MULTI-MODAL DISENTANGLEMENT OF SPATIAL TRANSCRIPTOMICS AND HISTOPATHOLOGY IMAGING

Hassaan Maan[1,2,3*], **Zongliang Ji**[2,4], **Elliot Sicheri**[6], **Tiak Ju Tan**[5,6], **Alina Selega**[2,6], **Ricardo Gonzalez**[6], **Rahul G. Krishnan** [2,4,9], **Bo Wang** [1,2,3,4,9,10*†], **Kieran R. Campbell** [2,4,5,6,7,8*†]

1. Peter Munk Cardiac Center, University Health Network, Toronto, ON, Canada
2. Vector Institute, Toronto, ON, Canada
3. Department of Medical Biophysics, University of Toronto, Toronto, ON, Canada
4. Department of Computer Science, University of Toronto, Toronto, ON, Canada
5. Department of Molecular Genetics, University of Toronto, Toronto, ON, Canada
6. Lunenfeld-Tanenbaum Research Institute, Toronto, ON, Canada
7. Department of Statistical Sciences, University of Toronto, Toronto, ON, Canada
8. Ontario Institute for Cancer Research, Toronto, ON, Canada
9. Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, ON, Canada
10. University Health Network AI Hub, Toronto, ON, Canada

* Corresponding authors
† Supervised this work equally

ABSTRACT

Spatially-resolved expression profiling data has revolutionized biological research with multiple emerging clinical applications. Spatial transcriptomic assays are often jointly measured with histopathology imaging data, which is frequently used for diagnosing and staging various diseases. However, determining the extent to which the spatial transcriptomic and histopathology data represent overlapping or unique sources of variation is challenging, particularly given the myriad of factors influencing both, including expression variation, spatial context, tissue morphology, and batch effects. Here, we view this challenge as multi-modal disentanglement and develop an evaluation framework. We introduce SpatialDIVA, a disentanglement technique for jointly measured spatially resolved transcriptomics and histopathology data. We demonstrate that SpatialDIVA outperforms baseline techniques in disentangling salient factors of variation in curated pathologist-annotated multi-sample colorectal and pancreatic cancer cohorts. Further, SpatialDIVA removes batch effects from multi-modal data, allows for factor covariance analysis, and yields actionable biological insights through a novel conditional multi-modal generation method. The SpatialDIVA model, evaluation code, and datasets are available at `https://github.com/hsmaan/SpatialDIVA`.

1 INTRODUCTION

Spatial context is an important measurement in the study of biological systems, as it dictates the organization of tissues, the flow of information in the form of cell to cell communication, transport of biomolecules and nutrients, and many other factors (Rao et al., 2021; Tian et al., 2023). To incorporate spatial context in molecular measurements, researchers have developed spatial transcriptomics (ST) technologies, which quantify mRNA expression in small numbers of cells (1-10) at specific locations in a tissue, commonly referred to as *spots* (Moses & Pachter, 2022). Barcoding then allows for identifying the spatial position of captured RNA molecules (Moses & Pachter, 2022). After determining cell-type identity through the transcriptomic information, the spatial context can be used to perform additional analysis, such as annotating spatial domains and determining patterns of cell to cell communication (Rao et al., 2021).

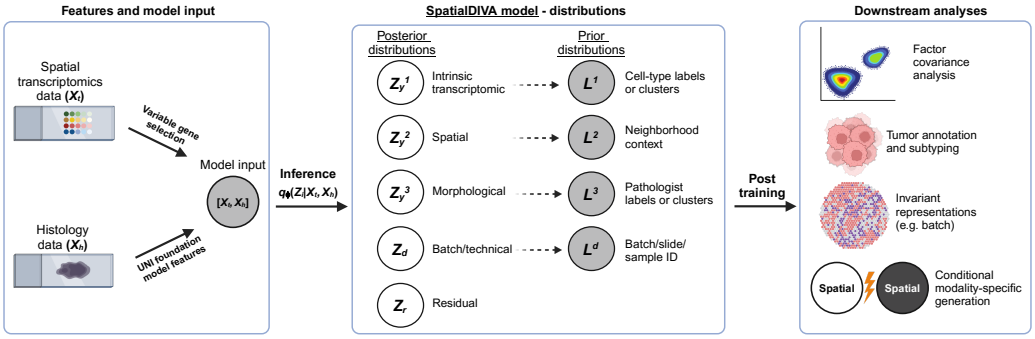


Figure 1: **Overview of contributions and the *SpatialDIVA* model.** Our approach is the first to frame the challenge of evaluating modality and factor-specific contributions as a multi-modal disentanglement problem, for which we present the SpatialDIVA model as a solution. SpatialDIVA allows for biologically and clinically relevant downstream analyses, such as factor covariance, tumor annotation in cancer biopsy samples, removing batch effects, and determining modality-specific information through conditional generation.

In addition to quantifying spatially-resolved gene expression, spot-based ST assays such as 10X Visium (Ståhl et al., 2016) are able to simultaneously image a hematoxylin and eosin (H&E) stain (Fischer et al., 2008; Janesick et al., 2023) of the tissue section under study for histological analysis. H&E staining of tissue and examination of associated morphological features is routinely used for diagnosing and staging many malignancies (He et al., 2012). The addition to H&E staining to ST technologies allows researchers to associate morphological features with RNA expression of different genes in the underlying tissue, as well as predict gene expression from H&E (Janesick et al., 2023).

Although there is a myriad of existing approaches to predict RNA expression from histopathology images (Appendix A), these are limited, as histopathology data is often only weakly correlated with RNA expression (Zeng et al., 2022; Xie et al., 2023). Further, this correlation is highly variable across genes (Zeng et al., 2022; Xie et al., 2023). One way to interpret these methods is that they are approximating the upper bound on mutual information between the two modalities. However, this upper bound does not answer the question of *which information* is present or absent in each modality, and this is more biologically relevant. For example, knowing which variation is exclusively found in RNA expression and not in histology data allows us to determine which factors relevant to disease states are being missed through only performing histology analysis, and vice versa.

Many other factors also contribute to modality-specific variation, such as spatial context, and nuisance factors including batch effects. Delineating the contribution of these factors allows us to obtain an interpretable model of the multi-modal data generating distribution, infer which gene-programs and morphological patterns are dictated by specific factors, and remove the effects of nuisance variation.

In this work, we determine the precise effects of different biological and technical factors in multi-modal ST and histology data. Our contributions are as follows:

- In Sec. 2.1-2.3, we frame the challenge of determining the contribution of distinct biological factors to multi-modal assays as a multi-modal disentanglement problem. We propose a general framework that incorporates biological knowledge (Appendix B.1), based on prior work that demonstrates fully unsupervised approaches to disentanglement are suboptimal (Locatello et al., 2018).
- We introduce *SpatialDIVA* (Fig. 1) in Sec. 3, which builds upon previous work in disentangled representation learning (Ilse et al., 2019) by introducing prior-constrained multi-modal disentanglement in a ST and histopathology setting, with continuous label distributions.
- We show that SpatialDIVA outperforms previous techniques and exhibits strong disentanglement of colorectal and pancreatic cancer data (Sec. 4.1 and 4.2). To benchmark disentanglement in this setting, we develop a framework incorporating several metrics and two multi-patient/sample pathologist-annotated datasets, which can be used by the community to further our contributions (Appendices D and I).
- To determine modality-specific effects of latent factors, we develop a conditional generation framework for multi-modal data using the SpatialDIVA model (Fig. 1), and use this

framework to infer modality-specific gene regulatory programs in a multi-patient pancreatic cancer cohort (Sec. 4.3 and 4.4).

2 BACKGROUND AND PROBLEM FORMULATION

2.1 ST AND HISTOLOGY DATA

Given a slide with a slice of tissue and jointly measured ST and histology (Ståhl et al., 2016), we obtain the following data representation per spot i on the slide:

$$S_{(i)} : \{X_{t(i)}, X_{h(i)}, L_{(i)}^{1..m}, L_{(i)}^{d}, P_{x,y-(i)}\} \tag{1}$$

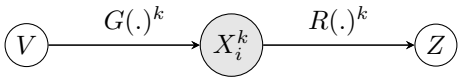
Where X_t indicates the transcriptomic counts across genes measured in the spot, X_h indicates the histopathology image distribution, $L^{1..m}$ indicates any m label groups for the spot (such as pathologist annotations and cell-type labels), L^d indicates a label for the specific slide or tissue section (e.g. slide 1, slide 2, ...), and $P_{x,y}$ are the spatial coordinates of that spot on the slide.

The distribution of expression of each gene j per spot i is given by $g_{i,j} \sim \text{Poisson}(\lambda_{i,j})$. Log-normalization of this count data can be performed for modelling (Hao et al., 2024), or the untransformed counts can be used directly (Zhao et al., 2022) (Appendices E and F). The distribution of the histopathology data follows a 3-channel RGB image, and histopathology foundation models can be used to extract features for this data (Chen et al., 2024) (Appendix E).

2.2 LATENT VARIABLE MODEL OF ST AND HISTOLOGY

The underlying biology that drives the variation of ST and the paired histology data is often shared. However, the different biological and technical factors, and the extent to which they contribute to each modality, is unresolved.

Assume a set of k generative factors $V = \{v_1, v_2, ..., v_k\}$ account for the data distributions of an arbitrary number of views (X_i^k) through an arbitrary number of generative processes $G(.)^k$:



The underlying generative processes $G(.)^k$, as well as the generative factors V, are unobservable. For an arbitrary number of observed views, our goal is to approximate the generative factors jointly for all views through processes $R(.)^k$ and infer m latent variables ($Z = z_1, z_2, \ldots, z_m$):

$$R(.)^k \approx p_\theta(Z|X_1, X_2, \ldots, X_n) \tag{2}$$

In the context of spatial transcriptomics, we know that both ST ($\mathbf{X_t}$) and histology ($\mathbf{X_h}$) are generated by overlapping biological factors, which we can approximate through m latent variables and the $R(.)^t$ and $R(.)^h$ processes :

$$\underset{q \in Q}{\text{argmin}}\, D_{KL}(q(\mathbf{Z}) \parallel p_\theta(\mathbf{Z}|X_t, X_h)) \tag{3}$$

Note that any number of views can be considered in this framework, but we restrict it to two based on our joint ST and histology setting. We aim to learn parameterizations for the functions $R(.)^t$ and $R(.)^h$. Assuming that these functions are parameterized by ψ_1 and ψ_2, collectively described as ψ, the objective becomes:

$$\psi^* = \underset{\psi}{\text{argmin}}\, D_{KL}(q(\mathbf{Z}|\psi) \parallel p_\theta(\mathbf{Z}|X_t, X_h)) \tag{4}$$

2.3 DISENTANGLING EXPLANATORY FACTORS FOR ST AND HISTOLOGY

Given that we want to learn m explanatory factors for both the ST ($\mathbf{X_t}$) and histology ($\mathbf{X_h}$) data that best approximate the underlying generative distribution corresponding to distinct biological processes, we aim to learn *disentangled* representations of the approximated latent distribution Z. In general, we want to minimize the total correlation between each learned latent covariate:

$$\text{corr}(Z_1, Z_2, \ldots, Z_m) = D_{KL}((p(\mathbf{Z}) \parallel \prod_k^m p_k(Z_k)) \tag{5}$$

The other constraint in our setting is that we want the learned latent factors to correspond to both known and novel biological sources of variation.

It has been shown by Locatello et al. (2018) that unsupervised approaches to learning disentangled representations, such as through the Beta variational autoencoder (β-VAE) model (Higgins et al., 2016), are under-defined, and entangled representations can lead to the same marginal distributions as disentangled representations in this setting. This renders unsupervised identification of non-linear latent variables difficult. Therefore, we incorporate relevant prior biological knowledge (Appendix B.1). Further, we add a residual latent factor that accounts for any variation not captured by prior-constrained factors, similar to previous work (Ilse et al., 2019).

3 THE SPATIALDIVA MODEL

Given the problem description outlined in Sec. 2, we introduce a generative framework for disentangling ST and histology data: the Spatial Domain Invariant Variational Autoencoder (SpatialDIVA) model (Fig. 1 and 2). SpatialDIVA is a deep latent variable model, that aims to infer latent distributions for m biological covariates (Z_y^k), batch effects (Z_d), and residual variation (Z_r), through maximizing the marginal likelihood of the histology (X_h) and ST (X_t) data, as well as prior knowledge (L):

$$p_\theta(X_t, X_h, L) = \int p_\theta(X_t|Z)p_\theta(X_h|Z)p_\theta(Z|L)p(L)dz \tag{6}$$

However, this marginalization is intractable, and therefore we learn a lower bound on the log-likelihood (Appendix G):

$$\begin{aligned}\log p_\theta(X_t, X_h, L) \geq \mathbb{E}_{q(Z|X_t,X_h)}\left[\log p(X_t|Z)\right] + \mathbb{E}_{q(Z|X_t,X_h)}\left[\log p(X_h|Z)\right] \\ + \mathbb{E}_{q(Z|X_t,X_h)}[\log p(Z|L) - \log q(Z|X_t, X_h)]\end{aligned} \tag{7}$$

Incorporating SpatialDIVA's prior-constrained latent variables (Appendices B.1 and G), we obtain our evidence lower bound (ELBO), where θ and ϕ are neural network parameters (Appendix F):

$$\begin{aligned}L(\theta, \phi) = \mathbb{E}_{q_\phi(Z_y^k, Z_d, Z_r|X_t,X_h)}[\log p_\theta(X_t|Z_y^k, Z_d, Z_r)] \\ + \mathbb{E}_{q_\phi(Z_y^k, Z_d, Z_r|X_t,X_h)}[\log p_\theta(X_h|Z_y^k, Z_d, Z_r)] - \sum_{k=1}^{m} D_{KL}(q_\phi(Z_y^k|X_t, X_h) \parallel p_\theta(Z_y^k|L^k)) \\ - D_{KL}(q_\phi(Z_d|X_t, X_h) \parallel p_\theta(Zd|L^d)) - D_{KL}(q_\phi(Z_r|X_t, X_h) \parallel p_\theta(Z_r))\end{aligned} \tag{8}$$

Similar to DIVA (Ilse et al., 2019) and CCVAE (Joy et al., 2020), we incorporate classification heads parameterized by ψ, to ensure that the posterior distributions contain the relevant labeled knowledge (L), giving the overall objective:

$$L(\theta, \phi, \psi) = L(\theta, \phi) + \sum_{k=1}^{m} \mathbb{E}_{q_\phi(Z_y^k|X_t,X_h)}[\log q_\psi(L_y^k|Z_y^k)] + \mathbb{E}_{q_\phi(Z_d|X_t,X_h)}[\log q_\psi(L^d|Z_d)]. \tag{9}$$

We encode the **batch labels (L^d)** using an index of the slide from which each spot originates, enabling generalization across slide contexts. **Prior biological knowledge** is encoded based on labels that contain biologically relevant information from both modalities (L^k) (Appendices B.1 and H).

For **intrinsic transcriptomic variation**, we use a categorical distribution of expert-annotated cell-type labels for each spot (Appendix H). For **morphology variation**, we use pathologist annotations of morphology features on histology slides corresponding to each spot, through a categorical distribution (Appendix H). In cases where prior annotations might not be available, unsupervised clustering on the ST and histology data individually can be used to derive modality-specific labels.

The last biologically informative prior that we consider is **spatial context**. Each spot on a spatial transcriptomics slide has spatial coordinates P_x, P_y. However, these coordinates are not generalizable across slides. Therefore, we developed a **spatially aware context distribution** for each spot (Appendix H). For each spot i, we determine the N nearest neighbors using the spatial coordinates (P_x, P_y). For both the transcriptomic (X_t) and histology (X_h) features, we decompose their information for all spots in a slide through principal component analysis (PCA):

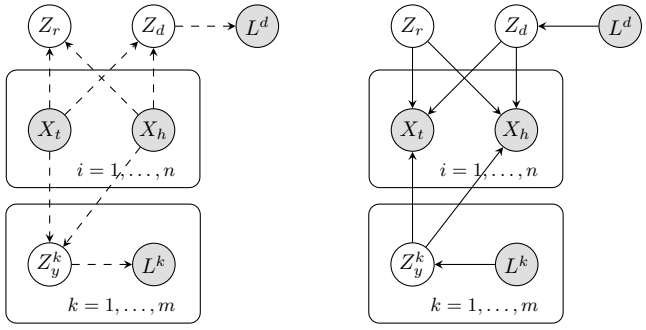
$$\Sigma = \frac{1}{N-1} X^\top X$$
$$\widetilde{X} = U \, \Sigma_{\text{SVD}} \, V^\top \tag{10}$$

We use the concatenated decomposed representations of X_t and X_h ($\widetilde{X}_{th} = [\widetilde{X_t}|\widetilde{X_h}] \in \mathbb{R}^{N \times (d+d)}$) of the N neighbors for spot i to obtain a representation of spatial context (Y^s):

$$Y_i^S = \frac{1}{N} \sum_{j \in N_i}^{N} \widetilde{X}_{th_j} \tag{11}$$

A key advantage of this approach is that the model does not have to predict the high-dimensional ST and histology distributions, leading to faster training when used in the classification loss and prior distribution (Eqn. 9) ($Y^S = L_y^S$ for spatial context). This generalizable representation of spatial context can be used across slides/samples, and allows for multi-sample training. Complete details on all of the biological priors can be found in Appendices B.1 and H.

Within this problem context, SpatialDIVA offers significant advantages compared to other approaches for modelling multi-modal ST and histology data (Appendix Table 2, Appendix A).



(a) Inference (b) Generation

Figure 2: **SpatialDIVA model overview. (a)** The observed ST (X_t) and histology (X_h) representations are used to infer latent residual variation (Z_r), batch/technical variation (Z_d), and variation of key biological factors (Z_y^k), including intrinsic transcriptomic, morphological, and spatial factors. **(b)** The X_t and X_h likelihoods are generated by conditioning on learned factors, where the conditioning is controllable.

4 Results

4.1 Evaluating disentanglement of multi-modal ST and histology data

To assess how well SpatialDIVA can disentangle factors of variation affecting both modalities, we compared it with baseline models using a series of disentanglement metrics across pathologist-annotated datasets.

Datasets: We collated multi-patient pancreatic ductal adenocarcinoma (PDAC, 13 slides) and colorectal cancer (CRC, 14 slides) data profiled with ST and H&E imaging (Cui Zhou et al., 2022; Valdeolivas et al., 2024). We created pathologist annotations for tumour/normal tissue regions for the PDAC dataset and used existing pathologist annotations for the CRC dataset (Appendix D). We preprocess the data such that we obtain patches of the H&E image that correspond to the area around the spots that capture ST data (Appendix D). Image features for a per-spot representation are obtained through zero-shot inference of the UNI histopathology foundation model (Chen et al., 2024). ST data is processed uniformly for all datasets (Appendix D), and highly-variable gene (HVG) selection is performed for all spots across all slides (Appendix E).

Baselines: We compared SpatialDIVA to an array of baselines including PCA, a standard VAE model (Kingma & Welling, 2013), and an unsupervised disentanglement approach in the β-VAE model (Higgins et al., 2016) (Appendix I). Comparison with the DIVA model (Ilse et al., 2019) and similar

approaches (Joy et al., 2020) was not possible as they cannot handle multi-modal data and continuous label distributions for spatial context (Appendix F).

Evaluation: A disentangled representation would result in each factor containing specific information about morphology, transcriptomic, spatial, and batch variation. For morphology we used pathologist annotations of the data at a per-patch level. For transcriptomic information, we used expert-annotated cell-types within each spot. For batch variation, we used the slide label that each spot originates from. It was unclear how to evaluate continuous spatial variation in this case, so this was omitted. The categorical labels for the morphological, transcriptomic, and batch variation were one-hot encoded and embeddings from the baselines and SpatialDIVA model were compared with the encoded labels to assess disentanglement (Appendix I).

Models were trained on randomly selected 90% subsets of the datasets for PDAC and CRC (one model per cancer type), then evaluated on the held-out 10% of the data for 10 iterations (Appendix I). Disentanglement was assessed using multiple metrics for categorical factors and continuous embeddings (Appendix J) (Carbonneau et al., 2020).

Overall, SpatialDIVA performed the best considering an aggregate ranking across all disentanglement metrics used in the benchmark, for both the PDAC and CRC cohorts (Table 1). These results demonstrate that within a multi-modal disentanglement setting, the conclusions from Locatello et al. (2018) hold, and that supervision based on prior information leads to better disentanglement.

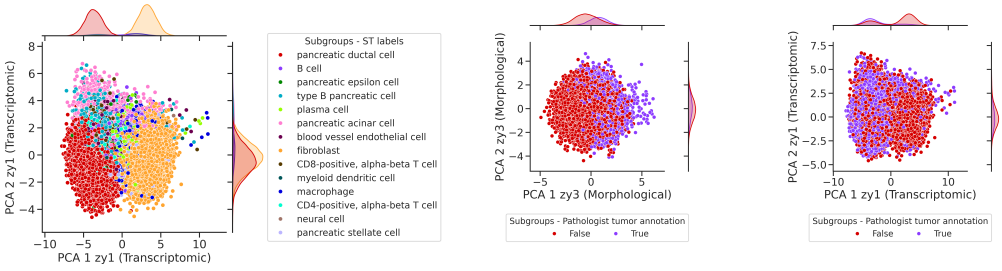
Table 1: Disentanglement benchmark results for the pancreatic (PDAC) and colorectal cancer (CRC) datasets. Results are the mean scores across 10 random subsamples of the datasets. The best results for each metric are in **bold** and the second best results are underlined. Average rank is based on rankings across all metrics for each dataset (Appendix I). Standard deviations across iterations can be found in Appendix C (Table 3).

	Colorectal cancer cohort				**Pancreatic cancer cohort**			
Metric \ Method	**PCA**	**VAE**	**β-VAE**	**SpatialDIVA**	**PCA**	**VAE**	**β-VAE**	**SpatialDIVA**
JEMMING (↑)	0.2011	0.7690	**0.9200**	0.3537	0.2579	0.7922	**0.9032**	0.5327
SAP (↑)	**0.0049**	0.0000	0.0000	0.0024	**0.0203**	0.0000	0.0000	0.0034
MIG (↑)	0.0075	0.0049	0.0041	**0.0128**	0.0225	0.0057	0.0038	**0.0893**
MIG-SUP (↑)	0.0102	**0.0541**	0.0328	0.0336	0.0186	**0.0668**	0.0246	0.0577
Modularity (↑)	0.9473	0.9600	0.9562	**0.9647**	**0.9421**	0.9298	0.8938	0.9357
DCI-MIG (↑)	0.0337	0.0007	0.0003	**0.1231**	0.0700	0.0001	0.0001	**0.2131**
Explicitness (↑)	**0.9505**	0.1032	0.0409	0.9375	0.9010	0.0293	0.0035	**0.9433**
IRS (↑)	0.5008	**0.5196**	0.5001	0.5085	0.4459	0.4951	0.5000	**0.5351**
Average Rank (↓)	3	2	4	**1**	2	3	4	**1**

4.2 Assessing disentanglement and covariance of latent factors

To determine how the disentanglement properties of SpatialDIVA affect the learned latent spaces, we trained the model on the 13 PDAC slides and evaluated samples from the posterior distributions $q_\phi(Z_y^k|X_t, X_h)$ and $q_\phi(Z_d|X_t, X_h)$, conditioned on the observed data X_t and X_h (Appendix I). We performed PCA on high-dimensional samples from one factor and visualized the first two PCs. We then overlaid the label distributions onto the PCA reduction of each learned factor, to visually examine how the annotated labels covary in the factor-specific latent spaces (Fig. 3 and Appendix Fig. 5).

Examining the first 2 PCs demonstrates $q_\phi(Z_y^1|X_t, X_h)$ captures variation in the transcriptomic subgroups, specifically separating the fibroblast and pancreatic ductal cell populations (Fig. 3a). The posterior distribution for morphology ($q_\phi(Z_y^3|X_t, X_h)$) captures distinction between pathologist-annotated tumor/normal areas in the histology data (Fig. 3b). Interestingly, overlaying these labels from the histology data onto the transcriptomics latent space ($q_\phi(Z_y^1|X_t, X_h)$) demonstrates the transcriptomics information also distinguishes the tumor and normal pathologist annotations (Fig. 3c), indicating that there is a high degree of mutual information between the pathologist annotations and transcriptomics information in this setting. This type of analysis can be done with any number of factors using SpatialDIVA.



(a) **(b)** **(c)**

Figure 3: **Transcriptomic and morphology-associated posterior distributions.** High-dimensional samples from the posterior distributions ($q_\phi(Z_y^k|X_t, X_h)$) of the transcriptomic-associated (a, c) and morphology-associated (b) latent spaces, reduced using PCA and the two axes associated with the highest variation (x, y) are shown. The cell-type (a) and pathologist annotations (b, c) are overlaid, with the density of the labels shown across the axes.

We next investigated how SpatialDIVA removes batch and technical effects (Fig. 1). To our knowledge, SpatialDIVA is the first method that explicitly models batch effects from both the ST expression data and histology image data in a joint manner (Fig. 2a). Batch correction is a challenging task in both the transcriptomics and histology spaces (Kothari et al., 2014; Guo et al., 2023). Visualizing the batch covariate and associated posterior samples ($q_\phi(Z_d|X_t, X_h)$) shows that this latent factor captures the batch variation, as most batches can be distinguished even in the first two PCs of this distribution (Appendix Fig. 5a). When we examine the transcriptomic and morphological latent distributions ($q_\phi(Z_y^1|X_t, X_h)$ and $q_\phi(Z_y^3|X_t, X_h)$ respectively) and overlay the batch/slide labels, we find that this variation has been removed from these factors (Appendix Fig. 5b,c), validating that Z_d effectively captures batch variation.

We compared the batch-correction ability of SpatialDIVA to a conditional VAE (cVAE) model that conditions on batch (slide number), removing all respective across-slide variation (Appendix I). We find that the different posterior distributions of SpatialDIVA minimize batch effects, but do not remove inter-batch variation completely, as biological variation for the relevant posterior distributions (Z_y^k) differs between batches (Appendix Table 4). As expected, the batch-associated posterior (Z_d) has the lowest batch-correction score, as it captures batch variation (Appendix Table 4). The residual variation (Z_r) has the highest score, indicating it is strongly independent of slide context (Appendix Table 4). Furthermore, SpatialDIVA preserves biological information in its latent spaces better than the cVAE baseline (Appendix Table 5).

4.3 Conditional generation of multi-modal data

As SpatialDIVA is a generative model, it is possible to generate new data (X_t, X_h) while intervening on disentangled factors. Specifically, we can sample from the likelihoods by conditioning on specific factors sampled from the posterior, while setting others to a constant value. For instance, if we condition on only transcriptomic context (Z_y^1):

$$\begin{aligned} Z_y^1 &\sim q_\phi(Z_y^1|X_t, X_h) \\ p_\theta(X_t|Z_y^1, Z_y^{i\neq 1} = C, Z_r = C, Z_d = C) \\ p_\theta(X_h|Z_y^1, Z_y^{i\neq 1} = C, Z_r = C, Z_d = C) \end{aligned} \tag{12}$$

Essentially, Eqn. 12 indicates that the generated samples for X_t and X_h will vary according to Z_y^1, but not the other factors, as they are held constant. This allows us to quantify how one disentangled factor influences variation in the multi-modal data. For example, if we want to evaluate the information that morphology-associated variation $q_\phi(Z_y^3|X_t, X_h)$ encodes in the ST data (X_t), we can condition on the morphology factor and hold others constant, when generating the transcriptomic likelihood (Appendix Fig. 6).

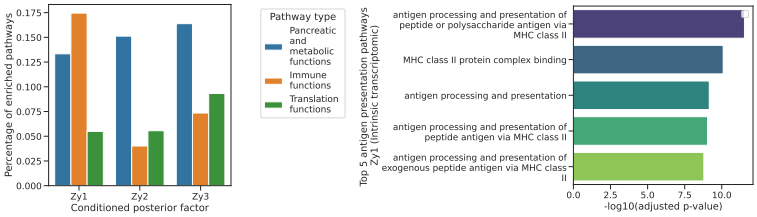
4.4 Conditional generation analysis of PDAC cancer biopsy samples

Using this setup (Sec. 4.3), we sought to understand which gene programs and pathways can be exclusively associated with transcriptomic variation, spatial context, and morphology information,

and which are shared, in pancreatic cancer. Specifically, we trained the SpatialDIVA model on all 13 slides of the multi-patient PDAC cohort and used the generated counts for further analysis (Appendix I). After training the model using a negative binomial likelihood for the ST counts (X_t), we sampled the shape (θ) and mean (μ) parameters for each spot and gene across the PDAC slides, based on each conditional factor: $\theta_{i,g}, \mu_{i,g} \sim p_\theta(X_{t,i,g}|Z_{y,i}^{\text{Cond}}, Z_{y,i}^{i \notin [Cond]} = 0, Z_{r,i} = 0, Z_{d,i} = 0)$. After obtaining these parameters for each spot on a per-gene level, we obtained negative binomial distribution parameterizations which we then sampled counts from: $X_{t,i,g} \sim \text{NB}_{i,g}(\theta_{i,g}, \frac{\mu_{i,g}}{\theta_{i,g}})$. This was done for a random subset of 10000 spots sampled from all PDAC slides.

We first determined whether the differentially expressed genes (DEGs) in the counts generated are specific or shared across factors (Appendix I). Differential expression quantifies which genes exhibit statistically significant variance between clusters of spots, and are indicative of cell-types, cell-states and gene programs that can be captured by transcriptional counts (Wolf et al., 2018) (Appendix I).

The top 500 DEGs in the conditionally generated counts from each factor showed mutual exclusivity (Appendix Fig. 7a). This indicates that the gene programs associated with these factors are likely to be mutually exclusive, and is another result that shows the SpatialDIVA model has successfully performed multi-modal disentanglement.



(a) **(b)**

Figure 4: **Pathway enrichment of top 500 DEGs from conditional generation.** (a) The number of enriched pathways specific to a functional group for each posterior factor, as a percentage of total enriched pathways from that posterior factor. (b) The top 5 antigen presentation-specific pathways with the highest enrichment based on corrected p-value for the Z_y^1 (intrinsic transcriptomic variation) conditioned generated transcriptomic counts.

We analyzed more concretely the higher-level gene programs encoded by the top DEGs based on pathway enrichment analysis. Pathway analysis assesses statistical overrepresentation for a set of genes in biological pathways curated by experts (Kolberg et al., 2023) (Appendix I). As expected based on the DEG analysis, there were many factor-specific enriched pathways, with spatial (Z_y^2) and morphological (Z_y^3) pathways exhibiting the highest overlap (Appendix Fig. 7b). Further examination of the pathways based on keyword matching (Appendix I) reveals that all three factors encode general pancreatic and metabolic functions (Fig. 4a). Transcriptomic variation (Z_y^1) was found to capture the majority of the immunological signal (Fig. 4a,b). This result is important as PDAC characterization and progression is significantly influenced by immune cells in the tumor microenvironment (Karamitopoulou, 2019). The morphological variation (Z_y^3) was enriched for translation-specific functions (Fig. 4a).

This analysis offers a vignette showing the capabilities of SpatialDIVA in determining the contribution of disentangled factors to specific information in the observed data. Depending on the clinical importance of certain gene programs, specific assays and analyses can be prioritized, such as assays for transcriptomic variation, which was found to capture most of the immunological signal in the PDAC data.

5 Conclusion

Here, we framed the challenge of determining the generative factors of ST and histopathology, and their overlap, as multi-modal disentanglement. We introduced *SpatialDIVA*, the first technique to perform multi-modal disentanglement in this setting, which leads to an interpretable and flexible set of posterior distributions that are able to generate novel biological insights through multi-modal conditional generation. The problem formulation and evaluation framework we developed is an important resource for the community in creating models that learn factors relevant to each modality.

MEANINGFULNESS STATEMENT

A meaningful representation of life should be interpretable, representative of the population, biologically/clinically useful, and be invariant to artifacts that may arise during data acquisition, such as batch effects. Within these constraints, SpatialDIVA is the first approach to jointly model multi-modal histopathology and spatial transcriptomics data in a manner that removes batch-effects, while returning a disentangled and interpretable model of the data generating distributions for both modalities. Further, SpatialDIVA allows biologists and clinicians to draw insights on the type of information present in histopathology and spatial transcriptomics, which can help guide experimental and clinical workflows.

ACKNOWLEDGMENTS

The authors thank Michael Geuenich and Chengxin Yu for providing curated lists of PDAC marker genes. HM thanks Julia Greissl, Paidamoyo (Ash) Chapfuwa, Ted Meeds, Melanie F. Pradier, and Niranjani Prasad from Microsoft Research for insightful discussions on DIVA and related models.

REFERENCES

Bruce Alberts, Rebecca Heald, Alexander Johnson, David Morgan, Martin Raff, Keith Roberts, and Peter Walter. *Molecular biology of the cell (seventh edition)*. WW Norton, New York, NY, July 2022.

Michael Barresi and Scott Gilbert. *Developmental biology*. Oxford University Press, New York, NY, 13 edition, April 2023.

Tommaso Biancalani, Gabriele Scalia, Lorenzo Buffoni, Raghav Avasthi, Ziqing Lu, Aman Sanger, Neriman Tokcan, Charles R Vanderburg, Asa Segerstolpe, Meng Zhang, Inbal Avraham-Davidi, Sanja Vickovic, Mor Nitzan, Sai Ma, Ayshwarya Subramanian, Michal Lipinski, Jason Buenrostro, Nik Bear Brown, Duccio Fanelli, Xiaowei Zhuang, Evan Z Macosko, and Aviv Regev. Deep learning and alignment of spatially resolved single-cell transcriptomes with tangram. *Nat. Methods*, 18(11):1352–1362, November 2021.

Johann Brehmer, Pim de Haan, Phillip Lippe, and Taco Cohen. Weakly supervised causal representation learning. *arXiv [stat.ML]*, March 2022.

Dario Bressan, Giorgia Battistoni, and Gregory J Hannon. The dawn of spatial omics. *Science*, 381 (6657):eabq4964, August 2023.

Shenghao Cao and Ye Yuan. A framework for gene representation on spatial transcriptomics. *bioRxiv*, September 2024.

Marc-André Carbonneau, Julian Zaidi, Jonathan Boilard, and Ghyslain Gagnon. Measuring disentanglement: A review of metrics. *arXiv [cs.LG]*, December 2020.

Richard J Chen, Tong Ding, Ming Y Lu, Drew F K Williamson, Guillaume Jaume, Andrew H Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, Mane Williams, Lukas Oldenburg, Luca L Weishaupt, Judy J Wang, Anurag Vaidya, Long Phi Le, Georg Gerber, Sharifa Sahai, Walt Williams, and Faisal Mahmood. Towards a general-purpose foundation model for computational pathology. *Nat. Med.*, 30(3):850–862, March 2024.

Ricky T Q Chen, Xuechen Li, Roger Grosse, and David Duvenaud. Isolating sources of disentanglement in variational autoencoders. *arXiv [cs.LG]*, February 2018.

Ozan Ciga, Tony Xu, and Anne Louise Martel. Self supervised contrastive learning for digital histopathology. *Mach. Learn. Appl.*, 7(100198):100198, March 2022.

Micaela E Consens, Cameron Dufault, Michael Wainberg, Duncan Forster, Mehran Karimzadeh, Hani Goodarzi, Fabian J Theis, Alan Moses, and Bo Wang. To transformers and beyond: Large language models for the genome. *arXiv [q-bio.GN]*, November 2023.

Daniel Cui Zhou, Reyka G Jayasinghe, Siqi Chen, John M Herndon, Michael D Iglesia, Pooja Navale, Michael C Wendl, Wagma Caravan, Kazuhito Sato, Erik Storrs, Chia-Kuei Mo, Jingxian Liu, Austin N Southard-Smith, Yige Wu, Nataly Naser Al Deen, John M Baer, Robert S Fulton, Matthew A Wyczalkowski, Ruiyang Liu, Catrina C Fronick, Lucinda A Fulton, Andrew Shinkle, Lisa Thammavong, Houxiang Zhu, Hua Sun, Liang-Bo Wang, Yize Li, Chong Zuo, Joshua F McMichael, Sherri R Davies, Elizabeth L Appelbaum, Keenan J Robbins, Sara E Chasnoff, Xiaolu Yang, Ashley N Reeb, Clara Oh, Mamatha Serasanambati, Preet Lal, Rajees Varghese, Jay R Mashl, Jennifer Ponce, Nadezhda V Terekhanova, Lijun Yao, Fang Wang, Lijun Chen, Michael Schnaubelt, Rita Jui-Hsien Lu, Julie K Schwarz, Sidharth V Puram, Albert H Kim, Sheng-Kwei Song, Kooresh I Shoghi, Ken S Lau, Tao Ju, Ken Chen, Deyali Chatterjee, William G Hawkins, Hui Zhang, Samuel Achilefu, Milan G Chheda, Stephen T Oh, William E Gillanders, Feng Chen, David G DeNardo, Ryan C Fields, and Li Ding. Spatially restricted drivers and transitional cell populations cooperate with the microenvironment in untreated and chemo-resistant pancreatic cancer. *Nat. Genet.*, 54(9):1390–1405, September 2022.

Kien Do and Truyen Tran. Theory and evaluation metrics for learning disentangled representations. In *International Conference on Learning Representations*, 2020.

Sirio Dupont, Leonardo Morsut, Mariaceleste Aragona, Elena Enzo, Stefano Giulitti, Michelangelo Cordenonsi, Francesca Zanconato, Jimmy Le Digabel, Mattia Forcato, Silvio Bicciato, Nicola Elvassore, and Stefano Piccolo. Role of YAP/TAZ in mechanotransduction. *Nature*, 474(7350): 179–183, June 2011.

Andrew H Fischer, Kenneth A Jacobson, Jack Rose, and Rolf Zeller. Hematoxylin and eosin staining of tissue and cell sections. *CSH Protoc.*, 2008(5):db.prot4986, May 2008.

David S Fischer, Anna C Schaar, and Fabian J Theis. Modeling intercellular communication in tissues using spatial graphs of cells. *Nat. Biotechnol.*, 41(3):332–336, March 2023.

Adam Gayoso, Romain Lopez, Galen Xing, Pierre Boyeau, Valeh Valiollah Pour Amiri, Justin Hong, Katherine Wu, Michael Jayasuriya, Edouard Mehlman, Maxime Langevin, Yining Liu, Jules Samaran, Gabriel Misrachi, Achille Nazaret, Oscar Clivio, Chenling Xu, Tal Ashuach, Mariano Gabitto, Mohammad Lotfollahi, Valentine Svensson, Eduardo da Veiga Beltrame, Vitalii Kleshchevnikov, Carlos Talavera-López, Lior Pachter, Fabian J Theis, Aaron Streets, Michael I Jordan, Jeffrey Regier, and Nir Yosef. A python library for probabilistic analysis of single-cell omics data. *Nat. Biotechnol.*, 40(2):163–166, February 2022.

Sean Gillies et al. Shapely: manipulation and analysis of geometric objects, 2007. URL `https://github.com/shapely/shapely`.

Tiantian Guo, Zhiyuan Yuan, Yan Pan, Jiakang Wang, Fengling Chen, Michael Q Zhang, and Xiangyu Li. SPIRAL: integrating and aligning spatially resolved transcriptomics data across different experiments, conditions, and technologies. *Genome Biol.*, 24(1):241, October 2023.

Douglas Hanahan. Hallmarks of cancer: New dimensions. *Cancer Discov.*, 12(1):31–46, January 2022.

Yuhan Hao, Tim Stuart, Madeline H Kowalski, Saket Choudhary, Paul Hoffman, Austin Hartman, Avi Srivastava, Gesmira Molla, Shaista Madad, Carlos Fernandez-Granda, and Rahul Satija. Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat. Biotechnol.*, 42(2): 293–304, February 2024.

Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020. doi: 10.1038/s41586-020-2649-2. URL `https://doi.org/10.1038/s41586-020-2649-2`.

Bryan He, Ludvig Bergenstrahle, Linnea Stenbeck, Abubakar Abid, Alma Andersson, Ake Borg, Jonas Maaskola, Joakim Lundeberg, and James Zou. Integrating spatial gene expression and breast tumour morphology via deep learning. *Nat. Biomed. Eng.*, 4(8):827–834, August 2020.

Lei He, L Rodney Long, Sameer Antani, and George R Thoma. Histology image analysis for carcinoma detection and grading. *Comput. Methods Programs Biomed.*, 107(3):538–556, September 2012.

I Higgins, L Matthey, Arka Pal, Christopher P Burgess, Xavier Glorot, M Botvinick, S Mohamed, and Alexander Lerchner. Beta-VAE: Learning basic visual concepts with a constrained variational framework. *Int Conf Learn Represent*, November 2016.

Jian Hu, Xiangjie Li, Kyle Coleman, Amelia Schroeder, Nan Ma, David J Irwin, Edward B Lee, Russell T Shinohara, and Mingyao Li. SpaGCN: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nat. Methods*, 18(11):1342–1351, November 2021.

Maximilian Ilse, Jakub M Tomczak, Christos Louizos, and Max Welling. DIVA: Domain invariant variational autoencoders. *arXiv [stat.ML]*, pp. 322–348, May 2019.

Amanda Janesick, Robert Shelansky, Andrew D Gottscho, Florian Wagner, Stephen R Williams, Morgane Rouault, Ghezal Beliakoff, Carolyn A Morrison, Michelli F Oliveira, Jordan T Sicherman, Andrew Kohlway, Jawad Abousoud, Tingsheng Yu Drennon, Seayar H Mohabbat, 10x Development Teams, and Sarah E B Taylor. High resolution mapping of the tumor microenvironment using integrated single-cell, spatial and in situ analysis. *Nat. Commun.*, 14(1):8353, December 2023.

Guillaume Jaume, Paul Doucet, Andrew H Song, Ming Y Lu, Cristina Almagro-Pérez, Sophia J Wagner, Anurag J Vaidya, Richard J Chen, Drew F K Williamson, Ahrong Kim, and Faisal Mahmood. HEST-1k: A dataset for spatial transcriptomics and histology image analysis. *arXiv [cs.CV]*, June 2024.

Tom Joy, Sebastian M Schmon, Philip H S Torr, N Siddharth, and Tom Rainforth. Capturing label characteristics in VAEs. *arXiv [cs.LG]*, June 2020.

Eva Karamitopoulou. Tumour microenvironment of pancreatic cancer: immune landscape is dictated by molecular and histopathological features. *Br. J. Cancer*, 121(1):5–14, July 2019.

A A Kiger, B Baum, S Jones, M R Jones, A Coulson, C Echeverri, and N Perrimon. A functional genomic analysis of cell morphology using RNA interference. *J. Biol.*, 2(4):27, October 2003.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv [cs.LG]*, December 2014.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv [stat.ML]*, December 2013.

Vitalii Kleshchevnikov, Artem Shmatko, Emma Dann, Alexander Aivazidis, Hamish W King, Tong Li, Rasa Elmentaite, Artem Lomakin, Veronika Kedlian, Adam Gayoso, Mika Sarkin Jain, Jun Sung Park, Lauma Ramona, Elizabeth Tuck, Anna Arutyunyan, Roser Vento-Tormo, Moritz Gerstung, Louisa James, Oliver Stegle, and Omer Ali Bayraktar. Cell2location maps fine-grained cell types in spatial transcriptomics. *Nat. Biotechnol.*, 40(5):661–671, May 2022.

Liis Kolberg, Uku Raudvere, Ivan Kuzmin, Priit Adler, Jaak Vilo, and Hedi Peterson. g:profiler-interoperable web service for functional enrichment analysis and gene identifier mapping (2023 update). *Nucleic Acids Res.*, 51(W1):W207–W212, July 2023.

Sonal Kothari, John H Phan, Todd H Stokes, Adeboye O Osunkoya, Andrew N Young, and May D Wang. Removing batch effects from histopathological images for enhanced cancer diagnosis. *IEEE J. Biomed. Health Inform.*, 18(3):765–772, May 2014.

Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. Variational inference of disentangled latent concepts from unlabeled observations. In *International Conference on Learning Representations*, 2018.

Jeffrey T Leek, Robert B Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W Evan Johnson, Donald Geman, Keith Baggerly, and Rafael A Irizarry. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.*, 11(10):733–739, October 2010.

Zhiyuan Li, Jaideep Vitthal Murkute, Prashnna Kumar Gyawali, and Linwei Wang. Progressive learning and disentanglement of hierarchical representations. In *International Conference on Learning Representations*, 2020.

Alexander Lin and Alex X Lu. Incorporating knowledge of plates in batch normalization improves generalization of deep learning for microscopy images. *bioRxiv*, October 2022.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. *arXiv [cs.LG]*, November 2018.

Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. Weakly-supervised disentanglement without compromises. *arXiv [cs.LG]*, February 2020.

Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nat. Methods*, 15(12):1053–1058, 2018.

Malte D Luecken, M Büttner, K Chaichoompu, A Danese, M Interlandi, M F Mueller, D C Strobl, L Zappia, M Dugas, M Colomé-Tatché, and Fabian J Theis. Benchmarking atlas-level data integration in single-cell genomics. *Nat. Methods*, 19(1):41–50, January 2022.

Ying Ma and Xiang Zhou. Spatially informed cell-type deconvolution for spatial transcriptomics. *Nat. Biotechnol.*, 40(9):1349–1359, September 2022.

Lambda Moses and Lior Pachter. Museum of spatial transcriptomics. *Nat. Methods*, 19(5):534–546, May 2022.

Giovanni Palla, Hannah Spitzer, Michal Klein, David Fischer, Anna Christina Schaar, Louis Benedikt Kuemmerle, Sergei Rybakov, Ignacio L Ibarra, Olle Holmberg, Isaac Virshup, Mohammad Lotfollahi, Sabrina Richter, and Fabian J Theis. Squidpy: a scalable framework for spatial omics analysis. *Nat. Methods*, 19(2):171–178, February 2022.

Minxing Pang, Kenong Su, and Mingyao Li. Leveraging information in spatial transcriptomics to predict super-resolution gene expression from histology images in tumors. *bioRxiv*, November 2021.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Junya Peng, Bao Fa Sun, Chuan Yuan Chen, Jia Yi Zhou, Yu Sheng Chen, Hao Chen, Lulu Liu, Dan Huang, Jialin Jiang, Guan Shen Cui, Ying Yang, Wenze Wang, Dan Guo, Menghua Dai, Junchao Guo, Taiping Zhang, Quan Liao, Yi Liu, Yong Liang Zhao, Da Li Han, Yupei Zhao, Yun Gui Yang, and Wenming Wu. Single-cell RNA-seq highlights intra-tumoral heterogeneity and malignant progression in pancreatic ductal adenocarcinoma. *Cell Res.*, 29(9):725–738, 2019.

Duy Pham, Xiao Tan, Brad Balderson, Jun Xu, Laura F Grice, Sohye Yoon, Emily F Willis, Minh Tran, Pui Yeng Lam, Arti Raghubar, Priyakshi Kalita-de Croft, Sunil Lakhani, Jana Vukovic, Marc J Ruitenberg, and Quan H Nguyen. Robust mapping of spatiotemporal trajectories and cell-cell interactions in healthy and diseased tissues. *Nat. Commun.*, 14(1):7739, November 2023.

Anjali Rao, Dalia Barkley, Gustavo S França, and Itai Yanai. Exploring tissue architecture using spatial transcriptomics. *Nature*, 596(7871):211–220, 2021.

Karl Ridgeway and Michael C Mozer. Learning deep disentangled embeddings with the f-statistic loss. In *Advances in Neural Information Processing Systems*, 2018.

Anna Christina Schaar, Alejandro Tejada-Lapuerta, Giovanni Palla, Robert Gutgesell, Lennard Halle, Mariia Minaeva, Larsen Vornholz, Leander Dony, Francesca Drummer, Mojtaba Bahrami, and Fabian J Theis. Nicheformer: a foundation model for single-cell and spatial omics. *bioRxiv*, April 2024.

Uwe Schmidt, Martin Weigert, Coleman Broaddus, and Gene Myers. Cell detection with star-convex polygons. *arXiv [cs.CV]*, June 2018.

Anna Sepliarskaia, Julia Kiseleva, and Maarten de Rijke. Evaluating disentangled representations. *arXiv:1910.05587*, 2020.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Neural Inf Process Syst*, pp. 3483–3491, December 2015.

Patrik L Ståhl, Fredrik Salmén, Sanja Vickovic, Anna Lundmark, José Fernández Navarro, Jens Magnusson, Stefania Giacomello, Michaela Asp, Jakub O Westholm, Mikael Huss, Annelie Mollbrink, Sten Linnarsson, Simone Codeluppi, Ake Borg, Fredrik Ponten, Paul Igor Costea, Pelin Sahlén, Jan Mulder, Olaf Bergmann, Joakim Lundeberg, and Jonas Frisén. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353(6294): 78–82, 2016.

Raphael Suter, Djordje Miladinovic, Bernhard Schölkopf, and Stefan Bauer. Robustly disentangled causal mechanisms: Validating deep representations for interventional robustness. In *International Conference on Machine Learning*, 2019.

Luyi Tian, Fei Chen, and Evan Z Macosko. The expanding vistas of spatial transcriptomics. *Nat. Biotechnol.*, 41(6):773–782, June 2023.

Hoa Thi Nhu Tran, Kok Siong Ang, Marion Chevrier, Xiaomeng Zhang, Nicole Yee Shin Lee, Michelle Goh, and Jinmiao Chen. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol.*, 21(1):1–32, 2020.

Alberto Valdeolivas, Bettina Amberg, Nicolas Giroud, Marion Richardson, Eric J C Gálvez, Solveig Badillo, Alice Julien-Laferrière, Demeter Túrós, Lena Voith von Voithenberg, Isabelle Wells, Benedek Pesti, Amy A Lo, Emilio Yángüez, Meghna Das Thakur, Michael Bscheider, Marc Sultan, Nadine Kumpesa, Björn Jacobsen, Tobias Bergauer, Julio Saez-Rodriguez, Sven Rottenberg, Petra C Schwalie, and Kerstin Hahn. Profiling the heterogeneity of colorectal cancer consensus molecular subtypes using spatial transcriptomics. *NPJ Precis. Oncol.*, 8(1):10, January 2024.

Viola Vogel and Michael Sheetz. Local force and geometry sensing regulate cell functions. *Nat. Rev. Mol. Cell Biol.*, 7(4):265–275, April 2006.

Lina Wadi, Mona Meyer, Joel Weiser, Lincoln D Stein, and Jüri Reimand. Impact of outdated gene annotations on pathway enrichment analysis. *Nat. Methods*, 13(9):705–706, August 2016.

Ross Wightman. Pytorch image models. `https://github.com/rwightman/pytorch-image-models`, 2019.

F Alexander Wolf, Philipp Angerer, and Fabian J Theis. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.*, 19(1):15, December 2018.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M Rush. HuggingFace's transformers: State-of-the-art natural language processing. *arXiv [cs.CL]*, October 2019.

Ronald Xie, Kuan Pang, Sai W Chung, Catia T Perciani, Sonya A MacParland, Bo Wang, and Gary D Bader. Spatially resolved gene expression prediction from H&E histology images via bi-modal contrastive learning. *arXiv [cs.CV]*, June 2023.

Chang Xu, Xiyun Jin, Songren Wei, Pingping Wang, Meng Luo, Zhaochun Xu, Wenyi Yang, Yideng Cai, Lixing Xiao, Xiaoyu Lin, Hongxin Liu, Rui Cheng, Fenglan Pang, Rui Chen, Xi Su, Ying Hu, Guohua Wang, and Qinghua Jiang. DeepST: identifying spatial domains in spatial transcriptomics by deep learning. *Nucleic Acids Res.*, 50(22):e131, December 2022.

Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, Sheng Zhang, Rajesh Rao, Tristan Naumann, Cliff Wong, Zelalem Gero, Javier González, Yu Gu, Yanbo Xu, Mu Wei, Wenhui Wang, Shuming Ma, Furu Wei, Jianwei Yang, Chunyuan Li, Jianfeng Gao, Jaylen Rosemon, Tucker Bower, Soohee Lee, Roshanthi Weerasinghe, Bill J Wright, Ari Robicsek, Brian Piening, Carlo Bifulco, Sheng Wang, and Hoifung Poon. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 630(8015):181–188, June 2024.

Chengxin Yu, Michael J Geuenich, Sabrina Ge, Gun-Ho Jang, Tan Tiak Ju, Amy Zhang, Grainne M O'Kane, Faiyaz Notta, and Kieran R Campbell. Decoding multicellular niche formation in the tumour microenvironment from nonspatial single-cell expression data. *bioRxiv*, August 2024.

Yuansong Zeng, Zhuoyi Wei, Weijiang Yu, Rui Yin, Yuchen Yuan, Bingling Li, Zhonghui Tang, Yutong Lu, and Yuedong Yang. Spatial transcriptomics prediction from histology jointly through transformer and graph neural networks. *Brief. Bioinform.*, 23(5), September 2022.

Chongyue Zhao, Zhongli Xu, Xinjun Wang, Shiyue Tao, William A MacDonald, Kun He, Amanda C Poholek, Kong Chen, Heng Huang, and Wei Chen. Innovative super-resolution in spatial transcriptomics: a transformer model exploiting histology images and spatial gene expression. *Brief. Bioinform.*, 25(2), January 2024.

Peiyao Zhao, Jiaqiang Zhu, Ying Ma, and Xiang Zhou. Modeling zero inflation is not necessary for spatial transcriptomics. *Genome Biol.*, 23(1):118, May 2022.

A RELATED WORK

ST and H&E models: Existing task-specific models for ST data include inferring cell-cell communication in a given tissue context (Fischer et al., 2023), general representation learning (Xu et al., 2024), deconvolving cell-types in ST spots (Ma & Zhou, 2022), and aligning ST data with single-cell RNA sequencing data (Biancalani et al., 2021).

Advances in foundation models have naturally led to an increasing interest in their uses for biological data (Consens et al., 2023). Large pretrained models for histopathology data have been developed since the rise of self-supervised pretraining in computer vision models (Ciga et al., 2022; Chen et al., 2024; Xu et al., 2024). In our work, we use the UNI model (Chen et al., 2024) to obtain informative representations of the histology images (X_h). ST foundation models have also been introduced (Schaar et al., 2024; Cao & Yuan, 2024), and can be used to generate embeddings of the ST counts (X_t) for use in the SpatialDIVA model.

Several models have been developed that aim to learn joint representations of the ST and histology data, for various downstream tasks (Hu et al., 2021; Xu et al., 2022; Pham et al., 2023; Zhao et al., 2024). However, these models do not learn disentangled representations of factors of variation for both modalities, as we do in this work.

Histology to gene expression prediction models: Many models exploit paired ST and histology to predict gene expression for a given region of an H&E slide (He et al., 2020; Pang et al., 2021; Zeng et al., 2022; Xie et al., 2023). Although this approach can impute gene expression from morphology alone, it is bounded by the maximum mutual information between the two modalities and does not return any relevant information on the specific type of information in each, as well as their associated generative factors.

Disentangled representation learning: There have been several advances in unsupervised disentanglement (Higgins et al., 2016; Chen et al., 2018). However, Locatello et al. (2018) demonstrated empirically and theoretically that without inductive biases, there are no guarantees for learning disentangled representations. Consequently, multiple methods adopted strong and weak supervision, as well as semi-supervised approaches (Ilse et al., 2019; Joy et al., 2020; Locatello et al., 2020; Brehmer et al., 2022).

The two approaches that form the basis for our work are the **domain invariant variational autoencoder (DIVA)** (Ilse et al., 2019) and the **characteristic capturing VAE (ccVAE)** (Joy et al., 2020) models. These models were not designed to handle multi-modal data and continuous label distributions, as is the case for our spatial supervision covariate. As such, we explicitly build upon their work for disentanglement of spatial transcriptomics and histopathology imaging.

B SPATIALDIVA - EXTENDED BACKGROUND AND PROBLEM FORMULATION

Table 2: SpatialDIVA compared to other methods for jointly modeling ST and H&E image data.

Method	Joint embeddings?	Disentangled representations?	Interpretable latent space?	Conditional multi-modal generation?
SpatialDIVA	✓	✓	✓	✓
Unsupervised disentanglement	✓	✓	✗	✗
Histology to gene expression	✓	✗	✗	✗

B.1 BIOLOGICALLY INFORMATIVE PRIORS

Through outlining the flow of information within and across cells, we can determine some of the important factors that generate the ST and histology data distributions.

Intrinsic transcriptomic variation: Variation in the RNA expression of different genes caused by cell intrinsic genetic or epigenetic factors, is an important driving factor for both the ST and histology data. This affects both ST expression variation as well as morphological variation, driven by changes in expression of morphology-associated genes (Kiger et al., 2003).

Morphological variation: Morphological variation on a per-cell level clearly influences the morphological variation on a per-spot level. However, changes in protein expression can also influence transcriptomic variation (Vogel & Sheetz, 2006; Dupont et al., 2011) - there is not a simple linear flow of information from DNA to RNA to protein expression. Therefore, this factor of variation can affect both ST and histology readouts.

Spatial variation: Aside from intrinsic cellular variation at the per-spot level, variation in both transcriptomic and morphological distributions can be driven by spatial context (Bressan et al., 2023). The most concrete example of this is cell to cell communication across spatial domains. This type of communication can influence transcriptomic variation, which in turn can affect both the ST and histology contexts.

Technical/batch variation: Biological factors can influence changes in spot-level representations of the data, however, technical variation can also drive these changes. Batch effects are variations caused by assay differences, experimental differences, or even ambient conditions (Leek et al., 2010). This variation is often conflated with biological variation, unless specifically accounted for (Leek et al., 2010).

Residual variation: Any variation in the ST and histology data distributions that is not accounted for by the aforementioned factors is considered residual variation. If batch/technical effects have been accounted for, residual variation should capture biological effects not captured by the prior information injected through the other factors.

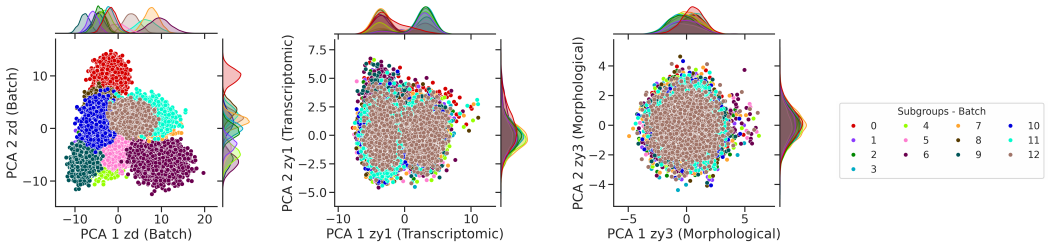
C SpatialDIVA - extended results

C.1 Results Section 4.1

Table 3: Disentanglement benchmark results for the pancreatic and colorectal cancer datasets, **standard deviation results from Table 1**. Results are shown from 10 random subsamples of the datasets.

	Colorectal cancer cohort				Pancreatic cancer cohort			
Metric \ Method	**PCA**	**VAE**	**β-VAE**	**SpatialDIVA**	**PCA**	**VAE**	**β-VAE**	**SpatialDIVA**
JEMMING	0.0063	0.2730	0.0481	0.0212	0.0098	0.2400	0.0030	0.0254
SAP	0.0004	0.0000	0.0000	0.0013	0.0008	0.0000	0.0000	0.0023
MIG	0.0005	0.0054	0.0023	0.0020	0.0009	0.0082	0.0025	0.0058
MIG-SUP	0.0012	0.0728	0.0231	0.0019	0.0006	0.0585	0.0181	0.0023
Modularity	0.0051	0.0262	0.0102	0.0077	0.0045	0.0440	0.0311	0.0129
DCI-MIG	0.0044	0.0008	0.0003	0.0166	0.0021	0.0001	0.0001	0.0166
Explicitness	0.0022	0.1582	0.0971	0.0029	0.0042	0.0581	0.0020	0.0047
IRS	0.0014	0.0616	0.0001	0.0016	0.0596	0.0183	0.0000	0.0400

C.2 Results Section 4.2



(a) **(b)** **(c)**

Figure 5: **Batch-associated posterior distribution and invariance in other factors.** High-dimensional samples from the posterior distributions of the batch-associated (a), transcriptomic-associated (b), and morphology-associated (c) latent spaces, reduced using PCA with the two axes of highest variation presented (x, y). The batch labels, based on the slide number of the samples, are overlaid, with the density of the labels shown across the axes.

Table 4: Batch correction evaluation for the SpatialDIVA model. The average batch silhouette width (ASW_{batch}) measures the degree of batch mixing where 0 indicates no batch mixing and 1 indicates perfect batch mixing. Results are shown for 5 random seeds for the conditional variational autoencoder (cVAE) model and the different posterior latent distributions of SpatialDIVA. The best results in each dataset are **bolded**.

Method	Colorectal cancer cohort **batch ASW**	Pancreatic cancer cohort **batch ASW**
cVAE	$\mathbf{1.000 \pm 0.000}$	0.949 ± 0.114
SpDIVA Z_y^1	0.729 ± 0.028	0.704 ± 0.008
SpDIVA Z_y^2	0.610 ± 0.031	0.393 ± 0.084
SpDIVA Z_y^3	0.721 ± 0.009	0.571 ± 0.018
SpDIVA Z_r	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
SpDIVA Z_d	0.200 ± 0.049	0.144 ± 0.020

Table 5: Biological conservation evaluation for the SpatialDIVA model. The average biology conservation score, which is an aggregate of several metrics (Appendix I) is shown. Results are shown for 5 random seeds for the conditional variational autoencoder (cVAE) model and the different posterior latent distributions of SpatialDIVA. The best results in each dataset are **bolded**, and the second-best results are underlined.

Method	Colorectal cancer cohort **biology conservation**	Pancreatic cancer cohort **biology conservation**
cVAE	0.388 ± 0.000	0.359 ± 0.042
SpDIVA Z_y^1	$\mathbf{0.433 \pm 0.006}$	$\mathbf{0.519 \pm 0.011}$
SpDIVA Z_y^2	0.328 ± 0.017	0.296 ± 0.021
SpDIVA Z_y^3	$\underline{0.412 \pm 0.005}$	$\underline{0.462 \pm 0.009}$
SpDIVA Z_r	0.388 ± 0.000	0.378 ± 0.000
SpDIVA Z_d	0.336 ± 0.007	0.388 ± 0.007

C.3 Results Section 4.3

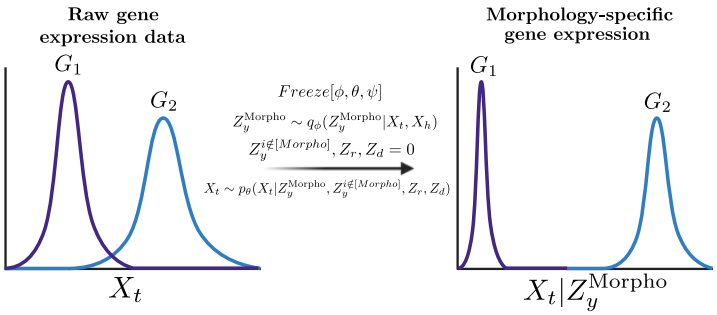
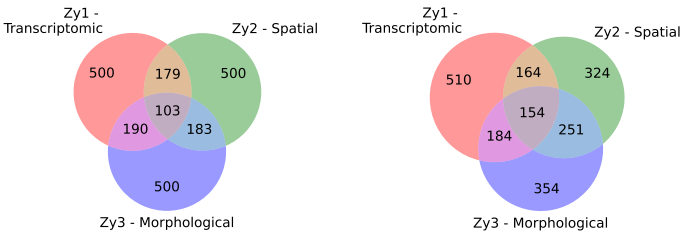


Figure 6: **Conditional generation of transcriptomic data.** The trained model can be frozen and posterior samples for the morphology-specific latent can be used to generate morphology-specific ST gene expression values, as the other factors are held constant (0) during generation.

C.4 Results Section 4.4



(a) **(b)**

Figure 7: **Overlap of DEGs and pathways for conditional generation.** For PDAC, conditional generation of the ST counts (X_t) was performed based on the intrinsic transcriptomic (Z_y^1), spatial (Z_y^2) and morphological (Z_y^3) factors of variation. (a) The overlap of the top 500 DEGs for the generated counts conditioned on each factor. (b) Overlap of the enriched pathways in the top 500 DEGs for each factor.

D DATASETS AND DATASET FORMATTING

The SpatialDIVA model is dependent on having spot-level alignment of ST data and patches from the histology image. Given this constraint, we used data that was either already processed by the HEST-1k pipeline (Jaume et al., 2024), or used their preprocessing scripts to process other paired ST/histology data.

D.1 ST DATA PREPARATION

D.1.1 COLORECTAL CANCER DATA

The spatial transcriptomics data for the colorectal cancer cohort (Valdeolivas et al., 2024) was obtained from the HEST-1k dataset of jointly measured ST and histology data using the Visium platform (Jaume et al., 2024). The HEST-1k uses a specific pipeline for segmentation of tissue and alignment of ST and histology data (Jaume et al., 2024). Given the data as it was procecessed by HEST-1k, we aimed to match the annotations by the authors in the original paper for the ST spot-level cell-type deconvolution and expert pathologist annotations (Valdeolivas et al., 2024), to the data as processed by HEST-1k. We used the barcode information from each ST sample to achieve this, and subset the data from HEST-1k to spots that have both pathologist and ST cell-type annotations from the supplementary information from Valdeolivas et al. (2024). This resulted in 14 tissue sections, corresponding to 14 samples for paired ST and histology measurements.

D.1.2 PANCREATIC CANCER DATA

The pancreatic cancer data (Cui Zhou et al., 2022) was reprocessed and reannotated with updated single-cell references (Peng et al., 2019; Yu et al., 2024).

The dataset consists of 15 sections from 10 patients available on the Human Tumour Atlas Network (HTAN) under the atlas code WUSTL. To estimate the number of cells per spot, we performed nuclear segmentation on the full resolution histology image. Briefly, image crops of each Visium spot were obtained and nuclear segmentation was performed using squidpy (Palla et al., 2022) and stardist (Schmidt et al., 2018) to obtain cell counts per spot. An average cell count per spot was derived for each Visium slide after segmentation of all spots. To infer cell-type proportions for each spot, we utilized cell2location (Kleshchevnikov et al., 2022) and cell-type labels from publicly available single cell adenocarcinoma datasets from Peng et al. (Peng et al., 2019) and Cheng et al (Yu et al., 2024). Briefly, the single cell reference was trained in cell2location using the following parameters for gene selection: (cell_count_cutoff: 5, cell_percentage_cutoff2: 0.03, nonz_mean_cutoff: 1.12, non-mitochondrial genes) and training (num_samples: 1000, batch_size: 2500, num_epochs: 250, batch_key: sample_id). To predict cell-type proportions for each Visium sample, we utilized the average cells per spot from segmentation for each slide and the following parameters (detection_alpha: 20, max_epochs: 30000). Lastly, the 5% quantile of the posterior distribution was utilized as the inferred cell-type abundance per spot and projected onto the histology image for visualization. We used 13 of the 15 sections due to annotation limitations for the histology data (Sec. D.2).

For both the colorectal and pancreatic cancer data ST-labels, which are a proportion of cell-types per spot, we selected the cell-type that had the highest proportion in each spot to use as a categorical label when training and evaluating the SpatialDIVA model.

End-users of the SpatialDIVA method can use their own datasets for ST data, provided there are labels for both the cell-type (ST-derived) and pathologist annotations. In cases where these labels have not yet been derived, unsupervised clustering can be used within each modality to obtain labels.

D.2 HISTOLOGY DATA PREPARATION

Alignment of ST data at the level of spots and the histology data which is split up into patches, is necessary for training the SpatialDIVA model. The Visium data from HEST-1k (Jaume et al., 2024) was processed using an end-to-end pipeline that the authors developed for Visium datasets that performs automatic tissue segmentation, alignment, and resolution detection. The pipeline results in processed ST data with a measurement of gene expression per spot on the ST slide, and histology patches that are centered around each ST spot. The pipeline does this by creating 224x224 px patches

at 20X magnification around each spot. This results in patches of histology images that approximately correspond to each of the ST spots in the slide.

From here, we extract image features for each spot-aligned patch using the UNI foundation model (Chen et al., 2024). The UNI model is loaded via the timm library (Wightman, 2019) and the Huggingface Hub (Wolf et al., 2019). The pretrained model is loaded via the following parameters: {pretrained = True, init_values = 1e-5, dynamic_img_size = True}. Specifically, we used the ViT-L/16 model from the first version of UNI (Chen et al., 2024). Using UNI in inference mode with no further fine-tuning, we obtained 1024 dimensional embeddings for each spot-level patch for both the colorectal cancer and pancreatic cancer datasets.

D.2.1 COLORECTAL CANCER DATA

The colorectal cancer data (Valdeolivas et al., 2024) had already undergone the respective processing using the HEST-1k pipeline to obtain spot-level patch representations of the histology image. As such, there was no further processing necessary and we used these patches and the corresponding ST spots (Sec. D.1) with ST cell-type labels and pathologist annotations from the original study (Valdeolivas et al., 2024) for further analysis. The pathologist annotations were done at the spot-level, so aligning these with the spot-level patches used the same code as aligning cell-type labels.

D.2.2 PANCREATIC CANCER DATA

The pancreatic cancer data (Cui Zhou et al., 2022) was not present in the HEST-1k dataset. As such, we used the HEST-1k pipeline to perform tissue segmentation, upscaling, and patching at the spot-level for this data (Jaume et al., 2024). Specifically, we used the VisiumReader() function from HEST-1k, which loads the histopathology image, feature matrix for ST, and the spatial coordinates. Then we used a built-in function for OTSU segmentation, followed by patching at a size of 224 px. This resulted in patches corresponding to ST spots for the segmented tissue from HEST-1k. The ST data that we had from (Cui Zhou et al., 2022) was inner-joined with the spots that were segmented through the HEST-1k pipeline.

For the samples of the pancreatic cancer data (Cui Zhou et al., 2022), of the 15 samples, 13 of them were annotated by a clinical pathologist in our team, for tumor versus normal regions of the slide. As such, we only used the 13 pathologist-annotated samples from the original study for subsequent analysis. The pathologist annotations for the 13 samples were transferred onto the ST spots through the Shapely library (Gillies et al., 2007).

After processing both datasets, the resulting representations had spot-aligned UNI features (1024 dimensional), pathologist annotations at the spot-level, deconvolved cell-type annotations at the spot-level, as well as expression across all mapped genes, spatial coordinates on each slide for each spot, and other metadata that was used (such as batch/slide number).

E ST AND HISTOLOGY DATA PREPROCESSING

E.1 ST DATA PREPROCESSING

After data preparation, the raw counts from the ST data can be processed further depending on the experiments.

In general, for the ST-derived labels, which comprised of deconvolved cell-type proportions for each spot, we took the cell-type with the highest proportion (representation) for each spot as a soft categorical label. This label was used for model training and evaluation across all results sections.

For evaluation of the SpatialDIVA model versus baselines in quantification of disentanglement (Results Sec. 4.1, Table 1) and the quantitative evaluation of batch-correction effects (cVAE and SpatialDIVA, Results Sec. 4.2, Appendix Table 4), the SpatialDIVA model and baselines were trained after the following preprocessing procedures to the ST counts:

- Count normalization per spot to a fixed value (10000)
- Log1p, equivalent to $ln(x + 1)$, transformation of the counts
- Highly-variable gene selection using the 'seurat' method

Count normalization and Log1p transformations were done for each sample/slide individually, and samples for the complete datasets (pancreatic and colorectal cancer) were concatenated to perform a joint highly-variable gene selection. Joint highly-variable gene selection ensures that cell-types and states across all slides are best captured. These transformations were done using the Scanpy library (Wolf et al., 2018).

For the qualitative disentanglement results in Figs. 3 and 5, the SpatialDIVA model was trained after count and Log1p normalization, but no highly-variable gene selection was done and all genes were used in training the model and evaluating the empirical posterior distributions.

E.2 HISTOLOGY DATA PREPROCESSING

The extracted UNI (Chen et al., 2024) features for the spot-level patches of histopathology image data (Sec. D.2) were standardized at the feature-level (1024 dimensions = 1024 features) for all experiments. Standardization of these features was done for each slide (sample) individually. The reasoning for this was to ensure that batch-effects are not introduced by standardization of the UNI features across samples/slides (Lin & Lu, 2022).

E.3 PREPROCESSING AND EXPERIMENTAL SETUP

A potential challenge in the across-slide selection of highly-variable genes and within-slide standardization of UNI features, is when evaluating disentanglement quantitatively (Sec. 4.1). However, we consider all samples of one cohort to be part of a single observed data distribution, and effectively evaluate samples that are i.i.d. in the disentanglement benchmark. As such, factors such as generalization and data leakage are not considered, because this is a statistical learning problem. This is in line with previous work on evaluating disentanglement (Locatello et al., 2018). The approach of training and evaluating on different subsets of a known data distribution can be interpreted as bootstrapping our estimates of disentanglement performance (Sec. 4.1).

F SPATIALDIVA MODEL DETAILS

F.1 SPATIALDIVA DISTRIBUTIONS

The SpatialDIVA model architecture is that of a neural network with separate encoders for each posterior covariate ($q_\phi(Z_i)$) that is considered in the model, linear decoders that use samples from the posterior distribution to reconstruct the likelihoods for the ST (X_t) and histology (X_h) data distributions ($p_\theta(X_i)$), encoders for the prior distributions ($p_\theta(Z_i)$), and classification heads for the label distributions ($q_\psi(L^i)$). Table 6 summarizes the distributions used in the model, as well as the data required for each.

Table 6: Distributions used in the SpatialDIVA model and details.

Distribution	Specification	Dependency
$p_\theta(X_t \mid Z_y^k, Z_d, Z_r)$	Gaussian/Negative Binomial likelihood	Posterior distributions of Z_y^k, Z_d, Z_r
$p_\theta(X_h \mid Z_y^k, Z_d, Z_r)$	Gaussian likelihood	Posterior distributions of Z_y^k, Z_d, Z_r
$q_\phi(Z_y^1 \mid X_t, X_h)$	Gaussian posterior for transcriptomic variation	Empirical data distributions for X_t and X_h
$q_\phi(Z_y^2 \mid X_t, X_h)$	Gaussian posterior for spatial variation	Empirical data distributions for X_t and X_h
$q_\phi(Z_y^3 \mid X_t, X_h)$	Gaussian posterior for morphological variation	Empirical data distributions for X_t and X_h
$q_\phi(Z_d \mid X_t, X_h)$	Gaussian posterior for batch variation	Empirical data distributions for X_t and X_h
$q_\phi(Z_r \mid X_t, X_h)$	Gaussian posterior for residual variation	Empirical data distributions for X_t and X_h
$p_\theta(Z_y^1 \mid L^1)$	Gaussian prior for transcriptomic variation	Categorical cell-type labels/clusters from ST
$p_\theta(Z_y^2 \mid L^2)$	Gaussian prior for spatial variation	Continuous neighborhood labels (X_t and X_h)
$p_\theta(Z_y^3 \mid L^3)$	Gaussian prior for morphological variation	Categorical pathologist labels/clusters from H&E
$p_\theta(Z_d \mid L^d)$	Gaussian prior for batch variation	Categorical batch, sample, or slide labels
$p_\theta(Z_r)$	Standard normal Gaussian prior for residual variation	N/A
$q_\psi(L^1 \mid Z_y^1)$	Categorical of ST cell-type/cluster labels	Posterior distribution of Z_y^1
$q_\psi(L^2 \mid Z_y^2)$	Gaussian of neighborhood representations (X_t, X_h)	Posterior distribution of Z_y^2
$q_\psi(L^3 \mid Z_y^3)$	Categorical of pathologist/cluster H&E labels	Posterior distribution of Z_y^3
$q_\psi(L^d \mid Z_d)$	Categorical of batch labels	Posterior distribution of Z_d

F.2 SPATIALDIVA ARCHITECTURE

For each of the given distributions outlined in the previous section, Tables 7, 8, 9, 10, and 11 summarize the SpatialDIVA architecture choices for the different experiments in the paper. Exceptions based on experiments are noted after the tables.

Table 7: Architecture for data likelihood distribution for ST - $p_\theta(X_t \mid Z_y^k, Z_d, Z_r)$

Module number	Component
1	nn.Linear(100, 64)
2	mu = Softplus(nn.Linear(64, 36601))
3	logvar = nn.Linear(64, 36601)

Table 8: Architecture for data likelihood distribution for histology - $p_\theta(X_h \mid Z_y^k, Z_d, Z_r)$

Module number	Component
1	Linear(100, 64)
2	mu = Linear(64, 1024)
3	logvar = Linear(64, 1024)

Table 9: Architecture for posterior distributions - $q_\phi(Z_i|X_t, X_h)$

Module number	Component
1	Linear(37625, 64)
2	BatchNorm1D(64)
3	ReLU()
4	mu_zy = Linear(64, 20)
5	logvar_zy = Linear(64, 20)

Table 10: Architecture for prior distributions - $p_\theta(Z_i|L^i)$. **Variable** here indicates the variable input length for the prior labels, which depends on the posterior/label combination and dataset.

Module number	Component
1	Linear(**Variable**, 64)
2	ReLU()
3	mu_zy = Linear(64, 20)
4	logvar_zy = Linear(64, 20)

Table 11: Architecture for label distributions - $q_\psi(L^i|Z_i)$. **Variable** here indicates the variable output length for the prior labels, which depends on the posterior/label combination and dataset.

Module number	Component
1	ReLU()
2	Linear(20, **Variable**)

In terms of **exceptions** to the architectures outlined in Tables 7, 8, 9, 10, and 11:

- For the experiments benchmarking disentanglement (Table 1) and batch-correction (Appendix Table 4), the input size for ST counts (X_t) was 2500 (instead of 36601), as highly-variable gene selection was done for these experiments (changes Table 7 output size, Table 9 input size)
- For the latent space covariance analysis, the dimensionality of output for the posterior of batch variability ($q_\phi(Z_d|X_t, X_h)$) was changed from 20 to 5 (Table 9). Also for this analysis, two hidden layers were used for the ST and histology data distributions ($p_\theta(X_t|..)$, $p_\theta(X_h|..)$) of sizes 256 followed by 128 (Tables 7 and 8). Lastly, an extra hidden layer was used for the encoders of the posterior distributions ($q_\phi(Z_i|X_t, X_h)$), of size 32 (two sequential hidden layers, size 64 and 32 with ReLU() non-linearities after each) (Table 9)
- For the conditional generation experiments (Results Sec. 4.3 and 4.4), the likelihood distribution for ST (Table 7) outputs θ and μ to parametrize a Negative Binomial distribution, which are both constrained to be positive via Softplus, instead of μ and logvar for a Gaussian distribution. Further, the dimensionality of output for the posterior of batch variability ($q_\phi(Z_d|X_t, X_h)$) was changed from 20 to 5 (Table 9). Two hidden layers were used for the ST and histology data distributions ($p_\theta(X_t|..)$, $p_\theta(X_h|..)$) of sizes 256 followed by 128 (Tables 7 and 8). Lastly, an extra hidden layer was used for the encoders of the posterior distributions ($q_\phi(Z_i|X_t, X_h)$), of size 32 (two sequential hidden layers, size 64 and 32 with ReLU() non-linearities after each) (Table 9).

G SPATIALDIVA OBJECTIVE DERIVATION

For the SpatialDIVA model (Fig. 2), we want to maximize the following marginal likelihood of the ST counts (X_t), histology features (X_h) and observed labels (L):

$$p_\theta(X_t, X_h, L) = \int p_\theta(X_t|Z)p_\theta(X_h|Z)p_\theta(Z|L)p(L)dz$$

However, this marginalization over all possibilities of Z is intractable. We can instead learn a lower bound on the log-likelihood:

$$\begin{aligned}
\log\left(p_\theta(X_t, X_h, L)\right) &= \log \int p_\theta(X_t|Z)p_\theta(X_h|Z)p_\theta(Z|L)p(L)dz \\
&= \log \int p_\theta(X_t|Z)p_\theta(X_h|Z)p_\theta(Z|L)p(L)\frac{q(Z|X_t, X_h)}{q(Z|X_t, X_h)}dz \\
&= \mathbb{E}_{q(Z|X_t,X_h)}[\log p(X_t|Z) + \log p(X_h|Z) + \log p(Z|L) \\
&\quad + \cancel{\log p(L)}^{\text{not dependent on } q(Z|.)} - \log q(Z|X_t, X_h)]
\end{aligned}$$

Using Jensen's inequality, which indicates that $\log(\mathbb{E}(..)) \geq \mathbb{E}[\log(..)]$:

$$\begin{aligned}
\log\left(p_\theta(X_t, X_h, L)\right) &\geq \mathbb{E}_{q(Z|X_t,X_h)}[\log p(X_t|Z) + \log p(X_h|Z) + \log p(Z|L) - \log q(Z|X_t, X_h)] \\
&\geq \mathbb{E}_{q(Z|X_t,X_h)}[\log p(X_t|Z)] + \mathbb{E}_{q(Z|X_t,X_h)}[\log p(X_h|Z)] \\
&\quad + \mathbb{E}_{q(Z|X_t,X_h)}[\log p(Z|L) - \log p(Z|X_t, X_h)]
\end{aligned}$$

We have two types of label distributions - L^k for k biologically informative prior labels, L^d for the batch/sample. These correspond to m Z_y^k posterior distributions, a Z_d posterior. There is also a posterior for residual variation in Z_r. We can break down this bound further based on these distinct factors, without explicitly summing over the k posteriors and labels for biologically informative labels. The posterior for residual variation (Z_r) is not constrained by a label distribution.

$$\begin{aligned}
&\geq \mathbb{E}_{q(Z_y^k,Z_d,Z_r|X_t,X_h)}[\log p(X_t|Z_y^k, Z_d, Z_r)] \\
&\quad + \mathbb{E}_{q(Z_y^k,Z_d,Z_r|X_t,X_h)}[\log p(X_h|Z_y^k, Z_d, Z_r)] \\
&\quad + \mathbb{E}_{q(Z_y^k,Z_d,Z_r|X_t,X_h)}[\log p(Z_y^k|L^k) + \log p(Z_d|L^d) + \log p(Z_r) \\
&\quad - \log q(Z_y^k|X_t, X_h) - \log q(Z_d|X_t, X_h) - \log q(Z_r|X_t, X_h)] \\
&\geq \mathbb{E}_{q(Z_y^k,Z_d,Z_r|X_t,X_h)}[\log p(X_t|Z_y^k, Z_d, Z_r)] \\
&\quad + \mathbb{E}_{q(Z_y^k,Z_d,Z_r|X_t,X_h)}[\log p(X_h|Z_y^k, Z_d, Z_r)] \\
&\quad + \mathbb{E}_{q(Z_y^k|X_t,X_h)}[\log p(Z_y^k|L^k) - \log q(Z_y^k|X_t, X_h)] \\
&\quad + \mathbb{E}_{q(Z_d|X_t,X_h)}[\log p(Z_d|L^d) - \log q(Z_d|X_t, X_h)] \\
&\quad + \mathbb{E}_{q(Z_r|X_t,X_h)}[\log p(Z_r) - \log q(Z_r|X_t, X_h)]
\end{aligned}$$

Through the definition of the KL-divergence:

$$\begin{aligned}
&\geq \mathbb{E}_{q(Z_y^k,Z_d,Z_r|X_t,X_h)}[\log p(X_t|Z_y^k, Z_d, Z_r)] \\
&\quad + \mathbb{E}_{q(Z_y^k,Z_d,Z_r|X_t,X_h)}[\log p(X_h|Z_y^k, Z_d, Z_r)] \\
&\quad - D_{KL}(q(Z_y^k|X_t, X_h) \parallel p(Z_y^k|L^k)) - D_{KL}(q(Z_d|X_t, X_h) \parallel p(Z_d|L^k)) \\
&\quad - D_{KL}(q(Z_r|X_t, X_h) \parallel p(Z_r))
\end{aligned}$$

The first two expectations correspond to the likelihoods of the ST (X_t) and histology (X_h) data. The last three KL divergence terms penalize the learned posterior distributions based on the learned prior distributions. The exception is $q(Z_r|X_t, X_h)$, which is penalized based on a standard normal Gaussian prior $p(Z_r)$. We can decompose the KL divergences for all of the Z_y terms we consider, including intrinsic transcriptomic variation Z_y^1, spatial variation Z_y^2, and morphological variation Z_y^3:

$$
\begin{aligned}
&\geq \mathbb{E}_{q(Z_y^k, Z_d, Z_r|X_t, X_h)}[\log p(X_t|Z_y^k, Z_d, Z_r)] \\
&\quad + \mathbb{E}_{q(Z_y^k, Z_d, Z_r|X_t, X_h)}[\log p(X_h|Z_y^k, Z_d, Z_r)] \\
&\quad - \sum_{k=1}^{3}\Big(D_{KL}(q(Z_y^k|X_t, X_h) \parallel p(Z_y^k|L^k))\Big) \\
&\quad - D_{KL}(q(Z_d|X_t, X_h) \parallel p(Z_d|L^k)) \\
&\quad - D_{KL}(q(Z_r|X_t, X_h) \parallel p(Z_r))
\end{aligned}
$$

Similar to ccVAE (Joy et al., 2020) and DIVA (Ilse et al., 2019), we incorporate classification losses for the posterior samples based on the labeled data (Appendix 6):

$$
\begin{aligned}
&\geq \mathbb{E}_{q(Z_y^k, Z_d, Z_r|X_t, X_h)}[\log p(X_t|Z_y^k, Z_d, Z_r)] \\
&\quad + \mathbb{E}_{q(Z_y^k, Z_d, Z_r|X_t, X_h)}[\log p(X_h|Z_y^k, Z_d, Z_r)] \\
&\quad - \sum_{k=1}^{3}\Big(D_{KL}(q(Z_y^k|X_t, X_h) \parallel p(Z_y^k|L^k))\Big) \\
&\quad - D_{KL}(q(Z_d|X_t, X_h) \parallel p(Z_d|L^k)) \\
&\quad - D_{KL}(q(Z_r|X_t, X_h) \parallel p(Z_r)) + \sum_{k=1}^{m}\mathbb{E}_{q_\phi(Z_y^k|X_t, X_h)}[\log q_\psi(L_y^k|Z_y^k)] \\
&\quad + \mathbb{E}_{q_\phi(Z_d|X_t, X_h)}[\log q_\psi(L^d|Z_d)]
\end{aligned}
$$

This leads to the following objective, with β's corresponding to hyperparameters. We minimize the loss, and hence the signs flip:

$$
\begin{aligned}
-L(\theta, \phi, \psi) = &-\beta_1\mathbb{E}_{q(Z_y^k, Z_d, Z_r|X_t, X_h)}[\log p(X_t|Z_y^k, Z_d, Z_r)] \\
&- \beta_2\mathbb{E}_{q(Z_y^k, Z_d, Z_r|X_t, X_h)}[\log p(X_h|Z_y^k, Z_d, Z_r)] \\
&+ \beta_3\sum_{k=1}^{3}\Big(D_{KL}(q(Z_y^k|X_t, X_h) \parallel p(Z_y^k|L^k))\Big) \\
&+ \beta_4 D_{KL}(q(Z_d|X_t, X_h) \parallel p(Z_d|L^k)) \\
&+ \beta_5 D_{KL}(q(Z_r|X_t, X_h) \parallel p(Z_r)) \\
&- \beta_6\sum_{k=1}^{m}\mathbb{E}_{q_\phi(Z_y^k|X_t, X_h)}[\log q_\psi(L_y^k|Z_y^k)] \\
&- \beta_7\mathbb{E}_{q_\phi(Z_d|X_t, X_h)}[\log q_\psi(L^d|Z_d)]
\end{aligned}
$$

The parameterizations of θ, ϕ, and ψ for the different distributions are outlined in Appendix F.

In general, across experiments, we use hyperparameters where all β's are set to 1. Exceptions are explicitly indicated in Appendix I.

H BIOLOGICAL FACTORS OF VARIATION IN ST AND HISTOLOGY

For the key biological factors of variation that we considered in our analysis that are initially described in Appendix B.1, we provide more detail on the rationale behind their selection, framing in the problem, and preprocessing below.

H.1 INTRINSIC TRANSCRIPTOMIC VARIATION - Z_y^1

Intrinsic variation of gene expression takes into account the biological processes that are not strongly influenced by cellular organization and cell-cell communication. Within a tissue context there is a large-scale orchestration of cell-cell communication that occurs via signalling, both short and long-range (Alberts et al., 2022). Even longer range signals can arrive from entirely different organs and parts of the body that influence the transcriptomic state of individual cells, for functional effects such as differentiation (Alberts et al., 2022). Variation in gene expression that is intrinsic, therefore, should capture effects and functional changes to gene expression that are dictated by the internal state of a cell. An example of this could be immune evasion functions governed by gene expression changes that are activated in cancer cells due to mutations (Hanahan, 2022).

We consider intrinsic transcriptomic variation to be an important factor, as this variation will affect both the transcriptomic counts (X_t) in a trivial manner, and the histology features (X_h) due to the relationship between transcription and protein expression (more details provided in the Morphology section). A potential challenge is determining how to isolate intrinsic versus extrinsic transcriptomic variation, where extrinsic variation is due to factors such as cell-cell communication. From a supervision perspective, the labels we utilize, whether they are clusters or expert-annotated cell-types derived from the ST data, will also likely be influenced by extrinsic variation as this will affect the ST counts (X_t) as well. In SpatialDIVA, we aim to remove the extrinsic variation captured in the posterior distribution for Z_y^1 by introducing a spatial variation covariate (Z_y^2). Although we constrain intrinsic transcriptomic variation on labels that may contain information from both intrinsic and extrinsic sources, by encouraging disentanglement through the Z_y^2 posterior, we aim to remove as much extrinsic variation from the posterior distribution of Z_y^1 as possible.

In our analysis, we use expert-annotated cell-type labels derived from the ST counts, for both the colorectal and pancreatic cancer datasets (L^1). These labels are derived using a single-cell reference, as the Visium ST protocol does not yield single-cells per spot but a mixture of cells (Ståhl et al., 2016). The process is referred to as cell-type deconvolution, and returns a proportion of cell-types estimated to be in each spot of the ST slide (Ma & Zhou, 2022). As we use a categorical distribution of labels to constrain our posterior for intrinsic transcriptomic variation (Appendix 6), we take the cell-type that has the *highest proportion in each spot as a label*, and we consider the entire set of cell-type labels derived in this manner as a categorical distribution. This was done for both the colorectal cancer data and re-annotated pancreatic cancer datasets.

H.2 SPATIAL VARIATION - Z_y^2

Spatial variation considers the influence of extracellular signalling as well as more direct means of cell-cell communication, such as at cellular junctions (Alberts et al., 2022). Spatial context is relevant across many biological scenarios, including development and cancer. Within development, cellular signalling gradients which are organized spatially, dictate how certain cells differentiate and what cells and tissues they will give rise to (Barresi & Gilbert, 2023). In cancer, as we highlight in the case of PDAC, spatial context is important for characterization of a tumor as well as potential diagnostic and therapeutic avenues (Karamitopoulou, 2019; Cui Zhou et al., 2022). We consider both intrinsic transcriptomic and morphological variation as separate covariates, and therefore, the spatial variation posterior (Z_y^2) should aim to capture spatial variation that will affect both the ST (X_t) and histology profiles (X_h). Therefore, for our spatial context covariate (Eqns. 10, 11), we consider the concatenation of reduced features for both the transcriptomic (X_t) and histology data (X_h).

Essentially, through our neighborhood decomposition, we aim to represent a label distribution (L^2) that captures the *context* of all the cells in each spot i, by having a representation that is predictive of the X_t and X_h features of the spatial neighbors of i. These neighbors are defined by locality, and we are using the spatial coordinates available in the data ($P_{x,y}$) to create this predictive representation. If the prior and posterior distributions for Z_y^2 retain this context, and if disentanglement in the model is

working appropriately, the spatial context should be captured by Z_y^2, while being minimized in all of the other posterior distributions.

H.3 MORPHOLOGICAL VARIATION - Z_y^3

The last factor of variation that we considered was morphological variation. Cellular morphology is dictated primarily by protein expression, cellular shape, and organization of subcellular organelles like the nucleus (Alberts et al., 2022). Morphological features have a long history of being utilized to differentiate the state of cells, such as the use of H&E staining in histopathology to differentiate cancer cells from normal cells (He et al., 2012). Morphological features are distinguished through the histology aspect of the multi-modal data we consider (X_h), and therefore have a direct effect on this modality, much like intrinsic transcriptomic variation has a direct effect on the transcriptomic counts per spot (X_t). However, as described in Appendix B.1, both morphological and intrinsic transcriptomic variation can affect both the transcriptomic (X_t) and histology (X_h) readouts.

Further, we have observed a direct case where transcriptomic variation is correlated with morphological features, in that intrinsic transcriptomic variation (Z_y^1) was found to be predictive of a histopathology derived label (tumor/normal) in PDAC (Fig. 3).

For labels that constrain the morphology posterior and prior distributions (L^3), we used pathologist annotations of the histopathology slides based on regions. For the colorectal cancer data (Valdeolivas et al., 2024), this comprised of annotations for distinctive tumor, stromal, and normal regions of cells stained by H&E. For the pancreatic cancer data (Cui Zhou et al., 2022), the slides were annotated by a pathologist on our team, delineating tumor and normal epithelial regions. As such, we had a categorical distribution of labels for the colorectal cancer data, and a binary distribution of labels for the pancreatic cancer data. These labels were derived exclusively through the histology features, and they should thus be significantly correlated with the morphological variation we aimed to capture in this covariate (Z_y^3).

I EXPERIMENT DETAILS AND CONFIGURATIONS

This appendix section provides details on the experiment settings, as well as baselines that were used and their setup. The different experiments and their details are contained in their own sections.

I.1 RESULTS SEC. 4.1

I.1.1 QUANTITATIVE DISENTANGLEMENT BENCHMARK

For the quantitative evaluation of disentanglement, we used a modified version of SpatialDIVA with highly-variable genes (Sec. F.2). For both the colorectal and pancreatic cancer datasets, preprocessing was done in a uniform manner for all baselines and the SpatialDIVA method:

- ST count (X_t) normalization to 10000 counts per spot
- Log1p transformation per spot-gene pair across all spots and genes
- Highly-variable gene selection using all slides in a cohort (2500 genes)
- Standardization of UNI features (1024) for the histology (X_h) data, per slide

For both the colorectal cancer and pancreatic cancer datasets, we sampled 90% of the spots from the combined slides and held-out 10% for evaluation of disentanglement. These splits were done randomly using the numpy (Harris et al., 2020) library, and 10 random seeds.

The **baselines** for disentanglement benchmarking were set up as follows:

PCA: For principal component analysis, the transcriptomic counts (X_t) and standardized UNI histology features (X_h) were concatenated and the transcriptomic counts were also standardized. A PCA reduction (Pedregosa et al., 2011) was done on the concatenated representation, and the first 20 dimensions were used as embeddings for disentanglement quantification.

VAE: For a base variational autoencoder, we considered a VAE (Kingma & Welling, 2013) with a 20 dimensional latent posterior ($q_\phi(Z|X_t, X_h)$), with a 64 dimensional hidden layer for the encoder and decoder. ReLU activations were done after the hidden layers, but not before the likelihood and posterior mean and logvar output steps. Similar to SpatialDIVA, a Gaussian posterior and likelihood were used.

β-VAE: The configuration for the β-VAE and VAE were the exact same, the only difference was that the weight on the KL divergence term for the posterior distribution was increased to 100, as outlined in Higgins et al. (Higgins et al., 2016):

$$\mathbf{1} * \mathbb{E}_{q_\phi(Z|X_t,X_h)}[\log p_\theta(X_t, X_h|Z)] + \mathbf{100} * D_{KL}(q_\phi(Z|X_t, X_h) \parallel p(Z))$$

SpatialDIVA used the configuration outlined previously, with a Gaussian likelihood for the log-normalized transcriptomic counts (X_t) as well as the standardized UNI histology features (X_h). The trained SpatialDIVA model is used on the test data (frozen model) to extract the mean parameters (μ_i) from the following posterior distributions: $q_\phi(Z_y^1|X_t, X_h)$, $q_\phi(Z_y^3|X_t, X_h)$, $q_\phi(Z_d|X_t, X_h)$. The means from these distributions were then concatenated to obtain test embeddings.

The **PCA** baseline was trained directly on the training set for each iteration using a singular value decomposition (SVD) solver (Pedregosa et al., 2011). The test data is then projected onto the principal components (n=20) calculated via the training data. This 20 dimensional embedding is used for further testing.

The **VAE** and **β-VAE** baselines were trained on each training subset and the test embeddings from the latent spaces ($q_\phi(Z_d|X_t, X_h)$) were extracted after the models were frozen. The 20 dimensional means μ_i of the latent embeddings were used for further testing.

The **VAE, β-VAE, and SpatialDIVA** models were trained using the Adam optimizer (Kingma & Ba, 2014) at a learning rate of 0.001 for 100 epochs and batch size of 64.

The extracted embeddings for **SpatialDIVA** and each of the **baseline** methods were used to compute the disentanglement metrics (Appendix J). The factors used in the metrics comprised of the

cell-type labels (from the ST data), pathologist annotations from the histology, and batch labels. Essentially, these metrics measure how well the latent spaces of the baselines, as well as the SpatialDIVA method, capture variance in ST-labelled cell-types, pathologist-annotated histology data, and batch/technical effects. These factors were one-hot encoded and were used with the embeddings extracted from each method to quantify the disentanglement scores across metrics (Appendix J).

The assessment of disentanglement for this setup was done for both the pancreatic and colorectal cancer datasets (Table 1).

Using the metric values, the methods were *ranked* based on performance, where a 1 indicated the best rank in a dataset for a given metric and 4 indicated the worst rank. Ranks were added up across metrics for the methods and the method with the lowest aggregate score was the best and was ranked first overall, and the other methods followed and were ranked using the same aggregation procedure.

I.2 Results Sec. 4.2

I.2.1 Disentangled latent space and covariance analysis

To assess the latent spaces of the posterior distributions, the SpatialDIVA model was trained on all of the genes quantified in the ST (X_t) data for the pancreatic cancer cohort, with all slides used for training the model. In terms of architecture, the architecture indicated in Sec. F was used with the indicated changes (Sec. F.2).

The following preprocessing steps were done for the PDAC data before training:

- ST count (X_t) normalization to 10000 counts per spot
- Log1p transformation per spot-gene pair across all spots and genes
- Standardization of UNI features (1024) for the histology (X_h) data, per slide

The model was trained with a batch size of 256, for 50 epochs with the Adam (Kingma & Ba, 2014) optimizer at a learning rate of 0.001.

After training the model with all of the PDAC data, embeddings were extracted for the posterior distributions of the model: $q_\phi(Z_y^1|X_t, X_h)$, $q_\phi(Z_y^2|X_t, X_h)$, $q_\phi(Z_y^3|X_t, X_h)$, $q_\phi(Z_d|X_t, X_h)$, $q_\phi(Z_r|X_t, X_h)$. In this case, the means were not only used, and distributions for each posterior were sampled once per datapoint. For example, for Z_y^1 and sample s for spot i in the observed data:

$$\begin{aligned} \mu_i &\sim q_\phi(Z_y^1|X_t, X_h) \\ \sigma_i &\sim q_\phi(Z_y^1|X_t, X_h) \\ \epsilon &\sim N(0, 1) \\ s_i &= \mu_i + \epsilon * \sigma_i \end{aligned}$$

The samples for these posterior distributions were high dimensional (all 20 except for Z_d which was 5). Therefore, for visualization, these were reduced to 2-dimensions using PCA (Pedregosa et al., 2011), and the two axes of highest variation were visualized (Fig. 3 and Appendix 5).

The available cell-type labels derived from the ST data, pathologist annotations, and batch labels were overlaid on the posterior samples for each spot in the data, for the respective plots.

For this experiment, the SpatialDIVA model we trained had modified ELBO hyperparameters (β's as defined in Appendix G). Specifically, β_1 and β_2, corresponding to the likelihood expectations, were set to 100. All other β's remained 1.

I.2.2 Multi-modal batch correction benchmark

The batch-correction benchmark used a conditional variational autoencoder (cVAE) (Sohn et al., 2015) as a baseline, which conditioned on batch/slide label during training.

We trained SpatialDIVA and the cVAE using the exact same architecture and training setup as the quantitative disentanglement experiments (Sec. I.1.1). The key difference is that we performed this analysis for 5 iterations and in each iteration we used the entire dataset (colorectal or pancreatic

cancer) for training and then subsequently evaluated batch-correction by freezing the network and getting embeddings for all datapoints. Training on the full dataset and obtaining batch-corrected embeddings is standard practice (Tran et al., 2020; Luecken et al., 2022).

Randomization in this case corresponded to different torch seeds for initialization of the models to capture training variability. The architecture of the cVAE model is the same as that of the VAE model in Sec. I.1.1, other than an added input of one-hot encoded batch-labels. This is done to reflect the supervision for batch-correction available to the SpatialDIVA model.

For evaluation, we extracted the 20 dimensional posterior means from the cVAE latent space ($q_\phi(Z|X_t, X_h, L^d)$), and the following 20 dimensional posterior means from SpatialDIVA: $q_\phi(Z_y^1|X_t, X_h)$, $q_\phi(Z_y^3|X_t, X_h)$, $q_\phi(Z_d|X_t, X_h)$, $q_\phi(Z_r|X_t, X_h)$.

The posterior means from each latent outlined from SpatialDIVA was used in the analysis, as well as the posterior means from the cVAE. We used the the scib-metrics package for benchmarking batch correction (Luecken et al., 2022). Specifically, we used the average silhouette width (ASW) (Luecken et al., 2022), which measures how well the batches mix across cells in given cell-types, and averages this value across cell-types. For each cell-type label C_j with n total cells:

$$batch\, ASW_j = \frac{1}{n} \sum_{i \in C_j} 1 - |silhouette(i)|$$

This value effectively measures how well the batches are mixed in the embeddings for a given cell-type C_j, where 1 indicates perfect mixing and 0 indicates the most suboptimal mixing of batches possible. After obtaining this quantity per cell-type, we can average across M cell-types:

$$average\, batch\, ASW = \frac{1}{M} \sum_{i \in M} batch\, ASW_i$$

For cell-type labels, we combined the ST-celltypes and pathologist annotations for both the colorectal and pancreatic cancer datasets, into one string value which was then encoded for use with this metric. We calculated this metric for each of the latent subspaces indicated from SpatialDIVA and the latent subspace from cVAE. The calculation was done for 5 iterations for both cVAE and SpatialDIVA, as indicated.

I.2.3 MULTI-MODAL BIOLOGY CONSERVATION BENCHMARK

For the results in Table 5, we used the exact same pipeline as in Sec. I.2.2, but instead of using the ASW metric, we used an aggregated score that measures how well the embeddings preserve biological signal with respect to the combined pathologist annotation and ST celltype labels. The metrics we used were also from the scib-metrics package (Luecken et al., 2022), and included the following bio-conservation metrics:

- Isolated biology label score
- Cell-type local inverse simpson index (cLISI)
- NMI using k-means clusters and biology labels
- ARI using k-means clusters and biology labels
- Biology label ASW

Full details on all of these metrics can be found in the scib-metrics documentation and the original scib publication (Luecken et al., 2022).

The values of these metrics for cVAE and SpatialDIVA were averaged per iteration to determine the biology conservation score.

I.3 RESULTS SEC. 4.4

I.3.1 CONDITIONAL MULTI-MODAL GENERATION OF PDAC

For this analysis, the PDAC data was processed using the following steps:

- Standardization of UNI features (1024) for the histology (X_h) data, per slide

The model was trained on all of the genes, hence no highly-variable gene selection. Further, normalization and log transformation of the counts was not done as we considered a negative binomial likelihood for the ST data (X_t), which works best with untransformed transcriptomic counts (Lopez et al., 2018).

The SpatialDIVA model was set up based on the changes to the default architecture as indicated in Sec. F.2. A negative binomial parametrization of the ST (X_t) likelihood allowed for resampling of the counts after training the model, based on conditioning of certain covariates. We maximized a Negative Binomial likelihood, with the formulation taken from the scVI model (Lopez et al., 2018; Gayoso et al., 2022). The same formulation was used for subsequent sampling of Negative Binomial distributions using the obtained parameters.

The model was trained on all of the PDAC data, with a batch size of 256, the Adam optimizer (Kingma & Ba, 2014) at a learning rate of 0.001, and for 50 epochs.

After training, conditional multi-modal generation was done, as outlined in Sec. 4.3, after conditioning on transcriptomic (Z_y^1), spatial (Z_y^2) and morphological variation (Z_y^3), while holding other factors constant. This process was the same across all 3 covariates. As an example, for the intrinsic transcriptomic-conditioned generation for spot i:

Algorithm 1 Conditional generation of $X_t|Z_y^1$

Input: N spots with X_t and X_h features, frozen (θ, ϕ, ψ) SpatialDIVA model
for $Spot_i$ in range(N) **do**
  $\mu_i, \sigma_i \sim q_\phi(Z_{yi}^1|X_{ti}, X_{hi})$
  $\epsilon \sim N(0,1)$
  $S_i = \mu_i + \sigma_i * \epsilon$
  $\theta_i, \mu_i \sim p_\theta(X_{ti}|Z_y^1 = S_i, Z_y^2 = 0, Z_y^3 = 0, Z_d = 0, Z_r = 0)$
  $\hat{X_{ti}} \sim \text{NB}_i(\theta_i, \frac{\mu_i}{\theta_i})$
end for

This conditional generation was done by conditioning on the three indicated covariates, for 10000 (N) randomly selected spots across the PDAC data. This resulted in three distributions of the transcriptomic counts, conditioned on each factor.

For each of these three distributions [$(\hat{X}_t|Z_y^1)$, $(\hat{X}_t|Z_y^2)$, $(\hat{X}_t|Z_y^3)$], we performed the following processing steps to obtain the top 500 differentially expressed genes (DEGs) (Wolf et al., 2018):

- ST count ($\hat{X}_t$) normalization to 10000 counts per spot
- Log1p transformation per spot-gene pair across all spots and genes
- Highly-variable gene selection using all 10000 spots (2500 genes)
- Count standardization and principal component reduction for the top 50 PCs
- Nearest-neighbor graph construction using the PCA embedding
- Leiden clustering of the data using the nearest-neighbor graph
- Differential gene expression across Leiden-derived clusters using the Wilcoxon rank-sum test

The scanpy (Wolf et al., 2018) library was used for these steps (v1.10.0). Default parameters were used, except where indicated. After determining the DEGs for each of the conditionally generated distributions, the top 500 were selected based on those exhibiting the highest Log-fold change

in expression between clusters (Wolf et al., 2018). An all-versus one differential expression test was done, meaning that the expression in each cluster is compared with all of the other clusters to determine DEGs. Therefore, sorting by the highest Log-fold change can be interpreted as sorting genes that exhibit the highest specificity to their respective clusters in the reconstructed counts.

Using the top 500 DEGs from each of the conditionally generated count distributions [($\hat{X}_t|Z_y^1$), ($\hat{X}_t|Z_y^2$), ($\hat{X}_t|Z_y^3$)], we used the gProfiler online platform (Kolberg et al., 2023) to perform pathway enrichment analysis. Default parameters were used, except for the following changes:

- 'Ordered query' was used
- The pathway databases were subset to GO molecular function, GO biological process, and REACTOME

Ordered query was utilized as the top 500 DEGs are ordered, and the subsetting of pathway databases was done to ensure that outdated and uncurated databases did not affect the analysis (Wadi et al., 2016).

The resulting enriched pathways were further analyzed for overlap. Pathways were sorted based on multiple-testing corrected p-values (Kolberg et al., 2023). The functional grouping of the pathways (Fig. 4a,b) was done based on presence of any of the following keywords:

- **Pancreas and metabolic functions**: keywords = ("pancreas", "pancreatic", "islet", "beta cell", "alpha cell", "acinar cell", "ductal cell", "exocrine", "endocrine", "insulin", "glucagon", "somatostatin", "delta cell", "langerhans", "pdac", "secretin", "cck", "cholecystokinin", "metabolic", "catabolic", "biosynthetic", "oxidation", "mito", "glycolysis", "glucose", "pyruvate", "metabolism")
- **Immune functions**: keywords = ("immune", "immuno", "immune-mediated", "immunity", "inflammation", "inflammatory", "inflammasome", "antigen", "antigen presentation", "mhc", "hla", "immunoglobulin", "antibody", "adaptive immunity", "innate immunity", "t cell", "t lymphocyte", "b cell", "b lymphocyte", "nk cell", "natural killer", "tcr", "bcr", "chemokine", "chemokine receptor", "cytokine", "interleukin", "il-", "ifn", "interferon", "complement", "fc receptor", "immunoregulation", "immunosuppression", "lymphocyte", "dendritic cell", "macrophage", "monocyte", "phagocytosis", "phagosome", "influenza", "viral)
- **Translation functions**: keywords = ("ribosome", "ribosomal", "rRNA", "ribosomal subunit", "polysome", "polyribosome", "translation", "translational", "protein biosynthesis", "elongation factor", "peptide chain", "tRNA","rna processing")
- **Antigen presentation functions**: keywords = ("antigen", "antigen presentation", "mhc", "hla")

J METRICS TO ASSESS DISENTANGLEMENT IN A CONTINUOUS SPACE

In this section, we provide a detailed explanation of the metrics outlined in Table 1. These metrics and the disentanglement evaluation framework are adapted from Carbonneau et al. (2020).

We define a set of N observations as $X = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$. Each observation is assumed to be fully determined by a set of M factors $\mathcal{V} = \{v_1, v_2, ..., v_M\}$ through a generative process $g(\mathbf{v}) \mapsto \mathbf{x}$. Let $V = \{\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_N\}$ represent the factor realizations that generate X. A representation learning algorithm maps $r(\mathbf{x}) \mapsto \mathbf{z}$, where $\mathbf{z} \in \mathbb{R}^d$ is a point in the learned latent space $\mathcal{Z} = \{z_1, z_2, ..., z_d\}$. The set $Z = \{\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_N\}$ contains all points in X projected into the latent space by $r(.)$. The disentanglement metrics evaluate the relationship between V and Z to compute a disentanglement score.

In our setup, we consider Z to be the continuous learned latent space from embeddings of the baseline models and SpatialDIVA, and the one-hot encoded labels from the data (Appendix I) to be $\mathcal{V}$.

The details of the metrics, originally outlined in Carbonneau et al. (2020), are indicated below. The terms latent space and latent codes are used interchangeably for $Z = \{\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_N\}$.

J.1 EXPLICITNESS SCORE

As proposed in Ridgeway & Mozer (2018), explicitness is measured by training a classifier on the entire latent space to predict factor classes, assuming discrete factor values. A simple logistic regression classifier is employed, and its performance is evaluated using the area under the ROC curve (AUC-ROC). The final explicitness score is the average AUC-ROC across all classes and factors. Since the minimum AUC-ROC value is 0.5, the scores are normalized to fall within the range [0, 1]. The logistic regression loss is adjusted due to class imbalance.

J.2 ATTRIBUTE PREDICTABILITY SCORE (SAP)

SAP (Kumar et al., 2018) assigns a score S_{ij} for every factor-code pair (v_i, z_j). For categorical factors, a decision tree classifier is used, and balanced accuracy is returned. Scores for codes below a user-defined energy threshold (*dead-codes*) are set to 0. The complete SAP score is calculated as the average difference between the two highest scores S_{ij} for each factor:

$$\text{SAP} = \frac{1}{M} \sum_{i=1}^{M} (S_{i\star} - S_{i\circ}) \tag{13}$$

Here, $S_{i\star}$ and $S_{i\circ}$ are the highest and second-highest scores for factor v_i, respectively. Large differences indicate better disentanglement.

J.3 MODULARITY SCORE

Modularity measures whether each code dimension z_j is associated with only one factor. Following Ridgeway & Mozer (2018), the factor $v_\star$ with the highest mutual information (MI) for each code dimension is identified, and the MI values with other factors are penalized:

$$\text{modularity} = 1 - \frac{\sum_{i \in \mathcal{V}_{\neq\star}} I(i, z_j)^2}{I(v_\star, z_j)^2 (M-1)} \tag{14}$$

Here, $\mathcal{V}_{\neq\star}$ is the set of all factors except $v_\star$, and M is the number of factors. The modularity score is averaged over all code dimensions.

J.4 MUTUAL INFORMATION GAP (MIG) AND MIG-SUP

MIG (Chen et al., 2018) evaluates the compactness of the representation by computing the MI between each factor and code dimension. The difference between the highest and second-highest MI for each factor is normalized by the factor entropy:

$$\text{MIG} = \frac{I(v_i, z_\star) - I(v_i, z_\circ)}{H(v_i)} \tag{15}$$

The MIG score is the average gap across all factors.

MIG-sup (Li et al., 2020) extends MIG to include modularity. It computes the MI gap from the perspective of the code:

$$\text{MIG-sup} = I(z_j, v_\star) - I(z_j, v_\circ) \tag{16}$$

The meaningful code dimensions are determined by a threshold on $I(z_j, v_\star)$. All code dimensions are considered to avoid thresholding.

J.5 Joint Entropy Minus Mutual Information Gap (JEMMIG)

JEMMIG (Do & Tran, 2020) addresses MIG's inability to measure modularity by incorporating the joint entropy of the factor and its most related code dimension:

$$\text{JEMMIG} = H(v_i, z_\star) - I(v_i, z_\star) + I(v_i, z_\circ) \tag{17}$$

The score is normalized to lie between 0 and 1:

$$\widehat{\text{JEMMIG}} = 1 - \frac{H(v_i, z_\star) - I(v_i, z_\star) + I(v_i, z_\circ)}{H(v_i) + \log(B_z)} \tag{18}$$

The average score across all factors is reported.

J.6 DCI-MIG

DCIMIG (Sepliarskaia et al., 2020) combines elements of DCI and MIG. It computes MI gaps for each factor and code dimension and aggregates the scores into a single disentanglement measure:

$$\text{DCIMIG} = \frac{\sum_{i=1}^{M} S_i}{\sum_{i=1}^{M} H(v_i)} \tag{19}$$

Here, S_i is derived from the maximum MI gap associated with each factor.

J.7 Interventional Robustness Score (IRS)

IRS (Suter et al., 2019) quantifies the robustness of code dimensions to changes in nuisance factors. Sets of codes are compared before and after targeted interventions, and the maximum observed distances are used to compute the final score, weighted by factor realization frequencies.