End-to-end training of deep probabilistic CCA on paired biomedical observations

Gregory Gundersen [*]
ggundersen@princeton.edu

Bianca Dumitrascu [†]
biancad@princeton.edu

Jordan T. Ash [*]
jordanta@cs.princeton.edu

Barbara E. Engelhardt [*]
bee@princeton.edu

Abstract

Medical pathology images are visually evaluated by experts for disease diagnosis, but the connection between image features and the state of the cells in an image is typically unknown. To understand this relationship, we develop a multimodal modeling and inference framework that estimates shared latent structure of joint gene expression levels and medical image features. Our method is built around probabilistic canonical correlation analysis (PCCA), which is fit to image embeddings that are learned using convolutional neural networks and linear embeddings of paired gene expression data. We train the model end-to-end so that the PCCA and neural network parameters are estimated simultaneously. We demonstrate the utility of this method in constructing image features that are predictive of gene expression levels on simulated data and the Genotype-Tissue Expression data. We demonstrate that the latent variables are interpretable by disentangling the latent subspace through shared and modality-specific views.

1. INTRODUCTION

Many diseases are diagnosed by pathologists using morphological features in tissue imaging data. But the genes that capture the internal state of cells and are associated with a specific tissue morphology are typically unknown and hard to assay in a particular sample. The Genotype-Tissue Expression (GTEx) Consortium [Consortium et al., 2017, Carithers et al., 2015] has collected data from over 948 autopsy subjects (donors), including standardized whole tissue histology slides, giving us images of each sample, and bulk RNA-sequencing, giving us gene expression levels for each sample, from approximately 50 different human tissues (Figure 1). These multi-subject, multi-view data provide an opportunity to develop computational tools that capture the relationship between cell state (observable in gene expression data) and morphological features (observable in histology images).

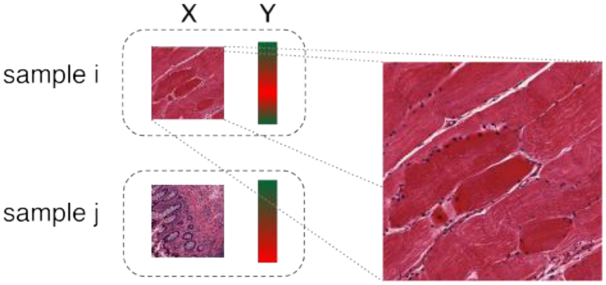

Figure 1: In our experiments, a *paired* GTEx sample consists of a whole tissue slide (cropped) along with gene expression levels from the same tissue and donor.

Historically, modeling data across two views with the goal of extracting shared signal has been performed by some version of canonical correlation analysis (CCA) [Hotelling, 1936]. Given two random vectors, CCA aims to find the linear projections into a shared latent space for which the projected vectors are maximally correlated. Probabilistic CCA (PCCA) is particularly attractive for medical applications with small sample sizes but large feature spaces by explicitly modeling uncertainty [Ghahramani, 2015].

In its most general form, PCCA will ignore possibly important nonlinear structure in data such as images; this nonlinear structure could be extracted first with computer vision techniques such as convolutional neural networks [LeCun et al., 1998] that have achieved excel-

[*]Department of Computer Science, Princeton University
[†] Graduate Program in Quantitative and Computational Biology, Princeton University

lent performance on medical imaging tasks [Bar et al., 2015, Shah et al., 2017, Esteva et al., 2017, Geras et al., 2017, Gulshan et al., 2016]. To this end, two recent studies trained models in a two-stage approach, first embedding the imaging data using convolutional models and then fitting CCA to the lower-dimensional embedded data [Ash et al., 2018, Subramanian et al., 2018]. Another two-stage approach computed the principal components of single views and then computed cross-view correlations [Barry et al., 2018], while Cooper et al. [2012] first clustered image features and then looked for genomic associations. However, two-stage approaches decouple the image feature learning from estimating the shared latent subspace of the data views, leading to image features that capture minimal variation in the shared subspace, and projections of the two views that are difficult to interpret. While "interpretability" is a broad concept [Lipton, 2016], here we mean specifically that we can identify small subsets of correlated features across modalities, here, gene expression levels representing cell state that are associated with specific image features.

To address these challenges, we propose a multimodal modeling and inference framework, deep probabilistic CCA (DPCCA). DPCCA estimates end-to-end the nonlinear embeddings and shared latent structure of paired high-dimensional data sets—here, histology images and paired gene expression data. Our model is *probabilistic* in that it is generative and models parameter uncertainty, *interpretable* in that it uses sparse PCCA to associate sets of genes with image features, and *nonlinear* in that it learns convolutional features for high-dimensional observations. Our training procedure makes use of automatic differentiation of a single loss function; this avoids the difficulties of implementing joint inference over probabilistic models and neural networks via conjugacy and message passing [Johnson et al., 2016, Lin et al., 2018].

The impact of solving these challenges is twofold. First, our DPCCA model will include latent factors capturing shared variation across the paired data modalities but also latent factors that capture modality-specific variation; our end-to-end inference will substantially increase the shared variation captured in the latent space by identifying embeddings for each data view that maximize the shared variation. Second, this generative framework allows cross-mode imputation: given a fitted model, we can guess at the values of gene expression data for a held-out histology image, for example. This is particularly important given a discrepancy in cost between the two data modalities—it is much more expensive to assay gene expression in a tissue sample than to stain and image that sample.

We illustrate the behavior of our method for simultaneous embedding and latent space modeling of multiple data modalities on both the MNIST data [LeCun et al., 2010], where we paired Gaussian-distributed vectors with specific digits, and on the GTEx v6 data. We compare the results from our approach against results from related multimodal approaches in order to illustrate the additional gains in the variation captured in the shared latent space and also the interpretability of the end-to-end inference of embeddings and shared space. We validate these results in the GTEx v6 data using additional held out biological data that correlate with signal identified in the inferred latent space. We conclude with thoughts on further improvements to our model.

1.1 CONTRIBUTIONS

The fundamental contributions of DPCCA to the field of simultaneous embedding and joint modeling of high-dimensional multimodal data include addressing three main methodological and domain-specific challenges:

- In our end-to-end training procedure, the shared latent subspace drives the convolutional image embeddings. Compared with a standard autoencoder that learns embeddings that minimize a reconstruction loss, the PCCA backend encourages image embeddings that maximally explain variation in the other data modalities.

- The shared and modality-specific latent variables provide three views into variation that can be mined for domain-specific patterns of interest, making our model interpretable with respect to the data domain.

- The shared latent variables represent a *composite phenotype* between tissue morphology and gene expression—sets of genes representing cell state and image features that covary together. These composite phenotypes can be used for many downstream tasks, including identifying paired phenotypic differences between sick and healthy patients and testing for associations with other modalities, such as genotype.

2. RELATED WORK

The original realization of CCA was more recently reframed as a probabilistic model. This model is known as inter-battery factor analysis in the statistics community [Browne, 1979] and was rederived as probabilistic CCA [Bach and Jordan, 2005, Murphy, 2012] in the machine learning community. An important feature of PCCA is the allowance for view-specific noise. If PCCA assumed independent noise, that would mean that any view-specific variation in the data would have to be modeled as shared variation. The model would have no other way of explaining that variance given that it assumes noise is independent.

CCA has also been extended to nonlinear settings using kernel methods [Akaho, 2006, Hardoon et al., 2004]. Variants of combining CCA with neural networks also exist. Deep CCA (DCCA) estimates linear projections on the outputs of neural networks [Andrew et al., 2013]. Deep variational CCA (DVCCA) is a variational approximation of CCA using a single encoder, while we learn pairs of embeddings with view-specific encoders [Wang et al., 2016]. While DCCA learns embeddings that capture shared structure, it does not explicitly model view-specific noise as in PCCA. We demonstrate that this is an important benefit of our model in Section 4. Furthermore, learning linear maps as in CCA and PCCA is key to the interpretation of covarying data features from a given latent variable.

Deep multimodal learning without the notion of correlation maximization is another related body of work [Ngiam et al., 2011]. However, a multimodal autoencoder that learns a shared lower-dimensional representation explicitly optimizes a reconstruction loss, but it does not disentangle the latent space across views, which is an essential component of our model. Another related model worth mentioning is oi-VAE [Ainsworth et al., 2018], which uses multiple decoders over the same latent variables, with the goal of having interpretable factors of the same data view, not accounting for multiple views.

We note that our domain-specific problem is related to other domains such as image captioning in computer vision and neural machine translation in natural language processing. A major distinction in language-based modeling is that they cannot make the same Gaussian assumptions about their data.

3. DEEP PROBABILISTIC CCA

3.1 PROBLEM SETUP

We index n paired samples using $i \in \{1, 2, \ldots, n\}$, and we index two data views a and b using $j \in \{a, b\}$. The ith paired sample is a tuple $(\mathbf{x}_i^a, \mathbf{x}_i^b)$. Here, an image $\mathbf{x}_i^a$ is a multidimensional array with dimensions for the number of channels, image height, and image width; for ease of notation we can flatten this multidimensional array into a vector with dimensionality $\mathbb{R}^{q^a}$. A gene expression sample $\mathbf{x}_i^b$ is a vector $\mathbb{R}^{q^b}$ for q^b genes. We embed each data view before performing PCCA, and we refer to these view-specific embeddings as p^a- and p^b-dimensional vectors $\mathbf{y}^a$ and $\mathbf{y}^b$. Here we use a convolutional autoencoder for the image vector and a linear embedding for the gene expression vector. Each paired sample comes from a single donor and one of 29 human tissues after preprocessing (Table 1). The sample tissue labels are held out to be used as biological signal to validate the model, which we explore in Section 4.

TISSUE	COUNT	TISSUE	COUNT
Adipose Tissue	5	Nerve	9
Adrenal Gland	134	Ovary	88
Bladder	4	Pancreas	166
Blood Vessel	47	Pituitary	51
Brain	172	Prostate	53
Breast	5	Salivary Gland	10
Cervix Uteri	7	Skin	28
Colon	81	Small Intestine	59
Esophagus	134	Spleen	103
Fallopian Tube	4	Stomach	106
Heart	188	Testis	44
Kidney	12	Thyroid	65
Liver	115	Uterus	69
Lung	76	Vagina	17
Muscle	369	TOTAL	2221

Table 1: We preprocessed the GTEx v6 data to only include samples from which 1000 × 1000 pixels crops could be taken and that had both tissue slides and gene expression levels. After preprocessing, we obtained 2221 paired samples from 29 tissue types. The data are both small and class-imbalanced.

3.2 CANONICAL CORRELATION ANALYSIS

Consider two paired data views, $\mathbf{Y}^a \in \mathbb{R}^{n \times p^a}$ and $\mathbf{Y}^b \in \mathbb{R}^{n \times p^b}$. We assume the data are mean-centered. The objective of CCA is to learn two linear maps $\mathbf{H}^a \in \mathbb{R}^{p^a \times k}$ and $\mathbf{H}^b \in \mathbb{R}^{p^b \times k}$ such that the ith pair of *canonical variables*, $\mathbf{z}_i^a = \mathbf{Y}^a \mathbf{h}_i^a$ and $\mathbf{z}_i^b = \mathbf{Y}^b \mathbf{h}_i^b$, are maximally correlated. These canonical variables are further constrained to have unit length ($\|\mathbf{z}_i^a\| = \|\mathbf{z}_i^b\| = 1$) and to be orthogonal ($\langle \mathbf{z}_i^a, \mathbf{z}_k^a \rangle = \langle \mathbf{z}_i^b, \mathbf{z}_k^b \rangle = 0$ for all $i \neq k$ pairs). The solution to this optimization problem can be found analytically by solving the standard eigenvalue problem [Hotelling, 1936, Hardoon et al., 2004]. The geometric interpretation is that we estimate two linear maps that project both views into a shared subspace.

3.3 PROBABILISTIC CCA

A probabilistic interpetation of CCA (PCCA) extends these ideas to a model that shares properties with factor analysis [Browne, 1979, Bach and Jordan, 2005]. In

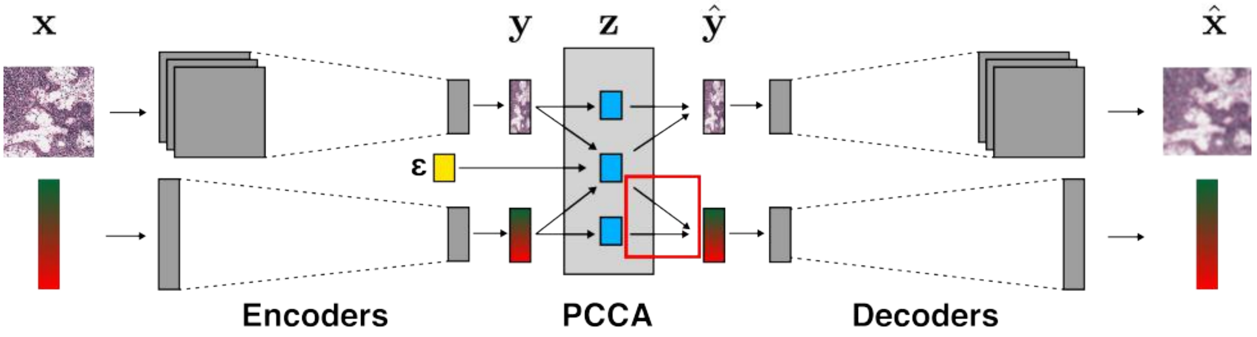

Figure 2: The input is a paired set of histology images and gene expression levels. The model is trained by fitting PCCA to embeddings from convolutional (images) and linear (gene expression) autoencoders (gray). Then we sample from the PCCA model using the reparameterization trick ($\boldsymbol{\epsilon} \sim p(\boldsymbol{\epsilon})$ is in yellow), and then we backpropagate through the model using the reconstruction loss. The model learns shared and modality-specific latent variables (blue). Sparsity is induced on the PCCA parameters for the gene expression levels (red box).

particular, PCCA can be written as

$$\begin{aligned}\mathbf{z}^c &\sim \mathcal{N}(\mathbf{0}_k; \mathbf{I}_k)\\ \mathbf{z}^a, \mathbf{z}^b &\sim \mathcal{N}(\mathbf{0}_k; \mathbf{I}_k)\\ \mathbf{y}^a &\sim \mathcal{N}(\boldsymbol{\Lambda}^a \mathbf{z}^c + \mathbf{B}^a \mathbf{z}^a; \boldsymbol{\Psi}^a)\\ \mathbf{y}^b &\sim \mathcal{N}(\boldsymbol{\Lambda}^b \mathbf{z}^c + \mathbf{B}^b \mathbf{z}^b; \boldsymbol{\Psi}^b),\end{aligned} \tag{1}$$

where $\boldsymbol{\Lambda}^j \in \mathbb{R}^{p^j \times k}$, $\mathbf{B}^j \in \mathbb{R}^{p^j \times k}$, and $\boldsymbol{\Psi}^j \in \mathbb{R}^{p^j \times p^j}$. Note that the *view-specific* latent variables $\mathbf{z}^a$ and $\mathbf{z}^b$ account for view-specific variation, while the *shared* latent variable $\mathbf{z}^c$ captures shared variation (covariation) across the two views.

Note that PCCA can be viewed as factor analysis with appropriately tiled data and parameters,

$$\mathbf{y} = \begin{bmatrix}\mathbf{y}^a\\ \mathbf{y}^b\end{bmatrix} \qquad \mathbf{z} = \begin{bmatrix}\mathbf{z}^c\\ \mathbf{z}^a\\ \mathbf{z}^b\end{bmatrix}$$

$$\boldsymbol{\Lambda} = \begin{bmatrix}\boldsymbol{\Lambda}^a & \mathbf{B}^a & \mathbf{0}\\ \boldsymbol{\Lambda}^b & \mathbf{0} & \mathbf{B}^b\end{bmatrix} \qquad \boldsymbol{\Psi} = \begin{bmatrix}\boldsymbol{\Psi}^a & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\Psi}^b\end{bmatrix}, \tag{2}$$

where $\mathbf{0}$ denotes appropriately sized matrices of all zeros. This immediately suggests expectation-maximization (EM) for inference in PCCA, drawing from the EM parameter updates for factor analysis given this tiling [Ghahramani et al., 1996]:

$$\begin{aligned}\boldsymbol{\Lambda}^\star &= \sum_i \Big(\mathbf{y}_i \mathbb{E}_{\mathbf{z}|\mathbf{y}_i}[\mathbf{z} \mid \mathbf{y}_i]^\top\Big)\Big(\mathbb{E}_{\mathbf{z}|\mathbf{y}_i}[\mathbf{z}\mathbf{z}^\top \mid \mathbf{y}_i]\Big)^{-1}\\ \boldsymbol{\Psi}^\star &= \sum_i \frac{1}{n}\operatorname{diag}\Big(\mathbf{y}_i \mathbf{y}_i^\top - \boldsymbol{\Lambda}^\star \mathbb{E}_{\mathbf{z}|\mathbf{y}_i}[\mathbf{z} \mid \mathbf{y}_i]\mathbf{y}_i^\top\Big).\end{aligned} \tag{3}$$

In this framing, $\mathbf{y} \in \mathbb{R}^p$ where $p = p^a + p^b$ and $\mathbf{z} \in \mathbb{R}^k$ where $k = k^c + k^a + k^b$, the dimensions of $\mathbf{z}^c$, $\mathbf{z}^a$, and $\mathbf{z}^b$ respectively. Thus, $\mathbf{y} \in \mathbb{R}^p$, $\boldsymbol{\Lambda} \in \mathbb{R}^{p \times k}$, and $\boldsymbol{\Psi} \in \mathbb{R}^{p \times p}$. In contrast to CCA, PCCA does not constrain the latent variables to be orthonormal.

Algorithm 1 End-to-end training of DPCCA

```
Initialize PCCA parameters, image encoder and decoder parameters, and gene encoder and decoder parameters ($\boldsymbol{\Lambda}$, $\boldsymbol{\Psi}$, $\mathbf{W}_e^a$, $\mathbf{W}_d^a$, $\mathbf{W}_e^b$, $\mathbf{W}_d^b$, respectively).
while epoch < # epochs do
    For $m$ paired samples, $B = \{(\mathbf{x}^a, \mathbf{x}^b)_i\}_{i=1}^m$.
    for $(\mathbf{x}^a, \mathbf{x}^b)_i \in B$ do
        Encode the $j$th view as $E^j(\mathbf{x}^j) \rightarrow \mathbf{y}^j$.
        Compute $\boldsymbol{\Lambda}^\star$ and $\boldsymbol{\Psi}^\star$ using Equation 3.
        Sample $\hat{\mathbf{y}}^j \sim \mathcal{N}(\boldsymbol{\Lambda}^{j\star}\mathbf{z}^c + \mathbf{B}^{j\star}\mathbf{z}^j; \Psi_j^\star)$ using the reparameterization trick.
        Decode the $j$th view as $D^j(\hat{\mathbf{y}}^j) \rightarrow \hat{\mathbf{x}}^j$.
    end for
    Compute reconstruction loss (Equation 4) and backpropagate to compute $\nabla \mathcal{L}_{\boldsymbol{\Theta}}$.
end while
```

3.4 END-TO-END TRAINING OF DPCCA

DPCCA is a deep generative model that fits PCCA to the embeddings of two autoencoders. Additionally, the model has an ℓ_1 penalty on the PCCA gene weights ($\boldsymbol{\Lambda}^b$ and $\mathbf{B}^b$) to encourage sparsity in the factors for the gene expression levels. The DPCCA model is trained end-to-end with backpropogation through the reconstruction loss (Figure 2).

In detail, given a paired sample $(\mathbf{x}_i^a, \mathbf{x}_i^b)$, each encoder $E^j(\cdot)$ with parameters $\mathbf{W}_e^j$ embeds its respective views into a vector, $\mathbf{y}^i \in \mathbb{R}^{p^i}$. Each embedding is view-specific: here we use a convolutional encoder for the images and a linear projection for the genes. The embedded vectors $\mathbf{y}^a$ and $\mathbf{y}^b$ are then fit by PCCA using the parameter updates in Equation 3 with a sparsity-inducing prior on the b-specific parameters. This results in shared and view-specific latent variables $\mathbf{z} = \begin{bmatrix}\mathbf{z}^c & \mathbf{z}^a & \mathbf{z}^b\end{bmatrix}^\top$.

Embedding samples $\hat{\mathbf{y}}^j$ are obtained from the generative process of the model through sampling from the low-dimensional PCCA representation $\hat{\mathbf{y}}^j \sim \mathcal{N}(\mathbf{\Lambda}^{j\star}\mathbf{z}^c + \mathbf{B}^{j\star}\mathbf{z}^j; \mathbf{\Psi}^{j\star})$ using the reparameterization trick similar to Kingma and Welling [2013]. This reparameterization is needed so that the Monte Carlo estimate of the expectation is differentiable with respect to the encoders' parameters.

Each sampled PCCA embedding $\hat{\mathbf{y}}^a$ and $\hat{\mathbf{y}}^b$ is then decoded into reconstructions $\hat{\mathbf{x}}^a$ and $\hat{\mathbf{x}}^b$ using view-specific decoders with parameters $\mathbf{W}_d^j$ (Figure 2). Finally, let $\mathcal{L}$ be the reconstruction loss and $\mathbf{\Theta}$ be both the PCCA and neural network parameters, or

$$\mathbf{\Theta} = \{\mathbf{\Lambda}, \mathbf{\Psi}\} \cup \{\mathbf{W}_e^j, \mathbf{W}_d^j\}_{j \in \{a,b\}}.$$

To estimate the parameters $\mathbf{\Theta}$, we perform stochastic gradient descent, where the gradient at each step is $\nabla_{\mathbf{\Theta}}\mathcal{L}$ with

$$\mathcal{L} = \frac{1}{n}\sum_{i=1}^{n}\Big(\|\hat{\mathbf{x}}_i^a - \mathbf{x}_i^a\|_2^2 + \|\hat{\mathbf{x}}_i^b - \mathbf{x}_i^b\|_2^2\Big) + \gamma\big(\|\mathbf{\Lambda}^b\|_1 + \|\mathbf{\Lambda}^{bc}\|_1\big). \tag{4}$$

The hyperparameter γ is the ℓ_1 coefficient. This procedure is summarized in Algorithm 1.

4. EXPERIMENTS

In this section, we explore the strengths of our model in two settings: an expanded version of the MNIST handwritten digit data [LeCun et al., 2010], and the GTEx v6 data [Consortium et al., 2017, Carithers et al., 2015] that includes publicly available paired histology images and gene expression data. We implemented our model in PyTorch [Paszke et al., 2017] and used the Adam optimizer for all experiments [Kingma and Ba, 2014]. Our code is available online[1] to encourage more work in this important area.

4.1 BASELINES AND MULTIMODAL MNIST

We first wanted to study the performance of our model, and compare our results with results from related work using a simple data set. To do this, we built a *multimodal MNIST* data set using the MNIST handwritten digits. MNIST consists of 60,000 training and 10,000 testing images, each with 28×28 pixels with values ranging between 0 (black) and 255 (white). The images are handwritten digits between 0 and 9 and have corresponding class labels $y_i \in \{0, 1, \ldots 9\}$.

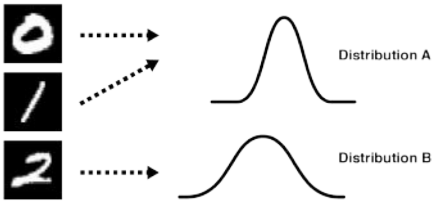

Figure 3: Multimodal MNIST: Each image from one of three classes $(0, 1, 2)$ is paired with a continuous random variable (*pseudogene*) drawn from one of two multivariate normal distributions with separate means and separate diagonal covariance matrices. Digits of 0s and 1s are paired with samples from the first distribution. Digits of 2s are paired with samples from the second distribution.

We augmented MNIST in the following way[2]. First, we removed all images with labels not in $[0-2]$. For each remaining image, we created an associated *pseudogene* expression vector by sampling from one of two multivariate normal distributions, depending on the image label. The distributions had separate means and separate diagonal covariance matrices. If the image was a 0 or 1, we sampled from the first distribution. If the image was a 2, we sampled from the second distribution (Figure 3).

Our model should ideally be able to reconstruct both modalities using the latent variables, including the modality-specific variation and the shared variation. We can also examine $\mathbf{z}$ to ensure that it captures the shared information we encoded in the data: namely, the relationship between the (0,1) images and the 2 images with their respective multivariate normal distributions rather than image digit label, which, for (0,1), are not distinguished by pseudogenes.

As baselines, we fit a single-view autoenoder (AE) on just images, a multimodal autoencoder (MAE) on both data views, and standard PCCA to both data views. We found that DPCCA can reconstruct both modalities well relative to these baselines (Table 2). The AE and MAE

	Image MSE	**Pseudogene MSE**
Image AE	0.0196 (0.0019)	NA
MAE	0.0435 (0.0015)	2.287 (0.0117)
PCCA	0.1207 (0.0032)	33.749 (0.648)
DPCCA	0.0518 (0.0121)	2.3098 (0.0137)

Table 2: Baseline experiments comparing an image-only autoencoder (AE), a multimodal autoencoder (MAE), PCCA, and DPCCA on image and pseudogene reconstructions of multimodal MNIST. Each error is an average of five independent trials; standard deviations are shown parenthetically. Our method performs comparably to an MAE and outperforms PCCA at reconstructing both views.

[1] `https://github.com/gwgundersen/dpcca`

[2] See `data/mnist/generate.py` in the repository.

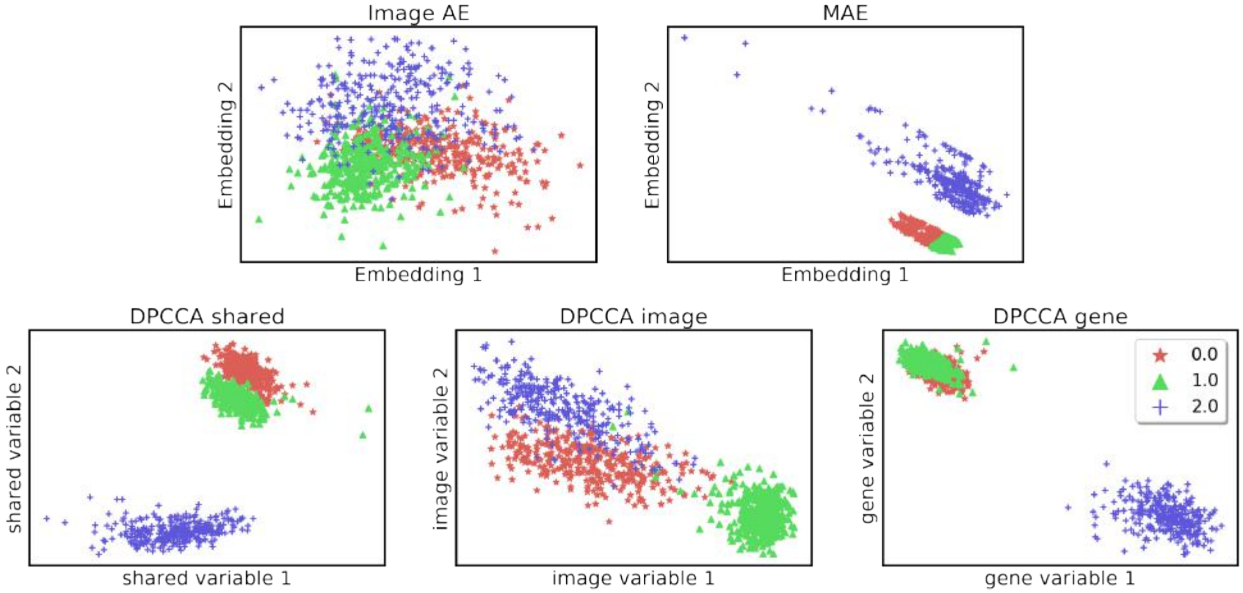

Figure 4: Two-dimensional embeddings from models trained on multimodal MNIST. **Top row.** Embeddings from an autoencoder (left) and multimodal autoencoder (right). **Bottom row.** The shared (left), image-specific (center), and gene-specific (right) embeddings from DPCCA.

are better than DPCCA at reconstruction, which is expected since our method also must optimize PCCA in an inner loop. Standard PCCA performs worse in reconstructing both views. However, neither the MAE nor the AE incorporate both shared and view-specific latent variables, which is crucial to the interpretability of our framework.

Second, we found that the shared and view-specific latent variables contained appropriate shared and view-specific information. To illustrate this, we compared the latent space of DPCCA to the embeddings of the single-view AE and an MAE (Figure 4, top). In this experiment, we set k, the dimensionality of the embeddings, to 2 because we have empirically found that the AE with two-dimensional embeddings can reconstruct MNIST well. Recall that, in our model, $\mathbf{z} = \begin{bmatrix}\mathbf{z}^c & \mathbf{z}^a & \mathbf{z}^b\end{bmatrix}^\top$. DPCCA's shared latent variables $\mathbf{z}^c$ primarily capture the relationship between the two views rather than distinguishing between digits (Figure 4, bottom left). For comparison, the AE trained on images alone distinguishes digit label, while the MAE captures the shared view without distinguishing 0s and 1s (Figure 4, top).

We compared the shared and view-specific latent variables of our model to understand the signal captured by each set. The shared latent variables do not distinguish digits 0 and 1 but instead distinguish 0s and 1s versus 2s (Figure 4, top right). The image-specific latent variables capture information that distinguishes the three digits; this makes sense since the MNIST images are digits (Figure 4, bottom center). This view is similar to the image-only AE. The pseudogene-specific latent variables, like the shared latent variables, do not distinguish 0s and 1s because the pseudogene variables corresponding to both 0s and 1s are drawn from the same distribution (Figure 4, bottom right). These results suggest that DPCCA can estimate embeddings that maximize the correlation of the two views, and that together these shared and view-specific embeddings capture meaningful signals and information contained in held-out class labels better than autoencoders alone.

4.2 GTEX PAIRED DATA

In experiments on the GTEx data [Consortium et al., 2017, Carithers et al., 2015], we wanted to show that our model can be applied to these data, that it captures interesting held-out biological information such as tissue type, and that the shared latent variables model variation in both images and gene expression levels. To show this, we analyzed the latent factors of our model—a "factor" being a row vector of $\mathbf{Z} \in \mathbb{R}^{k' \times n}$ where k' may be k, k^c, k^a, or k^b depending on context—and found tissue-specific information and variation in images that covaries with changes in factor value. We used held-out genotypes known to be associated with specific genes to identify genotypes associated with tissue morphology using the shared factors. While these results are preliminary from a biological perspective, they are evidence that our model may be a useful tool for joint analysis of paired

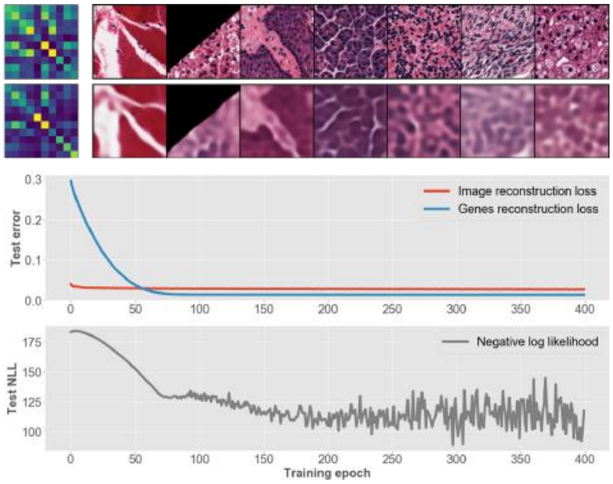

Figure 5: Top. Examples of original data (top row) and reconstructions (bottom row) of gene expression covariance matrices (left) and histology slides (right). For clarity, we show the top 10 columns with the highest variance in the original data. **Bottom.** (Top) View-specific test error over training on the test data from the GTEx data. (Bottom) Negative log likelihood (Test NLL) of DPCCA over training on the test data from the GTEx data.

data.

To this end, we trained our model on 2221 samples from the GTEx v6 study. Each whole tissue slide was subsampled once down to a 1000×1000 pixel RGB image. The crops were chosen as follows. A slide was scanned for tiles in which the mean gray values of the tile and its neighboring tiles were darker than 180 out of 255. The final crop was chosen uniformly at random from suitable tiles. To augment the data, the model was trained on 128×128 pixel crops with random rotations and reflections. The image autoencoder is based on the DCGAN architecture [Radford et al., 2015], and the gene autoencoder is two linear layers. The gene expression measurements are approximately 18,000-dimensional real-valued vectors [Hubbard et al., 2002]. For the number of latent variables for each of the three latent variables types, we swept over $k \in \{2, 10, 20, 50, 100, 500\}$ and used the smallest number, 10, that resulted in high-quality image and gene reconstructions. Thus, $k^c = k^a = k^b = 10$ and $k = 30$.

Before using our model for biomedical data analysis, we wanted to verify two important properties. First, we wanted to show that our model could reconstruct both data modalities from a shared latent variable. To show this, we saved reconstructions of the images and reconstructions of the gene covariance matrices during training. We found that our model was able to reconstruct both modalities (Figure 5, top) and that the test error for both views decreases throughout training (Figure 5, bottom). This suggests that the shared latent variables carry sufficient information to recapitulate both views.

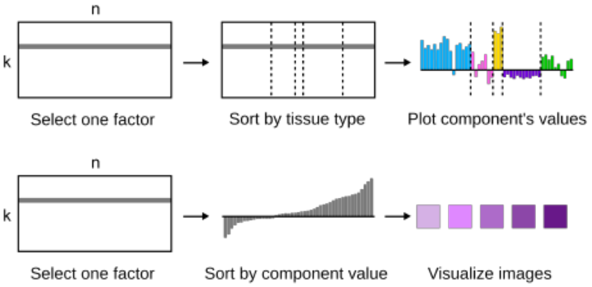

Figure 6: Analyzing a latent factor. **Top.** Given a single n-dimensional latent factor (left), we sort the samples or columns by tissue type (middle) and then plot the value of the factor for each sample. **Bottom.** Given a single n-dimensional latent factor (left), we sort the samples based on the factor's value (middle). Then we visualize each sample by its associated histology slide in the same order as the sorted factors.

Second, we wanted to verify our end-to-end training procedure. With a model composed of both neural networks and PCCA, we might ask whether one of the sub-models is ignored due to an imbalance in the numbers of parameters. To test this, we computed the expected complete negative log-likelihood of held-out test data and found that it decreased over training (Figure 5, bottom). Taken together, these results suggest that the neural networks and PCCA are jointly learning parameters for embedding and reconstructing data from nonlinear observations while minimizing the negative log-likelihood of the generative model.

4.3 TISSUE-SPECIFIC ASSOCIATIONS

Next, we wanted to see if our latent factors captured meaningful, held-out biological information: the tissue type of the sample. We did this by sorting the samples (latent variables) by tissue type and plotting the value of a latent factor for each sample (Figure 6, top). We found that the model's latent factor capture tissue-specific information, and quantified this using a one-sample two-sided t-test (Figure 7, top). This measures the extent to which the different subsets of the latent variable's factors pull out tissue-specific information.

Our analysis demonstrates that tissue-specific structure is shared across images and genes, and is captured both in the shared factors and also in the gene-specific factors, but less so in the image specific factors. We hypothesize that the tissue-specific signal in images is captured in the shared latent space, which is why no tissue-specific signal is observed in the image-specific latent space.

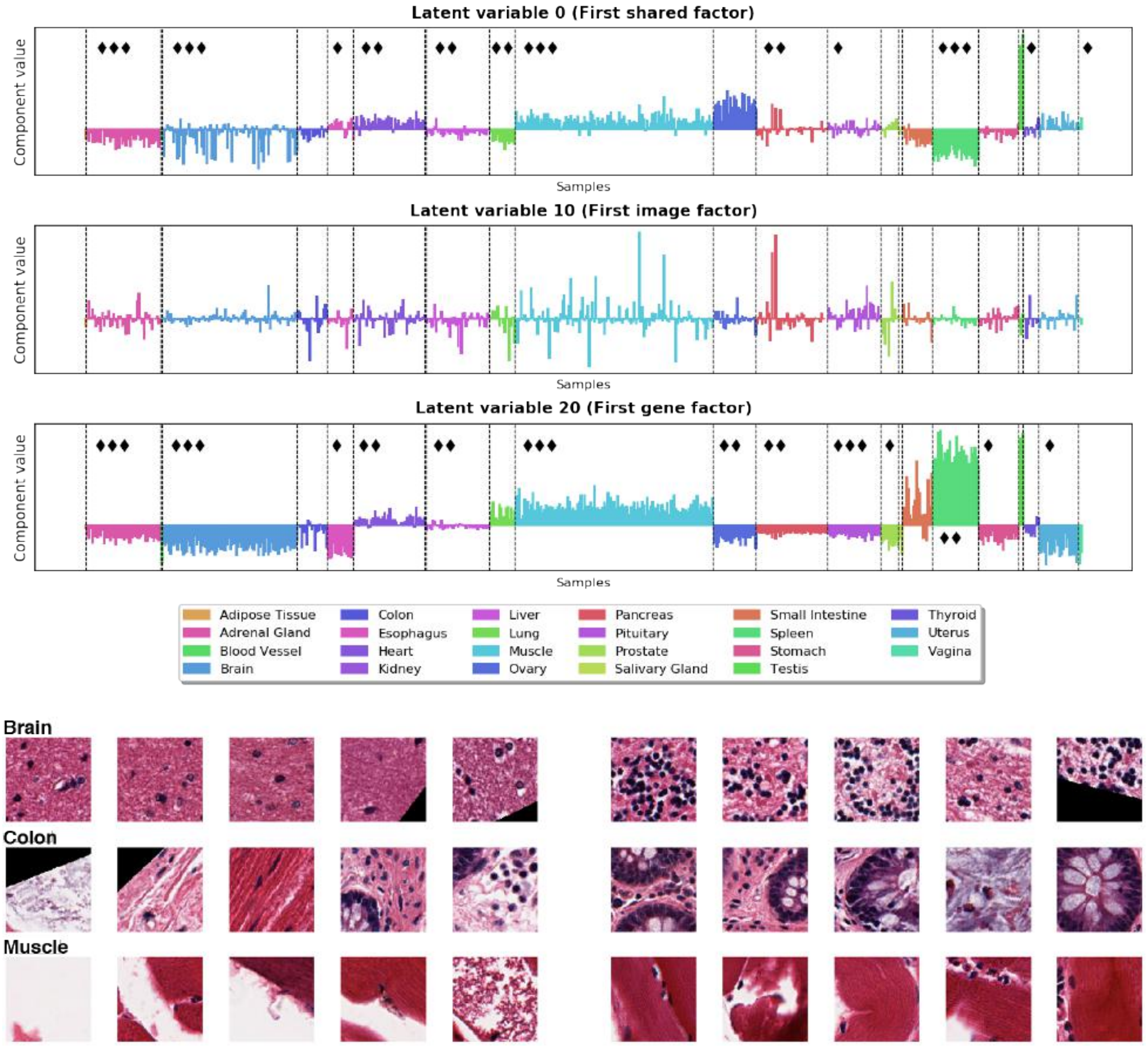

Figure 7: Top. Analysis of shared (top row), image-specific (middle row), and gene-specific (bottom row) latent factors. The x-axis (samples) is sorted by tissue type and the y-axis is factor value. We performed a one-sample two-sided t-test on the latent factor values for all samples of the same tissue type. We applied Bonferroni correction to a p-value threshold of 0.05. Tissue samples that reject the null hypothesis are marked with diamonds. We ranked the significant p-values and then uniformly partitioned them from most (3 diamonds) to least (1 diamond) significant. **Bottom.** Visualization of the variation in the latent factors. The x-axis (samples) is sorted by factor value, and the images associated with the five most extreme positive and negative values are shown for three tissues.

4.4 IMAGE-SPECIFIC VARIATION

Next, we wanted to see if DPCCA captures interpretable morphological information about the images. We did this by visualizing the image associated with each sample after sorting by a single latent factor (Figure 6, bottom). We found that our model's latent factors capture variations in images that are visible to the human eye (Figure 7, bottom). In some cases, this variation is a feature of the image itself. For example, cropped images with black chunks are toward one end of the spectrum. But in other cases, this variation is related to tissue morphology. For example, more striated muscle tissue and cerebellar granule cells are both captured by the factor value.

4.5 DOWNSTREAM ANALYSIS: IMAGE QTLS

The shared latent variables from our method can be integrated into established genomics pipelines such as quantitative trait loci (QTL) mapping. A cornerstone of

TISSUE	SNP	P-VALUE	FDR
Adrenal Gland	chr10_5330574_G_T_b38	3.49E-10	2.40E-3
Adrenal Gland	chr13_33247934_C_T_b38	7.56E-10	3.81E-3
Adrenal Gland	chr9_18386575_C_G_b38	1.65E-9	9.57E-3
Brain Cerebellum	chr1_14376802_A_G_b38	1.82E-10	1.83E-3
Brain Cerebellum	chr9_133727220_C_A_b38	4.43E-10	2.57E-3
Colon Sigmoid	chr10_49206009_T_G_b38	2.02E-10	1.37E-3
Colon Sigmoid	chr12_82520469_A_G_b38	2.29E-9	5.30E-3
Esophagus Mucosa	chr13_68558539_A_T_b38	3.45E-10	4.51E-5
Muscle Skeletal	chr6_150866897_G_A_b38	6.14E-11	5.62E-4
Uterus	chr11_6991683_G_A_b38	1.89E-9	6.20E-3

Table 3: Top genotypes hypothesized to affect the composite phenotype capturing gene expression and morphology across different tissues.

quantitative genomic analysis [Consortium et al., 2017], this method aims to identify associations between genetic variants (genotypes) and quantitative human traits such as height, weight, or gene expression levels (phenotypes), using false discovery rate (FDR)-corrected linear regression within each tissue. The shared latent factors estimated using DPCCA constitute a phenotype describing how the morphology of cells in a tissue (how cells appear) covaries with gene expression levels (characterizing cellular state). These resulting composite phenotypes allow researchers to study the close relationship between a cell's appearance and a cell's state at a macroscopic scale, with the goal of using a cell's appearance to infer its state at a high resolution. QTL analysis takes these ideas one step further to query whether population variation, in the form of differences in genotypes at a particular genetic locus, leads to differences in cellular morphology or cellular state.

To do this, we performed QTL analysis using linear regression (MatrixEQTL [Shabalin, 2012]) with a Benjamini-Hochberg corrrected FDR threshold of 0.05 between the 30 composite phenotypes and over $400,000$ genomic loci per tissue across 635 individuals. We found over $20,000$ associations (Table 3). While validating these associations and elucidating the biological mechanisms behind them is beyond the scope of this work, we note that some of these associations are recurrent in the biological literature. For example, the genotype on chromosome 1 at position 14376802 appears to regulate expression levels of the gene *KAZN* or Kazrin in cerebellum in the brain. This gene has previously been found to affect changes in cell shape across various species [Cho et al., 2011].

5. DISCUSSION AND FUTURE WORK

In this paper, we developed a model and associated end-to-end inference method for learning shared structure in paired samples, specifically, histology images and gene expression levels. While our framework combines the power of neural networks for nonlinear embeddings with probabilistic models for interpretable dimension reduction, inference is gradient-based and can be implemented using frameworks leveraging automatic differentiation such as PyTorch [Paszke et al., 2017] and TensorFlow [Abadi et al., 2016].

We demonstrated that the latent factors estimated by DPCCA revealed tissue-specific structure, despite withholding tissue labels from the model, as well as view-specific structure such as color and tissue attenuation for the images. We further validated our results using QTL analysis.

Future work will address a unique modeling opportunity arising from the availability of single cell data, namely the challenge of annotating images of cells with predicted gene expression levels at pixel-level resolution. In the GTEx v6 data, the gene expression levels are assayed in bulk, meaning that gene expression levels are assayed across the thousands of cells in the sample that are captured in a histological image. In this case, we hypothesize that our fitted model could estimate the shared latent variables from a single region of the image and output predicted gene expression labels. This would allow a dense labeling or annotation of test images, in which each region of an image were overlaid with the predicted expression values for a gene of interest. But validation of these densely labeled images requires single cell technologies. As single cell data sets are expanding under the auspice of cell atlas projects [Regev et al., 2017], multi-view biomedical data sets in which images of expression levels in single cells are paired with cell-specific gene expression measurements will be increasingly available. Other than their use in validation, exciting multi-view models can be adapted for use in single cell data sets.

Acknowledgements

We thank Daniel Munro and Greg Darnell for helpful conversations.

References

M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th $\{$USENIX$\}$ Symposium on Operating Systems Design and Implementation ($\{$OSDI$\}$ 16)*, pages 265–283, 2016.

S. K. Ainsworth, N. J. Foti, A. K. Lee, and E. B. Fox. oi-vae: Output interpretable vaes for nonlinear group factor analysis. In *International Conference on Machine Learning*, pages 119–128, 2018.

S. Akaho. A kernel method for canonical correlation analysis. *arXiv preprint cs/0609071*, 2006.

G. Andrew, R. Arora, J. Bilmes, and K. Livescu. Deep canonical correlation analysis. In *International Conference on Machine Learning*, pages 1247–1255, 2013.

J. Ash, G. Darnell, D. Munro, and B. Engelhardt. Joint analysis of gene expression levels and histological images identifies genes associated with tissue morphology. *bioRxiv*, 2018. doi: 10.1101/458711.

F. R. Bach and M. I. Jordan. A probabilistic interpretation of canonical correlation analysis. 2005.

Y. Bar, I. Diamant, L. Wolf, S. Lieberman, E. Konen, and H. Greenspan. Chest pathology detection using deep learning with non-medical training. In *Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on*, pages 294–297. IEEE, 2015.

J. D. Barry, M. Fagny, J. N. Paulson, H. J. Aerts, J. Platig, and J. Quackenbush. Histopathological image qtl discovery of immune infiltration variants. *iScience*, 5:80–89, 2018.

M. W. Browne. The maximum-likelihood solution in inter-battery factor analysis. *British Journal of Mathematical and Statistical Psychology*, 32(1):75–86, 1979.

L. J. Carithers, K. Ardlie, M. Barcus, P. A. Branton, A. Britton, S. A. Buia, C. C. Compton, D. S. DeLuca, J. Peter-Demchok, E. T. Gelfand, et al. A novel approach to high-quality postmortem tissue procurement: the gtex project. *Biopreservation and biobanking*, 13(5): 311–319, 2015.

K. Cho, M. Lee, D. Gu, W. A. Munoz, H. Ji, M. Kloc, and P. D. McCrea. Kazrin, and its binding partners arvcf-and delta-catenin, are required for xenopus laevis craniofacial development. *Developmental Dynamics*, 240(12): 2601–2612, 2011.

G. Consortium et al. Genetic effects on gene expression across human tissues. *Nature*, 550(7675):204, 2017.

L. A. Cooper, J. Kong, D. A. Gutman, F. Wang, J. Gao, C. Appin, S. Cholleti, T. Pan, A. Sharma, L. Scarpace, et al. Integrated morphologic analysis for the identification and characterization of disease subtypes. *Journal of the American Medical Informatics Association*, 19(2): 317–323, 2012.

A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115, 2017.

K. J. Geras, S. Wolfson, S. Kim, L. Moy, and K. Cho. High-resolution breast cancer screening with multi-view deep convolutional neural networks. *arXiv preprint arXiv:1703.07047*, 2017.

Z. Ghahramani. Probabilistic machine learning and artificial intelligence. *Nature*, 521(7553):452, 2015.

Z. Ghahramani, G. E. Hinton, et al. The em algorithm for mixtures of factor analyzers. Technical report, Technical Report CRG-TR-96-1, University of Toronto, 1996.

V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama*, 316(22):2402–2410, 2016.

D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12): 2639–2664, 2004.

H. Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.

T. Hubbard, D. Barker, E. Birney, G. Cameron, Y. Chen, L. Clark, T. Cox, J. Cuff, V. Curwen, T. Down, et al. The ensembl genome database project. *Nucleic acids research*, 30(1):38–41, 2002.

M. Johnson, D. K. Duvenaud, A. Wiltschko, R. P. Adams, and S. R. Datta. Composing graphical models with neural networks for structured representations and fast inference. In *Advances in neural information processing systems*, pages 2946–2954, 2016.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Y. LeCun, C. Cortes, and C. Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann. lecun. com/exdb/mnist*, 2, 2010.

W. Lin, N. Hubacher, and M. E. Khan. Variational message passing with structured inference networks. *arXiv preprint arXiv:1803.05589*, 2018.

Z. C. Lipton. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*, 2016.

K. P. Murphy. *Machine learning: a probabilistic perspective*. Cambridge, MA, 2012.

J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 689–696, 2011.

A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. 2017.

A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

A. Regev, S. A. Teichmann, E. S. Lander, I. Amit, C. Benoist, E. Birney, B. Bodenmiller, P. Campbell, P. Carninci, M. Clatworthy, et al. Science forum: the human cell atlas. *Elife*, 6:e27041, 2017.

A. A. Shabalin. Matrix eqtl: ultra fast eqtl analysis via large matrix operations. *Bioinformatics*, 28(10):1353–1358, 2012.

M. Shah, D. Wang, C. Rubadue, D. Suster, and A. Beck. Deep learning assessment of tumor proliferation in breast cancer histological images. In *Bioinformatics and Biomedicine (BIBM), 2017 IEEE International Conference on*, pages 600–603. IEEE, 2017.

V. Subramanian, B. Chidester, J. Ma, and M. N. Do. Correlating cellular features with gene expression using cca. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 805–808. IEEE, 2018.

W. Wang, X. Yan, H. Lee, and K. Livescu. Deep variational canonical correlation analysis. *arXiv preprint arXiv:1610.03454*, 2016.