MIXTURE REPRESENTATION LEARNING WITH COUPLED AUTOENCODING AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Jointly identifying a mixture of discrete and continuous factors of variability can help unravel complex phenomena. We study this problem by proposing an unsupervised framework called coupled mixture VAE (cpl-mixVAE), which utilizes multiple interacting autoencoding agents. The individual agents operate on augmented copies of training samples to learn mixture representations, while being encouraged to reach consensus on the categorical assignments. We provide theoretical justification to motivate the use of a multi-agent framework, and formulate it as a variational inference problem. We benchmark our approach on MNIST and dSprites, achieving state-of-the-art categorical assignments while preserving interpretability of the continuous factors. We then demonstrate the utility of this approach in jointly identifying cell types and type-specific, activity-regulated genes for a single-cell gene expression dataset profiling over 100 cortical neuron types.

1 INTRODUCTION

Complex phenomena can be attributed to a mixture of discrete and continuous factors of variability. Such complexity is crucial to resolve in a variety of different contexts, from learning models for image datasets to identifying factors underlying neuronal identity. A common approach to study such phenomena is clustering, which can benefit from utilizing representations that jointly capture the dependence on discrete and continuous factors. Generative models can learn such representations, and this has recently received attention from the deep learning community. Deep Gaussian mixture models are among the first deep generative models to jointly represent discrete and continuous factors, in which a continuous representation is decomposed into discrete clusters (Johnson et al., 2016; Dilokthanakul et al., 2016; Jiang et al., 2017). Such models have mainly focused on clustering without regard to interpretability. Adversarial and variational methods have been proposed to learn mixture representations that can identify interpretable continuous factors. While adversarial learning, e.g. InfoGAN (Chen et al., 2016) is susceptible to stability issues (Kim & Mnih, 2018; Dupont, 2018; Jeong & Song, 2019), variational approaches, e.g. JointVAE and CascadeVAE have produced promising and more stable results (Dupont, 2018; Jeong & Song, 2019). However, such variational methods utilizing a single autoencoding agent rely either on a heuristic data-dependent embedding capacity, or on solving a separate optimization problem for the discrete variable. Thus, learning interpretable and stable mixture representations remains challenging.

To address these issues, we introduce a multi-agent variational framework to jointly infer discrete and continuous factors through collective decision making. Coupling of autoencoding agents has been previously studied in the context of multi-modal recordings, where each agent learns a continuous latent representation for one of the data modalities (Feng et al., 2014; Gala et al., 2019). Here, we propose pairwise-coupled autoencoders to learn mixture representations for a single data modality in an unsupervised fashion. Each agent receives an augmented copy of the given sample with the same cluster label. To achieve this, we design a novel type-preserving augmentation that generates noisy copies of the data using within-cluster variabilities, while preserving its cluster identity. Coupling across the agents is achieved by encouraging discrete variables to be invariant under type-preserving augmentation, which regularizes the agents to learn interpretable representations. We demonstrate that such a coupled multi-agent architecture can increase inference accuracy and robustness by exploiting within-cluster variabilities, without requiring a prior distribution on the relative cluster abundances.

Our contributions can be summarized as follows: (i) We first provide a theoretical justification to motivate the advantage of collective decision making for learning more accurate discrete factors,

which utilizes noisy copies of the same sample. To achieve such samples, we propose an unsupervised type-preserving augmentation method. (ii) We formulate collective decision making as a variational inference problem. In this formulation, we introduce an approximation of Aitchison distance in the simplex to avoid mode collapse in the categorical variables. (iii) We benchmark our method and display its superiority in categorical assignment using the MNIST and dSprites datasets. (iv) Finally, we apply the method to a single cell gene expression dataset. We demonstrate that our method can be used to investigate the role of genes in the context of the discrete neuronal cell types and continuous regulatory activities, known as cell state.

2 PRELIMINARIES

For an observation $\mathbf{x} \in \mathbb{R}^D$, a variational autoencoder (VAE) learns a generative model $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$ and a variational distribution $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$, where $\mathbf{z} \in \mathbb{R}^M$ for $M \ll D$ is a latent variable with a parameterized distribution $p(\mathbf{z})$ (Kingma & Welling, 2013). *Disentangling* different sources of variability into different dimensions of $\mathbf{z}$ enables an interpretable selection of latent factors (Higgins et al., 2017; Locatello et al., 2018). However, in many real-world applications the inherent mixture distribution of continuous and discrete variations is often overlooked. This problem can be addressed within the VAE framework in an unsupervised fashion by introducing a categorical latent variable $\mathbf{c} \in \mathcal{S}^K$, denoting the class label defined in a K-simplex, alongside the continuous latent variable $\mathbf{s} \in \mathbb{R}^M$. Here, we refer to the continuous variable $\mathbf{s}$ as the *state* or *style* variable interchangeably. Assuming $\mathbf{s}$ and $\mathbf{c}$ are independent random variables (RVs), the evidence lower bound (ELBO) (Blei et al., 2017) for a single autoencoding agent with the distributions parameterized by $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ is given by,

$$\mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\theta}) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{s},\mathbf{c}|\mathbf{x})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{s},\mathbf{c})\right] - D_{KL}\left(q_{\boldsymbol{\phi}}(\mathbf{s}|\mathbf{x}) \| p(\mathbf{s})\right) - D_{KL}\left(q_{\boldsymbol{\phi}}(\mathbf{c}|\mathbf{x}) \| p(\mathbf{c})\right). \quad (1)$$

Maximizing ELBO in Eq. 1 to jointly learn $q(\mathbf{s}|\mathbf{x})$ and $q(\mathbf{c}|\mathbf{x})$ is challenging due to the mode collapse problem, where the network ignores a subset of latent variables. Akin to β-VAE (Higgins et al., 2017; Burgess et al., 2018), JointVAE assigns controlled information capacities to both continuous and categorical factors to prevent mode collapse (Dupont, 2018). A drawback of this method is that the capacities are dataset-dependent, and need to be tuned empirically over training iterations. As an alternative, CascadeVAE (Jeong & Song, 2019) maximizes the ELBO by iterating over two separate optimizations for the continuous and categorical variables after a warm-up period, instead of a fully gradient-based optimization. Although the computational cost for the suggested optimization for the categorical variable has an approximately linear dependence on the number of categories and batch size, it can still be a deterrent for problems with numerous categories and unbalanced datasets requiring larger batch sizes. Thus, single-agent VAEs fall short of efficiently learning interpretable mixture representations, either due to their reliance on a heuristic embedding capacity, or lacking a fully variational approach.

In addition to the issues discussed above, the performance and interpretability of the such VAE approaches are further limited by a common assumption, which is that the continuous variable representing the style of the data is independent of the class label. In practice, style often depends on class label. For instance, even for the well-studied MNIST dataset, the histograms of common digit styles, e.g. "width" markedly vary for different digits (Figure 1.a). Moreover, in these approaches, unsupervised clustering of only the continuous variable achieves a relatively high classification accuracy ($\sim 66\%$, see supplementary material, Sections F and G), underscoring that the independence assumption is not valid.

3 COUPLED MIXTURE VAE FRAMEWORK

The key intuition behind multi-agent networks is cooperation for decision making to improve the probabilistic estimation. Utilizing collective decision is also known as "wisdom of the crowd" (Surowiecki, 2005). Assuming that decisions made by a crowd (multiple agents) form a probability distribution, multiple estimates from the same sample averages to a collective decision, which is closer to the ground truth than the individual decisions. This is theoretically justified by Proposition 1 in the context of categorical assignment.

Definition 1. (Confidence) *Suppose the sample* $\mathbf{x}$ *is generated by a multivariate mixture distribution, so that* $p(\mathbf{x}) = \sum_{\phi} p(\phi)p(\mathbf{x}|\phi)$*, where* $p(\phi)$ *denotes an arbitrary prior for categories abundance. Utilizing a Bayes framework, the confidence of assignment for category* k *can be expressed as:*

$$\mathcal{C}(k) = \mathbb{E}_{p(\mathbf{x})}\left[\log p(\phi = k|\mathbf{x})\right]. \quad (2)$$

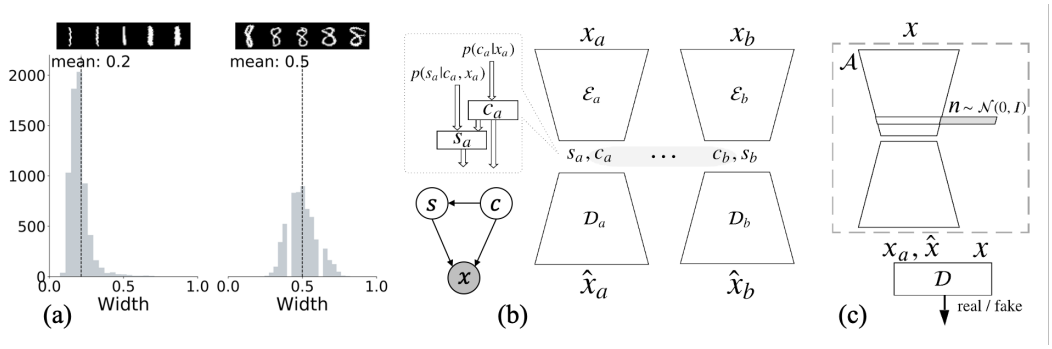

Figure 1: (a) Type-dependent distribution for the continuous state variable, for digits 1 and 8 in MNIST. Subfigures show empirical distributions of the digit's width which depends on the digit type. (b) The proposed architecture for the type-preserving augmentation; the augmenter is composed of an autoencoder that generates an augmented sample, $\mathbf{x}_a$ by projecting down the the sample $\mathbf{x}$ to a lower dimension and concatenating it with a Gaussian noise $\mathbf{n}$; the discriminator that regularizes the generated sample discriminate between $\mathbf{x}_a$ and $\mathbf{x}$. (c) Schematic of the cpl-mixVAE model with the graphical model of a single mixture VAE network. Each agent learns a mixture representation. Coupling of the networks is achieved by imposing a penalty on mismatches in the categorical assignments $\mathbf{c}_a$ and $\mathbf{c}_b$ for augmented copies $\mathbf{x}_a$ and $\mathbf{x}_b$ of the sample $\mathbf{x}$.

Proposition 1. *Consider the problem of mixture representation learning using a multi-agent framework. In the presence of sufficient number of agents, confidence of the categorical assignment is higher than the.* (Proof in supplementary material, Section A)

Proposition 1 reveals that for a dataset with an arbitrary prior discrete distribution, collective decision made by enough number of agents enhances the categorical assignment. To estimate the minimum number of agents, we need to have access to the prior and cluster-dependent distributions. While in an unsupervised framework class-dependent information is unavailable, in case of uniform prior distribution over categories, we have the following:

Corollary 1. *In case of uniformly distributed categories, one pair of agents is sufficient to increase confidence in the categorical assignment.* (Proof in supplementary material, Section A)

According to Corollary 1, for datasets with uniformly distributed classes, having two agents is sufficient to improve the categorical assignment.

3.1 Multi-agent VAE

Employing the insight obtained from Proposition 1, we formulate collective decision making for an A-agent VAE network as the following constrained optimization,

$$\begin{aligned} \max \quad & \mathcal{L}_{\mathbf{s}_1|\mathbf{c}_1}(\boldsymbol{\phi}_1, \boldsymbol{\theta}_1) + \cdots + \mathcal{L}_{\mathbf{s}_A|\mathbf{c}_A}(\boldsymbol{\phi}_A, \boldsymbol{\theta}_A) \\ & \text{s.t. } \mathbf{c}_1 = \cdots = \mathbf{c}_A \end{aligned} \tag{3}$$

where $\mathcal{L}_{\mathbf{s}_a|\mathbf{c}_a}(\boldsymbol{\phi}_a, \boldsymbol{\theta}_a)$ is the variational loss function for agent a that receives an augmented sample $\mathbf{x}_a$ to approximate two discriminative models, $q(\mathbf{c}_a|\mathbf{x}_a)$ and $q(\mathbf{s}_a|\mathbf{x}_a, \mathbf{c}_a)$, representing the type and the state variables respectively (see supplementary material for derivation, Section B).

$$\begin{aligned} \mathcal{L}_{\mathbf{s}|\mathbf{c}}(\boldsymbol{\phi}, \boldsymbol{\theta}) \quad = \quad & \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{s},\mathbf{c}|\mathbf{x})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{s},\mathbf{c})\right] - \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{c}|\mathbf{x})}\left[D_{KL}\left(q_{\boldsymbol{\phi}}(\mathbf{s}|\mathbf{c},\mathbf{x}) \| p(\mathbf{s}|\mathbf{c})\right)\right] \\ & - \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{s}|\mathbf{c},\mathbf{x})}\left[D_{KL}\left(q_{\boldsymbol{\phi}}(\mathbf{c}|\mathbf{x}) \| p(\mathbf{c})\right)\right]. \end{aligned} \tag{4}$$

Not only is it challenging to solve the maximization in Eq. 3 due to the equality constraint, but the objective remains a function of the prior $p(\mathbf{c})$ which is unknown, and typically non-uniform. To overcome this, we introduce an equivalent formulation for Eq. 3 based on the pairwise coupling paradigm as follows (see supplementary material Section C for derivation).

$$\begin{aligned} \max \quad & \sum_{a=1}^{A} (A-1)\left(\mathbb{E}_{q(\mathbf{s}_a,\mathbf{c}_a|\mathbf{x}_a)}\left[\log p(\mathbf{x}_a|\mathbf{s}_a,\mathbf{c}_a)\right] - \mathbb{E}_{q(\mathbf{c}_a|\mathbf{x}_a)}\left[D_{KL}\left(q(\mathbf{s}_a|\mathbf{c}_a,\mathbf{x}_a) \| p(\mathbf{s}_a|\mathbf{c}_a)\right)\right]\right) \\ & - \sum_{a<b} \mathbb{E}_{q(\mathbf{s}_a|\mathbf{c}_a,\mathbf{x}_a)}\mathbb{E}_{q(\mathbf{s}_b|\mathbf{c}_b,\mathbf{x}_b)}\left[D_{KL}\left(q(\mathbf{c}_a|\mathbf{x}_a)q(\mathbf{c}_b|\mathbf{x}_b) \| p(\mathbf{c}_a,\mathbf{c}_b)\right)\right] \\ & \text{s.t. } \mathbf{c}_a = \mathbf{c}_b \;\; \forall a,b \in [1,A],\; a<b \end{aligned} \tag{5}$$

Here, in the last term, the KL divergence across coupled agents is a function of the joint distribution $p(\mathbf{c}_a, \mathbf{c}_b)$, rather than $p(\mathbf{c})$. We relax Eq. 5 into an unconstrained problem by assuming a differentiable form for $p(\mathbf{c}_a, \mathbf{c}_b)$ (full derivation in supplementary material, Section D).

$$\max \quad \sum_{a=1}^{A} (A-1) \left(\mathbb{E}_{q(\mathbf{s}_a, \mathbf{c}_a|\mathbf{x}_a)} \left[\log p(\mathbf{x}_a|\mathbf{s}_a, \mathbf{c}_a) \right] - \mathbb{E}_{q(\mathbf{c}_a|\mathbf{x}_a)} \left[D_{KL} \left(q(\mathbf{s}_a|\mathbf{c}_a, \mathbf{x}_a) \| p(\mathbf{s}_a|\mathbf{c}_a) \right) \right] \right) + \sum_{a<b} H(\mathbf{c}_a|\mathbf{x}_a) + H(\mathbf{c}_b|\mathbf{x}_b) - \lambda \mathbb{E}_{q(\mathbf{c}_a|\mathbf{x}_a)} \mathbb{E}_{q(\mathbf{c}_b|\mathbf{x}_b)} \left[d^2(\mathbf{c}_a, \mathbf{c}_b) \right] \tag{6}$$

In Eq. 6, we refer to λ as the *coupling hyperparameter*, which controls pairwise coupling between agents. Note that this formulation can be easily extended to include an additional hyperparameter to encourage disentanglement of continuous variables as in β-VAE Higgins et al. (2017). Distance between categorical variables $\mathbf{c}_a$ and $\mathbf{c}_b$ is denoted by $d(\mathbf{c}_a, \mathbf{c}_b)$ and is defined as $d(\mathbf{c}_a, \mathbf{c}_b) = \|clr(\mathbf{c}_a) - clr(\mathbf{c}_b)\|_2, \forall \mathbf{c}_a, \mathbf{c}_b \in \mathcal{S}^K$. Here, $clr(.)$ denotes the isometric centered-log-ratio transformation and d satisfies the conditions of a mathematical metric according to Aitchison geometry (Aitchison, 1982; Egozcue et al., 2003). To sample from $q(\mathbf{c}_a|\mathbf{x}_a)$ in a gradient descent framework, we use the Gumbel-softmax distribution (Jang et al., 2016; Maddison et al., 2014) with a temperature parameter $0 < \tau \leq 1$.

To benefit from collective decision making, the VAE agents share their categorical latent variables with each other, (Figure 1.b) and try to achieve identical assignments while independently learning their own style variables. In the rest of this manuscript, we refer to the model in Eq. 6 as *cpl-mixVAE*.

3.2 Type-preserving augmentation

To enable such collective decision making in an unsupervised fashion, we need a type-preserving augmentation to generate a noisy copies of a given sample for each agent. Type-preserving augmentation can be considered as a generative model that not only learns the data distribution to regenerate the original data, but it also learns a transformation on a per-class distribution representing the within-class variations. Achieving such augmentations is not always straightforward and requires prior knowledge about the underlying factors of invariance. For image datasets conventional data augmentation, such as rotation, scaling, or translation can provide type-preserving augmentations. However, these affine operations face two principal challenges; they cannot applied to non-image datasets; and they often fail to capture the richness of the underlying process. There are alternative methods that address some shortcomings of conventional data augmentation (Hauberg et al., 2016; Jaiswal et al., 2018; Antoniou et al., 2017). However, they are not general enough. They either learn a set of transformations that are specific for images or they need class label information.

Figure 1.c displays the proposed design for the unsupervised type-preserving augmentation using an autoencoder model that is imposed by a Gaussian noise (network $\mathcal{A}$). Since autoencoders are often fail to generate sharp samples, we include a discriminator to improve the quality of the augmented samples (network $\mathcal{D}$). For each training iteration, network $\mathcal{A}$ generates two samples: $\mathbf{x_n}$ and $\mathbf{x_{\not n}}$. The former denotes an augmented sample, i.e. $\mathbf{x_a}$, and the later presents a generated sample in the absence of noise, i.e. $\hat{\mathbf{x}}$. To train the whole settings, we are using the following minmax optimization,

$$\min_{\mathcal{A}} \max_{\mathcal{D}} \mathcal{V}(\mathcal{D}, \mathcal{A}) - \mathcal{R}(\mathcal{A}) + \mathcal{T}_\alpha(\mathcal{A}) + \gamma d(\mathcal{A}) \tag{7}$$

where, $\mathcal{V}(\mathcal{D}, \mathcal{A}) = \mathbb{E}_{\mathbf{x}} [\log \mathcal{D}(\mathbf{x})] + \mathbb{E}_{\mathbf{x}} [\log (1 - \mathcal{D}(\mathbf{x_{\not n}}))] + \mathbb{E}_{\mathbf{x},\mathbf{n}} [\log (1 - \mathcal{D}(\mathbf{x_n}))]$ is the value function for the joint training of the discriminator and augmenter; $\mathcal{R}(\mathcal{A}) = \mathbb{E}_{q(\mathbf{z}|\mathbf{x})} [\log p(\mathbf{x}|\mathbf{z})]$ is the reconstruction loss; $\mathcal{T}_\alpha(\mathcal{A}) = \max(\|\mathbf{x} - \mathbf{x_{\not n}}\|_2 - \|\mathbf{x} - \mathbf{x}_n\|_2 + \alpha, 0)$ is the triplet loss that prevents network $\mathcal{A}$ to generate identical samples by neglecting noise; and $d(\mathcal{A}) = D_{KL}(q(\mathbf{z}|\mathbf{x}) \| q(\mathbf{z}|\mathbf{x}, \mathbf{n}))$ is the distance loss between the latent variables generated in the absence and presence of noise. The distance measure, indeed, encourages type-preserving augmentation by assuring both noisy and original samples are close to each other in the latent space. This measure is controlled by hyperparameter $\gamma \ll 1$, to reduce the risk of generating identical samples.

3.3 Stabilizing the training by mini-batch variance

The solution to the maximization problem in Eq. 6, which includes minimization of $d(\mathbf{c}_a, \mathbf{c}_b)$, has trivial local optima that result in the mode collapse issue (Lucas et al., 2019), in which the network learns to ignore a subset of the discrete latent space. For instance, in the extreme case of $\mathbf{c}_{a_n} = \mathbf{c}_{b_n} = \mathbf{c}, \ \forall \ \mathbf{x}_{a_n}, \mathbf{x}_{b_n} \in \mathbf{X}$, where n denotes the sample index, the distance is trivially minimized. In this scenario, the state variable is compelled to act as the latent variable of a classical

variational autoencoder, while the model fails to deliver an interpretable representation despite achieving an overall low loss value. To avoid this collapse issue, we regularize the Aitchison distance $d(\mathbf{c}_{a_n}, \mathbf{c}_{b_n})$ between the categorical assignments of the n-th samples of agents a and b by using mini-batch statistics, which is a common technique to stabilize the training process: $d^2_\sigma(\mathbf{c}_{a_n}, \mathbf{c}_{b_n}) = \sum_k \left(\sigma^{-1}_{a_k} \log c_{a_{n_k}} - \sigma^{-1}_{b_k} \log c_{b_{n_k}}\right)^2$, where $\sigma^2_{a_k}$ indicates the variance of the k-th category of agent a. Here, we show that d^2_σ is an approximation of the Aitchison distance in the probability simplex.

Proposition 2. *Suppose $\mathbf{c}_{a_n}, \mathbf{c}_{b_n} \in \mathcal{S}^K$, where $\mathcal{S}^K$ is a simplex of K parts and n is the sample index. If $d\left(\mathbf{c}_{a_n}, \mathbf{c}_{b_n}\right)$ denotes the Aitchison distance, then*

$$d^2_\sigma\left(\mathbf{c}_{a_n}, \mathbf{c}_{b_n}\right) - d^2\left(\mathbf{c}_{a_n}, \mathbf{c}_{b_n}\right) \leq \frac{1}{K}\left(\left(\tau_{\mathbf{c}} + \tau_\sigma\right)^2 + K^2\tau^2_\sigma - \Delta^2_\sigma\right)$$

where $\tau_{\mathbf{c}} = \max_k\{\log c_{a_{n_k}} - \log c_{b_{n_k}}\}$, $\tau_\sigma = \max_k\{(\sigma^{-1}_{a_k} - 1)\log c_{a_{n_k}} - (\sigma^{-1}_{b_k} - 1)\log c_{b_{n_k}}\}$, and $\Delta_\sigma = \sum_k (\sigma^{-1}_{a_k} - 1)\log c_{a_{n_k}} - (\sigma^{-1}_{b_k} - 1)\log c_{b_{n_k}}$. (Proof in supplementary material, Section E)

Accordingly, as the Gumbel-softmax approximations of the categorical variable of the agents move closer to each other on the simplex, d_σ converges to d.

4 Experiments

We assess the performance of cpl-mixVAE for three different datasets. To facilitate comparisons with other methods, first we conducted experiments on two benchmark datasets: MNIST and dSprites. Additionally, we used a single cell RNA-sequencing dataset (FACS) (Tasic et al., 2018), to evaluate the utility of our approach in identifying cell types and cell states. The network architecture and training details are reported in the supplementary material, Section H.

MNIST: To learn interpretable representations for the MNIST dataset, we used a 10-dimensional categorical variable representing digits (type), and a 10-dimensional continuous random variable representing the state variable, which can be perceived as the writing style. During training, each agent receives an augmented copy of the original image consisting of random rotation, brightness and saturation changes. To interpret the role of the continuous factor, we fix the discrete latent variable and change the state variable according to the conditional state distribution learned for each category. Figure 2 illustrates these continuous latent traversal results for four dimensions of the state variable obtained from a bi-agent cpl-mixVAE with shared categorical variables. Each row corresponds to a different dimension of the categorical variable, and the state variable monotonically changes across the columns. Panels (a), (b), and (c) represent commonly-identified continuous factors with global attributes, while panel (d) represents roundness, all in a digit-dependent manner.

Table 1 displays the classification performance of the discrete latent variable (as the predicted class label) for different methods including JointVAE and CascadeVAE, where we report the loss value based on the reconstruction objective, $\log p(\mathbf{x}|\mathbf{s}, \mathbf{c})$, and the accuracy of the categorical assignments (ACC) across 10 initializations with different random seeds. For CascadeVAE, we used the numbers reported in (Jeong & Song, 2019) and for JointVAE, we used the publicly available implementation and hyperparameters detailed in (Dupont, 2018). We report the performance of the proposed coupled VAE for two cases: (i) cpl-mixVAE($\mathbf{s} \not| \mathbf{c}$), in which the state variable is independent of the discrete variable, and (ii) cpl-mixVAE($\mathbf{s} \mid \mathbf{c}$), which is the proposed VAE model in Eq. 6. The reported numbers for cpl-mixVAE models belong to only one of the agents (Further accuracy/stability improvements could be achieved by considering the results of both agents, but that is besides the focus of the present work). For a fair comparison, in all experiments, the selected Gumble-softmax temperature, τ and latent dimensionality are identical to those of JointVAE and CascadeVAE. To understand whether architecture differences put JointVAE at a disadvantage, we report the results for JointVAE†, which uses the same architecture for the basic encoder/decoder networks as the one used in cpl-mixVAE. That is, JointVAE† uses the same learning procedure as JointVAE, but its convolutional layers are replaced by fully-connected layers (see supplementary material for implementation details). Comparison of the results obtained with JointVAE and JointVAE† suggests that the superiority of cpl-mixVAE is not due to the network architecture. Additionally, to detach the impact of augmentation in the training, we report the results for JointVAE‡, in which the JointVAE model has been trained with an augmented copy of the original MNIST dataset. The reported

	ACC ↑ (mean ± s.d.)	
Method	MNIST	dSprites
InfoGAN	77.87 ± 21.68	-
CascadeVAE	81.41 ± 09.54	78.84 ± 15.65
JointVAE	68.99 ± 11.76	44.79 ± 03.88
JointVAE†	62.19 ± 05.73	-
JointVAE‡	68.21 ± 09.58	-
cpl-mixVAE($\mathbf{s} \not| \mathbf{c}$)	xx.xx ± xx.xx	-
cpl-mixVAE^{a}($\mathbf{s} \mid \mathbf{c}$)	82.92 ± 04.64	96.30 ± 09.15
cpl-mixVAE($\mathbf{s} \mid \mathbf{c}$)	**84.56 ± 06.47**	xxxx ± xxxx

Table 1: Clustering results for the MNIST and dSprites datasets. Reported values denote the average accuracy of categorical assignment, over 10 runs. For MNIST, all models use $|\mathbf{c}| = 10$, $|\mathbf{s}| = 10$, and for dSprites, the reported ones use $|\mathbf{c}| = 3$, $|\mathbf{s}| = 6$. For CascadeVAE, all JointVAEs and cpl-mixVAEs, we have $\tau = 0.67$.

clustering performance for JointVAE‡ denotes that including data augmentation by itself does not enhance the categorical assignment. Finally, we note that cpl-mixVAE marginally outperforms CascadeVAE, despite the latter using an inner discrete optimization loop whose time complexity scales with the number of categories, hence being significantly slower.

In many real-life applications, the true number of types is unknown. Accordingly, in a separate set of experiments, we studied the performance of cpl-mixVAE($\mathbf{s}|\mathbf{c}$) for different cardinalities of the categorical variable, $\mathbf{c}$. Figure 3.a shows the performance of the proposed cpl-mixVAE in terms of ACC and AMP as a function of $|\mathbf{c}| = K \in [7, 13]$. Here, AMP denotes the average of maximum probability of categories i.e., $1/K \sum_{k=1}^{K} \max p(\mathbf{c}_k|\mathbf{x})$. Our experimental results show that the best performance is obtained for $K = 10$. Figure 3.b demonstrates the sorted probabilistic assignment of $\mathbf{c}$ for all digits. As illustrated, an inadequate number of categorical variables (left panel) results in some dimensions being used for more than one digit, while extra $\mathbf{c}_k$ leaves some categories under-utilized (right panel).

dSprites: Similar to JointVAE and CascadeVAE, we used a 3-dimensional categorical variable for learning the shape (type), and a 6-dimensional state variable representing the style of each shape. During training, each agent received an augmented copy of the original image consisting of random rotation, scaling, and translation. To interpret the role of the continuous factor, we fix the discrete latent variable and change the state variable, as in the MNIST experiment. Figure 4 illustrates these traversal results for three dimensions of the state variable obtained from a 2-agent cpl-mixVAE. Each row corresponds to a different dimension of the categorical variable, and the state variable monotonically changes across the columns. In addition, Table 1 shows the degree to which cpl-mixVAE outperforms the other two methods in clustering.

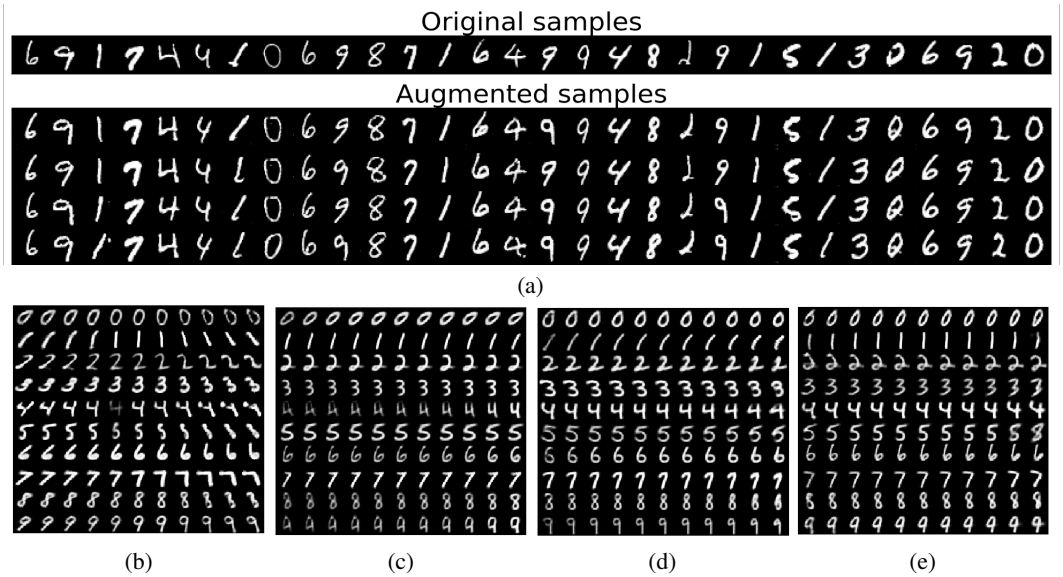

Figure 2: (a) Augmented samples generated by the proposed type-preserving augmentation. (b-e) Continuous latent traversals of cpl-mixVAE with 10-dimensional continuous and 10-dimensional categorical variables. Examples of (b) rotation angle, (c) stroke thickness, (d) character width and (e) roundness of looped features are presented. The discrete variable $\mathbf{c}$ is constant for all reconstructions in the same row.

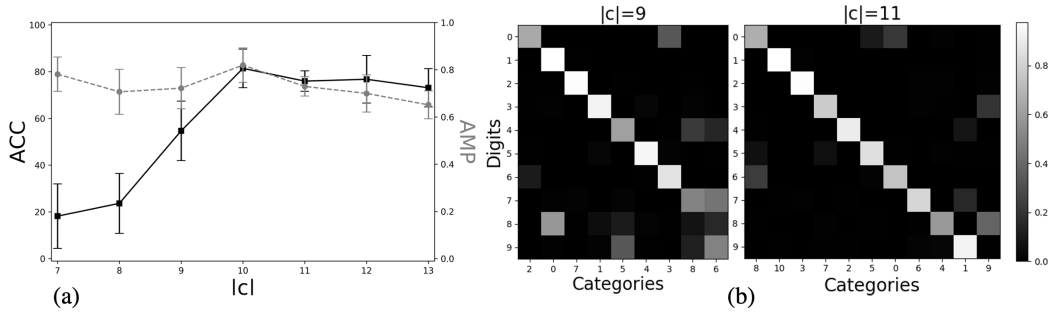

Figure 3: Clustering performance of cpl-mixVAE for MNIST, when the number of clusters ($|\mathbf{c}|$) is unknown.

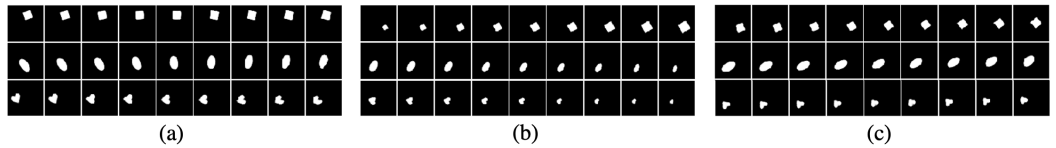

Figure 4: Continuous latent traversals of the trained cpl-mixVAE model with 6-dimensional continuous variable and 3-dimensional categorical variable. Examples of (a) rotation, (b) scale, (c) position features are presented. The discrete variable $\mathbf{c}$ is held as fixed for all reconstructions in the same row.

FACS: Compared to typical machine learning datasets, the FACS dataset is very high-dimensional, with over 10,000 genes. It includes over 100 cell types with sizeable difference between the most and the least abundant clusters. This dataset profiles gene expression of 22,365 cortical neurons for adult mouse (Tasic et al., 2018). Here, we excluded non-neuronal cells and used a subset of 5,000 most expressed genes based on their peak values. The original publication partitioned the neurons into 115 clusters referred to as cell types using a biologically-motivated, hierarchical clustering of marker gene expression (Tasic et al., 2018). From the perspective of neuroscience, it is not yet clear whether these genetic types correspond to meaningful circuit components with matching physiological and/or anatomical characteristics, or if the detailed categorization is a case of over-splitting. Moreover, individual cells can be different either because they belong to different types or different cell states (Trapnell, 2015; Andrews & Hemberg, 2018), the latter perhaps better explained by continuous factors of variability around cell type representatives. One final challenging aspect of categorical assignment in this dataset is its substantial imbalance, where for the most- and the least-abundant types, there exist 1404 and 16 samples, respectively.

We used a 115-dimensional discrete and a 2-dimensional continuous variable for cell type and cell state representations, respectively. We applied gene-wise dropout augmentation to train each VAE agent (Srivastava et al., 2014). Figure 5.a illustrates the performance of the 2-agent cpl-mixVAE model trained over 45,000 iterations. The dendrograms on the y-axis displays the hierarchical relationship between neuron types according to (Tasic et al., 2018). For inhibitory neuronal cells,

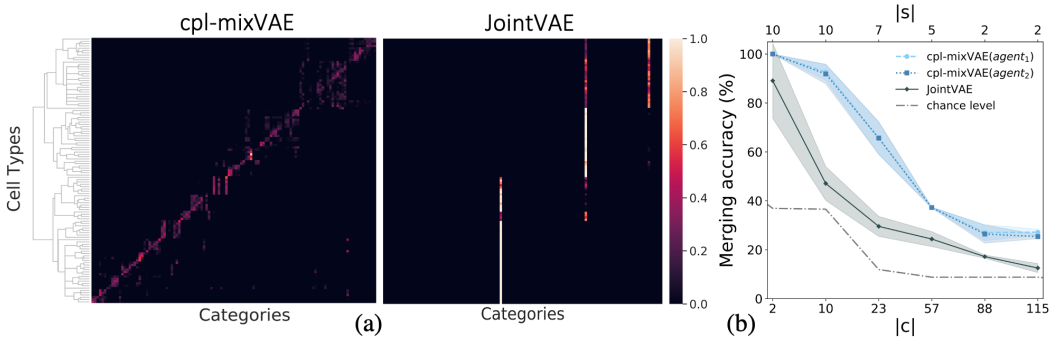

Figure 5: Categorical assignments for the FACS dataset. (a) Confusion matrices of cpl-mixVAE and JointVAE using $|\mathbf{c}| = 115$, $|\mathbf{s}| = 2$, $\tau = 1$, and $\lambda = 1$. The dendrogram on the y-axis shows marker-based hierarchical classification with 115 cell types. (b) Classification accuracy, over 3 runs with 2M training iterations, as a function of number of categories by progressively merging the 115 types according to the hierarchical dendrogram.

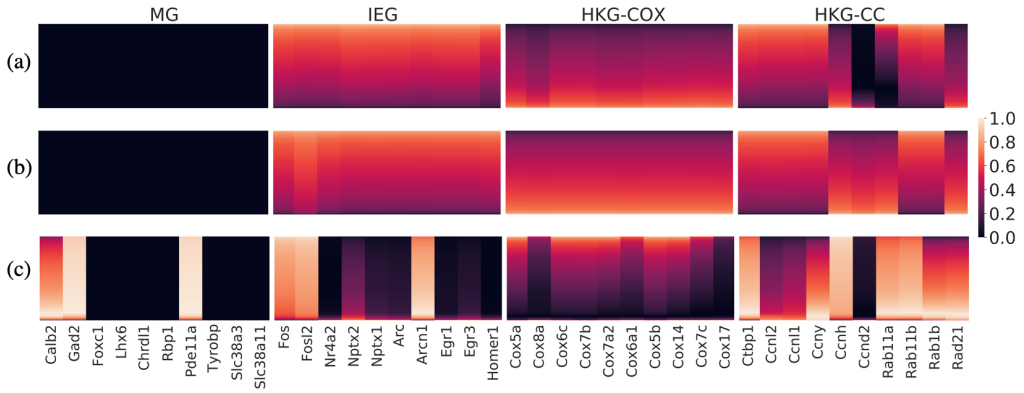

Figure 6: Continuous latent traversals of cpl-mixVAE for two excitatory cells (a) "L5 NP ALM Trhr Nefl" and (b) "L6 CT Nxph2 Sla", and one inhibitory cell (c) "Pvalb Akr1c18 Ntf3". For each cell, the continuous latent traversal is color-mapped to a normalized reconstructed gene expression (colorbar values) as a function of the state variable for four gene subsets from left to right: marker genes (MG), immediate early genes (IEG), and two subgroups of house keeping genes, cytochrome c oxidase (HKG-COX) and cell cycle regulators (HKG-CC).

whether the observed diversity corresponds to discrete variability or a continuum is an ongoing debate (Ascoli et al., 2008; Harris et al., 2018). Consistently, while the classification accuracy of our model is lower for inhibitory neurons and 32% across the whole dataset (chance level is $\sim 6\%$, based on the most abundant type), the accuracy significantly increases as clusters are merged according to the hierarchy defined by (Tasic et al., 2018) (see supplementary material). Figure 5.b compares the clustering performance of JointVAE and cpl-mixVAE for different numbers of categories obtained for merged types. For instance, $|\mathbf{c}| = 2$ corresponds to the highest node in the dendrogram, where there are only two neuronal types, excitatory and inhibitory. The top x-axis denotes the state variable dimension, which is gradually increased by merging neuronal types that leads to an increment in within-cluster variations.

To examine the role of the continuous latent variable (the cell state), we applied a similar traversal analysis to that used for the MNIST and dSprites datasets. For a given cell sample and its discrete type, we changed every dimension of the continuous state variable using conditional distribution $q(\mathbf{s}|\mathbf{c})$, and inspected gene expression changes caused by state variable alterations. Figure 6 shows the results of the state traversal experiment for two excitatory neurons belonging to the "L5 NP" (near-projecting) and "L6 CT" (corticothalamic) sub-classes, and an inhibitory neuron belonging to the "PV" (parvalbumin) class. In each sub-figure, the latent traversal is color-mapped to normalized reconstructed expression values, where the y-axis corresponds to one dimension of the state variable, and the x-axis corresponds to four gene subsets, namely (i) marker genes (MG) for the two excitatory types identified by (Tasic et al., 2018), (ii) immediate early genes (IEG), and two house keeping gene (HKG) subgroups (iii) cytochrome c oxidase (COX), and (iv) cell cycle (CC) regulators (Hrvatin et al., 2018; Tarasenko et al., 2017). MGs are ideally expected to function as indicators for particular cell types such that their normalized expression is not affected by the state of the cell. Indeed, the expression of the reported excitatory MGs remains constant for excitatory traversals but not necessarily for the inhibitory traversal (i.e., Calb2, Gad2, Pde11a in Fig. 6). In contrast, the expression of IEGs and HKGs should depend strongly on the metabolic and environmental conditions. Indeed, we find that the expression changes of IEGs and HKGs are for the most part monotonically linked to the state variable, confirming that the state variable captures relevant, interpretable continuous variability, as in the MNIST and dSprites examples. Lastly, the expression of IEGs and HKGs (state-dependent genes) depends on the cell type e.g., not all IEGs are activated for all cell types. Notably, for the excitatory "L5 NP" and "L6 CT" cells that are proximate in the hierarchy (suggested in (Tasic et al., 2018)), state traversal is quite similar. These results suggest that the state variable inferred by cpl-mixVAE provides insight when deciphering the molecular mechanisms shaping the landscape of cell states.

5 Conclusion

We have introduced cpl-mixVAE as a general framework to apply the power of collective decision making in unsupervised joint learning of discrete and continuous generative factors, without imposing

a prior distribution on the discrete variable. The proposed framework utilizes multiple pairwise-coupled autoencoding agents with a shared categorical variable, while independently learning the continuous variables. We derived a variational formulation for this joint learning machine that conditions the continuous factor on the discrete latent variable, and showed that the resulting joint representations are more accurate, more stable, and yet interpretable. Our experimental results for all three datasets show that cpl-mixVAE outperforms the comparable VAE models. Finally, our results for a challenging gene expression dataset showed that the proposed framework can identify annotated neuronal types and differentiate between type-dependent and state-dependent genes. (See supplementary material for a discussion on future directions.)

REFERENCES

John Aitchison. The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2):139–160, 1982.

Tallulah S Andrews and Martin Hemberg. Identifying cell populations with scrnaseq. *Molecular aspects of medicine*, 59:114–122, 2018.

Antreas Antoniou, Amos Storkey, and Harrison Edwards. Data augmentation generative adversarial networks. *arXiv preprint arXiv:1711.04340*, 2017.

Giorgio A Ascoli, Lidia Alonso-Nanclares, Stewart A Anderson, German Barrionuevo, Ruth Benavides-Piccione, Andreas Burkhalter, György Buzsáki, Bruno Cauli, Javier DeFelipe, Alfonso Fairén, et al. Petilla terminology: nomenclature of features of gabaergic interneurons of the cerebral cortex. *Nature Reviews Neuroscience*, 9(7):557, 2008.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-vae. *arXiv preprint arXiv:1804.03599*, 2018.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in neural information processing systems*, pp. 2172–2180, 2016.

Nat Dilokthanakul, Pedro AM Mediano, Marta Garnelo, Matthew CH Lee, Hugh Salimbeni, Kai Arulkumaran, and Murray Shanahan. Deep unsupervised clustering with gaussian mixture variational autoencoders. *arXiv preprint arXiv:1611.02648*, 2016.

Emilien Dupont. Learning disentangled joint continuous and discrete representations. In *Advances in Neural Information Processing Systems*, pp. 710–720, 2018.

Juan José Egozcue, Vera Pawlowsky-Glahn, Glòria Mateu-Figueras, and Carles Barcelo-Vidal. Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, 35(3): 279–300, 2003.

Fangxiang Feng, Xiaojie Wang, and Ruifan Li. Cross-modal retrieval with correspondence autoencoder. In *Proceedings of the 22nd ACM international conference on Multimedia*, pp. 7–16, 2014.

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 *Advances in Neural Information Processing Systems*, pp. 9263–9272, 2019.

Kenneth D Harris, Hannah Hochgerner, Nathan G Skene, Lorenza Magno, Linda Katona, Carolina Bengtsson Gonzales, Peter Somogyi, Nicoletta Kessaris, Sten Linnarsson, and Jens Hjerling-Leffler. Classes and continua of hippocampal ca1 inhibitory neurons revealed by single-cell transcriptomics. *PLoS biology*, 16(6):e2006387, 2018.

Søren Hauberg, Oren Freifeld, Anders Boesen Lindbo Larsen, John Fisher, and Lars Hansen. Dreaming more data: Class-dependent distributions over diffeomorphisms for learned data augmentation. In *Artificial Intelligence and Statistics*, pp. 342–350, 2016.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. *Iclr*, 2(5):6, 2017.

Sinisa Hrvatin, Daniel R Hochbaum, M Aurel Nagy, Marcelo Cicconet, Keiramarie Robertson, Lucas Cheadle, Rapolas Zilionis, Alex Ratner, Rebeca Borges-Monroy, Allon M Klein, et al. Single-cell analysis of experience-dependent transcriptomic states in the mouse visual cortex. *Nature neuroscience*, 21(1):120–129, 2018.

Ayush Jaiswal, Rex Yue Wu, Wael Abd-Almageed, and Prem Natarajan. Unsupervised adversarial invariance. In *Advances in Neural Information Processing Systems*, pp. 5092–5102, 2018.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

Yeonwoo Jeong and Hyun Oh Song. Learning discrete and continuous factors of data via alternating disentanglement. *arXiv preprint arXiv:1905.09432*, 2019.

Zhuxi Jiang, Yin Zheng, Huachun Tan, Bangsheng Tang, and Hanning Zhou. Variational deep embedding: an unsupervised and generative approach to clustering. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pp. 1965–1972, 2017.

Matthew J Johnson, David K Duvenaud, Alex Wiltschko, Ryan P Adams, and Sandeep R Datta. Composing graphical models with neural networks for structured representations and fast inference. In *Advances in neural information processing systems*, pp. 2946–2954, 2016.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. *arXiv preprint arXiv:1802.05983*, 2018.

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

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. *arXiv preprint arXiv:1811.12359*, 2018.

James Lucas, George Tucker, Roger B Grosse, and Mohammad Norouzi. Don't blame the elbo! a linear vae perspective on posterior collapse. In *Advances in Neural Information Processing Systems*, pp. 9403–9413, 2019.

Chris J Maddison, Daniel Tarlow, and Tom Minka. A* sampling. In *Advances in Neural Information Processing Systems*, pp. 3086–3094, 2014.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

James Surowiecki. *The wisdom of crowds*. Anchor, 2005.

Tatyana N Tarasenko, Susan E Pacheco, Mary Kay Koenig, Julio Gomez-Rodriguez, Senta M Kapnick, Francisca Diaz, Patricia M Zerfas, Emanuele Barca, Jessica Sudderth, Ralph J DeBerardinis, et al. Cytochrome c oxidase activity is a metabolic checkpoint that regulates cell fate decisions during t cell activation and differentiation. *Cell metabolism*, 25(6):1254–1268, 2017.

Bosiljka Tasic, Zizhen Yao, Lucas T Graybuck, Kimberly A Smith, Thuc Nghi Nguyen, Darren Bertagnolli, Jeff Goldy, Emma Garren, Michael N Economo, Sarada Viswanathan, et al. Shared and distinct transcriptomic cell types across neocortical areas. *Nature*, 563(7729):72–78, 2018.

Cole Trapnell. Defining cell types and states with single-cell genomics. *Genome research*, 25(10): 1491–1498, 2015.