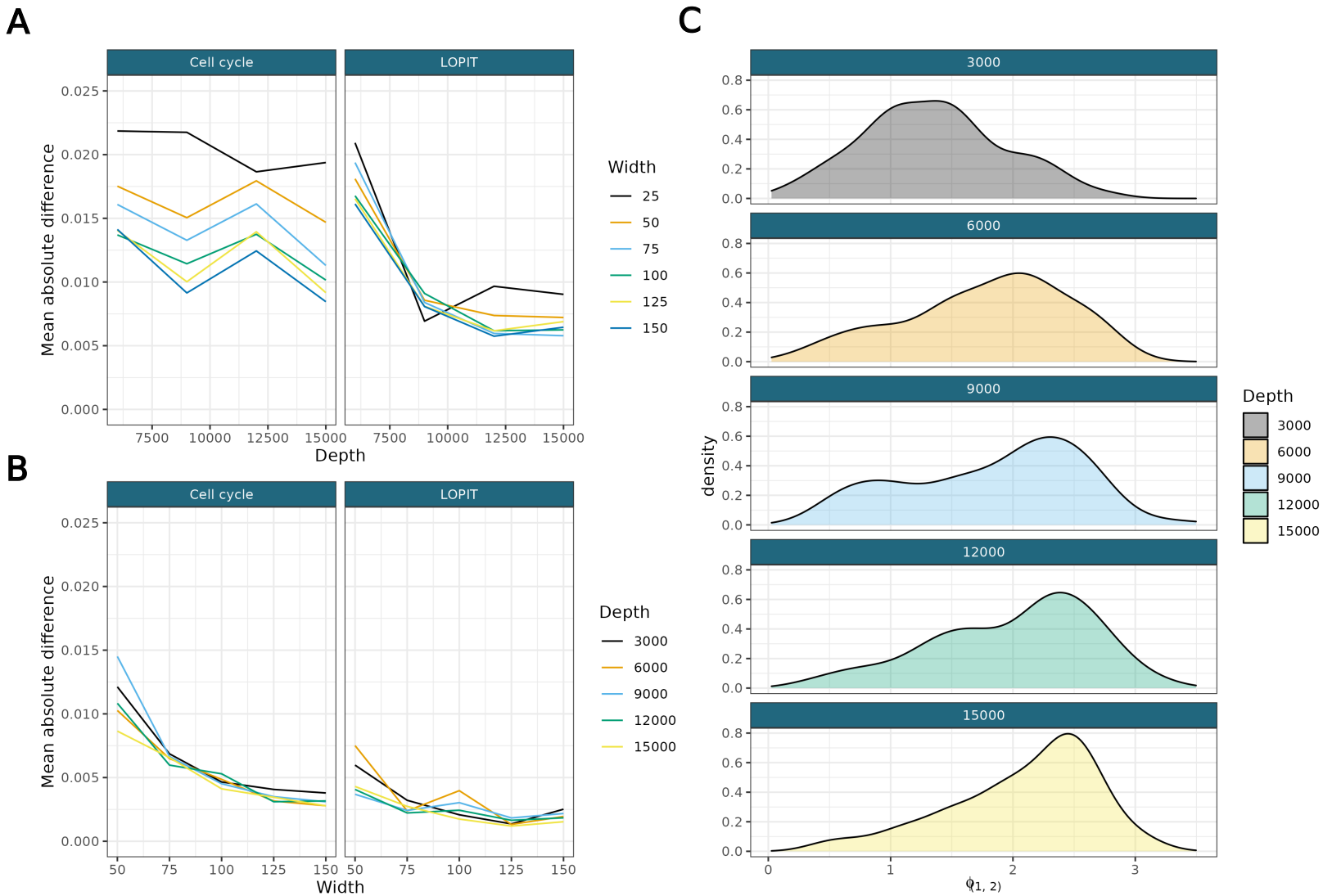

Figure S.2: A-B) Mean absolute difference between consensus matrices of increasing depth (A) and width (B). C) The sampled $\phi_{(1,2)}$ density for increasing depths.

References

Anjishnu Banerjee, David B. Dunson, and Surya T. Tokdar. Efficient Gaussian process regression for large datasets. *Biometrika*, 100(1):75–89, 12 2012. ISSN 0006-3444. doi: 10.1093/biomet/ass068. URL https://doi.org/10.1093/biomet/ass068.

Stephen Coleman, Paul D. W. Kirk, and Chris Wallace. Consensus clustering for bayesian mixture models. *BMC Bioinformatics*, 23(1):290, Jul 2022. ISSN 1471-2105. doi: 10.1186/s12859-022-04830-8. URL https://doi.org/10.1186/s12859-022-04830-8.

Oliver M. Crook, Kathryn S. Lilley, Laurent Gatto, and Paul D. W. Kirk. Semi-supervised nonparametric Bayesian modelling of spatial proteomics. *The Annals of Applied Statistics*, 16(4):2554 – 2576, 2022. doi: 10.1214/22-AOAS1603. URL https://doi.org/10.1214/22-AOAS1603.

Gene H Golub and Charles F Van Loan. *Matrix computations*. John Hopkins University press, 2013.

Paul Kirk, Jim E Griffin, Richard S Savage, Zoubin Ghahramani, and David L Wild. Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics*, 28(24):3290–3297, 2012.

Haitao Liu, Yew-Soon Ong, Xiaobo Shen, and Jianfei Cai. When Gaussian Process Meets Big Data: A Review of Scalable GPs. *IEEE Transactions on Neural Networks and Learning Systems*, 31(11):4405–4423, 2020. doi: 10.1109/TNNLS.2019.2957109.

Joaquin Quinonero-Candela and Carl Edward Rasmussen. A unifying view of sparse approximate Gaussian process regression. *The Journal of Machine Learning Research*, 6:1939–1959, 2005.

Christopher KI Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

Yunong Zhang, W.E. Leithead, and D.J. Leith. Time-series gaussian process regression based on toeplitz computation of o(n2) operations and o(n)-level storage. In *Proceedings of the 44th IEEE Conference on Decision and Control*, pages 3711–3716, 2005. doi: 10.1109/CDC.2005.1582739.