Bioinformatics, 2024, **40**, i189–i198
https://doi.org/10.1093/bioinformatics/btae216
ISMB 2024

OXFORD

Probabilistic pathway-based multimodal factor analysis

Alexander Immer[1,2,*], Stefan G. Stark[1,3], Francis Jacob[4], Ximena Bonilla[1,3], Tinu Thomas[1,3], André Kahles [1,3], Sandra Goetze[3,5,6], Emanuela S. Milani[3,5], Bernd Wollscheid[3,5], The Tumor Profiler Consortium, Gunnar Rätsch [1,3,7,8,9,*], Kjong-Van Lehmann [1,10,11,*]

[1]Biomedical Informatics Group, Department of Computer Science, ETH Zurich, 8092 Zurich, Switzerland
[2]Max Planck Institute for Intelligent Systems, 72076 Tübingen, Germany
[3]Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland
[4]Ovarian Cancer Research, Department of Biomedicine, University Hospital Basel and University of Basel, 4031 Basel, Switzerland
[5]Institute of Translational Medicine, Department of Health Sciences and Technology, ETH Zurich, 8093 Zurich, Switzerland
[6]ETH PHRT Swiss Multi-Omics Center (SMOC), 8093 Zurich, Switzerland
[7]AI Center at ETH Zurich, 8092 Zurich, Switzerland
[8]Biomedical Informatics Research, University Hospital Zurich, 8006 Zurich, Switzerland
[9]Department of Biology, ETH Zurich, 8049 Zurich, Switzerland
[10]Cancer Research Center Cologne Essen, University Hospital Cologne, 50937 Cologne, Germany
[11]Joint Research Center for Computational Biomedicine, University Hospital RWTH Aachen, 52074 Aachen, Germany

*Corresponding author. Biomedical Informatics Group, Department of Computer Science, ETH Zurich, Zurich, Switzerland.
E-mails: alexander.immer@inf.ethz.ch (A.I.), gunnar.ratsch@ratschlab.org (G.R.), and kjlehmann@ukaachen.de (K.-V.L.)

Abstract

Motivation: Multimodal profiling strategies promise to produce more informative insights into biomedical cohorts via the integration of the information each modality contributes. To perform this integration, however, the development of novel analytical strategies is needed. Multimodal profiling strategies often come at the expense of lower sample numbers, which can challenge methods to uncover shared signals across a cohort. Thus, factor analysis approaches are commonly used for the analysis of high-dimensional data in molecular biology, however, they typically do not yield representations that are directly interpretable, whereas many research questions often center around the analysis of pathways associated with specific observations.

Results: We develop PathFA, a novel approach for multimodal factor analysis over the space of pathways. PathFA produces integrative and interpretable views across multimodal profiling technologies, which allow for the derivation of concrete hypotheses. PathFA combines a pathway-learning approach with integrative multimodal capability under a Bayesian procedure that is efficient, hyper-parameter free, and able to automatically infer observation noise from the data. We demonstrate strong performance on small sample sizes within our simulation framework and on matched proteomics and transcriptomics profiles from real tumor samples taken from the Swiss Tumor Profiler consortium. On a subcohort of melanoma patients, PathFA recovers pathway activity that has been independently associated with poor outcome. We further demonstrate the ability of this approach to identify pathways associated with the presence of specific cell-types as well as tumor heterogeneity. Our results show that we capture known biology, making it well suited for analyzing multimodal sample cohorts.

Availability and implementation: The tool is implemented in python and available at https://github.com/ratschlab/path-fa

1 Introduction

A current trend in biomedical research is to obtain multiple types of molecular data for a given sample (Boehm *et al.* 2022). These modalities represent different views of the molecular landscape, each measuring different aspects, resolutions, and scales to provide us with complementary information that helps us understand the relationships between molecular mechanisms. A common question regarding multimodal data is to identify factors that explain the range of observed measurements (e.g. gene or protein expression, copy number variations, phosphorylation, etc.) across a given sample population (Ritchie *et al.* 2015). In molecular oncology, related approaches are used to quantify tumor composition and immune cell content from bulk measurement technologies, leveraging shared signals observed across a larger cohort (Chen *et al.* 2018).

Thus, extracting useful representations of samples from omics data that are biologically relevant, correlate with cell-type abundance, or even clinical variables, remains an important challenge. PLIER (Mao *et al.* 2019) and MultiPLIER (Taroni *et al.* 2019) leverage pathway level information to analyze factors that drive the differences in biology observed within a cohort that significantly improve interpretability compared to gene level approaches, but only apply to a single modality of transcriptomics data. MOFA (Argelaguet *et al.* 2018) conversely, is a multimodal factor analysis approach but operates on a potentially unidentifiable latent space instead of pathways, which can make interpretation challenging. By making less assumptions about the latent factors, it also requires more samples to infer them.

Here, we propose a novel approach for a multimodal factor analysis that operates on the level of biological pathways to integrate the information from multiple modalities. We leverage the concepts from PLIER and MOFA into a novel single framework, PathFA. We use pathway information to integrate multiple modalities into the same factor analysis model. This new multi-omics factor analysis is able to join *proteomics and transcriptomics* (RNA) data in the space of pathways (PLIER is designed for RNA) instead of unknown

latent factors (MOFA). To effectively integrate and balance different observed markers and modalities, we propose a probabilistic pathway-based factor analysis model and efficient Bayesian inference method.

Using PathFA, we show that both (1) the addition of another data modality and (2) utilization of prior knowledge improve reconstruction in a simulation and downstream performance on real data, for example correlation with cell types. Specifically, we improve over MOFA when it is possible to incorporate prior knowledge in form of pathways and over PLIER when multiple omics are available. On matched proteomics and transcriptomics data of the Tumor Profiler Consortium (Irmisch *et al.* 2021; The Tumor Profiler Consortium 2024a, 2024b), the latent factors of PathFA align with relevant pathways for clinical outcome, tumor heterogeneity, and cell-type composition. Further, we find that proteomics data by itself can be as useful as transcriptomics in our case, although pathways are originally derived on RNA markers.

In summary, in this study, we make the following main contributions:

1) We introduce PathFA, a novel probabilistic factor analysis model integrating multimodal data (transcriptomics and proteomics) using biological pathways.
2) We implemented an efficient Bayesian inference method for automatic hyper-parameter optimization, enhancing the analysis of high-dimensional molecular data.
3) We demonstrate improved interpretability of molecular datasets by anchoring analysis in known biological pathways.
4) We validate PathFA with real-world patient data from the Tumor Profiler consortium, successfully identifying key biological insights related to cell-type composition, survival, and tumor heterogeneity.

2 Methods

2.1 Data and preprocessing

We use the transcriptomics (RNA-Seq) and proteomics data (DIA-MS) data (Xuan *et al.* 2020) generated from Melanoma (The Tumor Profiler Consortium 2024b) and Ovarian patients (The Tumor Profiler Consortium 2024a), as part of the Tumor Profiler(TuPro) study (Irmisch *et al.* 2021). TuPro is a multicenter study that has deeply phenotyped metastatic tumor across multiple indications (Irmisch *et al.* 2021). All data will be made available upon release of the Tumor Profiler Marker papers.

2.1.1 Sample selection for transcriptomics and proteomics

Based on the transcriptomics and proteomics data from tumor profiler, we selected a subset of samples for ovarian cancer and melanoma. RNA-seq library preparation was done using the Illumina TruSeq Standed Total RNA Sample Preparation kit (Ribo-Zero Gold). Nova-Seq 6000 was used to sequence the samples. Proteomics data was generated via a Data Independent Acquisition Mass Spectrometry approach. We selected a subcohort of patients that had passed the internal quality control of both transcriptomics and proteomics nodes. RNA-seq data was filtered based on RIN score and fastqc metrics. Samples failed QC when the RIN score was below six, or if three or more FASTQC modules were failed. For proteomics data, quality control was assessed in context of reference control samples [mix of three Ovarian cell lines (Kuramochi, OVCA3, SKOV3) or for melanoma (MKFY6-1-P15/MKFY6-1-P15/MTG5K-1-P19)]. Samples with more than 70%, as well as peptides with more than 75%, missing values were removed. We also filtered for samples that had both modalities profiled, and a complete set of the relevant metadata. This left us with 42 ovarian samples and 34 melanoma samples.

2.1.2 Data preprocessing

Both transcriptomics and proteomics data are processed similarly and according to standard practice. In both cases, the raw data are first standardized by the library size, i.e. the total number of counts per sample. For RNA, this is equivalent to reads per million mapped reads (RPM). The counts are transformed with the following function: $\log(1+x)$, then, quantile normalized, and z-scored. The quantile normalization ensures that the transformed counts for each sample roughly follow a normal distribution and z-scoring standardizes per marker. The preprocessing is identical to Mao *et al.* (2019) with the difference that we use $\log(1+x)$ instead of filtering genes with low counts and applying $\log(x)$. The melanoma dataset had 19 612 genes and 4651 proteins. The ovarian dataset had 19 965 genes and 4068 proteins.

The prior knowledge about pathways is extracted from the Molecular Signatures Database (MSigDB; Liberzon *et al.* 2011) in addition to a *curated* set of pathways, developed by the Tumor Profiler Consortium for the analysis of melanoma pathways (Irmisch *et al.* 2021). From MSigDB, we use Hallmark (Liberzon *et al.* 2015) and cell-type signature genesets. However, these are specified in terms of contained genes only. To use the same pathways for proteomics data, we translate these pathways into the corresponding markers, mapping to Emsembl gene ids (Ruffier *et al.* 2017) via UniProt (Consortium 2019).

2.1.3 Synthetic data-generating process

To generate synthetic data for simulation experiments, we use MSigDB Hallmark pathways, which provide us with 50 pathways and corresponding transcriptomics and translated (Section 2.1.2) proteomics markers. Further, each pathway is associated to one of eight process categories, e.g. development, DNA damage, or immune, which we use as ground truth latent variables. This allows us to generate data by simulating loadings from a mixture of three isotropic Gaussians, simulating clusters of samples. The mean of each cluster is itself sampled from an isotropic Gaussian with higher variance. Further, we use a heteroscedastic observation noise on the markers that follows the shape of observed data in Tumor Profiler data. The average transcriptomics observation noise is 0.95 and that of Proteomics is 0.98 while the spread of noises on the marker level is smaller for Proteomics as shown in Fig. 1B.

2.2 Probabilistic pathway-based factor analysis

PathFA uses prior information in terms of associations of pathways to either gene or protein markers. This allows our model to use the pathways as a latent space based on prior knowledge. PLIER (Mao *et al.* 2019) uses the same prior information in their model but only for transcriptomics data and in combination with a standard factor analysis (FA) model. Below, we introduce our method that is both applicable to transcriptomics and proteomics data and we extend it

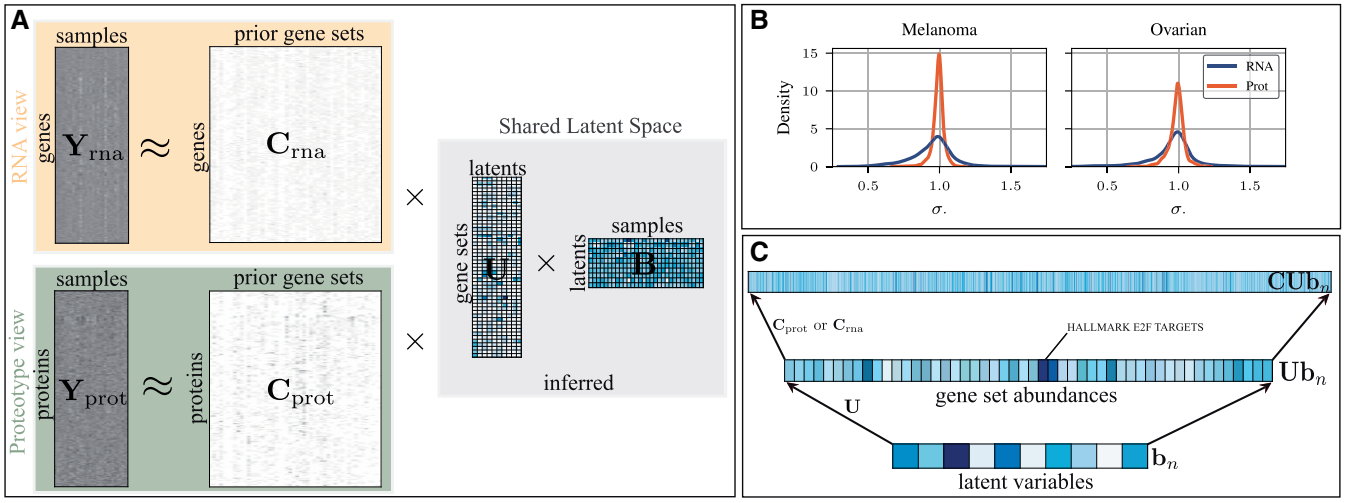

Figure 1. Schematic overview of our method. (A) Samples are hierarchically represented via pathways and latents inferred jointly from transcriptomics (RNA) and proteomics observations. Corresponding prior pathways translate into the space of both observed modalities. (B) Density plot of observation noises for both RNA and proteomics data shows heteroscedasticity within and across modalities. Proteomics markers have less varying precision while RNA has more spread. (C) Hierarchy of representations for a single sample. A sample is represented by a low-dimensional latent variable, projected into pathway abundances, which can be reconstructed into both proteomics and RNA observations.

to the multimodal setting (observing both per sample) in Section 2.4.

We denote the data of m gene or protein markers for n samples as matrix $\mathbf{Y} \in \mathbb{R}^{m \times n}$. Further, we assume prior knowledge about p relevant pathways that each consist of several markers (gene or protein) as $\mathbf{C} \in \{0,1\}^{m \times p}$ with $C_{m,p} = 1$ if the mth marker is part of the pth pathway and otherwise $C_{mp} = 0$. PathFA infers two parameter matrices: $\mathbf{U} \in \mathbb{R}_{\geq 0}^{p \times k}$ associates the pathways to latent variables such that $U_{p,k}$ denotes the relevance of the pth pathway to the kth latent variable. Associating latent variables through pathways allows to interpret them better than in a standard factor analysis. Finally, each column of the matrix $\mathbf{B} \in \mathbb{R}^{k \times n}$ specifies the k-dimensional loadings per sample. Informally, we model the observations $\mathbf{Y}$ with $\mathbf{CUB}$. Typically, we have m markers in the order of thousands, p pathways in the order of hundreds, and tens of k latent dimensions.

PathFA uses a Gaussian likelihood function with learnable observation noises per marker. This allows PathFA to ignore noisy markers and focus on the informative ones as they are actually heteroscedastic (Argelaguet *et al.* 2018). The observation noise is parameterized by $\boldsymbol{\sigma} \in \mathbb{R}_+^m$ with σ_m denoting the observation noise of the mth marker. The likelihood function is then given by

$$p(\mathbf{Y}|\mathbf{U},\mathbf{B},\boldsymbol{\sigma}) = \prod_{m,n} \mathcal{N}(Y_{m,n}|\mathbf{c}_m^{\mathrm{T}}\mathbf{U}\mathbf{b}_n, \sigma_m^2), \tag{1}$$

where $\mathbf{c}_m$ is the vector denoting the mth row of $\mathbf{C}$ and $\mathbf{b}_n$ the nth column of $\mathbf{B}$. The likelihood governs that each marker of a sample is reconstructed by the pathways it belongs to ($\mathbf{c}_m$) summed over the samples' abundance of the respective pathways ($\mathbf{Ub}_n$). This reconstruction is visualized in Fig. 1C. In comparison to standard factor analysis, all observations are reconstructed through the pathways. We additionally still maintain lower-dimensional latent variables in $\mathbf{B}$ corresponding to the *loadings* in factor analysis.

We place independent zero-mean Gaussian priors over the parameter matrices to ensure sparsity of the pathway-latent association $\mathbf{U}$ and to identify the relevant latent dimensions in $\mathbf{B}$. For $\mathbf{U}$, we use the identically-sized matrix of prior precisions $\Lambda \in \mathbb{R}_+^{\mathrm{p} \times \mathrm{k}}$ with entries $\Lambda_{\mathrm{p,k}}$ denoting the regularization strength of the entry $U_{p,k}$ akin to automatic relevance determination (ARD). ARD is used commonly for biological latent-variable models to achieve sparsity in a probabilistic framework (Li *et al.* 2002; Tan and Févotte 2012). Its advantage is that we do not have to set a sparsity level or regularization hyperparameter *a priori* but can infer it automatically. The precision of $\mathbf{B}$ is controlled per latent variable with vector $\boldsymbol{\delta} \in \mathbb{R}_+^k$. Mathematically, we have the priors

$$p(\mathbf{U}|\Lambda) = \prod_{\mathrm{p,k}} \mathcal{N}(\mathrm{U_{p,k}}; 0, \Lambda_{\mathrm{p,k}}^{-1}), \;\; \mathrm{p}(\mathbf{B}|\boldsymbol{\delta}) = \prod_{\mathrm{k,n}} \mathcal{N}(\mathrm{B_{k,n}}; 0, \delta_{\mathrm{k}}^{-1}). \tag{2}$$

To infer $\mathbf{U}$ and $\mathbf{B}$, we find their *maximum a-posteriori* (MAP) values by minimizing their negative log joint distribution with the observations,

$$\begin{aligned} \mathcal{L}(\mathbf{U},\mathbf{B}) &= -\log p(\mathbf{Y},\mathbf{U},\mathbf{B}|\boldsymbol{\sigma},\Lambda,\boldsymbol{\delta}) \\ &= -\log p(\mathbf{Y}|\mathbf{U},\mathbf{B},\boldsymbol{\sigma}) - \log p(\mathbf{U}|\Lambda) - \log \mathrm{p}(\mathbf{B}|\boldsymbol{\delta}). \end{aligned} \tag{3}$$

To optimize this objective, we use alternating least-squares (ALS) (Hastie *et al.* 2015), which uses alternating closed-form updates for the parameters. The inference procedure is detailed in Section 2.3.

To select the hyperparameters of the model that control the observation noises $\boldsymbol{\sigma}$, and relevance parameters $\Lambda, \boldsymbol{\delta}$, we use a Bayesian model selection scheme, which, perhaps surprisingly, does not increase the complexity of our algorithm. For example with cross-validation, choosing relevance parameters would be intractable due to the curse of dimensionality as we have $m + pk + k$ hyperparameters. In our experiments, this can be more than 10^4. For our hyperparameter optimization, we use a Laplace approximation to the marginal likelihood that has been successfully used in automatic relevance determination (MacKay 1994) and can

jointly optimize observation noises $\boldsymbol{\sigma}$ and relevance parameters $\Lambda, \boldsymbol{\delta}$. In particular, we maximize the evidence of the hyperparameters,

$$\begin{aligned}\log p(\mathbf{Y}|\boldsymbol{\sigma}, \Lambda, \boldsymbol{\delta}) \approx\ & \log \mathrm{p}(\mathbf{Y}, \mathbf{U}_*, \mathbf{B}_*|\boldsymbol{\sigma}, \Lambda, \boldsymbol{\delta}) \\ -\frac{1}{2}\log|\frac{1}{2\pi}\nabla^2_{\mathrm{U,B}}\mathcal{L}(\mathbf{U}_*, \mathbf{B}_*)|,\end{aligned} \tag{4}$$

where the Hessian is evaluated at MAP estimates $\mathbf{U}_*$ and $\mathbf{B}_*$. We further approximate the Hessian as block-diagonal individually for $\mathbf{U}$ and $\mathbf{B}$, which typically performs as well but has significant computational benefits (Immer *et al.* 2021). The hyperparameters are updated in a closed-form procedure similar to MacKay (1994) and Tipping (2001). Interestingly, the closed-form updates require computation of the same quantities as required for the alternating least-squares updates and thus do not increase overall computational complexity. That is, ARD is for free in our model when we use ALS updates for the parameters.

2.3 Efficient inference of path-FA

We optimize the parameters of the model using *alternating least-squares* (ALS), which guarantees improvement in each step of the algorithm due to closed-form updates. Since both ALS and the Laplace approximation rely on the Hessian, computation is shared making hyperparameter optimization asymptotically free. Our updates resemble a Bayesian ALS that jointly optimizes parameters and hyperparameters by repeated integration as done for neural networks (Immer *et al.* 2021), and might be relevant outside the context of fitting PathFA, for example, in general matrix factorization or factor analysis problems. We first derive the gradients and Hessians of the parameter objective in Equation (3), which then allow to derive the parameter and hyperparameter updates.

2.3.1 Gradients and Hessians

Defining $\mathbf{C}_{\boldsymbol{\sigma}} \stackrel{\text{def}}{=} \boldsymbol{\sigma}^{-2} \circ \mathbf{C}$, where $\circ$ is the pointwise hadamard product, the gradients of the objective are given by

$$\begin{aligned}\mathbf{G}(\mathbf{U}) &\stackrel{\text{def}}{=} \nabla_{\mathbf{U}}\mathcal{L}(\mathbf{U}, \mathbf{B}) = -\mathbf{C}^{\mathrm{T}}_{\boldsymbol{\sigma}}(\mathbf{Y}-\mathbf{CUB})\mathbf{B} + \Lambda \circ \mathbf{U}, \\ \mathbf{G}(\mathbf{B}) &\stackrel{\text{def}}{=} \nabla_{\mathbf{B}}\mathcal{L}(\mathbf{U}, \mathbf{B}) = -\mathbf{U}^{\mathrm{T}}\mathbf{C}^{\mathrm{T}}_{\boldsymbol{\sigma}}(\mathbf{Y}-\mathbf{CUB}) + (\boldsymbol{\delta} \otimes 1_n) \circ \mathbf{B},\end{aligned} \tag{5}$$

where $\otimes$ denotes the Kronecker product and 1_n the n-dimensional vector of 1 s. For alternating least-squares and the marginal likelihood approximation, we further need the Hessians

$$\begin{aligned}\mathbf{H}(\mathbf{U}) &\stackrel{\text{def}}{=} \nabla^2_{\mathrm{vec}(\mathbf{U})}\mathcal{L}(\mathbf{U}, \mathbf{B}) = (\mathbf{C}^{\mathrm{T}}_{\boldsymbol{\sigma}}\mathbf{C}) \otimes (\mathbf{BB}^{\mathrm{T}}) + \mathrm{diag}(vec(\boldsymbol{\Lambda})), \\ \mathbf{H}(\mathbf{B}) &\stackrel{\text{def}}{=} \nabla^2_{\mathrm{vec}(\mathbf{B})}\mathcal{L}(\mathbf{U}, \mathbf{B}) = (\mathbf{U}^{\mathrm{T}}\mathbf{C}^{\mathrm{T}}_{\boldsymbol{\sigma}}\mathbf{CU}) \otimes \mathbf{I}_n + \mathrm{diag}(\boldsymbol{\delta}) \otimes \mathbf{I}_n \\ &= (\mathbf{U}^{\mathrm{T}}\mathbf{C}^{\mathrm{T}}_{\boldsymbol{\sigma}}\mathbf{CU} + \mathrm{diag}(\boldsymbol{\delta})) \otimes \mathbf{I}_n,\end{aligned} \tag{6}$$

where $\mathrm{vec}(\Lambda)$ concatenates the columns of Λ into a vector and $\mathrm{diag}(\cdot)$ turns a vector into a diagonal matrix with elements given by the argument. The Kronecker-factored structure of the Hessian w.r.t. $\mathbf{B}$ allows for efficient computation of inverse and log-determinant required for ALS and marginal likelihood, respectively, by computing these quantities on the individual factors. However, the Hessian w.r.t. $\mathbf{U}$ requires evaluation of the Kronecker product for inversion due to the ARD prior controlled by Λ. This is not a problem because $\mathbf{U}$'s shape is the number of pathways times number of latents, and therefore, small, i.e. below 10^3.

2.3.2 Updates and algorithm

We optimize the model parameters $\mathbf{U}$ and $\mathbf{B}$ efficiently using ALS, or equivalently Newton's method, by preconditioning the gradient with the inverse of the Hessian. The hyperparameters are updated using the same quantities, therefore incurring no additional cost, and use fixed-point updates proposed by MacKay (1992) for neural network regression but we apply these updates already during training, which requires less time to converge (Immer *et al.* 2021). $\mathbf{U}$ is initialized to a zero-matrix and constrained to be non-negative and $\mathbf{B}$ is initialized using the right factor of a k-truncated singular-value decomposition of the observation matrix. The observation noises are initialized to a vector of 1 s and prior precisions logarithmically spaced between 10^{-2} and 10^2 along the k latent variables.

The parameter updates are given by a Newton update with step size γ:

$$\mathbf{M}_{t+1} \leftarrow \mathbf{M}_t - \gamma\mathbf{H}(\mathbf{M}_t)^{-1}\mathrm{vec}(\mathbf{G}(\mathbf{M}_t)), \tag{7}$$

where $\mathbf{M}$ is either parameter $\mathbf{U}$ or $\mathbf{B}$. For $\mathbf{B}$ this further simplifies due to the Kronecker-factored structure of its Hessian:

$$\mathbf{B}_{t+1} \leftarrow \mathbf{B}_t - \gamma(\mathbf{U}_t^{\mathrm{T}}\mathbf{C}^{\mathrm{T}}_{\boldsymbol{\sigma}}\mathbf{CU}_t + \mathrm{diag}(\boldsymbol{\delta}))^{-1}\mathbf{G}(\mathbf{B}), \tag{8}$$

which is efficient to compute since the left-multiplied Kronecker factor of the Hessian is quadratic in latents ($k \times k$) and cheap to invert. To update $\mathbf{U}$, we need to invert a matrix of size $pk \times pk$, which is typically tractable and fast with $pk \approx 10^3$.

The hyperparameter updates rely on the inverse of the Hessian as well and therefore require no overhead computational complexity. The updates are similar to the ones for sparse Bayesian learning (MacKay 1992; Tipping 2001) but applied to our particular case of matrix factorization, which has considerably different priors and likelihoods. In particular, we have for the prior precisions Λ:

$$\Lambda_{\mathrm{p,k}} = \frac{1 - \Lambda_{\mathrm{p,k}}\mathrm{H}(\mathbf{U})^{-1}_{pk,pk}}{\mathbf{U}^2_{p,k}}, \tag{9}$$

which requires the diagonal elements of the inverse of $\mathbf{U}$s Hessian like the ALS update. For the prior precisions of $\mathbf{B}$ and observation noises, we have:

$$\begin{aligned}\delta_k &= \frac{n(1 - \delta_k(\mathbf{U}^{\mathrm{T}}\mathbf{C}^{\mathrm{T}}_{\boldsymbol{\sigma}}\mathbf{CU} + \mathrm{diag}(\boldsymbol{\delta}))^{-1}_{k,k})}{\sum_n \mathrm{B}^2_{k,n}}, \\ \sigma^2_m &= \frac{\sum_n (\mathbf{c}_m\mathbf{U}\mathbf{b}_n - \mathrm{Y}_{m,n})^2}{n - p + \mathrm{tr}(\delta\mathbf{H}^{-1})}.\end{aligned} \tag{10}$$

Computing these updates poses no overhead in computation over the ALS updates and enables a hyperparameter-free algorithm. This is because only ALS requires a step size γ,

which can be simply set to a common default of 0.1, and the hyperparameter updates are entirely in closed-form.

2.4 Multi-omics PathFA

To extend PathFA to simultaneous proteomics and transcriptomics data per sample, we only extend the likelihood but keep the same latent variables $\mathbf{U}$ and $\mathbf{B}$, which construct a shared hierarchical latent space (see Fig. 1). Specifically, we have another view and thus likelihood for the proteomics data with translated pathway prior knowledge $\mathbf{C}_p$. In principle, analyzing both modalities concurrently should enhance the precision of latent variable inference, even with a reduced number of samples. Moreover, the complementary nature of these modalities can potentially lead to more robust and effective latent representations, which may be beneficial for downstream tasks.

2.4.1 Probabilistic model of multi-omics PathFA

Mathematically, we have prior knowledge about the gene-set-to-marker relationships in form of $\mathbf{C}_p \in \{0,1\}^{m_p \times p}$ for m_p protein markers and $\mathbf{C}_r \in \{0,1\}^{m_r \times p}$ for m_r RNA markers, respectively. Importantly, both cover the same pathways, which function as representation space. With proteomics observations $\mathbf{Y}_p \in \mathbb{R}^{m_p \times n}$ and transcriptomics observations $\mathbf{Y}_r \in \mathbb{R}^{m_r \times n}$, we have the likelihood

$$p(\mathbf{Y}_p, \mathbf{Y}_r | \mathbf{U}, \mathbf{B}, \boldsymbol{\sigma}_p, \boldsymbol{\sigma}_r) = p(\mathbf{Y}_p | \mathbf{U}, \mathbf{B}, \boldsymbol{\sigma}_p) \times p(\mathbf{Y}_r | \mathbf{U}, \mathbf{B}, \boldsymbol{\sigma}_r), \quad (11)$$

which is simply the product of two unimodal likelihoods, which are defined in Equation (1). This means, we potentially observe more evidence for latent variables $\mathbf{U}$ and $\mathbf{B}$. We model both observations as conditionally independent and with their own per-marker noise levels denoted by $\boldsymbol{\sigma}_r$ and $\boldsymbol{\sigma}_p$. Inferring noise levels is important to correctly weight the signal and noise coming from either modality. Further, modality-specific scalar factors c_r and c_p are multiplied to reconstruct the respective modalities, which can absorb potential inconsistencies in preprocessing or normalization of transcriptomics and proteomics data, respectively. Both scalars can be updated in a closed form and have no prior. The Gaussian priors with zero mean and precision parameters λ_U, λ_B on $\mathbf{U}$ and $\mathbf{B}$ are as in Equation (2).

2.4.2 Inference for multi-omics PathFA

The alternating least-squares parameter updates for multimodal PathFA and the corresponding hyperparameter updates are only slightly different from the ones of the unimodal variant described above. In particular, the additional observation in the multimodal likelihood [Equation (11)] gives an additional summand in the gradient and Hessian computation, which are given in Equations (5) and (6) for the unimodal case, respectively. This gives us only slightly modified gradients $\mathbf{G}_m(\mathbf{U}), \mathbf{G}_m(\mathbf{B})$ and Hessians $\mathbf{H}_m(\mathbf{U}), \mathbf{H}_m(\mathbf{B})$ with the subscript $_m$ denoting the multimodal model. This does not complicate the computation and the same parameter and hyperparameter updates hold as detailed in Section 2.3.2 by simply replacing the Hessian with the multimodal variants. The final algorithm in pseudo-code is given in Algorithm 1.

2.4.3 Hyperparameters and computational considerations

The overall algorithm can be entirely hyperparameter-free as step sizes γ below 1 lead to convergence using ALS within tens to hundreds of steps and ARD can shrink unnecessary latent variables. In practice, we use $\gamma = 0.1$ and $k = 10$ latents in our experiments. Alternatively, the computed marginal likelihood can be used to select the number of latent variables by running the algorithm for different numbers of latents k as in Bayesian PCA (Bishop 1998). The algorithm scales cubically in the Hessian sizes, which is $\mathcal{O}(k^3)$ for the Hessian of $\mathbf{B}$ due to the Kronecker-factored structure and $\mathcal{O}(p^3k^3)$ for the Hessian of $\mathbf{U}$. The latter can be expensive when the number of latents and pathways is large, in which case a diagonal approximation is still viable and implemented in our code.

Algorithm 1. Probabilistic Multi-Path-FA

Require: observations $\mathbf{Y}_r \in \mathbb{R}^{m_r \times n}, \mathbf{Y}_p \in \mathbb{R}^{m_p \times n}$, pathway knowledge $\mathbf{C}_r \in \{0,1\}^{m_r \times p}, \mathbf{C}_p \in \{0,1\}^{m_p \times p}$, $N > 0$ iterations, $\gamma > 0$ step size, $k > 0$ number of latents

$\mathbf{U} \leftarrow 0 \in \mathbb{R}_+^{p \times k}$; $\mathbf{B} \leftarrow \texttt{truncSVD}(\mathbf{Y}_r, k)$ ▷ init. params

$\mathbf{r} \leftarrow \texttt{logspace}(-4, 4, k)$; $c_r \leftarrow 1$; $c_p \leftarrow 1$

$\boldsymbol{\delta} \leftarrow \mathbf{r} \in \mathbb{R}_+^k$; $\Lambda \leftarrow 1_p \otimes \mathbf{r} \in \mathbb{R}_+^{p \times k}$ ▷ init. prior precision

$\boldsymbol{\sigma}_r \leftarrow 1 \in \mathbb{R}_+^{m_r}$; $\boldsymbol{\sigma}_p \leftarrow 1 \in \mathbb{R}_+^{m_p}$ ▷ init. obs. noise

for $n \in [1, \ldots, N]$ **do**

  if n is even **then** ▷ Alternating least-squares

    Compute gradient and Hessian $\mathbf{G}_m(\mathbf{B}), \mathbf{H}_m(\mathbf{B})$

    $\mathbf{B} \leftarrow \mathbf{B} - \gamma \mathbf{H}_m(\mathbf{B})^{-1} \text{vec}(\mathbf{G}_m(\mathbf{B}))$

  else

    Compute gradient and Hessian $\mathbf{G}_m(\mathbf{U}), \mathbf{H}_m(\mathbf{U})$

    $\mathbf{U} \leftarrow \mathbf{U} - \gamma \mathbf{H}_m(\mathbf{U})^{-1} \text{vec}(\mathbf{G}_m(\mathbf{U}))$

    Project $\mathbf{U}$ to non-negative by clamping to zero

Update $\Lambda, \boldsymbol{\delta}, \boldsymbol{\sigma}_r, \boldsymbol{\sigma}_p$ as in Equations (9) and (10) using $\mathbf{H}_m(\mathbf{U}), \mathbf{H}_m(\mathbf{B})$

Closed-form update for scalar factors c_r, c_p

3 Results

We developed and implemented a novel model for factor analysis, PathFA that leverages pathway information for multimodal data integration and incorporates a probabilistic framework with automatic hyperparameter optimization. It bridges the gap between two existing approaches, MOFA and PLIER. PathFA enables pathway-based interpretation of multimodal factor analysis, specifically suited for small cohorts.

In the sections that follow, we demonstrate how the inclusion of prior information enhances the reconstruction log-likelihood. We observe improvements in the reconstruction of each individual modality and in the convergence behavior when a second data modality is incorporated. Additionally, we showcase the practical utility of our approach through its application to a subset of patient samples from the Swiss Tumor Profiler cohort in a real-world setting.

3.1 Including prior information improves reconstruction

On the synthetic data generated using MSigDB Hallmark as described in Section 2.1.3, we observe that PathFA has clear advantages over PLIER and MOFA in terms of reconstruction performance, sample efficiency, and inferring observation noises on the markers of both proteomics and

transcriptomics markers. We use the public implementations and propose default parameters of PLIER and MOFA and set the number of latent variables for all methods to the ground truth 8.

With access to the ground truth parameters, we can measure the reconstruction performance on test data and fitting of heteroscedastic marker-level observation noises. That is, we use the method in question on a set of m samples to infer the model and then fix the model before inferring loadings for a held-out set of m_{test} samples, $\mathbf{Y}_r^{(\text{test})}$ and $\mathbf{Y}_p^{(\text{test})}$. With inferred loadings, we reconstruct the test data and obtain the estimates $\hat{\mathbf{Y}}_r^{(\text{test})}$ and $\hat{\mathbf{Y}}_p^{(\text{test})}$. The reconstruction log-likelihood,

$$\log p(\hat{\mathbf{Y}}_p^{(\text{test})} | \mathbf{C}, \mathbf{U}, \mathbf{B}, \boldsymbol{\sigma}) = \log \mathcal{N}(\hat{\mathbf{Y}}_p^{(\text{test})}; \mathbf{CUB}, \boldsymbol{\sigma}\mathbf{I}), \qquad (12)$$

denotes an entry-wise Gaussian distribution on the matrices and holds for either proteomics or transcriptomics observation and reconstruction. This allows to assess how well the fitted model can infer representations from noisy observations. This is only possible to measure in a simulation. We have an upper bound on the performance with the ground truth reconstruction $\hat{\mathbf{Y}}_x^{(\text{test})} = \mathbf{C}_x\mathbf{UB}$ for any data modality x. With these parameters, we can reconstruct the noise-free observation, $\hat{\mathbf{Y}}_r$ and $\hat{\mathbf{Y}}_p$, and assess this reconstruction under the true model in terms of log-likelihood. Further, we assess how well our probabilistic model can recover the true observation noise of the model in comparison to MOFA, which does not use prior knowledge $\mathbf{C}$.

We compare the reconstruction log-likelihood across different models using one modality, synthetic proteomics, or synthetic RNA, in Fig. 2. The trends of the reconstruction log-likelihood across different sample sizes are consistent between both modalities. MOFA without pathway information shows the lowest reconstruction log-likelihood. PathFA further improves over PLIER, especially for small sample sizes, due to automatic hyperparameter optimization and focus on pathway-level factor analysis. Figure 3 further shows that using both omics improves reconstruction of them individually, especially for proteomics, which appears to provide less evidence for the loadings. Already around $m = 30$ samples suffice for the loadings to converge with PathFA. Like MOFA, PathFA can infer marker-level heteroscedastic observation noises and converges faster in terms of the samples needed due to the prior information on pathways, as apparent in Fig. 4.

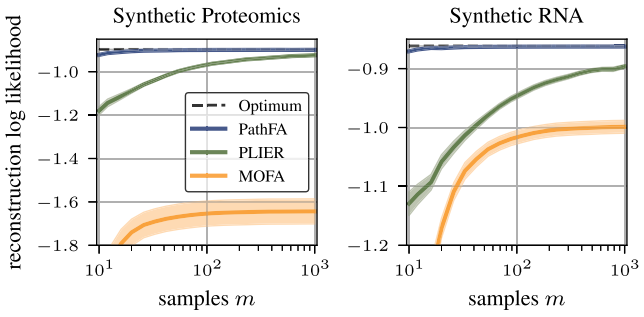

Figure 2. Reconstruction log-likelihood on the synthetic benchmark data as a function of samples for proteomics and RNA averaged over 20 runs. Shaded regions denote twice the standard error about the mean. Unimodal PathFA, unimodal PLIER, and MOFA are compared to optimal attainable performance (dotted line).

3.2 PathFA performs well in small-sample cohorts

For the purpose of investigating how multimodality contribute towards the performance of PathFA, we compare the change in reconstruction likelihood over sample number between PathFA in a unimodal setting with synthetic tranc-scriptomics or proteomics data only (blue) versus PathFA with multimodal data (purple) in Fig. 3. We then use the unimodal or multimodal model, to evaluate the reconstruction log-likelihood and the observation noise, respectively, on transcriptomics and proteomics data separately.

Including both modalities strictly improves the reconstruction performance, especially for small sample sizes, which are common in realistic biomedical datasets, as also in our case in Section 3.3.

To show the effect of prior information, we also compare PathFA against multi-omics factor-analysis (MOFA). Unlike MOFA, we integrate modalities on a pathway level. This provides not only prior information but also enables a pathway-based multimodal factor analysis that should simplify the interpretation of the results. Generally, PathFA demonstrates similar behaviour to MOFA with respect to fitting observation noise. However, in the range of small sample numbers, the utility of prior information takes effect leading to much smaller observation noise error in PathFA.

3.3 Comparing pathway loadings to cell type composition

To demonstrate the utility of our approach, we also benchmark PathFA on a subset of patient samples from the TuPro cohort. This cohort is well suited for a benchmark study as it provides paired bulk proteomics and transcriptomics measurements for each sample. In addition, CyTOF data, also

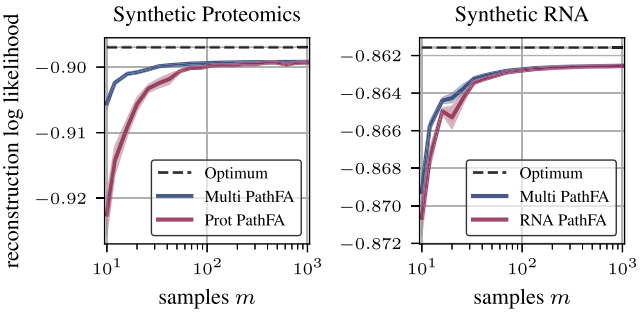

Figure 3. Reconstruction log-likelihood of multimodal PathFA in comparison to the unimodal variant as depicted in Fig. 2.

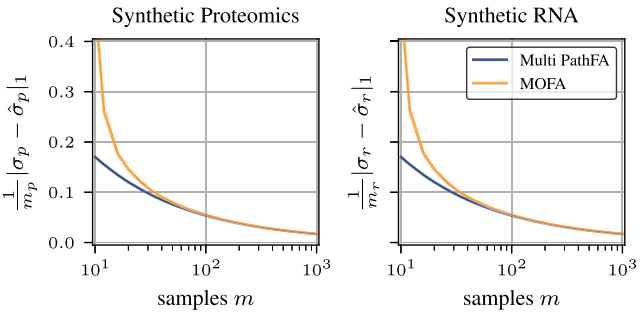

Figure 4. Average absolute error of marker-level observation noises on synthetic benchmark of Multi-PathFA and MOFA for both proteomics and RNA. The lines show the average over 20 runs and shaded regions two standard errors.

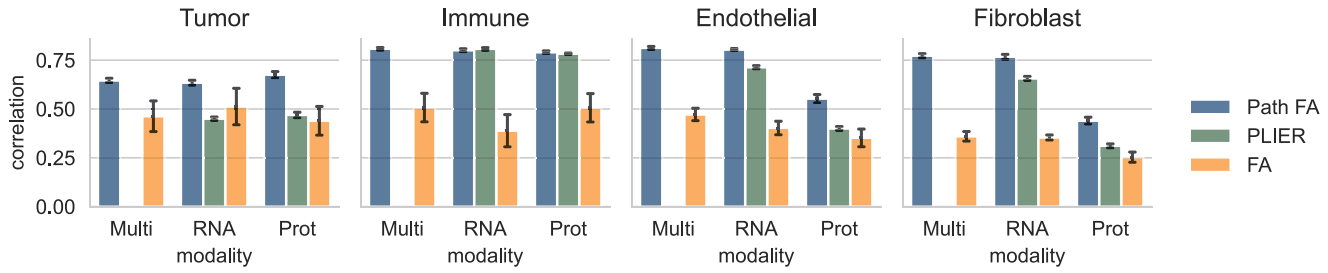

Figure 5. This figure shows the Pearson correlation coefficients of pathway loadings with cell-type content (based on ground truth computed from CyTOF data) across 42 ovarian cancer samples for the four most common cell-types. MSigDB cell-type pathways were used for PathFA across all experiments. Multirefers to the multimodal setting, while RNA and Prot refers towards results based on transcriptomic and proteomic data only. FA refers to factor analysis respectively MOFA in the multisetting. For both, correlation with latent factors is computed. PLIER can only be used in unimodal setting.

generated for all samples, provides single-cell measurements of several protein markers that we able to use as an independent estimate of the ground-truth cell-type composition. We use a subset of 42 samples from ovarian cancer patients as well as 34 samples from the melanoma cancer patients that have available CyTOF compositions, clinical information of the melanoma cohort, and tumor heterogeneity scores of the ovarian cancer cohort.

Here, we train PathFA using the MSigDB Hallmark pathways and evaluate the quality of learned representations by their ability to correlate with cell-type composition. Figure 5 shows the highest correlating pathway loadings from PathFA and baseline approaches with CyTOF estimates of cell-type content in a unimodal and multimodal setting. In the unimodal and multimodal setting, PathFA consistently outperforms standard factor analysis (respectively, MOFA in the multimodal setting), for which we use standard instead of pathway loadings, as well as PLIER with respect to deriving factors that are highly correlated with cell-type content. These results are similar with the corresponding analysis in the melanoma cancer cohort (see supplemental Fig. S1), however, with systematically lower correlation in endothelial and fibroblasts across all approaches. The gain in the multimodal setting over the unimodal approach is small with respect to this evaluation. While the multimodal approach does improve over a proteomic-only approach, the performance over an RNA-only approach is similar. The difference in marker resolution between transcriptomics and proteomics data, may explain this trend.

Next, we identify pathways associated with cell-type composition by computing correlations between the CyTOF estimate and pathway loadings. Thus, we run PathFA using cell-type marking pathways and report the top five highest correlated pathways to each cell type in supplemental Fig. S3 in the Melanoma case. These pathways seem indicative of their respective cell types, melanoma pathway for tumor cell type, T cell pathways for immune cell type, Cancer associated fibroblasts (CAFs) pathways for fibroblast cell type, and endothelial melanoma pathways for endothelial cell types. We observe similar results in the Ovarian cancer cohort (see supplemental Fig. S2).

We further test the efficacy of PathFA when the pathway masks are perturbed and thus the prior information might not be fully accurate. In supplemental Fig. S6, we find that the average cell-type correlation over all four types remains high even for a substanial amount of flipped pathway-marker associations. The same figure shows that on the synthetic data, this is not necessarily the case and the reconstruction log likelihood suffers from flipping associations.

3.4 Associating pathway loadings to patient survival in a melanoma cohort

Conceptually, PathFA is designed to enable easier interpretation of the results of a multimodal factor analysis on the level of pathways. To investigate this further, we look at survival data as well as time to progression of the melanoma patient cohort. We analyzed the association of the pathway loadings of PathFA in the multimodal setting (transcriptomics and proteomics) with survival and patient progression. Each patient is given a binary label for four categories, if the patient has survived for at least 6 and 12 months, and if the patient has survive without progression for 6 and 12 months. Then, a *t*-test is performed between each label and the pathway loadings. A clustermap over the *t*-test statistics for associated pathway loadings are shown in Fig. 6

We observe three major clusters in the melanoma gene set analysis. Cluster 1, top 5 pathways, consists out of upgregulated gene sets associated with proliferation and DNA damage, while cluster 2, middle 12 pathways, encompasses various process categories, such as pathways, metabolism, and development. The remaining cluster 3, bottom 6 pathways, shows enrichment in downregulated gene sets, primarily related to immune responses, DNA damage, and pathways. Furthermore, in line with current literature reporting downregulated gene sets associated with worse clinical outcomes (Garg *et al.* 2021), hallmark gene sets like hallmark_allograft_rejection and hallmark_interferon_gamma_response are enriched in the group of patients with poor outcomes, including those who either succumbed to the disease or progressed within 12 months (see Fig. 6). When compared to a single-modal setting using transcriptomics and proteomics data separately, it becomes apparent that most of the observed signal originates from the transcriptomics data, with two pathways being uniquely identified through proteomics signal (see supplemental Fig. S4).

3.5 Pathway loadings are associated with tumor heterogeneity in ovarian cancer

We proceed to evaluate the performance of PathFA in the ovarian cancer cohort, focusing on the challenge of tumor heterogeneity (Vázquez-García *et al.* 2022). The tumor heterogeneity is computed based on the Jenson-Shannon divergence from CyTOF data on tumor cells. Similar to the previous tasks, PathFA achieves the highest correlation between latent factors and tumor heterogeneity (see supplemental Fig. S5). When comparing heterogeneity estimates derived from tumor cell populations using CyTOF, we observe significant correlations. Notably, OXIDATIVE_PHOSPHORYLATION and UN FOLDED_PROTEIN_RESPONSE shows positive correlations (see Table 1) including mTORC1 (see Fig. 7). This suggests

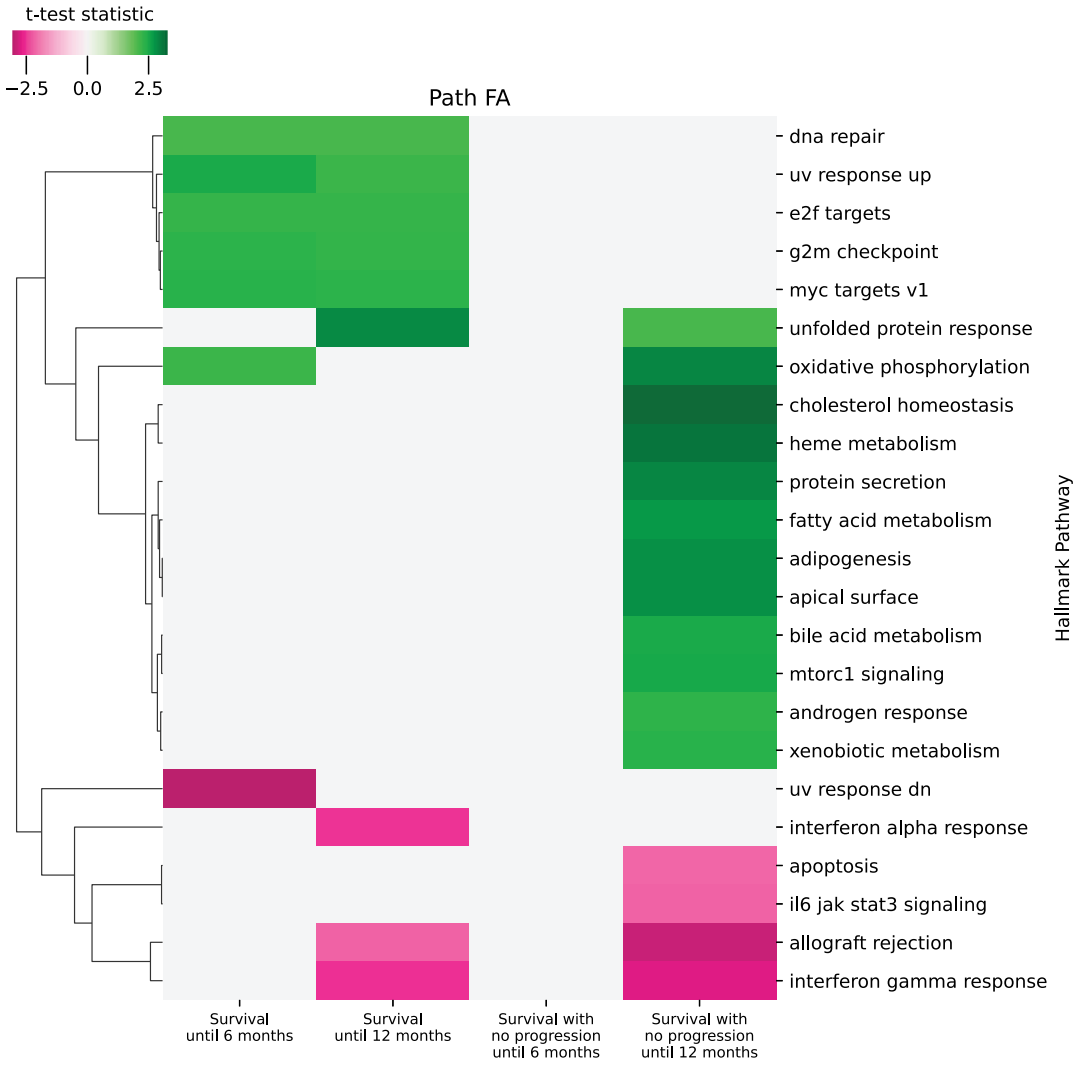

Figure 6. Significant associations between survival and progression in relationship to pathway loadings from PathFA. The color refers to the normalized mean difference between the two patient groups in each column, for significant associations only. The clustering tree implies the presence of three clusters, two of which seem to be higher expressed [top (5 pathways) and middle (12 pathways)], and the remaining one appears to be down-regulated in reference to the according patient group (*y*-axis).

Table 1. This is a list of 10 MSigDB Hallmark pathways that have the highest Pearson correlation with tumor heterogeneity score on 42 ovarian cancer samples.

Rank	MSigDB Hallmark pathway	Correlation	*P*-value
1	MTORC1_SIGNALING	0.41	.01
2	GLYCOLYSIS	0.34	.03
3	G2M_CHECKPOINT	0.32	.04
4	E2F_TARGETS	0.31	.04
5	MYC_TARGETS_V1	0.29	.06
6	OXIDATIVE_PHOSPHORYLATION	0.29	.06
7	UNFOLDED_PROTEIN_RESPONSE	0.24	.12
8	DNA_REPAIR	0.23	.14
9	MYC_TARGETS_V2	0.23	.15
10	PEROXISOME	0.22	.16

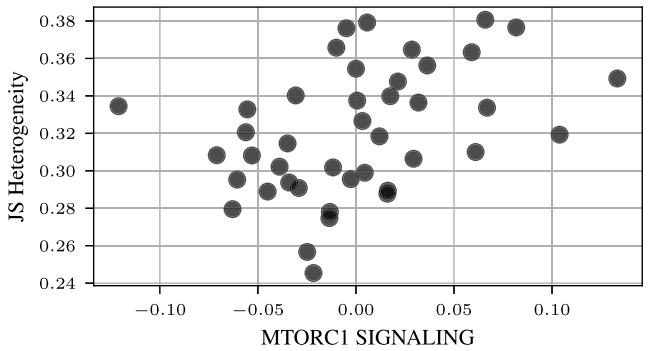

Figure 7. Scatter plot of Jensen–Shannon tumor heterogeneity computed on CyTOF measurements of tumor cells (*y*-axis) and mTORC1 pathway loading generated by PathFA. Each dot represents a ovarian cancer patient sample. The Pearson's correlation is 0.41 (*P*-value = .0075).

associations between endoplasmic reticulum stress, metabolic processes and tumor heterogeneity in ovarian cancer and is consistent with previous reports on pro-survival mechanisms (Madden *et al.* 2019) and metabolic processes (Sancho *et al.* 2016).

4 Discussion and conclusion

In this study, we introduced PathFA, a novel multimodal factor analysis approach tailored for genomic data, employing Bayesian hyperparameter optimization to seamlessly

integrate transcriptomic and proteomic data using pathway sets. This method innovatively addresses heteroscedasticity in markers and modalities, a prevalent challenge in experimental measurements. Through various experiments, we established that incorporating known prior information not only enhances reconstruction quality but also improves data efficiency through Bayesian hyperparameter updates. Our probabilistic formulation, especially the automatic optimization of hyperparameters, marks an improvement over PLIER, which lacks this adaptability.

A notable advancement of PathFA is its ability to provide immediate interpretability of pathway activities, in contrast to traditional approaches that often rely on post-hoc gene set enrichment analysis. This immediate interpretability is exemplified in our analysis of a cancer patient cohort from the Tumor Profiler Project, where PathFA successfully correlates pathway activities with crucial biological aspects such as cell-type composition, survival, and tumor heterogeneity. These capabilities illustrate the potential of PathFA in elucidating complex biological relationships and hypothesis generation.

PathFA is implemented using a Gaussian likelihood. This provides a robust foundation for different data modalities. In the future, it may be desirable to expand this framework to allow other likelihood functions such as the negative binomial, often used in context of read count data. PathFA may also expand across other biological data-types for which groupings (e.g. pathways masks) of individual markers (e.g. genes) is possible across the modalities. Furthermore, our approach extends beyond the typical limitations of analyzing dual modalities. It is readily adaptable to include additional data types, promising broader applicability as genomic profiling technologies evolve. A potential complication of such multimodal datasets lie in the differences in cellular composition between parallel tissues slices. Looking forward, we envision enhancements to PathFA that incorporate modality-specific prior information that are in addition flexible to these tissues effects, further augmenting its analytical power.

In summary, PathFA represents a significant step forward in multimodal genomic analysis, offering robustness, flexibility, and direct interpretability. Its development aligns with the ongoing advancements in profiling technologies, positioning it as a valuable tool in understanding complex biological systems and disease mechanisms.

Author contributions

Alexander Immer, Stefan G. Stark, Francis Jacob, Gunnar Rätsch, and Kjong-Van Lehmann (Wrote the Manuscript), Kjong-Van Lehmann, Alexander Immer, Stefan G. Stark and Gunnar Rätsch (Conceptualized the Idea), Alexander Immer (Developed the Approach and Simulation), Stefan G. Stark (Prepared and Analyzed the Tumor Profiler Data), Francis Jacob (Provided Biological Interpretation of the Results), Sandra Goetze, Emanuela S. Milani and Bernd Wollscheid (Provided the Proteomics Data), and Ximena Bonilla, Tinu Thomas and André Kahles (Processed the RNA-seq Data). All authors read and approved the manuscript

Supplementary data

Supplementary data are available at *Bioinformatics* online.

Data availability

Data will be made available with the upcoming Tumor Profiler Marker Publications. Details will be made available under http://tu-pro.ch/download/pathFA/.

Conflict of interest

None declared.

Funding

This work was supported by ETH core funding to G.R. (funding A.I., S.G.S., K.V.L., X.B., A.K.). A.I. is partially funded by the Max Planck ETH Center for Learning Systems. S.G.S. and K.V.L. were also funded by the Tumor Profiler program. K.V.L. was also supported by the funding programme Cancer Center Cologne Essen of the Ministry of Culture and Science of the State of North Rhine-Westphalia. S.G. and B.W. were supported by funding through the Personalized Health and Related Technologies (PHRT) strategic focus area of ETH.

References

Argelaguet R, Velten B, Arnol D *et al.* Multi-omics factor analysis—a framework for unsupervised integration of multi-omics data sets. *Mol Syst Biol* 2018;**14**:e8124.

Bishop C. Bayesian PCA. *Adv Neural Inform Process Syst* 1998;11.

Boehm KM, Khosravi P, Vanguri R *et al.* Harnessing multimodal data integration to advance precision oncology. *Nat Rev Cancer* 2022; **22**:114–26.

Chen B, Khodadoust MS, Liu CL *et al.* Profiling tumor infiltrating immune cells with cibersort. In: *Cancer Systems Biology: Methods and Protocols*, 2018, 243–59.

Consortium, U Uniprot: a worldwide hub of protein knowledge. *Nucleic Acids Res* 2019;**47**:D506–15.

Garg M, Couturier D-L, Nsengimana J *et al.* Tumour gene expression signature in primary melanoma predicts long-term outcomes. *Nat Commun* 2021;**12**:1137.

Hastie T, Mazumder R, Lee JD *et al.* Matrix completion and low-rank svd via fast alternating least squares. *J Mach Learn Res* 2015; **16**:3367–402.

Immer A, Bauer M, Fortuin V *et al.* Scalable marginal likelihood estimation for model selection in deep learning. In: *International Conference on Machine Learning*, p. 4563–4573. PMLR, 2021.

Irmisch A, Bonilla X, Chevrier S *et al.* The tumor profiler study: integrated, multi-omic, functional tumor profiling for clinical decision support. *Cancer Cell* 2021;**39**:288–93.

Li Y, Campbell C, Tipping M *et al.* Bayesian automatic relevance determination algorithms for classifying gene expression data. *Bioinformatics* 2002;**18**:1332–9.

Liberzon A, Birger C, Thorvaldsdóttir H *et al.* The molecular signatures database hallmark gene set collection. *Cell Syst* 2015;**1**:417–25.

Liberzon A, Subramanian A, Pinchback R *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* 2011;**27**:1739–40.

MacKay DJ. Bayesian interpolation. *Neural Comput* 1992;**4**:415–47.

MacKay DJC. Bayesian nonlinear modeling for the prediction competition. *ASHRAE Trans* 1994;**100**:1053–62.

Madden E, Logue SE, Healy SJ *et al.* The role of the unfolded protein response in cancer progression: from oncogenesis to chemoresistance. *Biol Cell* 2019;**111**:1–17.

Mao W, Zaslavsky E, Hartmann BM *et al.* Pathway-level information extractor (PLIER) for gene expression data. *Nat Methods* 2019; **16**:607–10.

Ritchie MD, Holzinger ER, Li R *et al.* Methods of integrating data to uncover genotype–phenotype interactions. *Nat Rev Genet* 2015; **16**:85–97.

Ruffier M, Kähäri A, Komorowska M *et al.* Ensembl core software resources: storage and programmatic access for DNA sequence and genome annotation. *Database* 2017;**2017**:bax020.

Sancho P, Barneda D, Heeschen C *et al.* Hallmarks of cancer stem cell metabolism. *Br J Cancer* 2016;**114**:1305–12.

Tan VY, Févotte C. Automatic relevance determination in nonnegative matrix factorization with the/spl beta/-divergence. *IEEE Trans Pattern Anal Mach Intell* 2012;**35**:1592–605.

Taroni JN, Grayson PC, Hu Q *et al.* Multiplier: a transfer learning framework for transcriptomics reveals systemic features of rare disease. *Cell Syst* 2019;**8**:380–94.e4.

The Tumor Profiler Consortium. Multimodal profiling identifies chemotherapy-associated phenotypic divergence in high-grade serous ovarian cancer. *In preparation* 2024a.

The Tumor Profiler Consortium. Single-cell molecular and functional landscapes of metastatic melanoma converge on clinically actionable features. *In preparation* 2024b.

Tipping ME. Sparse bayesian learning and the relevance vector machine. *J Mach Learn Res* 2001;**1**:211–44.

Vázquez-García I, Uhlitz F, Ceglia N *et al.* Ovarian cancer mutational processes drive site-specific immune evasion. *Nature* 2022; **612**:778–86.

Xuan Y, Bateman NW, Gallien S *et al.* Standardization and harmonization of distributed multi-center proteotype analysis supporting precision medicine studies. *Nat Commun* 2020;**11**:5248.

Bioinformatics, 2024, 40, 189–198
https://doi.org/10.1093/bioinformatics/btae216
ISMB 2024