frontiers | Frontiers in Neuroinformatics

TYPE Methods
PUBLISHED 14 December 2023
DOI 10.3389/fninf.2023.1265079

Check for updates

OPEN ACCESS

EDITED BY
Georgios N. Dimitrakopoulos,
Ionian University, Greece

REVIEWED BY
Aris Vrahatis,
Ionian University, Greece
Ioannis Kakkos,
National Technical University of Athens, Greece

*CORRESPONDENCE
Mu Qiao
✉ muqiao0626@gmail.com

†PRESENT ADDRESS
Mu Qiao,
LinkedIn, Mountain View, CA, United States

RECEIVED 21 July 2023
ACCEPTED 06 November 2023
PUBLISHED 14 December 2023

CITATION
Qiao M (2023) Factorized discriminant analysis for genetic signatures of neuronal phenotypes. *Front. Neuroinform.* 17:1265079.
doi: 10.3389/fninf.2023.1265079

Factorized discriminant analysis for genetic signatures of neuronal phenotypes

Mu Qiao*†

Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA, United States

Navigating the complex landscape of single-cell transcriptomic data presents significant challenges. Central to this challenge is the identification of a meaningful representation of high-dimensional gene expression patterns that sheds light on the structural and functional properties of cell types. Pursuing model interpretability and computational simplicity, we often look for a linear transformation of the original data that aligns with key phenotypic features of cells. In response to this need, we introduce factorized linear discriminant analysis (FLDA), a novel method for linear dimensionality reduction. The crux of FLDA lies in identifying a linear function of gene expression levels that is highly correlated with one phenotypic feature while minimizing the influence of others. To augment this method, we integrate it with a sparsity-based regularization algorithm. This integration is crucial as it selects a subset of genes pivotal to a specific phenotypic feature or a combination thereof. To illustrate the effectiveness of FLDA, we apply it to transcriptomic datasets from neurons in the Drosophila optic lobe. We demonstrate that FLDA not only captures the inherent structural patterns aligned with phenotypic features but also uncovers key genes associated with each phenotype.

KEYWORDS

single-cell transcriptomics, linear discriminant analysis, dimensionality reduction algorithm, generalized eigenvalue problem, pheno- and geno-typing

1 Introduction

The analysis of gene expression data in single cells presents an intriguing and complex problem. Each cell's gene expression data can be viewed as a high-dimensional vector, allowing each cell to be represented as a single point in the vast space of gene expression. Clusters form within this space, each identifiable and associated with a particular cell type, thanks to the verification from the molecular markers of cell types (Macosko et al., 2015; Shekhar et al., 2016; Tasic et al., 2016, 2018; Peng et al., 2019).

When the phenotypic traits of each cell type are known, either from past studies or direct measurement (Sanes and Masland, 2015; Cadwell et al., 2016; Zeng and Sanes, 2017; Strell et al., 2019), we can label each cell type according to its unique characteristics. For example, differentiation of neuronal cell types could be achieved through analyzing a variety of features, such as dendritic and axonal laminations, electrophysiological properties, and connectivity (Sanes and Masland, 2015; Zeng and Sanes, 2017; Gouwens et al., 2019). These features are often categorical in nature.

A critical challenge arises when we attempt to factorize the high-dimensional gene expression data into modules that align with these phenotypes. In simple terms, we aim to find a low-dimensional embedding of gene expression where each axis signifies a single factor. This factor might correspond to a specific phenotypic feature or potentially, the combination of several.

Ideally, variation along one axis in the embedding space would exclusively affect one phenotypic feature. However, due to inevitable noise in the data, this is challenging to achieve. As a workaround, we allow for data projected along one axis to vary primarily with one phenotypic feature and minimally with others. Simultaneously, we want to preserve cell type identities in the low-dimensional space. This means that cells of the same type should remain in close proximity within the embedding space, while cells of different types remain distinct.

In order to address this issue, we propose the method of factorized linear discriminant analysis (FLDA). This is a supervised dimensionality reduction technique, rooted in the concepts of multi-way analysis of variance (ANOVA; Fisher, 1918). FLDA enables the factorization of data into components that correspond to phenotypic features and their combinations. It then seeks a linear transformation that is highly variable with one component, yet stable with others. The power of this approach lies in its simplicity and interpretability. To further leverage our analysis, we introduce a sparse variant of this method. This variant restricts the number of non-zero elements contributing to each linear projection, thereby identifying a subset of genes crucial to each phenotype. The efficacy of FLDA is demonstrated through its application to a single-cell RNA-Seq dataset of T4/T5 neurons in Drosophila (Kurmangaliyev et al., 2019), focusing particularly on two phenotypes: dendritic location and axonal lamination.

2 Factorized linear discriminant analysis (FLDA)

Let's consider a situation where each cell type can be characterized by two phenotypic features, both of which are categorical. This essentially means that the sample space for cell types is a Cartesian product of the sample spaces of the two phenotypic features I and J:

$$I \times J = \{(i,j) | i \in I, j \in J\} \tag{1}$$

In this equation, i, j represent different categories of the two phenotypic features. Suppose we have observed n_{ij} cells for each cell type (i,j). This information can be visualized with a contingency table, as shown in Figures 1A, B. Note here we account for the scenario where the table might be only partially filled.

We denote the expression values of g genes measured in the kth cell of the cell type (i,j) as $\boldsymbol{x}_{ijk}(k \in 1, 2, ...n_{ij})$ $(\boldsymbol{x}_{ijk} \in \mathbf{R}^g)$. Our task is to find linear projections $y_{ijk} = \boldsymbol{u}^T\boldsymbol{x}_{ijk}$ $(\boldsymbol{u} \in \mathbf{R}^g)$ and $z_{ijk} = \boldsymbol{v}^T\boldsymbol{x}_{ijk}$ $(\boldsymbol{v} \in \mathbf{R}^g)$ that align with features i and j, respectively (see Figure 1C).

To address this, we explored whether we could factorize, for example, y_{ijk}, into components dependent on features i and j. By employing the principles of linear factor models from multi-way ANOVA and the concept of variance partitioning, we formulated an objective function to find $\boldsymbol{u}$ that maximizes this objective (for a detailed analysis, refer to Appendix A).

$$\boldsymbol{u}^* = \underset{\boldsymbol{u} \in \mathbf{R}^g}{\arg\max} \frac{\boldsymbol{u}^T\boldsymbol{N}_A\boldsymbol{u}}{\boldsymbol{u}^T\boldsymbol{M}_e\boldsymbol{u}} \tag{2}$$

With a complete table, where a and b are the number of categories for feature i and j, we have:

$$\boldsymbol{N}_A = \boldsymbol{M}_A - \lambda_1\boldsymbol{M}_B - \lambda_2\boldsymbol{M}_{AB} \tag{3}$$

Here, $\boldsymbol{M}_A$, $\boldsymbol{M}_B$, and $\boldsymbol{M}_{AB}$ denote the covariance matrices explained by feature i, feature j, and their combination, respectively. The hyper-parameters λ_1 and λ_2 determine the relative weights of $\boldsymbol{M}_B$ and $\boldsymbol{M}_{AB}$ in comparison to $\boldsymbol{M}_A$. The residual covariance matrix, $\boldsymbol{M}_e$, represents variance within cell type clusters and signifies noise in gene expressions. The formal definitions of these terms are as follows:

$$\boldsymbol{M}_A = \frac{1}{a-1}\sum_{i=1}^{a}(\boldsymbol{m}_{i.} - \boldsymbol{m}_{..})(\boldsymbol{m}_{i.} - \boldsymbol{m}_{..})^T \tag{4}$$

$$\boldsymbol{M}_B = \frac{1}{b-1}\sum_{j=1}^{b}(\boldsymbol{m}_{.j} - \boldsymbol{m}_{..})(\boldsymbol{m}_{.j} - \boldsymbol{m}_{..})^T \tag{5}$$

$$\boldsymbol{M}_{AB} = \frac{1}{(a-1)(b-1)}\sum_{i=1}^{a}\sum_{j=1}^{b}(\boldsymbol{m}_{ij} - \boldsymbol{m}_{i.} - \boldsymbol{m}_{.j} + \boldsymbol{m}_{..})(\boldsymbol{m}_{ij} - \boldsymbol{m}_{i.} - \boldsymbol{m}_{.j} + \boldsymbol{m}_{..})^T \tag{6}$$

$$\boldsymbol{M}_e = \frac{1}{N-ab}\sum_{i=1}^{a}\sum_{j=1}^{b}[\frac{1}{n_{ij}}\sum_{k=1}^{n_{ij}}(\boldsymbol{x}_{ijk} - \boldsymbol{m}_{ij})(\boldsymbol{x}_{ijk} - \boldsymbol{m}_{ij})^T] \tag{7}$$

and

$$\boldsymbol{m}_{..} = \frac{1}{ab}\sum_{i=1}^{a}\sum_{j=1}^{b}\boldsymbol{m}_{ij} \tag{8}$$

$$\boldsymbol{m}_{i.} = \frac{1}{b}\sum_{j=1}^{b}\boldsymbol{m}_{ij} \tag{9}$$

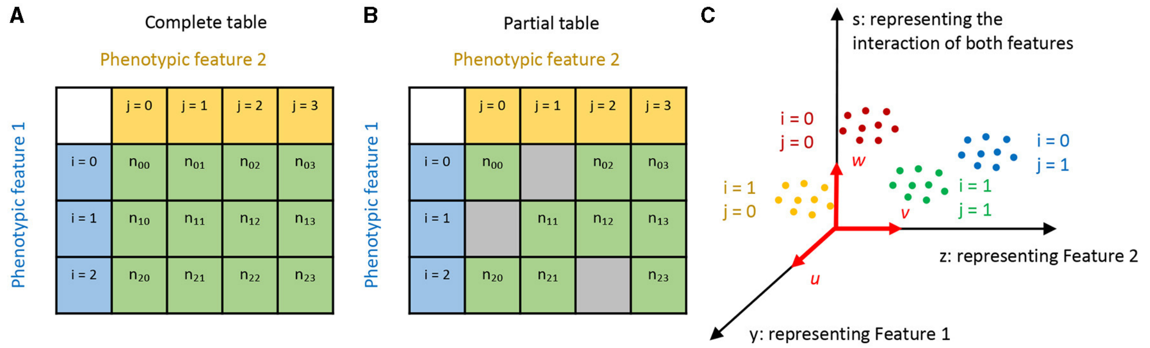
$$\boldsymbol{m}_{.j} = \frac{1}{a}\sum_{i=1}^{a}\boldsymbol{m}_{ij} \tag{10}$$

with

$$\boldsymbol{m}_{ij} = \frac{1}{n_{ij}}\sum_{k=1}^{n_{ij}}\boldsymbol{x}_{ijk} \tag{11}$$

Analogously, the linear projections $\boldsymbol{v}$ for feature j and $\boldsymbol{w}$ for the combination of both features i and j can be determined by similar formulas. By applying the same rationale to a partial table, we can derive $\boldsymbol{u}$ or $\boldsymbol{v}$ as the linear projection for feature i or j (see Appendix B for a detailed mathematical discussion).

Note that $\boldsymbol{N}_A$ is symmetric and $\boldsymbol{M}_e$ is positive definite, transforming the optimization problem into a generalized eigenvalue problem (Ghojogh et al., 2019). When $\boldsymbol{M}_e$ is invertible, $\boldsymbol{u}^*$ is the eigenvector associated with the highest eigenvalue of $\boldsymbol{M}_e^{-1}\boldsymbol{N}_A$. Generally, if we aim to embed $\boldsymbol{x}_{ijk}$ into a d-dimensional subspace aligned with feature i $(d < a)$, we take the eigenvectors corresponding to the d largest eigenvalues of $\boldsymbol{M}_e^{-1}\boldsymbol{N}_A$, which we term as the top d factorized linear discriminant components (FLDs).



	j = 0	j = 1	j = 2	j = 3
i = 0	n_{00}	n_{01}	n_{02}	n_{03}
i = 1	n_{10}	n_{11}	n_{12}	n_{13}
i = 2	n_{20}	n_{21}	n_{22}	n_{23}

	j = 0	j = 1	j = 2	j = 3
i = 0	n_{00}		n_{02}	n_{03}
i = 1		n_{11}	n_{12}	n_{13}
i = 2	n_{20}	n_{21}		n_{23}

FIGURE 1
Illustration of our approach. **(A, B)** Here, cell types are represented by two phenotypic features, labeled with *i* and *j*, respectively. If only some combinations of the two features are observed, we have a partial contingency table **(B)** rather than a complete one **(A)**. **(C)** We aim to find linear projections of the data that separate the cell types in a manner factorized according to the two features. In this diagram, $\boldsymbol{u}$, $\boldsymbol{v}$, and $\boldsymbol{w}$ are aligned with Feature 1, Feature 2, and their combination respectively, with the projected coordinates *y*, *z*, and *s*.

In situations where the number of genes greatly exceeds the number of cells, $\boldsymbol{M}_e$ becomes singular and non-invertible. In such cases, we resort to solutions suggested in Friedman (1989), Dudoit et al. (2002), and Bickel and Levina (2004) that uses a diagonal estimate of $\boldsymbol{M}_e$: $diag(\hat{\sigma}_1^2, \hat{\sigma}_2^2, ..., \hat{\sigma}_p^2)$, where $\hat{\sigma}_i^2$ is the ith diagonal element of $\boldsymbol{M}_e$. This solution has been employed in multiple computational biology studies (Tibshirani et al., 2003; Butler et al., 2018; Stuart et al., 2019).

As multi-way ANOVA can handle contingency tables with more than two dimensions, our analysis can be easily extended to handle more than two phenotypic features (Hahn et al., 2023). In summary, FLDA is well-suitable for data whose labels form a Cartesian product of multiple features.

3 Sparse regularization of FLDA

In computational biology applications, we are often interested in identifying a small subset of genes that effectively characterizes a specific phenotypic feature. This leads to the identification of axes with a few non-zero elements. To find such a sparse solution, we address the following optimization problem:

$$\boldsymbol{u}^* = \underset{\boldsymbol{u} \in \mathbb{R}^g}{\arg\max} \frac{\boldsymbol{u}^T \boldsymbol{N}_A \boldsymbol{u}}{\boldsymbol{u}^T \boldsymbol{M}_e \boldsymbol{u}} \quad \text{subject to} \quad ||\boldsymbol{u}||_0 \leq l \qquad (12)$$

where the number of non-zero elements of $\boldsymbol{u}^*$ is constrained to be less or equal to l.

This problem, also known as a sparse generalized eigenvalue problem, presents three challenges (Tan et al., 2018): Handling extremely high-dimensional data, $\boldsymbol{M}_e$ can be singular and non-invertible; Working with the normalization term $\boldsymbol{u}^T \boldsymbol{M}_e \boldsymbol{u}$, which restricts the application of many sparse eigenvalue solutions; Maximizing a convex objective over a non-convex set, a problem known to be NP-hard.

To overcome these challenges, we employ the truncated Rayleigh flow (Rifle) method, which was designed specifically for solving sparse generalized eigenvalue problems. The Rifle algorithm is a two-step process (Tan et al., 2018): First, it acquires an initial vector $\boldsymbol{u}_0$ that is close to $\boldsymbol{u}^*$. For this, we use the non-sparse FLDA solution as an initial estimate for $\boldsymbol{u}_0$; Second, it iteratively performs a gradient ascent on the objective function. This is followed by a truncation step that retains the l entries of $\boldsymbol{u}$ with the highest values and sets the remaining entries to zero. The step-by-step process of applying the Rifle method to solve our problem is detailed in the following pseudo-code:

```
procedure RIFLE($\boldsymbol{N}_A, \boldsymbol{M}_e, \boldsymbol{u}_0, l, \eta$)      ▷ $\eta$ is the step size
    $t = 1$                     ▷ $t$ indicates the iteration number
    while not converge do      ▷ Converge when $\boldsymbol{u}_t \simeq \boldsymbol{u}_{t-1}$
        $\rho_{t-1} \leftarrow \frac{\boldsymbol{u}_{t-1}^T \boldsymbol{N}_A \boldsymbol{u}_{t-1}}{\boldsymbol{u}_{t-1}^T \boldsymbol{M}_e \boldsymbol{u}_{t-1}}$
        $\boldsymbol{C} \leftarrow \boldsymbol{I} + (\frac{\eta}{\rho_{t-1}})(\boldsymbol{N}_A - \rho_{t-1} \boldsymbol{M}_e)$
        $\boldsymbol{u}_t \leftarrow \frac{\boldsymbol{C} \boldsymbol{u}_{t-1}}{||\boldsymbol{C} \boldsymbol{u}_{t-1}||_2}$
        Truncate $\boldsymbol{u}_t$ by preserving the top $l$ entries of $\boldsymbol{u}$ with the largest values and setting the remaining entries to 0
        $\boldsymbol{u}_t \leftarrow \frac{\boldsymbol{u}_t}{||\boldsymbol{u}_t||_2}$
        $t \leftarrow t + 1$
    end while
    return $\boldsymbol{u}_t$
end procedure
```

As previously demonstrated in Tan et al. (2018), the Rifle method can effectively converge to the unique sparse leading generalized eigenvector, assuming it exists, at the optimal statistical rate of convergence. The computational complexity of the second step in each iteration is $O(lg + g)$, indicating that the Rifle algorithm scales linearly with g, the number of genes in the input data.

In terms of hyperparameter selection, the step size η should be small enough to ensure convergence, specifically $\eta \lambda_{max}(\boldsymbol{M}_e) < 1$, where $\lambda_{max}(\boldsymbol{M}_e)$ is the largest eigenvalue of $\boldsymbol{M}_e$. This is akin to taking small steps to ensure that we don't overshoot the optimal solution. The other hyperparameter, l, which determines the number of genes to be preserved, is chosen empirically based on the design of the subsequent experiment. This parameter can be adjusted depending on the specific requirement of a biological study.

4 Related work: dimensionality reduction

FLDA is one method for linear dimensionality reduction (Cunningham and Ghahramani, 2015). In formal terms, linear dimensionality reduction can be defined as follows: Given n data points, each of g dimensions, $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n] \in \mathbf{R}^{g \times n}$, and a chosen reduced dimensionality $r < g$, an objective function $f(.)$ is optimized to produce a linear projection $\boldsymbol{U} \in \mathbf{R}^{r \times g}$. The result is a low-dimensional transformed dataset $\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{X} \in \mathbf{R}^{r \times n}$.

Leading methods for linear dimensionality reduction include Principal Component Analysis (PCA), Factor Analysis (FA), Linear Multidimensional Scaling (MDS), Linear Discriminant Analysis (LDA), Canonical Correlation Analysis (CCA), Maximum Autocorrelation Factors (MAF), Slow Feature Analysis (SFA), Sufficient Dimensionality Reduction (SDR), Locality Preserving Projections (LPP), and Independent Component Analysis (ICA; Cunningham and Ghahramani, 2015). These approaches are important in single-cell transcriptomics for dissecting cellular heterogeneity, understanding cellular differentiation trajectories, and identifying correspondences between cells in different experiments (Trapnell et al., 2014; Stuart et al., 2019; Xiang et al., 2021).

4.1 Unsupervised methods for linear dimensionality reduction

Unsupervised linear dimensionality reduction methods, including PCA (Jolliffe, 2002), ICA (Hyvärinen et al., 2001), and FA (Spearman, 1904), project data into a low-dimensional space without the use of supervision labels. These methods are crucial in the initial stages of single-cell data analysis to reduce dimensionality and noise, and have been used in numerous studies to identify subpopulations of cells and understand the variance structure of the data (Stuart et al., 2019; Xiang et al., 2021). The shortcoming of these unsupervised methods is that the axes of the low-dimensional space often fail to represent the underlying structure of the data, rendering them uninterpretable. This issue is particularly pronounced with gene expression data due to its high dimensionality (usually encompassing tens of thousands of genes) and the noisy expressions of many genes. These noisy expressions result in significant variance among individual cells, albeit without a structured pattern. In the absence of supervisory signals from phenotypic features, unsupervised methods tend to select these genes to construct the low-dimensional space, which does not provide the desired alignment or effective separation of cell type clusters. To illustrate this, we compared the performance of PCA on the gene expression data with that of FLDA. In brief, we solved the following objective to find the linear projection:

$$\boldsymbol{u}^* = \underset{\boldsymbol{u} \in \mathbf{R}^g}{\arg\max} \frac{\boldsymbol{u}^T \boldsymbol{X}\boldsymbol{X}^T \boldsymbol{u}}{\boldsymbol{u}^T \boldsymbol{u}} \tag{13}$$

The results of this comparison are detailed in the Results Section.

4.2 Supervised methods for linear dimensionality reduction

Supervised linear dimensionality reduction techniques, such as LDA (Fisher, 1936; McLachlan, 2004) and CCA (Hotelling, 1936; Wang et al., 2016), can overcome the aforementioned issues. By incorporating supervised signals of phenotypic features, genes whose expressions do not inform on the phenotypes can be de-emphasized.

4.2.1 Linear discriminant analysis (LDA)

LDA models the differences among data organized in predetermined classes. Formally, the optimization problem solved by LDA is as follows:

$$\boldsymbol{u}^* = \underset{\boldsymbol{u} \in \mathbf{R}^g}{\arg\max} \frac{\boldsymbol{u}^T \boldsymbol{\Sigma}_b \boldsymbol{u}}{\boldsymbol{u}^T \boldsymbol{\Sigma}_e \boldsymbol{u}} \tag{14}$$

where $\boldsymbol{\Sigma}_b$ and $\boldsymbol{\Sigma}_e$ are estimates of the between-class and within-class covariance matrices, respectively.

Unlike FLDA, LDA doesn't explicitly formulate the representation of these classes as a contingency table composed of multiple features. As a result, when applied to an example problem where cell types are organized into a two-dimensional contingency table with phenotypic features i and j, the axes from LDA are generally not aligned with these two phenotypic features.

However, it is possible to perform two separate LDAs for the two features. This modification allows the axes from each LDA to align with its specific feature. We refer to this approach as "2LDAs." There are two main limitations of this approach: first, it discards information about the component depending on the combination of the two features; second, it explicitly maximizes the segregation of cells with different feature levels, which sometimes is not consistent with a good separation of cell type clusters. Detailed comparisons between LDA, "2LDAs," and FLDA are provided in the Results Section.

4.2.2 Canonical correlation analysis (CCA)

CCA projects two datasets $\boldsymbol{X}_a \in \mathbf{R}^{g \times n}$ and $\boldsymbol{X}_b \in \mathbf{R}^{d \times n}$ to $\boldsymbol{Y}_a \in \mathbf{R}^{r \times n}$ and $\boldsymbol{Y}_b \in \mathbf{R}^{r \times n}$, such that the correlation between $\boldsymbol{Y}_a$ and $\boldsymbol{Y}_b$ is maximized. Formally, it tries to maximize this objective:

$$(\boldsymbol{u}, \boldsymbol{v}) = \underset{\boldsymbol{u} \in \mathbf{R}^g, \boldsymbol{v} \in \mathbf{R}^d}{\arg\max} \frac{\boldsymbol{u}^T (\boldsymbol{X}_a \boldsymbol{X}_a^T)^{-\frac{1}{2}} \boldsymbol{X}_a \boldsymbol{X}_b^T (\boldsymbol{X}_b \boldsymbol{X}_b^T)^{-\frac{1}{2}} \boldsymbol{v}}{(\boldsymbol{u}^T \boldsymbol{u})^{-\frac{1}{2}} (\boldsymbol{v}^T \boldsymbol{v})^{-\frac{1}{2}}} \tag{15}$$

To apply CCA to our problem, we designate $\boldsymbol{X}_a$ as the gene expression matrix, and $\boldsymbol{X}_b$ as the matrix of d phenotypic features ($d = 2$ for two features as demonstrated later). Unlike FLDA, CCA identifies a transformation of gene expressions that is aligned with a linear combination of phenotypic features, instead of a factorization of gene expressions corresponding to individual phenotypic features. The differences in these approaches are quantified and discussed in the Results Section.

4.3 Non-linear dimensionality reduction methods

Apart from linear dimensionality reduction, non-linear methods have emerged as popular choices for analyzing single-cell transcriptomic datasets due to their ability to capture complex, non-linear relationships inherent in the data (Xiang et al., 2021). Notable among these methods are t-Distributed Stochastic Neighbor Embedding (t-SNE) (Maaten and Hinton, 2008) and Uniform Manifold Approximation and Projection (UMAP; McInnes et al., 2018). Unlike linear methods, these algorithms can unravel intricate structures in the data by modeling non-linear manifold structures.

t-SNE minimizes the divergence between two distributions over pairs of data points, one in the high-dimensional space and one in the low-dimensional space, to create a map that reflects the structure of the data. UMAP assumes that the data is uniformly distributed on a locally-connected Riemannian manifold and seeks to find a similar uniform distribution in lower dimensions.

The comparison of FLDA with t-SNE and UMAP hinges on the trade-off between linear and non-linear dimensionality reductions. While non-linear methods excel in capturing complex data structures and modeling dropout effects (Qiu, 2020), and often produce visually appealing embeddings, they exhibit certain limitations compared to linear methods, such as:

- Interpretability: Linear methods offer a clear and direct relationship between the original features and the reduced dimensions, which facilitates interpretability. In contrast, the embeddings produced by non-linear methods are often challenging to interpret due to the complex and non-linear transformation functions involved.
- Computational Efficiency: Linear methods are generally more computationally efficient compared to non-linear methods, which can become computationally intensive, especially as the size of the dataset increases.

In single-cell transcriptomics applications, the choice between linear and non-linear dimensionality reduction hinges on balancing the capture of complex data structures with the maintenance of interpretability and computational efficiency. In the context of this paper, our proposed FLDA method is designed to address the challenges associated with single-cell data by offering a structured and interpretable low-dimensional space aligned with neuronal phenotypes. Therefore, we constrained our comparisons of FLDA with other linear dimensionality reduction methods that share the objective of interpretability.

5 Experimental design

5.1 Datasets

To quantitatively evaluate FLDA against other linear dimensionality reduction methods such as PCA, CCA, LDA, and the "2LDAs" approach, we initially opted for synthetic datasets. The primary rationale behind this choice lay in the controlled environment synthetic data afford, enabling a precise and standardized comparison of the methods under varying conditions. These datasets consisted of four types of cells, each containing 250 examples, generated from a 2 × 2 Cartesian product of two features i and j (Figure 2A). We generated expressions for 1000 genes of each cell, with gene levels being either purely noise-driven or correlated with feature i, feature j, or the combination of both. Detailed information about the data generation can be found in Appendix C.

To bridge the gap between the controlled synthetic environment and real-world biological scenarios, we employed a dataset of Drosophila T4/T5 neurons (Kurmangaliyev et al., 2019) to demonstrate the applicability and advantages of FLDA in analyzing single-cell transcriptome datasets. T4 and T5 neurons, while similar in general morphology and physiological properties, differ in the location of their dendrites in the medulla and lobula, which are two separate brain regions. Both T4 and T5 neurons comprise four subtypes, each pair demonstrating axonal lamination in a specific layer within the lobula plate (Figure 3A). Thus, we identified these neurons using two phenotypic features: feature i indicating the dendritic location in either the medulla or lobula, and feature j signifying axonal lamination at one of the four layers (a/b/c/d) (Figure 3B). In this study, we concentrated on a dataset containing expression data for 17,492 genes from 3,833 cells, all collected at a predefined time during brain development.

5.2 Data preprocessing

The preprocessing of the T4/T5 neuron dataset adhered to previously documented procedures (Shekhar et al., 2016; Kurmangaliyev et al., 2019; Peng et al., 2019; Tran et al., 2019). Briefly, the transcript counts within each column of the count matrix (genes×cells) were normalized to equate to the median number of transcripts per cell, leading to normalized counts, or Transcripts-per-million (TPM_{gc}), for Gene g in Cell c. We used the log-transformed expression data, denoted by $E_{gc} = \ln(TPM_{gc} + 1)$, for subsequent analysis. We selected highly variable genes for further FLDA application based on a common approach in single-cell RNA-Seq studies. This approach is based on establishing a relationship between mean and coefficient of variation (Chen et al., 2016; Pandey et al., 2018; Kurmangaliyev et al., 2019). For this particular experiment, we set the hyper-parameters λs in Equation (3) to 1.

5.3 Evaluation metrics

The dimensionality reduction process should satisfy two primary goals: (1) to identify axes that efficiently segregate distinct cell types, and (2) to discover axes that are well-aligned with the respective labels. Consequently, to evaluate the effectiveness of FLDA and various alternative methodologies, we implemented the following metrics (Detailed information of implementing these metrics can be found in Appendix D):

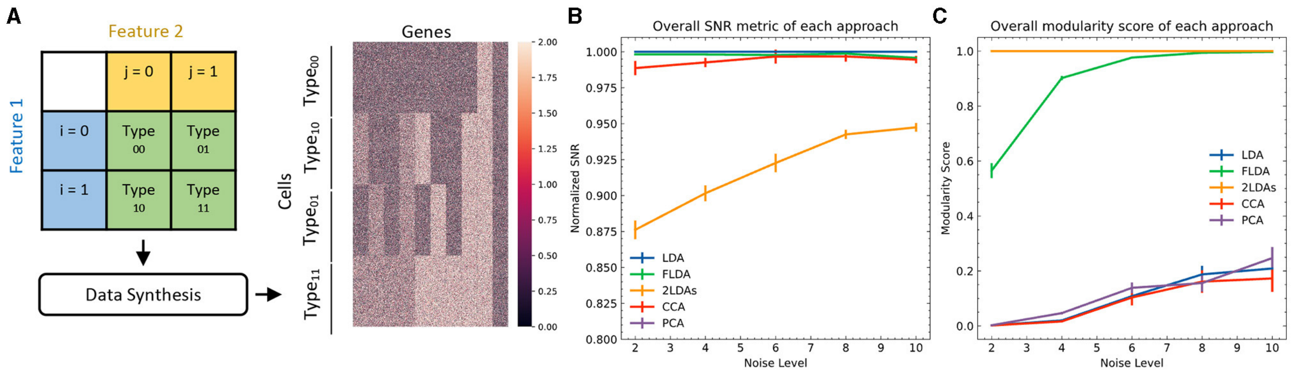


FIGURE 2

Quantitative comparison between FLDA and other models. **(A)** Illustration of data synthesis. For implementation details, see Appendix C. The color bar represents the expression values of the 1000 generated genes. **(B)** Normalized overall Signal-to-Noise Ratio (SNR) metric for each analysis, normalized with respect to that of LDA. The normalized SNR metric of PCA is below 0.8. **(C)** Overall modularity score for each analysis. The error bars in **(B, C)** denote standard errors calculated from 10 repeated simulations.

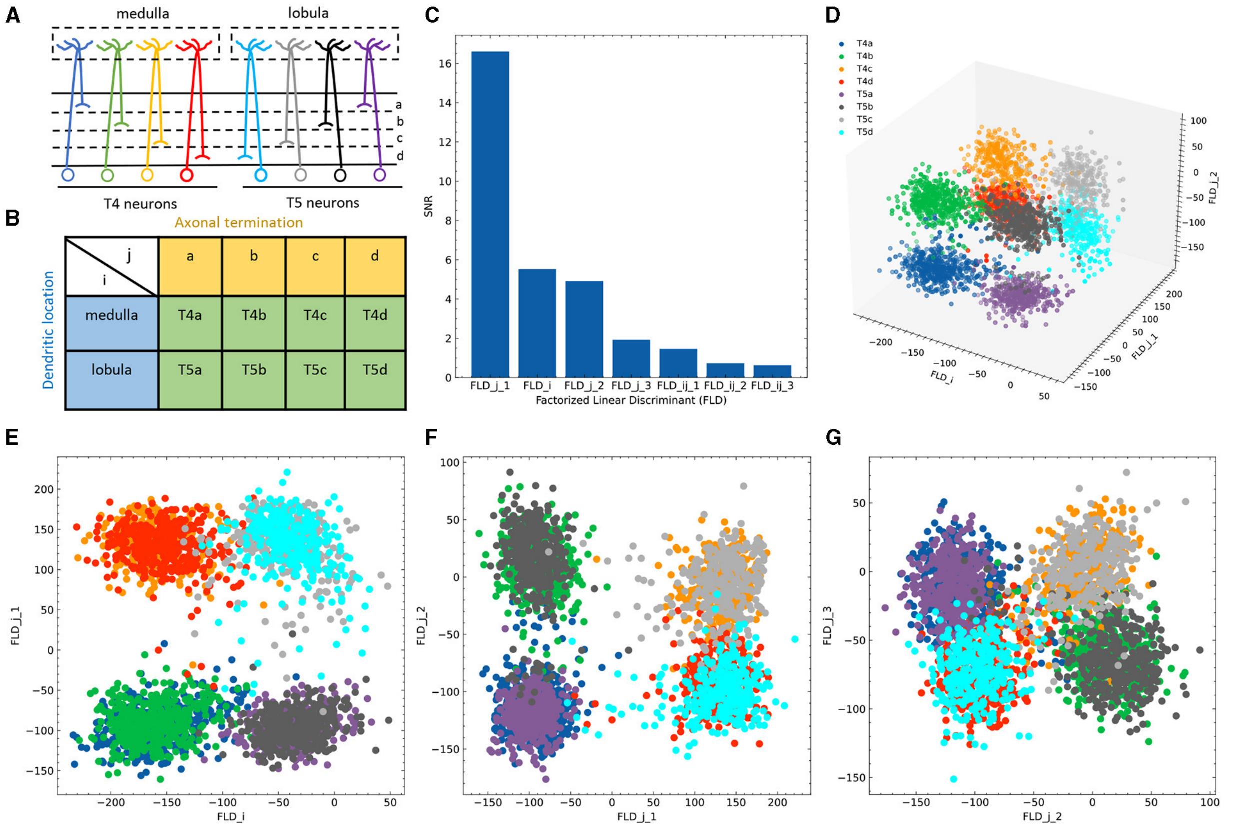


FIGURE 3

Application of FLDA to the dataset of T4/T5 neurons. **(A)** T4/T5 neuronal cell types and their dendritic and axonal location phenotypes. **(B)** The organization of T4/T5 neurons in a complete contingency table, where *i* indicates dendritic location and *j* indicates axonal termination. **(C)** SNR metric for each discriminant axis. **(D)** Data projection into the three-dimensional space consisting of the discriminant axis for feature *i* (FLD_i) and the first and second discriminant axes for feature *j* (FLD_{j_1} and FLD_{j_2}). **(E–G)** Data projection into the two-dimensional space comprised of FLD_i and FLD_{j_1} **(E)**, FLD_{j_1} and FLD_{j_2} **(F)**, or FLD_{j_2} and FLD_{j_3} (the third discriminant axis for feature *j*) **(G)**. Different cell types are represented by different colors as depicted in **(A, D)**.

- Signal-to-Noise Ratio (SNR): This metric measures the efficacy of each discriminant axis in distinguishing distinct cell types. Higher SNR suggests better separation of different cell types. This metric is relevant to the first goal.
- Explained Variance (EV): This metric gauges the proportion of variance of the feature *i* or *j* that a discriminant axis explains. Higher EV indicates that the dimensionality reduction method effectively encapsulates

TABLE 1 Average Silhouette scores for FLDA and other models.

Sigma	FLDA	2LDAs	LDA	CCA	PCA
2	0.905050961	0.899968945	0.904654957	0.904642723	0.862205691
4	0.809044235	0.799580694	0.80809655	0.808670943	0.70561478
6	0.70898633	0.697440319	0.707543347	0.708353597	0.524329239
8	0.624518364	0.613810816	0.622878691	0.624528348	0.33756613
10	0.535243253	0.526429522	0.532720359	0.534676146	0.145181255

the feature information. This metric is relevant to the second goal.

- Mutual Information (MI): This metric calculates the association between each discriminant axis and each feature, providing insights into how much information an axis provides about a specific feature. A higher MI score suggests better ability of the dimensionality reduction method to capture essential characteristics. This metric is relevant to the second goal.
- Modularity Score: This metric assesses whether each axis is predominantly dependent on a single feature (Ridgeway and Mozer, 2018). A higher modularity score indicates successful disentanglement of features, which is crucial for interpreting biological data. This metric is relevant to the second goal.
- Silhouette Score: This metric computes the average Silhouette value of all samples, which is a measure of how similar a cell is to its own cluster compared to other clusters. A higher Silhouette score indicates better cluster separation and tighter clustering, This metric is relevant to the first goal.

In addition, we evaluated the execution times of FLDA and alternative methodologies.

6 Results

6.1 Comparative analysis of FLDA with other linear dimensionality reduction methods

To provide a quantitative comparison between FLDA and other dimensionality reduction methods such as PCA, CCA, LDA, and "2LDAs," we measured the proposed metrics on the synthesized datasets as shown in Figure 2A. Given that the synthesized data was organized in a 2 × 2 table, each LDA of the "2LDAs" approach could only identify one dimension for the specific features i or j. Therefore, as a fair comparison, we only included the corresponding dimensions in FLDA (FLD$_i$ and FLD$_j$) and the top two components of PCA, CCA, and LDA. The overall SNR values normalized by that of LDA and the modularity scores across different noise levels are depicted in Figures 2B, C. The performance of PCA is the worst due to its unsupervised approach, which cannot effectively mitigate the impact of noise on the signal. While supervised approaches generally demonstrate superior SNR, LDA, and CCA suffer from low modularity scores. This outcome aligns with our expectation, as LDA maximizes cell type cluster separation without necessarily aligning axes to individual features i or j, and CCA maximizes the correlation to a linear combination of phenotypic features rather than individual ones. Conversely, "2LDAs" achieves the highest modularity scores but exhibits the lowest SNR among supervised approaches, as it aims to maximize the separation of cells with different feature levels, which does not necessarily coincide with maximizing cell type segregation. Both the SNR and modularity score of FLDA approach optimal values because it considers both the alignment of axes to different features and the constraint of variance within cell types. Consistent with the SNR metric, the average Silhouette score for FLDA is close to those of LDA and CCA, outperforms "2LDAs", and significantly surpasses PCA, as detailed in Table 1. Consistent with the modularity score, a robust axis alignment to either feature i or j is observed in FLDA and "2LDAs," but not in the other methods, as shown in a representative plot of the EV and MI metrics across these models in Figure 4.

We further analyzed the execution times of FLDA and other models and summarized the findings in Table 2. The execution time of FLDA is on par with that of LDA, albeit longer than PCA's, attributed to the handling of the covariance matrix in the denominator. In contrast, the execution times for "2LDAs" and CCA are considerably extended, nearly doubling those of FLDA and LDA. This increment is due to "2LDAs" requiring two LDA operations, while CCA necessitates the computation of covariance matrices for both input and phenotypic features, thereby doubling the execution time.

6.2 Real-world application in computational biology

A significant question in biology is whether diverse cell type phenotypes are generated by modular transcriptional programs, and if so, what the gene signature for each program is. To demonstrate the potential of our approach in addressing this question, we applied FLDA to the Drosophila T4/T5 neuron dataset.

Given that the data is organized in a 2 × 4 contingency table, we chose to project the expression data into a seven-dimensional subspace. This subspace was structured such that one FLD was aligned with dendritic location i (FLD$_i$), three FLDs were aligned with axonal termination j (FLD$_{j_{1-3}}$), and the remaining three were tailored to represent the combination between both phenotypes (FLD$_{ij_{1-3}}$). Ranking these axes based on their SNR metrics revealed that FLD$_{j_1}$, FLD$_i$, and FLD$_{j_2}$ had considerably higher SNRs than the others (Figure 3C). Indeed, data representations in the subspace comprising these three dimensions clearly separated the eight

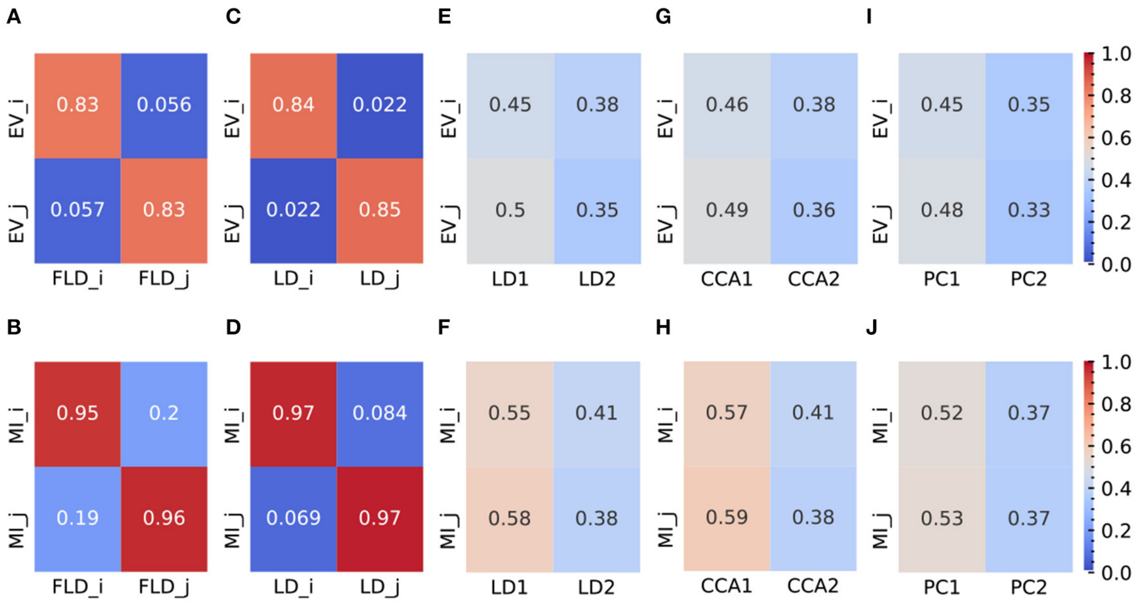


FIGURE 4
Representative plots (at $\sigma = 6$) of EV and MI metrics for FLDA and other models. **(A, B)** EV **(A)** and MI **(B)** metrics of FLDA. FLD$_i$ and FLD$_j$ indicate the factorized linear discriminants for features *i* and *j*. **(C, D)** EV **(C)** and MI **(D)** metrics of 2LDAs. LD$_i$ and LD$_j$ indicate the linear discriminant components for features *i* and *j*. **(E, F)** EV **(E)** and MI **(F)** metrics of LDA. LD$_1$ and LD$_2$ indicate the first two linear discriminant components. **(G, H)** EV **(G)** and MI **(H)** metrics of CCA. CCA$_1$ and CCA$_2$ indicate the first two canonical correlation axes. **(I, J)** EV **(I)** and MI **(J)** metrics of PCA. PC$_1$ and PC$_2$ indicate the first two principal components. EV$_i$ and EV$_j$ are the explained variance of features *i* and *j* along an axis, and MI$_i$ and MI$_j$ indicate the mutual information between an axis and features *i* and *j*, respectively. Values of EV and MI metrics are also indicated by the color bars on the right side.

TABLE 2 **Average execution time (in seconds) for FLDA and other models.**

Sigma	FLDA	2LDAs	LDA	CCA	PCA
2	0.67947216	1.540404677	0.760401511	1.453615189	0.045516276
4	0.673864269	1.524357891	0.768786931	1.277295494	0.045400095
6	0.670872188	1.523296094	0.763518047	1.225833297	0.046381855
8	0.674873614	1.526811552	0.761365056	1.207562256	0.045538688
10	0.679605055	1.521809649	0.759853601	1.190890384	0.045457077

neuronal cell types (Figure 3D). As expected, FLD$_i$ differentiated T4 from T5 neurons, which have dendrites located in different brain regions (Figure 3E). Interestingly, FLD$_{j_1}$ separated T4/T5 neurons into two groups, a/b vs. c/d, according to the upper or lower lobula place, while FLD$_{j_2}$ divided them into another two groups, a/d vs. b/c, indicating whether their axons laminated at the middle or lateral part of the lobula plate (Figures 3E, F). Among these three dimensions, FLD$_{j_1}$ has a much higher SNR than FLD$_i$ and FLD$_{j_2}$, suggesting a hierarchical structure in the genetic organization of T4/T5 neurons: they are first separated into either a/b or c/d types, and subsequently divided into each of the eight subtypes. In fact, this matches the sequence of their cell fate determination, as revealed in a previous genetic study (Pinto-Teixeira et al., 2018). Lastly, the final discriminant axis of the axonal feature FLD$_{j_3}$ separates the group a/c from b/d, suggesting its role in fine-tuning the axonal depth within the upper or lower lobula plate (Figure 3G).

To identify gene signatures for the discriminant components in FLDA, we applied sparsity-based regularization to constrain the number of genes with non-zero weight coefficients. We set the number to 20, a reasonable number of candidate genes that could be tested in a follow-up biological study. We extracted a list of 20 genes each for the axis of FLD$_i$ or FLD$_{j_1}$. The relative importance of these genes to each axis is directly informed by their weight values (Figures 5A, C). Alongside, we plotted expression profiles of these genes in the eight neuronal cell types (Figures 5B, D). For both axes, the genes critical in separating cells with different feature levels are differentially expressed in corresponding cell types. Finally, FLDA allowed us to examine the component that depends on the combination of both features and identify its gene signature, providing insights into transcriptional regulation of gene expressions in the T4/T5 neuronal cell types (Figures 6, 7).

6.3 Perturbation analysis

As FLDA, like other supervised methods, relies on accurate phenotype labels to extract meaningful information, we sought to investigate how it might behave in real-world scenarios where

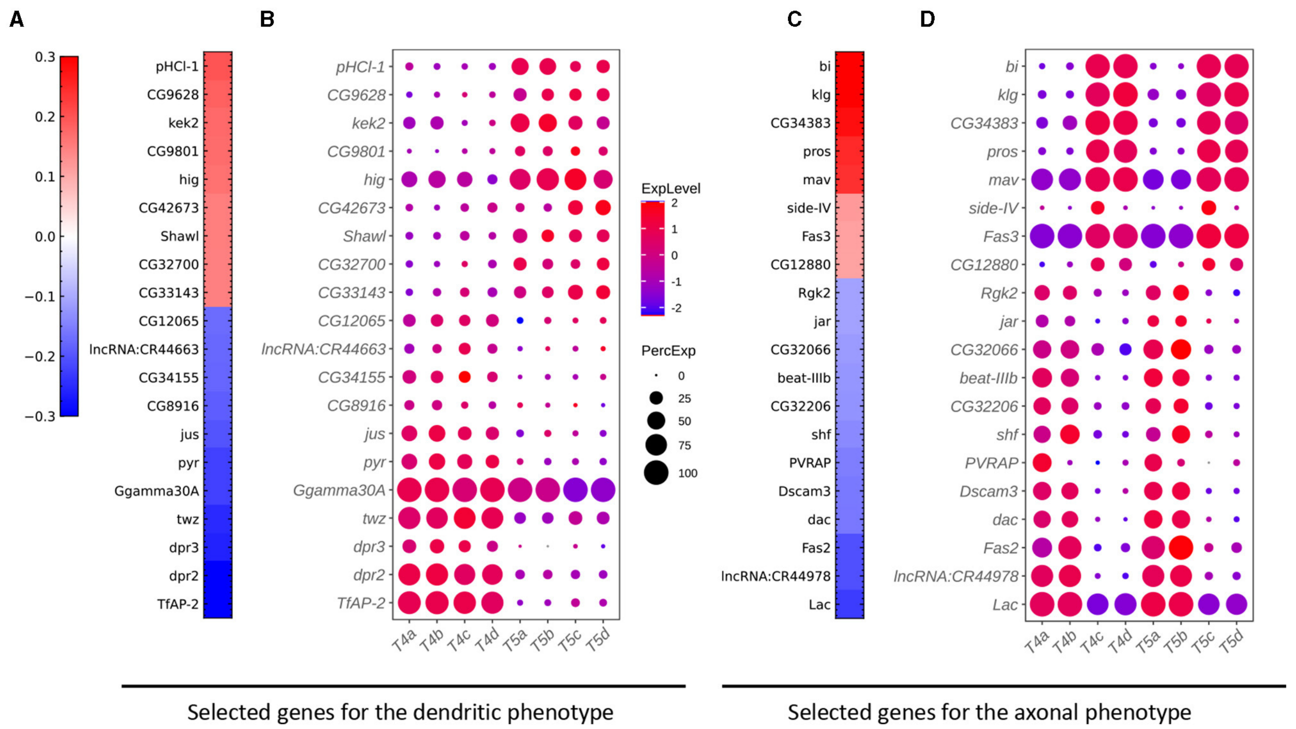


FIGURE 5
Critical genes extracted from the sparse algorithm. **(A)** Weight vector of the 20 genes selected for the dendritic phenotype (FLD$_i$). The weight value is indicated in the color bar with color indicating direction (red: positive and green: negative) and saturation indicating magnitude. **(B)** Expression patterns of the 20 genes from **(A)** in eight types of T4/T5 neurons. Dot size indicates the percentage of cells in which the gene was expressed, and color represents average scaled expression. **(C)** Weight vector of the 20 genes selected for the axonal phenotype (FLD$_{j_1}$). Legend as in **(A)**. **(D)** Expression patterns of the 20 genes from **(C)** in eight types of T4/T5 neurons. Legend as in **(B)**.

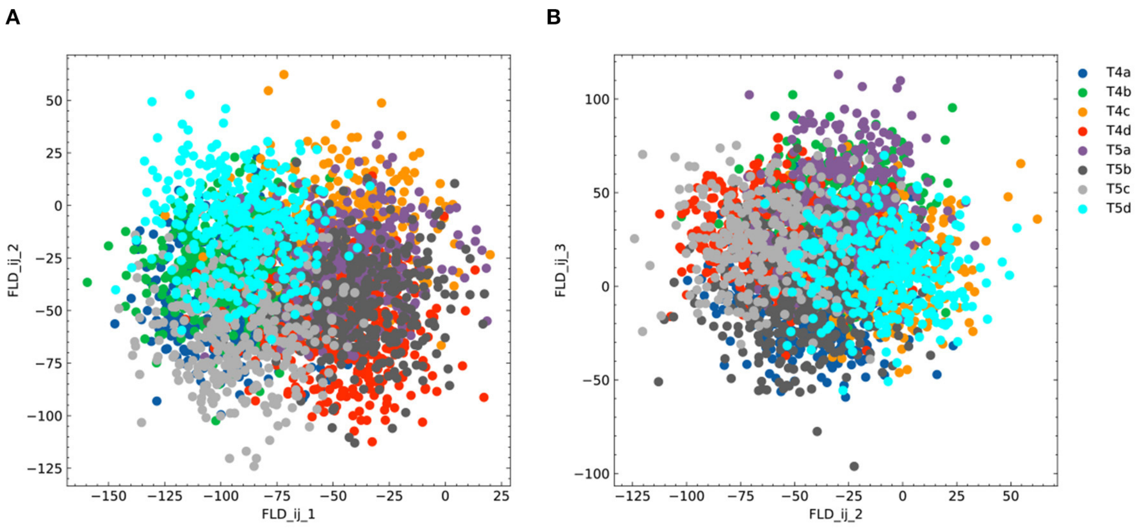


FIGURE 6
Additional plots for FLDA on the dataset of T4/T5 neurons. **(A, B)** Projection of the original gene expression data into the two-dimensional space made of the first and second (FLD$_{ij_1}$ and FLD$_{ij_2}$) **(A)** or the second and third (FLD$_{ij_2}$ and FLD$_{ij_3}$) **(B)** discriminant axes for the component that depends on the combination of both features *i* and *j*. Different cell types are indicated in different colors as in **(B)**.

inaccuracies are bound to occur. If the phenotypes are annotated incorrectly, can we use FLDA to raise a flag? To address this, we propose a perturbation analysis of FLDA, based on the assumption that among possible phenotype annotations, the projection of gene expression data with correct labels leads to better metric measurements than incorrect ones. As detailed in Appendix E, we deliberately generated three kinds of incorrect labels for the T4/T5 neuron dataset, simulating common errors that could occur during

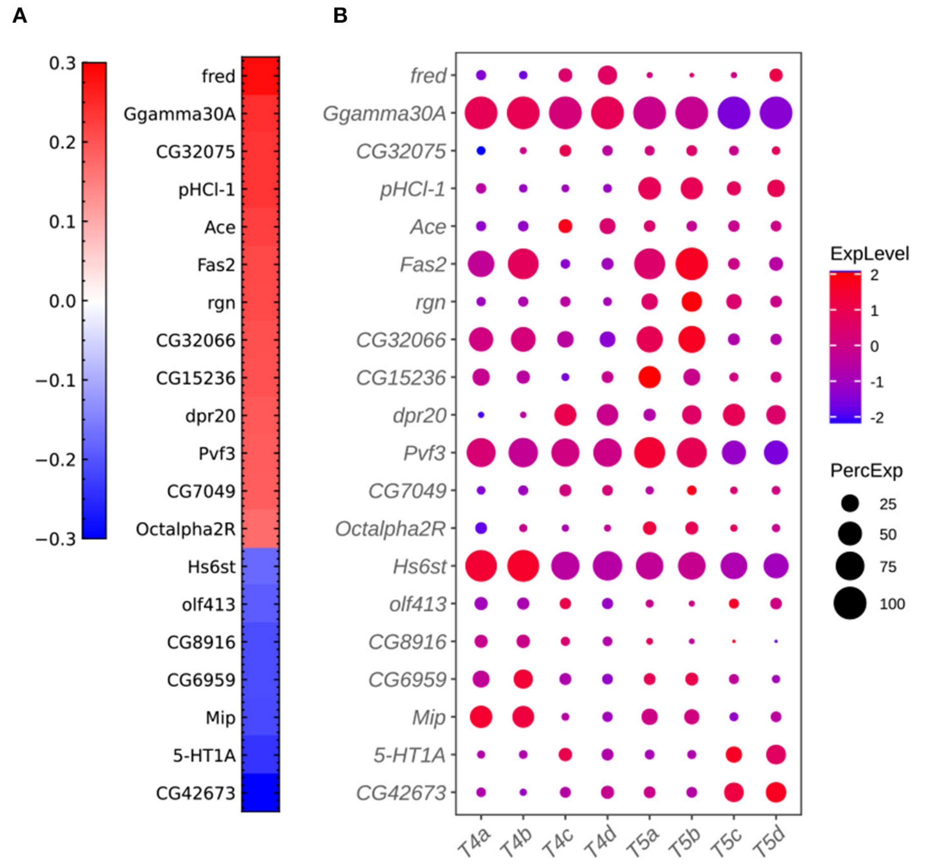


FIGURE 7
Additional plots for critical genes extracted from the sparse algorithm. **(A)** Weight vector of the 20 genes selected for the combination of dendritic and axonal features (FLD_{ij_1}). The weight value is indicated in the color bar with color indicating direction (red: positive and green: negative) and saturation indicating magnitude. **(B)** Expression patterns of the 20 genes from **(A)** in eight types of T4/T5 neurons. Dot size indicates the percentage of cells in which the gene was expressed, and color represents average scaled expression.

labeling: the phenotypes of a cell type were mislabeled with those of another type; a singular phenotypic category was incorrectly split into two; two phenotypic categories were incorrectly merged into one. We applied FLDA to gene expressions of T4/T5 neurons using these perturbed annotations, and found that proposed metrics, such as SNR and the modularity score, were best when the labels were correct (Figure 8), suggesting that this type of perturbation analysis can be used to flag potential errors in labeling.

In summary, our findings demonstrate that FLDA is a powerful tool for identifying and interpreting gene expressions that correspond to particular phenotypic features, even in the face of potential data mislabeling. This makes it a valuable tool for understanding complex biological systems. The perturbation analysis provides a robust method for validating the accuracy of phenotype annotations, thereby increasing the reliability of subsequent analyses and conclusions.

7 Discussion

We have introduced FLDA, a novel dimensionality reduction method that linearly projects high-dimensional data, such as gene expressions, into a low-dimensional space. The axes of this space are aligned with predefined features like phenotypes, making it an intuitive representation. Furthermore, we incorporated sparse regularization into FLDA, allowing us to select a small set of critical genes that are most informative about the phenotypes. Our application of FLDA in a computational biology context, particularly in the analysis of gene expression data from Drosophila T4/T5 neurons with two phenotypic labels, not only illuminated data structures aligned with the phenotypic labels, but also unveiled previously unreported genes associated with each phenotype. A comparison of our gene lists with those from the previous study (Kurmangaliyev et al., 2019) unveiled consistent genes including indicator genes for dendritic location like $TfAP-2$, $dpr2$, $dpr3$, twz, $CG34155$, and $CG12065$, and those for axonal lamination such as klg, bi, $pros$, mav, $beat-IIIb$, and $Fas2$. Remarkably, we identified genes not reported in the previous study. For example, our results suggest that the gene $pHCl-1$ is important to the dendritic phenotype, and the gene Lac is critical to axonal lamination. These genes are promising genetic targets for subsequent experimentation.

FLDA's potential extends beyond the dataset explored in this study. In a separate work, we applied FLDA to another real-world single-cell transcriptomic dataset, showcasing its ability to discern a low-dimensional representation of neuronal types

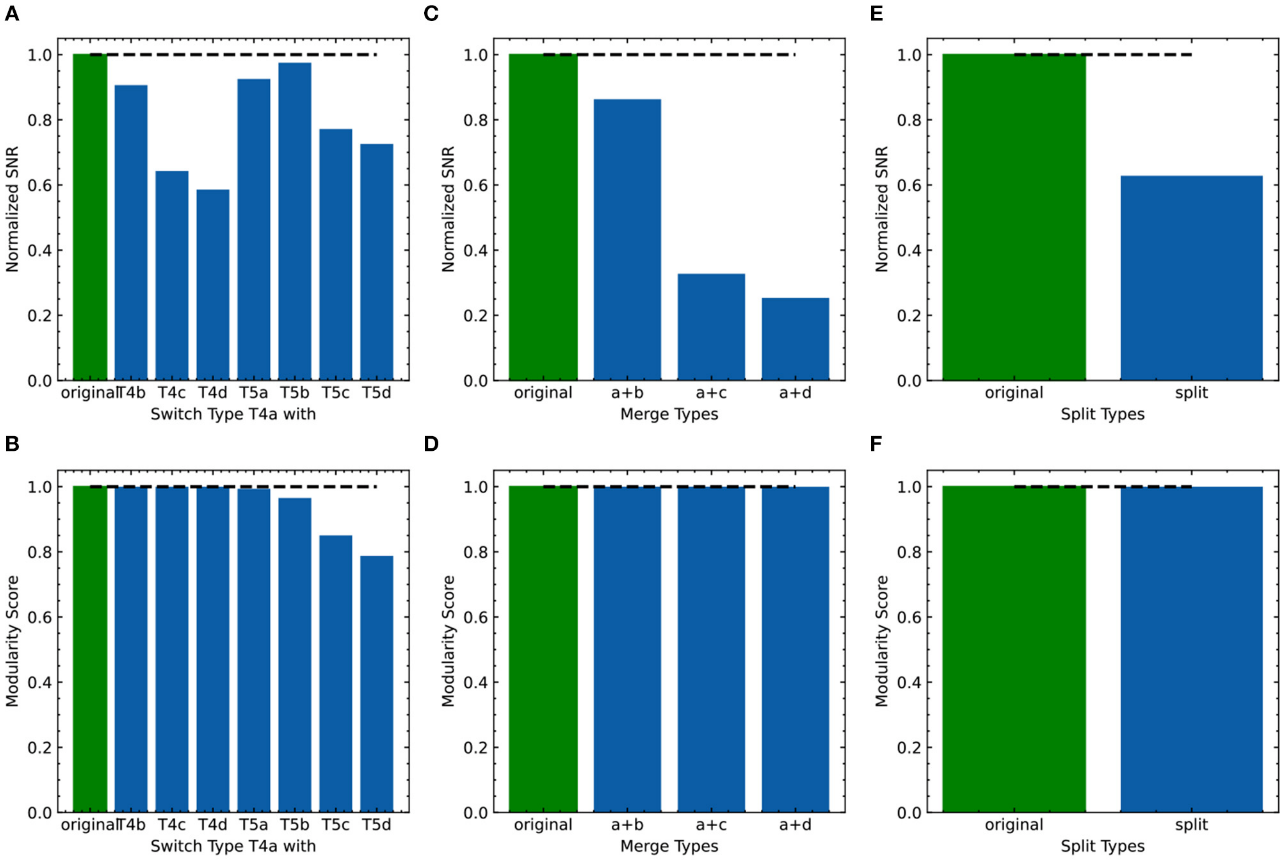


FIGURE 8
Evaluation of the effect of incorrect phenotype annotation on the dataset of T4/T5 neurons. **(A, B)** Normalized overall SNR metric **(A)** and overall modularity score **(B)** of FLDA after switching labels of T4a type with another neuronal type. **(C, D)** Normalized overall SNR metric **(C)** and overall modularity score **(D)** of FLDA after merging the axonal phenotypic level a with another phenotypic level (b/c/d). **(E, F)** Normalized overall SNR metric **(E)** and overall modularity score **(F)** of FLDA after splitting each axonal phenotypic level into two. Metrics under the original annotation are colored in green, and their values are indicated by the dashed lines. Here the SNR values are normalized with respect to that of the original annotation.

aligned with phenotypic and species attributes, thereby revealing evolutionary counterparts of primate retinal ganglion cells (Hahn et al., 2023). This further substantiates FLDA's applicability across diverse datasets and its promise in unveiling biologically meaningful insights.

The method could also play a role in the discovery of cell types. For example, the known phenotypes in a population might only form a partial table with missing entries (Figure 1B). Like the empty cells in Mendeleev's Periodic Table led to the prediction of new elements, these gaps could indicate predictions of new cell types (Mendelejew, 1869). FLDA can help pinpoint the region of the gene expression space that corresponds to the predicted new type, potentially revealing rare cell populations that might otherwise be overlooked due to insignificance.

Beyond computational biology, FLDA's application can extend to any labeled dataset with labels forming a Cartesian product of multiple attributes. This ability to separate attribute-specific factors makes FLDA invaluable in creating disentangled representations (Karaletsos et al., 2016; Ridgeway and Mozer, 2018). The potential of FLDA extends to these areas, and its performance can be optimized for diverse applications.

While our work offers significant advancements, it is not without limitations. The inherent linearity of FLDA, though providing an explicit and easily interpretable model, also presupposes a linear relationship between input features, which may not always hold true. Future work could involve a non-linear version of FLDA. For example, the input features can be projected into an embedding space using a neural network, where the axes align with each label attribute.

8 Code availability statement

FLDA analysis was performed in Python, and the code and documentation are available at: https://github.com/muqiao0626/FLDA-in-ComputBiol.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found at: Data used in the paper can be found in NCBI GEO under accession: GSE126139.

Author contributions

MQ developed the method, validated it, and wrote the manuscript.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Acknowledgments

We thank Jialong Jiang, Yuxin Chen, Oisin Mac Aodha, Matt Thomson, Lior Pachter, Andrew McMahon, Christof Koch, Yu-li Ni, Karthik Shekhar, Joshua R. Sanes, Tony Zhang, and Markus Meister for their helpful discussions and comments. MQ was supported by a postdoctoral fellowship from the Swartz Foundation.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fninf.2023.1265079/full#supplementary-material

References

Bickel, P. J., and Levina, E. (2004). Some theory for Fisher's linear discriminant function, 'Naive Bayes', and some alternatives when there are many more variables than observations. *Bernoulli* 10, 989–1010. doi: 10.3150/bj/1106314847

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* 36, 411–420. doi: 10.1038/nbt.4096

Cadwell, C. R., Palasantza, A., Jiang, X., Berens, P., Deng, Q., Yilmaz, M., et al. (2016). Electrophysiological, transcriptomic and morphologic profiling of single neurons using Patch-seq. *Nat. Biotechnol.* 34, 199–203. doi: 10.1038/nbt.3445

Chen, H.-I. H., Jin, Y., Huang, Y., and Chen, Y. (2016). Detection of high variability in gene expression from single-cell RNA-seq profiling. *BMC Genomics* 17(Suppl. 7):508. doi: 10.1186/s12864-016-2897-6

Cunningham, J. P., and Ghahramani, Z. (2015). Linear dimensionality reduction: survey, insights, and generalizations. *J. Mach. Learn. Res.* 16, 2859–2900.

Dudoit, S., Fridlyand, J., and Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Am. Stat. Assoc.* 97, 77–87. doi: 10.1198/016214502753479248

Fisher, R. A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *R. Soc. Edinburgh.* doi: 10.1017/S0080456800012163

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Ann. Eugen.* 7, 179–188. doi: 10.1111/j.1469-1809.1936.tb02137.x

Friedman, J. H. (1989). Regularized discriminant analysis. *J. Am. Stat. Assoc.* 84, 165–175. doi: 10.1080/01621459.1989.10478752

Ghojogh, B., Karray, F., and Crowley, M. (2019). Eigenvalue and generalized eigenvalue problems: tutorial. *arXiv:1903.11240.*

Gouwens, N. W., Sorensen, S. A., Berg, J., Lee, C., Jarsky, T., Ting, J., et al. (2019). Classification of electrophysiological and morphological neuron types in the mouse visual cortex. *Nat. Neurosci.* 22, 1182–1195. doi: 10.1038/s41593-019-0417-0

Hahn, J., Monavarfeshani, A., Qiao, M., Kao, A., Kölsch, Y., Kumar, A., et al. (2023). Evolution of neuronal cell classes and types in the vertebrate retina. *bioRxiv.* doi: 10.1101/2023.04.07.536039

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika* 28, 321–377. doi: 10.1093/biomet/28.3-4.321

Hyvärinen, A., Karhunen, J., and Oja, E. (2001). *Independent Component Analysis, 1st* Edn. New York, NY: Wiley-Interscience. doi: 10.1002/0471221317

Jolliffe, I. T. (2002). *Principal Component Analysis, 2nd Edn.* Springer Series in Statistics. New York, NY: Springer-Verlag.

Karaletsos, T., Belongie, S., and Rätsch, G. (2016). Bayesian representation learning with oracle constraints. *arXiv:1506.05011 [cs, stat*].

Kurmangaliyev, Y. Z., Yoo, J., LoCascio, S. A., and Zipursky, S. L. (2019). Modular transcriptional programs separately define axon and dendrite connectivity. eLife, 8:e50822. doi: 10.7554/eLife.50822

Maaten, L. v. d., and Hinton, G. (2008). Visualizing data using t-SNE. *J. Mach. Learn. Res.* 9, 2579–2605.

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., et al. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* 161, 1202–1214. doi: 10.1016/j.cell.2015.05.002

McInnes, L., Healy, J., and Melville, J. (2018). Umap: uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426.* doi: 10.21105/joss.00861

McLachlan, G. (2004). *Discriminant Analysis and Statistical Pattern Recognition.* Hoboken, NJ: Wiley-Interscience.

Mendelejew, D. (1869). Über die beziehungen der eigenschaften zu den atomgewichten der elemente. *Zeitsch. Chem.* 12, 405–406.

Pandey, S., Shekhar, K., Regev, A., and Schier, A. F. (2018). Comprehensive identification and spatial mapping of habenular neuronal types using single-cell RNA-Seq. *Curr. Biol.* 28, 1052–1065.e7. doi: 10.1016/j.cub.2018.02.040

Peng, Y.-R., Shekhar, K., Yan, W., Herrmann, D., Sappington, A., Bryman, G. S., et al. (2019). Molecular classification and comparative taxonomics of foveal and peripheral cells in primate retina. *Cell* 176, 1222–1237.e22. doi: 10.1016/j.cell.2019.01.004

Pinto-Teixeira, F., Koo, C., Rossi, A. M., Neriec, N., Bertet, C., et al. (2018). Development of concurrent retinotopic maps in the fly motion detection circuit. *Cell* 173, 485–498.e11. doi: 10.1016/j.cell.2018.02.053

Qiu, P. (2020). Embracing the dropouts in single-cell RNA-seq analysis. *Nat. Commun.* 11:1169. doi: 10.1038/s41467-020-14976-9

Ridgeway, K., and Mozer, M. C. (2018). "Learning deep disentangled embeddings with the F-statistic loss," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems* (Red Hook, NY: Curran Associates, Inc.), 185–194.

Sanes, J. R., and Masland, R. H. (2015). The types of retinal ganglion cells: current status and implications for neuronal classification. *Annu. Rev. Neurosci.* 38, 221–246. doi: 10.1146/annurev-neuro-071714-034120

Shekhar, K., Lapan, S. W., Whitney, I. E., Tran, N. M., Macosko, E. Z., Kowalczyk, M., et al. (2016). Comprehensive classification of retinal bipolar neurons by single-cell transcriptomics. *Cell* 166, 1308–1323.e30. doi: 10.1016/j.cell.2016.07.054

Spearman, C. (1904). "General intelligence," objectively determined and measured. *Am. J. Psychol.* 15, 201–292. doi: 10.2307/1412107

Strell, C., Hilscher, M. M., Laxman, N., Svedlund, J., Wu, C., Yokota, C., et al. (2019). Placing RNA in context and space –methods for spatially resolved transcriptomics. *FEBS J.* 286, 1468–1481. doi: 10.1111/febs.14435

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M., et al. (2019). Comprehensive integration of single-cell data. *Cell* 177, 1888–1902.e21. doi: 10.1016/j.cell.2019.05.031

Tan, K. M., Wang, Z., Liu, H., and Zhang, T. (2018). Sparse generalized eigenvalue problem: optimal statistical rates via truncated Rayleigh flow. *J. R. Stat. Soc. Ser. B* 80, 1057–1086. doi: 10.1111/rssb.12291

Tasic, B., Menon, V., Nguyen, T. N., Kim, T. K., Jarsky, T., Yao, Z., et al. (2016). Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nat. Neurosci.* 19, 335–346. doi: 10.1038/nn.4216

Tasic, B., Yao, Z., Graybuck, L. T., Smith, K. A., Nguyen, T. N., Bertagnolli, D., et al. (2018). Shared and distinct transcriptomic cell types across neocortical areas. *Nature* 563, 72–78. doi: 10.1038/s41586-018-0654-5

Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2003). Class prediction by nearest shrunken centroids, with applications to DNA microarrays. *Stat. Sci.* 18, 104–117. doi: 10.1214/ss/1056397488

Tran, N. M., Shekhar, K., Whitney, I. E., Jacobi, A., Benhar, I., Hong, G., et al. (2019). Single-cell profiles of retinal neurons differing in resilience to injury reveal neuroprotective genes. *bioRxiv* 2019:711762. doi: 10.1101/711762

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., et al. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* 32, 381–386. doi: 10.1038/nbt.2859

Wang, W., Arora, R., Livescu, K., and Bilmes, J. (2016). On deep multi-view representation learning: objectives and optimization. *arXiv:1602.01024 [cs*].

Xiang, R., Wang, W., Yang, L., Wang, S., Xu, C., and Chen, X. (2021). A comparison for dimensionality reduction methods of single-cell RNA-seq data. *Front. Genet.* 12:646936. doi: 10.3389/fgene.2021.646936

Zeng, H., and Sanes, J. R. (2017). Neuronal cell-type classification: challenges, opportunities and the path forward. *Nat. Rev. Neurosci.* 18, 530–546. doi: 10.1038/nrn.2017.85