JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

PLPCA: Persistent Laplacian-Enhanced PCA for Microarray Data Analysis

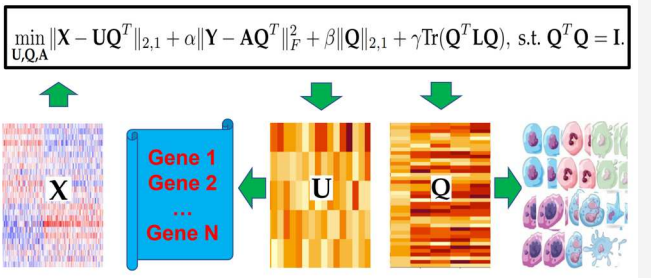
Sean Cottrell, Rui Wang, and Guo-Wei Wei*

Cite This: J. Chem. Inf. Model. 2024, 64, 2405−2420 Read Online

ACCESS | Metrics & More | Article Recommendations

Downloaded via MICHIGAN STATE UNIV on June 3, 2024 at 19:05:20 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

ABSTRACT: Over the years, Principal Component Analysis (PCA) has served as the baseline approach for dimensionality reduction in gene expression data analysis. Its primary objective is to identify a subset of disease-causing genes from a vast pool of thousands of genes. However, PCA possesses inherent limitations that hinder its interpretability, introduce class ambiguity, and fail to capture complex geometric structures in the data. Although these limitations have been partially addressed in the literature by incorporating various regularizers, such as graph Laplacian regularization, existing PCA based methods still face challenges related to multiscale analysis and capturing higher-order interactions in the data. To address these challenges, we propose a novel approach called Persistent Laplacian-enhanced Principal Component Analysis (PLPCA). PLPCA amalgamates the advantages of earlier regularized PCA methods with persistent spectral graph theory, specifically persistent Laplacians derived from algebraic topology. In contrast to graph Laplacians, persistent Laplacians enable multiscale analysis through filtration and can incorporate higher-order simplicial complexes to capture higher-order interactions in the data. We evaluate and validate the performance of PLPCA using ten benchmark microarray data sets that exhibit a wide range of dimensions and data imbalance ratios. Our extensive studies over these data sets demonstrate that PLPCA provides up to 12% improvement to the current state-of-the-art PCA models on five evaluation metrics for classification tasks after dimensionality reduction.



1. INTRODUCTION

Biological processes heavily depend on the different expression levels of genes over time. Thus, it is no surprise that analyzing gene expression data holds an important place in the field of biological and medical research, particularly in tasks such as identifying characteristic genes strongly correlated with various cancer types, as well as classifying tissue samples into cancerous and normal categories.[2]

In microarray analysis, mRNA molecules are collected from a tissue sample and converted into complementary DNA (cDNA). These cDNA molecules are subsequently labeled with a fluorescent dye and hybridized onto a microarray. The microarray is then scanned to measure the expression level of each gene. This process generates gene expression data, which represents the intensity of each gene in a sample, typically at the RNA production level. The continuous advancements in microarray technology have led to the generation of large-scale gene expression data sets.[3]

Gene expression data is commonly represented in matrix form, where rows correspond to genes and columns denote tissue samples. Each matrix element indicates the expression level of a specific gene in a particular sample. Given the high dimensionality of gene expression data, often encompassing over 10,000 genes but merely a few hundred samples, considering all genes in a tumor classification analysis could introduce noise and notably augment computational complexity. Therefore, it is common practice to perform gene filtering or dimensionality reduction prior to applying classification methods in order to extract meaningful insights with reduced noise.[1] Specifically, by reducing the dimensionality in a way which emphasizes the most significant features, or genes, for overall variance in the data, one can identify these genes as important contributors to various biological processes, disease mechanisms, and potential therapeutic targets, and interpret their roles via downstream analysis.

There are several methods available in the literature for achieving effective dimensionality reduction, categorizable as either linear or nonlinear based on the chosen distance metric. Notably, Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) are examples of linear methods.[4]

Special Issue: Machine Learning in Bio-cheminformatics

Received: July 6, 2023
Published: September 22, 2023

© 2023 American Chemical Society

https://doi.org/10.1021/acs.jcim.3c01023
J. Chem. Inf. Model. 2024, 64, 2405−2420

PCA remains widely employed, which considers the global Euclidean structure of the original data. On the other hand, LDA aims to find a linear combination of features that maximizes class separability and minimizes interclass variance, particularly for multiclass classification problems.[5] Nonlinear methods divide into two subgroups: those preserving global pairwise distances and those maintaining local distances. Kernel PCA falls into the former category, while Laplacian Eigenmaps serve as an example of the latter (Laplacian Eigenmaps will be extensively discussed later).[6,7] As the name suggests, Kernel PCA builds upon traditional PCA. While PCA may not perform well on data sets with complex algebraic/manifold structures that cannot be adequately represented in a linear space, Kernel PCA addresses this limitation by employing kernel functions in a reproducing kernel Hilbert space, thereby accommodating nonlinearity.

Furthermore, there exist several Machine Learning (ML) based dimensionality reduction methods that are popular for high dimensional biomedical data.[8] Filter methods are used to determine the significance of different features in the data and can classify as either univariate or multivariate categories. The univariate filter method first employs a specific criterion to pinpoint the most significant feature rankings, after which each attribute is evaluated and given a distinct rating. The multivariate filter, meanwhile, considers the relationships between various qualities rather than one specific criterion. Wrapper methods rely on classifiers. They operate by choosing a subset of characteristics from a given learning model that yield the best outcomes for ML-based classification. However, dimensionality reduction via PCA better controls overfitting problems that are common with these procedures, and can represent the genes in a more informative space by combining them into Eigen-Genes.[8] There are also a variety of nature inspired algorithms for dimensionality reduction, such as the genetic algorithm, which seeks to mimic natural selection, or the Bat Algorithm, which was inspired by echolocation behavior of microbats.[9,10] We note, however, that generally these nature inspired algorithms require a complete mathematical framework for understanding their robustness, and generally there are issues with repeatability.[8]

While these procedures have found broad applications in the mathematical and statistical sciences, they all possess inherent limitations. In our specific context of gene expression data analysis, we aim to build upon recent advancements in PCA to mitigate various limitations intrinsic to advanced PCA techniques. By leveraging these improvements, we strive to enhance the analysis and interpretation of gene expression data. Although PCA is a widely used procedure for dimensionality reduction, it has several associated weaknesses. These include a lack of interpretability of the principal components due to the dense loadings and issues with class ambiguities. Various methods have been proposed to address these issues. Most notably, Feng et al. proposed Supervised Discriminative Sparse PCA (SDSPCA) to include class information and sparse constraints by introducing a class label matrix and optimizing the $L_{2,1}$ norm.[11]

As we mentioned earlier, an additional desirable aspect of dimensionality reduction is the capability to identify low-dimensional structures that are embedded within higher-dimensional spaces. Graph theory has offered solutions to this issue. In particular, Jiang et al. (2013) incorporated a graph Laplacian term into PCA (gLPCA),[12] while Zhang et al. (2022) combined this approach with SDSPCA to integrate structural information, interpretability, and class information.[13] Additionally, they developed a robust variant by employing the $L_{2,1}$ norm instead of the Frobenius norm in the loss function, and they proposed an iterative optimization algorithm.[12,13] However, it is important to note that the graph Laplacian utilized in this method is only defined at a single scale and lacks topological considerations.

Eckmann introduced topological graphs using simplicial complexes, leading to the development of topological Laplacians on graphs.[14] Topological Laplacians generalize the traditional pairwise graph relations into many-body relations, and their kernel dimensions are identical to those of corresponding homological groups.[15] This can be regarded as a discrete generalization of the Hodge Laplacian on manifolds. Recently, we have introduced persistent Hodge Laplacians on manifolds[16] and persistent combinatorial Laplacians on graphs.[17] The latter is also known as persistent spectral graphs or persistent Laplacians (PLs).[18,19] PLs can be viewed as a generalization of persistent homology.[20−22] The fundamental idea of persistent homology is to represent data as a topological space, such as a simplicial complex. We can then use tools from algebraic topology to reveal the topological features of our data, such as holes and voids. Additionally, persistent homology employs filtration to perform a multiscale analysis of the data and thus creates a family of topological invariants to characterize data in a unique manner. Nevertheless, persistent homology cannot capture the homotopic shape evolution of data. PLs were designed to address this limitation.[17] PLs have both harmonic spectra and nonharmonic spectra. The harmonic spectra recover all the topological invariants from persistent homology, whereas the nonharmonic spectra reveal the homotopic shape evolution. PLs have been employed to facilitate machine learning-assisted protein engineering predictions,[23] accurately forecast future dominant SARS-CoV-2 variants BA.4/BA.5,[24] and predict protein−ligand binding affinity.[25]

Our objective is to introduce PL-enhanced PCA theory (PLPCA). PLPCA can better capture multiscale geometrical structure information than standard graph regularization does. Specifically, PLs enhance our ability to recognize the stability of topological features in our data at multiple scales. This is achieved via filtration, which induces a sequence of simplicial complexes. We can study the spectra of each corresponding Laplacian matrix for each complex in the sequence to extract this topological and geometric information.[19] We will then validate our novel method and demonstrate its performance by microarray data classification after dimensionality reduction.

Our work proceeds as follows: first, we delve into the mathematics behind PCA and its previous relevant improvements, such as sparseness, label information, and graph regularization. Next, we will discuss the tools from PL theory, which we believe may improve the effectiveness of dimensionality reduction. Then, we incorporate these tools to formulate two new PCA methods. The first one, denoted as pLPCA, is a simple persistent Laplacian-enabled PCA model. The persistent Laplacians introduce multiscale nonlinear geometric information to the gLPCA method. The second method, denoted PLPCA, is persistent Laplacian-enhanced robust supervised discriminative sparse PCA, combining sparseness, label information, and robustness with our PL-

https://doi.org/10.1021/acs.jcim.3c01023
J. Chem. Inf. Model. 2024, 64, 2405−2420

enhancement for capturing topological information. Lastly, we validate the proposed methods by a comparison of their results with those in the literature for ten tumor/cancer classifications after dimensionality reduction. Extensive results indicate that the proposed methods are the-state-of-art models for dimensionality reduction. Specifically, we demonstrate that, on average, our procedure outperforms the next best PCA enhancement by 8.01% in accuracy, 7.49% in recall, 8.15% in precision, 11.89% in F1, and 5.14% in AUC. We also note that an improvement in performance was found to be had on all 10 of our tested data sets, which differed widely in their dimensionality, underscoring the comprehensiveness of our method.

2. METHODS

2.1. Principal Component Analysis. Recognizing the importance of dimensionality reduction for tumor classification given gene sequencing data, we formally introduce the notion of PCA. The purpose of PCA is to map M-dimensional data (with N samples), $\mathbf{X} \in \mathbb{R}^{M\times N}$, into a m-dimensional space such that $m < M$. This is accomplished via computing the principal components, which can be used to perform a change of basis into a lower dimensional space. The principal components are obtained via an eigendecomposition of the data covariance matrix, for which the principal components are eigenvectors. Equivalently, principal components are expressed as linear combinations of the original variables which explain the most variance. The goal is then to describe the maximal amount of variation from the original data using a subset of our principal components.[30]

When we normalize the values in our dataset, the optimal m-dimensional space can also be obtained by solving:

$$\min_{\mathbf{U},\mathbf{Q}} \|\mathbf{X} - \mathbf{U}^T\mathbf{Q}\|_F^2, \quad \text{s.t. } \mathbf{U}^T\mathbf{U} = \mathbf{I}_m \tag{1}$$

where $\mathbf{U} = \{\vec{u}_1, ..., \vec{u}_m\}$, $\mathbf{U} \in \mathbb{R}^{m\times M}$, represents the principal directions in order of explained variation and $\left\|\mathbf{X}\right\|_F^2 = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m}|x_{ij}|^2}$ is the Frobenius norm of $\mathbf{X}$. In classical PCA, we take $\mathbf{U}$ to be orthogonal, though we can also apply the orthogonality constraint to $\mathbf{Q} = \{\vec{q}_1, ..., \vec{q}_N\}$, $\mathbf{Q} \in \mathbb{R}^{N\times m}$, which represents the projected data points in our new space.

2.2. Sparseness and Discriminative Information. Traditional PCA requires that the principal components be obtained via a linear combination of all features with nonzero weightings (called loadings.) In the context of gene selection, each feature would then represent a specific gene.[31] There is then an unnecessarily added layer of complexity by enforcing that the loadings be nonzero, as most of the genes would be irrelevant to our analysis and we may wish to focus on only a select few. Thus, the interpretation of our principal components would be aided by the allowance of zero weights via the introduction of sparse PCA. The mathematical formulation of sparse PCA can take several forms, and the inclusion of an $L_{2,1}$ norm penalty term on the projected data matrix is the method chosen for solving SDSPCA.[11] The $L_{2,1}$ norm is defined as $\|\mathbf{Q}\|_{2,1} = \sum_{i=1}^{n}\|\mathbf{q}_i\|_2$. First, we calculate the L_2 norm of each gene, and then compute the L_1 norm of gene-based L_2 norms. Or rather, we sum over each of the tissue samples. In minimizing this term, we are encouraging cells belonging to the same cluster to share a similar representation in the reduced space. Thus, genes which correlate little with the presence of cancer are pushed toward zero, inducing sparseness in our principal components.

One can further build upon this by also introducing discriminative information to reduce class ambiguity. Supervised discriminative sparse PCA (SDSPCA) obtains principal components by introducing supervised label information as well as a sparse constraint.[11] This is realized via the following optimization formula:

$$\min_{\mathbf{U},\mathbf{Q},\mathbf{A}} \|\mathbf{X} - \mathbf{U}\mathbf{Q}^T\|_F^2 + \alpha\|\mathbf{Y} - \mathbf{A}\mathbf{Q}^T\|_F^2 + \beta\|\mathbf{Q}\|_{2,1}, \quad \text{s.t. } \mathbf{Q}^T\mathbf{Q} = \mathbf{I} \tag{2}$$

Here, α and β are scale weights balancing the class label and sparse constraint terms. We arbitrarily initialize matrix $\mathbf{A} \in \mathbb{R}^{c\times m}$ to obtain a solution. Matrix $\mathbf{Y} \in \mathbb{R}^{c\times N}$ represents the one-hot coding class indicator matrix, and c then represents the number of classes in the data. The class indicator matrix consists of 0s and 1's, with the position of element 1 in each column representing the class label. The matrix can be defined as follows, with $s_{i,j}$ representing class labels

$$\mathbf{Y}_{i,j} = \begin{cases} 1, & \text{if } s_{i,j} = i,\ j = 1, ..., n,\ i = 1, ..., c \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

SDSPCA incorporates both label information and sparsity into PCA, with the second and third terms guaranteeing discriminative ability and interpretability, respectively.

2.3. Intrinsic Geometrical Structure. While SDSPCA improves performance relative to traditional PCA, one still wishes to capture and preserve the geometric structure of our gene sequence data during dimensionality reduction, motivating the introduction of graph regularization.[12]

Graph Laplacian-based embedding preserves local geometric relationships while maximizing the smoothness with respect to the intrinsic manifold of the data set in the low embedding space. Equivalently, one wishes to construct a representation for the data sampled from a low-dimensional manifold embedded in the original higher dimensional space. It has been shown that this can be accomplished with graphs with pairwise edges, specifically the Laplacian operator.[32]

The core algorithm proceeds as follows. First, for N points $\vec{x}_1, ..., \vec{x}_N \in \mathbb{R}^M$ we construct a weighted graph W with the set of nodes for each point, and the set of edges connecting neighboring points to one another. We put an edge between points if they are adjacent, which we can choose to determine according to K-nearest-neighbors (KNNs) or some distance threshold.[32] While less geometrically intuitive, the KNN framework tends to be simpler.[33] We then weigh the edges via the Gaussian kernel and can obtain the following matrix:

$$\mathbf{W}_{ij} = \begin{cases} e^{-\|\mathbf{x}_i,\mathbf{x}_j\|^2/\eta} & \text{if } \mathbf{x}_j \in \mathcal{N}_k(\mathbf{x}_i) \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

The matrix defined above, which we associate with our graph, is the adjacency matrix, and it encodes our connectivity information. Also note the introduction of a scale parameter $\eta \in \mathbb{R}$, which defines the geodesic distance, or the width of the Gaussian kernel. Alternatively, we can weight each edge

https://doi.org/10.1021/acs.jcim.3c01023
J. Chem. Inf. Model. 2024, 64, 2405−2420

$\mathbf{W}_{ij} = 1$ if points i and j are connected, but the choice of Gaussian Kernel weighting can be justified.[32] Our goal can now be viewed as mapping the weighted connected graph to a line in a way such that connected, or similar, points remain close together after the embedding. This means choosing $q_i \in \mathbb{R}$ to minimize:

$$\sum_{i,j} \|q_i - q_j\|^2 \mathbf{W}_{ij} \tag{5}$$

It can be shown that this problem reduces to computing eigenvalues and eigenvectors for the generalized eigenvector problem:

$$\mathbf{L}\vec{q} = \lambda \mathbf{D}\vec{q} \tag{6}$$

where $\mathbf{D}$ is a diagonal weight matrix with entries equaling the row sums of the adjacency matrix, or the degree of each vertex.[32] $\mathbf{L} = \mathbf{D} - \mathbf{W}$ is the weighted Laplacian matrix. Let $\vec{q}_1$, ..., $\vec{q}_N$ be the solutions of the eigenvector problem ordered according to their eigenvalues. The image of $\vec{X}_i$ under the embedding into the lower dimensional space $\mathbb{R}^m$ is then given by $\mathbf{Q} = (\vec{q}_1(i), ...,\vec{q}_k(i))$. Thus, we need to minimize:

$$\sum_{i,j} \left\|\mathbf{Q}_i - \mathbf{Q}_j\right\|^2 \mathbf{W}_{ij} = \mathrm{tr}(\mathbf{Q}^T\mathbf{L}\mathbf{Q}),\ \mathbf{Q}^T\mathbf{Q} = \mathbf{I} \tag{7}$$

Thus, given our data matrix $\mathbf{X}$ and weighted graph $\mathbf{W}$, we seek a low dimensional representation that is regularized by the data manifold encoded in $\mathbf{W}$. Because $\mathbf{Q}$ in PCA and Laplacian embedding serve the same purpose, we set them equal and combine eqs 1 and 7, giving rise to graph Laplacian PCA (gLPCA), which is implemented according to the following optimization formula:[12]

$$\min_{\mathbf{U},\mathbf{Q}} \|\mathbf{X} - \mathbf{U}\mathbf{Q}^T\|_F^2 + \gamma \mathrm{Tr}(\mathbf{Q}^T\mathbf{L}\mathbf{Q}), \quad \text{s.t. } \mathbf{Q}^T\mathbf{Q} = \mathbf{I} \tag{8}$$

where the γ parameter scales the geometrical structure capture. Next, Zhang et al. combined this methodology with SDSPCA to incorporate sparseness, structural information, and discriminative information into one procedure. This new method, called Laplacian Supervised Discriminative Sparse PCA is solved for by combining Equations 7 and 2:[13]

$$\min_{\mathbf{U},\mathbf{Q},\mathbf{A}} \|\mathbf{X} - \mathbf{U}\mathbf{Q}^T\|_F^2 + \alpha\|\mathbf{Y} - \mathbf{A}\mathbf{Q}^T\|_F^2 + \beta\|\mathbf{Q}\|_{2,1} + \gamma \mathrm{Tr}(\mathbf{Q}^T\mathbf{L}\mathbf{Q}), \quad \text{s.t. } \mathbf{Q}^T\mathbf{Q} = \mathbf{I} \tag{9}$$

However, Zhang et al. also noted that the Frobenius norm regularization in LSDSPCA is sensitive to outliers. For robustness, one can replace the Frobenius norm regularization with $L_{2,1}$-norm regularization, which results in Robust Laplacian Supervised Discriminative Sparse PCA (RLSDSPCA):[13]

$$\min_{\mathbf{U},\mathbf{Q},\mathbf{A}} \|\mathbf{X} - \mathbf{U}\mathbf{Q}^T\|_{2,1} + \alpha\|\mathbf{Y} - \mathbf{A}\mathbf{Q}^T\|_F^2 + \beta\|\mathbf{Q}\|_{2,1} + \gamma \mathrm{Tr}(\mathbf{Q}^T\mathbf{L}\mathbf{Q}), \quad \text{s.t. } \mathbf{Q}^T\mathbf{Q} = \mathbf{I} \tag{10}$$

Having integrated robustness, interpretability, class information, and geometric structural information, we now turn to replacing the graph regularization with persistent spectral graphs[17] to introduce multiscale analysis.

2.4. Persistent Laplacians. Motivated by the success of persistent homology and multiscale graphs in analyzing biomolecular data, we turn to persistent spectral graph theory to enhance our ability to capture the multiscale geometric structure.[34] Like persistent homology, persistent spectral graph theory tracks the birth and death of topological features of a dataset as they change over scales.[16,18] We carry out this analysis by using the filtration procedure on our data set to construct a family of geometric configurations, or simplicial complexes.[17] We then can study the topological properties of each configuration by its corresponding Laplacian matrix. The topological persistence can be studied through multiple successive configurations.

We first must introduce the notion of a simplex. A 0-simplex is a node, a 1-simplex is an edge, a 2-simplex is a triangle, a 3-simplex is a tetrahedron, and so on. Generally, we consider q-simplices which we label σ_q. A simplicial complex is a way of approximating a topological space by gluing together lower-dimensional simplices in a specific way. More formally, a simplicial complex K is a collection of simplices such that

1. If $\sigma_q \in K$ and σ_p is a face of σ_q then $\sigma_p \in K$.
2. The nonempty intersection of any two simplices is a face of both simplices.

A q-chain is then a formal sum of q-simplices in a simplicial complex K with coefficients in $\mathbb{Z}_2$. The set of all q-chains has a basis of the set of q-simplicies in K. This set forms a finitely generated free Abelian group $C_q(K)$. We then define the boundary operator to be a group homomorphism that relates the chain groups, ∂_q: $C_q(K) \to C_{q-1}(K)$.[19]

We denote the q-simplex by its vertices v_i: $\sigma_q = [v_0, v_1, ..., v_q]$. The boundary operator is then defined as

$$\partial_q\sigma_q := \sum_{i=0}^{q} (-1)^i \sigma_{q-1}^i \tag{11}$$

where $\sigma_{q-1}^i = [v_0,\cdots,\hat{v}_i,\cdots,v_q]$ is the $(q-1)$ simplex with v_i removed. The sequence of chain groups connected by boundary operators is then called a chain complex:

$$... \xrightarrow{\partial_{q+2}} C_{q+1}(K) \xrightarrow{\partial_{q+1}} C_q(K) \xrightarrow{\partial_q} ...$$

The chain complex associated with a simplicial complex defines the q^{th} homology group $H_q = \ker \partial_q/\mathrm{Im}\ \partial_q$. The dimension of H_q is then the q^{th} Betti number β_q, which captures the number of q-dimensional holes in the simplicial complex. We can also define a dual chain complex through the adjoint operator of ∂_q defined on the dual spaces $C^q(K) \cong C_q{}^*(K)$. The coboundary operator $\partial_q{}^*$: $C^{q-1}(K) \to C^q(K)$ is defined as

$$\partial^*\omega^{q-1}(c_q) \equiv \omega^{q-1}(\partial c_q) \tag{12}$$

where $\omega^{q-1} \in C^{q-1}(K)$ is a $(q-1)$ cochain, or a homomorphism mapping a chain to the coefficient group, and $c_q \in C_q(K)$ is a q-chain. The homology of the dual chain complex is referred to as the cohomology. Now we can define the q-combinatorial Laplacian operator, Δ_q: $C^q(K) \to C^q(K)$ as

$$\Delta_q := \partial_{q+1}\partial_{q+1}^* + \partial_q^*\partial_q \tag{13}$$

Now, denote the matrix representation of the q-boundary operator with respect to the standard basis for $C_q(K)$ and $C_{q-1}(K)$ as $\mathcal{B}_q$ and the matrix representation of the q-coboundary operator as $\mathcal{B}_q^T$. We then can define the matrix

https://doi.org/10.1021/acs.jcim.3c01023
J. Chem. Inf. Model. 2024, 64, 2405−2420

representation of the qth-order combinatorial Laplacian operator as $\mathcal{L}_q$:

$$\mathcal{L}_q = \mathcal{B}_{q+1}\mathcal{B}_{q+1}^T + \mathcal{B}_q^T\mathcal{B}_q \tag{14}$$

It is well-known that β_q is also the multiplicity of zero in the spectrum of the Laplacian matrix which corresponds to that simplicial complex (the harmonic spectrum). Specifically:

β_0 = Number of connected components in K

β_1 = Number of holes in K

β_2 = Number of two-dimensional voids in K

and so on. The nonharmonic spectrum also contains other topological and shape information.

However, a single simplicial complex offers very limited information to our understanding of the structure of our data. We then could consider the creation of a sequence of simplicial complexes which we induce by a filtration parameter: $\{\theta\} = K_0 \subseteq K_1 \subseteq \cdots \subseteq K_p = K$. The filtration is illustrated in Figure 1.

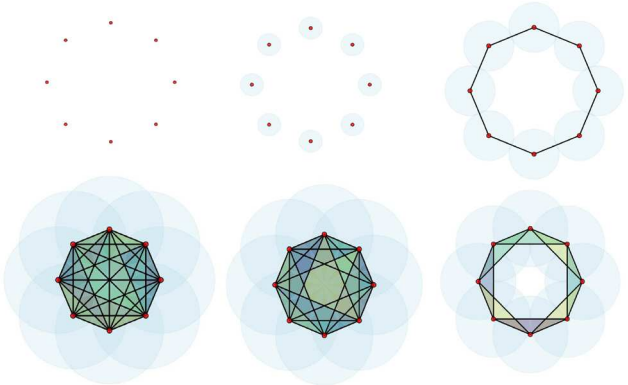


Figure 1. Illustration of the filtration of a point cloud by varying a distance threshold. The Vietoris-Rips complex is used.

For each subcomplex K_t we can denote its chain group to be $C_q(K_t)$, and the q-boundary operator ∂_q^t: $C_q(K_t) \to C_{q-1}(K_t)$. By convention, we define $C_q(K_t) = \{0\}$ for $q < 0$ and the q-boundary operator to then be the zero map. We then have

$$\partial_q^t \sigma_q = \sum_i^q (-1)^i \sigma_{q-1}^i, \quad \forall\, \sigma_q \in K_t \tag{15}$$

Which is essentially the same construction as before. Likewise, the adjoint operator of ∂_q^t is the coboundary operator ∂_q^{t*}: $C^{q-1}(K_t) \to C^q(K_t)$, which we regard as a map from $C_{q-1}(K_t) \to C_q(K_t)$ through the isomorphism between cochain and chain groups. We can then define a sequence of chain complexes.

Next, we introduce persistence to the Laplacian spectra. Define the subset of C_q^{t+p} whose boundary is in C_{q-1}^t as $\mathbb{C}_q^{t,p}$, assuming the natural inclusion map from $C_{q-1}^t \to C_{q-1}^{t+p}$.

$$\mathbb{C}_q^{t,p} := \{\beta \in C_q^{t+p} | \partial_q^{t,p}(\beta) \in C_{q-1}^t\} \tag{16}$$

On this subset, one may define the p-persistent q-boundary operator denoted by $\hat{\partial}_q^{t,p}$: $\mathbb{C}_q^{t,p} \to C_{q-1}^t$, and corresponding adjoint operator $(\hat{\partial}_q^{t,p})^*$: $C_{q-1}^t \to \mathbb{C}_q^{t,p}$, as before. The q-order p-persistent Laplacian operator is then $\Delta_q^{t,p}$: $C_q^t \to C_q^t$:

$$\Delta_q^{t,p} := \hat{\partial}_{q+1}^{t,p}(\hat{\partial}_{q+1}^{t,p})^* + (\hat{\partial}_q^t)^*\hat{\partial}_q^t \tag{17}$$

And the matrix representation in simplicial basis is again:

$$\mathcal{L}_q^{t,p} := \mathcal{B}_{q+1}^{t,p}(\mathcal{B}_{q+1}^{t,p})^T + (\mathcal{B}_q^t)^T\mathcal{B}_q^t \tag{18}$$

We may again recognize the multiplicity of zero in the spectrum of $\mathcal{L}_q^{t,p}$ as the qth order p-persistent Betti number $\beta_q^{t,p}$ which counts the number of (independent) q-dimensional holes in K_t that still exists in K_{t+p}. We can then see how the qth-order Laplacian is actually just a special case of the qth-order 0-persistent Laplacian at a simplicial complex K_t. In other words, the spectrum of $\mathcal{L}_q^{t,0}$ is simply associated with a snapshot of the filtration at some step t.[19]

We can capture a more thorough view of the spatial features of our data by focusing on the 0-persistent Laplacian. Specifically, by inducing a family of subgraphs through the varying of a distance threshold ϵ, as seen in Figure 1. This is known as the Vietoris Rips Complex. The edges of the complex connect pairs of vertices that are within our distance threshold ϵ, which we vary to construct our sequence of complexes. Alternatively, we connect vertices according to K-nearest-neighbors and then weight the edges according to some notion of distance, such as the Gaussian Kernel.[32] We then filter out edges by increasing our ϵ value. In the next section we will see a convenient method for computation, and also a description of the PLPCA procedure.

2.5. PLPCA. Now, the generation of the Vietoris Rips Complex can be achieved through implementing a filtration procedure on our weighted Laplacian matrix based on an increasing threshold. Observe:

$$\mathbf{L} = (l_{ij}),\ l_{ij} = \begin{cases} l_{ij},\ i \neq j,\ i, j = 1, \ldots, n \\ l_{ii} = -\sum_{j=1}^{n} l_{ij} \end{cases} \tag{19}$$

For $i \neq j$, let $l_{\max} = \max(l_{ij})$, $l_{\min} = \min(l_{ij})$, $d = l_{\max} - l_{\min}$. Set the t^{th} persistent Laplacian $\mathbf{L}^t$, t = 1,⋯,p:

$$\mathbf{L}^t = (l_{ij}^t),\ l_{ij}^t = \begin{cases} 0, & \text{if } l_{ij} \leq (t/p)d + l_{\min} \\ -1, & \text{otherwise} \end{cases} \tag{20}$$

$$l_{ii}^t = -\sum_{j=1}^{n} l_{ij}^t \tag{21}$$

This procedure results in the generation of a family of persistent Laplacians derived from our weighted graph Laplacian. However, due to the Gaussian Kernel weighting of the edges, it is more appropriate to filter values above the threshold rather than below. This is because more negative values indicate connections between data points that are closer together, or more similar, which are the features we want to emphasize. To consolidate this family of graphs into a single term, we assign weights to each of them and then sum them together to construct an accumulated spectral graph, **PL**.

$$\mathbf{PL} := \sum_{t=1}^{p} \zeta_t \mathbf{L}^t \tag{22}$$

https://doi.org/10.1021/acs.jcim.3c01023
J. Chem. Inf. Model. 2024, 64, 2405−2420

Incorporating this new term into the gLPCA algorithm in place of the graph Laplacian should better retain geometrical structure information by emphasizing features that are persistent at multiple scales and providing a more thorough spatial view. The optimal space is now obtained via

$$\min_{\mathbf{U},\mathbf{Q}} \|\mathbf{X} - \mathbf{U}\mathbf{Q}^T\|_F^2 + \gamma \mathrm{Tr}(\mathbf{Q}^T(\mathbf{PL})\mathbf{Q}), \quad \text{s.t.} \mathbf{Q}^T\mathbf{Q} = \mathbf{I} \tag{23}$$

We call this method pLPCA. We can then combine this procedure with RLSDSPCA to formulate the most optimal procedure, which, for convenience, we refer to simply as PLPCA, and which we solve for via

$$\min_{\mathbf{U},\mathbf{Q},\mathbf{A}} \|\mathbf{X} - \mathbf{U}\mathbf{Q}^T\|_{2,1} + \alpha\|\mathbf{Y} - \mathbf{A}\mathbf{Q}^T\|_F^2 + \beta\|\mathbf{Q}\|_{2,1} + \gamma \mathrm{Tr}(\mathbf{Q}^T(\mathbf{PL})\mathbf{Q}), \quad \text{s.t.} \mathbf{Q}^T\mathbf{Q} = \mathbf{I} \tag{24}$$

We should however recognize the issues with PLPCA regarding the inclusion of a class indicator matrix, most notably in the context of K-Means clustering via PCA.[35] In this case, the label information would presumably not be known beforehand and therefore pLPCA would be the preferable method, although slightly less robust.

Figure 2 provides an overview of the PLPCA framework for both tumor classification and characteristic gene selection.[13]

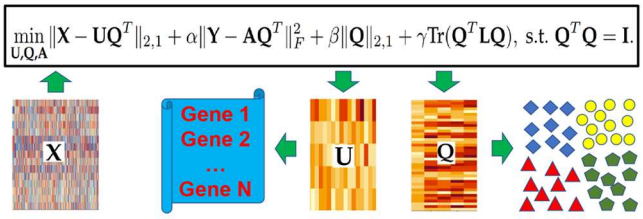


Figure 2. Outline of the PLPCA procedure for dimensionality reduction, feature selection, and classification.

The process begins with the input gene expression matrix $\mathbf{X}$, and dimensionality reduction is performed using eq 24. The resulting outputs are the projected data matrix $\mathbf{Q}$ and the principal directions matrix $\mathbf{U}$. The projected data matrix can then be utilized for tumor classification. Previous studies have also utilized the projected data matrix for feature selection, as it contains valuable information about each gene's contribution to the overall variance of the data.[13]

Obtaining a closed form solution for PLPCA is difficult, but we can iteratively optimize the model. The optimization of PLPCA can be performed using the alternating direction method of multipliers (ADMM) algorithm. ADMM is a variant of the augmented Lagrangian method, which is employed to solve constrained optimization problems.[13] The augmented Lagrangian method transforms constrained optimization problems into a series of unconstrained problems by introducing a penalty term to the cost function. It also incorporates an additional term resembling a Lagrangian multiplier.[36] The penalty function approach solves this problem iteratively by updating each parameter at each step.[37]

ADMM, meanwhile, is a method which uses partial updates for dual variables.[38] Consider the generic problem:

$$\min_{\mathbf{Q}} f(\mathbf{Q}) + g(\mathbf{U}) \tag{25}$$

which is equivalent to

$$\min_{\mathbf{Q},\mathbf{U}} f(\mathbf{Q}) + g(\mathbf{U}), \quad \text{s.t.} \mathbf{Q}\mathbf{Q}^T = \mathbf{I} \tag{26}$$

The ADMM technique allows us to approximately solve this problem by first solving for $\mathbf{Q}$ with $\mathbf{U}$ fixed via the augmented Lagrangian method, and then vice versa. Specific to our problem, this approach can be taken for approximately solving for the optimal $\mathbf{Q}$, $\mathbf{U}$, $\mathbf{A}$, $\mathbf{E}$, $\mathbf{G}$, and $\mathbf{C}$ matrices in our algorithm. This is the method chosen for optimizing RLSDSPCA, and we implemented it as well for PLPCA.[13]

The optimization procedure for RLSDSPCA is described in detail in the paper by Zhang et al.[13] Their study demonstrated that the RLSDSPCA objective function exhibits a monotonically decreasing trend with each iteration. The same proof applies to PLPCA as well, with the substitution of the persistent Laplacian for the Laplacian term. In Algorithm 1, we present a summary of the updated optimization algorithm for PLPCA.

Algorithm 1 PLPCA procedure

Input:

Data matrix $\mathbf{X} \in \mathbb{R}^{M\times N}$; OHE Label Matrix $\mathbf{Y} \in \mathbb{R}^{c\times N}$;
Weight Parameters: α, β, γ; Number of Subspace Dimensions: k;
Convergence Parameter: Θ, Number of Iterations: *MaxIter*;
Weight Parameters ζ_i $(i = 1, ..., p)$; Number of Subgraphs: p

Output:

Principal Directions Matrix $\mathbf{U}$ and Projected Data Matrix $\mathbf{Q}$

Initialize:

Initialize Matrices $\mathbf{G}$, $\mathbf{E}$, $\mathbf{C}$ to identity matrix;
Randomly intialize matrices $\mathbf{A}$, $\mathbf{Q}_1$ (An auxillary matrix to check convergence);
Construct Weighted Adjacency Matrix $\mathbf{W}$ according to K-Nearest-Neighbors;
Compute Weighted Laplacian $\mathbf{L}$;
Compute family of subgraphs $\mathbf{L}^i$;
Construct Persistent Laplacian $\mathbf{PL}$;
Initialize $\mu = 1$

ADMM

for i=1 to *MaxIter* **do**
  Compute $\mathbf{Q}$
  Compute $\mathbf{U}$
  Compute $\mathbf{A}$
  Compute $\mathbf{E}$
  Compute $\mathbf{G}$
  Compute $\mathbf{C}$
  Compute μ
  Check the Convergence Condition:
  if $i > 1$ and $\|\mathbf{Q} - \mathbf{Q}_1\|_{2,1} < \Theta$ **then**
    Break
  end if
end for
Let $\mathbf{Q}_1 = \mathbf{Q}$

After obtaining the optimal dimensionality reduction, the classification of cancerous tumors begins by normalizing the gene expression data and randomly splitting it into training and testing sets. The testing set accounts for 20% of the data, while the training set constitutes the remaining 80%. To mitigate the impact of data distribution, we employed a 5-fold cross-validation approach. The classification accuracy was calculated as the average performance over five repetitions. For classification purposes, we utilized the K-nearest-neighbors algorithm.[39] The mean accuracy of the classification was recorded for subspace dimensions in the range {100, 95, ..., 5, 1}.

3. RESULTS AND DISCUSSION

3.1. Data Summary. Previous studies have utilized benchmark data sets obtained from The Cancer Genome Atlas, which is a project aimed at cataloging genetic mutations that contribute to cancer through genome sequencing.[26−29] The Cancer Genome Atlas specifically focuses on 33 cancer types that fulfill the following criteria: poor prognosis, significant public health impact, and availability of samples.

In our study, we expand upon this by focusing on the two standard data sets tested in the literature: the MultiSource data set and the COAD data set,[26−29] as well as eight

https://doi.org/10.1021/acs.jcim.3c01023
J. Chem. Inf. Model. 2024, 64, 2405−2420

Table 1. Datasets Summary

data set	category	no. of samples	no. of features	cancer description
MultiSource	CHOL	36	20502	cholangiocarcinoma
	HNCC	398	20502	head and neck squamous cell carcinoma
	PAAD	176	20502	pancreatic adenocarcinoma
	normal	33	20502	normal tissue
COAD	COAD	262	20502	colon adenocarcinoma
	normal	19	20502	normal tissue
GSE44076	normal	98	49385	normal tissue
	cancer	98	49385	colon adenocarcinoma
GSE14020	lung	4	54675	lung metastasizes from breast cancer
	bone	10	54675	bone metastasizes from breast cancer
	brain	15	54675	brain metastasizes from breast cancer
GSE39582	normal	19	54675	normal tissue
	cancer	566	54675	colorectal adenocarcinoma
GSE18842	normal	45	54675	normal tissue
	cancer	46	54675	lung cancer
GSE35988	normal	28	92529	normal tissue
	cancer	94	92529	prostate cancer
GSE29272	normal	134	22283	normal tissue
	cancer	134	22283	gastric cancers
GSE21034	normal	29	43418	normal tissue
	cancer	150	43418	prostate cancer
GSE28735	normal	45	28868	normal tissue
	cancer	45	28868	pancreatic cancer

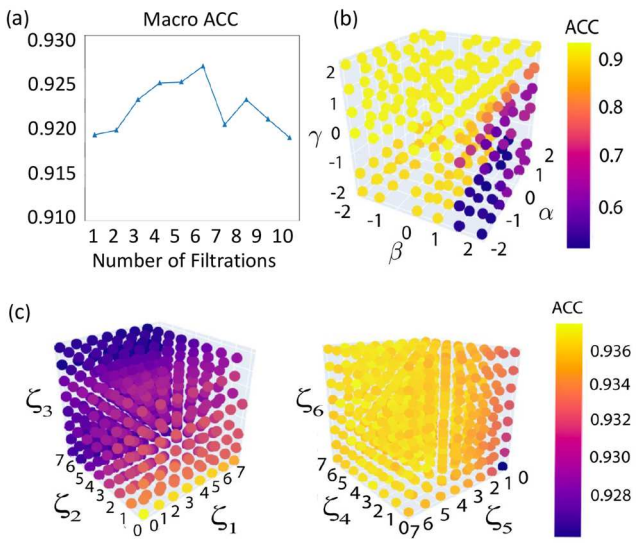


Figure 3. (a) Effect of different numbers of filtrations on classification accuracy. The *x*-axis represents the number of filtrations and the *y*-axis represents the classification accuracy. (b) Effect of different close, middle, and long-range connectivity weights on mean macro-ACC. (c) Effect of different regularization scale weights on classification accuracy. The three coordinates represent each scale weight; color represents accuracy for each parameter combination. Axis ticks denote powers of 10.

additional datasets obtained from the Gene Expression Omnibus. The Gene Expression Omnibus is a public repository that archives high-throughput functional genomics data submitted by members of the research community.

The MultiSource data set comprises normal tissue samples and three different cancer types along with their corresponding gene expression data. The included cancer types are cholangiocarcinoma (CHOL), head and neck squamous cell carcinoma (HNSCC), and pancreatic adenocarcinoma (PAAD). Specifically, the CHOL data set consists of 45 samples (9 normal tissue samples and 36 cancer samples), the HNSCC data set contains 418 samples (20 normal and 398 cancer samples), and the PAAD data set consists of 180 samples (4 normal and 176 cancer samples). Each data set encompasses 20,502 genes. Meanwhile, the COAD data set consists of 281 samples (19 normal samples and 262 colon adenocarcinoma samples) and also spans 20,502 genes.

A description of each data set obtained from the Gene Expression Omnibus, including the number of features, samples, accession numbers, and cancer types, can be found in Table 1 along with a description of the COAD and MultiSource datasets.

To further validate the proposed methods, we consider eight cancer data sets besides the two standard TCGA datasets utilized in previous works. The dimensions of these datasets reaches as high as 92,529 genes, which is especially challenging to manage, showcasing the robustness of our

Table 2. Comparison of pLPCA and gLPCA Performance on the COAD Dataset

method	mean ACC	mean macro-REC	mean macro-PRE	mean macro-F1	macro-AUC
gLPCA[12,13]	0.9777	0.9463	0.9138	0.9249	0.9470
gLPCA[a12]	0.9756	0.9429	0.8841	0.9002	0.9429
pLPCA	**0.9788**	**0.9450**	**0.8996**	**0.9115**	**0.9450**

[a]Reproduced in the present work.

https://doi.org/10.1021/acs.jcim.3c01023
J. Chem. Inf. Model. 2024, 64, 2405−2420

Table 3. Comparison of pLPCA and gLPCA Performance on the MultiSource Dataset

method	mean ACC	mean macro-REC	mean macro-PRE	mean macro-F1	macro-AUC
gLPCA[12,13]	0.9139	0.8147	0.8768	0.8316	0.8909
pLPCA	**0.9267**	**0.8318**	**0.8857**	**0.8471**	**0.8991**

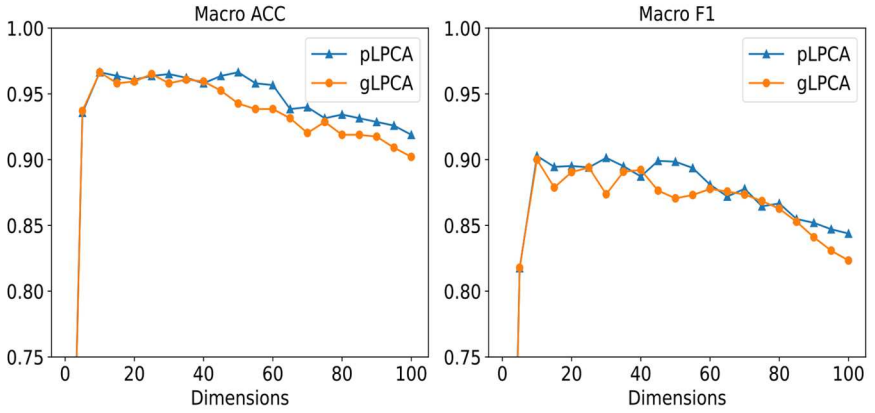


Figure 4. Macro-ACC and -F1 score on the MultiSource data set across different reduced dimensions (gLPCA vs pLPCA).

Table 4. Comparison of PLPCA and RLSDSPCA Performance on the COAD Dataset

method	mean ACC	mean macro-REC	mean macro-PRE	mean macro-F1	macro-AUC
RLSDSPCA[13]	0.9875	0.9614	0.9504	0.9530	0.9614
PLPCA	**0.9886**	**0.9680**	**0.9517**	**0.9578**	**0.9680**

Table 5. Comparison of PLPCA and RLSDSPCA Performance on the MultiSource Dataset

method	mean ACC	mean macro-REC	mean macro-PRE	mean macro-F1	macro-AUC
RLSDSPCA[13]	0.9273	0.8343	0.8972	0.8527	0.9024
PLPCA	**0.9371**	**0.8393**	**0.9089**	**0.8619**	**0.9065**

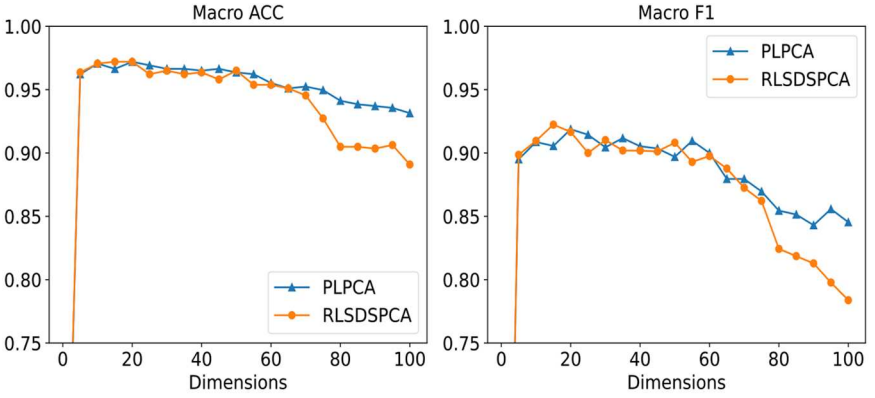


Figure 5. Macro ACC and F1 score on the MultiSource data set across different reduced dimensions (RLSDSPCA vs PLPCA).

proposed procedure. Gene expression matrices and class labels were obtained from GEO's Series Matrix Files. Duplicate genes were dropped from the analysis, and values were log transformed. Then normalization was carried out to have zero mean and a standard deviation of one.

Once again, we emphasize the significant imbalance between the number of samples and the number of features, as well as the number of samples among different classes in our gene expression data, which underscores the performance of our dimensionality reduction.

3.2. Parameter Analysis. Note with this method the introduction of several new hyper-parameters which we must optimize. Namely, each of the weights ζ_t as well as the number of subgraphs must be chosen. We imposed the constraint $\sum_{t=1}^{p} \zeta_t = 1$ and performed grid search to understand whether we should favor long-range, middle-range, or close-range connectivity. Ultimately, we achieved the best results with six filtrations, or a six-scale scheme ($p = 6$). Figure 3a shows the variations in mean macro-ACC as we vary the number of filtrations. For $p < 6$, there is not enough additional information incorporated to substantially improve performance, while for $p > 6$, too great a number of hyper-parameters to choose could have somewhat hurt performance. It is also possible that the additional filtrations hurt the performance

https://doi.org/10.1021/acs.jcim.3c01023
J. Chem. Inf. Model. 2024, 64, 2405−2420

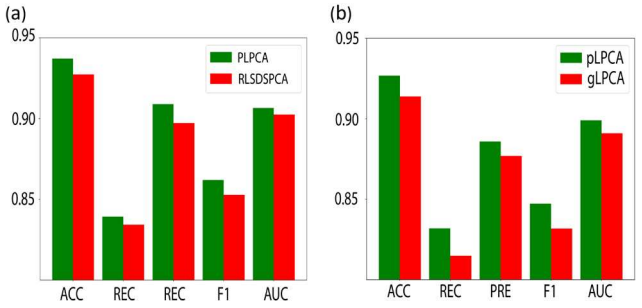


Figure 6. Comparison of each of the evaluation metrics (MultiSource data set): (a) between RLSDSPCA and PLPCA; (b) between pLPCA and gLPCA.

when mapping into lower subspace dimensions, thereby hurting the performance on average.

For the MultiSource data set, optimal results were achieved by emphasizing the long-range connections, while for the COAD data set, the connectivities were all roughly equally weighted, with the close range and long-range connectivity being slightly favored. Figure 3b displays the optimization results for different combinations of weights in $\{\zeta_t\}$. Higher values of ζ_t correspond to placing greater emphasis on the connectivity at that scale. For the other data sets obtained from GEO, optimal parameter values can be found in the PLPCA GitHub.

We tested ζ_t values ranging from 0 to 10, and then scaled them to satisfy the constraint $\sum_{t=1}^{p} \zeta_t = 1$. The difference was then factorized into the γ parameter, which we test with each different combination of ζ_t. For PLPCA on the COAD data set, $\{\zeta_t\} = \{0.5, 3, 1, 2, 2, 1\}$. For pLPCA on the COAD data set, $\{\zeta_t\} = \{2, 3, 0, 0, 2, 1\}$. For PLPCA on the MultiSource data set, $\{\zeta_t\} = \{0.5, 0, 0, 3, 0, 6\}$. For pLPCA on the MultiSource data set, $\{\zeta_t\} = \{0.5, 0, 0, 3, 2, 6\}$. After optimizing $\{\zeta_t\}$ and γ, we perform a grid search again to revise our choices of α and β as shown in Figure 3c.

We revised our choice of parameters to $\alpha = 10^4$, $\beta = 0.5$, and $\gamma = 10^1$ for the MultiSource data set and $\alpha = 10^5$, $\beta = 0.5$, and $\gamma = 1000$ for the COAD data set and pLPCA. For PLPCA, however, better results were obtained using $\gamma = 10^4$ in both cases. We can now examine the efficacy of our model by testing on several benchmark data sets.

3.3. Evaluation Metrics. Our study demonstrates that the incorporation of persistent graph regularization enhances classification performance after dimensionality reduction, surpassing the achievements of other state-of-the-art methods such as RLSDSPCA. We summarize the outcomes of our analysis conducted on the COAD and MultiSource data sets, sourced from the Cancer Genome Atlas, multiple additional data sets obtained from the Gene Expression Omnibus, as well as several simulated outlier data sets obtained from the RLSDSPCA GitHub repository.[13]

First, we discuss the evaluation metrics used to measure performance. While accuracy is a commonly employed metric, in the context of cancer diagnosis, additional emphasis is often placed on the F1-score. The F1-score represents the harmonic mean of precision and recall, providing a balanced assessment of a classifier's performance.

$$\text{Recall} = \text{True Positive}/(\text{True Positive} + \text{False Negative}) \quad (27)$$

$$\text{Precision} = \text{True Positive}/(\text{True Positive} + \text{False Positive}) \quad (28)$$

$$\text{F1-Score} = 2 \times (\text{Precision} \times \text{Recall})/(\text{Precision} + \text{Recall}) \quad (29)$$

By considering the cost of false negatives in our classification task, we acknowledge the significance of accurately identifying cases of a potentially life-threatening disease. The F1-score is particularly relevant in this context as it takes into account both precision and recall, making it more robust to class imbalances within the data. In our case, there are noticeable imbalances, particularly in the MultiSource data set.

Given that our data consist of multiple categories, the evaluation criterion we employ is the mean of each category indicator. This evaluation approach is commonly known as a macrometric, where performance measures are calculated for each category individually and then averaged to obtain an overall score.

$$\text{Macro-Recall} = 1/c \times \sum_{i=1}^{c} \text{Recall}_i \quad (30)$$

Table 6. Comparison of PLPCA and RLSDSPCA on Datasets Obtained from Gene Expression Omnibus

data set	method	mean ACC	mean macro-REC	mean macro-PRE	mean macro-F1	macro-AUC
GSE44076	**PLPCA**	**0.8507**	**0.8555**	**0.8762**	**0.8459**	**0.8210**
	RLSDSPCA[13]	0.6121	0.6236	0.5712	0.5198	0.5383
GSE14020	**PLPCA**	**0.8228**	**0.7240**	**0.6679**	**0.6709**	**0.6811**
	RLSDSPCA[13]	0.8152	0.7219	0.6692	0.6681	0.6790
GSE39582	**PLPCA**	**0.9960**	**0.9749**	**0.9520**	**0.9603**	**0.9700**
	RLSDSPCA[13]	0.9959	0.9748	0.9473	0.9573	0.9650
GSE18842	**PLPCA**	**0.8688**	**0.8648**	**0.8864**	**0.8559**	**0.8500**
	RLSDSPCA[13]	0.8166	0.8041	0.8542	0.7837	0.8401
GSE35988	**PLPCA**	**0.7159**	**0.7640**	**0.7229**	**0.6710**	**0.7111**
	RLSDSPCA[13]	0.7054	0.7589	0.7108	0.6616	0.7105
GSE29272	**PLPCA**	**0.7925**	**0.7901**	**0.8289**	**0.7814**	**0.7923**
	RLSDSPCA[13]	0.6738	0.6676	0.7445	0.6038	0.7601
GSE21034	**PLPCA**	**0.8214**	**0.7356**	**0.7044**	**0.6989**	**0.7283**
	RLSDSPCA[13]	0.8190	0.7335	0.7026	0.6968	0.7236
GSE28735	**PLPCA**	**0.6476**	**0.6558**	**0.6599**	**0.6416**	**0.6449**
	RLSDSPCA[13]	0.5990	0.6105	0.6169	0.5871	0.6328

https://doi.org/10.1021/acs.jcim.3c01023
J. Chem. Inf. Model. 2024, 64, 2405−2420

Table 7. Comparison of PLPCA and Other Notable Methods Performances on the MultiSource Data

method	mean ACC	mean macro-REC	mean macro-PRE	mean macro-F1	macro-AUC
PLPCA	**0.9371**	**0.8393**	**0.9089**	**0.8619**	**0.9065**
pLPCA	0.9267	0.8318	0.8857	0.8471	**0.8991**
RLSDSPCA[13]	0.9273	0.8343	0.8972	0.8527	0.9024
SDSPCA[11,13]	0.9124	0.8144	0.8917	0.8333	0.8891
RgLPCA[12,13]	0.9197	0.8210	0.8748	0.8353	0.8945
gLSPCA[13,41]	0.9195	0.8148	0.8769	0.8318	0.8910
gLPCA[12,13]	0.9193	0.8147	0.8768	0.8316	0.8909
PCA[13,42]	0.9108	0.7957	0.8389	0.8025	0.8726

Table 8. Comparison of PLPCA and Other Notable Methods Performances on the COAD Data

method	mean ACC	mean macro-REC	mean macro-PRE	mean macro-F1	macro-AUC
PLPCA	**0.9886**	**0.9680**	**0.9517**	**0.9578**	**0.9680**
pLPCA	0.9788	0.9450	0.8996	**0.9115**	0.9450
RLSDSPCA[a13]	0.9797	0.9443	0.9081	0.9149	0.9444
SDSPCA[a11]	0.9643	0.9533	0.8740	0.8918	0.9533
RgLPCA[a12]	0.9734	0.9015	0.8990	0.8750	0.8990
gLSPCA[a41]	0.9761	0.9250	0.8969	0.8958	0.9250
gLPCA[a12]	0.9756	0.9429	0.8841	0.9002	0.9429
PCA[a42]	0.9593	0.8988	0.8599	0.8799	0.8980

[a]Reproduced in the present work.

Table 9. Comparison of PLPCA and Other Methods on Datasets Obtained from Gene Expression Omnibus

data set	method	mean ACC	mean macro-REC	mean macro-PRE	mean macro-F1	macro-AUC
GSE44076	**PLPCA**	**0.8507**	**0.8555**	**0.8762**	**0.8459**	**0.8210**
	pLPCA	0.6223	0.6340	0.5940	0.5312	0.6072
	RLSDSPCA[13]	0.6121	0.6236	0.5712	0.5198	0.5883
	SDSPCA[11,13]	0.6161	0.6273	0.5767	0.5225	0.5989
	RgLPCA[12]	0.6150	0.6263	0.5867	0.5231	0.5834
	gLSPCA[41]	0.6216	0.6332	0.5936	0.5299	0.5977
	gLPCA[12]	0.6219	0.6334	0.5938	0.5301	0.6026
	PCA[42]	0.6209	0.6324	0.5887	0.5286	0.5910
GSE21034	**PLPCA**	**0.8214**	**0.7356**	**0.7044**	**0.6989**	**0.7283**
	pLPCA	0.8185	0.7333	0.7022	0.6963	0.7230
	RLSDSPCA[13]	0.8190	0.7335	0.7026	0.6968	0.7236
	SDSPCA[11,13]	0.8185	0.7333	0.7022	0.6963	0.7230
	RgLPCA[12]	0.8190	0.7335	0.7026	0.6968	0.7236
	gLSPCA[41]	0.5999	0.4577	0.4571	0.4367	0.4499
	gLPCA[12]	0.8185	0.7333	0.7022	0.6963	0.7230
	PCA[42]	0.8185	0.7333	0.7022	0.6963	0.7230
GSE28735	**PLPCA**	**0.6476**	**0.6558**	**0.6599**	**0.6416**	**0.6449**
	pLPCA	0.5809	0.5905	0.5990	0.5669	0.6211
	RLSDSPCA[13]	0.5990	0.6105	0.6169	0.5871	0.6328
	SDSPCA[11,13]	0.5914	0.6026	0.6111	0.5787	0.6309
	RgLPCA[12]	0.5866	0.5971	0.6035	0.5738	0.6287
	gLSPCA[41]	0.5666	0.5699	0.5736	0.5566	0.5777
	gLPCA[12]	0.5806	0.5907	0.5966	0.5684	0.6205
	PCA[42]	0.5803	0.5905	0.5959	0.5678	0.6200

$$\text{Macro-Precision} = 1/c \times \sum_{i=1}^{c} \text{Precision}_i \tag{31}$$

$$\text{Macro-F1} = 2 \times (\text{Macro-Precision} \times \text{Macro-Recall}) / (\text{Macro-Precision} + \text{Macro-Recall}) \tag{32}$$

To enhance visualization, residue similarity (R-S) scores can be computed.[40] Traditional visualization techniques often involve reducing the data to two or three dimensions, which may result in the loss of structure and integrity in multiclass data. R-S plots were introduced as a method to visualize results while better preserving the underlying structure of the data.

An R-S plot consists of two main components: the residue score and the similarity score. The residue score is calculated as the sum of distances between classes, capturing the dissimilarity between them. On the other hand, the similarity score represents the average similarity within each class, indicating the degree of similarity between instances belonging to the same class. By considering both scores, R-S plots

https://doi.org/10.1021/acs.jcim.3c01023
J. Chem. Inf. Model. 2024, 64, 2405−2420

Table 10. Comparison of PLPCA and Other Methods on Datasets Obtained from Gene Expression Omnibus

data set	method	mean ACC	mean macro-REC	mean macro-PRE	mean macro-F1	macro-AUC
GSE14020	**PLPCA**	**0.8228**	**0.7240**	**0.6679**	**0.6709**	**0.6811**
	pLPCA	0.8038	0.7102	0.6564	0.6546	0.6698
	RLSDSPCA[13]	0.8152	0.7219	0.6692	0.6681	0.6790
	SDSPCA[11,13]	0.7876	0.6964	0.6375	0.6354	0.6500
	RgLPCA[12]	0.8123	0.7145	0.6619	0.6612	0.6733
	gLSPCA[41]	0.5399	0.3788	0.4000	0.3415	0.5237
	gLPCA[12]	0.8133	0.7177	0.6649	0.6633	0.6745
	PCA[42]	0.7800	0.6930	0.6346	0.6306	0.6466
GSE39582	**PLPCA**	**0.9960**	**0.9749**	**0.9520**	**0.9603**	**0.9700**
	pLPCA	0.9927	0.9732	0.9182	0.9385	0.9609
	RLSDSPCA[13]	0.9959	0.9748	0.9473	0.9573	0.9650
	SDSPCA[11,13]	0.9947	0.9742	0.9742	0.9570	0.9623
	RgLPCA[12]	0.9932	0.9734	0.9182	0.9387	0.9611
	gLSPCA[41]	0.9400	0.5824	0.5614	0.5701	0.6123
	gLPCA[12]	0.9926	0.9731	0.9173	0.9380	0.9608
	PCA[42]	0.9923	0.9730	0.9118	0.9342	0.9600
GSE18842	**PLPCA**	**0.8688**	**0.8648**	**0.8864**	**0.8559**	**0.8500**
	pLPCA	0.8171	0.8047	0.8505	0.7831	0.8438
	RLSDSPCA[13]	0.8166	0.8041	0.8542	0.7837	0.8401
	SDSPCA[11,13]	0.8176	0.8052	0.8552	0.7845	0.8449
	RgLPCA[12]	0.8166	0.8043	0.8542	0.7837	0.8452
	gLSPCA[41]	0.6300	0.6188	0.6357	0.6078	0.6210
	gLPCA[12]	0.8183	0.8060	0.8554	0.7856	0.8470
	PCA[42]	0.8185	0.8062	0.8556	0.7859	0.8474
GSE35988	**PLPCA**	**0.7159**	**0.7640**	**0.7229**	**0.6710**	**0.7111**
	pLPCA	0.6697	0.7307	0.6824	0.6267	0.6995
	RLSDSPCA[13]	0.7054	0.7589	0.7108	0.6616	0.7105
	SDSPCA[11,13]	0.7009	0.7520	0.7102	0.6550	0.7098
	RgLPCA[12]	0.6864	0.7426	0.6953	0.6434	0.7051
	gLSPCA[41]	0.7050	0.6450	0.6103	0.6136	0.6800
	gLPCA[12]	0.6735	0.7334	0.6888	0.6306	0.7002
	PCA[42]	0.6764	0.7353	0.6900	0.6336	0.7002
GSE29272	**PLPCA**	**0.7925**	**0.7901**	**0.8289**	**0.7814**	**0.7923**
	pLPCA	0.6729	0.6672	0.7279	0.6057	0.7591
	RLSDSPCA[13]	0.6738	0.6676	0.7445	0.6038	0.7601
	SDSPCA[11,13]	0.6747	0.6686	0.7412	0.6040	0.7614
	RgLPCA[12]	0.6700	0.6644	0.7277	0.6023	0.7552
	gLSPCA[41]	0.5120	0.5021	0.5555	0.3492	0.5156
	gLPCA[12]	0.6710	0.6653	0.7280	0.6028	0.7587
	PCA[42]	0.6707	0.6649	0.7290	0.6025	0.7585

provide a comprehensive representation of the data's structure in a visualization.

Given data of the form $\{(\vec{x}_m, y_m) | \vec{x}_m \in \mathbb{R}^N, y_m \in \mathbb{Z}_l\}_{m=1}^M$, we have y_m representing the class label of our mth data point $\vec{x}_m \in \mathbf{X}$. Say that our data have N samples, M features, and L classes. We can then partition our data set $\mathbf{X}$ into subsets containing each of the classes by taking $C_l = \{\vec{x}_m \in \mathbf{X} | y_m = l\}$. For each class l we then define the residue score as follows:

$$R_m := R(\vec{x}_m) = \frac{1}{R_{\max}} \sum_{\vec{x}_j \notin C_l} \|\vec{x}_m - \vec{x}_j\| \tag{33}$$

where $\|\cdot\|$ denotes the Euclidean distance between vectors and $R_{\max}$ is the maximal residue score for that subset. The similarity score, meanwhile, is given as

$$S_m := S(\vec{x}_m) = \frac{1}{|C_l|} \sum_{\vec{x}_j \in C_l} \left(1 - \frac{\|\vec{x}_m - \vec{x}_j\|}{d_{\max}}\right) \tag{34}$$

where $d_{\max}$ is the maximal pairwise distance of the data set. For constructing R-S plots, we then take $R(\vec{x})$ to be the x-axis and $S(\vec{x})$ to be the y-axis.

3.4. Comparison of pLPCA and gLPCA. The classification of benchmark tumor datasets provides an opportunity to evaluate the performance of our method in comparison to other state-of-the-art approaches. First, we validate our claim that persistent Laplacian-based regularization surpasses graph Laplacian regularization by comparing pLPCA and gLPCA. Following this validation, we proceed to compare our PLPCA with RLSDSPCA, which has demonstrated the best performance in the existing literature.[13]

To summarize the comparison between pLPCA and gLPCA on the COAD dataset, please refer to Table 2. Note that the gLPCA results reported in the early work[13] do not appear to be reasonable because they are higher than those of the improved model gLSPCA, which should not happen according to Feng et al.[41] Therefore, we have reproduced these results for gLPCA using the parameters specified by

https://doi.org/10.1021/acs.jcim.3c01023
J. Chem. Inf. Model. 2024, 64, 2405−2420

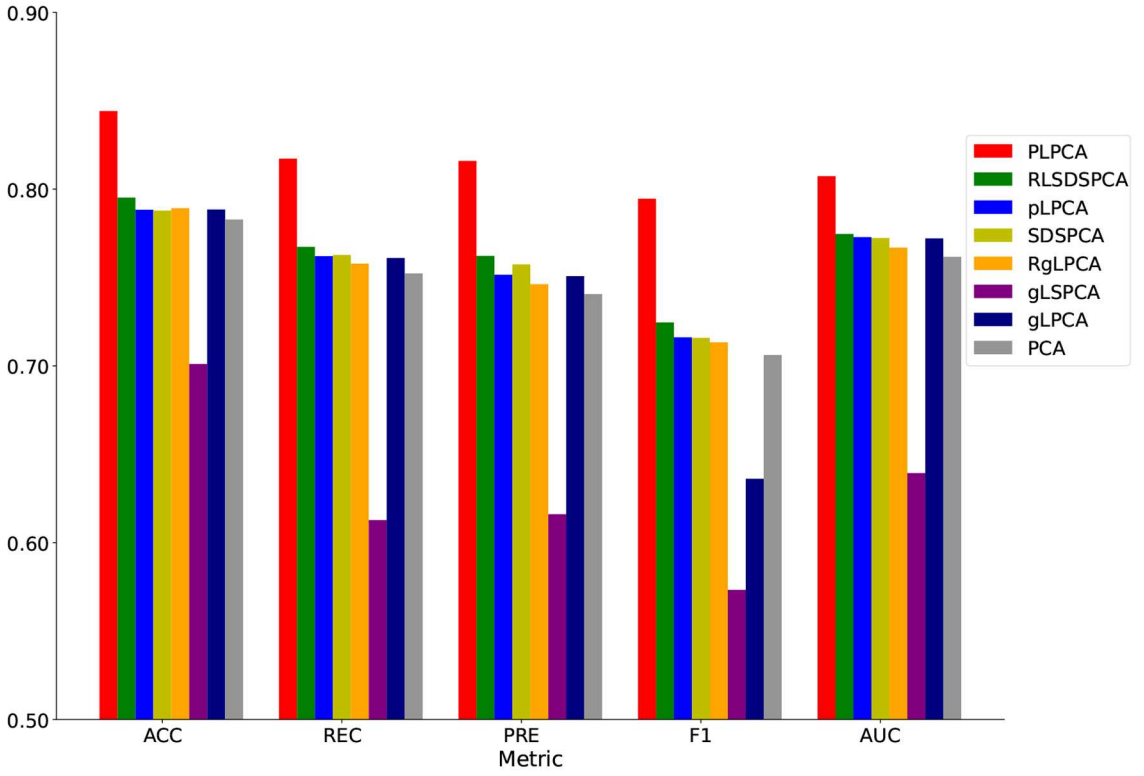


Figure 7. Comparison across five performance metrics, on average, for PCA-based methods across all 10 data sets.

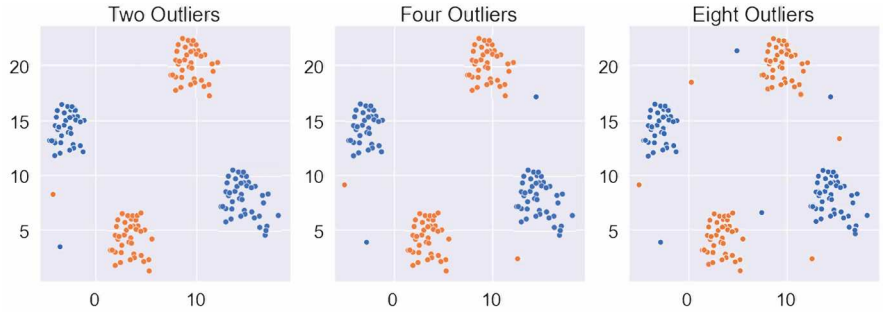


Figure 8. PCA plot of simulated data sets with two, four, and eight outliers for robustness testing. Each data set has been mapped to $m = 2$ dimensions for visualization.

Zhang et al.[13] Our results are listed in Table 5 for a comparison.

This table clearly demonstrates that pLPCA outperforms gLPCA in all performance metrics, highlighting the superior ability of persistent spectral graphs to retain topological and geometrical information during dimensionality reduction. Specifically, we observe an improvement in mean accuracy from 0.9756 to 0.9788. Similarly, the macro-F1 score improved from 0.9002 to 0.9115. These results indicate that pLPCA not only achieves higher overall classification accuracy but also reduces the number of false negatives.

To further validate the performance of pLPCA, we conducted tests on the MultiSource dataset, and the results are presented in Table 3.

Once again, it is important to highlight the significant improvement in performance across all five evaluation metrics achieved by incorporating persistent Laplacian-based regularization instead of graph Laplacian. The mean accuracy has shown a substantial improvement of 1.28%, while the macro-F1 score has improved by an even greater 1.55%.

To provide a visual representation of this performance enhancement, Figure 4 illustrates the improvement in mean accuracy across each of the tested subspace dimensions ranging from 1 to 100. This visualization clearly demonstrates the superior accuracy achieved by pLPCA compared to gLPCA across almost every reduced dimension. This reinforces the intuitiveness of pLPCA in achieving better accuracy across a wide range of dimensional reductions.

In a similar manner, Figure 4 presents a graphical analysis of the macro-F1 score. It is evident that, similar to the accuracy results, the macro-F1 score demonstrates improvement across almost all tested subspace dimensions. This emphasizes the positive impact of incorporating persistent spectral graphs to augment performance.

Moving forward, let us explore how the integration of discriminative information and sparseness can further enhance performance.

https://doi.org/10.1021/acs.jcim.3c01023
J. Chem. Inf. Model. 2024, 64, 2405−2420

Table 11. Verifying Robustness on Simulated Outlier Datasets

data set	method	mean ACC	mean macro-REC	mean macro-PRE	mean macro-F1	macro-AUC
two outliers	PLPCA	1.0	1.0	1.0	1.0	1.0
	RLSDSPCA[13]	1.0	1.0	1.0	1.0	1.0
	SDSPCA[11,13]	0.9939	0.9923	0.9952	0.9935	0.9923
	RgLPCA[12,13]	0.9939	0.9923	0.9952	0.9935	0.9923
	gLSPCA[13,41]	0.9939	0.9923	0.9952	0.9935	0.9923
	gLPCA[12,13]	0.9939	0.9923	0.9952	0.9935	0.9923
	PCA[13,42]	0.9939	0.9923	0.9952	0.9935	0.9923
four outliers	PLPCA	0.9939	0.9952	0.9923	0.9936	0.9952
	RLSDSPCA[13]	0.9939	0.9952	0.9923	0.9936	0.9952
	SDSPCA[11,13]	0.9878	0.9889	0.9867	0.9874	0.9889
	RgLPCA[12,13]	0.9818	0.9813	0.9806	0.9808	0.9813
	gLSPCA[13,41]	0.9878	0.9869	0.9869	0.9869	0.9869
	gLPCA[12,13]	0.9818	0.9813	0.9806	0.9808	0.9813
	PCA[13,42]	0.9818	0.9813	0.9806	0.9808	0.9813
eight outliers	PLPCA	0.9939	0.9947	0.9933	0.9938	0.9947
	RLSDSPCA[13]	0.9939	0.9947	0.9933	0.9938	0.9947
	SDSPCA[11,13]	0.9818	0.9823	0.9811	0.9815	0.9823
	RgLPCA[12,13]	0.9757	0.9756	0.9758	0.9754	0.9756
	gLSPCA[13,41]	0.9818	0.9823	0.9811	0.9815	0.9823
	gLPCA[12,13]	0.9757	0.9756	0.9758	0.9754	0.9756
	PCA[13,42]	0.9757	0.9756	0.9758	0.9754	0.9756

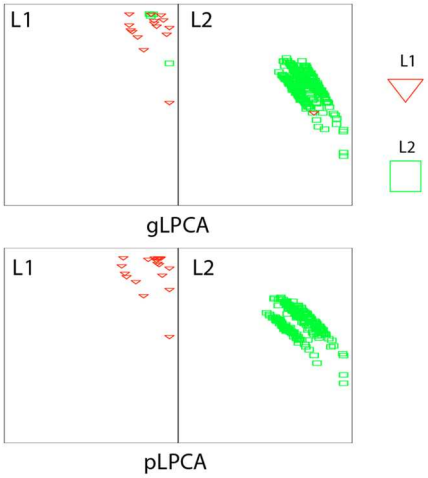


Figure 9. R-S plots of clusters generated from gLPCA and pLPCA-based dimensionality reduction. The x-axis is the residual score, and the y-axis is the similarity score. Each section corresponds to one cluster, and the data were colored according to the predicted labels from KNN on the COAD data set at k = 100.

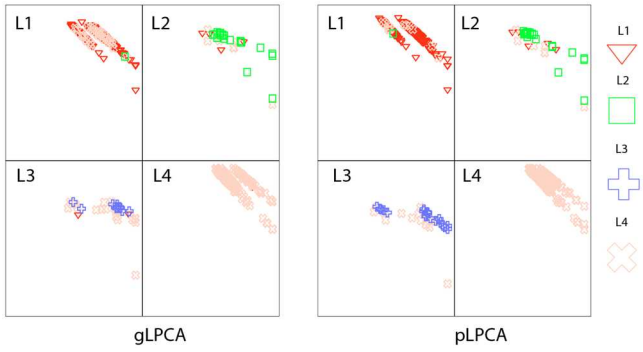


Figure 10. R-S plots of clusters generated from gLPCA and pLPCA-based dimensionality reduction. The x-axis is the residual score, and the y-axis is the similarity score. Each section corresponds to one cluster and the data were colored according to the predicted labels from KNN on the MultiSource data set at k = 65.

3.5. Comparison of PLPCA and RLSDSPCA. After observing the superior performance of pLPCA compared to gLPCA, we conducted a similar study to compare PLPCA and RLSDSPCA, which has been identified as the top-performing method among all PCA-related approaches in the literature.[13] Table 4 provides a summary of our results on the COAD data set for both RLSDSPCA and PLPCA.

The results clearly demonstrate that PLPCA outperforms RLSDSPCA in every major category. Specifically, the mean accuracy across all subspaces was 0.9875 for RLSDSPCA and 0.9886 for PLPCA. Additionally, the macro-F1-score, which highlights the impact of false negatives, improved from 0.9530 to 0.9578, representing a 0.48% improvement.

The performance improvement is even more significant when considering the results of the MultiSource data set, as presented in Table 5.

PLPCA exhibits an improvement in mean accuracy on this benchmark data set, increasing from 0.9273 to 0.9371, which corresponds to a 0.98% improvement. Similarly, the F1 score shows an improvement from 0.8527 to 0.8619. To visually demonstrate the comparison between the two methods, Figure 5 again presents the distribution of performance on the MultiSource data set across different reduced dimensions for both procedures. This intuitive visualization provides a clear illustration of how the two methods compare in terms of performance.

From the depicted figure, we can observe that while the performance of RLSDSPCA tends to significantly decline as the subspace dimension (m value) increases, this decline is considerably mitigated in the new procedure (PLPCA). As a result, PLPCA exhibits better overall performance. Particularly noteworthy is the even greater improvement in the F1 score, as illustrated in Figure 5. This demonstrates the ability of

https://doi.org/10.1021/acs.jcim.3c01023
J. Chem. Inf. Model. 2024, 64, 2405−2420

PLPCA to consistently outperform the next best method across almost all tested subspace dimensions. This is especially encouraging considering the F1 score's crucial role in tumor classification.

To facilitate a more straightforward comparison of the two procedures, we provide a barplot in Figure 6(a). This barplot allows for a comprehensive evaluation of the relative performances of both procedures across all five evaluation metrics for the MultiSource data set: accuracy, recall, precision, AUC, and F1. We include a similar plot comparing the performances of gLPCA and pLPCA as well in Figure 6(b).

The figure clearly illustrates the superiority of PLPCA over RLSDSPCA in all the tested metrics, surpassing the previously identified next best method. Notably, the most substantial improvement is observed in the F1 and recall scores, which are considered particularly important. These findings provide strong evidence of the effectiveness of PLPCA.

Now, to further supplement and confirm these results, in Table 6 we compare the performances of PLPCA and RLSDSPCA on each of the data sets obtained from the Gene Expression Omnibus.

Next, we strengthen our findings by examining and comparing our procedure with some other PCA methods from the literature.

3.6. Comparisons with Other Methods. Zhang et al.[13] demonstrated that RLSDSPCA achieves superior results compared to other PCA-based approaches. However, after observing how PLPCA enhances classification performance in comparison to RLSDSPCA, it is necessary to further evaluate the performance of our method against other existing approaches.

To validate the performance of PLPCA, we can refer to Table 7 for the MultiSource data set, where a comprehensive comparison of different methods is presented.

This table provides a comprehensive overview of the performance differences between our procedure and other PCA enhancements, including pLPCA. The results clearly demonstrate the significant impact of PLPCA's ability to capture geometrical structure information through persistent spectral graphs, while also incorporating label information and sparseness.

Notably, there is a considerable improvement in mean accuracy when transitioning from PCA to PLPCA, with the metric increasing from 0.9108 to 0.9371, representing a 2.63% improvement. Similarly, the F1 score shows an even greater improvement, increasing from 0.8025 to 0.8619, which corresponds to a remarkable improvement of 5.94%. These findings underscore the importance of not only capturing geometrical information but also addressing class ambiguities and enforcing sparseness.

Additionally, we can compare the performance of PLPCA with other notable PCA enhancements on the COAD data set, as depicted in Table 8.

Once again, it is important to highlight the consistently superior performance across all five evaluation metrics, with particular emphasis on the macro- F1 score. The results clearly demonstrate that the PLPCA procedure outperforms other PCA methods by a significant margin.

To further underscore this point, let us compare our method to traditional PCA. The comparison reveals noteworthy improvements in mean accuracy, increasing from 0.9593 to 0.9886, and mean F1 score, improving from 0.8799 to 0.9578. It is crucial to acknowledge though that PLPCA also exhibits superior performance across all major evaluation metrics, not just accuracy and F1 score.

Now, we can compare the performance of PLPCA against the other methods on the data sets obtained from the Gene Expression Omnibus in Tables 9 and 10.

To visually represent the superior performance of our method, we provide a barplot in Figure 7, comparing the performance metrics of the mentioned procedures averaged over each of the 10 tested data sets.

From the depicted image, it is clearly evident that PLPCA surpasses other PCA enhancements, including RLSDSPCA, in all aspects, particularly in terms of F1 and recall.

After confirming the efficacy of our procedure on real gene expression data, we can proceed to evaluate our method on various simulated outlier datasets sourced from the RLSDSPCA GitHub repository. The objective is to assess whether the enhanced robustness of RLSDSPCA is compromised by the inclusion of persistent spectral graphs.

3.7. Robustness to Outliers. Earlier studies show that inclusion of $L_{2,1}$ norm regularization for the error function induces robustness to outliers.[13] Here, we verify the continued effectiveness of this method on the PLPCA procedure by testing several simulated outlier data sets. The data sets have two, four, and eight outliers, respectively. In each case, there are two classes. We include the PCA plots of each simulated data set in Figure 8.

We now summarize the classification results of each algorithm on each data set in Table 11.

The inclusion of persistent spectral graph regularization in RLSDSPCA does not significantly affect its robustness, indicating that our new method remains robust to outliers to a certain extent. However, it is important to note that an increased number of outliers can still have a negative impact on performance, although this is less problematic for both PLPCA and RLSDSPCA.

3.8. Residue-Similarity Analysis. To more effectively visualize our gene expression data after dimensionality reduction, we can generate residue-similarity plots for each of the tested data sets.[40] We can then compare the results for the gLPCA and pLPCA models.

Figures 9 and 10 compare the classification accuracy on the COAD and MultiSource data sets between dimensionality reductions using gLPCA and pLPCA. The chosen subspace dimensions for visualization were $m = 100$ and $m = 65$, respectively. These results, along with Figure 4, showcase the ability of persistent Laplacian-regularized PCA to outperform graph Laplcian-based PCA. In particular, we note the poor performance of gLPCA-based dimensionality reduction for classifying Labels 1 and 3 in the MultiSource data set, and the improvement seen when using pLPCA instead.

4. CONCLUDING REMARKS

As DNA sequencing technologies have advanced in timeliness and cost, they have greatly expanded our understanding of the pathogenic genes responsible for the development and progression of different cancers. These insights have led to the identification of diagnostic biomarkers and therapeutic targets. Given the need for dimensionality reduction to effectively analyze this data, it is crucial to maximize its accurate representation. In this regard, we propose a novel method called persistent Laplacian-enhanced PCA (PLPCA). This method incorporates robustness, label information, and

https://doi.org/10.1021/acs.jcim.3c01023
J. Chem. Inf. Model. 2024, 64, 2405−2420

sparsity, while also improving the capture of geometrical structure information using techniques derived from persistent topological Laplacian theory.[17]

While in this study we have examined the benefits of dimensionality reduction for microarray data classification using our topological technique, we should mention other important possible applications of our method. Most obvious is that dimensionality reduction can reveal underlying clusters in the data, making it a necessary prior step to any clustering analysis for cancerous tumor identification. Furthermore, our method may improve current feature selection techniques, where one wishes to identify genes which contribute the most to overall variance in the data, or correlate strongly with different tissue types. Lastly, our method could be used in conjunction with different visualization techniques as a data preprocessing step. For example, using tSNE or UMAP for Eigen-Gene visualization requires an aggressive reduction to only $m = 2$ dimensions, distorting the integrity of the data. If we first prepossess our data by reducing to, say, $m = 50$ dimensions prior to visualization with tSNE or UMAP, we may improve the quality of the visualization.

Our extensive computational results over ten diverse data sets demonstrate that by incorporating persistent topological regularization in the RLSDSPCA procedure, we achieve the highest level of classification performance after dimensionality reduction compared to previous methods. While the inclusion of a graph Laplacian contributes to capturing geometrical structure information, the current analysis is limited to a single-scale Laplacian. To overcome this limitation, one may generate a sequence of topological Laplacians through filtration, providing a more comprehensive multiscale perspective of the data and enabling one to emphasize features at different important scales. We have incorporated these enhancements alongside label information, sparseness, and robustness to outliers, resulting in a dimensionality reduction technique superior to other comparable procedures from the literature. Ultimately, we have showed that this additional regularization resulted in a significant improvement in performance. On average, we saw performance metrics increase by a margin of 8.01% in accuracy, 7.49% in recall, 8.15% in precision, 11.89% in F1, and 5.14% in AUC. Alternatively, we also achieve similarly superb results by incorporating the PL regularization to the original PCA approach. This method, called pLPCA, does not depend on the availability of data labels, and thus would be preferable for use in unsupervised machine learning tasks such as clustering.

Despite progress made by our proposed method, there is still ample room for improvement. First, it is interesting to examine the role of higher-dimensional Laplacians in dimensionality reduction. Furthermore, we note the weakness in our method associated with the extensive hyper-parameter search necessary to optimize performance. Future research efforts should focus on ways to develop a parameter-free method, or at least a method with a significantly narrowed parameter distribution. Additionally, further analysis is needed to evaluate the performance of this procedure for feature selection compared to other methods. Previous studies, including Zhang et al.,[13] have described a feature selection procedure that assumes linear relationships among genes, which may not be optimal.[43] It would be advantageous to explore more sophisticated feature selection techniques that account for the nonlinear relationships among genes. Moreover, integrating our new dimensionality reduction procedure into these methods could lead to further performance improvements.[44,45] Additionally, understanding the role and significance of the selected genes in driving or correlating with different cancer incidences is an important area for future research. Both aspects require continued efforts in the fields of computational and mathematical biology.

■ ASSOCIATED CONTENT

Data Availability Statement

The model and data used in this analysis is publicly available at the PLPCA GitHub repository: https://github.com/seanfcottrell/PLPCA

■ AUTHOR INFORMATION

Corresponding Author

Guo-Wei Wei − *Department of Mathematics, Michigan State University, East Lansing, Michigan 48824, United States; Department of Electrical and Computer Engineering and Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, Michigan 48824, United States;* orcid.org/0000-0002-5781-2937; Email: weig@msu.edu

Authors

Sean Cottrell − *Department of Mathematics, Michigan State University, East Lansing, Michigan 48824, United States*

Rui Wang − *Department of Mathematics, Michigan State University, East Lansing, Michigan 48824, United States;* orcid.org/0000-0002-7402-6372

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.3c01023

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported in part by NIH grants R01GM126189, R01AI164266, and R35GM148196, NSF grants DMS-2052983, DMS-1761320, and IIS-1900473, NASA grant 80NSSC21M0023, MSU Foundation, Bristol-Myers Squibb 65109, and Pfizer.

■ REFERENCES

(1) Hozumi, Y.; Tanemura, K.; Wei, G. W. Preprocessing of Single Cell RNA Sequencing Data Using Correlated Clustering and Projection. *J. Chem. Inf. Model.* **2023**, DOI: 10.1021/acs.jcim.3c00674.

(2) DeRisi, J.; Penland, L.; Bittner, M.; et al. Use of a Cdna Microarray to Analyse Gene Expression. *Nat. genet* **1996**, *14*, 457−460.

(3) Sarmah, C.; Samarasinghe, S. Microarray Gene Expression: A Study of Between-Platform Association of Affymetrix and Cdna Arrays. *Comput. Biol. Med.* **2011**, *41* (10), 980−986.

(4) Dunteman, G. *Principal Components Analysis*; Number 69; Sage, 1989.

(5) Xanthopoulos, P.; Pardalos, P.; Trafalis, T. Linear Discriminant Analysis. *Robust data mining* **2013**, 27−33.

(6) Belkin, M.; Niyogi, P. Laplacian Eigenmaps for Dimensionality Reduction and Data representation. *Neural computation* **2003**, *15* (6), 1373−1396.

(7) Schölkopf, B.; Smola, A.; Müller, K. Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural computation* **1998**, *10* (5), 1299−1319.

(8) Yaqoob, A.; Aziz, R.; Verma, N.; Lalwani, P.; Makrariya, A.; Kumar, P. A Review on Nature-Inspired Algorithms for Cancer Disease Prediction and Classification. *Mathematics* **2023**, *11*, 1081.
(9) Yang, X. A New Metaheuristic Bat-Inspired Algorithm. *Nature Inspired Cooperative Strategies for Optimization (NICSO 2010)* **2010**, *284*, 65.
(10) Holland, J. Genetic Algorithms. *Sci. Am.* **1992**, *267* (1), 66−73.
(11) Feng, C.; Xu, Y.; Liu, J.; Gao, Y.; Zheng, C. Supervised Discriminative Sparse PCA for Com-characteristic Gene Selection and Tumor Classification on Multiview Biological Data. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *30* (10), 2926−2937.
(12) Jiang, B.; Ding, C.; Luo, B.; Tang, J. Graph-Laplacian PCA: Closed-Form Solution and Robustness. *In CVPR* **2013**, 3492−3498.
(13) Zhang, L.; Yan, H.; Liu, Y.; Xu, J.; Song, J.; Yu, D. Enhancing Characteristic Gene Selection and Tumor Classification by the Robust Laplacian Supervised Discriminative Sparse PCA. *J. Chem. Inf. Model.* **2022**, *62* (7), 1794−1807.
(14) Eckmann, B. Harmonische Funktionen und Randwertaufgaben in Einem Komplex. *Commentarii Mathematici Helvetici* **1944**, *17* (1), 240−255.
(15) Horak, D.; Jost, J. Spectra of Combinatorial Laplace Operators on Simplicial Complexes. *Advances in Mathematics* **2013**, *244*, 303−336.
(16) Chen, J.; Zhao, R.; Tong, Y.; Wei, G. W. Evolutionary de Rham-Hodge Method. *Discrete Contin. Dyn. Syst. Series B* **2021**, *26* (7), 3785.
(17) Wang, R.; Nguyen, D.; Wei, G. W. Persistent Spectral Graph. *Int. J. Numer. Methods Biomed. Eng.* **2020**, *36* (9), No. e3376.
(18) Mémoli, F.; Wan, Z.; Wang, Y. Persistent Laplacians: Properties, Algorithms and Implications. *SIMODS* **2022**, *4* (2), 858−884.
(19) Wang, R.; Zhao, R.; Ribando-Gros, E.; Chen, J.; Tong, Y.; Wei, G. W. Hermes: Persistent Spectral Graph Software. *FoDS (Springfield, Mo.)* **2021**, *3*, 67.
(20) Zomorodian, A.; Carlsson, G. Computing Persistent Homology. *SoCG* **2005**, 347−356.
(21) Edelsbrunner, H.; Harer, J.; et al. Persistent Homology-A Survey. *Ser. Contemp. Appl. Math.* **2008**, *453* (26), 257−282.
(22) Cang, Z.; Wei, G. W. Topologynet: Topology Based Deep Convolutional and Multi-Task Neural Networks for Biomolecular Property Predictions. *PLoS computational biology* **2017**, *13* (7), No. e1005690.
(23) Qiu, Y.; Wei, G. W. Persistent Spectral Theory-Guided Protein Engineering. *Nat. Comput. Sci.* **2023**, *3*, 149.
(24) Chen, J.; Qiu, Y.; Wang, R.; Wei, G. W. Persistent Laplacian Projected Omicron BA. 4 and BA. 5 to Become New Dominating Variants. *Comput. Biol. Med.* **2022**, *151*, No. 106262.
(25) Meng, Z.; Xia, K. Persistent Spectral−Based Machine Learning (perspect ml) for Protein-ligand Binding Affinity Prediction. *Sci. Adv.* **2021**, *7* (19), No. eabc5329.
(26) Cancer Genome Atlas Network.; et al. Comprehensive Molecular Characterization of Human Colon and Rectal Cancer. *Nature* **2012**, *487* (7407), 330.
(27) Raphael, B.; et al. Integrated Genomic Characterization of Pancreatic Ductal Adenocarcinoma. *Cancer cell* **2017**, *32* (2), 185−203.
(28) Cancer Genome Atlas Network.; et al. Comprehensive Genomic Characterization of Head and Neck Squamous Cell Carcinomas. *Nature* **2015**, *517* (7536), 576.
(29) Farshidfar, F.; et al. Integrative Genomic Analysis of Cholangiocarcinoma Identifies Distinct idh-mutant Molecular Profiles. *Cell Rep.* **2017**, *18* (11), 2780−2794.
(30) Jolliffe, I.; Cadima, J. Principal Component Analysis: A Review and Recent Developments. *Trans. R. Soc., A* **2016**, *374* (2065), No. 20150202.
(31) Luecken, M.; Theis, F. Current Best Practices in Single-Cell RNA-seq Analysis: A Tutorial. *Mol. Syst. Biol.* **2019**, *15* (6), No. e8746.
(32) Belkin, M.; Niyogi, P. Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering. *Adv. Neural Inf. Process, 14*, **2001**.
(33) Cai, D.; He, X.; Han, J.; Huang, T. Graph Regularized Nonnegative Matrix Factorization for Data Representation. *TPAMI/PAMI* **2011**, *33* (8), 1548−1560.
(34) Nguyen, D.; Wei, G. Agl-Score: Algebraic Graph Learning Score For Protein-ligand Binding Scoring, Ranking, Docking, and Screening. *J. Chem. Inf. Model.* **2019**, *59* (7), 3291−3304.
(35) Ding, C.; He, X. K-means Clustering via Principal Component Analysis. *Proceedings of the 21 st International Conference on Machine Learning* **2004**, 29.
(36) Fortin, M., Glowinski, R. *Augmented Lagrangian Methods: Applications to the Numerical Solution of Boundary-Value Problems*; Elsevier, 2000.
(37) Echebest, N.; Sánchez, M.; Schuverdt, M. Convergence Results of an Augmented Lagrangian Method Using the Exponential Penalty Function. *J. Optim Theory Appl.* **2016**, *168*, 92−108.
(38) Boyd, S.; Parikh, N.; Chu, E.; Peleato, B.; Eckstein, J.; et al. Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. *Found. Trends Mach. Learn.* **2010**, *3* (1), 1−122.
(39) Peterson, L. K-Nearest Neighbor. *Scholarpedia* **2009**, *4* (2), 1883.
(40) Hozumi, Y.; Wang, R.; Wei, G. W. CCP: Correlated Clustering and Projection for Dimensionality Reduction. *ArXiv* **2022**, No. 2206.04189v1.
(41) Feng, C.; Xu, Y.; Hou, M.; Dai, L.; Shang, J. PCA via Joint Graph Laplacian and Sparse Constraint: Identification of Differentially Expressed Genes and Sample Clustering on Gene Expression Data. *BMC Bioinf.* **2019**, *20*, 1−11.
(42) Jolliffe, I. Principal Component Analysis. *Encyclo. Stat. Behav. Sci.* **2005**, No. bsa501, DOI: 10.1002/0470013192.bsa501.
(43) Huerta, M.; Cedano, J.; QUEROL, E.; et al. Analysis of Nonlinear Relations Between Expression Profiles by the Principal Curves of Oriented-Points Approach. *J. Bioinf. Comput. Biol.* **2008**, *6* (02), 367−386.
(44) Kiselev, V.; et al. Sc3: Consensus Clustering of Single-Cell RNA-seq Data. *Nat. Methods* **2017**, *14* (5), 483−486.
(45) Ren, X.; Zheng, L.; Zhang, Z. Sscc: A Novel Computational Framework for Rapid and Accurate Clustering Large-Scale Single Cell RNA-seq Data. *Genomics, Proteomics Bioinf.* **2019**, *17* (2), 201−210.

https://doi.org/10.1021/acs.jcim.3c01023
J. Chem. Inf. Model. 2024, 64, 2405−2420