Local Manifold Explanations with Tangent Space Regression

Jesse He JEH020@UCSD.EDU
Yusu Wang YUSUWANG@UCSD.EDU
Gal Mishne GMISHNE@UCSD.EDU
Halıcıoğlu Data Science Institute, University of California San Diego

Abstract

Low-dimensional manifold learning is used to embed and visualize high-dimensional data, revealing its underlying geometry. However, identifying which features drive local variation along the manifold remains difficult. Many post-hoc explanation methods target explaining extrinsic embedding coordinates rather than intrinsic manifold structure, or provide only global explanations. In this work, we introduce Local Tangent Space Regression Explanations (LTSREx), a method to explain the local structure of a manifold in terms of interpretable features by performing sparse linear regression in the tangent space of the manifold at each point, coupled with Tikhonov denoising via the connection Laplacian to ensure that explanations are consistent and vary smoothly across nearby points. We show that our method produces meaningful local explanations on synthetic data, rotated MNIST digits, and two single-cell gene expression datasets. Our code is available at https://github.com/he-jesse/LTSREx.

Keywords: manifold learning, explainability, interpretability

1. Introduction

Across scientific disciplines, dimensionality reduction has cemented itself as a prominent set of tools for exploratory data analysis. The field of manifold learning is dedicated to uncovering the underlying geometric structure of data, i.e. clusters, continuous manifolds, loops, local variation, which can inspire hypotheses for further study. However, discovering which features drive the geometry of the manifold often involves a series of educated guesses, manually examining the visual relationship between the low-dimensional embedding and each feature. This lack of interpretability poses a challenge in application domains, where the manifold structure may arise from a small subset of features, for example a few genes out of several thousand genes in single-cell RNA sequencing (scRNA-seq) or the collective responses of thousands of neurons to behavioral variables of an animal.

While prior methods have been proposed to automate the process of interpreting a low-dimensional manifold embedding, many of these prior methods explain the *extrinsic* embedding coordinates for the manifold rather than the *intrinsic* manifold structure. That is, their aim is to explain the embedding coordinates rather than attempting to explain the local variation within the manifold itself. Another important distinction is between *local* and *global* explanations: local methods aim to describe the geometry surrounding an individual sample, in contrast to global methods that attempt to identify features that explain the entire manifold structure. Global explanations may be suitable for some applications, but this depends on both the scientific domain and the embedding method. For example, in scRNA-seq data, samples typically cluster based on cell type, and each cluster may be

explained by a different set of genes. Local variation can also differ even within a cluster. Thus local—rather than global—explanations are necessary.

In this work, we present Local Tangent Space Regression Explanations (LTSREx), a method to produce consistent local explanations of intrinsic manifold structure. Our contribution is twofold: we provide a method to compute local intrinsic explanations of a manifold in terms of an arbitrary feature dictionary (i.e. the features may, but need not be the original features) by performing sparse linear regression in the estimated tangent space around each point (Section 3). We also enforce consistency of explanations across neighboring points using a signal denoising-inspired approach: we apply Tikhonov denoising using the connection Laplacian (Section 3.1). Section 4 demonstrates how our method produces meaningful local explanations on synthetic data, MNIST, and two scRNA-seq datasets.

1.1. Related Work

A number of methods have been proposed to explain low-dimensional manifold embeddings in terms of the original feature space or a dictionary of domain-relevant features. Marion et al. (2019) introduce a method to find the "Best Interpretable Rotation (BIR)," giving the orthogonal transformation of a multi-dimensional scaling embedding that is most suitable for interpretation with sparse regression. Several methods are inspired by the explainability method LIME (Ribeiro et al., 2016), explaining embeddings with a local surrogate model: Bibal et al. (2020) adapt LIME to produce local explanations for t-SNE embeddings; LXDR (Bardos et al., 2022; Mylonas et al., 2024) fits a separate linear surrogate for each coordinate in the manifold embedding space; and LIMEADE (Zikry and Allen, 2025) fits a multivariate sparse linear model in a group LASSO problem. LIMEADE also offers partition-based and local formulations (Zikry and Allen, 2025), although these may require prior knowledge to define regions. A similar approach is taken by ManifoldLasso (Koelle et al., 2022), where interpretable coordinates are provided as a dictionary of domain-relevant features. By projecting the gradients of these interpretable features onto the tangent spaces of the manifold, ManifoldLasso solves a group LASSO problem over the manifold to replace abstract embedding coordinates with an "equivalent" set of meaningful embedding functions. These methods provide *extrinsic* explanations, unlike our method.

A notable exception is TSLasso (Koelle et al., 2024), a refinement of the setting of ManifoldLasso. TSLasso also begins with a dictionary of interpretable features and projects their gradients onto the manifold's tangent spaces. Unlike ManifoldLasso, however, TSLasso directly computes the intrinsic features of the manifold rather than providing extrinsic embedding functions. However, a key difference between our method and TSLasso is that TSLasso is formulated globally rather than locally: given an embedding, TSLasso attempts to identify a globally consistent set of explanatory features for the *entire* manifold, producing global rather than local explanations.

2. Preliminaries

Let $X = \{x_1, \ldots, x_n\}$ be points sampled from a d-manifold $\mathcal{M}$ embedded in $\mathbb{R}^p$, possibly with noise. Let F be a dictionary of q functions $F = (f_1, \ldots, f_q)$, $f_r : \mathcal{M} \to \mathbb{R}$. The dictionary F may consist of the original features themselves, metadata attached to each point x_i (which are not used for embedding), or additional features engineered on the data

x_i, e.g., from domain knowledge. We wish to identify for a given x_i which functions f_r explain the manifold near x_i. That is, for a local neighborhood $U_i \ni x_i$, our aim is to provide a local parameterization of U_i in terms of a subset of the dictionary functions f_r.

Most previous work takes F to be the original features which produced the low-dimensional embedding, although Koelle et al. (2024) assume that F consists of several candidate interpretation functions provided by domain expertise. Whether local or global, previous work largely fits a sparse linear regression model of X onto F to provide the interpretation. Implicit in this approach, especially for local methods, is the intuition that the local neighborhood of each point is approximately linear because the data lies on a manifold. That is, local variation is described by the tangent spaces of the manifold. Koelle et al. (2024) use this assumption explicitly, regressing the *gradients* of each f_r onto an estimated basis for the local tangent space about each point in a group LASSO formulation.

Robust Tangent Space Estimation. Traditionally, estimating the tangent space around a point x_i involves performing a weighted PCA on the local neighborhood of x_i (Zhang and Zha, 2003; Singer and Wu, 2012; Koelle et al., 2024). While this is compatible with our method, noise and curvature both pose difficulties for accurately estimating a tangent space basis with local PCA. In the presence of heavy noise in directions normal to the manifold, a PCA decomposition of the local neighborhood can mistakenly identify noise directions as tangent space directions. In such cases, accurately estimating the tangent space may require a larger neighborhood for local PCA. However, sparse data or intense curvature can limit the number of neighbors which are suitable for tangent space estimation, as large neighborhoods may cease to be locally linear. To mitigate this effect, we adopt Laplacian Eigenvector Gradient Orthogonalization (LEGO) (Kohli et al., 2025) as our tangent space estimator for noisy data. A detailed description of LEGO is given in Appendix A.1.

3. Method

LTSREx is composed of three main steps: first, for each point we estimate the local tangent space based on its local neighborhood. Next we fit a sparse regression model in that tangent space, finally, we denoise the resulting coefficients across the neighborhood graph.

As in related work, a natural approach to express a neighborhood U_i around a point x_i in terms of several functions f_r is through sparse linear regression. In our implementation, we take U_i to be the k nearest neighbors $(x_{j_1}, \ldots, x_{j_k})$ of x_i. Let $\mathbf{X}^{(i)}$ be the neighborhood points centered at x_i, and we do the same for the feature values of each x_j:

$$\mathbf{X}^{(i)} = \begin{bmatrix} x_{j_1} - x_i \\ \vdots \\ x_{j_k} - x_i \end{bmatrix} \in \mathbb{R}^{k \times p}, \quad \mathbf{F}^{(i)} = \begin{bmatrix} f_1(x_{j_1}) - f_1(x_i) & \cdots & f_q(x_{j_1}) - f_q(x_i) \\ \vdots & \ddots & \vdots \\ f_1(x_{j_k}) - f_1(x_i) & \cdots & f_q(x_{j_k}) - f_q(x_i) \end{bmatrix} \in \mathbb{R}^{k \times q}. \quad (1)$$

Related work often fits $\mathbf{X}^{(i)}$ onto $\mathbf{F}^{(i)}$ via sparse regression, thereby explaining variation in the ambient coordinates of the neighborhood. Notice, however, that in fitting a linear model to U_i, we are implicitly making use of the manifold assumption: if U_i is sufficiently small, then it can be described linearly. To make this intuition explicit, what we are really interested in is the *tangent space* $T_{x_i}\mathcal{M}$ around x_i. Therefore, our goal is to identify which functions $f_r \in F$ *vary* along the tangent directions of $\mathcal{M}$ at x_i.

We may estimate $T_{x_i}\mathcal{M}$ using local PCA, weighted local PCA, or LEGO, obtaining a basis $\mathbf{B}^{(i)} \in \mathbb{R}^{d\times p}$ of vectors in $\mathbb{R}^p$ whose rows span $T_{x_i}\mathcal{M}$. We project each neighbor $x_j \in U_i$ onto span $\mathbf{B}^{(i)}$ and center at x_i to obtain projections $\overline{x}_j = \mathrm{Proj}_{\mathbf{B}^{(i)}}(x_j - x_i)$. Let

$$\overline{\mathbf{X}}^{(i)} = \begin{bmatrix} \overline{x}_{j_1} \\ \vdots \\ \overline{x}_{j_k} \end{bmatrix} \text{ and } \overline{\mathbf{F}}^{(i)} = \mathbf{F}^{(i)} \operatorname{diag} \begin{bmatrix} \sigma_1^{-1} \\ \vdots \\ \sigma_q^{-1} \end{bmatrix} \text{ where } \sigma_r = \mathrm{stddev}(f_r(U_i)). \tag{2}$$

$\overline{\mathbf{F}}^{(i)}$ contains the values of F restricted to U_i, centered at $F(x_i)$, and normalized column-wise by their standard deviation. We wish to find a sparse coefficient matrix $\mathbf{W}^{(i)} \in \mathbb{R}^{q\times d}$ such that $\overline{\mathbf{F}}^{(i)}\mathbf{W}^{(i)} \approx \overline{\mathbf{X}}^{(i)}$. Then at x_i, our objective is

$$\min_{\mathbf{W}^{(i)}} \left\| \overline{\mathbf{X}}^{(i)} - \overline{\mathbf{F}}^{(i)}\mathbf{W}^{(i)} \right\|_F^2 + \lambda_2 \left\| \mathbf{W}^{(i)} \right\|_F^2 + \lambda_1 \left\| \mathbf{W}^{(i)} \right\|_1. \tag{3}$$

The first term is simply the reconstruction loss of the underlying regression problem. The other terms are the L_2 and L_1 terms of the ElasticNet penalty (Zou and Hastie, 2005). We adopt ElasticNet here for two reasons: the first is to induce sparsity, as the automatic feature selection results in a sparse model which can be readily interpreted. The second reason is that the combination of neighborhood selection and the large space of possibilities for functions f_r mean that we may have $q \gg |U_i|$, even if $q < n$. In this high-dimensional regime, regularization is necessary to ensure that the regression problem is well-defined. Minimizing (3) yields a sparse approximation of $T_{x_i}\mathcal{M}$ in terms of F. Assuming the intrinsic dimension d is known, we may also tune λ_1, λ_2 so that the resulting coefficient matrix $\mathbf{W}$ has d nonzero columns corresponding to d explanatory features at x_i.

Notice that if U_i is small enough that it can be expressed in normal coordinates about x_i, (say within ε of x_i) then each point $x_j \in U_i$ can be expressed as the Riemannian exponential of some tangent vector v_j in $T_{x_i}\mathcal{M}$: $x_j = \exp_{x_i}(v_j)$. Then for a function $f_r \in F$,

$$f_r(x_j) = f_r(x_i) + \langle \nabla f_r(x_i), v_j \rangle + O(\varepsilon^2). \tag{4}$$

Thus the centered value $f_r(x_j) - f_r(x_i)$ approximates $df_{x_i}(v_j) = \langle \nabla f_r(x_i), v_j \rangle$. Since v_j approximates $\overline{x}_j$, and can be expressed in terms of the basis $\mathbf{B}^{(i)}$ of $T_{x_i}\mathcal{M}$, (3) implicitly provides an estimate of ∇f_r. To be more specific, notice that for each f_r, the r-th row $\mathbf{W}_r^{(i)}$ of $\mathbf{W}^{(i)}$ provides an estimate of ∇f_r expressed in terms of $\mathbf{B}^{(i)}$. That is, $\mathbf{W}_r^{(i)}$ is a vector in the tangent space at x_i. Then over the whole manifold, the local coefficients together define a vector field for each feature f_r, where at each point x_i the vector $\mathbf{W}_r^{(i)}$ expresses not just the *importance* of f_r but also the *direction* of f_r as an explanatory feature.

3.1. Local Consistency of Regression Coefficients

Because our goal is to provide local explanations at a single point, we do not seek a single set of explanatory features globally over the manifold. However, we expect that the coefficients for explanatory features should vary smoothly across nearby points. This is not guaranteed by minimizing (3). In particular, a feature may contain sparse outlier values that are suppressed in the low-dimensional embedding. These create high-leverage points that cause

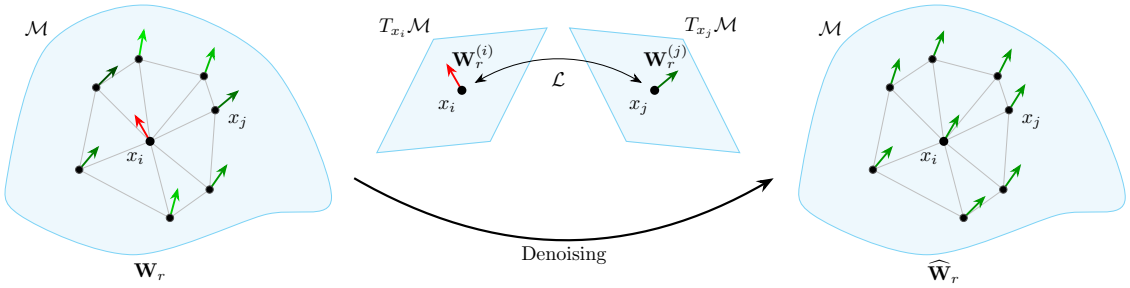

Figure 1: Tikhonov denoising with the connection Laplacian smooths coefficient vectors across nearby points in a parallel transport-aware manner.

the sparse regression objective in (3) to select features that have nearby points with outlier values rather than features that vary smoothly across neighbor instances. Thus a feature f_r with outlier values near a point x_i may be selected as an explanation for x_i, while its coefficients are zero for nearby points.

While robust regression could mitigate the effects of such outliers, we take a different approach based on graph signal processing. When a feature f_r is selected as an explanation for one point but not its neighbors, or conversely when a feature f_r is left out for one point but selected for its neighbors, the result is rather like spike noise in a graph signal corresponding to the selected features across the graph. This suggests *denoising* as a strategy to ensure consistency. In signal processing, *Tikhonov denoising* (Tikhonov and Arsenin, 1977) is a classic method to recover a noisy signal, and can be applied to graph signals via the graph Laplacian (Chen et al., 2014).

With this insight, we can formalize our intuition that the explanations for one point should be similar to the explanations for its neighbors by performing Tikhonov denoising using the *connection Laplacian*. Similar to the standard graph Laplacian, the connection Laplacian $\mathcal{L}$ provides a measure of smoothness for a signal over a graph. However, unlike the standard Laplacian, the connection Laplacian operates on *vector fields* over a graph, making it a natural choice for denoising $\mathbf{W}$. The connection Laplacian encodes the parallel transport operators between tangent spaces at neighboring points, allowing for a parallel transport-aware comparison of coefficient vectors. (We describe in detail how to construct the $\mathcal{L}$ in Appendix A.2.) By using the connection Laplacian, we are able to smooth $\mathbf{W}$ in a parallel transport-aware manner, ensuring that the importance *and direction* of an explanation vary smoothly over $\mathcal{M}$.

Over a neighborhood of x_i (which need not be the neighborhood used for tangent space estimation), we construct $\mathcal{L}$ for the local neighborhood graph. For efficiency, we restrict $\mathcal{L}$ to a small neighborhood subgraph of G. Over each neighboring x_j, we treat the coefficient matrices $\mathbf{W}^{(j)}$ as a noisy vector field signal $\mathbf{W} \in \mathbb{R}^{kd \times q}$. Then for a regularization parameter $\gamma > 0$, we solve

$$\widehat{\mathbf{W}}^* = \underset{\widehat{\mathbf{W}}}{\operatorname{argmin}} \left\| \widehat{\mathbf{W}} - \mathbf{W} \right\|_F^2 + \gamma \left\| \mathcal{L}\widehat{\mathbf{W}} \right\|_F^2. \tag{5}$$

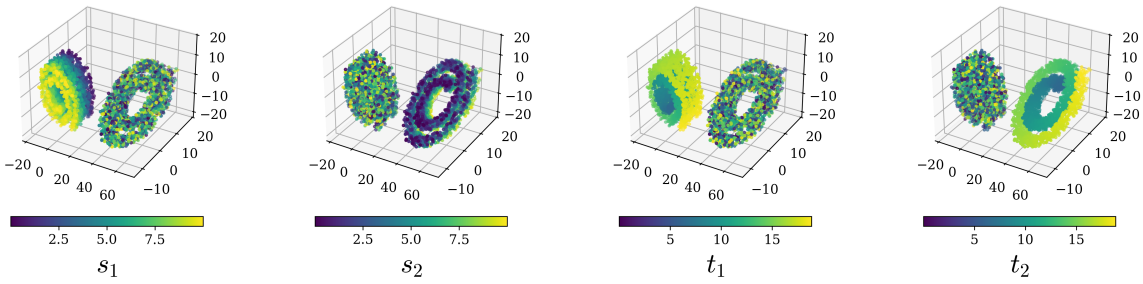

Figure 2: Rotated Swiss rolls. Colored from left to right: s_1, s_2, t_1, t_2

Equation (5) admits the simple closed-form solution $\widehat{\mathbf{W}}^* = (\mathbf{I} + \gamma\mathcal{L})^{-1}\mathbf{W}$. The Tikhonov regularization term uses $\mathcal{L}$ to penalize sharp differences in $\mathbf{W}$. Thus solving (5) for $\widehat{\mathbf{W}}^*$ gives a denoised collection of coefficient matrices near x_i. The i-th block $\widehat{\mathbf{W}}_i^*$ then provides an explanation near x_i which is consistent with its neighbors. Figure 1 shows how this denoising process brings an outlier coefficient vector in line with its neighbors.

Remark. While the smoothness penalty $\gamma \left\| \mathcal{L}\widehat{\mathbf{W}} \right\|_F^2$ could also be applied to (3) and optimized jointly over several neighboring points, we perform the regression and denoising steps separately so that (3) is formulated independently at each x_i without being coupled across neighboring points. This enables e.g. computing $\mathbf{W}^{(i)}$ for several points in parallel, and then simply applying the matrix inverse solution to (5).

4. Results

We first validate our method on synthetic data, where the ground truth explanatory features are known. We then demonstrate its utility on three real-world datasets: rotated MNIST digits, an scRNA-seq dataset of fruit fly clock neurons, and an scRNA-seq dataset of human immune cells. Additional details and results are provided in Appendix B.

Synthetic Data. We validate our method on a synthetic dataset of $n = 5000$ points sampled from two randomly rotated Swiss rolls $\mathcal{M}_1$ and $\mathcal{M}_2$ embedded in $\mathbb{R}^3$, with varying levels of uniform noise orthogonal to the manifold. Our feature dictionary consists of four functions $F = \{s_1, s_2, t_1, t_2\}$, where (s_1, t_1) are the true positions of points on $\mathcal{M}_1$ and random uniform noise on $\mathcal{M}_2$, and similarly for (s_2, t_2) (Fig. 2). Thus for points on $\mathcal{M}_1$ the ground truth explanation is (s_1, t_1) and for points on $\mathcal{M}_2$ it is (s_2, t_2).

We compare different methods of tangent space estimation for increasing noise and neighborhood size k. Detailed results are given in Figure 6, Appendix B.1. Without coefficient denoising, tangent space projection alone provides a strong benefit in the presence of noise, demonstrating the value of intrinsic explanation. Across methods, tangent space regression with LEGO is the most robust to high noise at $k = 8, 16$ with and without coefficient denoising, although at $k = 32$ LPCA can also provide a robust estimate of the tangent space. The denoising step (Section 3.1) substantially improves performance across all estimation methods.

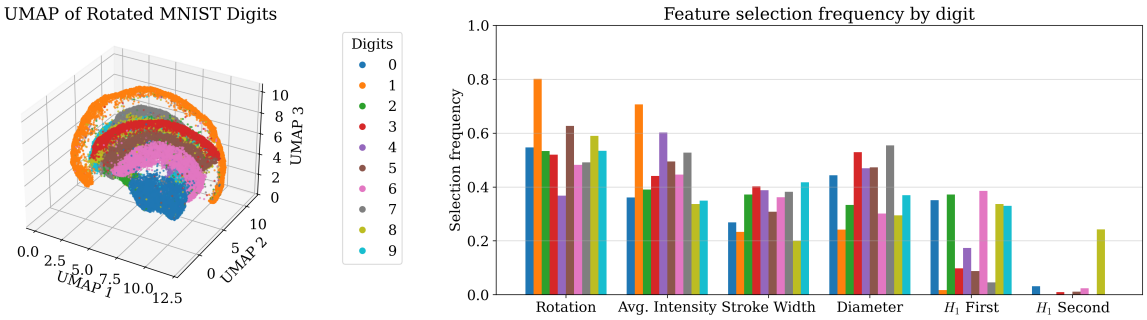

Figure 3: UMAP embedding of rotated MNIST digits (left), with frequency of explanatory features for 1000 randomly selected rotated digits (right).

Rotated MNIST. Here we demonstrate LTSREx on the MNIST dataset with random rotations (Lecun et al., 1998). Each digit is deskewed and then a random rotation between 0 and π is applied. We show that our method can handle arbitrary feature dictionaries by providing the following interpretable feature dictionary: (1) the ground-truth rotation of the digit; (2) the average intensity of the image; (3) the average stroke width of the digit; (4) the diameter of the digit; (5) the persistence of the most persistent H_1 homology class (measuring the intensity of the most intense loop); (6) the persistence of the second most persistent H_1 homology class (measuring the intensity of the second most intense loop). We describe how each feature is computed in Appendix B.2. We use UMAP (McInnes et al., 2020) to embed the data into $\mathbb{R}^3$ (Fig. 3). We then apply LTSREx to 1000 randomly selected points, selecting the top 2 features as the final explanation. Figure 3 shows how frequently each feature is selected. As we expect, the ground truth rotation is frequently chosen as an explanatory feature, as it varies continuously along the entire embedding. However, within each digit class the other features selected differ. For example, the persistence of the second H_1 class is almost exclusively selected for class 8, which is the only digit which regularly has two holes. Similarly, the persistence of the first H_1 class is chosen most often for classes 0, 6, and 9. However, It is also selected for some digits in classes 2 and 4, where there is clear within-class variation of first homology (Examples are shown in Fig. 10, Appendix B.2). These differences in explanations at each digit reflect the locality of our explanation method, as the local variation changes both between and within digits.

***Drosophila* Clock Neurons.** Here, we apply LTSREx to a gene expression dataset of fruit fly clock neurons from Ma et al. (2021), sampled from flies at different times of day to identify rhythmically expressed genes. After preprocessing with ScanPy (Appendix B.4) we embed the data into $\mathbb{R}^{10}$ using UMAP and identify a cluster of cells where the cell embedding varies with time. Fig. 4 shows the cluster in the first two UMAP coordinates as well as the top four explanatory genes identified by LTSREx. These genes are Con, a cell adhesion protein which has been implicated in sleep duration (Shafer, 2025); APPL, which has been observed to affect circadian regulation (Blake et al., 2015); CG45263, an uncharacterized gene predicted to code for cell-cell adhesion and observed in clock neurons

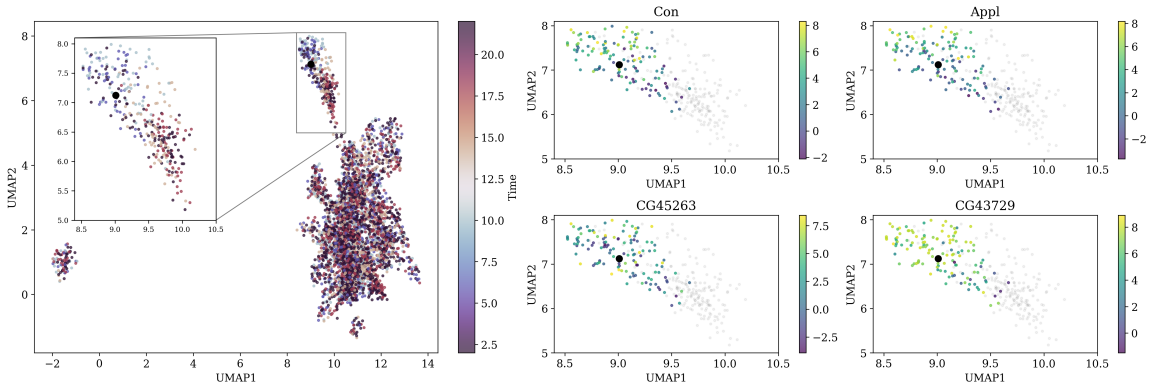

Figure 4: UMAP embedding of *Drosophila melanogaster* clock neurons colored by time (left), with time-varying cluster highlighted (inset). The neighborhood of a cell in this cluster is colored by four selected genes (right).

(Chen et al., 2026); and CG43729, a gene involved in calcium channel regulation which is required for normal circadian activity (Hsu et al., 2018).

PBMC3k. We examine the PBMC3k (3000 peripheral blood mononuclear cells) dataset from 10X Genomics, consisting of single-cell gene expression from about 3000 immune cells introduced in (Zheng et al., 2017), preprocessed by the ScanPy library (Wolf et al., 2018), and embedded in $\mathbb{R}^4$ using UMAP. We use expression of individual genes as features. Fig. 5 shows the top 2 explanatory features for two cells. The explanation for the first cell, a CD8+ T cell at the border of CD8+ T cells and CD4+ T cells, includes CCL5, a cytokine produced by CD8+ cells (Eberlein et al., 2020); and NKG7, which marks cytotoxicity in T cells and NK cells (Turiello et al., 2025). The explanation of the second cell, a CD14+ monocyte at the boundary between CD14+ and FCGR3A+ monocytes consists of LGALS2, which is characteristic of CD14+ monocytes (Wong et al., 2011) (and whose expression increases in the direction of the CD14+ cluster); and FCGR3A, which is characteristic of FCGR3A+ monocytes (and whose expression increases in the direction of FCGR3A+ cells).

5. Discussion

We introduce LTSREx, a method to explain the local intrinsic structure of a manifold embedding, enabling interpretation of low-dimensional manifold embeddings at fine resolutions. Our pipeline of robust tangent space estimation, local sparse regression, and coefficient denoising is both flexible and efficient, capable of extracting meaningful interpretations from noisy data. However, limitations and opportunities for future work remain. Because our approach relies on projecting nearby points onto the manifold's tangent space, we also require a reasonable estimate of the local manifold dimension. While we did not observe high sensitivity to manifold dimensionality in our experiments, estimating d is not trivial in general. In addition, we formulate our Tikhonov denoising step (5) separately from our local regression problem (3). While we decouple these steps for computational efficiency,

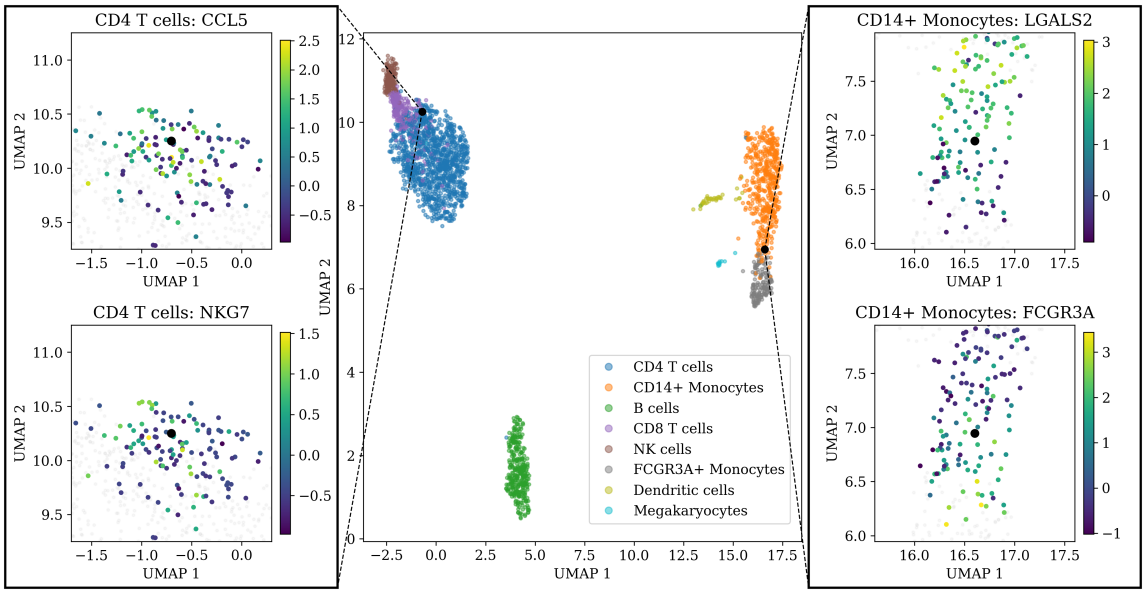

Figure 5: UMAP embedding of PBMC3k cells (center). Local neighborhood of CD8+ T cell (left) and CD14+ monocyte (right) are colored by selected explanatory genes.

future work could incorporate the smoothness penalty into (3) directly, jointly optimizing sparsity and smooth variation of feature coefficients. Finally, while we provide a flexible methodological pipeline, choosing the feature dictionary itself remains an important step which requires domain knowledge. While the raw input features are frequently a sensible option, such as in our gene expression experiments, this may not always the case, such as in image data.

Acknowledgments

We thank Dhruv Kohli for useful discussions about LEGO. This work was partially supported by NSF CCF-2112665, CCF-2217058, and CIF 2403452.

References

10X Geonomics. 3k PBMCs from a healthy donor. Universal 3' dataset analyzed using Cell Ranger 1.1.0. URL https://www.10xgenomics.com/datasets/3-k-pbm-cs-from-a-healthy-donor-1-standard-1-1-0.

Avraam Bardos, Ioannis Mollas, Nick Bassiliades, and Grigorios Tsoumakas. Local explanation of dimensionality reduction. In *Proceedings of the 12th Hellenic Conference on Artificial Intelligence*, SETN '22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450395977. doi: 10.1145/3549737.3549770. URL https://doi.org/10.1145/3549737.3549770.

Mikhail Belkin and Partha Niyogi. Towards a theoretical foundation for laplacian-based manifold methods. *Journal of Computer and System Sciences*, 74(8):1289–1308, 2008.

Adrien Bibal, Viet Minh Vu, Géraldin Nanfack, and Benoît Frénay. Explaining t-SNE embeddings locally by adapting lime. In *28th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning: ESANN2020*, pages 393–398. ESANN (i6doc. com), 2020.

Matthew R Blake, Scott D Holbrook, Joanna Kotwica-Rolinska, Eileen S Chow, Doris Kretzschmar, and Jadwiga M Giebultowicz. Manipulations of amyloid precursor protein cleavage disrupt the circadian clock in aging drosophila. *Neurobiology of disease*, 77: 117–126, 2015.

Chenghao Chen, Ratna Chaturvedi, Victoria Louis, Lauren E. North, Yongliang Xia, Maria Paz Gonzalez-Perez, Vikas Kumar, Qing Yu, and Patrick Emery. Membrane proteomics of the drosophila circadian neural network. *bioRxiv*, 2026. doi: 10.64898/2026.05.20.726616. URL https://www.biorxiv.org/content/early/2026/05/21/2026.05.20.726616.

Siheng Chen, Aliaksei Sandryhaila, José M. F. Moura, and Jelena Kovacevic. Signal denoising on graphs via graph filtering. In *2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 872–876, 2014. doi: 10.1109/GlobalSIP.2014.7032244.

Pawel Dlotko. Cubical complex. In *GUDHI User and Reference Manual*. GUDHI Editorial Board, 3.12.0 edition, 2026. URL https://gudhi.inria.fr/doc/3.12.0/group__cubical__complex.html.

Jens Eberlein, Bennett Davenport, Tom T Nguyen, Francisco Victorino, Kevin Jhun, Verena van der Heide, Maxim Kuleshov, Avi Ma'ayan, Ross Kedl, and Dirk Homann. Chemokine signatures of pathogen-specific T cells I: Effector T cells. *The Journal of Immunology*, 205(8):2169–2187, 10 2020. ISSN 0022-1767. doi: 10.4049/jimmunol.2000253. URL https://doi.org/10.4049/jimmunol.2000253.

Dibya Ghosh and Alvin Wan. A guide to MNIST: Deskewing, 2016. URL https://fsix.github.io/mnist/Deskewing.html.

Matthias Hein, Jean-Yves Audibert, and Ulrike von Luxburg. Graph laplacians and their convergence on random neighborhood graphs. *Journal of Machine Learning Research*, 8 (6), 2007.

I-Uen Hsu, Jeremy W. Linsley, Jade E. Varineau, Orie T. Shafer, and John Y. Kuwada. Dstac is required for normal circadian activity rhythms in drosophila. *Chronobiology International*, 35(7):1016–1026, 2018. doi: 10.1080/07420528.2018.1454937. URL https://doi.org/10.1080/07420528.2018.1454937. PMID: 29621409.

Samson J. Koelle, Hanyu Zhang, Marina Meila, and Yu-Chia Chen. Manifold coordinates with physical meaning. *Journal of Machine Learning Research*, 23(133):1–57, 2022. URL http://jmlr.org/papers/v23/19-644.html.

Samson J. Koelle, Hanyu Zhang, Octavian-Vlad Murad, and Marina Meila. Consistency of dictionary-based manifold learning. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li, editors, *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 4348–4356. PMLR, 02–04 May 2024. URL https://proceedings.mlr.press/v238/koelle24a.html.

Dhruv Kohli, Sawyer J. Robertson, Gal Mishne, and Alexander Cloninger. Robust tangent space estimation via laplacian eigenvector gradient orthogonalization, 2025. URL https://arxiv.org/abs/2510.02308.

Jan Lause, Philipp Berens, and Dmitry Kobak. Analytic Pearson residuals for normalization of single-cell RNA-seq UMI data. *Genome biology*, 22(1):258, 2021.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Dingbang Ma, Dariusz Przybylski, Katharine C Abruzzi, Matthias Schlichting, Qunlong Li, Xi Long, and Michael Rosbash. A transcriptomic taxonomy of *Drosophila* circadian neurons around the clock. *eLife*, 10:e63056, jan 2021. ISSN 2050-084X. doi: 10.7554/eLife.63056. URL https://doi.org/10.7554/eLife.63056.

Rebecca Marion, Adrien Bibal, and Benoît Frénay. BIR: A method for selecting the best interpretable multidimensional scaling rotation using external variables. *Neurocomputing*, 342:83–96, 2019. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2018.11.093. URL https://www.sciencedirect.com/science/article/pii/S0925231219301481. Advances in artificial neural networks, machine learning and computational intelligence.

Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction, 2020. URL https://arxiv.org/abs/1802.03426.

Nikolaos Mylonas, Ioannis Mollas, Nick Bassiliades, and Grigorios Tsoumakas. Exploring local interpretability in dimensionality reduction: Analysis and use cases. *Expert Systems with Applications*, 252:124074, 2024. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2024.124074. URL https://www.sciencedirect.com/science/article/pii/S0957417424009400.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322. doi: 10.1145/2939672.2939778. URL https://doi.org/10.1145/2939672.2939778.

Orie Thomas Shafer. 25 years of drosophila "sleep genes". *Fly*, 19(1):2502180, 2025.

Amit Singer. From graph to manifold laplacian: The convergence rate. *Applied and Computational Harmonic Analysis*, 21(1):128–134, 2006.

Amit Singer and Hau-tieng Wu. Vector diffusion maps and the connection Laplacian. *Communications on Pure and Applied Mathematics*, 65(8):1067–1144, 2012. doi: https://doi.org/10.1002/cpa.21395. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cpa.21395.

A.N. Tikhonov and V.I.A. Arsenin. *Solutions of Ill-posed Problems*. Halsted Press book. Winston, 1977. ISBN 9780470991244. URL https://books.google.com/books?id=ECrvAAAAMAAJ.

Roberta Turiello, Susanna S. Ng, Elisabeth Tan, Gemma van der Voort, Nazhifah Salim, Michelle C. R. Yong, Malika Khassenova, Johannes Oldenburg, Heiko Rühl, Jan Hasenauer, Laura Surace, Marieta Toma, Tobias Bald, Michael Hölzel, and Dillon Corvino. NKG7 is a stable marker of cytotoxicity across immune contexts and within the tumor microenvironment. *European Journal of Immunology*, 55(6):e51885, 2025. doi: https://doi.org/10.1002/eji.202551885. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/eji.202551885.

Stefan Van der Walt, Johannes L Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D Warner, Neil Yager, Emmanuelle Gouillart, and Tony Yu. scikit-image: image processing in python. *PeerJ*, 2:e453, 2014.

F Alexander Wolf, Philipp Angerer, and Fabian J Theis. SCANPY: large-scale single-cell gene expression data analysis. *Genome biology*, 19(1):15, 2018.

Kok Loon Wong, June Jing-Yi Tai, Wing-Cheong Wong, Hao Han, Xiaohui Sem, Wei-Hseun Yeap, Philippe Kourilsky, and Siew-Cheng Wong. Gene expression profiling reveals the defining features of the classical, intermediate, and nonclassical human monocyte subsets. *Blood*, 118(5):e16–e31, 2011. ISSN 0006-4971. doi: https://doi.org/10.1182/blood-2010-12-326355. URL https://www.sciencedirect.com/science/article/pii/S0006497120408389.

Zhenyue Zhang and Hongyuan Zha. Nonlinear dimension reduction via local tangent space alignment. In *International Conference on Intelligent Data Engineering and Automated Learning*, pages 477–481. Springer, 2003.

Grace XY Zheng, Jessica M Terry, Phillip Belgrader, Paul Ryvkin, Zachary W Bent, Ryan Wilson, Solongo B Ziraldo, Tobias D Wheeler, Geoff P McDermott, Junjie Zhu, et al. Massively parallel digital transcriptional profiling of single cells. *Nature communications*, 8(1):14049, 2017.

Tarek M Zikry and Genevera I. Allen. LIMEADE: Local interpretable manifold explanations for dimension evaluations. In *ICLR 2025 Workshop on Machine Learning for Genomics Explorations*, 2025. URL https://openreview.net/forum?id=kmLV911L80.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320, 04 2005.

ISSN 1369-7412. doi: 10.1111/j.1467-9868.2005.00503.x. URL https://doi.org/10.1111/j.1467-9868.2005.00503.x.

Appendix A. Additional Mathematical Details

A.1. Robust Tangent Space Estimation

Here we briefly describe how LEGO provides a robust estimate of the tangent space of a manifold $\mathcal{M}$ (Kohli et al., 2025). Let $\mathcal{T}^\varepsilon$ be a tubular neighborhood of thickness $\varepsilon > 0$ around $\mathcal{M}$. Then let $\Delta_{\mathcal{T}^\varepsilon}$ be the Laplace-Beltrami operator on $\mathcal{T}^\varepsilon$. The normalized Laplacian $\mathbf{L}$ constructed from X using a Gaussian kernel-weighted neighbor graph converges to $\Delta_{\mathcal{T}^\varepsilon}$, and the eigenvalues μ_m and eigenvectors φ_m of $\mathbf{L}$ converge to the eigenvectors and eigenfunctions of $\Delta_{\mathcal{T}^\varepsilon}$ (Singer, 2006; Hein et al., 2007; Belkin and Niyogi, 2008). Then around a point x_i, LEGO centers the k nearest neighbors $x_{j_1}, \ldots, x_{j_k}$ to x_i and the values of the first $M \ll n$ eigenvectors φ_m:

$$\mathbf{X}^{(i)} = \begin{bmatrix} x_{j_1} - x_i \\ \vdots \\ x_{j_k} - x_i \end{bmatrix} \text{ and } \varphi_m^{(i)} = \begin{bmatrix} \varphi_m(x_{j_1}) - \varphi_m(x_i) \\ \vdots \\ \varphi_m(x_{j_k}) - \varphi_m(x_i). \end{bmatrix} \tag{6}$$

Then taking an orthonormal basis $\mathbf{B}_\Phi$ of $\text{span}\{\varphi_1, \ldots, \varphi_M\}$, the gradients of the Laplacian eigenfunctions on $\mathcal{T}^\varepsilon$ can be estimated by $\widehat{\nabla}\varphi_m = \widehat{\mathbf{C}}_m \mathbf{B}_\Phi^\top$ where

$$\widehat{\mathbf{C}}_m = \min_{\mathbf{C}_m \in \mathbb{R}^{p \times M}} \frac{1}{n} \sum_{i=1}^{n} \left\| \mathbf{X}^{(i)} \mathbf{C}_m \mathbf{B}_\Phi^\top(x_i) - \varphi_m^{(i)} \right\|_2^2. \tag{7}$$

Since $\mathbf{B}_\Phi$ is an orthonormal basis, the solution to (7) is

$$\widehat{\mathbf{C}}_m = \begin{bmatrix} \mathbf{X}^{(1)\dagger} \varphi_m^{(1)} & \cdots & \mathbf{X}^{(n)\dagger} \varphi_m^{(n)} \end{bmatrix} \mathbf{B}_\Phi. \tag{8}$$

Here, $\mathbf{A}^\dagger$ denotes the pseudoinverse of $\mathbf{A}$. Then at x_i, we have

$$\widehat{\nabla}\Phi = \begin{bmatrix} \widehat{\nabla}\varphi_1(x_i) & \cdots & \widehat{\nabla}\varphi_M(x_i) \end{bmatrix}. \tag{9}$$

These estimated gradients span $T_{x_i}\mathcal{T}^\varepsilon$; taking the first d singular vectors provides a basis for the tangent space $T_{x_i}\mathcal{M}$ of the underlying manifold $\mathcal{M}$. Kohli et al. (2025) show that for a manifold $\mathcal{M}$ with tubular neighborhood $\mathcal{T}^\varepsilon$, the Laplacian eigenfunctions of $\mathcal{T}^\varepsilon$ whose gradients are normal to $\mathcal{M}$ lie deeper in the spectrum of $\mathcal{T}^\varepsilon$, so the singular vectors of (9) provide a tangent space estimate which is robust to noise. This also allows for accurate tangent space estimation using smaller neighborhoods, making it more robust to curvature as well.

A.2. The Connection Laplacian

Here we describe briefly how one constructs the connection Laplacian for a weighted neighbor graph $G = (X, A)$. For an edge between x_i, x_j, the weights are given by the Gaussian

kernel

$$K_{ij} = \exp\left\{-\frac{\|x_i - x_j\|^2}{\varepsilon}\right\}. \tag{10}$$

For a point x_i, let $\mathbf{B}^{(i)}$ be an orthogonal basis of $T_{x_i}\mathcal{M}$. We also assign each edge (i,j) an orthogonal transformation $\mathbf{O}_{ij} \in O(d)$, given by

$$\mathbf{O}_{ij} = \underset{\mathbf{O}\in O(d)}{\operatorname{argmin}} \left\|\mathbf{O} - \mathbf{B}^{(i)}{\mathbf{B}^{(j)}}^\top\right\|_{HS}, \tag{11}$$

where $O(d)$ is the group of $d \times d$ orthogonal matrices and $\|\mathbf{A}\|_{HS}$ is the Hilbert-Schmidt norm $\|\mathbf{A}\|_{HS}^2 = \operatorname{tr}(\mathbf{A}\mathbf{A}^\top)$. The transformations $\mathbf{O}_{ij}$ approximate the parallel transport operators between $T_{x_i}\mathcal{M}$ and $T_{x_j}\mathcal{M}$ (Singer and Wu, 2012), and can be computed using the singular value decomposition of $\mathbf{B}^{(i)}{\mathbf{B}^{(j)}}^\top$.

From $\mathbf{O}_{ij}$ we construct the $nd \times nd$ block matrix $\mathbf{S}$ whose (i,j)-th block is given by $\mathbf{S}(i,j) = K_{ij}\mathbf{O}_{ij}$. Let $\mathbf{D}$ be the diagonal $nd \times nd$ degree matrix whose i-th diagonal block is given by $\deg(x_i)\mathbf{I}_d$. Then the *(normalized) connection Laplacian* is the matrix

$$\mathcal{L} = \mathbf{I}_{nd} - \mathbf{D}^{-1}\mathbf{S}. \tag{12}$$

Appendix B. Additional Experimental Details

Here we provide additional experimental details for each dataset. Across all our experiments we take $\lambda_1 = \lambda_2$ in (3) and $\gamma = 1.0$ in (5). For denoising the explanation for a point x_i, we restrict our construction of the connection Laplacian to the 1-hop neighborhood of x_i in the k-nearest neighbor graph.

B.1. Two Swiss Rolls

We validate LTSREx on these Swiss rolls using $k = 8, 16, 32$ for the k-NN graph with four tangent space estimation methods: none (i.e. using the neighborhood in $\mathbb{R}^3$); unweighted local PCA (LPCA); weighted local PCA (WLPCA) with Gaussian kernel weights at $r = 1.0$; and LEGO using $M = 40$ eigenvectors. We also compare LTSREx without the denoising step and with denoising. In each run, we compute explanations for 500 randomly selected points. For the naive method, we automatically tune λ_1, λ_2 so that $\mathbf{W}^{(i)}$ has exactly $d = 2$ nonzero coefficient vectors at each point. For denoising, we threshold the coefficients so that only the top two features are retained. An explanation is correct if both features identified by the explanation match the ground truth features at that point, e.g. if a point is from $\mathcal{M}_1$ then its explanation should be (s_1, t_1). Results are averaged over five random seeds (Fig. 6).

B.2. Rotated MNIST

Dataset Generation. Using the train split of the MNIST dataset (Lecun et al., 1998), we generate rotated versions of each digit. First, we deskew each image, using code from Ghosh and Wan (2016). Then the deskewed image is assigned a random rotation between

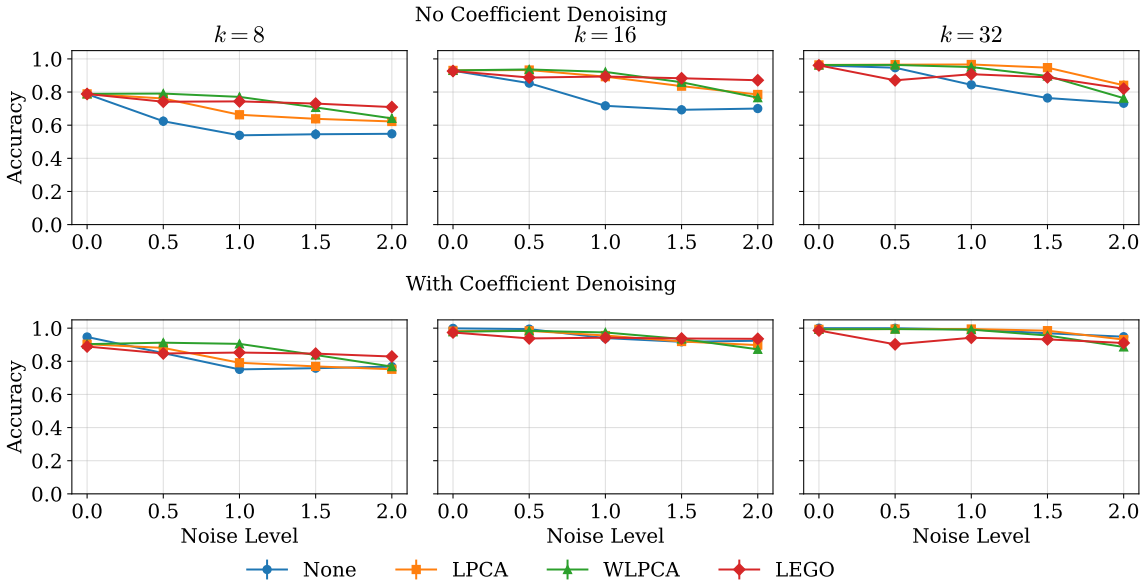

Figure 6: Accuracy using different tangent space estimation methods without denoising coefficients (top) and with denoising (bottom) on two rotated Swiss rolls.

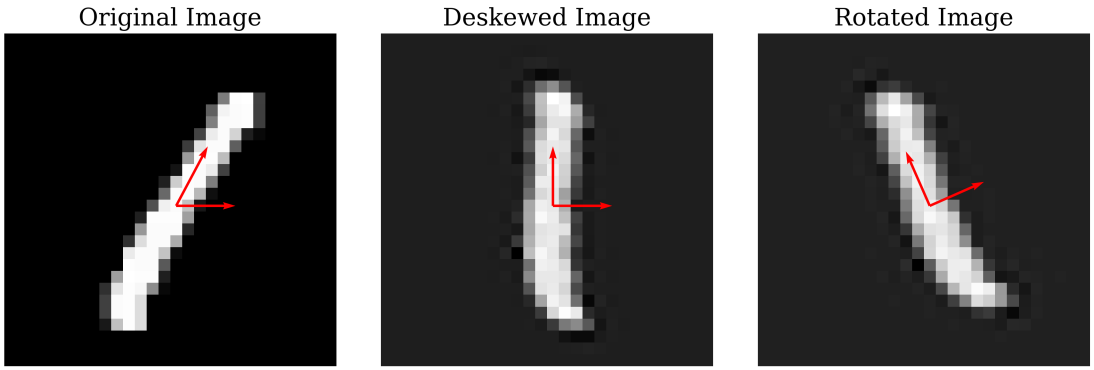

Figure 7: A deskewed and rotated "1" from MNIST

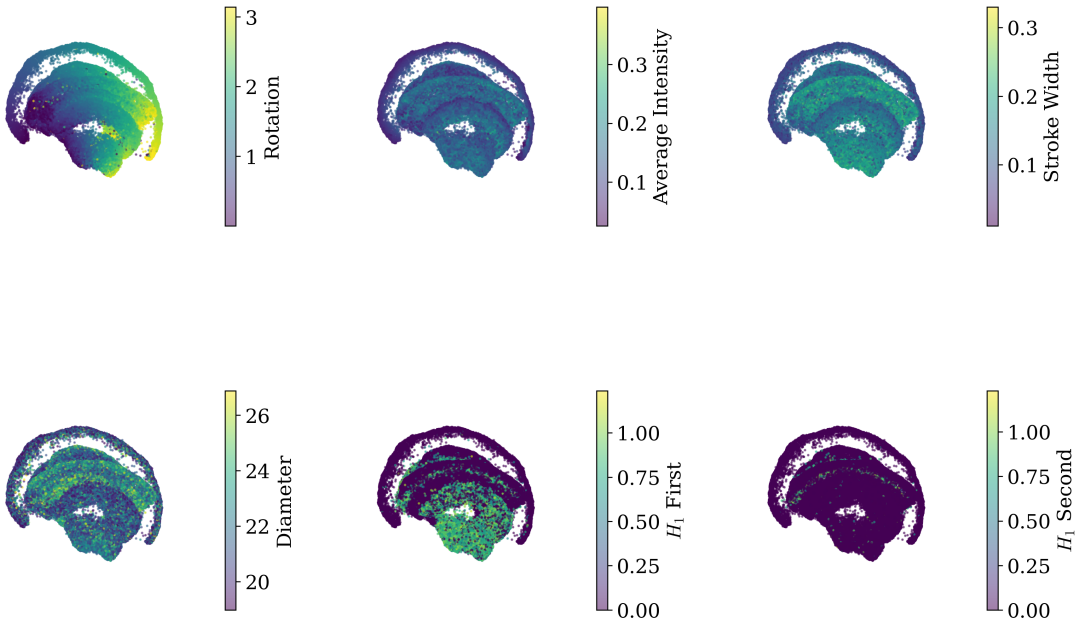

Figure 8: UMAP Embedding of rotated MNIST digits colored by interpretable features.

0 and π radians (Figure 7). We embed the rotated images in $\mathbb{R}^3$ with UMAP, using the $k = 16$ nearest-neighbor graph at a minimum distance of 0.3.

Then, on the rotated digits, we compute the following features as our interpretation candidates: rotation, average intensity, average stroke width, diameter, persistence of the two most persistent first homology classes. Figure 8 shows how the values of each feature vary over the UMAP embedding.

Rotation. This is given by the ground-truth rotation assigned to each image.

Average Intensity. We take the average pixel value for each image.

Stroke Width. Each deskewed image is binarized and skeletonized. The stroke width of each connected component in the skeleton is measured using the Scikit-image library (Van der Walt et al., 2014). If there are multiple connected components, we take the mean.

Diameter. We compute the convex hull of the deskewed image, and compute the diameter by taking the maximum distance between extremal points on the convex hull.

H_1 First and H_1 Second. Taking each image as a cubical complex and pixel intensity as a filtration value, we use the CubicalComplex library from gudhi (Dlotko, 2026) to compute the persistent H_1 classes of each image. We take the highest two persistence values among all detected H_1 classes to be the first and second H_1 feature values, respectively.

Experimental Parameters and Additional Results. For LTSREx, we use $k = 16$ nearest neighbors for tangent space estimation (computed from the UMAP embedding), $d = 2$, $\lambda_1 = \lambda_2 = 0.25$, and denoising parameter $\gamma = 1.0$.

We also present additional results. Figure 9 shows subsampled twos where the persistence of first homology was selected in the local explanation. The examples include twos with varying sizes of loop. Figure 10 shows the subsampled fours where the the most persis-

tent first homology class was selected in the local explanation. Notice that several appear near the border of the "4" and "9" clusters in the UMAP embedding. Several examples also have nontrivial first homology, or have very narrow openings at their tops. Finally, Figure 11 shows the subsampled eights whose *second* most persistent H_1 class was selected as explanatory. Many examples have only one closed loop, leaving the top loop open.

B.3. PBMC3k

We use the preprocessed version of the PBMC3k dataset provided by SCANPY (Wolf et al., 2018). We project preprocessed cell counts into $\mathbb{R}^4$ with UMAP ($k = 30$ neighbors, minimum distance 0.3) and apply LTSREx with $k = 128$, $d = 2$, $\lambda_1 = \lambda_2 = 0.05$, $\gamma = 1.0$. Here we also provide additional examples of cells with selected explanatory features (Figure 12).

B.4. *Drosophila* Clock Neurons

We obtain raw gene counts from Ma et al. (2021) (NCBI GEO GSE157504) and preprocess using SCANPY. Following the quality control metrics of (Ma et al., 2021), we remove cells which express too few (< 1000) or too many (> 6000) genes. We also filter genes, keeping only those with > 6000 Unique Molecular Identifiers (UMIs) and < 75000. We also filter out cells whose gene expression entropy is less than 5.5 nats. We then take the 3000 most highly variable genes and normalize the gene expression matrix by Pearson residuals (Lause et al., 2021). After further removing mitochondrial, ribisomal, and translational genes from consideration, we project the data into $\mathbb{R}^{50}$ using PCA. After preprocessing, we embed in $\mathbb{R}^{10}$ using UMAP ($k = 30$ neighbors, minimum distance 1.0). We apply our method to a cell within this cluster using $k = 128$, $d = 2$, $\lambda_1 = \lambda_2 = 0.05$, and $\gamma = 1.0$.

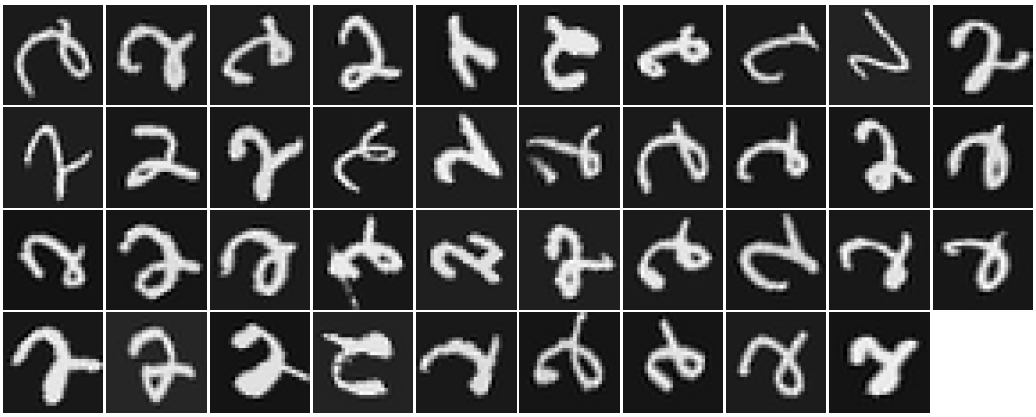

Figure 9: Subsampled twos where "H_1 First" was selected as an explanatory feature.

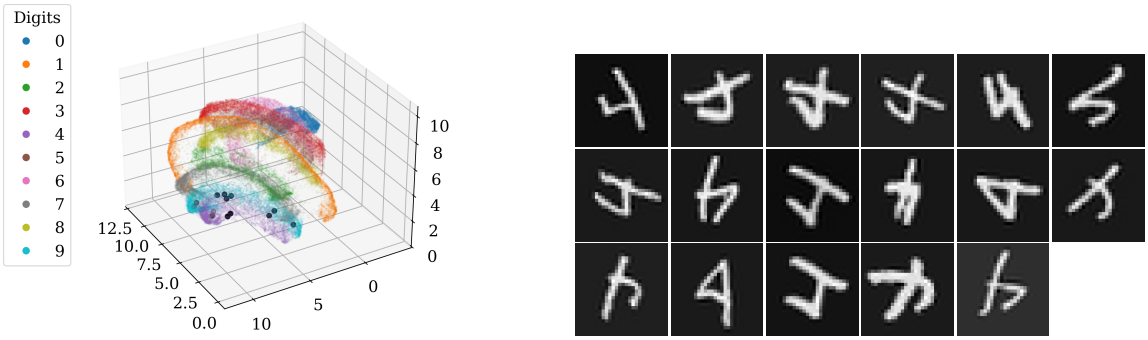

Figure 10: Location in UMAP embedding of fours where "H_1 First" was selected as an explanatory feature (left) and corresponding images (right).

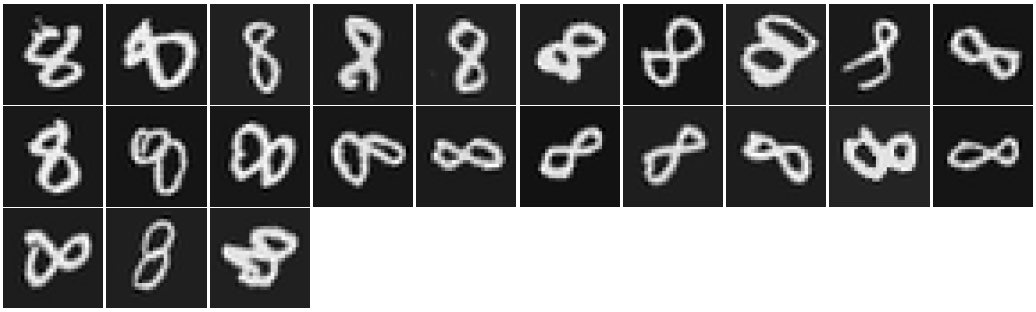

Figure 11: Subsampled eights where "H_1 Second" was selected as an explanatory feature.

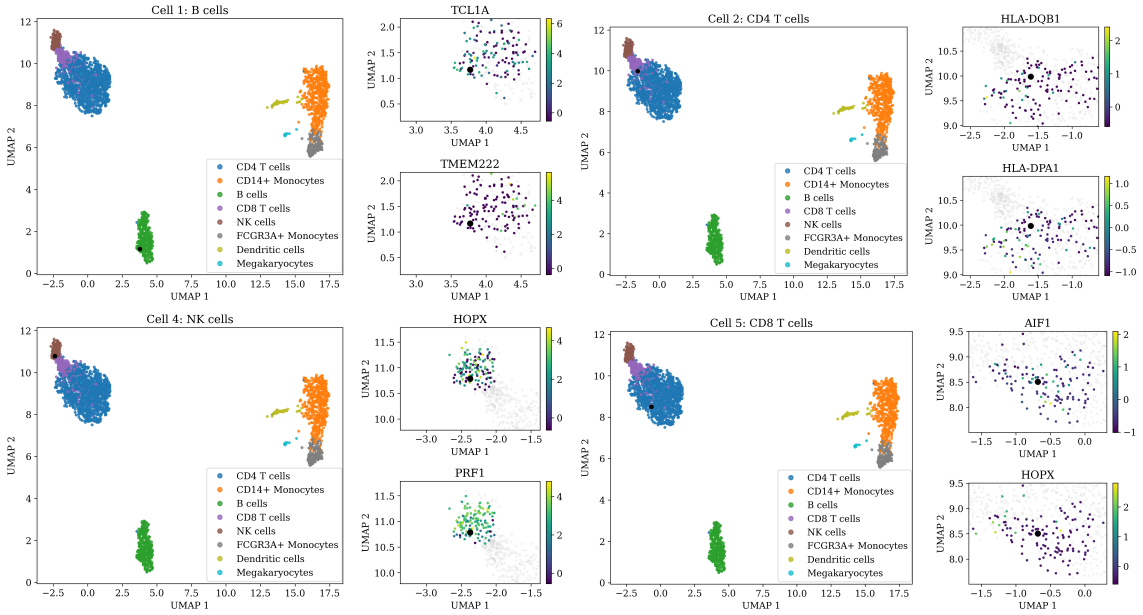

Figure 12: Additional results from PBMC3k dataset. (Cells 0 and 3 were used to create Figure 5.)