MIXTURE REPRESENTATION LEARNING WITH COUPLED AUTOENCODERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Latent representations help unravel complex phenomena. While continuous latent variables can be efficiently inferred, fitting mixed discrete-continuous models remains challenging despite recent progress, especially when the discrete factor dimensionality is large. A pressing application for such mixture representations is the analysis of single-cell omic datasets to understand neuronal diversity and its molecular underpinnings. Here, we propose an unsupervised variational framework using multiple interacting networks called *cpl-mixVAE* that significantly outperforms state-of-the-art in high-dimensional discrete settings. *cpl-mixVAE* introduces a consensus constraint on discrete factors of variability across the networks, which regularizes the mixture representations at the time of training. We justify the use of this framework with theoretical results and validate it with experiments on benchmark datasets. We demonstrate that our approach discovers interpretable discrete and continuous variables describing neuronal identity in two single-cell RNA sequencing datasets, each profiling over a hundred cortical neuron types.

1 INTRODUCTION

Fitting mixed discrete-continuous models arises in many contexts. While continuous latent variables can be efficiently inferred with variational and adversarial formulations, inference of continuous and discrete factors in generalized mixture models remains challenging despite recent progress (Jang et al., 2016; Chen et al., 2016; Dupont, 2018; Jeong & Song, 2019). A pressing domain of application for such models is quantifying factors of biological variability in single-cell omic studies. The high-throughput and high-dimensional datasets produced by these studies document a previously unappreciated diversity of gene expression. In neuroscience, this poses the identification of cell types and cell states as a key research area to understand how neuronal circuits function (Bargmann et al., 2014), where the notions of cell types and states can be considered as biological interpretations of discrete and continuous variability. While marker gene based studies suggest the existence of more than 100 neuronal cell types in just a single brain region, there is no agreement on such categorization and interpretation of the remaining continuous variability (Seung & Sümbül, 2014; Zeng & Sanes, 2017; Tasic et al., 2018). Moreover, existing unsupervised, joint continuous-discrete learning methods are tailored for problems with relatively few and equally-abundant discrete components, and accurate inference remains out of reach for these applications.

Deep generative models have previously been applied to single-cell datasets, where the focus is on the cluster identity and it is typically inferred by post-hoc analysis of a continuous factor (Lopez et al., 2018). Deep Gaussian mixture models (Dilokthanakul et al., 2016; Johnson et al., 2016; Jiang et al., 2017) also focus on the identification of categories and do not take interpretability of the remaining continuous variability into account. To address the need for joint inference of interpretable discrete and continuous factors, various adversarial and variational methods have been proposed. While existing adversarial generative models, e.g. InfoGAN (Chen et al., 2016), are susceptible to stability issues (Higgins et al., 2017; Kim & Mnih, 2018), variational autoencoders (VAEs) emerge as efficient and more stable alternatives (Tschannen et al., 2018; Zhang et al., 2018; Dupont, 2018; Jeong & Song, 2019). VAE-based approaches approximate the mixture model by assuming a family of distributions q_ϕ and select the member closest to the true model p. Popular choices in VAE implementations include (1) using KL divergence to compute discrepancy between q_ϕ and p, and (2) using a multivariate Gaussian mixture distribution with uniformly distributed discrete and isotropic Gaussian distributed continuous priors. However, such choices may lead to underestimating the

posterior variance (Minka et al., 2005; Blei et al., 2017). Solutions to resolve this issue are mainly applicable in low-dimensional spaces or for continuous factors alone (Deasy et al., 2020; Kingma et al., 2016; Ranganath et al., 2016; Quiroz et al., 2018).

Inspired by *collective decision making*, we introduce a variational framework using multiple *autoencoding arms* to jointly infer interpretable finite discrete (categorical) and continuous factors in the presence of high-dimensional discrete space. Coupled-autoencoders have been previously studied in the context of multi-modal recordings, where each arm learns only a continuous latent representation for one of the data modalities (Feng et al., 2014; Gala et al., 2019; Lee & Pavlovic, 2020). Here, we develop a novel pairwise-coupled autoencoder framework for a single data modality. The proposed framework imposes a consensus constraint on the categorical posterior at the time of training and allows dependencies between continuous and categorical factors. We define the consensus constraint based on the Aitchison geometry in the probability simplex, which avoids the mode collapse problem. We show that the coupled multi-arm architecture enhances accuracy, robustness, and interpretability of the inferred factors without requiring any priors on the relative abundances of categories. Finally, on datasets profiling different cortical regions in the mammalian brain, we show that our method can be used to discover neuronal types as discrete categories and type-specific genes regulating the continuous within-type variability, such as metabolic state or disease state.

Related work. There is an extensive body of research on clustering in mixture models (Dilokthanakul et al., 2016; Jiang et al., 2017; Tian et al., 2017; Guo et al., 2016; Locatello et al., 2018b). The idea of improving the clustering performance through seeking a consensus and *co-training* and *ensembling* across multiple observations has been explored in both unsupervised (Monti et al., 2003; Kumar & Daumé, 2011) and semi-supervised contexts (Blum & Mitchell, 1998). However, these methods do not consider the underlying continuous variabilities across observations. Moreover, unlike ensemble methods, which pool the results of different trained workers, autoencoding arms seek a consensus at the time of learning in our framework.

The proposed framework does not need any supervision since the individual arms provide a form of prior or weak supervision for each other. In this regard, our paper is related to a body of work that attempts to improve representation learning by using semi-supervised or group-based settings (Bouchacourt et al., 2017; Hosoya, 2019; Nemeth, 2020). Bouchacourt et al. (2017) demonstrated a multi-level variational autoencoder (MLVAE) as a semi-supervised VAE by revealing that observations within groups share the same type. Hosoya (2019) and Nemeth (2020) attempted to improve MLVAE by imposing a weaker condition to the grouped data. In recent studies (Shu et al., 2019; Locatello et al., 2020), a weakly supervised variational setting has been proposed for disentangled representation learning by providing pairs of observations that share at least one underlying factor. These studies rely on learning latent variables in continuous spaces, and have been applied only to image datasets with low-dimensional latent representations.

Recent advances in structured variational methods, such as imposing a prior (Ranganath et al., 2016) or spatio-temporal dependencies (Quiroz et al., 2018) on the latent distribution parameters, allow for scaling to larger dimensions. However, these solutions are not directly applicable to the discrete space, which will be addressed in our A-arm VAE framework.

2 Single mixture VAE framework

For an observation $\mathbf{x} \in \mathbb{R}^D$, a VAE learns a generative model $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$ and a variational distribution $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$, where $\mathbf{z} \in \mathbb{R}^M$ is a latent variable with a parameterized distribution $p(\mathbf{z})$ and $M \ll D$ (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., 2018a). However, the interplay between continuous and discrete variabilities present in many real-world datasets is often overlooked by existing methods. This problem can be addressed within the VAE framework in an unsupervised fashion by introducing a categorical latent variable $\mathbf{c}$ denoting the class label, alongside the continuous latent variable $\mathbf{s}$. 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, the evidence lower bound (ELBO) (Blei et al., 2017) for a single mixture VAE 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 imposes characteristics on $q(\mathbf{s}|\mathbf{x})$ and $q(\mathbf{c}|\mathbf{x})$ that can result in underestimation of posterior probabilities such as the mode collapse problem, where the network ignores a

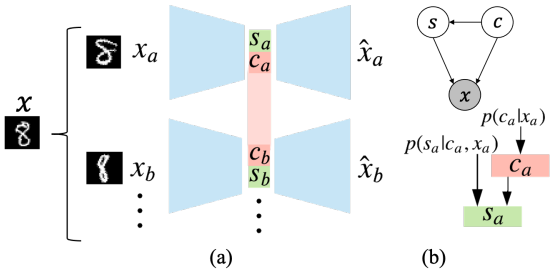


Figure 1: (a) Multi-arm autoencoder framework proposed as the cpl-mixVAE model. Individual arms receive non-identical noisy copies of given samples $\mathbf{x}$, i.e. $\{\mathbf{x}_a, \mathbf{x}_b, \dots\}$, where they all belong to the same category, to learn mixture representations, i.e. $\{q(\mathbf{c}_a, \mathbf{s}_a), q(\mathbf{c}_b, \mathbf{s}_b), \dots\}$. VAE arms cooperate to learn the categorical assignment, $p(\mathbf{c})$. (b) Graphical model of each autoencoder to learn type dependence of the state variable.

subset of latent variables (Minka et al., 2005; Blei et al., 2017). Recently, VAE-based solutions were proposed by imposing a uniform structure on $p(\mathbf{c})$: akin to β-VAE (Higgins et al., 2017; Burgess et al., 2018), JointVAE (Dupont, 2018) modifies the ELBO by assigning a pair of controlled information capacities for each variational factor, i.e. $\mathcal{C}_s \in \mathbb{R}^{|\mathbf{s}|}$ and $\mathcal{C}_c \in \mathbb{R}^{|\mathbf{c}|}$. The main drawback of JointVAE is that its performance is tied to heuristic tuning of $|\mathbf{s}| \times |\mathbf{c}|$ capacities over training iterations so that it is vulnerable to mode collapse in high-dimensional settings. Another recent VAE-based mixture model solution, CascadeVAE (Jeong & Song, 2019), maximizes the ELBO through a semi-gradient-based algorithm by iterating over two separate optimizations for the continuous and categorical variables. While the separation of the optimization steps avoids the mode collapse problem, this separation is valid only when the categorical variable is uniformly distributed. Therefore, its performance strongly depends on the clusters having similar abundances in the dataset. Thus, earlier solutions fall short of learning interpretable mixture representations with high-dimensional discrete variables in real-world applications.

In addition to the issues discussed above, the performance and interpretability of those approaches are further limited by the common assumption that the continuous variable representing the style of the data is independent of the categorical variable. In practice, style often depends on the 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 (Supplementary Section I). Moreover, further analysis of the identified continuous factor in the earlier approaches reveals that the independence assumption among $q(\mathbf{s}|\mathbf{x})$ and $q(\mathbf{c}|\mathbf{x})$ can be significantly violated (see Supplementary Sections H and I).

3 Coupled Mixture VAE Framework

The key intuition behind multi-arm networks is cooperation to improve posterior estimation. While the context is different, the popular phrase "wisdom of the crowd" (Surowiecki, 2005) can nevertheless be revealing: when a crowd (multiple arms) needs to make a decision, multiple estimates can increase the expected probability of a correct choice.

3.1 A-arm VAE Framework

We define the A-arm VAE as an A-tuple of independent and architecturally identical autoencoding arms, where the a-th arm parameterizes a mixture model distribution (Fig. 1a). In this framework, individual arms receive a collection of non-identical copies, $\{\mathbf{x}_a, \mathbf{x}_b, \dots\}$ of the given sample, $\mathbf{x}$, belonging to the same category. While each arm has its own mixture representation with potentially non-identical parameters, all arms cooperate to learn $q(\mathbf{c}_a|\mathbf{x}_a)$, where $\mathbf{c}_a = \mathbf{c}_b = \cdots$, via a cost function at the time of training. Accordingly, a crowd of VAEs with A arms can be formulated as a collection of constrained variational objectives as follows.

$$\begin{aligned} \max \quad & \mathcal{L}_{\mathbf{s}_1|\mathbf{c}_1}(\boldsymbol{\phi}_1, \boldsymbol{\theta}_1) + \cdots + \mathcal{L}_{\mathbf{s}_A|\mathbf{c}_A}(\boldsymbol{\phi}_A, \boldsymbol{\theta}_A) \\ & \text{s.t. } \mathbf{c}_1 = \cdots = \mathbf{c}_A \end{aligned} \tag{2}$$

where $\mathcal{L}_{\mathbf{s}_a|\mathbf{c}_a}(\boldsymbol{\phi}_a, \boldsymbol{\theta}_a)$ is the variational loss for arm a,

$$\begin{aligned} \mathcal{L}_{\mathbf{s}_a|\mathbf{c}_a}(\boldsymbol{\phi}_a, \boldsymbol{\theta}_a) = \mathbb{E}_{q(\mathbf{s}_a,\mathbf{c}_a|\mathbf{x}_a)}\left[\log p(\mathbf{x}_a|\mathbf{s}_a, \mathbf{c}_a)\right] - \mathbb{E}_{q(\mathbf{c}_a|\mathbf{x}_a)}\left[D_{KL}\left(q(\mathbf{s}_a|\mathbf{c}_a, \mathbf{x}_a) \| p(\mathbf{s}_a|\mathbf{c}_a)\right)\right] \\ - \mathbb{E}_{q(\mathbf{s}_a|\mathbf{c}_a,\mathbf{x}_a)}\left[D_{KL}\left(q(\mathbf{c}_a|\mathbf{x}_a) \| p(\mathbf{c}_a)\right)\right]. \end{aligned} \tag{3}$$

In Eq. 3, the variational loss for each arm is defined according to the graphical model in Fig. 1b, which is built upon the traditional ELBO in Eq. 1 by conditioning the continuous state on the categorical variable (derivation in Supplementary Section B). Therefore, learning an interpretable decomposition of the data relies on accurate assignment (inference) of the categorical latent factor. Propositions 1 and 2 below show that the shared categorical assignment inferred from $q(\mathbf{c}|\mathbf{x}_1, \cdots, \mathbf{x}_A)$, under the

$\mathbf{c} = \mathbf{c}_1 = \cdots = \mathbf{c}_A$ constraint of the multi-arm framework improves the accuracy of the categorical assignment on expectation.

Proposition 1. *Consider the problem of mixture representation learning in a multi-arm VAE framework. For independent samples from category* $\mathbf{m}$*, i.e.* $\mathbf{x}_i \sim p(\mathbf{x}|\mathbf{m})$*,*

$$\begin{aligned} \mathbb{E}_{q(\mathbf{x}|\mathbf{m})}\left[\log q(\mathbf{c}=\mathbf{m}|\{\mathbf{x}_i\}_{1:A})\right] \quad &> \quad \mathbb{E}_{q(\mathbf{x}|\mathbf{m})}\left[\log q(\mathbf{c}=\mathbf{m}|\{\mathbf{x}_i\}_{1:B})\right] \\ s.t.\ \mathbf{c}=\mathbf{c}_1=\cdots=\mathbf{c}_A \quad &\qquad\quad s.t.\ \mathbf{c}=\mathbf{c}_1=\cdots=\mathbf{c}_B \end{aligned} \tag{4}$$

if $q(\mathbf{m}|\mathbf{x}_i) < 1$ *and* $A > B \geq 1$ *denote the number of arms.* (Proof in Supplementary Section A)

Thus, having more arms increases the expected log posterior for the true categorical latent variable unless it is already at its maximum.

Proposition 2. *In the* A*-arm VAE framework, there exists an* A *that guarantees a true categorical assignment on expectation. That is,*

$$\mathbf{m} = \arg\max_{\mathbf{c}} \mathbb{E}_{q(\mathbf{x}|\mathbf{m})}\left[\log q(\mathbf{c}|\{\mathbf{x}_i\}_{1:A})\right], \ \ s.t.\ \mathbf{c}=\mathbf{c}_1=\cdots=\mathbf{c}_A\ . \tag{5}$$

(Proof in Supplementary Section A)

Accordingly, the consensus constraint is sufficient to enhance inference for mixture representations in the A-arm VAE framework. Our theoretical results show that the required number of arms satisfying Eq. 5 is a function of the categorical distribution and the likelihood (Eq. 15, Supplementary Section A). In the particular case of uniformly distributed categories, one pair of coupled arms is enough to satisfy Eq. 5 (see Corollary 1, Supplementary Section A).

We emphasize that the proposed framework does not require any weak supervision as in (Bouchacourt et al., 2017). Instead, it relies on representations that are invariant under non-identical copies of observations. Moreover, unlike (Bouchacourt et al., 2017; Shu et al., 2019; Locatello et al., 2020), the multi-arm framework is not restricted to the continuous space.

Arms observe non-identical copies of samples. In the A-arm VAE framework, arms receive non-identical observations that share the discrete variational factor. To achieve this in a fully unsupervised setting, we use *type-preserving* data augmentation that generates independent and identically distributed copies of data while preserving its categorical identity. For image datasets, conventional transformations such as rotation, scaling, or translation can serve as type-preserving augmentations. However, for non-image datasets, e.g. single-cell data, we seek a generative model that learns transformations representing within-class variability in an unsupervised manner. To this end, inspired by DAGAN (Antoniou et al., 2017) and VAE-GAN (Larsen et al., 2016), we develop a generative model to provide collections of observations for our multi-arm framework (Supplementary Section F). In Supplementary Section A, Remark 2, we further discuss an under-exploration scenario in data augmentation, in which the augmented samples are not independently distributed and are concentrated around the given sample.

3.2 CPL-MIXVAE: PAIRWISE COUPLING IN A-ARM VAE

In the A-arm VAE framework, the mixture representation is obtained through the optimization in Eq. 2. Not only is it challenging to solve the maximization in Eq. 2 due to the equality constraint, but the objective remains a function of $p(\mathbf{c})$ which is unknown, and typically non-uniform. To overcome this, we use an equivalent formulation for Eq. 2 by applying the pairwise coupling paradigm as follows (details of derivation in Supplementary Section C):

$$\begin{aligned} \max \quad & \sum_{a=1}^{A}(A-1)\Big(\mathbb{E}_{q(\mathbf{s}_a,\mathbf{c}_a|\mathbf{x}_a)}\left[\log p(\mathbf{x}_a|\mathbf{s}_a,\mathbf{c}_a)\right] - \mathbb{E}_{q(\mathbf{c}_a|\mathbf{x}_a)}\left[D_{KL}\left(q(\mathbf{s}_a|\mathbf{c}_a,\mathbf{x}_a)\|p(\mathbf{s}_a|\mathbf{c}_a)\right)\right]\Big) - \\ & \sum_{a<b}\mathbb{E}_{q(\mathbf{s}_a|\mathbf{c}_a,\mathbf{x}_a)}\mathbb{E}_{q(\mathbf{s}_b|\mathbf{c}_b,\mathbf{x}_b)}\left[D_{KL}\left(q(\mathbf{c}_a|\mathbf{x}_a)q(\mathbf{c}_b|\mathbf{x}_b)\|p(\mathbf{c}_a,\mathbf{c}_b)\right)\right] \\ & \text{s.t. } \mathbf{c}_a=\mathbf{c}_b \ \ \forall a,b\in[1,A],\ a<b \end{aligned} \tag{6}$$

We relax the optimization in Eq. 6 into an unconstrained problem by marginalizing the joint distribution over a mismatch measure between categorical variables (see Supplementary Section D):

$$\begin{aligned} \max \sum_{a=1}^{A}(A-1)\Big(\mathbb{E}_{q(\mathbf{s}_a,\mathbf{c}_a|\mathbf{x}_a)}\left[\log p(\mathbf{x}_a|\mathbf{s}_a,\mathbf{c}_a)\right] - \mathbb{E}_{q(\mathbf{c}_a|\mathbf{x}_a)}\left[D_{KL}\left(q(\mathbf{s}_a|\mathbf{c}_a,\mathbf{x}_a)\|p(\mathbf{s}_a|\mathbf{c}_a)\right)\right]\Big) + \\ \sum_{a<b} H(\mathbf{c}_a|\mathbf{x}_a) + H(\mathbf{c}_b|\mathbf{x}_b) - \lambda\mathbb{E}_{q(\mathbf{c}_a,\mathbf{c}_b|\mathbf{x}_a,\mathbf{x}_b)}\left[d^2(\mathbf{c}_a,\mathbf{c}_b)\right] \end{aligned} \tag{7}$$

In Eq. 7, in addition to entropy-based confidence penalties known as mode collapse regularizers (Pereyra et al., 2017), the distance measure $d(\mathbf{c}_a, \mathbf{c}_b)$ encourages a consensus on the categorical assignment controlled by $\lambda \geq 0$, the coupling hyperparameter.

We refer to the model in Eq. 7 as *cpl-mixVAE* (Fig. 1a). In cpl-mixVAE, VAE arms try to achieve identical categorical assignments while independently learning their own style variables. In experiments, we set $\lambda = 1$ universally. While the bottleneck architecture already encourages interpretable continuous variables, this formulation can be easily extended to include an additional hyperparameter to promote disentanglement of continuous variables as in β-VAE (Higgins et al., 2017). Additional analyses to assess the sensitivity of the cpl-mixVAE's performance to its coupling factor can be found in the Supplementary Section G.

It may be instructive to cast Eq. 7 in an equivalent constrained optimization form.

Remark 1. *The A-arm VAE framework is a collection of constrained variational models as follows:*

$$\max \sum_{a=1}^{A} \mathbb{E}_{q(\mathbf{s}_a, \mathbf{c}_a|\mathbf{x}_a)} \left[\log p(\mathbf{x}_a|\mathbf{s}_a, \mathbf{c}_a)\right] - \mathbb{E}_{q(\mathbf{c}_a|\mathbf{x}_a)} \left[D_{KL}\left(q(\mathbf{s}_a|\mathbf{c}_a, \mathbf{x}_a) \| p(\mathbf{s}_a|\mathbf{c}_a)\right)\right] + H(\mathbf{c}_a|\mathbf{x}_a)$$
$$\textit{s.t.}\ \mathbb{E}_{q(\mathbf{c}_a|\mathbf{x}_a)} \left[d^2(\mathbf{c}_a, \mathbf{c}_b)\right] < \epsilon \tag{8}$$

where ϵ denotes the strength of the consensus constraint. Here, $\mathbf{c}_b$ indicates the assigned category by any one of the arms, $b \in \{1, \ldots, A\}$, imposing structure on the discrete variable to approximate its prior distribution.

Distance between categorical variables. $d(\mathbf{c}_a, \mathbf{c}_b)$ denotes the distance between a pair of $|\mathbf{c}|$-dimensional un-ordered categorical variables, which are associated with probability vectors with non-negative entries and sum-to-one constraint that form a K-dimensional simplex, where $K = |\mathbf{c}|$. In the real space, a typical choice to compute the distance between two vectors is using Euclidean geometry. However, this geometry is not suitable for probability vectors. Here, we utilize *Aitchison geometry* (Aitchison, 1982; Egozcue et al., 2003), which defines a vector space on the simplex. Accordingly, the distance in the simplex, i.e. $d_{S^K}(\mathbf{c}_a, \mathbf{c}_b)$ is defined as $d_{S^K}(\mathbf{c}_a, \mathbf{c}_b) = \|clr(\mathbf{c}_a) - clr(\mathbf{c}_b)\|_2, \forall \mathbf{c}_a, \mathbf{c}_b \in \mathcal{S}^K$, where $clr(\cdot)$ denotes the *isometric centered-log-ratio* transformation in the simplex. This categorical distance satisfies the conditions of a mathematical metric according to Aitchison geometry.

3.3 Seeking consensus in the simplex

An instance of the mode collapse problem (Lucas et al., 2019) manifests itself in the minimization of $d_{S^K}(\mathbf{c}_a, \mathbf{c}_b)$ (Eq. 7): its trivial local optima encourages the network to abuse the discrete latent factor by ignoring many of the available categories. In the extreme case, the representations can collapse onto a single category; $\mathbf{c}_a = \mathbf{c}_b = \mathbf{c}_0$. In this scenario, the continuous variable is compelled to act as a primary latent factor, while the model fails to deliver an interpretable mixture representation despite achieving an overall low loss value. To avoid such undesirable local equilibria while training, we add perturbations to the categorical representation of each arm. If posterior probabilities in the simplex have small dispersion, the perturbed distance calculation overstates the discrepancies. Thus, instead of minimizing $d^2_{S^K}(\mathbf{c}_a, \mathbf{c}_b)$, we minimize a perturbed distance $d^2_\sigma(\mathbf{c}_a, \mathbf{c}_b) = \sum_k \left(\sigma^{-1}_{a_k} \log c_{a_k} - \sigma^{-1}_{b_k} \log c_{b_k}\right)^2$, which corresponds to the distance between additively perturbed $\mathbf{c}_a$ and $\mathbf{c}_b$ vectors in Aitchison geometry. Here, $\sigma^2_{a_k}$ and $\sigma^2_{b_k}$ indicate the mini-batch variances of the k-th category, for arms a and b. We next show that the perturbed distance $d_\sigma(\cdot)$ is bounded by $d_{S^K}(\cdot)$ and non-negative values ρ_u, ρ_l:

Proposition 3. *Suppose $\mathbf{c}_a, \mathbf{c}_b \in \mathcal{S}^K$, where $\mathcal{S}^K$ is a simplex of $K > 0$ parts. If $d_{S^K}(\mathbf{c}_a, \mathbf{c}_b)$ denotes the distance in Aitchison geometry and $d^2_\sigma(\mathbf{c}_a, \mathbf{c}_b) = \sum_k \left(\sigma^{-1}_{a_k} \log c_{a_k} - \sigma^{-1}_{b_k} \log c_{b_k}\right)^2$ denotes a perturbed distance, then*

$$d^2_{S^K}(\mathbf{c}_a, \mathbf{c}_b) - \rho_l \leq d^2_\sigma(\mathbf{c}_a, \mathbf{c}_b) \leq d^2_{S^K}(\mathbf{c}_a, \mathbf{c}_b) + \rho_u$$

where $\rho_u, \rho_l \geq 0$, $\rho_u = K\left(\tau^2_{\sigma_u} + \tau^2_{\mathbf{c}}\right) + 2\Delta_\sigma \tau_{\mathbf{c}}$, $\rho_l = \dfrac{\Delta^2_\sigma}{K} - K\tau^2_{\sigma_l}$, $\tau_{\mathbf{c}} = \max_k\{\log c_{a_k} - \log c_{b_k}\}$, $\tau_{\sigma_u} = \max_k\{g_k\}$, $\tau_{\sigma_l} = \max_k\{-g_k\}$, $\Delta_\sigma = \sum_k g_k$, and $g_k = (\sigma^{-1}_{a_k} - 1)\log c_{a_k} - (\sigma^{-1}_{b_k} - 1)\log c_{b_k}$. (Proof in Supplementary Section E)

Thus, when $\mathbf{c}_a$ and $\mathbf{c}_b$ are similar and their spread is not small, $d_\sigma(\mathbf{c}_a, \mathbf{c}_b)$ closely approximates $d_{S^K}(\mathbf{c}_a, \mathbf{c}_b)$. Otherwise, it diverges from $d_{S^K}(\cdot)$ to avoid mode collapse.

4 EXPERIMENTS

We used four datasets: dSprites, MNIST, and two single-cell RNA sequencing (scRNA-seq) datasets; Smart-seq ALM-VISp (Tasic et al., 2018) and 10X MOp (Yao et al., 2021). Although dSprites and MNIST datasets do not require high-dimensional settings for mixture representation, to facilitate comparisons of cpl-mixVAE with earlier methods, first we report the results for these benchmark datasets. We trained three unsupervised VAE-based methods for mixture modeling: JointVAE (Dupont, 2018), CascadeVAE (Jeong & Song, 2019), and ours (cpl-mixVAE). For MNIST, we additionally trained the popular InfoGAN (Chen et al., 2016) as the most comparable GAN-based model. To show the interpretability of the mixture representations, (i) for the discrete latent factor, we report the accuracy (ACC) of categorical assignments and the $D_{KL}(q(\mathbf{c})\|p(\mathbf{c})))$, (ii) for the continuous variable, we perform latent traversal analysis by fixing the discrete factor and changing the continuous variable according to $p(\mathbf{s}|\mathbf{c}, \mathbf{x})$. We calculated the accuracy by using minimum weight matching (Kuhn, 1955) to match the categorical variables obtained by cpl-mixVAE with the available cluster labels for each dataset. Additionally, we report the computational efficiency (number of iterations per second) to compare the training complexity of the multi-arm framework against earlier methods (Table 1). All reported numbers for cpl-mixVAE models are average accuracies calculated across arms. In VAE-based models, to sample from $q(\mathbf{c}_a|\mathbf{x}_a)$, we use the Gumbel-softmax distribution (Jang et al., 2016; Maddison et al., 2014). In cpl-mixVAE, each arm received an augmented copy of the original input generated by the deep generative augmenter (Supplementary Section F) during training. Details of the network architectures and training settings can be found in Supplementary Section L.

4.1 BENCHMARK DATASETS

dSprites. dSprites is procedurally generated from discrete (3 shapes) and continuous (6 style factors: scale, rotation, and position) latent factors. Based on the uniform distribution of classes, we used a 2-arm cpl-mixVAE with $|\mathbf{c}| = 3$ and $|\mathbf{s}| = 6$ to learn interpretable representations. Results in Table 1 show that our method outperforms the other methods in terms of categorical assignment accuracy. In addition to demonstrating the traversal results (Fig. 2, bottom row), we report disentanglement scores (DS in Table 1). Even though the continuous factors do not depend on the discrete factors in this synthetic dataset, we did not change the architecture and expected the network to infer this independence. For a fair comparison, we used the same disentanglement metric implemented for CascadeVAE (Jeong & Song, 2019).

MNIST. Similarly, due to the uniform distribution of digit labels in MNIST, we again used a 2-arm cpl-mixVAE model. Following the convention (Dupont, 2018; Jeong & Song, 2019; Bouchacourt

Table 1: Training results for all datasets. cpl-mixVAE uses 2 arms. $|\mathbf{c}|$ and $|\mathbf{s}|$ denote the cardinality of latent discrete and continuous spaces. Iter, ACC and DS denote number of iterations, the accuracy of the categorical assignment, and the disentanglement score, respectively. Computation denotes the training speed (iteration/second) on a GeForce RTX 2080 Ti GPU. The computation of cpl-mixVAE includes the entire execution time for training one pair of coupled networks, plus data augmentation.

Dataset	**Iter**	$\|\mathbf{c}\|$	$\|\mathbf{s}\|$	**Method**	**ACC** (%) $\uparrow$ (mean $\pm$ s.d.)	**Computation** $\uparrow$ (iteration/sec)		
dSprites	300K	3	6	JointVAE	44.79 ± 03.9	52.6	74.5 ± 5.2	DS
				CascadeVAE	78.84 ± 15.7	15.4	90.5 ± 5.3	
				cpl-mixVAE	$\mathbf{96.30} \pm \mathbf{09.2}$	20.6	89.9 ± 4.1	
MNIST	120K	10	2	InfoGAN	77.87 ± 21.7	12.2	2.15 ± 4.2	$D_{KL}(q(\mathbf{c})\|p(\mathbf{c}))$ ($\times$ 100)
		10	10	JointVAE	68.99 ± 11.8	74.1	3.12 ± 1.0	
				JointVAE†	68.21 ± 09.6	54.4	3.65 ± 2.7	
				CascadeVAE	74.83 ± 06.9	23.8	1.99 ± 1.1	
				CascadeVAE†	72.98 ± 13.4	17.8	2.44 ± 1.4	
				cpl-mixVAE	$\mathbf{84.56} \pm \mathbf{06.5}$	17.5	$\mathbf{1.48} \pm \mathbf{1.4}$	
Smart-seq ALM-VISp	40K	115	2	JointVAE	12.53 ± 02.9	28.6	459 ± 12	
				CascadeVAE	02.69 ± 00.1	03.4	54.8 ± 0.5	
				cpl-mixVAE	$\mathbf{38.78} \pm \mathbf{01.3}$	10.1	$\mathbf{43.9} \pm \mathbf{3.5}$	
10X MOp	220K	140	2	JointVAE	02.32 ± 00.1	14.8	542 ± 17	
				CascadeVAE	02.85 ± 00.5	00.6	150 ± 3.5	
				cpl-mixVAE	$\mathbf{23.64} \pm \mathbf{00.9}$	08.2	$\mathbf{135} \pm \mathbf{5.2}$	

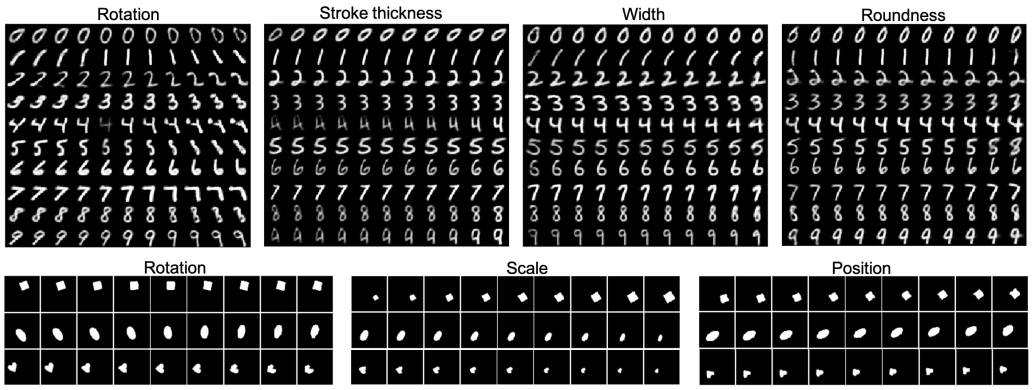


Figure 2: Interpretable continuous latent traversals of 1-st arm of the cpl-mixVAE framework with two autoencoders, for MNIST (top) and dSprites (bottom). The discrete variable $\mathbf{c}$ is constant for all reconstructions in the same row.

et al., 2017), each arm of cpl-mixVAE uses a 10-dimensional categorical variable representing digits (type), and a 10-dimensional continuous random variable representing the writing style (state). Table 1 displays the accuracy of the categorical assignment and the discrepancy between $q(\mathbf{c})$ and $p(\mathbf{c})$ for InfoGAN, two 1-arm VAE methods (JointVAE and CascadeVAE), and cpl-mixVAE with 2 arms. Additionally, to isolate the impact of data augmentation in training, we trained JointVAE† and CascadeVAE† where the models were trained with the same augmented copies of the original MNIST dataset as cpl-mixVAE. The results in Table 1 suggest that data augmentation by itself does not enhance the performance. Fig. 2 (top row) illustrates the continuous latent traversals for four dimensions of the state variable inferred by cpl-mixVAE, where each row corresponds to a different dimension of the categorical variable, and the state variable monotonically changes across columns. Both results in Table 1 and Fig. 2 show that cpl-mixVAE achieved an interpretable mixture representation with the highest categorical assignment accuracy.

Summary. cpl-mixVAE improves the discrete density approximation and infers better mixture representations. It outperforms earlier methods, without using extraneous optimization or heuristic channel capacities. Beyond performance and robustness, its computational cost is also comparable to that of the baselines.

4.2 Single-cell RNA sequencing data

In this dataset the observations are individual cells and each observation consists of expressions of thousands of genes. Here, we used two scRNA-seq datasets: (i) Smart-seq ALM-VISp (Tasic et al., 2018) and (ii) 10X MOp (Yao et al., 2021). The Smart-seq dataset includes transcriptomic profiles of more than $10,000$ genes for $\sim 22,000$ cells from the mouse anterior lateral motor cortex (ALM) and the primary visual cortex (VIPs). The 10X MOp dataset profiles ~ 123000 cells in the mouse primary motor cortex (MOp) with the droplet-based 10X Genomics Chromium platform. 10X-based data often display more gene dropouts, especially for genes with lower expression levels. scRNA-seq datasets are significantly more complex than typical benchmark datasets due to (i) large number of cell types (discrete variable), and (ii) class imbalance; in the 10X MOp dataset, for instance, the most- and the least-abundant cell types include $17,000$ and 20 samples, respectively. Moreover, whether the observed diversity corresponds to discrete variability or a continuum is an ongoing debate in neuroscience (Scala et al., 2020). While using genes that are differentially expressed in subsets of cells, known as *marker genes* (MGs) (Trapnell, 2015) is a common approach to define cell types, the identified genes rarely obey the idealized MG definition in practice. Here, we focus on *neuronal* cells and use a subset of $5,000$ highest variance genes. The original MG-based studies for each dataset suggested 115 (Smart-seq data) (Tasic et al., 2018) and 140 (10X-based data) (Yao et al., 2021) discrete neuronal types.

Neuron type identification. Based on the suggested taxonomies in (Tasic et al., 2018; Yao et al., 2021), for the Smart-seq ALM-VISp data, we used 115- and 2-dimensional discrete and continuous variables, and for the 10X MOp data, we used 140- and 2-dimensional discrete and continuous latent variables. We compared the suggested cell types in (Tasic et al., 2018; Yao et al., 2021) with the discrete representations that are inferred from VAE models. Table 1 and Fig. 3(a-b) demonstrate the

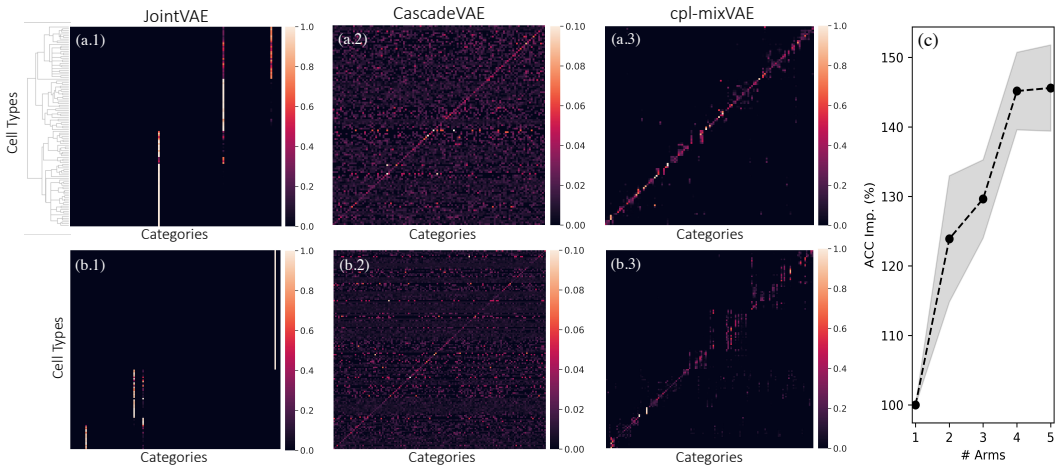
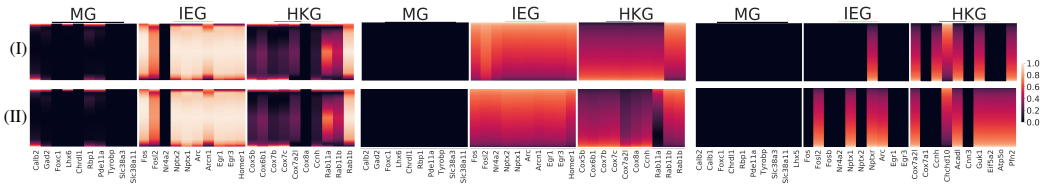


Figure 3: Categorical assignments for the scRNA-seq datasets. The top row shows confusion matrices of JointVAE (a.1), CascadeVAE (a.2), and cpl-mixVAE (a.3) trained by $|\mathbf{c}| = 115$, $|\mathbf{s}| = 2$, for the Smart-seq ALM-VISp dataset. The dendrogram on the y-axis shows MG-based hierarchical classification with 115 cell types, suggested by Tasic et al. (2018). The bottom row shows confusion matrices of JointVAE (b.1), CascadeVAE (b.2), and cpl-mixVAE (b.3) trained by $|\mathbf{c}| = 140$, $|\mathbf{s}| = 2$, for the 10X MOp dataset. Cell types on the y-axis are sorted based on a hierarchical classification suggested by Yao et al. (2021). (c) Improvement of the categorical representation (ACC) of cpl-mixVAE by adding more arms to the multi-arm framework. A-arm's performance for $A \geq 2$ is compared with the baseline 1-arm, JointVAE, across 3 randomly initialized runs.

performance of a 2-arm cpl-mixVAE model against JointVAE and CascadeVAE. In Fig. 3a.1 and 3b.1, we observe that for both datasets, JointVAE succeeds in identifying (sub)classes of neurons, e.g. excitatory class, or Sst subclass, but not neuronal types at leaf nodes. On the other hand, CascadeVAE learns an almost uniform distribution over all types despite a sizeable difference between the relative abundances of neuronal types (Fig. 3a.2 and 3b.2). Our results in Fig. 3(a-b) clearly show that cpl-mixVAE outperforms JointVAE and CascadeVAE in identifying meaningful known cell types. The confusion matrices in Fig. 3a.3 and 3b.3 demonstrate that even the inaccurate categorical assignments of cpl-mixVAE are still close to the matrix diagonals, suggesting a small cophenetic distance. That is, those cells are still assigned to nearby cell types in the dendrogram.

Using $\mathbf{A} > 2$. Unlike the discussed benchmark datasets, the neuronal types are not uniformly distributed. Accordingly, we also investigated the accuracy improvement for categorical assignment when more than two arms are used. Fig. 3c illustrates the accuracy improvement with respect to a single autoencoder model, i.e. JointVAE, in agreement with our theoretical findings.

Identifying genes regulating cell activity. To examine the role of the continuous latent variable, we applied a similar traversal analysis to that used for the benchmark datasets. For a given cell sample and its discrete type, we changed each dimension of the continuous variable using the conditional distribution, and inspected gene expression changes caused by continuous variable alterations. Fig. 4 shows the results of the continuous traversal study for JointVAE and cpl-mixVAE, for two excitatory neurons belonging to the "L5 NP" (cell type (I)) and "L6 CT" (cell type (II)) sub-classes in ALM and MOp regions. Note that here, JointVAE is equivalent to a 1-arm VAE, with the exception of the type dependence of the state variable. Since CascadeVAE did not learn meaningful clustering of cells, even at the subclass level, we did not consider it for the continuous factor analysis. 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 continuous variable, and the x-axis corresponds to three gene subsets, namely (i) MGs for the two excitatory types, (ii) immediate early genes (IEGs), and (iii) housekeeping gene (HKG) subgroups (Hrvatin et al., 2018; Tarasenko et al., 2017). For cpl-mixVAE (Fig. 4b), the normalized expression of the reported MGs as indicators for excitatory cell types (discrete factors) is unaffected by changes of identified continuous variables. In contrast, for JointVAE (Fig. 4a), we observed that the normalized expression of some MGs (5 out of 10) are changed due to the continuous factor traversal. Additionally, we found that the expression changes inferred by cpl-mixVAE for IEGs and HKGs are essentially *monotonically* linked to the continuous variable, confirming that the expression of IEGs and HKGs depends strongly on the cell activity variations under different metabolic and environmental conditions. Conversely, JointVAE



(a) JointVAE - Smart-seq ALM-VISp (b) cpl-mixVAE - Smart-seq ALM-VISp (c) cpl-mixVAE - 10x MOp

Figure 4: Continuous latent traversal analysis for two excitatory cell types: (I) "L5 NP" and (II) "L6 CT", in different brain regions. For each type, the traversal is color-mapped to a normalized reconstructed gene expression value (colorbar) as a function of the state variable for 3 gene subsets: marker genes (MG), immediate early genes (IEG), and housekeeping genes (HKG).

fails to reveal such activity-regulated monotonicity for IEGs and HKGs. Furthermore, our results for cpl-mixVAE reveal that the expression of activity-regulated genes depends on the cell type, i.e. IEGs and HKGs respond differently to activation depending on their cell types (compare rows I and II in Fig. 4b). However, in Fig. 4a, since the baseline JointVAE does not take into account the dependency of discrete and continuous factors, it fails to reveal the dependence of activity-regulated expression to the cell type, and therefore produces identical expressions for both types (I) and (II). These findings are consistent over multiple randomly initialized runs (Supplementary Section K.1). See Supplementary Section K.2 for more results on other cell types and gene subsets.

Summary. The cpl-mixVAE model successfully identified the majority of known excitatory and inhibitory neurons in multiple cortical regions. Our findings suggest that cpl-mixVAE, by acknowledging the dependencies of continuous and categorical factors, captures relevant and interpretable continuous variability that can provide insight when deciphering the molecular mechanisms shaping the landscape of biological states, e.g. due to metabolism or disease.

4.3 Ablation studies

To elucidate the success of the A-arm VAE framework in mixture modeling, we investigate the categorical assignment performance under different training settings. Since CascadeVAE does not learn the categorical factors by variational inference, here we mainly study JointVAE (as a 1-arm VAE) and cpl-mixVAE (as a 2-arm VAE). In Section 4.1, we show that data augmentation by itself does not enhance the categorical assignment (JointVAE†). To understand whether architectural differences put JointVAE at a disadvantage, we trained JointVAE‡ (Table S1), which uses the same architecture as the one used in cpl-mixVAE. JointVAE‡ uses the same learning procedure as JointVAE, but its convolutional layers are replaced by fully-connected layers (see Supplementary Section J and L for details). The result for JointVAE‡ suggests that the superiority of cpl-mixVAE is not due to the network architecture either. We also examined the performance changes of the proposed 2-arm cpl-mixVAE under three different settings: (i) cpl-mixVAE*, where coupled networks are not independent and network parameters are shared; (ii) cpl-mixVAE^a, where only affine transformations are used for data augmentation; and (iii) cpl-mixVAE$(\mathbf{s} \not| \mathbf{c})$, where the state variable is independent of the discrete variable (Table S1). Our results show that the proposed cpl-mixVAE obtained the best categorical assignments across all training settings. We also examined the accuracy of categorical assignments for the cpl-mixVAE model, under different dimensions of discrete latent variable, for both MNIST and scRNA-seq datasets (see Supplementary Section J). We experimentally observe that while JointVAE suffers from sensitivity to empirical choices of $|\mathbf{c}|$, cpl-mixVAE is more robust in encoding the discrete variability, without suffering from mode collapse (Fig. S9 and Fig. S10).

5 Conclusion

We have proposed cpl-mixVAE as a multi-arm framework to apply the power of collective decision making in unsupervised joint representation learning of discrete and continuous factors, scalable to the high-dimensional discrete space. This framework utilizes multiple pairwise-coupled autoencoding arms with a shared categorical variable, while independently learning the continuous variables. Our experimental results for all datasets support the theoretical findings, and show that cpl-mixVAE outperforms comparable models. Importantly, for challenging scRNA-seq dataset, we showed that the proposed framework identifies biologically interpretable cell types and differentiate between type-dependent and activity-regulated genes.

REFERENCES

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

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

Cornelia Bargmann, William Newsome, A Anderson, E Brown, K Deisseroth, J Donoghue, P MacLeish, E Marder, R Normann, J Sanes, et al. Brain 2025: a scientific vision. *Brain Research through Advancing Innovative Neurotechnologies (BRAIN) Working Group Report to the Advisory Committee to the Director, NIH*, 2014.

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.

Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pp. 92–100, 1998.

Diane Bouchacourt, Ryota Tomioka, and Sebastian Nowozin. Multi-level variational autoencoder: Learning disentangled representations from grouped observations. *arXiv preprint arXiv:1705.08841*, 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.

Jacob Deasy, Nikola Simidjievski, and Pietro Liò. Constraining variational inference with geometric jensen-shannon divergence. *arXiv preprint arXiv:2006.10599*, 2020.

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.

Fangjian Guo, Xiangyu Wang, Kai Fan, Tamara Broderick, and David B Dunson. Boosting variational inference. *arXiv preprint arXiv:1611.05559*, 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.

Haruo Hosoya. Group-based learning of disentangled representations with generalizability for novel contents. In *IJCAI*, pp. 2506–2513, 2019.

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.

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.

Diederik P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improving variational inference with inverse autoregressive flow. *arXiv preprint arXiv:1606.04934*, 2016.

Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

Abhishek Kumar and Hal Daumé. A co-training approach for multi-view spectral clustering. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pp. 393–400, 2011.

Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. In *International conference on machine learning*, pp. 1558–1566. PMLR, 2016.

Mihee Lee and Vladimir Pavlovic. Private-shared disentangled multimodal vae for learning of hybrid latent representations. *arXiv preprint arXiv:2012.13024*, 2020.

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*, 2018a.

Francesco Locatello, Gideon Dresdner, Rajiv Khanna, Isabel Valera, and Gunnar Rätsch. Boosting black box variational inference. In *Advances in Neural Information Processing Systems*, pp. 3401–3411, 2018b.

Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. Weakly-supervised disentanglement without compromises. *arXiv preprint arXiv:2002.02886*, 2020.

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

Tom Minka et al. Divergence measures and message passing. Technical report, Citeseer, 2005.

Stefano Monti, Pablo Tamayo, Jill Mesirov, and Todd Golub. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine learning*, 52(1-2):91–118, 2003.

Jozsef Nemeth. Adversarial disentanglement with grouped observations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 10243–10250, 2020.

Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548*, 2017.

Matias Quiroz, David J Nott, and Robert Kohn. Gaussian variational approximation for high-dimensional state space models. *arXiv preprint arXiv:1801.07873*, 2018.

Rajesh Ranganath, Dustin Tran, and David Blei. Hierarchical variational models. In *International Conference on Machine Learning*, pp. 324–333. PMLR, 2016.

Federico Scala, Dmitry Kobak, Matteo Bernabucci, Yves Bernaerts, Cathryn René Cadwell, Jesus Ramon Castro, Leonard Hartmanis, Xiaolong Jiang, Sophie Laturnus, Elanine Miranda, et al. Phenotypic variation of transcriptomic cell types in mouse motor cortex. *Nature*, pp. 1–7, 2020.

H Sebastian Seung and Uygar Sümbül. Neuronal cell types and connectivity: lessons from the retina. *Neuron*, 83(6):1262–1272, 2014.

Rui Shu, Yining Chen, Abhishek Kumar, Stefano Ermon, and Ben Poole. Weakly supervised disentanglement with guarantees. *arXiv preprint arXiv:1910.09772*, 2019.

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.

Kai Tian, Shuigeng Zhou, and Jihong Guan. Deepcluster: A general clustering framework based on deep learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 809–825. Springer, 2017.

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

Michael Tschannen, Olivier Bachem, and Mario Lucic. Recent advances in autoencoder-based representation learning. *arXiv preprint arXiv:1812.05069*, 2018.

Zizhen Yao, Cindy TJ van Velthoven, Thuc Nghi Nguyen, Jeff Goldy, Adriana E Sedeno-Cortes, Fahimeh Baftizadeh, Darren Bertagnolli, Tamara Casper, Megan Chiang, Kirsten Crichton, et al. A taxonomy of transcriptomic cell types across the isocortex and hippocampal formation. *Cell*, 2021.

Hongkui Zeng and Joshua R Sanes. Neuronal cell-type classification: challenges, opportunities and the path forward. *Nature Reviews Neuroscience*, 18(9):530, 2017.

Cheng Zhang, Judith Bütepage, Hedvig Kjellström, and Stephan Mandt. Advances in variational inference. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):2008–2026, 2018.