Multi-modal Differentiable Unsupervised Feature Selection

Junchen Yang[1] **Ofir Lindenbaum**[2] **Yuval Kluger**[1,4,5] **Ariel Jaffe**[3]

[1]Interdepartmental Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT, USA
[2]Faculty of Engineering, Bar-Ilan University, Israel
[3]Department of Statistics and Data Science, Hebrew University of Jerusalem, Israel
[4]Applied Math Program, Yale University, New Haven, CT, USA
[5]Department of Pathology, School of Medicine, Yale University, New Haven, CT, USA

Abstract

Multi-modal high throughput biological data presents a great scientific opportunity and a significant computational challenge. In multi-modal measurements, every sample is observed simultaneously by two or more sets of sensors. In such settings, many observed variables in both modalities are often nuisance and do not carry information about the phenomenon of interest. Here, we propose a multi-modal unsupervised feature selection framework: identifying informative variables based on coupled high-dimensional measurements. Our method is designed to identify features associated with two types of latent low-dimensional structures: (i) shared structures that govern the observations in both modalities, and (ii) differential structures that appear in only one modality. To that end, we propose two Laplacian-based scoring operators. We incorporate the scores with differentiable gates that mask nuisance features and enhance the accuracy of the structure captured by the graph Laplacian. The performance of the new scheme is illustrated using synthetic and real datasets, including an extended biological application to single-cell multi-omics.

1 INTRODUCTION

In an effort to study biological systems, researchers are developing cutting-edge techniques that measure up to tens of thousands of variables at single-cell resolution. In recent years, research into the interplay between complex biological processes has inspired the development of multi-modal technologies that enable the simultaneous collection of measurements from two or more sets of sensors. Examples of such multi-modal measurements include SHARE-seq [Ma et al., 2020], DBiT-seq [Liu et al., 2020], CITE-seq [Stoeckius et al., 2017], etc., which have provided biological insights and advancements in applications such as transcription factor characterization [Joung et al., 2023], cell type identification in human hippocampus [Xiao et al., 2022], and immune cell profiling [Leblay et al., 2020].

Multi-modal learning is a powerful tool widely used across multiple disciplines to extract latent information from high-dimensional measurements [Sun, 2013, Yan et al., 2021]. Humans use complementary senses when attempting to "estimate" spoken words or sentences [Raij et al., 2000]. For example, lip movements can help us distinguish between two syllables that sound similar. The same intuition has inspired statisticians and machine learning researchers to develop learning techniques that exploit information captured simultaneously by complementary measurement devices.

The applicability of multi-modal datasets in multiple domains, has motivated the development of computational approaches tailored to multi-modal settings. Algorithms such as Contrastive Language–Image Pre-training (CLIP) [Radford et al., 2021] and Audioclip [Guzhov et al., 2022] have pushed the performance boundaries of machine learning for image, text, audio, analysis, and synthesis. The multi-modal data fusion task dates back to Hotelling [1936], which proposed the celebrated Canonical Correlation Analysis (CCA). CCA has many extensions [Andrew et al., 2013, Lindenbaum et al., 2022] and applications in diverse scientific domains [Pimentel et al., 2018, Chen et al., 2017]. Despite their tremendous success, classical or advanced multi-modal schemes are often unsuitable for analyzing biological data.

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

The large number of nuisance variables, which often exceeds the number of measurements, often causes correlation-based methods to overfit.

To attenuate the influence of nuisance or noisy features, several authors proposed unsupervised feature selection (UFS) schemes [Solorio-Fernández et al., 2020]. UFS seeks small subsets of informative variables in order to improve downstream analysis tasks, such as clustering or manifold learning. Empirical results demonstrate that informative features are often smooth and reflect some latent structure [Degeest et al., 2018]. In practice, the smoothness of features can be evaluated based on how slowly they vary with respect to a graph [He et al., 2005]. Follow-up works exploited this idea to identify informative features [Zhao and Liu, 2012, Shaham et al., 2022]. An alternative paradigm for UFS seeks subsets of features that can be used to reconstruct the entire data effectively [Balın et al., 2019].

While most fusion methods focus on extracting information shared between modalities, we propose a multi-modal UFS framework to identify features associated both with structures that appear in both modalities, and structures that are *modality-specific*, and appear in only one modality. To capture the shared structure, we construct a symmetric shared graph Laplacian operator that enhances the shared geometry across modalities. We further propose differential graph operators that capture smooth structures that are not shared with the other modality. To perform multi-modal feature selection, we incorporate differentiable gates [Yamada et al., 2020] with the *shared* and *modality-specific* graph Laplacian scoring functions. This leads to a differentiable UFS scheme that attenuates the influence of nuisance features during training and computes a more accurate Laplacian matrix [Lindenbaum et al., 2021].

Our contributions are four folds: (i) Develop a *shared* and *modality-specific* Laplacian scoring operators. (ii) Motivate our operators using a product of manifolds model. (iii) develop and implement a differentiable framework for multi-modal UFS. (iv) Evaluate the merits and limitations of our approach with synthetic and real data and compare it to existing schemes.

2 PROBLEM SETTING AND PRELIMINARIES

We are given two data matrices $\boldsymbol{X} \in \mathbb{R}^{n\times d}, \boldsymbol{Y} \in \mathbb{R}^{n\times m}$ whose rows contain n observations captured simultaneously in two modalities. The two sets of observations can be, for example, two arrays of sensors, cameras with different angles, etc. We are interested in processing modalities with bijective correspondences, which implies that there is a registration between the observations in both modalities.

Though the observations are high-dimensional, we assume that there are a small number of parameters governing the physical processes that underlies the data. These parameters can be continuous such as in a developmental process, or discrete - for example, when the observations are separated into distinct clusters. However, the latent structure in both modalities may not be identical. For example, the two sets of observations may be generated by sets of sensors with different resolutions or sensitivity. For illustration, consider the observations shown in Fig. 1 (left). Both modalities follow a very similar tree structure. The bottom tree, however, has an additional bifurcating point that does not appear in the upper tree (green points).

Thus, we assume the latent parameters in each modality can be partitioned into two components. The first, denoted $\boldsymbol{\theta}_s$, captures the structures shared by both modalities. The second, denoted $\boldsymbol{\theta}_x$ for modality $\boldsymbol{X}$, and $\boldsymbol{\theta}_y$ for modality $\boldsymbol{Y}$, captures the modality-specific structures that only appear in one set of observations. For example, the additional branch in the bottom tree (modality $\boldsymbol{Y}$) in Fig. 1 is governed by a parameter in $\boldsymbol{\theta}_y$. Thus, the observations $\boldsymbol{X}$ and $\boldsymbol{Y}$ are nonlinear transformations of $\boldsymbol{\theta}_s, \boldsymbol{\theta}_x$ and $\boldsymbol{\theta}_s, \boldsymbol{\theta}_y$, respectively.

Many biological data modalities are high dimensional and contain noisy features, which hinders the discovery of the underlying shared or modality-specific structures. Here, our goal is to identify groups of features associated with the shared structures $\boldsymbol{\theta}_s$ (e.g., the groups of features that are smooth with respect to the shared bifurcated tree in Fig. 1) and groups of features associated with the modality-specific structures $\boldsymbol{\theta}_x$ and $\boldsymbol{\theta}_y$ (e.g., the features that are smooth with respect to the additional branch of modality $\boldsymbol{Y}$ in Fig. 1). To achieve this goal, we compute two graphs that correspond to the two modalities. We use a spectral method to uncover the shared and graph-specific structures and apply a feature selection method to detect variables relevant to these structures. To better understand our approach, we first introduce some preliminaries about graph representation in Sec. 2.1, and discuss related work on feature selection in Sec. 2.2.

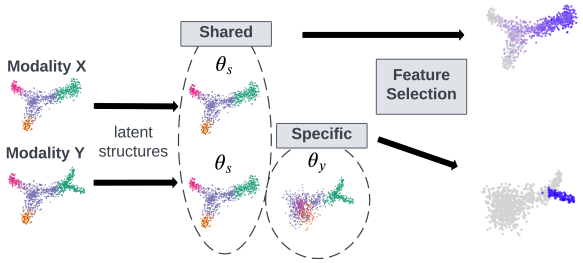


Figure 1: Overview of the goal: discovering features associated with shared and modality-specific latent structures

2.1 THE GRAPH LAPLACIAN AND LAPLACIAN SCORE

A common assumption when analyzing high-dimensional datasets is that their latent, underlying structure can be approximated by a low dimensional manifold [Linderman et al., 2019, Peterfreund et al., 2020]. Methods for manifold learning are often based on graphs that capture the affinities between data points. Let $\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}$ denote the i-th observation in the $\boldsymbol{X}$ and $\boldsymbol{Y}$ modalities and let $\boldsymbol{K}_x, \boldsymbol{K}_y$ be, respectively, their affinity matrices whose elements are computed by the following Gaussian kernel functions,

$$(\boldsymbol{K}_x)_{i,j} = \exp\Big(-\frac{\|\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)}\|^2}{2\sigma_x^2}\Big),$$
$$(\boldsymbol{K}_y)_{i,j} = \exp\Big(-\frac{\|\boldsymbol{y}^{(i)} - \boldsymbol{y}^{(j)}\|^2}{2\sigma_y^2}\Big),$$

where σ_x, σ_y are user-defined bandwidths that control the decay of each Gaussian kernel. Intuitively, the affinities decay exponentially with the distances between samples, thus capturing the local neighborhood structure in the high-dimensional space.

We compute the normalized Laplacian matrix by $\boldsymbol{L}_x = \boldsymbol{D}_x^{-\frac{1}{2}} \boldsymbol{K}_x \boldsymbol{D}_x^{-\frac{1}{2}}$, where $\boldsymbol{D}_x$ is a diagonal matrix of row sums of $\boldsymbol{K}_x$. Similarly, $\boldsymbol{L}_y$ is computed for modality $\boldsymbol{Y}$. An important property of the Laplacian matrix is that its eigenvectors corresponding to large eigenvalues reflect the underlying geometry of the data. The Laplacian eigenvectors are used for many applications, including data embeddings [Belkin and Niyogi, 2003], clustering [Von Luxburg, 2007], and feature selection [He et al., 2005]. For the latter, a popular metric for unsupervised identification of informative features is the Laplacian Score (LS) [He et al., 2005],

$$\boldsymbol{f}^T \boldsymbol{L}_x \boldsymbol{f} = \sum_{i=1}^{n} \lambda_i (\boldsymbol{f}^T \boldsymbol{u}_i)^2, \tag{1}$$

where $\boldsymbol{L}_x = \sum_{i=1}^{n} \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^T$ is the eigendecomposition of $\boldsymbol{L}_x$ and $\boldsymbol{f}$ is the normalized feature vector. Intuitively, when $\boldsymbol{f}$ varies slowly with respect to the underlying structure of $\boldsymbol{L}_x$, it will have a significant component projected onto the subspace of its top eigenvectors, and a higher score.

2.2 DIFFERENTIABLE UNSUPERVISED FEATURE SELECTION

A key limitation of the Laplacian score stems from the underlying assumption that the Laplacian matrix $\boldsymbol{L}_x$ accurately reflects the latent structure of the data. This assumption, however, may not be valid in the presence of many noisy features. In such cases the top eigenvectors of $\boldsymbol{L}_x$ may be heavily influenced by noise and would not capture the underlying structure accurately. A recent work [Lindenbaum et al., 2021] addresses this problem by developing Differentiable Unsupervised Feature Selection (DUFS), a framework that estimates the Laplacian matrix while simultaneously selecting informative features using Laplacian scores. Specifically, DUFS computes a binary vector $\boldsymbol{s} \in \{0,1\}^d$ that indicates which features are kept ($s_j = 1$) and which features are not ($s_j = 0$). Let $\Delta(\boldsymbol{s})$ denote a diagonal matrix with $\boldsymbol{s}$ on the diagonal. At each iteration of DUFS, the Laplacian is computed based on $\tilde{\boldsymbol{X}} = \boldsymbol{X}\Delta(\boldsymbol{s})$, while simultaneously updaing $\boldsymbol{s}$ by optimizing over the following loss function,

$$\mathcal{L} = -\frac{1}{n}\text{Tr}[\tilde{\boldsymbol{X}}^T \boldsymbol{L}_{\tilde{x}} \tilde{\boldsymbol{X}}] + \lambda \|\boldsymbol{s}\|_0, \tag{2}$$

where Tr[] denotes the matrix trace. The first term equals the sum of Laplacian Scores across all features normalized by the total number of samples n in a training batch. The second term is a ℓ_0 regularizer that imposes sparsity to the number of selected features, with λ being a tunable parameter that controls the sparsity level. The output of DUFS is a list of a small number of selected features, and the Laplacian matrix $\boldsymbol{L}_{\tilde{x}}$ learned from them.

However, the discrete nature of the ℓ_0 regularizer, makes the objective in (2) non differentiable, and thus finding the optimal vector $\boldsymbol{s}$ intractable. Following [Yamada et al., 2020], one can relax the ℓ_0 norm to a probabilistic differentiable counterpart, by replacing the binary indicator vector $\boldsymbol{s}$ with a relaxed Bernoulli vector $\boldsymbol{z}$. Specifically, $\boldsymbol{z}$ is a continuous Gaussian reparametrization of the discrete random variables, termed Stochastic Gates. It is defined for each feature i:

$$z_i = \max(0, \min(1, 0.5 + \mu_i + \epsilon_i)), \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2) \tag{3}$$

where μ_i is a learnable parameter, and σ is fixed throughout training. The loss function in Eq. (2) can now be reformulated as follows, which is the final objective of the DUFS:

$$\mathcal{L} = -\frac{1}{n}\text{Tr}[\tilde{\boldsymbol{X}}^T \boldsymbol{L}_{\tilde{x}} \tilde{\boldsymbol{X}}] + \lambda \|\boldsymbol{z}\|_0. \tag{4}$$

3 METHOD

We now derive our approach for unsupervised feature selection in multi-modal settings. Our method is designed to capture two types of features: (i) Features associated with latent structures that are *shared* between two modalities. (ii) Features associated with *differential latent structures*, that appear in only one modality. In Sec. 3.1 and 3.2, we derive two operators designed to capture shared and differential structures, respectively. To motivate our approach and illustrate the difference between shared and differential structures, we specifically address two examples: (i) shared and differential clusters and (ii) product of manifolds. We use the proposed operators in Sec. 3.3 to derive mmDUFS.

3.1 THE SHARED STRUCTURE OPERATOR

To motivate our approach, let us consider the artificial example illustrated in Fig. 2. The lower figure in the left panel shows the observations in modality $\boldsymbol{Y}$, which contains samples from a mixture of three distinct Gaussians. The upper figure shows modality $\boldsymbol{X}$, where one of the three clusters is partitioned again into three (less distinct) clusters.

It is instructive to study the *ideal setting* where we make the following assumptions: (i) The largest distance between two nodes within a cluster, denoted d_{within} is much smaller than the smallest distance between pairs of nodes of two clusters, denoted d_{between}. (ii) The bandwidth σ_x, σ_y is chosen such that $d_{\text{within}} \ll \sigma_x, \sigma_y \ll d_{\text{between}}$. In this setting, the three Gaussians constitute three main clusters, with no connections between pairs of nodes of different clusters and similar weights between pairs of nodes within clusters. Thus, the leading eigenvectors of $\boldsymbol{L}_y$ span the subspace of the three *indicator vectors*. That is, vectors that contain the square root of the degree of a node in a cluster and a zero value outside the cluster. See Von Luxburg [2007] and illustration in Fig. 2. The matrix $\boldsymbol{L}_x$ has two extra significant eigenvectors that span the separation of the third cluster, which appears only in $\boldsymbol{X}$. We denote by $\boldsymbol{V}_s$ a matrix that contains the indicator vectors of the three partitions that appear in $\boldsymbol{X}$ and $\boldsymbol{Y}$ and by $\boldsymbol{V}_x$ a matrix that contains the partitions that appear only in $\boldsymbol{X}$. Since there is no modality-specific structure in modality $\boldsymbol{Y}$, in our ideal setting the two Laplacian matrices $\boldsymbol{L}_x, \boldsymbol{L}_y$ can be approximated by

$$\boldsymbol{L}_x \approx \boldsymbol{V}_s \boldsymbol{V}_s^T + \boldsymbol{V}_x \boldsymbol{V}_x^T, \qquad \boldsymbol{L}_y \approx \boldsymbol{V}_s \boldsymbol{V}_s^T. \tag{5}$$

To capture *shared* latent structures we compute the operator $\boldsymbol{P}_{\text{shared}}$,

$$\boldsymbol{P}_{\text{shared}} = \boldsymbol{L}_x \boldsymbol{L}_y + \boldsymbol{L}_y \boldsymbol{L}_x. \tag{6}$$

For the cluster setting, the orthogonality between the matrices $\boldsymbol{V}_s, \boldsymbol{V}_x$ implies $\boldsymbol{P}_{\text{shared}} \approx 2\boldsymbol{V}_s \boldsymbol{V}_s^T$. Thus, the symmetric product of the two Laplacians captures clusters that appear in both modalities while removing modality-specific clusters; see right panel of Fig. 2. We note that related multimodal operators were previously proposed [Lindenbaum et al., 2020, Shnitzer et al., 2019] for computing low-dimensional representations. Here, we combine our operator with DUFS to develop a multi-modal feature selection pipeline. We illustrate the usefulness of the shared operator for the product of manifold setting.

Product of manifolds. Let $\mathcal{M}_a, \mathcal{M}_b$ and $\mathcal{M}_s$ be three low-dimensional manifolds embedded in high dimensional spaces. Here, we assume that the surface of the three manifolds is a smooth transformations of three sets of latent variables, denoted respectively by $\boldsymbol{\theta}_a, \boldsymbol{\theta}_b$ and $\boldsymbol{\theta}_s$. Consider the case where modalities $\boldsymbol{X}$ and $\boldsymbol{Y}$ each contains observations from the products $\mathcal{M}_y, \mathcal{M}_x$,

$$\mathcal{M}_y = \mathcal{M}_s \times \mathcal{M}_a, \qquad \mathcal{M}_x = \mathcal{M}_s \times \mathcal{M}_b.$$

Note that the dependence on $\mathcal{M}_s$ is shared between $\mathcal{M}_x, \mathcal{M}_y$, while the dependence on $\mathcal{M}_a, \mathcal{M}_b$ is modality-specific. In a product $\mathcal{M}_x = \mathcal{M}_s \times \mathcal{M}_b$, every point $\boldsymbol{x} \in \mathcal{M}_x$ is associated with two points $\boldsymbol{x}_s \in \mathcal{M}_s$ and $\boldsymbol{x}_b \in \mathcal{M}_b$. We define the projection operators $\pi_b^x(\boldsymbol{x}), \pi_s^x(\boldsymbol{x})$ that map a point $\boldsymbol{x}$ in $\mathcal{M}_x$ to points in $\mathcal{M}_b, \mathcal{M}_s$, respectively. With the projection operators, one can extend a function $f^b : \mathcal{M}_b \to \mathbb{R}$ to a function over the product $f^x : \mathcal{M}_x \to \mathbb{R}$ by $f^x(\boldsymbol{x}) = f^b(\pi_b^x(\boldsymbol{x}))$.

An important property of a product $\mathcal{M}_x$ is that the eigenfunctions $f_{l,m}^x$ of the Laplace Beltrami operator are equal to the pointwise product of the eigenfunctions of $\mathcal{M}_b, \mathcal{M}_s$, extended to $\mathcal{M}_x$.

$$f_{l,m}^x(\boldsymbol{x}) = f_l^s(\pi_s^x(\boldsymbol{x})) \cdot f_m^b(\pi_b^x(\boldsymbol{x})). \tag{7}$$

We refer to [Zhang et al., 2021b] for a detailed description of products of manifold properties. A simple example of a product of manifolds is a 2D rectangle area $(\theta_s, \theta_b) \in [0, l_s] \times [0, l_b]$. the projection π_s^x yields the first coordinate, while π_b^x yields the second. The eigenfunctions of the product with Neumann boundary conditions are equal to,

$$f_{l,m}(\theta_s, \theta_b) = \cos(\pi l \theta_s / l_s) \cos(\pi m \theta_b / l_b). \tag{8}$$

Observations generated uniformly at random over the product of manifolds. Here, we assume that the observations in the two modalities are generated by random and independent uniformly distributed samples over $\mathcal{M}_x, \mathcal{M}_y$. Let $\boldsymbol{\phi}_{l,m}^x(\boldsymbol{x}_i), \boldsymbol{\phi}_{l,k}^y(\boldsymbol{y}_i)$ denote the eigenvectors of $\boldsymbol{L}_x, \boldsymbol{L}_y$ evaluated at $\boldsymbol{x}_i, \boldsymbol{y}_i$ respectively. In the asymptotic regime where the number of points $n \to \infty$, the eigenvectors converge to the eigenfunctions as characterized in Eq. (7).

$$\begin{aligned} \boldsymbol{\phi}_{l,m}^x(\boldsymbol{x}_i) &= \boldsymbol{\phi}_l^s(\pi_s^x(\boldsymbol{x}_i)) \boldsymbol{\phi}_m^b(\pi_b^x(\boldsymbol{x}_i)) \\ \boldsymbol{\phi}_{l,k}^y(\boldsymbol{y}_i) &= \boldsymbol{\phi}_l^s(\pi_s^y(\boldsymbol{y}_i)) \boldsymbol{\phi}_k^a(\pi_a^y(\boldsymbol{y}_i)). \end{aligned} \tag{9}$$

Details about the definition and rate of convergence can be found, for example, in [Cheng and Wu, 2022, García Trillos et al., 2020], and reference therein. It is instructive to consider the ideal case, where due to their dependence on the independent projections π_b^x and π_a^x, the eigenvectors $\boldsymbol{\phi}_{l,m}^x, \boldsymbol{\phi}_{l,k}^y$ satisfy the following orthogonality property,

$$(\boldsymbol{\phi}_{l,m}^x)^T \boldsymbol{\phi}_{l',k}^y = \begin{cases} 1 & l = l', m = k = 0 \\ 0 & o.w. \end{cases} \tag{10}$$

It follows that the operator $\boldsymbol{P}_{\text{shared}}$ is equal to,

$$\boldsymbol{P}_{\text{shared}} = \boldsymbol{L}_x \boldsymbol{L}_y + \boldsymbol{L}_y \boldsymbol{L}_x = \sum_l (\boldsymbol{\phi}_l^s \otimes \boldsymbol{\phi}_0^a)(\boldsymbol{\phi}_l^s \otimes \boldsymbol{\phi}_0^b)^T, \tag{11}$$

where $\otimes$ denotes the Hadamard product. The vectors $\boldsymbol{\phi}_0^a, \boldsymbol{\phi}_0^b$ constitute the degree of the different observations and have little effect on the outcome. Thus, the leading eigenvectors of $\boldsymbol{P}_{\text{shared}}$ are associated with the shared component and not the differential components in the product of manifolds. Below, we illustrate this phenomenon with two examples.

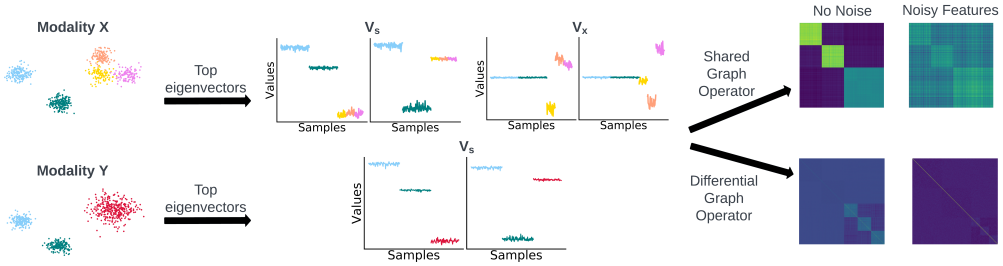


Figure 2: Visualization of the eigenvectors and the affinity matrix of the proposed operators on an artificial cluster example. Left: Visualization of the clusters. Middle: Leading eigenvectors of $\boldsymbol{L}_x$ and $\boldsymbol{L}_y$. Right: Affinity matrices of the proposed shared graph operator (top) and the differential graph operator (bottom) with/without the presence of noisy features.

Example 1: points in a 3D cube. Consider points generated uniformly at random over a 3D cube of dimensions $[0, l_s] \times [0, l_a] \times [0, l_b]$. Let $\boldsymbol{Y} \in \mathbb{R}^{n\times 2}$ constitute the first two coordinates of n independent observations, and let $\boldsymbol{X}$ constitute the first and third coordinates. This is a simple case of a product of manifolds, where the shared variable θ_s is the first coordinate, while the modality-specific variables θ_a, θ_b are the second and third coordinates. Following Eq. (8), the eigenvectors of the graph Laplacian matrices $\boldsymbol{L}_x, \boldsymbol{L}_y$, evaluated at (θ_s, θ_b) and (θ_s, θ_a) converge to,

$$\begin{aligned}\phi^x_{lm}(\theta_s, \theta_b) &= \cos(\pi l\theta_s/l_s)\cos(\pi m\theta_b/l_b)\\ \phi^y_{lk}(\theta_s, \theta_a) &= \cos(\pi l\theta_s/l_s)\cos(\pi k\theta_a/l_a).\end{aligned} \tag{12}$$

The first row of Fig. 1 (Appendix A) shows a scatter plot of the points in $\boldsymbol{X}$ (located according to the first two coordinates), colored by the values of the leading eigenvectors of $\boldsymbol{L}_x$. The second row shows the points in $\boldsymbol{X}$, but colored by the eigenvectors of $\boldsymbol{P}_{\text{shared}}$. As expected, all the eigenvectors of $\boldsymbol{P}_{\text{shared}}$ are functions of the shared coordinate θ_s.

Example 2: videos taken from different angles. Our second example is based on an experiment done in [Lederman and Talmon, 2014], where the two modalities constitute two videos of three dolls rotating at different angular speeds. The first camera (modality $\boldsymbol{X}$) captures the middle and left doll, while the second camera (modality $\boldsymbol{Y}$) captures the middle and right dolls (see Fig. 4a). Here, the shared variable $\boldsymbol{\theta}_s$ is the angle of the middle doll captured by both modalities. The modality-specific variables $\boldsymbol{\theta}_a, \boldsymbol{\theta}_b$ are the angles of the left and right dolls captured by each modality separately.

To illustrate Eq. (11) in this example, we first compute an approximation of the eigenvectors $\boldsymbol{\phi}^s_l$. To that end, we cropped each image in one of the videos such that only the middle doll (which appears in both modalities) is shown. One may think of this operation as a projection to the shared manifold. Next, we computed from the cropped images the leading eigenvectors $\boldsymbol{\phi}^s_l$ of the Laplacian matrix. Fig. 2 (Appendix A) shows the leading three eigenvectors of $\boldsymbol{P}_{\text{shared}}$ as a function of $\boldsymbol{\phi}^s_1, \boldsymbol{\phi}^s_2, \boldsymbol{\phi}^s_3$ as computed by the cropped images. The figure shows a linear dependency between the vectors, which implies that the shared operator retained only the shared component of the two modalities.

3.2 THE DIFFERENTIAL GRAPH OPERATORS

We design two operators $\boldsymbol{Q}_x$ and $\boldsymbol{Q}_y$ to infer latent structures that are *modality specific* to $\boldsymbol{X}, \boldsymbol{Y}$ respectively.

$$\boldsymbol{Q}_x = \tilde{\boldsymbol{L}}_y^{-1}\boldsymbol{L}_x\tilde{\boldsymbol{L}}_y^{-1}, \qquad \boldsymbol{Q}_y = \tilde{\boldsymbol{L}}_x^{-1}\boldsymbol{L}_y\tilde{\boldsymbol{L}}_x^{-1}, \tag{13}$$

where $\tilde{\boldsymbol{L}}_x = \boldsymbol{L}_x + c\boldsymbol{I}$, $\tilde{\boldsymbol{L}}_y = \boldsymbol{L}_y + c\boldsymbol{I}$, and c is a regularization constant. We address the cluster example used for the shared operator to motivate the use of these operators.

Differential clusters. In the synthetic cluster example in Fig. 2, modality $\boldsymbol{X}$ has three smaller clusters not observed in modality $\boldsymbol{Y}$. We show that one can detect the *differential clusters* of modality $\boldsymbol{X}$ via the leading eigenvectors of $\boldsymbol{Q}_x$. By Eq. (5), we can approximate $\tilde{\boldsymbol{L}}_y$ via,

$$\tilde{\boldsymbol{L}}_y = (1+c)\boldsymbol{V}_s\boldsymbol{V}_s^T + c\boldsymbol{V}_{\text{comp}}\boldsymbol{V}_{\text{comp}}^T, \tag{14}$$

where $\boldsymbol{V}_{\text{comp}} \in \mathbb{R}^{n\times(n-3)}$ contains, as columns, vectors that span the complementary subspace to $\boldsymbol{V}_s$. We write $\boldsymbol{Q}_x$ as:

$$\boldsymbol{Q}_x = \tilde{\boldsymbol{L}}_y^{-1}\boldsymbol{L}_x\tilde{\boldsymbol{L}}_y^{-1} = c^{-2}\boldsymbol{V}_x\boldsymbol{V}_x^T + (1+c)^{-2}\boldsymbol{V}_s\boldsymbol{V}_s^T. \tag{15}$$

The differential operator in Eq. (15) has two terms. The first spans the subspace corresponding to the differential structure $\boldsymbol{V}_x$, while the second spans the subspace of the shared structure $\boldsymbol{V}_s$. Since $c^{-2} > (1+c)^{-2}$, it follows that the leading eigenvectors of $\boldsymbol{Q}_x$ span the subspace of $\boldsymbol{V}_x$.

In theory, we can directly apply these operators to learn the structures. However, in many real-world applications, e.g., single-cell multi-omic technologies, both $\boldsymbol{X}$ and $\boldsymbol{Y}$ can be very noisy. In particular, abundant noisy features (e.g., genes) might dominate the data. The top eigenvectors of $\boldsymbol{L}_x$ and $\boldsymbol{L}_y$ might not capture the underlying structure, which would be detrimental to the learning of $\boldsymbol{P}_{\text{shared}}$, $\boldsymbol{Q}_x$, and

$\boldsymbol{Q}_y$. As shown in the affinity matrices on the right of Fig. 2, the structures are less clear when many noisy features are present. Therefore, it is necessary to have a feature selection framework that can effectively remove these noisy features in our multi-modal setting. With the aforementioned DUFS feature selection framework as the foundation, we will show in the next section how we can incorporate it into our proposed operators in the multi-modal setting.

3.3 MMDUFS

In this section, we describe our framework, termed multi-modal Differential Unsupervised Feature Selection (mmDUFS). We incorporates differentiable gates [Lindenbaum et al., 2021] with loss functions based on the shared and differential operators, detailed in Sec. 3.1 and 3.2. Our goal is to compute an accurate shared graph operator ($\boldsymbol{P}_{\text{shared}}$ in Eq. (6)) and differential graph operators ($\boldsymbol{Q}_x$ and $\boldsymbol{Q}_y$ in Eq. (13)) while simultaneously selecting the informative features. Let $\boldsymbol{f}_x, \boldsymbol{f}_y$ denote a feature vector in $\boldsymbol{X}, \boldsymbol{Y}$, respectively. To quantify how noisy or informative the features are with respect to the shared structure, we replace the Laplacian $\boldsymbol{L}$ in Eq. (1) with $\boldsymbol{P}_{\text{shared}}$, which yields the shared score $\boldsymbol{f}_x^T \boldsymbol{P}_{\text{shared}} \boldsymbol{f}_x$ and $\boldsymbol{f}_y^T \boldsymbol{P}_{\text{shared}} \boldsymbol{f}_y$. Similarly, $\boldsymbol{f}_x^T \boldsymbol{Q}_x \boldsymbol{f}_x$ and $\boldsymbol{f}_y^T \boldsymbol{Q}_y \boldsymbol{f}_y$ quantify the smoothness of these features with respect to the differential graph operators $\boldsymbol{Q}_x$ and $\boldsymbol{Q}_y$. The rationale behind these generalized Laplacian Scores is similar to the original score. For instance, let $\boldsymbol{P}_{\text{shared}} = \sum_{i=1}^{n} \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^T$ be the eigendecomposition of $\boldsymbol{P}_{\text{shared}}$. A feature vector $\boldsymbol{f}_x$ that varies slowly with respect to the underlying shared structure has a larger component within the subspace spanned by the leading eigenvectors of $\boldsymbol{P}_{\text{shared}}$, and thus a higher score.

To learn features with high generalized Laplacian Scores and accurate graph operators, mmDUFS learns two sets of Stochastic Gates $\boldsymbol{z}_x$ and $\boldsymbol{z}_y$ that filter irrelevant features in each modality. Similar to DUFS [Lindenbaum et al., 2021], these stochastic gates multiply the data matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$ to remove nuisance features, i.e., $\tilde{\boldsymbol{X}} = \boldsymbol{X}\Delta(\boldsymbol{z}_x)$ and $\tilde{\boldsymbol{Y}} = \boldsymbol{Y}\Delta(\boldsymbol{z}_y)$. At each iteration, the updated graph operators ($\tilde{\boldsymbol{P}}_{\text{shared}}$, $\tilde{\boldsymbol{Q}}_x$, $\tilde{\boldsymbol{Q}}_y$) are recomputed based on the gated inputs.

mmDUFS has two modes: (i) detecting shared structures using the shared graph operator $\tilde{\boldsymbol{P}}_{\text{shared}}$, and (ii) detecting modality-specific structures using the differential graph operators $\tilde{\boldsymbol{Q}}_x$, and $\tilde{\boldsymbol{Q}}_y$. To learn the shared structure and the corresponding features, we propose to optimize $\boldsymbol{z}_x$ and $\boldsymbol{z}_y$ by minimizing the following loss function:

$$\mathcal{L}_{\text{shared}} = -\frac{1}{n}\text{Tr}[\tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{P}}_{\text{shared}} \tilde{\boldsymbol{X}}] - \frac{1}{n}\text{Tr}[\tilde{\boldsymbol{Y}}^T \tilde{\boldsymbol{P}}_{\text{shared}} \tilde{\boldsymbol{Y}}] + \lambda_x \|\boldsymbol{z}_x\|_0 + \lambda_y \|\boldsymbol{z}_y\|_0,$$

where the first two terms are the Shared Laplacian Scores for each modality, and the regularizers $\lambda_x \|\boldsymbol{z}_x\|_0$ and $\lambda_y \|\boldsymbol{z}_y\|_0$ control the number of selected features for each modality. The parameters λ_x, λ_y are tunable, and determine the sparsity level. In Appendix B.1, we suggest a procedure to set these regularization parameters. To learn differential structures that appear in only one modality, we suggest the loss functions $\mathcal{L}_x, \mathcal{L}_y$,

$$\begin{aligned} \mathcal{L}_x &= -\frac{1}{n}\text{Tr}[\tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{Q}}_x \tilde{\boldsymbol{X}}] + \lambda_x \|\boldsymbol{z}_x\|_0, \\ \mathcal{L}_y &= -\frac{1}{n}\text{Tr}[\tilde{\boldsymbol{Y}}^T \tilde{\boldsymbol{Q}}_y \tilde{\boldsymbol{Y}}] + \lambda_y \|\boldsymbol{z}_y\|_0. \end{aligned} \tag{16}$$

The first term in each loss is the Differential Laplacian Score. Optimizing over Eq. (16) yields a set of features that are smooth with respect to the differential graph operators $\boldsymbol{Q}_x$ and $\boldsymbol{Q}_y$, with a sparsity level controlled by λ_x and λ_y. In Section 5 we show the usefulness of these score functions for detecting relevant features.

4 RELATED WORK

Learning the latent structures in multi-modal data has been studied extensively in the context of data fusion, where most existing methods aim to extract shared information from the modalities [Andrew et al., 2013, Lederman and Talmon, 2014, Zhou and Burges, 2007, Lindenbaum et al., 2015]. Only a few methods study the differences between modalities [Shnitzer et al., 2019]. However, these multi-modal learning methods become unsuitable when many nuisance or noisy features are present in the data. In [Cohen et al., 2022, Sristi et al., 2022], the authors use the manifold assumption to tackle feature selection and clustering in the supervised setting. In the unsupervised setting, several authors propose different Unsupervised Feature Selection (UFS) schemes to alleviate the influence of nuisance features. These methods aim to identify a subset of smooth features with respect to the underlying structure [Zhao and Liu, 2012, Lindenbaum et al., 2021, Shaham et al., 2022]. However, they focus on a single modality and are not applicable to multi-modal data.

5 RESULTS

We benchmark mmDUFS [1] using synthetic and real multi-modal datasets. For discovering the shared structures and associated features, we compare mmDUFS with the shared operator to the following variants of kernel fusion-based methods previously proposed for dimensionality reduction: (1) Matrix Concatenation (MC), where the Laplacian is computed based on a concatenated matrix of the two modalities. (2) Multi-modal Kernel Sum (mmKS) [Zhou and Burges, 2007], where the Laplacian is equal to $\boldsymbol{L}_x + \boldsymbol{L}_y$. (3) Multi-modal Kernel Product (mmKP) [Lindenbaum et al., 2015, 2016, 2020]. where the Laplacian is equal to $\boldsymbol{L}_x \boldsymbol{L}_y$.

To compare to the performance of mmDUFS on detecting differential features, we extended MC,mmKS and mmKP

[1] Codes are available at https://github.com/jcyang34/mmDUFS

by the following steps: (i) compute sets S_x, S_y of features that are smooth, separately, with respect to L_x and L_y via standard Laplacian scores. The selected features contain both shared and modality specific. (ii) Apply either MC, mmKS or mmKP to compute a set S_{xy} of shared features. (iii) remove the shared features S_{xy} from the sets S_x, S_y to obtain the features that are modality specific to $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively.

For each baseline, the k features with the highest Laplacian Scores are selected. For the synthetic datasets, we set k to be the correct number of informative features. We evaluate the performance of different methods by the F1-score $\text{F1} = \text{TP}/(\text{TP} + \frac{1}{2(\text{FP}+\text{FN})})$, where TP is the number of informative features selected by each method, FP is the number of uninformative selected features, and FN is the number of missed informative features. For the rescaled MNIST and rotating doll examples, the informative features are set to the 25% pixels with the highest standard deviation.

5.1 SYNTHETIC EXAMPLES

Rescaled MNIST. We designed a rescaled MNIST example with shared and modality-specific digits. We first randomly sample one image (28×28 pixels) of digits 0, 3, 8. Then, we rescale each digit randomly and independently 500 times resulting with 500 images of 0, 3, and 8. We concatenate pairs of 0 and 3 to create modality $\boldsymbol{X}$, and pairs of the same 3 and random 8 to create $\boldsymbol{Y}$, see example in Fig. 3a. Thus, this dataset consists of 500 samples and 28×56 pixels in each modality, with digit 3 shared between the modalities and digit 0 and 8 modality specific.

We apply mmDUFS with the shared operator to this example to select pixels corresponding to 3. The left column of Fig. 3b shows the pixels gate values from mmDUFS for modality $\boldsymbol{X}$ (top) and $\boldsymbol{Y}$ (bottom). We can see that selected pixels outline the shape of the digit 3 well. Table 1 compares the F1-score achieved by mmDUFS to three baselines. We can see that mmDUFS achieves a higher F1-score than all the baselines on both modalities, demonstrating its ability to identify informative features accurately.

Next, we apply mmDUFS with the differential operator to select modality-specific pixels. The right column of Fig. 3b shows the pixel gate values for both modality $\boldsymbol{X}$ (top) and $\boldsymbol{Y}$ (bottom). We can see that mmDUFS selects pixels that outline digits $0, 8$ for modalities $\boldsymbol{X}, \boldsymbol{Y}$, respectively. Additionally, mmDUFS achieves F1-score 0.8059 and 0.8832 for $\boldsymbol{X}$ and $\boldsymbol{Y}$, showcasing its effectiveness in identifying features contributing to the differential structures.

Lastly, we demonstrate that our model can be extended and applied to scenarios where there are more than 2 modalities. We extend this rescaled MNIST dataset by adding another modality ($\boldsymbol{Z}$), which contains 500 concatenated images of rescaled digits 3 and 4. Therefore, digit 3 is shared across all three modalities. We apply mmDUFS with this extended shared operator on this dataset to select pixels corresponding to 3. In Supplementary Table 1, we can see that mmDUFS outperforms all the baselines in terms of the F1-score, demonstrating its ability to accurately identify informative features in multimodal scenarios.

Dataset	Modality	MC	mmKS	mmKP	mmDUFS
Rescaled MNIST	X	0.3547	0.5291	0.5291	**0.7093**
	Y	0.4826	0.6219	0.6219	**0.8159**
Synthetic Developmental Tree	X	0.6000	0.7800	0.8400	**0.8800**
	Y	0.7800	0.8000	0.8200	**0.9000**
Original Gaussian	X	0.5000	0.7333	**1**	**1**
	Y	0.5500	0.6500	0.9500	**1**
Gaussian + 10 Noisy Feats	X	0.5000	0.7333	**1**	**1**
	Y	0.5000	0.6500	0.9000	**1**
Gaussian + 30 Noisy Feats	X	0.4667	0.7000	0.9667	**1**
	Y	0.4500	0.5500	0.8500	**1**
Gaussian + 50 Noisy Feats	X	0.4000	0.6333	0.9333	**0.9667**
	Y	0.4000	0.5500	0.8000	**0.8500**

Table 1: Comparing F1-score of the features associated with the shared structures between different methods on the rescaled MNIST example, the synthetic tree example, and the Gaussian mixture example with different numbers of additive noisy features.

Synthetic Developmental Tree. Tree structures are ubiquitous throughout different biological processes and data modalities in single-cell biology [Plass et al., 2018, Zhang et al., 2021a]. To understand the interplay of different mechanisms underlying the complex developmental process, it is vital to discover the genetic features that contribute to the tree structure shared across modalities and those that contribute to modality-specific structures.

We evaluate mmDUFS using a simulated developmental tree example generated via a tree simulator [2]. The original data has 1000 samples and 100 features. We divide the data into half, such that each modality has 50 informative features that contribute to the shared tree structure, as shown in the UMAP embeddings in Fig. 3c, where the samples in the tree are grouped into different branch groups (labeled G_1 to G_6). We then add 50 features drawn from negative binomial distributions to each modality to create differential branches, that are only observed in one modality. Specifically, branches G_1 and G_2 are bifurcated in modality $\boldsymbol{X}$ (top UMAP embeddings) but are mixed in modality $\boldsymbol{Y}$ (bottom UMAP embeddings), and G_3 and G_4 are bifurcated in modality $\boldsymbol{Y}$ but are mixed in modality $\boldsymbol{X}$ (see Supplementary section B.3 for further details). After log transformation and z-scoring the data, we concatenate 200 features drawn from $N(0, 1)$ to each modality as noisy features.

We apply our model with the shared and differential operators to recover the features that contribute to the overall tree structure and the set of features that contribute to the split branches, respectively. Fig. 3d shows the change, during training with the shared loss, in the Shared/Differential

[2] https://github.com/dynverse/dyntoy

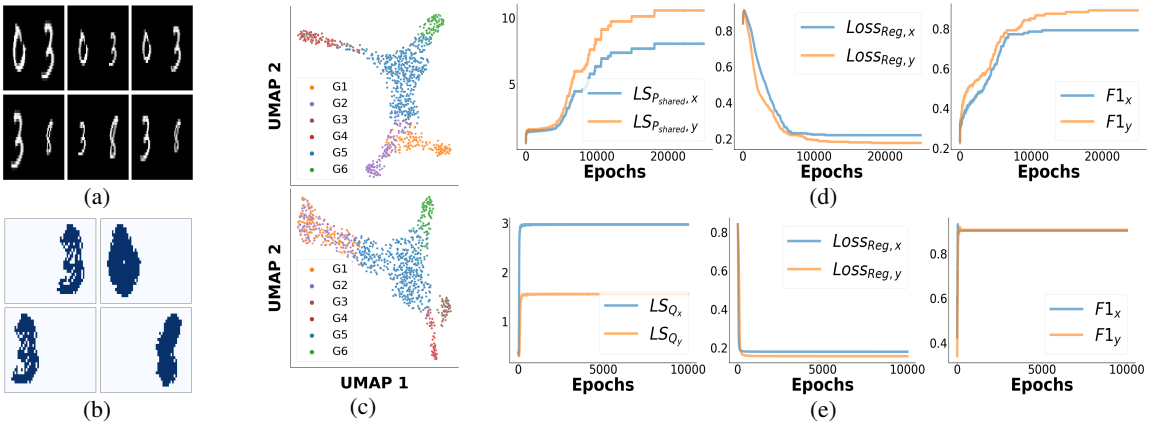


Figure 3: Left (a-b): Evaluation of the proposed approach on the rescaled MNIST dataset. (a): Random images from modality X (upper row) and modality Y (bottom row) in gray-scale. (b): Selected pixels (dark blue) for the shared operator (left column) and the differential operator (right column). Right (c-e): Synthetic developmental tree example. (c): UMAP embeddings of the tree using data from modality $\boldsymbol{X}$ (top) and modality $\boldsymbol{Y}$ (bottom). (d-e): Change of the Shared/Differential Laplacian Scores, regularization loss, and the F1-score of the selected features concerning the number of epochs (x-axis) for mmDUFS with the shared operator (panel (c)) and the differential operator (panel (e)).

Laplacian Scores, the regularization loss, and the F1-score. Fig. 3e shows the same properties for the differential loss. Table 1 compares the F1-score of the selected features between different methods. Here as well, mmDUFS clearly outperforms the other methods.

Synthetic Gaussian Mixtures. We generated a multimodal Gaussian mixture dataset, where $\boldsymbol{X}$ and $\boldsymbol{Y}$ each have three clusters. Two clusters are shared between modalities, and one cluster is specific to each modality. The observations in each modality include features informative of the clusters, along with noisy features (see Appendix B.2).

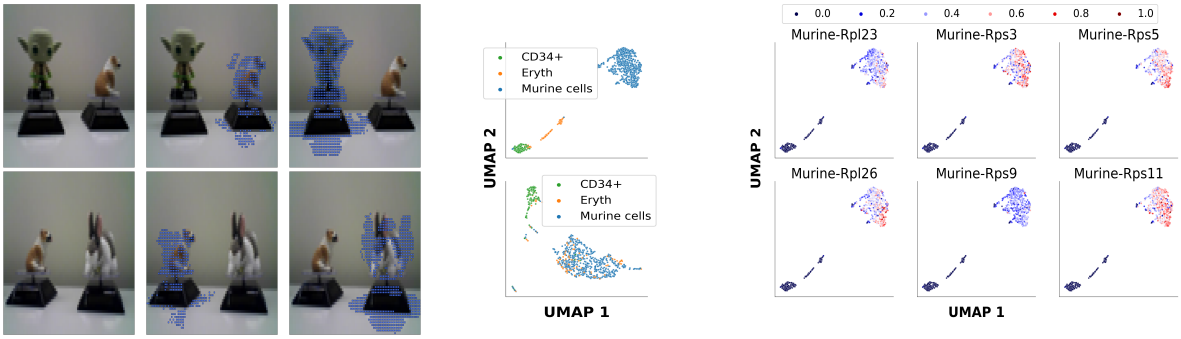
We apply mmDUFS to uncover the informative features of the shared clusters and the modality-specific clusters. In Fig. 3 of Supplementary section B.2, we plot the change of the average shared/differential Laplacian Scores across features, the regularization loss, and the F1-score of the selected features from mmDUFS with respect to the number of epochs. MmDUFS gradually selects the correct features while dropping the non-informative ones. To evaluate mmDUFS's feature selection capability in challenging regimes, we inject 10, 30, and 50 noisy features into each modality and compare the F1-score of features selected by different methods in each regime. Table 1 shows that mmDUFS consistently outperforms the baseline methods, and maintains its accuracy even in challenging regimes.

5.2 REAL DATA

Rotating Dolls. We evaluate mmDUFS's performance on the rotating doll video dataset described in Sec. 3.1 in which 2 cameras capture 2 dolls from different angles (Fig. 4a). By treating each video frame as one sample (4050 in total) and the gray-scaled pixels as features, we aim to uncover pixels that correspond to the shared doll (the dog) and the modality-specific dolls (Yoda and rabbit).

For mmDUFS with the shared operator, Fig. 4b shows selected pixels in both videos, as indicated by the blue dots. The shape of the dog is clearly delineated in both modalities. We further compute the F1-score of the selected pixels with respect to the underlying pixels that correspond to the dog. mmDUFS achieves F1-score of 0.7158 and 0.8033 for the two modalities, whereas MC achieves 0.2390 and 0.3822, and mmKS and mmKP achieve 0.5452 and 0.6868. Fig. 4c shows the selected pixels of mmDUFS with the differential operator in the two videos. In video 1, mmDUFS select mostly pixels corresponding to the Yoda (F1-score: 0.8861). For video 2, mmDUFS select mostly pixels corresponding to the rabbit (F1-score: 0.7446).

To demonstrate that our model can extract useful information from high-dimensional measurements, we use the selected features to estimate of rotation angles of the shared doll (the dog). For computing the *ground truth* rotation angles, we first compute the top 25% pixels with the highest standard deviation. Then, we keep only the features that belong to the dog, and compute the angle via the leading two Laplacian eigenvectors computed based on these pixels. Next, we compute the estimated angles that are based on features detected by mmDUFS and the other baseline methods. For comparison, we compute the mean squared error between the estimated angles and the ground truth, as shown below in Table 2. We can see that mmDUFS outperforms other methods, which shows that it can improve the capability of extracting latent information in multimodal data in unsupervised settings.



(a) (b) (c) (d) (e)

Figure 4: Left (a-c): Rotating dolls example. (a): Random images of the dolls from each video. (b-c): Selected pixels are marked in blue for mmDUFS with shared operator (b) and the differential operator (c). Right (d-e): CITE-seq data example. (d): UMAP embeddings using the RNA (top) and protein data (bottom), colored by cell type labels. (e): Similar UMAP embeddings colored by the expression level of several genes selected by mmDUFS with the differential operator.

Modality	MC	mmKS	mmKP	mmDUFS
X	0.4559	0.1146	0.1146	**0.0150**
Y	0.5353	0.0509	0.0509	**0.0426**

Table 2: MSE of the estimated doll rotation angles

CITE-seq Dataset. In single-cell biology, cell states are characterized by different features at different molecular levels. Identifying the contributing features is an open question crucial to understanding the underlying cell systems. We apply mmDUFS to a human cord blood mononuclear cells (CBMCs) CITE-seq dataset from [Stoeckius et al., 2017], in which cells are profiled at both transcriptomic and proteomic levels measuring expressions of genes and protein markers, to identify the genes and proteins that characterize the cell states in the multi-modal setting.

In this data, a group of murine cells is spiked-in as controls. Fig. 4d shows UMAP embeddings of the cells based on their RNA expression (top) and protein expression (bottom). From the full dataset, we analyzed 3 cell populations: murine cells (blue) and 2 CBMCs cell populations (Erythroids (orange) and CD34+ cells (green)). This dataset has 832 cells, with 500 top variable genes from modality 1 and 10 protein markers from modality 2. We can see that the murine cells are separable from the Erythroids in the RNA space but not in the proteomic space. We apply mmDUFS with the differential operator to this data to identify which gene markers contribute to the separation between cell groups.

To evaluate the quality of each set of selected features, we used each set to train an SVM model to classify Erythroids and murine cells (i.e., the differential structures). With an 5% / 95% training/test split, MC/mmKS/mmKP achieve 96.97% / 93.80% / 93.80% average balanced test accuracy, respectively, whereas mmDUFS achieve 97.52% average balanced test accuracy (repeated 10 times). Examining the selected genes by each model, we found that mmDUFS mostly selects murine genes. These murine genes are exclusively expressed in murine cells, as shown in Fig. 4e, thus we expect these genes can better separate the two cell types. In summary, this result shows that mmDUFS can better preserve modality-specific structure (two separable cell types) and the informative features that are relevant to the structure in single-cell multi-omic data.

6 DISCUSSION

We present mmDUFS, a feature selection method that learns two novel graph operators that capture the *shared* and the *modality-specific* structures in multi-modal data, while simultaneously selecting the features that are informative for these structures. MmDUFS can operate on small batches which makes it scalable to large datasets. On the other hand, finding the optimal regularization parameters for mmDUFS on real data may be challenging, for which we suggest an automatic procedure in Appendix B.1. A second potential limitation is the $\mathcal{O}(n^3)$ computational complexity required to compute $\tilde{\boldsymbol{L}}$ (Eq. (13)). A possible solution is to reduce the complexity by computing a sparse Laplacian matrix.

Acknowledgement

The authors thank Amit Moscovich for helpful discussions and feedback. Y.K. acknowledges support by grant R01GM131642, UM1PA051410, R33DA047037, U54AG076043, U54AG079759, U01DA053628, P50CA121974, R01GM135928

References

Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *Inter-*

national Conference on Machine Learning, pages 1247–1255, 2013.

Muhammed Fatih Balın, Abubakar Abid, and James Zou. Concrete autoencoders: Differentiable feature selection and reconstruction. In *International conference on machine learning*, pages 444–453. PMLR, 2019.

Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

Zhiwen Chen, Steven X Ding, Tao Peng, Chunhua Yang, and Weihua Gui. Fault detection for non-gaussian processes using generalized canonical correlation analysis and randomized algorithms. *IEEE Transactions on Industrial Electronics*, 65(2):1559–1567, 2017.

Xiuyuan Cheng and Nan Wu. Eigen-convergence of gaussian kernelized graph laplacian by manifold heat interpolation. *Applied and Computational Harmonic Analysis*, 61:132–190, 2022.

David Cohen, Tal Shnitzer, Yuval Kluger, and Ronen Talmon. Manifest: Manifold-based feature selection for small data sets. *arXiv preprint arXiv:2207.08574*, 2022.

Alexandra Degeest, Michel Verleysen, and Benoît Frénay. Smoothness bias in relevance estimators for feature selection in regression. In *Artificial Intelligence Applications and Innovations: 14th IFIP WG 12.5 International Conference, AIAI 2018, Rhodes, Greece, May 25–27, 2018, Proceedings 14*, pages 285–294. Springer, 2018.

Nicolás García Trillos, Moritz Gerlach, Matthias Hein, and Dejan Slepčev. Error estimates for spectral convergence of the graph laplacian on random geometric graphs toward the laplace–beltrami operator. *Foundations of Computational Mathematics*, 20(4):827–887, 2020.

Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Audioclip: Extending clip to image, text and audio. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 976–980. IEEE, 2022.

Xiaofei He, Deng Cai, and Partha Niyogi. Laplacian score for feature selection. *Advances in neural information processing systems*, 18, 2005.

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

Julia Joung, Sai Ma, Tristan Tay, Kathryn R Geiger-Schuller, Paul C Kirchgatterer, Vanessa K Verdine, Baolin Guo, Mario A Arias-Garcia, William E Allen, Ankita Singh, et al. A transcription factor atlas of directed differentiation. *Cell*, 186(1):209–229, 2023.

Noemie Leblay, Ranjan Maity, Elie Barakat, Sylvia McCulloch, Peter Duggan, Victor Jimenez-Zepeda, Nizar J Bahlis, and Paola Neri. Cite-seq profiling of t cells in multiple myeloma patients undergoing bcma targeting car-t or bites immunotherapy. *Blood*, 136:11–12, 2020.

Roy R Lederman and Ronen Talmon. Common manifold learning using alternating-diffusion. *submitted, Tech. Report YALEUIDCSITR1497*, 2014.

Ofir Lindenbaum, Arie Yeredor, and Moshe Salhov. Learning coupled embedding using multiview diffusion maps. In *Latent Variable Analysis and Signal Separation: 12th International Conference, LVA/ICA 2015, Liberec, Czech Republic, August 25-28, 2015, Proceedings 12*, pages 127–134. Springer, 2015.

Ofir Lindenbaum, Neta Rabin, Yuri Bregman, and Amir Averbuch. Multi-channel fusion for seismic event detection and classification. In *2016 IEEE International Conference on the Science of Electrical Engineering (ICSEE)*, pages 1–5. IEEE, 2016.

Ofir Lindenbaum, Arie Yeredor, Moshe Salhov, and Amir Averbuch. Multi-view diffusion maps. *Information Fusion*, 55:127–149, 2020.

Ofir Lindenbaum, Uri Shaham, Erez Peterfreund, Jonathan Svirsky, Nicolas Casey, and Yuval Kluger. Differentiable unsupervised feature selection based on a gated laplacian. *Advances in Neural Information Processing Systems*, 34, 2021.

Ofir Lindenbaum, Moshe Salhov, Amir Averbuch, and Yuval Kluger. L0-sparse canonical correlation analysis. In *International Conference on Learning Representations*, 2022.

George C Linderman, Manas Rachh, Jeremy G Hoskins, Stefan Steinerberger, and Yuval Kluger. Fast interpolation-based t-sne for improved visualization of single-cell rna-seq data. *Nature methods*, 16(3):243–245, 2019.

Yang Liu, Mingyu Yang, Yanxiang Deng, Graham Su, Archibald Enninful, Cindy C Guo, Toma Tebaldi, Di Zhang, Dongjoo Kim, Zhiliang Bai, et al. High-spatial-resolution multi-omics sequencing via deterministic barcoding in tissue. *Cell*, 183(6):1665–1681, 2020.

Sai Ma, Bing Zhang, Lindsay M LaFave, Andrew S Earl, Zachary Chiang, Yan Hu, Jiarui Ding, Alison Brack, Vinay K Kartha, Tristan Tay, et al. Chromatin potential identified by shared single-cell profiling of rna and chromatin. *Cell*, 183(4):1103–1116, 2020.

Erez Peterfreund, Ofir Lindenbaum, Felix Dietrich, Tom Bertalan, Matan Gavish, Ioannis G Kevrekidis, and Ronald R Coifman. Local conformal autoencoder for standardized data coordinates. *Proceedings of the National Academy of Sciences*, 117(49):30918–30927, 2020.

Harold Pimentel, Zhiyue Hu, and Haiyan Huang. Biclustering by sparse canonical correlation analysis. *Quantitative Biology*, 6(1):56–67, 2018.

Mireya Plass, Jordi Solana, F Alexander Wolf, Salah Ayoub, Aristotelis Misios, Petar Glažar, Benedikt Obermayer, Fabian J Theis, Christine Kocks, and Nikolaus Rajewsky. Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. *Science*, 360(6391): eaaq1723, 2018.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

Tommi Raij, Kimmo Uutela, and Riitta Hari. Audiovisual integration of letters in the human brain. *Neuron*, 28(2): 617–625, 2000.

Uri Shaham, Ofir Lindenbaum, Jonathan Svirsky, and Yuval Kluger. Deep unsupervised feature selection by discarding nuisance and correlated features. *Neural Networks*, 152:34–43, 2022.

Tal Shnitzer, Mirela Ben-Chen, Leonidas Guibas, Ronen Talmon, and Hau-Tieng Wu. Recovering hidden components in multimodal data with composite diffusion operators. *SIAM Journal on Mathematics of Data Science*, 1(3): 588–616, 2019.

Saúl Solorio-Fernández, J Ariel Carrasco-Ochoa, and José Fco Martínez-Trinidad. A review of unsupervised feature selection methods. *Artificial Intelligence Review*, 53(2):907–948, 2020.

Ram Dyuthi Sristi, Gal Mishne, and Ariel Jaffe. Disc: Differential spectral clustering of features. *arXiv preprint arXiv:2211.05314*, 2022.

Marlon Stoeckius, Christoph Hafemeister, William Stephenson, Brian Houck-Loomis, Pratip K Chattopadhyay, Harold Swerdlow, Rahul Satija, and Peter Smibert. Simultaneous epitope and transcriptome measurement in single cells. *Nature methods*, 14(9):865–868, 2017.

Shiliang Sun. A survey of multi-view machine learning. *Neural computing and applications*, 23:2031–2038, 2013.

Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

Yang Xiao, Graham Su, Yang Liu, Cheick A Sissoko, Yung-yu Huang, Adrienne N Santiago, Andrew J Dwork, Gorazd B Rosoklija, Underwood D Mark, Victoria Arango, et al. Spatially resolved transcriptomes in human hippocampus. *Biological Psychiatry*, 91(9):S18, 2022.

Yutaro Yamada, Ofir Lindenbaum, Sahand Negahban, and Yuval Kluger. Feature selection using stochastic gates. In *International Conference on Machine Learning*, pages 10648–10659. PMLR, 2020.

Xiaoqiang Yan, Shizhe Hu, Yiqiao Mao, Yangdong Ye, and Hui Yu. Deep multi-view learning methods: A review. *Neurocomputing*, 448:106–129, 2021.

Kai Zhang, James D Hocker, Michael Miller, Xiaomeng Hou, Joshua Chiou, Olivier B Poirion, Yunjiang Qiu, Yang E Li, Kyle J Gaulton, Allen Wang, et al. A single-cell atlas of chromatin accessibility in the human genome. *Cell*, 184(24):5985–6001, 2021a.

Sharon Zhang, Amit Moscovich, and Amit Singer. Product manifold learning. In *International Conference on Artificial Intelligence and Statistics*, pages 3241–3249. PMLR, 2021b.

Zheng Alan Zhao and Huan Liu. *Spectral feature selection for data mining*. Taylor & Francis, 2012.

Dengyong Zhou and Christopher JC Burges. Spectral clustering and transductive learning with multiple views. In *Proceedings of the 24th international conference on Machine learning*, pages 1159–1166, 2007.