JAMIE: Joint Variational Autoencoders for Multi-Modal Imputation and Embedding

Noah Cohen Kalafut[1,2], Xiang Huang[2], and Daifeng Wang[1,2,3]

[1] Department of Computer Sciences, Wisconsin, US
[2] Waisman Center, University of Wisconsin-Madison, Wisconsin, US
[3] Department of Biostatistics and Medical Informatics, Wisconsin, US

Corresponding Author: daifeng.wang@wisc.edu

Abstract. Single-cell multi-modal datasets have emerged to measure various characteristics of individual cells, enabling a deep understanding of cellular and molecular mechanisms. However, generating multi-modal data for many cells remains costly and challenging. Thus, missing modalities happens frequently, becoming a major obstacle to understanding mechanisms. Recently, machine learning approaches have been developed to impute cell data but typically use fully matched multi-modal data and learn common latent embeddings that potentially miss modality specificity. To address these issues, we developed a novel machine learning model with open-source tool, Joint variational Autoencoders for Multi-modal Imputation and Embedding (JAMIE). JAMIE takes single-cell multi-modal data that can have partially matched samples across modalities. Variational autoencoders learn the latent embeddings of each modality. Then, embeddings from matched samples across modalities are aggregated to identify joint cross-modal latent embeddings before reconstruction. To perform cross-modal imputation from one to another, the latent embeddings can be used with the opposite decoder. We also use Shapley values to prioritize input features for cross-modal imputation and known sample labels to make JAME interpretable. We applied JAMIE to both simulation data and emerging single-cell multi-modal data such as gene expression, chromatin accessibility, and electrophysiology in human and mouse brains for imputing various modalities and cell types. We found that JAMIE significantly outperforms the state-of-the-arts in most cases and prioritized important multi-modal features for imputation, providing potentially novel mechanistic insights at cellular resolution. JAMIE is open source at https://github.com/daifengwanglab/JAMIE.

Keywords: Multi-modal · Single-cell · Variational autoencoder · Deep learning · Data imputation · Data integration

1 Introduction

Understanding of molecular mechanisms at cellular resolution provides deeper insights into cellular function, development, and disease progression, but remains elusive. To this end, single-cell multi-modal datasets have recently emerged by new sequencing technologies to measure various characteristics of single cells and identify cell functions (e.g., cell types). For instance, patch-seq [2] simultaneously measures gene expression and identifies electrophysiology, and morphological features (beyond omics), e.g., for the mouse visual cortex over 4,000 cells from several brain cell types [8]. [18] profiles single-cell gene expression and chromatin accessibility in the human developing brain. Thus, an integration of single-cell multi-modal datasets can significantly aid in our understanding of biological mechanisms contributing to cell types and diseases through the automated discovery of inter-modal relationships.

Many methods have been developed to integrate multi-modal datasets to improve prediction of cell types and cellular phenotypes (an outline of the process can be seen in Fig. 1a). For instance, our recent DeepManReg [14] approach performs interpretable deep manifold learning which both improves phenotype prediction and is applicable to align single-cell multimodal data. Although these approaches primarily focus on prediction, the underlying idea can be extended to predict data present in separate modalities, referred to as cross-modal imputation going forward (a visual representation can be seen in Fig. 1b). Cross-modal imputation is not new, but has been increasingly explored with the advent of deep learning. For instance,

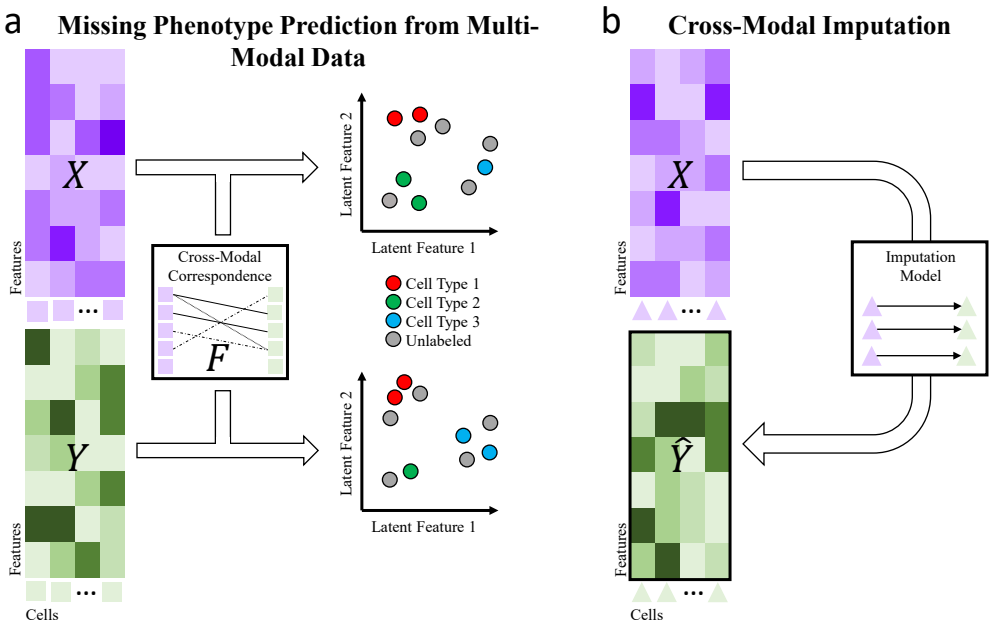

Fig. 1. Challenges for multi-modal data integration and imputation. (a) Using multi-modal data to effectively predict missing phenotypes (e.g. cell-types from multi-modal single-cell data) is difficult due to heterogeneous features across modalities. Identifying similar latent spaces across modalities enables cross-modal comparison and, by extension, missing phenotype prediction. Machine learning (ML) can be used to discover similar cross-modal latent spaces and enable comparison and phenotype prediction. (b) Certain modalities are cost-prohibitive but may lend significant insight into biological mechanisms. As examples, scATAC-seq data for cell-type epigenomics is expensive and electrophysiological data at single-cell resolution is difficult to produce. Imputing one modality from another using ML can alleviate these constraints.

Babel [19] focuses on conversion between RNA and ATAC-seq data matrices through the use of dual autoencoders, which have the particular advantage of being reusable. Moreover, autoencoders allow for flexibility in latent space formulation. Polarbear [20] utilizes a similar approach, focusing on a conversion network between independently generated latent spaces for each modality. Both methods, however, focus specifically on multi-omics data such as scRNA-seq and scATAC-seq and thus don't consider high nonlinearity from additional single-cell modalities like electrophysiology. Moreover, they require fully corresponding multimodal data (matched cells), limiting their capabilities with regards to data with missing modalities.

Also, analyzing multi-modal datasets incurs additional difficulties, including heterogeneous distributions, multicollinearity, and varying reliability. Several approaches have been thus utilized in aligning multi-modal datasets(an outline of the process can be seen in Fig. 1b). For instance, Unioncom [3] infers correspondence information from each modality's distance matrices, then use a modified tSNE method for the final mapping. MMD-MA [11] minimizes an objective function to maximize distribution similarity between datasets while minimizing distortion. GLUE [4] uses an autoencoder model to map differing modalities onto the same latent space. Besides unsupervised learning approaches without prior correspondence information, others rely on correspondence priors to help inform alignment. For example, CLUE [5] assumes completely aligned samples and introduces the impetus for aggregating corresponding latent-space embeddings. ManiNetCluster [13] takes a prior correspondence matrix as input and implements nonlinear manifold alignment by solving an eigenvalue problem. If only partial correspondence information is known, however, existing methods are limiting, and few are designed for such a use-case.

To address these issues, we developed Joint variational Autoencoders for Multi-modal Imputation and Embedding (JAMIE). An outline of JAMIE's capabilities can be found in Table 1. JAMIE trains a reusable dual variational autoencoder model to project available multi-modalities onto similar latent spaces (but still unique for each modality), allowing for enhanced inference of single-modality patterns [7]. To perform cross-modal imputation, data may be fed into an encoder, then the resultant latent space may be processed by the opposite decoder. JAMIE is able to use partial correspondence information. JAMIE combines the reusabil-

Table 1. Method comparison between various integration and imputation methods. JAMIE unifies features from several different integration and imputation methods in a single architecture.

	JAMIE	NLMA	UnionCom	CCA	Babel
Cell Type Separation	✓	✓	✓	✓	
Generalizable Model	✓			✓	✓
Latent Space	✓	✓	✓	✓	✓
Missing Modality Imputation	✓				✓
Model Type					
Autoencoder	✓				✓
Eigenanalysis		✓		✓	
Manifold Alignment		✓			
Projection			✓		
Non-Omics Data Compatibility	✓				
Partial Correspondence Priors	✓	✓		✓	

ity and flexible latent space generation of autoencoders with the automated correspondence estimation of alignment methods. We compared JAMIE to state-of-the-art methods on simulation and emerging single-cell multi-modal data using gene expression, chromatin accessibility, and electrophysiology in human and mouse brains. We found that JAIME significantly outperforms other methods in general and prioritized important multi-modal features for multi-modal imputation, also providing potentially novel mechanistic insights at cellular resolution.

2 Methods

Fig. 2a shows the training process of JAMIE. Two numeric data matrices of modalities X and Y are used as input. $\boldsymbol{X} \in \mathbb{R}^{d_{\boldsymbol{X}} \times n_{\boldsymbol{X}}}$ is a data matrix of modality X with $d_{\boldsymbol{X}}$ features and $n_{\boldsymbol{X}}$ samples. The ith column of $\boldsymbol{X}$, i.e. $\boldsymbol{x}_i \in \mathbb{R}^{d_{\boldsymbol{X}}}$, is the feature vector of the ith sample of modality X. $\boldsymbol{Y} \in \mathbb{R}^{d_{\boldsymbol{Y}} \times n_{\boldsymbol{Y}}}$ is a data matrix of modality Y with $d_{\boldsymbol{Y}}$ features and $n_{\boldsymbol{Y}}$ samples. Similarly, the ith column of $\boldsymbol{Y}$, i.e. $\boldsymbol{y}_i \in \mathbb{R}^{d_{\boldsymbol{Y}}}$, is the feature vector of the ith sample of modality Y. The features of $\boldsymbol{X}$ and $\boldsymbol{Y}$ are standardized to 0 mean and standard deviation 1. If available, a prior correspondence matrix $\boldsymbol{F} \in \mathbb{R}_{\geq 0}^{n_{\boldsymbol{X}} \times n_{\boldsymbol{Y}}}$ is used to aid construction of the similar latent spaces during training, where $\boldsymbol{F}_{ij} = 1$ implies one-to-one correspondence (i.e. same cells different modalities) between cell i in modality X and cell j in modality Y, $\boldsymbol{F}_{ij} = 0$ implies no known correspondences, and $0 < \boldsymbol{F}_{ij} < 1$ implies partial confidence of correspondence.

Jamie utilizes joint variational autoencoders to learn similar latent spaces for each of the two modalities, $\boldsymbol{L}^{\boldsymbol{X}} \in \mathbb{R}^{g \times n_{\boldsymbol{X}}}$ and $\boldsymbol{L}^{\boldsymbol{Y}} \in \mathbb{R}^{g \times n_{\boldsymbol{X}}}$, where g is the selected dimensionality of the latent space. Higher g will lead to better reconstruction loss, but the features will not be as dense information-wise. Lower g will condense information more efficiently, but at the cost of reconstruction and imputation performance. We experimentally found $g = 32$ to be robust across multiple datasets, and use this as the default for JAMIE. During training, we utilize correlation-based latent aggregation. In order to learn these latent spaces, JAMIE minimizes the following losses:

$$\ell_{\text{total}} = \ell_{\text{kl}} + \alpha \ell_{\text{combination}} + \beta \ell_{\text{reconstruction}} + \gamma \ell_{\text{alignment}}. \tag{1}$$

ℓ_{kl} takes the Kullback-Leibler (KL) divergence between the inferred distribution of our variational autoencoders (VAEs) and the multivariate standard normal, aiding with the continuity of the latent space. $\ell_{\text{combination}}$ enforces similarity of corresponding samples and takes the weight α. $\ell_{\text{reconstruction}}$ is the summed mean squared error between the reconstructions and original data matrices and is assigned the weight β. $\ell_{\text{alignment}}$ uses inferred cross-modal correspondence to shape the resultant latent space and is assigned the weight γ. The weights are user-defined and default to $(\alpha, \beta, \gamma) = (1000, 31.25, 31.25)$. An overview of the JAMIE algorithm can be seen in Algorithm 1.

2.1 Distribution of Latent Space Features

Intermediate feed-forward neural networks are utilized to generate distributions for each feature in the g-dimensional latent space for each modality. These will be henceforth referred to as encoders, denoted

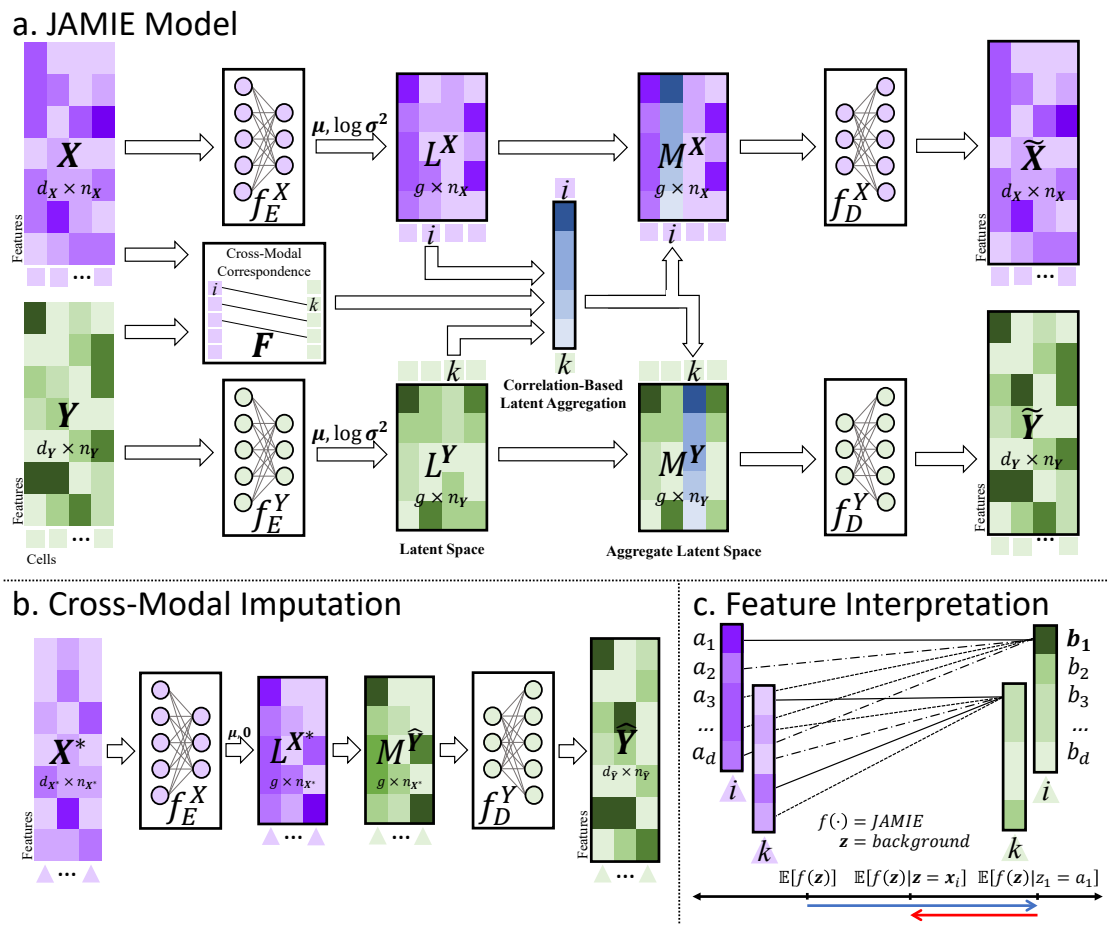

Fig. 2. JAMIE: Joint variational Autoencoders for Multi-modal Imputation and Embedding uses variational autoencoders with a novel latent space aggregation technique in order to generate similar latent spaces for each modality. (a) As input, two data matrices $\boldsymbol{X}, \boldsymbol{Y}$ are provided with optional *prior cross-modal correspondence* matrix $\boldsymbol{F}$. These data matrices are fed through encoding layers $f_E^X(\cdot), f_E^Y(\cdot)$ which provide $\boldsymbol{\mu}_i^X, \boldsymbol{\mu}_j^Y$ and $\boldsymbol{\sigma}_i^X, \boldsymbol{\sigma}_j^Y$ for all samples i in $\boldsymbol{X}$ and j in $\boldsymbol{Y}$. Then, $\boldsymbol{L}_i^X \sim \mathcal{N}(\boldsymbol{\mu}_i^X, \boldsymbol{\sigma}_i^X)$ and $\boldsymbol{L}_i^Y \sim \mathcal{N}(\boldsymbol{\mu}_i^Y, \boldsymbol{\sigma}_i^Y)$ are sampled for all i and j to produce the latent spaces $\boldsymbol{L}^{\boldsymbol{X}}, \boldsymbol{L}^{\boldsymbol{Y}}$. Latent spaces are aggregated using $\boldsymbol{F}$ to make $\boldsymbol{M}^{\boldsymbol{X}}, \boldsymbol{M}^{\boldsymbol{Y}}$. Finally, the decoding layers produce a reconstructed version of the original modalities $\widetilde{\boldsymbol{X}} = f_D^X(\boldsymbol{M}^{\boldsymbol{X}}), \widetilde{\boldsymbol{Y}} = f_D^Y(\boldsymbol{M}^{\boldsymbol{Y}})$. The latent embeddings $\boldsymbol{L}^{\boldsymbol{X}}, \boldsymbol{L}^{\boldsymbol{Y}}$ of the two modalities can be used in tandem for missing phenotype prediction. (b) The trained model can be reused for cross-modal imputation through coupling of opposing encoders and decoders. (c) The imputing function $f(\boldsymbol{x}) = f_D^Y(L^{\boldsymbol{X}}) = \boldsymbol{y}$, $\boldsymbol{x} \in \mathbb{R}^{d_{\boldsymbol{X}}}$, $\boldsymbol{y} \in \mathbb{R}^{d_{\boldsymbol{Y}}}$ is assessed using Shapley Additive Explanation values [12] which estimate contribution of each input feature by selectively masking the input feature vector with the background.

$f_E^X(\cdot), f_E^Y(\cdot)$ for each modality. The variational encoders have four hidden layers. The encoder transforms the feature dimension as follows: $n_{\boldsymbol{X}} \to 2n_{\boldsymbol{X}} \to n_{\boldsymbol{X}} \to g \to 2g$. For modality X, the top half of the fourth layer output corresponds to means and the bottom half corresponds to log variance per sample and feature as in Equation 2a. The encoder is similar for Y. Batch normalization and leaky ReLU activations are included wherever possible, and dropout layers are added ($p = 0.6$) for imputation models. For each sample $\boldsymbol{x}_i$ and $\boldsymbol{y}_j$, we sample the multivariate Gaussian

$$\left[f_E^X(\boldsymbol{X})\right]_i = \begin{bmatrix} \boldsymbol{\mu}_i^{\boldsymbol{X}} \\ \boldsymbol{\sigma}_i^{\boldsymbol{X}} \end{bmatrix} \quad \boldsymbol{L}_i^{\boldsymbol{X}} \sim \mathcal{N}(\boldsymbol{\mu}_i^{\boldsymbol{X}}, \mathrm{diag}(\boldsymbol{\sigma}_i^{\boldsymbol{X}})) \quad \boldsymbol{\mu}_i^{\boldsymbol{X}}, \boldsymbol{L}_i^{\boldsymbol{X}} \in \mathbb{R}^g \quad \boldsymbol{\sigma}_i^{\boldsymbol{X}} \in \mathbb{R}_{>0}^g \quad i \in \{1, 2, \ldots, n_{\boldsymbol{X}}\}, \tag{2a}$$

$$\left[f_E^Y(\boldsymbol{Y})\right]_j = \begin{bmatrix} \boldsymbol{\mu}_j^{\boldsymbol{Y}} \\ \boldsymbol{\sigma}_j^{\boldsymbol{Y}} \end{bmatrix} \quad \boldsymbol{L}_j^{\boldsymbol{Y}} \sim \mathcal{N}(\boldsymbol{\mu}_j^{\boldsymbol{Y}}, \mathrm{diag}(\boldsymbol{\sigma}_j^{\boldsymbol{Y}})) \quad \boldsymbol{\mu}_j^{\boldsymbol{Y}}, \boldsymbol{L}_j^{\boldsymbol{Y}} \in \mathbb{R}^g \quad \boldsymbol{\sigma}_j^{\boldsymbol{Y}} \in \mathbb{R}_{>0}^g \quad j \in \{1, 2, \ldots, n_{\boldsymbol{Y}}\}, \tag{2b}$$

where $\left[f_E^X(\boldsymbol{X})\right]_i$ and $\left[f_E^Y(\boldsymbol{Y})\right]_i$ are the encoded results of $\boldsymbol{x}_i$ and $\boldsymbol{y}_i$, respectively, and $\mathrm{diag}(\cdot)$ is a diagonal matrix generated from a vector. $\boldsymbol{L}_i^{\boldsymbol{X}}$ and $\boldsymbol{L}_i^{\boldsymbol{Y}}$ are sampled from the multivariate Gaussian distributions

indicated above. We define $\boldsymbol{L}_i^{\boldsymbol{X}} \in \mathbb{R}^g$ as the ith column of $\boldsymbol{L}^{\boldsymbol{X}}$, representing the feature vector of the ith sample. $\boldsymbol{L}_i^{\boldsymbol{Y}} \in \mathbb{R}^g$ is defined similarly.

2.2 Aggregation of the Latent Spaces

Using our prior correspondence matrix $\boldsymbol{F} \in \mathbb{R}^{n_{\boldsymbol{X}} \times n_{\boldsymbol{Y}}}$, we can perform an aggregate calculation to combine the latent space vectors of known aligned points. In particular,

$$\boldsymbol{M}^{\boldsymbol{X}} = \left(\boldsymbol{L}^{\boldsymbol{X}} + \delta \boldsymbol{L}^{\boldsymbol{Y}} \boldsymbol{F}^T\right) \left(\operatorname{diag}\left(1 + \delta \boldsymbol{F} \mathbf{1}_{n_y}\right)\right)^{-1}, \tag{3a}$$

$$\boldsymbol{M}^{\boldsymbol{Y}} = \left(\boldsymbol{L}^{\boldsymbol{Y}} + \delta^{-1} \boldsymbol{L}^{\boldsymbol{X}} F\right) \left(\operatorname{diag}\left(1 + \delta^{-1} \boldsymbol{F}^T \mathbf{1}_{n_x}\right)\right)^{-1}, \tag{3b}$$

where the aggregate feature vector $\boldsymbol{M}^{\boldsymbol{X}} \in \mathbb{R}^{g \times n_{\boldsymbol{X}}}$ has g features and $n_{\boldsymbol{X}}$ samples and $\boldsymbol{M}^{\boldsymbol{Y}} \in \mathbb{R}^{g \times n_{\boldsymbol{Y}}}$ has g features and $n_{\boldsymbol{Y}}$ samples. $\mathbf{1}_{n_x} \in \mathbb{R}^{n_x}$ and $\mathbf{1}_{n_y} \in \mathbb{R}^{n_y}$ are vectors of all ones. δ refers to the relative weighting between modalities X and Y in the aggregation function. By default, $\delta = 1$. A value $\delta < 1$ implies that $\boldsymbol{X}$ is weighted more than $\boldsymbol{Y}$ during aggregation, while $\delta > 1$ implies the opposite. The construction of $\boldsymbol{M}^{\boldsymbol{X}}, \boldsymbol{M}^{\boldsymbol{Y}}$ is shown visually in Fig. 2a in the formulation of the blue vector. Fig. 2a shows the case $\boldsymbol{F}_{ik} = 1$ with all other entries in row i and column k being 0, which results in simply averaging the feature vectors $\boldsymbol{x}_i$ and $\boldsymbol{y}_k$.

The linear aggregation technique is backed by the nonlinear transformation of the encoder, which is able to account for varying timings, magnitude changes, and distributions due to its deep structure. In Fig. 2, this is the blue latent space vector combination of corresponding samples. We term the technique *correlation-based latent aggregation*. This moves corresponding latent embeddings in similar directions over the course of training and is key in the formation of similar latent spaces. Correlation-based latent aggregation is the primary motivation for the use of the variational autoencoder framework, as it allows for this intuitive representation of partially aligned datasets. We can now adjust our low-dimensional data embeddings by using aggregates wherever possible to produce our final aggregate latent spaces $\boldsymbol{M}^{\boldsymbol{X}}$ and $\boldsymbol{M}^{\boldsymbol{Y}}$.

2.3 Continuity of the Aggregate Latent Spaces

Continuous $\boldsymbol{L}^{\boldsymbol{X}}$ and $\boldsymbol{L}^{\boldsymbol{Y}}$ facilitate imputation, thus, we adopt the VAE architecture [10] and the corresponding KL loss:

$$\ell_{\text{kl}} = \frac{\kappa}{g} \sum_{i=0}^{g} \boldsymbol{d}_i, \tag{4}$$

where $\boldsymbol{d}$ takes the KL divergence between the distribution of our latent space estimate and a multivariate standard normal as the targeted distribution:

$$\boldsymbol{d} = -\frac{1}{2n_{\boldsymbol{X}}} \sum_{i=0}^{n_{\boldsymbol{X}}} \left(1 + 2\log(\boldsymbol{\sigma}_i^{\boldsymbol{X}}) - (\boldsymbol{\mu}_i^{\boldsymbol{X}})^2 - (\boldsymbol{\sigma}_i^{\boldsymbol{X}})^2\right) - \frac{1}{2n_{\boldsymbol{Y}}} \sum_{i=0}^{n_{\boldsymbol{Y}}} \left(1 + 2\log(\boldsymbol{\sigma}_i^{\boldsymbol{Y}}) - (\boldsymbol{\mu}_i^{\boldsymbol{Y}})^2 - (\boldsymbol{\sigma}_i^{\boldsymbol{Y}})^2\right), \tag{5}$$

and κ is slowly increased from 0 to 1 over the course of training to accelerate convergence [1], further detailed in Subsection 2.7. KL divergence is described in Section 3.

ℓ_{kl} will lower to zero as the latent space's distribution approaches the standard multivariate Gaussian. This encourages a continuous latent space through the nonzero standard deviations. The mean approaching zero encourages a small, clustered latent space. This will negatively affect metrics such as silhouette distance, but can aid in imputing outlier single-cell data.

2.4 Similarity of the Original and Aggregate Latent Spaces

The structure alone with JAMIE already produces shared latent spaces $\boldsymbol{M}^{\boldsymbol{X}}$ and $\boldsymbol{M}^{\boldsymbol{Y}}$. To balance $\boldsymbol{L}^{\boldsymbol{X}}$ and $\boldsymbol{L}^{\boldsymbol{Y}}$ and keep both latent spaces the same scale, we implement a loss proportional to the square norm of the difference between each latent space and its aggregated matrix $\boldsymbol{M}$:

$$\ell_{\text{combination}} = \frac{1}{gn_{\boldsymbol{X}}} ||\boldsymbol{L}^{\boldsymbol{X}} - \boldsymbol{M}^{\boldsymbol{X}}||_F^2 + \frac{1}{gn_{\boldsymbol{Y}}} ||\boldsymbol{L}^{\boldsymbol{Y}} - \boldsymbol{M}^{\boldsymbol{Y}}||_F^2. \tag{6}$$

2.5 Similarity of the Original and Reconstructed Data Matrices

Decoders are structured in the reverse manner as encoders. It uses fully connected layers to transform the feature dimension as follows: $g \to n_{\boldsymbol{X}} \to 2n_{\boldsymbol{X}} \to n_{\boldsymbol{X}}$. The structure is similar for Y. Batch normalization and leaky ReLU are included where possible, and dropout layers are added ($p = 0.6$) for imputation models. We denote the decoders $f_D^X(\cdot), f_D^Y(\cdot)$ for each modality and use each to get an estimate of the original datasets $\widetilde{\boldsymbol{X}} = f_D^X(\boldsymbol{M}^{\boldsymbol{X}}) \in \mathbb{R}^{d_{\boldsymbol{X}} \times n_{\boldsymbol{X}}}$ and $\widetilde{\boldsymbol{Y}} = f_D^Y(\boldsymbol{M}^{\boldsymbol{Y}}) \in \mathbb{R}^{d_{\boldsymbol{Y}} \times n_{\boldsymbol{Y}}}$. The quality of the reconstruction, and thereby the information retained in the latent space, can be evaluated by:

$$\ell_{\text{reconstruction}} = \frac{1}{d_{\boldsymbol{X}} n_{\boldsymbol{X}}} ||\boldsymbol{X} - \widetilde{\boldsymbol{X}}||_F^2 + \frac{1}{d_{\boldsymbol{Y}} n_{\boldsymbol{Y}}} ||\boldsymbol{Y} - \widetilde{\boldsymbol{Y}}||_F^2. \tag{7}$$

2.6 Correspondence of the Aggregate Latent Spaces

We find the optimal inferred correspondence matrix $\widetilde{\boldsymbol{F}} \in \mathbb{R}_{\geq 0}^{n_{\boldsymbol{X}} \times n_{\boldsymbol{Y}}}$ by solving the following optimization problem:

$$\begin{aligned} &\min_{\alpha, \overline{\boldsymbol{F}}} ||\alpha \boldsymbol{K}_X - \overline{\boldsymbol{F}} \boldsymbol{K}_Y \overline{\boldsymbol{F}}^T||_F^2 \\ &\text{s.t.} \quad \overline{\boldsymbol{F}} \geq 0, \quad \alpha \geq 0, \\ &\overline{\boldsymbol{F}} \mathbf{1}_{n_y} = 1_{n_x}, \quad \overline{\boldsymbol{F}}^T \mathbf{1}_{n_x} \leq 1_{n_y}, \end{aligned} \tag{8}$$

where $[\boldsymbol{K}_X, \boldsymbol{K}_Y] = [dist(\boldsymbol{X}), dist(\boldsymbol{Y})]$ are the intra-dataset distance. This generally uses geodesic distance and is similar to the quadratic assignment problem.

The procedure originates from an extension on [6], which optimized a similar expression but assumed one-to-one correspondence on inter-dataset cell pairs. [3] extended this to the concept of one-to-many correspondence between modalities and created the implementation in Equation 8.

Once solved, $\widetilde{\boldsymbol{F}}_{ij}$ denotes the probability that sample $\boldsymbol{x}_i$ is matched with sample $\boldsymbol{y}_j$. Any unsupervised correspondence method may be substituted for $\widetilde{\boldsymbol{F}}$. The objective of $\widetilde{\boldsymbol{F}}$ is to provide complete correspondence information. $\boldsymbol{F}$ will generally contain only one-to-one relationships, while true inter-modal relationships are often layered and complex. In practice, we may wish to combine $\boldsymbol{F}, \widetilde{\boldsymbol{F}}$ if $\boldsymbol{F}$ is especially sparse. Then, we use $\boldsymbol{F} \leftarrow \rho \boldsymbol{F} + (1 - \rho)\widetilde{\boldsymbol{F}}$ for the mixture ratio ρ. This new $\boldsymbol{F}$ is then used in Equation 3a and Equation 3b. By default, ρ is 1.

To ensure that corresponding samples have similar aggregate latent spaces we introduce $\ell_{\text{alignment}}$. This metric relies on our correspondence estimate $\widetilde{\boldsymbol{F}}$ rather than the prior $\boldsymbol{F}$. Unlike our prior $\boldsymbol{F}$, our correspondence estimate $\widetilde{\boldsymbol{F}}$ contains information about samples which do not correspond. We can utilize this dense information to shape our latent space such that contributing samples are clustered:

$$\ell_{\text{alignment}} = \frac{1}{g n_{\boldsymbol{X}}} ||\boldsymbol{M}^{\boldsymbol{X}} - \boldsymbol{M}^{\boldsymbol{Y}} \widetilde{\boldsymbol{F}}^T||_F^2. \tag{9}$$

2.7 Early Stopping and KL Annealing

Early stopping is performed by keeping track of the best ℓ_{total} over the course of training. JAMIE is given a minimum number of epochs (default, $2,500$) over which time no early stopping can take place. Over this period, ℓ_{kl} is gradually increased from 0 to 1. The concept originates from [1].

Once this period is over, the training continues until the algorithm does not produce a new best loss for a set number of epochs (default, 500) or until the maximum number of epochs is achieved (default, $10,000$).

This process is further detailed in Algorithm 1.

2.8 JAMIE Algorithm

Datasets $\boldsymbol{X}$ and $\boldsymbol{Y}$ are provided pre-standardized along with an optional correspondence matrix $\boldsymbol{F}$ as described in Section 2. The provided variables are $\tau_{\min}, \tau_{\max}$ with default values $2,500$ and $10,000$, which are the minimum and maximum number of epochs for training. b_s is the batch size, defaulted to 512. (α, β, γ) are the weights for the model with defaults $(1000, 31.25, 31.25)$. ρ controls how much the inferred correspondence matrix is utilized over the priors. ρ controls the mixture between the inferred and prior correspondence matrices and defaults to 1 (full prior). $\tau_{\text{max_lapses}}, \ell_{\text{min_change}}$ control the behavior of the early-stopping algorithm. These have default values of 500 and 1e-8.

Algorithm 1 JAMIE

Input: $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{F}$
Params: $\tau_{\text{min}}, \tau_{\text{max}}$, b_s, (α, β, γ), ρ, $\tau_{\text{max_lapses}}, \ell_{\text{min_change}}$
Output: $f_E^X(\cdot), f_E^Y(\cdot), f_D^X(\cdot), f_D^Y(\cdot), \boldsymbol{L^X}, \boldsymbol{L^Y}$

$[\boldsymbol{K_X}, \boldsymbol{K_Y}] = [dist(\boldsymbol{X}), dist(\boldsymbol{Y})]$
$\widetilde{\boldsymbol{F}} = \arg\min_{\overline{\boldsymbol{F}},\alpha} ||\alpha \boldsymbol{K_X} - \overline{\boldsymbol{F}} \boldsymbol{K_Y} \overline{\boldsymbol{F}}^T||_F^2$
$\boldsymbol{F} \leftarrow (1-\rho)\boldsymbol{F} + \rho\widetilde{\boldsymbol{F}}$
$\ell_{\text{best}} \leftarrow \infty$, $\tau_{\text{lapses}} \leftarrow 0$
$\tau_{\text{current}} \leftarrow 0$
Initialize $f_E^X, f_D^X, f_E^Y, f_D^Y$ with random weights
while $\tau_{\text{current}} < \tau_{\text{max}}$ and $\tau_{\text{lapses}} < \tau_{\text{max_lapses}}$ **do**
  Randomly sample $\boldsymbol{X}_s \in \mathbb{R}^{d_{\boldsymbol{X}} \times b_s}$ from $\boldsymbol{X} \in \mathbb{R}^{d_{\boldsymbol{X}} \times n_{\boldsymbol{X}}}$ and $\boldsymbol{Y}_s \in \mathbb{R}^{d_{\boldsymbol{Y}} \times b_s}$ from $\boldsymbol{Y} \in \mathbb{R}^{d_{\boldsymbol{Y}} \times n_{\boldsymbol{Y}}}$
  Extract submatrix $\boldsymbol{F}_s \in \mathbb{R}^{b_s \times b_s}$ from $\boldsymbol{F} \in \mathbb{R}^{n_{\boldsymbol{X}} \times n_{\boldsymbol{Y}}}$ and $\widetilde{\boldsymbol{F}}_{\boldsymbol{s}} \in \mathbb{R}^{b_s \times b_s}$ from $\widetilde{\boldsymbol{F}} \in \mathbb{R}^{n_{\boldsymbol{X}} \times n_{\boldsymbol{Y}}}$
  Optionally transform $\boldsymbol{X}_s, \boldsymbol{Y}_s$ using PCA trained on $\boldsymbol{X}, \boldsymbol{Y}$
  Normalize columns of $\boldsymbol{F}_s$ and $\widetilde{\boldsymbol{F}}_{\boldsymbol{s}}$ to summation 1
  Randomly sample $L_i^{\mathbf{X_s}}$ from distributions $f_E^X(\boldsymbol{X}_s)$ and $L_j^{\mathbf{Y_s}}$ from distributions $f_E^Y(\boldsymbol{Y}_s)$ for $i, j \in \{1, 2, \ldots, b_s\}$
  $\boldsymbol{M}^{\boldsymbol{X}_s} \leftarrow \left(\boldsymbol{L}^{\boldsymbol{X}_s} + \delta \boldsymbol{L}^{\boldsymbol{Y}_s} \boldsymbol{F}_s^T\right) \left(\text{diag}\left(1 + \delta \boldsymbol{F}_s \mathbf{1}_{b_s}\right)\right)^{-1}$
  $\boldsymbol{M}^{\boldsymbol{Y}_s} \leftarrow \left(\boldsymbol{L}^{\boldsymbol{Y}_s} + \delta^{-1} \boldsymbol{L}^{\boldsymbol{X}_s} \boldsymbol{F}_s\right) \left(\text{diag}\left(1 + \delta^{-1} \boldsymbol{F}_s^T \mathbf{1}_{b_s}\right)\right)^{-1}$
  $\widetilde{\boldsymbol{X}}_{\boldsymbol{s}} \leftarrow f_D^X(\boldsymbol{M}^{\boldsymbol{X}_s})$ and $\widetilde{\boldsymbol{Y}}_{\boldsymbol{s}} \leftarrow f_D^Y(\boldsymbol{M}^{\boldsymbol{Y}_s})$
  If $\boldsymbol{X}_s, \boldsymbol{Y}_s$ were transformed using PCA, perform inverse transform on $\widetilde{\boldsymbol{X}}_{\boldsymbol{s}}, \widetilde{\boldsymbol{Y}}_{\boldsymbol{s}}$
  $\kappa \leftarrow \left(1 + \exp\left(\frac{-5(\tau_{\text{current}} - 0.5\tau_{\text{min}})}{0.5\tau_{\text{min}}}\right)\right)^{-1}$
  Calculate ℓ_{kl} via Equation 4
  Calculate $\ell_{\text{combination}}$ via Equation 6
  Calculate $\ell_{\text{reconstruction}}$ via Equation 7
  Calculate $\ell_{\text{alignment}}$ via Equation 9
  $\ell_{\text{total}} \leftarrow \ell_{\text{kl}} + \alpha\ell_{\text{combination}} + \beta\ell_{\text{reconstruction}} + \gamma\ell_{\text{alignment}}$
  if $\tau_{\text{current}} \geq \tau_{\text{min}}$ **then**
    if $\ell_{\text{total}} < \ell_{\text{best}} - \ell_{\text{min_change}}$ **then**
      $\ell_{\text{best}} \leftarrow \ell_{\text{total}}$
      $\tau_{\text{lapses}} \leftarrow 0$
    else
      $\tau_{\text{lapses}} \leftarrow \tau_{\text{lapses}} + 1$
    end if
    Update $f_E^X, f_D^X, f_E^Y, f_D^Y$ using gradient descent on ℓ_{total}
  end if
  $\tau_{\text{current}} \leftarrow \tau_{\text{current}} + 1$
end while
$\boldsymbol{L^X} \leftarrow f_E^X(\boldsymbol{X})$, $\boldsymbol{L^Y} \leftarrow f_E^Y(\boldsymbol{Y})$

3 Experimental Setup and Applications

3.1 Performance Evaluation

Label Transfer Accuracy (LTA): To assess cell type separation, we use label transfer accuracy (LTA) [9, 3]. LTA measures the cell-identification accuracy of a nearest neighbor classifier trained on the second modality and evaluated on the first. Higher values generally indicate better cell-type separation.

Fraction of Samples Closer to the True Match (FOSCTTM): We use this metric to evaluate the average fraction of samples closer to the truly aligned sample (in both modalities) than its actual pairing. This provides an estimate of how closely the two modalities are aligned. Lower values generally indicate closer cross-modal alignment.

Pearson Correlation: We use Pearson correlation mainly to compare imputed and measured feature values, providing an estimate of how well the imputation method captures the variance of the measured feature values. For the purposes of imputation, higher values usually indicate better performance.

Area under the ROC curve (AUROC): We use AUROC to evaluate imputed features for prediction of high or low measured expression values. To test this, we use AUROC on the median-binarized measured data, that is, each feature value above the feature median is treated as 1, and all others are treated as 0.

Jensen-Shannon Distance (JS distance): It is also important that an imputation method preserves individual cell-type distributions. We use JS distance, which is the square root of the Jensen-Shannon divergence. Jensen-Shannon divergence is the average KL divergence between two distributions and their average (a symmetrized version of KL divergence). KL divergence measures the difference between two probability distributions, and increases proportionally. As a baseline for imputation performance, a KNN model is used to generate new multi-modal data based on the average of nearest neighbors in the training set.

p-Values: When we compare two methods in terms of imputation performance, we utilize the two-tailed binomial test with the null hypothesis that both methods have an equal probability of imputing better on any given feature.

3.2 Hyperparameters and Validation

Principal Component Analysis (PCA) is used to lower the dimensionality of data with many features. The default value for this preprocessing is 512 features, and is automatically inverted when applying imputation. The hyperparameters used for all experiments in the paper are the default losses detailed in Section 2, the minimal and maximal iterations detailed in Subsection 2.7, and a batch size of 512.

When performing integration in Section 4, no data is withheld, hyperparameters are left at default, and dropout is set to $p = 0$ to facilitate maximal learning of single-modal nuances, regardless of dataset specificity. This is done because many alignment methods are unable to be reused, and perform integration only once before having to be recalculated. For imputation, hyperparameters are left unchanged but dropout is set at $p = 0.6$ to facilitate generalization. 20 percent of cells within each dataset are withheld at random. Any imputation performance calculations, such as those in Fig. 3f, are performed on this withheld data.

3.3 Datasets and Data Processing

We applied JAMIE to four single-cell multimodal datasets: (1) simulated multimodal data generated by sampling from a Gaussian distribution on a branching manifold in MMD-MA [11]; (2) Patch-seq gene expression and electrophysiology data of single neuronal cells in the mouse visual cortex (N=3,654, 6 cell types) [8] and mouse motor cortex (N=1,208, 9 cell types) [16]; (3) 10x Multiome gene expression and chromatin accessibility data of 8,981 cells in the human developing brain [18] (21 postconceptional weeks, covering 7 major cell types in human cereberal cortex); (4) scRNA-seq gene expression and scATAC-seq chromatin accessibility data of 4,301 cells from the COLO-320DM colon adenocarcinoma cell line [15] (which was also analyzed by Babel [17]). The input datasets to JAMIE were standardized (mean = 0, standard deviation = 1) across features and large numbers of input features were reduced dimensionality by PCA to 512 PCs as needed for efficient computation.

3.4 Applications

Phenotype Prediction from Multi-Modal Data: Integration on multi-modal datasets can improve classification, knowledge of phenotype, and understanding of complex biological mechanisms. Given two datasets $\boldsymbol{X}, \boldsymbol{Y}$ and correspondence $\boldsymbol{F}$, we generate $\boldsymbol{L^X}, \boldsymbol{L^Y}$ from the trained encoders $f_E^X(\cdot), f_E^Y(\cdot)$ as in Subsection 2.1. Clustering can be performed or classification methods may be directly built atop the reusable latent space. Clustering on these latent spaces confers several advantages, the main of which being the incorporation of both modalities in the process of feature generation. We can then actively predict which samples correspond (if not already known) and perform related tasks such as cell-type prediction. On partially-annotated datasets, cells in similar clusters should be of similar cell-type. There is generally no need to use a complicated clustering or classification algorithm, as JAMIE should do most of the separation as a part of the latent space generation. The networks may be further analyzed to provide a clearer picture of the relationship between data features and phenotype.

Cross-Modal Imputation: There are several methods to perform cross-modal imputation, but many methods are not able to prove that they have learned underlying biological mechanisms for the purpose of imputation. When utilized for cross-modal imputation, we can predict our missing data with a more rigorous foundation than that offered by a feed-forward network or linear regression. Given training data $\boldsymbol{X}, \boldsymbol{Y}$, we

can train the model. With new data $\boldsymbol{X}^* \in \mathbb{R}^{d_{\boldsymbol{X}} \times n_{\boldsymbol{X}^*}}$, we can predict its corresponding latent embedding $\widehat{\boldsymbol{Y}} = f_D^Y(L^{\boldsymbol{X}^*}) \in \mathbb{R}^{d_{\boldsymbol{Y}} \times n_{\boldsymbol{Y}^*}}$ which has true values Y^*. Then, we have predicted corresponding cells using a latent space which likely confers understanding of the relationship between data features and phenotype.

This has motivation within the given losses. The alignment loss incentivizes the low-dimensional embeddings for each modality to be on the same latent space. Additionally, since we are predicting corresponding cells, $\boldsymbol{F} = \mathbb{I}$. Ideally then, $\boldsymbol{L}^{\boldsymbol{X}^*} = \boldsymbol{M}^{\boldsymbol{X}^*} = \boldsymbol{M}^{\boldsymbol{Y}^*} = \boldsymbol{L}^{\boldsymbol{Y}^*}$. So, if we assume perfect and total alignment, $f_D^Y(\boldsymbol{L}^{\boldsymbol{X}^*}) = f_D^Y(\boldsymbol{L}^{\boldsymbol{Y}^*})$. This prediction is shown in Fig. 2b. Of course, this ideal scenario is not always the case. The efficacy of this cross-modal imputation method is explored within the results section of this paper, where 20 percent of the data is taken as validation, and the model is trained on the remainder.

Latent and Imputed Feature Interpretation: For interpretation of our trained model, we employ [12], henceforth referred to as SHAP. SHAP looks at individual predictions produced by a model and will assess importance of various input features through sample modulation. This can be used for a variety of interesting applications. If a target variable is easily separable by phenotype, SHAP can determine relevant features for further study. Additionally, if we perform imputation, SHAP can expose inter-modal linkages learned by the model. Given a model f and sample $\boldsymbol{x} \in \mathbb{R}^n$, SHAP values $\phi_0, \ldots, \phi_{n-1}$ are learned such that $\mathbb{E}[f(\boldsymbol{z})] + \sum_{i=\{j,\ldots,m\}} \phi_i = \mathbb{E}[f(\boldsymbol{z})|\boldsymbol{z}_j = x_j, \ldots, \boldsymbol{z}_m = x_m]$ for a background feature vector $\boldsymbol{z} \in \mathbb{R}^n$. If $\{j, \ldots, m\} = \{0, 1, \ldots, n-1\}$, then the sum of SHAP values and the background output will equal $f(\boldsymbol{x})$, with each ϕ_i proportional to the impact of x_i on the model output.

Another useful technique involves choosing a key metric (i.e. LTA, Correspondence between imputed and measured features) and evaluating the metric with each feature removed (replaced with its background) from the model. As an example, in Fig. 7, the DENND1B gene is imputed from chromatin accessibility data. Various regions are withheld and the correlation between the imputed and measured expression values are recorded. Then, the worse the key metric, the more important the removed feature is to the function of the model.

4 Results

4.1 Simulated Multimodal Data

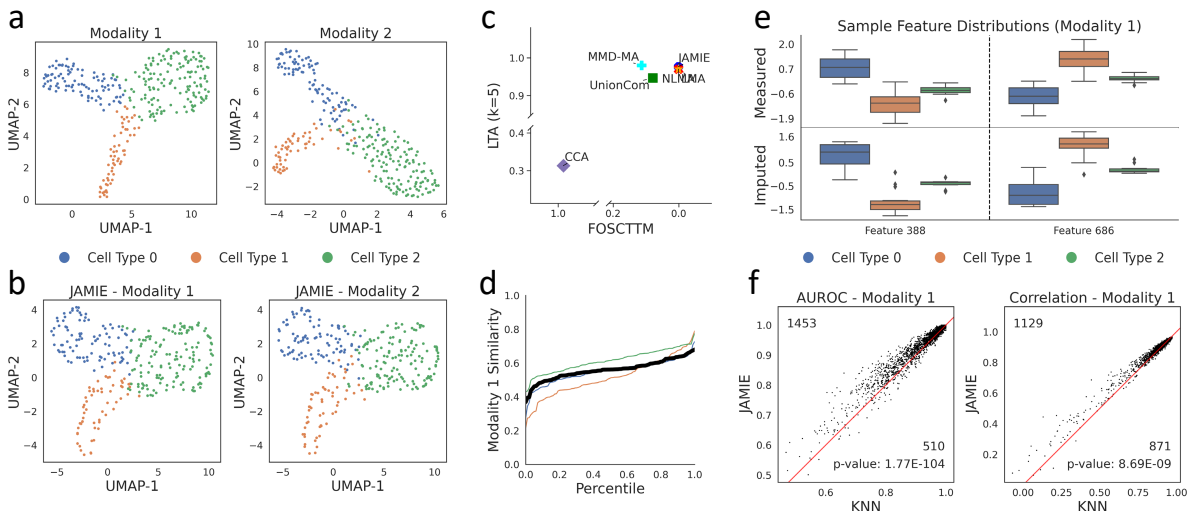

Fig. 3. Simulated multimodal data. (a) UMAPS of original spaces colored by cell types.(b) UMAPs of JAMIE latent spaces. (c) Fraction of samples closer to the true mean (x-axis) and label transfer accuracy (y-axis) of JAMIE and state-of-the-art methods for cell-type separation. (d) Cumulative distributions of similarity (1 − JS distance) between measured and imputed feature values in Modality 1. (e) Measured (top) and imputed (bottom) values of two select features in Modality 1 across cell types. (f) Imputation performance for Modality 1 of JAMIE versus a baseline KNN by AUROC and correlation Section 3.

We first tested JAMIE on simulated single-cell multi-modal data [11]. The simulation data was generated by sampling from a Gaussian distribution on a branching manifold (Fig. 3a). We found that the latent embeddings of two modalities in JAMIE preserve the branching structure of the manifold while aligning the cells of the same type in either modality and maintains cell-type separation (Fig. 3b). Moreover, forseparating cell types, JAMIE (LTA= 0.976, FOSCTTM= 0.001) outperforms state of the art alignment methods NLMA (LTA= 0.970, FOSCTTM= 0.001) and UnionCom (LTA= 0.947, FOSCTTM= 0.079) (Fig. 3c).

Also, we found that the imputed feature values by JAMIE are consistent with the measurements. For instance, as shown in Figure 2d, the imputed values of Modality 1 features have high distribution similarity with the measured values across cell types with average JS distance of 0.441 ± 0.096 (Fig. 3d). Two features with high similarity are highlighted with average JS distances 0.277 and 0.266 (Fig. 3e), also showing a preservation of expression changes across cell types (i.e., cell type 1 has a lower value than both other cell types). For imputing the first modality, we also see that JAMIE outperforms the baseline method KNN for 1,453 versus 510 features ($p <$ 1e-100) in terms of AUROC and 1,129 versus 871 features ($p <$ 9e-9) in terms of correlation (Fig. 3f). JAMIE exhibits similar outperformance for AUROC when imputing the second modality (Fig. S1).

4.2 Gene Expression and Electrophysiology of Neuronal Cells in the Mouse Visual and Motor Cortices

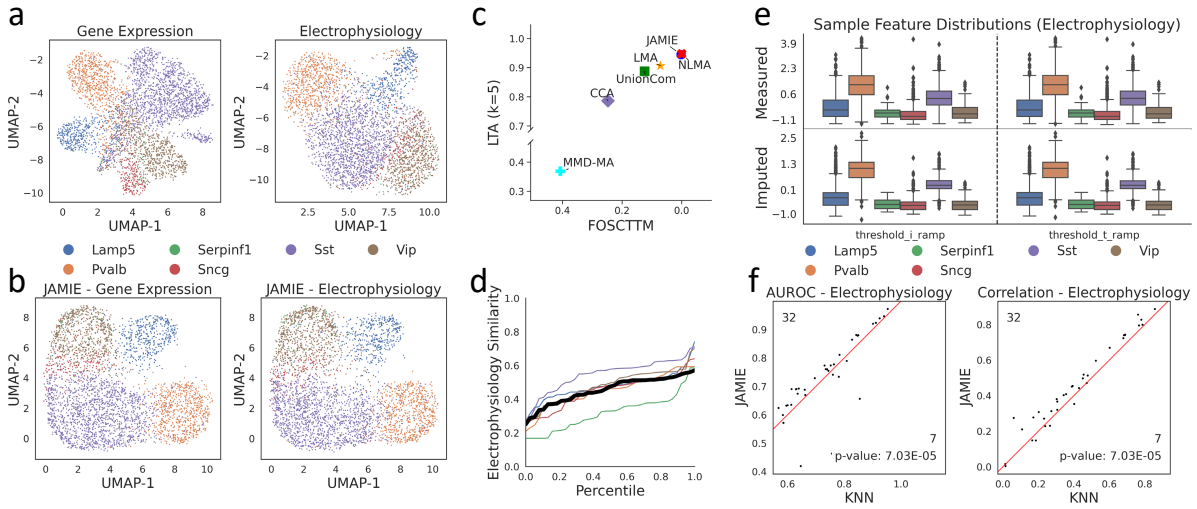

Fig. 4. Gene expression and ephys in the mouse visual cortex. (a) UMAP of single cells in by gene expression data (left) and ephys data (right), colored by cell types (inhibitory neuronal types). (b) JAMIE latent spaces of (a). (c) Fraction of samples closer to the true mean (x-axis) and label transfer accuracy (y-axis) of JAMIE and state-of-the-art methods for cell-type separation. (d) Cumulative distributions of similarity (1 − JS distance) between measured and imputed ephys features by gene expression. (e) Measured (top) and imputed (bottom) values of two ephys features across cell types. (f) Performance for imputing ephys features of JAMIE versus a baseline KNN by AUROC and correlation Section 3.

We applied JAMIE to infer cross-modal embeddings and impute gene expression and ephys data (inhibitory neuronal types: Lamp5, Serpinf1, Sst, Vip, Pvalb, Sncg) neuronal cells in the mouse visual cortex ([8], Fig. 4a).

We found that both gene and ephys embeddings identified by JAMIE can effectively separate those cell types (Fig. 4b). Further, JAMIE (LTA= 0.944, FOSCTTM= 0.002) outperforms other alignment methods, e.g., LMA (LTA= 0.907, FOSCTTM= 0.072) and UnionCom (LTA= 0.887, FOSCTTM= 0.124) (Fig. 4c).

Also, JAMIE performs imputation consistently across cell types and generally maintains ephys changes across cell types with an average JS distance of 0.546 ± 0.121 (Fig. 4d). Two ephys features with high similarity demonstrate this preservation with average JS distances 0.122 and 0.133 (Fig. 4e). Finally, ephys features are imputed significantly better than the baseline KNN, with 32 of 39 features ($p <$ 8e-5) performing better on JAMIE for both AUROC and correlation (Fig. 4f).

In addition to the mouse visual cortex, we also imputed gene expression and ephys features in the mouse motor cortex. JAMIE maintains cell type separation in the latent space (Fig. S4b). JAMIE (LTA= 0.899, FOSCTTM= 0.002) outperforms several methods, e.g., LMA (LTA= 0.897, FOSCTTM= 0.044) (Fig. S4c). Unioncom (LTA= 0.248, FOSCTTM= 0.445), CCA (LTA= 0.273, FOSCTTM= 0.366), and MMD-MA (LTA= 0.246, FOSCTTM= 0.478) are unable to align the datasets.

JAMIE also preserves ephys changes across cell types, with average JS distance 0.551 ± 0.172 (Fig. S4d). Two ephys features with high similarity demonstrate the preserving nature of JAMIE with JS distances 0.369 and 0.341 (Fig. S4e). Finally, JAMIE significantly outperforms the baseline KNN with 20 out of 28 features for AUROC ($p <$ 4e-2) and 24 of 29 features for correlation ($p <$ 6e-4) (Fig. S4f). Also, JAMIE similarly outperforms the imputation baseline for AUROC when imputing gene expression ($p <$ 6e-4) (Fig. S5c).

4.3 Gene Expression and Chromatin Accessibility in the Human Developing Brain

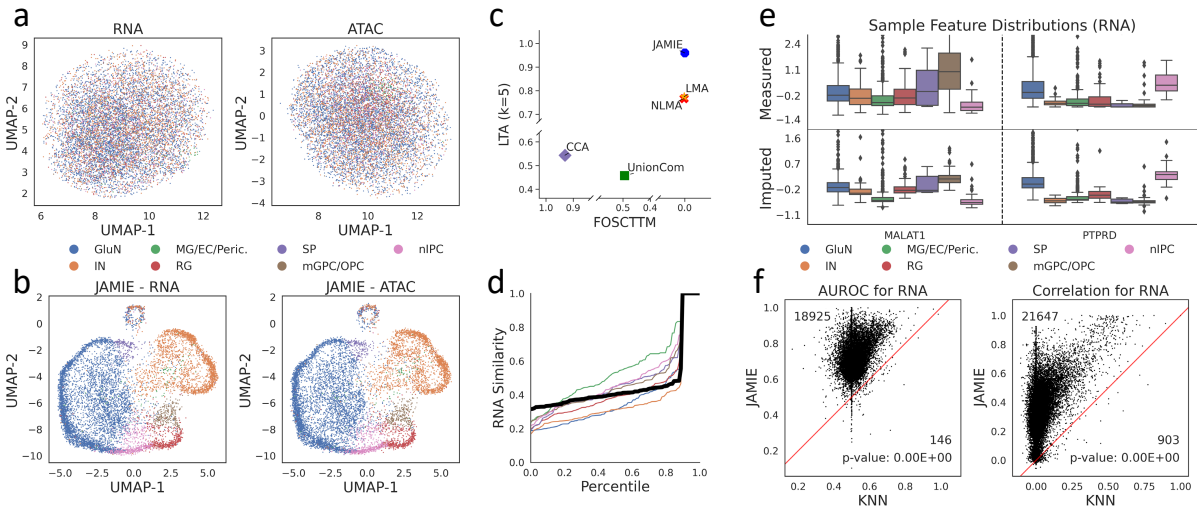

Fig. 5. Gene expression and chromatin accessibility of single cells in the developing brain. (a) UMAP of single cells in the human developing brain (21 postconceptional weeks) by gene expression (left) and the accessibility activities of open chromatin regions (right), colored by cell type. (b) JAMIE latent space of a. (c) Fraction of samples closer to the true mean (x-axis) and label transfer accuracy (y-axis) of JAMIE and state-of-the-art methods for cell type separation. (d) Cumulative distributions of similarity (1 − JS distance) between measured and imputed gene expression by chromatin accessibility. (e) Measured (top) and imputed (bottom) values of two genes across cell types. (f) Performance for imputing genes of JAMIE versus a baseline KNN by AUROC and correlation Section 3.

To further investigate the emerging single-cell multiomics data on gene regulation, we apply JAMIE to the gene expression and chromatin accessibility data of the developing human cerebral cortex (scRNA-seq and scATAC-seq by 10x Multiome) [18] (Fig. 5a). The chromatin accessibility data measure the accessibility of open chromatin regions (OCRs) by peak signals. OCRs play key epigenomic roles in regulating gene expression.

JAMIE embeddings separate cell types, in contrast to UMAP alone (Fig. 5a-b). In comparison of cell type separations, JAMIE also outperforms all other methods (LTA= 0.959, FOSCTTM= 0.000) compared with NLMA (LTA= 0.767, FOSCTTM= 0.002) and LMA (LTA= 0.775, FOSCTTM= 0..002) (Fig. 5c). CCA (LTA= 0.544, FOSCTTM= 0.930) and UnionCom (LTA= 0.458, FOSCTTM= 0.494) failed to align due to the complexity of the data.

Moreover, we used JAMIE to impute gene expression from OCRs (peaks) and vice versa. Imputed gene expression values are preserved across cell types with average JS distance 0.538 and standard deviation 0.224 (Fig. 5d). We observe wider distributions in the measured data when compared with the imputed data, likely leading to relatively high JS distance with a large number of cell types. We highlight two genes with specifically high similarity, having JS distances 0.304 and 0.266 (Fig. 5e). Finally, JAMIE significantly outperforms the imputation baseline, better imputing 18,925 of 19,071 genes for AUROC ($p <$ 1e-100) and

21,647 of 22,550 genes for correlation ($p <$ 1e-100) from OCRs (Fig. 5f). JAMIE exhibits similar imputation performance when imputing ATAC features from RNA (Fig. S3. Imputation of OCRs from gene expression also outperforms the KNN baseline ($p <$ 1e-100) (Fig. S3).

4.4 Gene Expression and Chromatin Accessibility on Colon Adenocarcinoma Cell Line

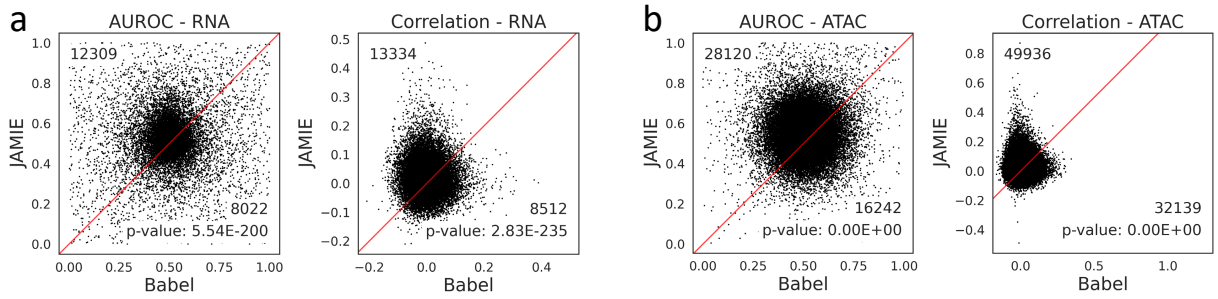

Fig. 6. JAMIE and BABEL compared by imputation performance for (a) ATAC to RNA and (b) RNA to ATAC imputation. AUROC and correlation by feature are shown for each comparison.

In addition, we compared JAMIE to BABEL [19] using scRNA-seq and scATAC-seq data in colon adenocarcinoma COLO-320DM cells in Fig. 6. JAMIE significantly outperforms BABEL in imputing both modalities. In particular, JAMIE was better on 12,309 ($p <$ 1e-100) and 13,334 ($p <$ 1e-100) genes by AUROC and correlation, respectively, for gene expression imputation from chromatin accessibility (scATAC-seq, OCR peaks) (Fig. 6a). For imputing OCRs from gene expression, JAMIE was better on 28,120 ($p <$ 1e-100) and 49,936 ($p <$ 1e-100) features by AUROC and correlation, respectively (Fig. 6b).

4.5 Biological Interpretability in JAMIE for Cross-Modal Imputation

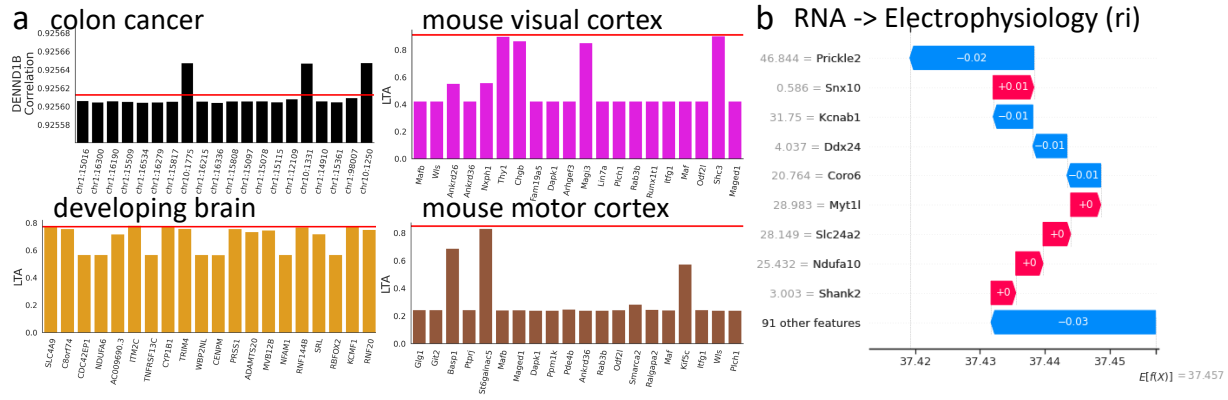

Fig. 7. Feature prioritization for cross-modal imputation and embedding. (a) Top left - prioritized open chromatin regions for imputing DENND1B gene in colon cancer data. Top right - LTA for JAMIE with select genes removed in the mouse visual cortex. Bottom left - LTA for JAMIE with select genes removed in the developing brain. Bottom right - LTA for JAMIE with select genes removed in the mouse motor cortex. For each, 14 of the lowest values are displayed along with 6 randomly sampled genes for background. Red Line: Baseline value with no chromatin regions or genes removed. (b) Waterfall plot of select important genes for imputing ephys feature *ri* in mouse visual cortex patch-seq dataset. All SHAP importance values scaled by 100 (x-axis). Red: Positive SHAP value, Blue: Negative SHAP value

To avoid the black-box nature of many deep learning models, we applied SHAP to prioritize features for cross-modal imputation (Section 3.4). Specifically, this analysis gives top features in one modality for imputing a given feature in another modality.

As shown in Fig. 7a, JAMIE prioritized open chromatin regions on Chromosome 1 for imputing the gene, DENND1B, which is a GI cancer related gene on the same chromosome [17]. Further, removing OCRs closer to the location of DENND1B generally results in worse imputation performance (e.g., removing chr1:163,361,107-163,362,488 produces lower correlation than chr1:78,047,142-78,048,370), suggesting that the possible capability of JAMIE to reveal the importance of genomic proximity from chromatin accessibility to gene expression.

For the developing brain, JAMIE identifies several important genes, including CDC42EP1, NFAM1, and RBFOX2 (all LTA= 0.888) for contributing JAMIE embeddings to separate the cell-types.

Additionally, JAMIE identifies the genes Ppm1k, Ankrd36, and Wls (LTA= 0.423 in mouse visual cortex, LTA= 0.239 in mouse motor cortex) as commonly important features in both brain regions. The commonality could indicate their importance in the generation of the *joint latent embedding* space, suggesting that these genes are important in regard to prediction of ephys features that refine brain cell type identification. For instance, many cell-type marker genes were found within the top 200 prioritized genes, with all possible found within the top 400. In particular, Prickle2 was also shown to be particularly predictive of ephys features as shown by their SHAP importance scores (Fig. 7b).

5 Discussion

JAMIE is a novel deep neural network model for cross-modal estimation. It works for complex, mixed, or partial correspondence multimodal data facilitated by a novel latent embedding aggregation methodology reliant on a dual variational autoencoder structure. In addition to its outperformances as above, JAMIE is also computationally efficient requiring less memory than other methods (Table 1). Moreover, the pretrained model along with learnt cross-modal latent embeddings may be reused for downstream analyses.

We evaluated JAMIE using the evaluation metrics that were designed for the alignment methods such as FOSCTTM and LTA. Namely, despite JAMIE not being designed to provide common latent spaces in each modality like alignment, JAMIE's FOSCTTM remains comparable to, and in many applications better than, existing methods in this paper. In addition, LTA appears to be better under complex use cases featuring unfiltered and noisy data. This suggests that the use of modality-specific and aggregated latent spaces in JAMIE allows for flexibility in forming the latent embeddings for better separate cell types from noisy samples. It is noted that JAMIE exhibits certain separation properties similar to that of UnionCom. Specifically, the silhouette coefficients of each cell type are generally similar in distribution. JAMIE slightly trades off these silhouette coefficients in favor of better LTA and FOSCTTM scores.

Training variational autoencoders is time consuming for larger datasets. Thus, prior feature selection such as the automated PCA in JAMIE helps alleviate time requirements. Data preprocessing is also crucial to avoid large or repeated features that can disproportionately shape the features of the low-dimensional embedding, due to the use of reconstruction loss. For cross-modal imputation specifically, the diversity of the training dataset has to be carefully considered to avoid biasing the final model, and negatively affecting the generalizability. JAMIE could also potentially be extended to align datasets from different sources rather than different modalities, such as gene expression measured under different conditions. Because the novel contributions of JAMIE are mainly in the processing of the latent space, many changes in structure and datatype are straightforward to achieve.

6 Acknowledgments

This work was supported by National Institutes of Health grants, R01MH128695-01A1, R21NS128761, R21NS127432, R01AG067025, R21CA237955, R01HD106197, U01MH116492, R03NS123969 and National Science Foundation Career Award 2144475.

7 Author Contributions

D.W. conceived and supervised the study. N.C.K. developed and implemented the methodology. X.H. and D.W. verified the methods. N.C.K. performed visualization and analysis. N.C.K, X.H., and D.W. edited and wrote the manuscript. All authors read and approved the final manuscript.

8 Competing interests

The authors declare no competing interests.

9 Data and Code Availability

All code was implemented in Python using PyTorch as the deep learning package and the source code is publicly available at https://github.com/daifengwanglab/JAMIE.

10 Financial Disclosure Statement

This work was supported by National Institutes of Health grants, R01MH128695-01A1, R21NS128761, R21NS127432, R01AG067025, R21CA237955, R01HD106197, U01MH116492, R03NS123969 to D.W., P50HD105353 to Waisman Center, National Science Foundation Career Award 2144475 to D.W., and the start-up funding for D.W. from the Office of the Vice Chancellor for Research and Graduate Education at the University of Wisconsin–Madison. The funders had no role in study design, data collection and analysis, decision to publish, or manuscript preparation.

11 Supplementary Materials

Supplementary Figures S1, S2, S3, S4, and S5 can be found in the accompanying supplement pdf.

References

[1] Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew M. Dai, Rafal Józefowicz, and Samy Bengio. "Generating Sentences from a Continuous Space". In: *CoRR* abs/1511.06349 (2015). arXiv: 1511.06349. URL: http://arxiv.org/abs/1511.06349.

[2] Cathryn R. Cadwell, Athanasia Palasantza, Xiaolong Jiang, Philipp Berens, Qiaolin Deng, Marlene Yilmaz, Jacob Reimer, Shan Shen, Matthias Bethge, Kimberley F. Tolias, Rickard Sandberg, and Andreas S. Tolias. "Electrophysiological, transcriptomic and morphologic profiling of single neurons using Patch-seq". In: *Nature Biotechnology* 34.2 (Feb. 2016), pp. 199–203. ISSN: 1546-1696. DOI: 10.1038/nbt.3445. URL: https://doi.org/10.1038/nbt.3445.

[3] Kai Cao, Xiangqi Bai, Yiguang Hong, and Lin Wan. "Unsupervised topological alignment for single-cell multi-omics integration". In: *Bioinformatics* 36.Supplement_1 (July 2020), pp. i48–i56. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btaa443. eprint: https://academic.oup.com/bioinformatics/article-pdf/36/Supplement_1/i48/33558199/btaa443.pdf. URL: https://doi.org/10.1093/bioinformatics/btaa443.

[4] Zhi-Jie Cao and Ge Gao. "Multi-omics single-cell data integration and regulatory inference with graph-linked embedding". In: *Nature Biotechnology* 40.10 (Oct. 2022), pp. 1458–1466. ISSN: 1546-1696. DOI: 10.1038/s41587-022-01284-4. URL: https://doi.org/10.1038/s41587-022-01284-4.

[5] Zhi-Jie Cao, Xin-Ming Tu, and Chen-Rui Xia. "Single-cell multi-omics data integration with GLUE/CLUE". In: (2022). URL: https://github.com/gao-lab/GLUE.

[6] Z. Cui, H. Chang, S. Shan, and X. Chen. "Generalized unsupervised manifold alignment". In: *Advances in Neural Information Processing Systems* 3 (Jan. 2014), pp. 2429–2437.

[7] Rohan Gala, Nathan Gouwens, Zizhen Yao, Agata Budzillo, Osnat Penn, Bosiljka Tasic, Gabe Murphy, Hongkui Zeng, and Uygar Sümbül. "A coupled autoencoder approach for multi-modal analysis of cell types". In: (2019). DOI: 10.48550/ARXIV.1911.05663. URL: https://arxiv.org/abs/1911.05663.

[8] Nathan W Gouwens, Staci A Sorensen, Fahimeh Baftizadeh, Agata Budzillo, Brian R Lee, Tim Jarsky, Lauren Alfiler, Katherine Baker, Eliza Barkan, Kyla Berry, Darren Bertagnolli, Kris Bickley, Jasmine Bomben, Thomas Braun, Krissy Brouner, Tamara Casper, Kirsten Crichton, Tanya L Daigle, Rachel Dalley, Rebecca A de Frates, Nick Dee, Tsega Desta, Samuel Dingman Lee, Nadezhda Dotson, Tom Egdorf, Lauren Ellingwood, Rachel Enstrom, Luke Esposito, Colin Farrell, David Feng, Olivia Fong, Rohan Gala, Clare Gamlin, Amanda Gary, Alexandra Glandon, Jeff Goldy, Melissa Gorham, Lucas Graybuck, Hong Gu, Kristen Hadley, Michael J Hawrylycz, Alex M Henry, Dijon Hill, Madie Hupp, Sara Kebede, Tae Kyung Kim, Lisa Kim, Matthew Kroll, Changkyu Lee, Katherine E Link, Matthew Mallory, Rusty Mann, Michelle Maxwell, Medea McGraw, Delissa McMillen, Alice Mukora, Lindsay Ng, Lydia Ng, Kiet Ngo, Philip R Nicovich, Aaron Oldre, Daniel Park, Hanchuan Peng, Osnat Penn, Thanh Pham, Alice Pom, Zoran Popović, Lydia Potekhina, Ramkumar Rajanbabu, Shea Ransford, David Reid, Christine Rimorin, Miranda Robertson, Kara Ronellenfitch, Augustin Ruiz, David Sandman, Kimberly Smith, Josef Sulc, Susan M Sunkin, Aaron Szafer, Michael Tieu, Amy Torkelson, Jessica Trinh, Herman Tung, Wayne Wakeman, Katelyn Ward, Grace Williams, Zhi Zhou, Jonathan T Ting, Anton Arkhipov, Uygar Sümbül, Ed S Lein, Christof Koch, Zizhen Yao, Bosiljka Tasic, Jim Berg, Gabe J Murphy, and Hongkui Zeng. "Integrated Morphoelectric and Transcriptomic Classification of Cortical GABAergic Cells". en. In: *Cell* 183.4 (Nov. 2020), 935–953.e19.

[9] Nelson Johansen and Gerald Quon. "scAlign: a tool for alignment, integration, and rare cell identification from scRNA-seq data". In: *Genome Biology* 20.1 (Aug. 2019), p. 166. ISSN: 1474-760X. DOI: 10.1186/s13059-019-1766-4. URL: https://doi.org/10.1186/s13059-019-1766-4.

[10] Diederik P Kingma and Max Welling. *Auto-Encoding Variational Bayes*. 2013. DOI: 10.48550/ARXIV.1312.6114. URL: https://arxiv.org/abs/1312.6114.

[11] Jie Liu, Yuanhao Huang, Ritambhara Singh, Jean-Philippe Vert, and William Stafford Noble. "Jointly Embedding Multiple Single-Cell Omics Measurements". en. In: *Algorithms Bioinform* 143 (Sept. 2019).

[12] Scott Lundberg and Su-In Lee. *A Unified Approach to Interpreting Model Predictions*. 2017. DOI: 10.48550/ARXIV.1705.07874. URL: https://arxiv.org/abs/1705.07874.

[13] Nam D Nguyen, Ian K Blaby, and Daifeng Wang. "ManiNetCluster: a novel manifold learning approach to reveal the functional links between gene networks". In: *BMC Genomics* 20.12 (Dec. 2019), p. 1003.

[14] Nam D. Nguyen, Jiawei Huang, and Daifeng Wang. "A deep manifold-regularized learning model for improving phenotype prediction from multi-modal data". In: *Nature Computational Science* 2.1 (Jan.

2022), pp. 38–46. ISSN: 2662-8457. DOI: 10.1038/s43588-021-00185-x. URL: https://doi.org/10.1038/s43588-021-00185-x.

[15] L A Quinn, G E Moore, R T Morgan, and L K Woods. "Cell lines from human colon carcinoma with unusual cell products, double minutes, and homogeneously staining regions". en. In: *Cancer Res* 39.12 (Dec. 1979), pp. 4914–4924.

[16] Federico Scala, Dmitry Kobak, Matteo Bernabucci, Yves Bernaerts, Cathryn René Cadwell, Jesus Ramon Castro, Leonard Hartmanis, Xiaolong Jiang, Sophie Laturnus, Elanine Miranda, Shalaka Mulherkar, Zheng Huan Tan, Zizhen Yao, Hongkui Zeng, Rickard Sandberg, Philipp Berens, and Andreas S. Tolias. "Phenotypic variation of transcriptomic cell types in mouse motor cortex". In: *Nature* 598.7879 (Oct. 2021), pp. 144–150. ISSN: 1476-4687. DOI: 10.1038/s41586-020-2907-3. URL: https://doi.org/10.1038/s41586-020-2907-3.

[17] Jianing Shi, Changming Cheng, Jun Ma, Choong-Chin Liew, and Xiaoping Geng. "Gene expression signature for detection of gastric cancer in peripheral blood". en. In: *Oncol Lett* 15.6 (Mar. 2018), pp. 9802–9810.

[18] Alexandro E. Trevino, Fabian Müller, Jimena Andersen, Laksshman Sundaram, Arwa Kathiria, Anna Shcherbina, Kyle Farh, Howard Y. Chang, Anca M. Pașca, Anshul Kundaje, Sergiu P. Pașca, and William J. Greenleaf. "Chromatin and gene-regulatory dynamics of the developing human cerebral cortex at single-cell resolution". In: *Cell* 184.19 (Sept. 2021), 5053–5069.e23. ISSN: 0092-8674. DOI: 10.1016/j.cell.2021.07.039. URL: https://doi.org/10.1016/j.cell.2021.07.039.

[19] Kevin E. Wu, Kathryn E. Yost, Howard Y. Chang, and James Zou. "BABEL enables cross-modality translation between multiomic profiles at single-cell resolution". In: *Proceedings of the National Academy of Sciences* 118.15 (2021), e2023070118. DOI: 10.1073/pnas.2023070118. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.2023070118. URL: https://www.pnas.org/doi/abs/10.1073/pnas.2023070118.

[20] Ran Zhang, Laetitia Meng-Papaxanthos, Jean-Philippe Vert, and William Stafford Noble. "Semi-supervised Single-Cell Cross-modality Translation Using Polarbear". In: *Research in Computational Molecular Biology*. Ed. by Itsik Pe'er. Cham: Springer International Publishing, 2022, pp. 20–35. ISBN: 978-3-031-04749-7.