

# Dual Advancement of Representation Learning and Clustering for Sparse and Noisy Images

## Supplementary Material

### 1 Supplementary Note

#### 1.1 MAP-EM Inference of SMM Parameters

We first list the mathematical notations used in the inference below:

$N$ : the number of images.

$K$ : the number of gene clusters (components in the SMM).

$D$ : the dimension of embeddings.

$\mathbf{Z} \in \mathbb{R}^{N \times D}$ : the *JLRCA*-generated image embeddings.

$\phi_k(\mu_k, \Sigma_k, v_k)$ : the *pdf* of the  $k$ -th SMM component.

$\Pi = \{\pi_k, \forall k \in [1, K]\}$ : the weights of SMM components.

$\Theta = \{\theta_k : \mu_k, \Sigma_k, v_k, \pi_k, \forall k \in [1, K]\}$ : parameters of the SMM components.

$\xi_i \in [1, K]$ : the SMM component membership of  $\mathbf{z}_i$ .

We re-write the Student's  $t$  distribution of the  $k$ -th component of the SMM as a Gaussian scale mixture:

$$\phi(\mathbf{z}_i | \mu_k, \Sigma_k, v_k) = \int N\left(\mathbf{z}_i \mid \mu_k, \frac{\Sigma_k}{\zeta_{i,k}}\right) \Gamma\left(\zeta_{i,k} \mid \frac{v_k}{2}, \frac{v_k}{2}\right) d\zeta_{i,k} \quad (1)$$

Therefore, under the EM framework, all hidden variables are  $H = \{h_i : \xi_i, \zeta_{i,k}, \forall i \in [1, N], \forall k \in [1, K]\}$ . The complete data log likelihood is:

$$\ell_c(\Theta) = (\mathbf{Z}, H | \Theta) = \sum_i \sum_k l(\xi_i = k) (\log \pi_k + \log f(\mathbf{z}_i, \zeta_{i,k} | \mu_k, \Sigma_k, v_k)) \quad (2)$$

$$\log f(\mathbf{z}_i, \zeta_{i,k} | \mu_k, \Sigma_k, v_k) \propto \log \Gamma\left(\zeta_{i,k} \mid \frac{v_k}{2}, \frac{v_k}{2}\right) + \frac{D}{2} \log \zeta_{i,k} - \frac{1}{2} \log |\Sigma_k| - \frac{1}{2} \zeta_{i,k} \sigma_{i,k} \quad (3)$$

$$\sigma_{i,k} = (\mathbf{z}_i - \mu_k)^T \Sigma_k^{-1} (\mathbf{z}_i - \mu_k) \quad (4)$$

*E step.* In the  $t$ -th iteration, we have the auxiliary function  $Q$  as:

$$\begin{aligned} Q(\Theta, \Theta^{(t-1)}) &= E(\ell_c(\Theta) | \Theta^{(t-1)}) \\ &= \sum_i \sum_k p(\xi_i = k | \mathbf{z}_i, \Theta^{(t-1)}) \left( \log \pi_k^{(t-1)} + E\left(\left(\zeta_{i,k} \mid \mu_k^{(t-1)}, \Sigma_k^{(t-1)}, v_k^{(t-1)}\right)\right) \right) \end{aligned} \quad (5)$$

The expected sufficient statistic (ESS) are:

$$\overline{\xi_{i,k}}^{(t)} = p(\xi_i = k | \mathbf{z}_i, \Theta^{(t-1)}) = \frac{\pi_k^{(t-1)} \phi(\mathbf{z}_i | \mu_k^{(t-1)}, \Sigma_k^{(t-1)}, v_k^{(t-1)})}{\sum_{k'} \pi_{k'}^{(t-1)} \phi(\mathbf{z}_i | \mu_{k'}^{(t-1)}, \Sigma_{k'}^{(t-1)}, v_{k'}^{(t-1)})} \quad (6)$$

$$\overline{\zeta_{i,k}}^{(t)} = E\left(p\left(\zeta_{i,k} \mid \mathbf{z}_i, \theta_k^{(t-1)}\right)\right) = E\left(\Gamma\left(\zeta_{i,k} \mid \frac{v_k^{(t-1)} + D}{2}, \frac{v_k^{(t-1)} + \sigma_{i,k}^{(t-1)}}{2}\right)\right) = \frac{v_k^{(t-1)} + D}{v_k^{(t-1)} + \sigma_{i,k}^{(t-1)}} \quad (7)$$

Then the complete data log likelihood of  $\mathbf{z}_i$  becomes:

$$E(\log f(\mathbf{z}_i, \zeta_{i,k} | \mu_k, \Sigma_k, \nu_k)) \propto G(\mathbf{z}_i, \mu_k, \Sigma_k)^{(t)} + F(\zeta_{i,k}, \nu_k)^{(t)} \quad (8)$$

$$G(\mathbf{z}_i, \mu_k, \Sigma_k)^{(t)} = -\frac{1}{2} \log |\Sigma_k| - \frac{\overline{\zeta_{i,k}}^{(t)}}{2} \sigma_{i,k} \quad (9)$$

$$F(\zeta_{i,k}, \nu_k)^{(t)} = \frac{\nu_k \log(\nu_k/2)}{2} - \Gamma\left(\frac{\nu_k}{2}\right) + \frac{\nu_k}{2} \left( \overline{\log \zeta_{i,k}}^{(t)} - \overline{\zeta_{i,k}}^{(t)} \right) \quad (10)$$

*M step.* In the  $t$ -th iteration, we maximize  $Q$  with respect to  $\forall \theta_k \in \Theta$ . Rather than achieving the MLE, we introduce a prior distribution on  $\theta_k$  and solve for MAP of  $\theta_k$  to alleviate model overfitting. Specifically, we introduce a conjugate prior on  $\Pi$  as a Dirichlet distribution:

$$\text{Dir}(\Pi | \alpha^0) \equiv \frac{1}{B(\alpha^0)} \prod_k \pi_k^{\alpha_k^0 - 1} \quad (11)$$

a conjugate prior on  $\{\mu_k, \Sigma_k\}$  as a normal-inverse Wishart (NIW) distribution:

$$\begin{aligned} & \text{NIW}(\mu_k, \Sigma_k | m_0, \kappa_0, S_0, \rho_0) \\ & \propto |\Sigma_k|^{-\frac{1}{2}} \exp\left(-\frac{\kappa_0}{2} (\mu_k - m_0)^T \Sigma_k^{(-1)} (\mu_k - m_0)\right) \\ & \times |\Sigma_k|^{-\frac{(\rho_0 + D + 1)}{2}} \exp\left(-\frac{1}{2} \text{tr}(S_0 \Sigma_k^{-1})\right) \end{aligned} \quad (12)$$

Here, we have weaker priors as  $\alpha_0 = \vec{1}, \rho_0 = D + 2, S_0 = K^{-\frac{1}{D}} \text{diag}((\mathbf{Z} - \frac{1}{N} \Omega \mathbf{Z})^T (Z - \frac{1}{N} \Omega \mathbf{Z})), \Omega = \vec{1}^T \times \vec{1}, \kappa_0 = 0, m_0 = \frac{\sum_i \mathbf{z}_i}{N}$ . The posterior distribution of  $\Pi$  and  $\{\mu_k, \Sigma_k\}$  are:

$$\begin{aligned} p(\Pi | X) & \sim \text{Dir}(\Pi | \alpha^N) \equiv \frac{1}{B(\alpha^N)} \prod_k \pi_k^{\alpha_k^N - 1} \\ \alpha_k^N & = \alpha_k^0 + \sum_i \overline{\xi_{i,k}}^{(t)}, \forall k \in [1, K] \\ p(\mu_k, \Sigma_k | X) & \sim \text{NIW}(\mu_k, \Sigma_k | m_N, \kappa_N, S_N, \rho_N) \\ \kappa_N & = \kappa_0 + \overline{\omega_k}^{(t)} = \overline{\omega_k}^{(t)} \\ \overline{\omega_{i,k}}^{(t)} & = \overline{\xi_{i,k}}^{(t)} \overline{\zeta_{i,k}}^{(t)} \\ \overline{\omega_k}^{(t)} & = \sum_i \overline{\omega_{i,k}}^{(t)} \\ \rho_N & = \rho_0 + \overline{\xi_k}^{(t)} \\ \overline{\xi_k}^{(t)} & = \sum_i \overline{\xi_{i,k}}^{(t)} \\ \mathbf{m}_N & = \frac{\overline{\omega_k}^{(t)} \overline{\mathbf{z}_k}^{(t)} + \kappa_0 m_0}{\kappa_N} = \overline{\mathbf{z}_k}^{(t)} \\ \overline{\mathbf{z}_k}^{(t)} & = \frac{\sum_i (\overline{\omega_{i,k}}^{(t)} \mathbf{z}_i)^{(t)}}{\overline{\omega_k}^{(t)}} \\ S_N & = S_0 + \sum \overline{\omega_{i,k}}^{(t)} \mathbf{z}_i \mathbf{z}_i^T + \kappa_0 m_0 m_0^T - \kappa_N m_N m_N^T = S_0 + \sum \overline{\omega_{i,k}}^{(t)} \mathbf{z}_i \mathbf{z}_i^T - \kappa_N m_N m_N^T \end{aligned} \quad (13)$$

Then we have the MAP estimates of  $\pi_k$  and  $\{\mu_k, \Sigma_k\}$  as  $\pi_k^{(t)}$  and  $\mu_k^{(t)}, \Sigma_k^{(t)}$ :

$$\begin{aligned}\pi_k^{(t)} &= \frac{\alpha_k^N - 1}{\sum_{k'} \alpha_{k'}^N - K} \\ \mu_k^{(t)} &= m_N = \bar{z}_k^{(t)} \\ S_N &= S_0 + \sum_i \left[ \overline{\omega_{i,k}}^{(t)} (\mathbf{z}_i - m_N)(\mathbf{z}_i - m_N)^T \right] = S_0 + S_{\text{mle},k}^{(t)} \\ S_{\text{mle},k}^{(t)} &= \sum_i \left[ \overline{\omega_{i,k}}^{(t)} \left( \mathbf{z}_i - \bar{\mathbf{z}}_k^{(t)} \right) \left( \mathbf{z}_i - \bar{\mathbf{z}}_k^{(t)} \right)^T \right] \\ \Sigma_k^{(t)} &= \frac{S_N}{\rho_N + D + 2} = \frac{S_0 + S_{\text{mle},k}}{\hat{\rho}_0 + \bar{\xi}_k^{(t)}} = \frac{\hat{\rho}_0}{\hat{\rho}_0 + \bar{\xi}_k^{(t)}} \cdot \frac{S_0}{\hat{\rho}_0} + \frac{\bar{\xi}_k^{(t)}}{\hat{\rho}_0 + \bar{\xi}_k^{(t)}} \cdot \frac{S_{\text{mle},k}}{\bar{\xi}_k^{(t)}} = \beta \Sigma_0 + (1 - \beta) \Sigma_{\text{mle},k}^{(t)}\end{aligned}\tag{14}$$

Then we have:

$$\Sigma_k^{(t)}(i, j) = \begin{cases} \beta \Sigma_0(i, j) + (1 - \beta) \Sigma_{\text{mle},k}^{(t)}(i, j), & \text{if } i = j \\ (1 - \beta) \Sigma_{\text{mle},k}^{(t)}(i, j), & \text{otherwise} \end{cases}\tag{15}$$

The off-diagonal entries in  $\Sigma_k^{(t)}$  are shrunk toward 0 to promote its sparsity, thereby reducing the computational load and possibility of overfitting.  $v_k^{(t)}$  can be derived by maximizing  $\Sigma_i \left( \overline{\xi_{i,k}}^{(t)} \cdot F \left( \zeta_{i,k}^{(t)}, v_k^{(t)} \right) \right)$ . However, there is no closed-form solution, so we apply the generalized EM (GEM) to approximate the solution as follows:

$$\begin{aligned}\zeta_{i,k}^{(t)} &\sim \Gamma \left( \frac{v_k^{(t-1)} + D}{2}, \frac{v_k^{(t-1)} + \sigma_{i,k}^{(t-1)}}{2} \right) \\ \implies \overline{\log \zeta_{i,k}^{(t)}} &= E \left( \log \zeta_{i,k}^{(t)} \right) = \Psi \left( \frac{v_k^{(t-1)} + D}{2} \right) - \log \left( \frac{v_k^{(t-1)} + \sigma_{i,k}^{(t-1)}}{2} \right)\end{aligned}\tag{16}$$

Here,  $\Psi(x) \equiv \frac{d}{dx} \Gamma(x)$  is the digamma function. Then we have:

$$\frac{d}{dv_k^{(t)}} \sum_i \left( \overline{\xi_{i,k}}^{(t)} \cdot F(\zeta_{i,k}^{(t)}, v_k^{(t)}) \right) = \sum_i \overline{\xi_{i,k}}^{(t)} \left( \frac{1}{2} \log \left( \frac{v_k^{(t)}}{2} \right) + \frac{1}{2} - \frac{1}{2} \Psi \left( \frac{v_k^{(t)}}{2} \right) + \frac{1}{2} \left( \overline{\log \zeta_{i,k}^{(t)}} - \overline{\zeta_{i,k}^{(t)}} \right) \right)\tag{17}$$

Then  $v_k^{(t)} \leftarrow v_k^{(t)} - \lambda \cdot \frac{d}{dv_k^{(t)}} F \left( \zeta_{i,k}^{(t)}, v_k^{(t)} \right)$  is repeated for several times to achieve a ‘‘partial’’ improvement to  $v_k^{(t)}$ , which still guarantees the convergence to a local optimum. Next, the EM algorithm continues to  $E$  step of the  $(t + 1)$ -th iteration to update  $H^{(t+1)} = [h_{ik}^{(t+1)} : \overline{\xi_{i,k}}^{(t+1)}, \overline{\zeta_{i,k}}^{(t+1)}, \forall i \in [1, N], \forall k \in [1, K]]$  until either convergence is achieved or a prespecified number of iterations is reached. Finally, the score and soft assignment of  $\mathbf{z}_i$  to the  $k$ -th component can be calculated by plugin  $\theta_k$  as:

$$q_{i,k} = \pi_k \phi(\mathbf{z}_i | \mu_k, \Sigma_k, v_k)\tag{18}$$

$$\mathbb{Q}_{i,k} = \frac{q_{i,k}}{\sum_j q_{i,j}}, \forall i \in [1, N], \forall k \in [1, K]\tag{19}$$

## 1.2 Deriving the Seeding Similarity Matrix

During the training of the model, an initial gene-gene similarity matrix is incorporated into the loss function  $\mathcal{L}_1$  as a regularization term to inform the training phase of the model. We leverage multiple image recognition operators to extract feature descriptors from images of gene spatial expression maps, based on which the seeding gene-gene similarity matrix is calculated. Specifically, on the gray-scale level of the image, we utilize Sobel operator to extract gradient magnitude and orientation descriptors, Laplacian operator to extract gradient divergence descriptor and Canny operators to extract the gradient continuity descriptor. Meanwhile, three average and standard deviation pooling filters of different sizes are used to extract patch brightness descriptors. The normalized spatial expression matrix of gene  $u$  is denoted as  $\mathbf{X}^u \in \mathbb{R}^{N_x \times N_y}$ , where  $N_x$  and  $N_y$  denote the number of spatial spots along the horizontal and vertical directions of the spatial map Out-of-tissue spatial spots

are all padded with 0s.  $X^u$  is smoothed with a convolutional Gaussian kernel  $H \in \mathcal{R}^{d \times d}$ , obtaining a denoised expression matrix  $\tilde{X}^u$ :

$$H_{i,j} = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{(i - (k-1)/2)^2 + (j - (k-1)/2)^2}{2\sigma^2}\right), 1 \leq i, j \leq d \quad (20)$$

$$\tilde{X}_{i,j}^u = \text{sum}(H \odot X_{s(k),t(k)}^u); \quad (21)$$

Next, Sober, Laplacian and Canny operators are applied on  $\tilde{X}^u$  to generate matrices of corresponding descriptors. For a specific gene  $u$ , let  $G^u$  and  $\Theta^u$  denote the matrices of gradient magnitude and orientation descriptors,  $L^u$  the matrix of gradient divergence descriptor, and  $C^u$  the matrix of gradient continuity descriptor. The pooling filters are applied by segmenting  $\tilde{X}^u$  into small patches containing  $k \times k$  spots, where  $k \in \{1, 3, 5\}$ , from which three patch brightness mean matrices  $A^u(k)$  and two variance matrices  $S^u(k)$  (eligible only for  $k = 3, 5$ ) are calculated:

$$\mathcal{A}_{i,j}(k) = \text{avg}\left(X_{s(k),t(k)}^u\right) \quad (22)$$

$$\mathcal{S}_{i,j}^u(k) = \text{std}\left(X_{s(k),t(k)}^u\right) \quad (23)$$

$$1 \leq i \leq N_x, \quad 1 \leq j \leq N_y$$

$$s(k) = \left[ i - \frac{(k-1)}{2} : i + \frac{(k-1)}{2} \right], \quad t(k) = \left[ j - \frac{(k-1)}{2} : j + \frac{(k-1)}{2} \right] \quad (24)$$

Finally, the initial similarity matrix  $\mathcal{S}$  is calculated as the average Pearson correlation between gene pairs' descriptor matrices:

$$\mathcal{S}_{u,v} = \begin{cases} \arg(\rho_{u,v}(\Xi^u, \Xi^v)); \Xi \in \{\mathcal{A}(k), \mathcal{S}(k), G, \Theta, L, C\}, k \in \{1, 3, 5\}, & \text{if } u \neq v \\ 0, & \text{if } u = v \end{cases} \quad (25)$$

Here,  $u, v$  represent any two images in the dataset.

### 1.3 The Joint Optimization of the Image Representation Learning Model and SMM via a Discriminative Boosted Clustering

In this section, we focus on deriving the gradients of  $\mathcal{L}_{\ell\ell}$  and  $\mathcal{L}_{size}$  with respect to  $\mathbf{Z}$ , and the gradients of  $\mathcal{L}_{rec}$  with respect to  $\mathbf{Z}$  and  $\Theta$ . The derivations of the gradients of  $\mathcal{L}_{lap}$  and  $\mathcal{L}_{rec}$  with respect to  $\mathbf{Z}$ , the gradients of  $\mathcal{L}_{rec}$  with respect to  $\hat{X}$ , and the gradients of  $\mathcal{L}_{clr}$  with respect to  $\mathbf{e}$  (see Equation 10 in the main text) [?] are relatively trivial and therefore ignored. First,  $\mathcal{L}_{\ell\ell}$ ,  $\mathcal{L}_{kl}$  and  $\mathcal{L}_{size}$  can be expressed as:

$$\mathcal{L}_{\ell\ell} = \log\mathcal{P}(Z|\Theta) = \sum_{i=1}^N \log\left[\sum_k q_{i,k}\right] \quad (26)$$

$$\mathcal{L}_{kl} = KL(P|Q) = \sum_i^N \sum_j^K \mathbb{P} \log \frac{\mathbb{P}_{i,j}}{\mathbb{Q}_{i,j}}, \text{ where } \mathbb{P}_{i,k} = \frac{\mathbb{Q}_{i,k}^2 / \sum_i \mathbb{Q}_{i,j}}{\sum_j (\mathbb{Q}_{i,j}^2 / \sum_i \mathbb{Q}_{i,j})} \quad (27)$$

$$\mathcal{L}_{size}(Z, \Theta) = \sum_{k=1}^K -J_k \log J_k, \text{ where } J_k = \begin{cases} \frac{\sum_i^N \mathbb{Q}_{i,k}}{N}, & \text{if } J_k \leq \tau \\ 1, & \text{otherwise} \end{cases} \quad (28)$$

The density function of  $\mathbf{z}_i$  given  $\{\mu_k, \Sigma_k, \nu_k\}$  is:

$$\begin{aligned} \phi(\mathbf{z}_i | \mu_k, \Sigma_k, \nu_k) &\propto \frac{\Gamma\left(\frac{\nu_k + D}{2}\right)}{\Gamma\left(\frac{\nu_k}{2}\right)} \nu_k^{-\frac{D}{2}} |\Sigma_k|^{-\frac{1}{2}} \left[ 1 + \frac{1}{\nu_k} (\mathbf{z}_i - \mu_k)^T \Sigma_k^{-1} (\mathbf{z}_i - \mu_k) \right]^{-\left(\frac{\nu_k + D}{2}\right)} \\ &= \ln(\nu_k) |\Sigma_k|^{-\frac{1}{2}} \left[ 1 + \frac{\sigma_{i,k}}{\nu_k} \right]^{-\left(\frac{\nu_k + D}{2}\right)} = \ln(\nu_k) |\Sigma_k|^{-\frac{1}{2}} \mu_{i,k}^{-\left(\frac{\nu_k + D}{2}\right)} \end{aligned} \quad (29)$$

The partial derivative of  $q_{i,k}$  with respect to  $\mathbf{z}_i$  can be represented as:

$$\begin{aligned}\frac{\partial q_{i,k}}{\partial \mathbf{z}_i} &= \frac{\partial q_{i,k}}{\partial u_{i,k}} \cdot \frac{\partial u_{i,k}}{\partial \mathbf{z}_i} = -\pi_k \mathbb{h}(v_k) |\Sigma_k|^{-\frac{1}{2}} \left( \frac{v_k + D}{2} \right)^{-\left(\frac{v_k + D}{2} + 1\right)} u_{i,k}^{-1} \cdot \frac{2}{v_k} \Sigma_k^{-1} (\mathbf{z}_i - \mu_k) \\ &= - \left( \frac{v_k + D}{v_k} \right) \mu_{i,k}^{-1} q_{i,k} \cdot \Sigma_k^{-1} (\mathbf{z}_i - \mu_k)\end{aligned}\quad (30)$$

The partial derivative of  $q_{i,k}$  with respect to  $\mu_k$  can be expressed as follows:

$$\frac{\partial q_{i,j}}{\partial \mu_k} = \begin{cases} \frac{\partial q_{i,k}}{\partial u_{i,k}} \cdot \frac{\partial u_{i,k}}{\partial \mu_k} = - \left( \frac{v_k + D}{v_k} \right) u_{i,k}^{-1} q_{i,k} \cdot \Sigma_k^{-1} (\mu_k - \mathbf{z}_i), & j = k \\ 0, & j \neq k \end{cases}\quad (31)$$

The expression for the partial derivative of  $q_{i,k}$  with respect to  $\Sigma_k$  can be articulated as follows:

$$\begin{aligned}\frac{\partial q_{i,j}}{\partial \Sigma_k} &= \begin{cases} \frac{\partial q_{i,k}}{\partial u_{i,k}} \cdot \frac{\partial u_{i,k}}{\partial \Sigma_k} + \frac{\partial q_{i,k}}{\partial |\Sigma_k|} \cdot \frac{\partial |\Sigma_k|}{\partial \Sigma_k}, & j = k \\ 0, & j \neq k \end{cases} \\ &= \begin{cases} q_{i,k} \left( \left( \frac{v_k + D}{2v_k} \right) u_{i,k}^{-1} \cdot \Sigma_k^{-1} (\mathbf{z}_i - \mu_k) (\mathbf{z}_i - \mu_k)^T \Sigma_k^{-1} - \frac{1}{2} \Sigma_k^{-1} \right) = q_{i,k} f(\mathbf{z}_i, \mu_k, v_k, \Sigma_k, D), & j = k \\ 0, & j \neq k \end{cases}\end{aligned}\quad (32)$$

The formula for the partial derivative of  $q_{i,k}$  with respect to  $v_k$  is as follows:

$$\begin{aligned}\frac{\partial q_{i,j}}{\partial v_k} &= \begin{cases} q_{i,k} \frac{\partial \ln(q_{i,k})}{\partial v_k}, & j = k \\ 0, & j \neq k \end{cases} \\ &= \begin{cases} q_{i,k} \left( \frac{v_k + D}{2} u_{i,k}^{-1} \frac{\sigma_{i,k}}{v_k^2} - \frac{1}{2} \ln u_{i,k} + \frac{1}{2} \Gamma \left( \frac{v_k + D}{2} \right) \Psi \left( \frac{v_k + D}{2} \right) - \frac{1}{2} \Gamma \left( \frac{v_k}{2} \right) \Psi \left( \frac{v_k}{2} \right) - \frac{D}{2v_k} \right) = q_{i,k} g(\mathbf{z}_i, \mu_k, v_k, \Sigma_k, D), & j = k \\ 0, & j \neq k \end{cases}\end{aligned}\quad (33)$$

The calculation of the partial derivative of  $q_{i,k}$  with regard to  $\pi_i$  is as follows:

$$\frac{\partial q_{i,j}}{\partial \pi_k} = \begin{cases} \frac{q_{i,k}}{\pi_k}, & j = k \\ 0, & j \neq k \end{cases}\quad (34)$$

By chain rules of derivatives, we have the partial derivative of  $\mathbb{Q}_{i,k}$  with respect to  $\mathbf{z}_k$  can be represented as:

$$\frac{\partial \mathbb{Q}_{i,k}}{\partial \mathbf{z}_i} = \frac{\frac{\partial q_{i,k}}{\partial \mathbf{z}_i} \cdot \sum_j q_{i,j} - q_{i,k} \cdot \sum_j \frac{\partial q_{i,j}}{\partial \mathbf{z}_i}}{\left( \sum_j q_{i,j} \right)^2} = - \left( \frac{v_k + D}{v_k} \right) \mathbb{Q}_{i,k} \left( \mu_{i,k}^{-1} \cdot \Sigma_k^{-1} (\mathbf{z}_i - \mu_k) - \sum_j \mu_{i,j}^{-1} \mathbb{Q}_{i,j} \cdot \Sigma_j^{-1} (\mathbf{z}_i - \mu_j) \right)\quad (35)$$

The partial derivative of  $\mathbb{Q}_{i,k}$  with respect to  $\mu_k$  is given by:

$$\begin{aligned}\frac{\partial \mathbb{Q}_{i,j}}{\partial \mu_k} &= \begin{cases} \frac{\frac{\partial q_{i,j}}{\partial \mu_k} \cdot \sum_j q_{i,j} - q_{i,j} \cdot \sum_j \frac{\partial q_{i,j}}{\partial \mu_k}}{\left( \sum_j q_{i,j} \right)^2}, & j \neq k \\ \frac{\partial q_{i,k}}{\partial \mu_k} \cdot \sum_j q_{i,j} - q_{i,k} \cdot \sum_j \frac{\partial q_{i,j}}{\partial \mu_k}, & j = k \end{cases} \\ &= \begin{cases} \left( \frac{v_k + D}{v_k} \right) \mathbb{Q}_{i,j} \left( \mu_{i,k}^{-1} \mathbb{Q}_{i,k} \cdot \Sigma_k^{-1} (\mu_k - \mathbf{z}_i) \right), & j \neq k \\ - \left( \frac{v_k + D}{v_k} \right) \mathbb{Q}_{i,k} \left( \mu_{i,k}^{-1} \cdot \Sigma_k^{-1} (\mu_k - \mathbf{z}_i) - \mu_{i,k}^{-1} \mathbb{Q}_{i,k} \cdot \Sigma_k^{-1} (\mu_k - \mathbf{z}_i) \right), & j = k \end{cases}\end{aligned}\quad (36)$$

The derivative of  $\mathbb{Q}_{i,k}$  with respect to  $\Sigma_k$  is expressed as:

$$\begin{aligned}\frac{\partial \mathbb{Q}_{i,j}}{\partial \Sigma_k} &= \begin{cases} \frac{\frac{\partial q_{i,j}}{\partial \Sigma_k} \cdot \sum_j q_{i,j} - q_{i,j} \cdot \sum_j \frac{\partial q_{i,j}}{\partial \Sigma_k}}{\left( \sum_j q_{i,j} \right)^2}, & j \neq k \\ \frac{\partial q_{i,k}}{\partial \Sigma_k} \cdot \sum_j q_{i,j} - q_{i,k} \cdot \sum_j \frac{\partial q_{i,j}}{\partial \Sigma_k}, & j = k \end{cases} \\ &= \begin{cases} -\mathbb{Q}_{i,j} \left( \mathbb{Q}_{i,k} \cdot \mathbb{f}(\mathbf{z}_i, \mu_k, v_k, \Sigma_k, D) \right), & j \neq k \\ \mathbb{Q}_{i,k} \left( \mathbb{f}(\mathbf{z}_i, \mu_k, v_k, \Sigma_k, D) - \mathbb{Q}_{i,k} \cdot \mathbb{f}(\mathbf{z}_i, \mu_k, v_k, \Sigma_k, D) \right), & j = k \end{cases}\end{aligned}\quad (37)$$

The partial derivative of  $q_{i,k}$  with respect to  $v_k$  can be denoted as:

$$\frac{\partial q_{i,j}}{\partial v_k} = \begin{cases} \frac{\frac{\partial q_{i,j}}{\partial v_k} \cdot \sum_j q_{i,j} - q_{i,j} \cdot \sum_j \frac{\partial q_{i,j}}{\partial v_k}}{(\sum_j q_{i,j})^2} = -q_{i,j} q_{i,k} g(\mathbf{z}_i, \mu_k, v_k, \Sigma_k, D), j \neq k \\ \frac{\frac{\partial q_{i,k}}{\partial v_k} \cdot \sum_j q_{i,j} - q_{i,k} \cdot \sum_j \frac{\partial q_{i,j}}{\partial v_k}}{(\sum_j q_{i,j})^2} = q_{i,k} (1 - q_{i,k}) g(\mathbf{z}_i, \mu_k, v_k, \Sigma_k, D), j = k \end{cases} \quad (38)$$

The expression for the partial derivative of  $q_{i,k}$  with respect to  $\pi_k$  is as follows:

$$\frac{\partial q_{i,j}}{\partial \pi_k} = \begin{cases} \frac{\partial q_{i,j}}{\partial \pi_k} \cdot \sum_j q_{i,j} - q_{i,j} \cdot \sum_j \frac{\partial q_{i,j}}{\partial \pi_k} = -\frac{q_{i,j} q_{i,k}}{\pi_k}, j \neq k \\ \frac{\partial q_{i,k}}{\partial \pi_k} \cdot \sum_j q_{i,j} - q_{i,k} \cdot \sum_j \frac{\partial q_{i,j}}{\partial \pi_k} = \frac{q_{i,k} (1 - q_{i,k})}{\pi_k}, j = k \end{cases} \quad (39)$$

Then the derivatives of  $\mathcal{L}_{\ell\ell}$  and  $\mathcal{L}_{size}$  with respect to  $\mathbf{z}_i$  are:

$$\frac{\partial \mathcal{L}_{\ell\ell}}{\partial \mathbf{z}_i} = \frac{\sum_j \frac{\partial q_{i,j}}{\partial \mathbf{z}_i}}{\sum_j q_{i,j}} = -\left(\frac{v_k + D}{v_k}\right) \frac{\mu_{i,k}^{-1} q_{i,k} \cdot \Sigma_k^{-1} (\mathbf{z}_i - \mu_k)}{\sum_j q_{i,j}} \quad (40)$$

$$\frac{\partial \mathcal{L}_{size}}{\partial \mathbf{z}_i} = \sum_{j \in \{J_j \leq \tau\}} \frac{-(1 + \log J_j)}{N} \sum_{i=1}^N \frac{\partial q_{i,j}}{\partial \mathbf{z}_i} \quad (41)$$

As the target distribution  $P$  is fixed during the joint optimization within an epoch, we have the derivatives of  $\mathcal{L}_{kl}$  with respect to  $\mathbf{z}_i$  and  $\theta_k = \{\mu_k, \Sigma_k, v_k, \pi_k\}$  as:

$$\frac{\partial \mathcal{L}_{kl}}{\partial \mathbf{z}_i} = -\sum_j \left[ \frac{\mathbb{P}_{i,j}}{q_{i,j}} \cdot \frac{\partial q_{i,j}}{\partial \mathbf{z}_i} \right] = \left(\frac{v_k + D}{v_k}\right) \sum_j (\mathbb{P}_{i,k} - q_{i,k}) \mu_{i,k}^{-1} \cdot \Sigma_k^{-1} (\mathbf{z}_i - \mu_k) \quad (42)$$

$$\frac{\partial \mathcal{L}_{kl}}{\partial \mu_k} = -\sum_i \sum_j \left[ \frac{\mathbb{P}_{i,j}}{q_{i,j}} \cdot \frac{\partial q_{i,j}}{\partial \mu_k} \right] = \left(\frac{v_k + D}{v_k}\right) \sum_i (\mathbb{P}_{i,k} - q_{i,k}) \mu_{i,k}^{-1} \cdot \Sigma_k^{-1} (\mu_k - \mathbf{z}_i) \quad (43)$$

$$\frac{\partial \mathcal{L}_{kl}}{\partial \Sigma_k} = -\sum_i \sum_j \left[ \frac{\mathbb{P}_{i,j}}{q_{i,j}} \cdot \frac{\partial q_{i,j}}{\partial \Sigma_k} \right] = \sum_i (q_{i,k} - \mathbb{P}_{i,k}) f(\mathbf{z}_i, \mu_k, v_k, \Sigma_k, D) \quad (44)$$

$$\frac{\partial \mathcal{L}_{kl}}{\partial v_k} = -\sum_i \sum_j \left[ \frac{\mathbb{P}_{i,j}}{q_{i,j}} \cdot \frac{\partial q_{i,j}}{\partial v_k} \right] = -\sum_i (\mathbb{P}_{i,k} - q_{i,k}) g(\mathbf{z}_i, \mu_k, v_k, \Sigma_k, D) \quad (45)$$

$$\frac{\partial \mathcal{L}_{kl}}{\partial \pi_k} = -\sum_i \sum_j \left[ \frac{\mathbb{P}_{i,j}}{q_{i,j}} \cdot \frac{\partial q_{i,j}}{\partial \pi_k} \right] = -\sum_i \frac{(\mathbb{P}_{i,k} - q_{i,k})}{\pi_k} \quad (46)$$

## 1.4 Creating the Training and Testing Datasets for Evaluating the Prediction on Gene-gene Interactions

A pair of genes is considered interacting (positive) if they share GO terms obtained using the R package ‘‘org.Hs.eg.db’’[?]; otherwise, they are considered noninteracting (negative). To reduce the number of false positive gene pairs, we omit the highly over-represented GO terms including ‘‘single transduction’’ (GO:0007165), three terms related to phosphorylation (‘‘protein amino acid phosphorylation’’, GO:0006468; ‘‘protein amino acid autophosphorylation’’, GO:0046777; ‘‘protein amino acid dephosphorylation’’, GO:0006470), as well as all terms at the first three layers of the GO hierarchy [?].

The experiment defines its universal gene set as the intersection of genes found in the hDLPFC datasets and the GO database, totaling 9,187 genes. This set includes 820,519 positive samples from 9,030 human genes and 41,375,372 negative samples from 9,187 human genes. The positive dataset incorporates all positive samples, and a matching number of randomly chosen negative samples forms the negative dataset. For the gene-gene interaction predictor neural network (GGIPNN), which is a basic MLP-based network as described in Du et al., 2019 [?], we sample 2% of gene pairs from both datasets for training, while the remaining 98% serve as the test set. These gene pairs are then employed for training and evaluation through linear probing across six datasets (151671-151676).

## 2 Supplementary Table

### 2.1 Representation Learning of Artificially Sparsified Images

Method	Clustering (ACC%)				Classification (ACC%)			
	CIFAR-10	CIFAR-10*	STL-10	STL-10*	CIFAR-10	CIFAR-10*	STL-10	STL-10*
MAE	30.97	17.85	33.29	21.01	45.46	23.74	46.42	31.99
SimCLR	60.32	21.99	59.34	28.89	76.17	32.18	67.01	35.13

Table 1: We evaluate the impacts of image sparsity on the efficacy of MIM (i.e., MAE) and CL (i.e., SimCLR) methods for yielding image representations. Images from CIFAR-10 and STL-10 datasets are sparsified using a 90% random pixel masking, creating CIFAR-10\* and STL-10\* datasets. The quality of representations learned from original and sparsified images are evaluated via downstream clustering and classification tasks, using K-means and logistic regression model, respectively. For clustering, the cluster labels are matched to ground truth image labels using the Hungarian algorithm [?]. For classification, the logistic regression model is trained via five-fold cross validation. The performance of both clustering and classification are measured in accuracy (ACC). Our results indicate degraded representations learned from sparsified images for both MAE and SimCLR, as evidenced by the significantly declined clustering and classification accuracy, compared to those learned from original images.

### 2.2 Effectiveness of Conventional Data Augmentation in Contrastive Learning with Sparsified Images

Data augmentation	Clustering (ACC%)				Classification (ACC%)			
	CIFAR-10	CIFAR-10*	STL-10	STL-10*	CIFAR-10	CIFAR-10*	STL-10	STL-10*
full	60.32	21.99	59.34	28.89	76.17	32.18	67.01	35.13
w/o rotating	55.74	22.18	61.08	29.63	76.70	31.92	64.14	37.59
w/o cropping	42.27	22.07	59.25	29.01	73.83	32.96	63.17	36.93
w/o greyscaling	40.30	21.08	47.41	27.86	68.98	32.08	62.24	37.42

Table 2: We evaluate the effectiveness of three conventional data augmentation techniques—rotation, cropping, and greyscaling—for contrastive learning (i.e., SimCLR) on sparsified images in the CIFAR-10 and STL-10 datasets, which are obtained in the same manner as in Supplementary 2.1. Experiment setups include “full” (all methods combined) and “w/o \*” (excluding “\*”, which can be rotating, cropping or greyscaling). The quality of learned representations are evaluated by clustering and classification, following the same procedure as in Supplementary 2.1. With original images, the significant drops in accuracy for the “w/o” setups, as compared to the “full” experiment, underscore the crucial role of each augmentation method in the learning process. Conversely, such declines are absent with sparsified images, indicating negligible effects of these augmentations on contrastive learning in this context. Meanwhile, the significant underperformance of SimCLR on sparsified images indicates the failure of data augmentation in facilitating contrastive learning.

### 3 Supplementary Figure

#### 3.1 Spatial Co-functional Gene images

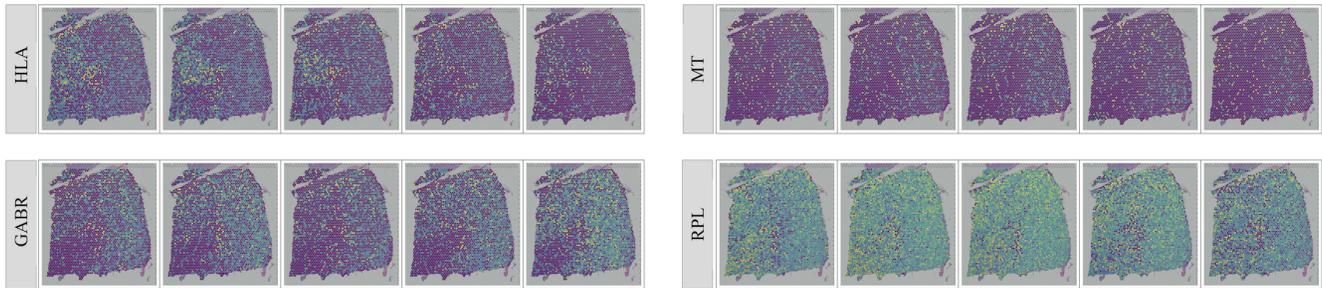


Figure 1: Spatial gene expression images for five member genes from each of the four gene families (i.e., HLA, MT, CABR, and RPL) in the dataset 151676.

### References