Dimension Reduction with Locally Adjusted Graphs

Yingfan Wang*, Yiyang Sun*, Haiyang Huang*, Cynthia Rudin

Duke University
{yingfan.wang, yiyang.sun, haiyang.huang, cynthia.rudin}@duke.edu

Abstract

Dimension reduction (DR) algorithms have proven to be extremely useful for gaining insight into large-scale high-dimensional datasets, particularly finding clusters in transcriptomic data. The initial phase of these DR methods often involves converting the original high-dimensional data into a graph. In this graph, each edge represents the similarity or dissimilarity between pairs of data points. However, this graph is frequently suboptimal due to unreliable high-dimensional distances and the limited information extracted from the high-dimensional data. This problem is exacerbated as the dataset size increases. If we reduce the size of the dataset by selecting points for a specific sections of the embeddings, the clusters observed through DR are more separable since the extracted subgraphs are more reliable. In this paper, we introduce LocalMAP, a new dimensionality reduction algorithm that dynamically and locally adjusts the graph to address this challenge. By dynamically extracting subgraphs and updating the graph on-the-fly, LocalMAP is capable of identifying and separating real clusters within the data that other DR methods may overlook or combine. We demonstrate the benefits of LocalMAP through a case study on biological datasets, highlighting its utility in helping users more accurately identify clusters for real-world problems.

Code — https://github.com/williamsyy/LocalMAP
Appendix — https://arxiv.org/abs/2412.15426

1 Introduction

Dimension reduction (DR) is an important data visualization strategy for understanding the structure of complicated high-dimensional datasets. DR is used extensively in image, text, and biomedical datasets, particularly -omics (Cao et al. 2019; Becht et al. 2019; Amezquita et al. 2020; Dries et al. 2021; Atitey, Motsinger-Reif, and Anchang 2024; Böhm, Berens, and Kobak 2023; Mu, Bhat, and Viswanath 2017; Raunak, Gupta, and Metze 2019). Beginning in the original high-dimensional space, DR methods typically first abstract the data into a *graph*, with each edge representing either similarity (positive edge) or dissimilarity (negative edge). This graph is then used to optimize the low-dimensional embedding by applying attractive forces along the positive edges and repulsive forces along the negative edges. Clearly, the quality of the graph, which connects the original high-dimensional data to the low-dimensional embedding, is critical to this process. However, given the complex nature of high-dimensional data, it is challenging to construct a graph that accurately represents all patterns and structures within the data, and graphs are usually sub-optimal.

In this work, we provide two key insights as to why the graphs could be flawed. First, we found that high-dimensional distances become less informative due to the curse of dimensionality, where measurements of similarity and dissimilarity determined using high-dimensional distances do not necessarily reflect similarities and dissimilarities along the data manifold. This issue is more pronounced in DR methods that select a small group of nearest neighbors (NNs) to form positive edges of the graph and apply strong attractive forces along these edges. When a "false" positive edge is constructed based on an inaccurate similarity measure between a pair of points that are actually dissimilar, strong attractive forces will pull them closer. If there are many such false positive edges, the DR method will generate overlapping clusters without clear boundaries, even if the data have distinct clusters. Since DR methods are unsupervised, the user would not be able to determine that distinct clusters exist – they would be blurred together in the DR result. As we show in this work, eliminating these false positive edges can dramatically improve DR projections.

Our second insight is that missing edges (lack of negative "further pair" edges between far points in high-dimensional space) also contribute to unwanted overlapping of clusters, since such edges help to define boundaries between clusters. However, as the scale of the dataset increases, the set of negative edges become both insufficient and less effective, as discussed in Section 4. Some of the most important applications of DR (single cell, transcriptomics) generate large high-dimensional datasets with many clusters, so it is important that there are enough further pair (FP) edges to separate them.

Based on the two insights discussed above, we propose LocalMAP, a new DR algorithm that dynamically and locally adjusts the graph *during dimension reduction* to address the aforementioned issues with graph construction. LocalMAP *detects false positive edges and removes them*, using ideas

*These authors contributed equally.

similar to outlier detection in robust statistics. This allows clusters to separate. It also *adds more further pair edges* dynamically, allowing crisper cluster boundaries to appear. Together, these ideas within LocalMAP produce clear, crisp separated clusters where other methods fail.

Figure 1 shows the results of several high-quality DR approaches, including t-SNE (van der Maaten and Hinton 2008), UMAP (McInnes, Healy, and Melville 2018) and PaCMAP (Wang et al. 2021), on the MNIST dataset where there are 10 separated clusters in the high-dimensional space. Algorithms t-SNE, UMAP and PaCMAP generate DR embeddings without clear boundaries between clusters, while LocalMAP generates a high-quality DR embedding with well-defined boundaries that are visible even without class information provided to the algorithm.

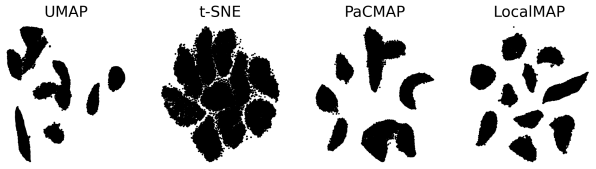

Figure 1: DR embeddings on MNIST dataset which contains 10 digit classes. Our LocalMAP method is on the right. The colored embeddings with true labels are shown in Figure 5.

In this work, we provide more evidence for the insights discussed above, in terms of both theory based on simple clustering models and empirical evidence. We also provide LocalMAP and a comprehensive evaluation of it, based on Huang et al. (2022). We provide case studies showing that LocalMAP is able to find true clusters – and not false clusters – more reliably than other DR approaches.

2 Related Work

While DR methods that preserve *global structure* date back to 1901 with PCA (Pearson 1901), multidimensional scaling (Torgerson 1952), and non-negative matrix factorization (Lee and Seung 1999), methods that preserve *local structure* have become indispensable and ubiquitous in numerous applications, particularly in -omics research. Local methods include Isomap (Tenenbaum, de Silva, and Langford 2000), Local Linear Embedding (LLE) (Roweis and Saul 2000), Laplacian Eigenmap (Belkin and Niyogi 2001), and more recent Neighborhood Embedding (NE) algorithms like t-SNE (van der Maaten and Hinton 2008), LargeVis (Tang et al. 2016), UMAP (McInnes, Healy, and Melville 2018), PaCMAP (Wang et al. 2021), TriMAP (Amid and Warmuth 2019), NCVis (Artemenkov and Panov 2020), h-NNE (Sarfraz et al. 2022), Neg-t-SNE (Damrich et al. 2023) and InfoNC-t-SNE (Damrich et al. 2023) and many others (Van Assel et al. 2022; Zu and Tao 2022). Global methods typically are not able to preserve separations between clusters or clearly display the data manifold, whereas local methods can sometimes do so; thus, they provide a unique perspective on the data that is difficult to gain in any other way.

Because DR methods are unsupervised, there is no common objective function like there would be for supervised learning (e.g., classification error). However, there are principles that reliable DR loss functions typically obey (Wang et al. 2021). We will discuss those in Appendix C, as LocalMAP obeys these principles based on what it inherits from PaCMAP.

There are works that discuss how DR methods' behavior is affected by different components of the DR algorithm. For example, there are several papers discussing the challenges of tuning parameters and applying these methods in practice (Wattenberg, Viégas, and Johnson 2016; Cao and Wang 2017; Nguyen and Holmes 2019; Belkina et al. 2019; Kobak and Linderman 2021). Wang et al. (2021); Böhm, Berens, and Kobak (2020) also discuss how changing different graph components and loss functions affect DR methods' behavior.

None of the above approaches adjusts the graph itself during DR optimization. The graph is typically considered to be fixed as ground truth; the graph information is analogous to the labels for classification tasks that are also considered ground truth. However, recent work has found improvements in classification performance by identifying possibly mislabeled points and omitting them (Chen et al. 2019; Han et al. 2018; Northcutt, Athalye, and Mueller 2024), showing that there is value in identifying and removing labels that appear to be wrong *during the classification process*. There is also a field of robust statistics that identifies and omits outliers that would wreck performance, and aims to ensure results are robust to assumptions on the data distribution (Martin et al. 2018; Li, Socher, and Hoi 2020). Our work is thus unique in extending these types of idea to DR, and not trusting every graph element as if it were correct. As discussed, we know the graph is probably wrong when the Euclidean distance is not the same as the distance along the data manifold (geodesic distance).

3 Review of PaCMAP

The proposed LocalMAP algorithm starts from a (faulty) embedding that is already formed. We first review PaCMAP since LocalMAP can start after its first two phases. Consider a dataset $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ consisting of n points in a high-dimensional space. PaCMAP seeks to find a low-dimensional embedding $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n\}$, where each $\mathbf{y}_i$ corresponds to $\mathbf{x}_i$. PaCMAP first constructs a graph in high-dimensional space with three kinds of pairs: NN pairs (near neighbors in the high-dimensional space), MN pairs (mid-near pairs, close but not as close as neighbors), and FP pairs (further point pairs, far in the high-dimensional space). Optimization of the low-dimensional embedding $\mathbf{Y}$ is performed using a simple objective:

$$Loss^{\text{PaCMAP}} = w_{\text{NN}} \cdot \sum_{(i,j):\,\text{NN}} \frac{\tilde{d}_{ij}}{C_{Med} + \tilde{d}_{ij}} + w_{\text{MN}} \cdot \sum_{(i,k):\,\text{MN}} \frac{\tilde{d}_{ik}}{C_{Lg} + \tilde{d}_{ik}} + w_{\text{FP}} \cdot \sum_{(i,l):\,\text{FP}} \frac{1}{1 + \tilde{d}_{il}} \tag{1}$$

where $\tilde{d}_{ij} = d_{ij}^2 + 1 = \|\mathbf{y}_i - \mathbf{y}_j\|^2 + 1$. C_{Lg} and C_{Med} are nontunable parameters controlling the scale of the embed-

ding, set at $\sim$10K and $\sim$10. The weights w_{NN}, w_{MN}, and w_{FP} are adjusted to balance attraction and repulsion. Neither the weights nor the parameters should be modified by users; they perform well across datasets (Wang et al. 2021).

4 Dynamically and Locally Adjusted Graph

Insight 1: False Positive Nearest Neighbor Edges Pull Nearby Clusters Together, Losing Clear Boundaries Between Them

Let us consider an experiment to illustrate this. DR methods apply attractive forces along the NN edges (between points that are close in the high-dimensional space) and apply repulsive forces along FP edges (points that are far in the high-dimensional space). An NN edge between two points is *preserved* if these points are placed nearby in the low-dimensional embedding. Given the limited capacity of low-dimensional spaces and the complex nearest-neighbor relationships in high-dimensional datasets, it is clear that not all NN edges can be preserved during dimensionality reduction. Further, we do not want to preserve all high-dimensional NNs, since the Euclidean distance neighbors are not always the true neighbors along the actual data manifold. Therefore, the question becomes how to identify which NNs to preserve.

The preservation of specific NN edges depends on intricate dynamics during the embedding optimization, which utilizes the graph extracted from the high-dimensional data. For a pair of NNs (i, j), if they share a greater number of common neighbors, which provide attraction along NN paths, it is more likely that these two points will be positioned close to each other in the low-dimensional space. We hypothesize that these dynamics sometimes correct erroneous edges generated by unreliable high-dimensional distances. If true, this hypothesis implies we can adjust the graph dynamically based on this information and further optimize the embedding using the revised graph.

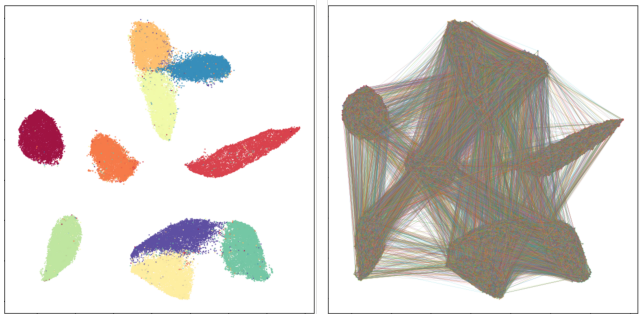

Figure 2: Visualization of NN edge connections of PaCMAP embedding on MNIST dataset.

Figure 2 presents a PaCMAP embedding of the MNIST dataset on the left and the corresponding NN edges' visualization on the right. The figure highlights a substantial number of NN edges bridging distinct clusters that are widely separated. When two points connected by such an NN edge are distant in the final embedding, it suggests that the dynamics of the optimization process have separated them, indicating that the NN edge might be erroneous. Despite this, the erroneous NN edge continues to exert an attractive force, pulling these two points towards each other. If numerous such NN edges exist between two clusters, they can lead to clusters being falsely connected, rather than clearly delineated with distinct boundaries.

Insight 2: Insufficient and Ineffective Negative Further Point Edges Fail To Create Clear Boundaries Between Nearby Clusters, Particularly in Large Datasets

Ideally, the repulsive forces along the FP edges should separate nearby clusters. However, when the dataset becomes larger, with more clusters and larger samples, it is more likely that the sampled FP edges in DR algorithms are insufficient.

Theorem 1. *Assume all points between two clusters are approximately equidistant, so that the probability of constructing a positive pair between points from these clusters is constant. The ratio between the number of NN edges to the number of FP edges of PaCMAP between two clusters increases with the number of data points in a dataset.*

Proof. Consider a dataset with n data points distributed across m clusters $C_1, C_2, ..., C_m$, where each cluster C_i contains n_i data points $(n_1 + n_2 + ... + n_m = n)$. For any two clusters C_i and C_j, by assumption, we have:

$$\forall x_i \in C_i, x_j \in C_j, \quad P(x_i, x_j \text{ are NNs}) = p_{ij}$$

where p_{ij} is constant.

FP edges for a given point are sampled randomly from all non-NN points. For each point, n_{FP} FP points are randomly selected, where n_{FP} is a constant defaulting to 20 for PaCMAP. Thus, the expected number of NNs and FPs between C_i and C_j are

$$\begin{aligned} \mathbb{E}(\#\text{ of NNs between } C_i, C_j) &= n_i n_j p_{ij} \\ \mathbb{E}(\#\text{ of FPs between } C_i, C_j) &= \frac{2 n_i n_j n_{\text{FP}}}{n}, \end{aligned} \tag{2}$$

because each point $i \in C_i$ selects $\frac{n_j}{n} \cdot n_{\text{FP}}$ FP pairs, the total number of points in C_i sampled from C_i to C_j is $\frac{n_i \cdot n_j \cdot n_{\text{FP}}}{n}$ FP pairs. Similarly, $\frac{n_i \cdot n_j \cdot n_{\text{FP}}}{n}$ FP pairs are sampled from all points in C_j between C_i and C_j. Therefore, $\frac{2 n_i n_j \cdot n_{\text{FP}}}{n}$ total FP pairs are sampled between these two clusters. Therefore, the ratio between the number of NN edges and the number of FP edges is

$$\frac{\mathbb{E}(\#\text{ of NNs between } C_i, C_j)}{\mathbb{E}(\#\text{ of FPs between } C_i, C_j)} = \frac{n \cdot p_{ij}}{2 n_{\text{FP}}}.$$

Considering that p_{ij} is unaffected by n and n_{FP}, n_i and n_j are constants, the ratio increases with n. This result is true for all clusters C_i and C_j. Thus, for each pair of clusters, the ratio of NN edges to FP edges grows linearly in n. This completes the proof. □

This phenomenon is further illustrated in Figure 3. When DR is applied to the entire MNIST dataset, images of six

different digits are grouped into two large clusters, with each cluster containing three digit classes. However, when DR is applied separately to each of these large clusters, they are successfully separated into distinct clusters, each representing a single digit class. It shows that when the sample size increases, usually more structure is involved in the dataset, which indicates higher complexity and larger number of classes.

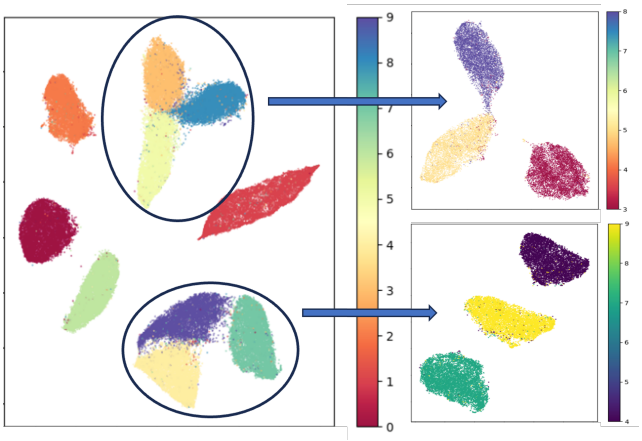

Figure 3: *Left:* PaCMAP embedding on the entire MNIST dataset where six clusters are groups into two large ones (each with three digit classes). *Right*: PaCMAP embeddings on each of the two groups of three digit classes. The right embeddings work when the left do not because the partial datasets are smaller and thus do not suffer from the problem identified in Insight 2.

5 LocalMAP

The insights above suggest a new approach to DR that keeps information *local*, so as to mitigate the problem of Insight 2, and to decrease the impact of false positive edges, avoiding the pitfall of Insight 1. The way LocalMAP handles this is to *replicate a small-scale DR within its large-scale computation.* Specifically, we increase both NN attractive forces and FP repulsive forces locally compared to standard DR approaches, adjusting weights *dynamically* as we learn more about which data points are false positives.

This approach has few downsides, if any. Because it sees more local information, and because it reduces the impact of false positive edges, LocalMAP is able to better capture local structure. It is slightly more computationally expensive than some of the regular DR approaches and faster than others, even for large datasets, as we show in Section 8.4.

LocalMAP Algorithm

As discussed, LocalMAP handles Insight 1 and 2 by increasing local computation. LocalMAP is shown in Algorithm 1 with a detailed version in Appendix A. Major differences from previous DR approaches are:

LocalMAP Computation 1. Adjusting the weights for NN edges dynamically and locally according to the low-dimensional distance during optimization. This is done according to a set of principles listed below.

LocalMAP Computation 2. Resampling FP edges to be local. This allows LocalMAP to separate nearby clusters. In practice, local FP edges are randomly selected non-neighbor pairs with distance no larger than the average low-dimension distance among all nearest clusters pairs; we call this the *proximal cluster distance commons* $\bar{d}_{\text{adj}}$.

Both LocalMAP Computation 1 and LocalMAP Computation 2 apply to the final stage of optimization. LocalMAP uses earlier stages from another DR approach to handle global structure placement, which is sufficient as set up for LocalMAP's computations.

LocalMAP's Principles for NN Edge Weighting in LocalMap Computation 1

LocalMAP creates several central aspects for DR. Specifically, these principles are:

1. Increased weights for NNs that are close in low-dimensional space. (These are estimated to be true positive pairs.)
2. Decreased weights for NNs that are far in low-dimensional space. (These are estimated to be false positive pairs.)
3. Avoid extremely large forces, ensuring stable convergence.
4. The weighting function needs to be simple, easy to combine with the NN loss terms, and fast to compute.

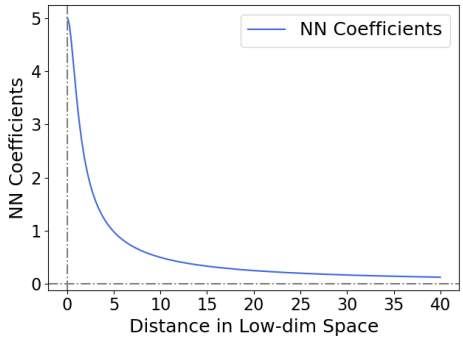

Figure 4: Curve of $\text{Coefficient}_{\text{NN}}$.

A natural choice to achieve Principles 1 and 2 and 4 is to use a weighting function that is inversely proportional to the low-dimensional distance. However, this could lead to very large values for small low-dimensional distances, making the convergence unstable, violating Principle 3.

Moreover, to define "close" in Principles 1 and 2, the embedding scale should be considered. We consider the average distance between adjacent clusters in PaCMAP embedding, $\bar{d}_{\text{adj}}$, which is approximately 10 based on observations. The

Algorithm 1: Implementation of LocalMAP

Require: $\mathbf{X}$ - data matrix, $\bar{d}_{\text{adj}}$ from Section 5, parameter $C_{Med}\sim 10$.
Ensure:

- Initialized low-dimensional embedding $\mathbf{Y}$.
- Construct neighbor pairs, mid-near pairs and further pairs according to high-dimensional distance.
- Optimize $\mathbf{Y}$ using the first two training phases of PaCMAP.
- Optimize $\mathbf{Y}$ using loss:

$$Loss(\mathbf{Y}) := \sum_{(i,j):\,\text{NN}} \frac{\bar{d}_{\text{adj}} \cdot \sqrt{\tilde{d}_{ij}}}{2(C_{Med} + \tilde{d}_{ij})} + \sum_{(i,l):\,\text{FP}} \frac{1}{1+\tilde{d}_{il}},$$

where $\tilde{d}_{ij} = \|\mathbf{y}_i - \mathbf{y}_j\|^2 + 1$.
The FP pairs are resampled every 10 iterations to stay *local* so that all FPs satisfies $\|\mathbf{y}_i - \mathbf{y}_l\| \le \bar{d}_{\text{adj}}, \forall (i,l) \in$ *FP pairs*.
return $\mathbf{Y}$

midpoint between the two clusters ($\frac{\bar{d}_{\text{adj}}}{2}$) could serve as a threshold to determine whether to increase or reduce the attractive force. The coefficient should be greater than 1 when the attraction force needs to be increased and less than 1 when it needs to be reduced.

To achieve all four principles, we choose a weighting function as follows:

$$\text{Coefficient}_{\text{NN}}(d_{ij}) = \frac{\frac{\bar{d}_{\text{adj}}}{2}}{\sqrt{d_{ij}^2+1}} = \frac{\bar{d}_{\text{adj}}}{2\sqrt{\tilde{d}_{ij}}}.$$

Based on the above equation, when $\frac{\bar{d}_{\text{adj}}}{2} > \sqrt{\tilde{d}_{ij}} \approx d_{ij}$, the attractive force along the (i,j) pair increases, and this force would decrease when $\frac{\bar{d}_{\text{adj}}}{2} < \sqrt{\tilde{d}_{ij}} \approx d_{ij}$.

Figure 4 shows how $\text{Coefficient}_{\text{NN}}$ changes with the low dimensional distance between pairs of NN. Implementing this by adapting PaCMAP's loss and setting C_{Med} to 10 to fix the scale of the embedding, its NN loss term becomes:

$$\begin{aligned} Loss_{\text{NN}} &= w_{\text{NN}} \cdot \sum_{(i,j):\,\text{NN}} \frac{\tilde{d}_{ij}}{C_{Med}+\tilde{d}_{ij}} \cdot \frac{\bar{d}_{\text{adj}}}{2\sqrt{\tilde{d}_{ij}}} \\ &= w_{\text{NN}} \cdot \sum_{(i,j):\,\text{NN}} \frac{\bar{d}_{\text{adj}} \cdot \sqrt{\tilde{d}_{ij}}}{2(C_{Med}+\tilde{d}_{ij})}. \end{aligned} \tag{3}$$

As stated in Section 2, DR methods are unsupervised and no common objective function exists; we proved in Theorem 2 that LocalMAP also satisfies the six principles mentioned in Wang et al. (2021).

Theorem 2. *LocalMAP's loss function obeys the six principles of Wang et al. (2021) for any choices of $\bar{d}_{adj}$, excluding NNs that have low dimensional distances larger than a threshold with value $\sqrt{C_{Med}-1}$ (i.e., possible false positives), where we set $C_{Med}\sim 10$ across datasets to determine the scale of the embedding.*

The proof is given in Appendix C.

6 Case Study

Figure 5 shows the results for several DR methods on three datasets. Two of the datasets are handwritten digit datasets (LeCun, Cortes, and Burges 2010; Hull 1994), which means that there are distinct clusters in high dimensions, but also these datasets are special in that Euclidean distance (which is used for defining the graph for DR) is not the same as the geodesic distance along the data manifold, meaning there will be plenty of false positive edges, similar to those shown in Figure 2. The last dataset is a biological dataset (Kang et al. 2018) that contains multiple cell types that are labeled.

We observe that *LocalMAP does a far better job in separating clusters on all three datasets*. A quantitative result is given by the Silhouette score shown in the figures. The definition of Silhouette score is in Appendix E. LocalMAP's scores are superior, confirming what we see visually in these DR plots. There are 5 additional datasets analyzed in Section 8 and and visualized in Appendix H, all with similarly impressive visual results.

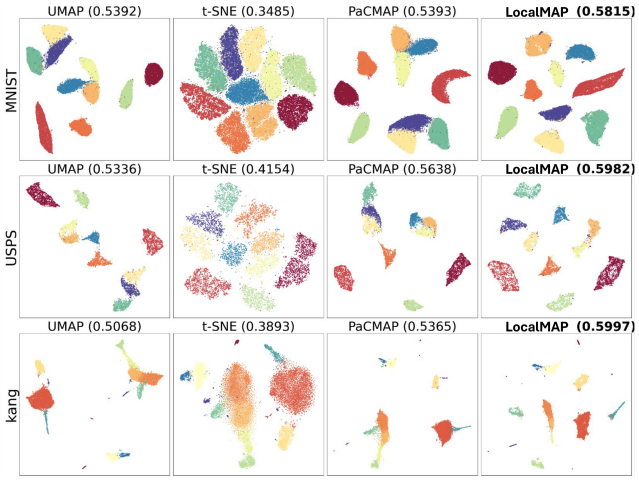

Figure 5: Case study on MNIST (LeCun, Cortes, and Burges 2010), USPS (Hull 1994) and Kang (Kang et al. 2018). The Silhouette scores are shown in parentheses.

7 Ablation Study

To understand how each part of the modification contributes to the DR results — an adjustment of NN edge weights and local FP edge resampling — we conducted an ablation study, as shown in Figure 6. The left plot in the figure displays the embedding generated by LocalMAP with adjusted NN edge weights but without local FP resampling. In this embedding, the clusters have higher density, yet the separation between nearby clusters remains inadequate. The right plot illustrates the embedding generated by LocalMAP with local FP resampling but without adjusted NN edge weights. Although the clusters are more dispersed, the separation between nearby clusters is still insufficient. Therefore, these results indicate that both NN edge weight adjustment and local FP edge resampling are necessary to achieve clear separation between clusters. These observations are clear with MNIST, which is why it is useful to work with this dataset; the same observations persist with other datasets.

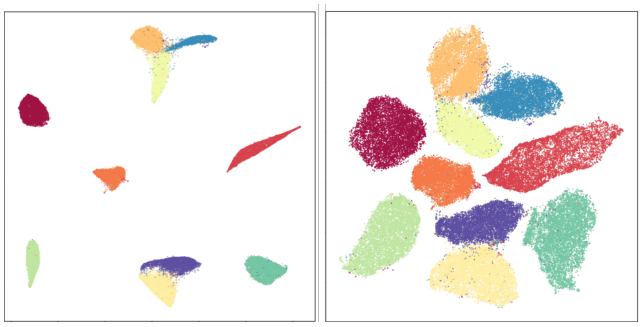

Figure 6: Ablation study of LocalMAP on MNIST dataset. **Left**: LocalMAP with adjusted NN edge weights but without local FP resampling. **Right**: LocalMAP with local FP resampling but without adjusted NN edge weights.

8 Experiments

8.1 Experimental Setup

Datasets. Inspired by other DR studies (Tang et al. 2016; McInnes, Healy, and Melville 2018; Amid and Warmuth 2019; Huang et al. 2022), the following datasets are used to evaluate and compare DR methods: MNIST (LeCun, Cortes, and Burges 2010), FMNIST (Xiao, Rasul, and Vollgraf 2017), USPS (Hull 1994), COIL20 (Nene, Nayar, and Murase 1996), 20NG (Lang 1995), Seurat (Stuart et al. 2019), Kang (Kang et al. 2018), Human Cortex (Zhu et al. 2023) and CBMC (Stoeckius et al. 2017).

Algorithms. We evaluate LocalMAP in comparison with several other DR techniques: t-SNE (van der Maaten and Hinton 2008), UMAP (McInnes, Healy, and Melville 2018), PaCMAP (Wang et al. 2021), TriMap (Amid and Warmuth 2019), PHATE (Moon et al. 2017), NCVis (Artemenkov and Panov 2020), h-NNE (Sarfraz et al. 2022), Neg-t-SNE (Damrich et al. 2023), and InfoNC-t-SNE (Damrich et al. 2023). For the t-SNE algorithm, we utilize the openTSNE implementation (Poličar, Stražar, and Zupan 2023).

Computational Environment. All experiments were run on a 12 Core Intel(R) Xeon(R) CPU E5640 @ 2.67GHz with a GPU RTX2080Ti, with memory limit 32G. All experiments were run 10 times for error bar computation.

Biological Dataset Preprocessing. For the single-cell datasets, all variables were normalized and log-transformed by the `SCANPY` package (Wolf, Angerer, and Theis 2018). The top 1000 variance genes within each dataset were selected before applying LocalMAP. For the datasets that have different batches, the ComBat algorithm (Johnson, Li, and Rabinovic 2007) was used to remove batch effects. The detailed dataset description are shown within Appendix D.

8.2 Experimental Results: Silhouette Score

Evaluation of DR approaches is challenging because DR methods are unsupervised. There are numerous evaluation methods that characterize different aspects of performance; an overview is given by Huang et al. (2022), though several DR papers propose their own evaluation metrics (e.g., Amid and Warmuth 2019). In this work, we use the silhouette score (Rousseeuw 1987) as our evaluation metric because it captures how clearly the clusters are separated, which no other metric captures. The definition of the sihouette score is in Appendix E.

All the methods are evaluated based on their default hyperparameters. The result in Table 1 shows that LocalMAP separates clusters better than other approaches. This table simply quantifies the high quality clusters we see visually in the LocalMAP plots throughout this paper. We have also shown the performance of UMAP, LargeVis, t-SNE, and PHATE with tuned hyperparameters, and additional results are in Table 6 in Appendix J. *LocalMAP shows a better separation among clusters.*

8.3 Experimental Results: Posthoc Classification

DR methods are unsupervised. Here, we consider an evaluation of whether class information, which is not presented to the DR algorithm, is preserved during the process of DR. This is a type of local structure evaluation, but unlike the silhouette score, it does not consider the margins between classes and has several other problems as an evaluation metric, discussed in Appendix F. Essentially, clusters that visually appear merged or are broken up into subclusters (i.e., poor DR plots) can still yield high posthoc classification scores. Table 3 and 7 in Appendix J show the results, which is that LocalMAP achieves similar posthoc classification performance to 11 state-of-the-art DR methods. (But, as discussed, this evaulation measure is problematic.)

8.4 Experimental Results: Runtime

Table 2 shows runtimes for several algorithms. LocalMAP is slightly more computationally intensive than other methods because it resamples the graph dynamically, but it is still efficient. Efficiency is typically not as important as DR quality, as evaluated in Section 8.2.

	MNIST	FMNIST	USPS	COIL20	20NG	Kang	Seurat	Human Cortex	CBMC
PCA	0.02±0.00	-0.03±0.00	0.10±0.00	0.01±0.00	-0.19±0.00	0.12±0.00	-0.06±0.00	-0.08±0.00	-0.11±0.00
t-SNE	0.35±0.00	0.12±0.00	0.42±0.00	0.41±0.00	-0.11±0.00	0.40±0.01	0.22±0.01	0.11±0.01	0.15±0.01
UMAP	0.52±0.01	**0.19±0.00**	0.53±0.00	**0.58±0.01**	-0.15±0.01	0.51±0.00	0.30±0.00	0.12±0.02	0.22±0.00
PaCMAP	0.54±0.01	**0.19±0.00**	0.56±0.00	0.51±0.02	-0.11±0.01	0.53±0.00	0.31±0.00	0.13±0.01	0.22±0.00
LargeVis	0.49±0.05	0.11±0.03	0.41±0.12	0.38±0.01	-0.13±0.01	0.44±0.01	0.25±0.01	0.10±0.02	0.17±0.00
TriMAP	0.41±0.00	0.17±0.00	0.48±0.00	0.47±0.00	-0.13±0.00	0.55±0.00	**0.32±0.00**	0.07±0.00	0.21±0.00
PHATE	0.26±0.02	0.11±0.01	0.27±0.01	0.33±0.00	-0.21±0.01	0.48±0.02	0.27±0.01	-0.09±0.01	0.06±0.01
HNNE	0.21±0.03	0.06±0.04	0.23±0.00	0.03±0.00	-0.34±0.03	0.39±0.06	-0.00±0.03	-0.09±0.06	0.12±0.05
Neg-t-SNE	0.48±0.00	**0.19±0.00**	0.48±0.00	0.44±0.01	-0.11±0.00	0.53±0.00	**0.32±0.00**	0.12±0.00	**0.24±0.00**
NCVis	0.38±0.02	**0.19±0.00**	0.44±0.00	0.53±0.00	-0.15±0.00	0.51±0.00	0.27±0.00	0.10±0.00	0.20±0.00
InfoNC-t-SNE	0.33±0.00	0.13±0.00	0.37±0.00	0.43±0.01	-0.11±0.00	0.46±0.00	0.26±0.00	0.10±0.00	0.21±0.00
LocalMAP	**0.58±0.00**	**0.19±0.00**	**0.60±0.00**	0.56±0.01	**-0.10±0.00**	**0.60±0.00**	**0.32±0.00**	**0.14±0.00**	0.22±0.00

Table 1: Silhouette scores for different Algorithms. **Bold** is best, underline is second best or statistically insignificant from the best. Each row is an algorithm, each column is a dataset.

	MNIST	FMNIST	USPS	COIL20	20NG	Kang	Seurat	Human Cortex	CBMC
PCA	00:01.77	00:01.64	00:00.13	00:02.77	00:00.11	00:00.07	00:00.07	00:00.22	00:00.20
t-SNE	03:02.38	02:36.97	00:35.92	00:10.36	01:10.69	00:40.33	01:05.32	01:05.68	01:49.60
UMAP	00:28.05	00:33.94	00:08.80	00:08.76	00:08.87	00:08.38	00:13.51	00:27.18	00:29.33
PaCMAP	00:52.39	00:34.54	00:04.26	00:03.28	00:08.29	00:05.59	00:13.16	00:18.72	00:39.59
LargeVis	14:51.71	14:02.86	06:45.63	07:09.97	06:44.39	07:12.39	08:15.04	08:31.17	11:18.19
TriMAP	00:53.64	00:48.16	00:06.59	00:01.29	00:13.05	00:13.26	00:19.31	00:26.75	01:03.43
PHATE	04:15.00	01:45.05	00:11.40	00:05.15	00:19.44	00:20.12	00:42.26	01:25.71	06:26.80
HNNE	00:10.05	00:06.41	00:00.60	00:02.35	00:02.19	00:01.33	00:04.21	00:03.02	00:04.27
Neg-t-SNE	01:02.65	01:10.92	00:12.62	00:03.86	00:23.06	00:18.32	00:28.25	00:42.77	00:56.06
NCVis	02:36.92	01:39.70	00:12.92	00:14.16	00:26.69	00:31.72	00:42.20	01:03.09	02:16.05
InfoNC-t-SNE	01:06.19	01:01.58	00:14.30	00:05.63	00:25.31	00:18.85	00:36.02	00:46.74	01:04.94
LocalMAP	01:47.35	00:47.51	00:06.23	00:06.77	00:13.28	00:07.63	00:20.17	00:29.02	01:12.35

Table 2: Running Time for Different Algorithms. Each row is an algorithm, each column is a dataset.

Experimental Results: Robustness and Sensitivity

DR results cannot be trusted if they change under random initialization. Hence, an important characteristic of DR algorithms is that they produce consistent results with different initial conditions. Figure 17 in Appendix I shows the result of LocalMAP over several runs, showing that it is capable of consistently producing correct clustering results, whereas other methods do not. In fact, *there are no runs of any other methods that distinctly separate the 10 clusters that LocalMAP finds every time.*

9 Discussion

While in the past, DR methods aimed to maintain the local and global structure inherent in high-dimensional data, these methods trusted its underlying graph structure, usually derived from an untrustworthy distance metric. LocalMAP does not do this, instead it dynamically (during run time) discovers what parts of the graph are untrustworthy, and what parts of the graph are not sampled well enough to be able to tell clusters apart. This is why its results are visually and quantitatively better than other DR methods – it essentially cleans up the data while it is running.

One direction for future work would be to combine the insights of LocalMAP with new parametric approaches to DR that have just begun to yield successful results (Huang, Wang, and Rudin 2024).

Our work could have substantial societal impact, for instance, if it is able to find clusters of patients that have different immune system properties (Semenova et al. 2024; Falcinelli et al. 2023). Our experiments indicate that LocalMAP has a higher chance of accomplishing this than past DR approaches.

Acknowledgments

We acknowledge funding from the National Science Foundation under grants DGE-2022040, CMMI-2323978, and IIS-

2130250 and the National Institutes of Health under grants R01-DA054994, NIH-241802, NIH-2022040.

References

Amezquita, R. A.; Lun, A. T.; Becht, E.; Carey, V. J.; Carpp, L. N.; Geistlinger, L.; Marini, F.; Rue-Albrecht, K.; Risso, D.; Soneson, C.; et al. 2020. Orchestrating Single-Cell Analysis with Bioconductor. *Nature Methods*, 17(2): 137–145.

Amid, E.; and Warmuth, M. K. 2019. TriMAP: Large-scale Dimensionality Reduction Using Triplets. *arXiv:1910.00204*.

Artemenkov, A.; and Panov, M. 2020. NCVis: Noise Contrastive Approach for Scalable Visualization. In *Proceedings of The Web Conference 2020*, 2941–2947. New York, NY, USA: Association for Computing Machinery.

Atitey, K.; Motsinger-Reif, A. A.; and Anchang, B. 2024. Model-based evaluation of spatiotemporal data reduction methods with unknown ground truth through optimal visualization and interpretability metrics. *Briefings in Bioinformatics*, 25(1): bbad455.

Becht, E.; McInnes, L.; Healy, J.; Dutertre, C.-A.; Kwok, I. W.; Ng, L. G.; Ginhoux, F.; and Newell, E. W. 2019. Dimensionality reduction for visualizing single-cell data using UMAP. *Nature Biotechnology*, 37(1): 38–44.

Belkin, M.; and Niyogi, P. 2001. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Advances in Neural Information Processing Systems*, volume 14, 585–591. MIT Press.

Belkina, A. C.; Ciccolella, C. O.; Anno, R.; Halpert, R.; Spidlen, J.; and Snyder-Cappione, J. E. 2019. Automated optimized parameters for T-distributed stochastic neighbor embedding improve visualization and analysis of large datasets. *Nature Communications*, 10(5415).

Böhm, J. N.; Berens, P.; and Kobak, D. 2020. A Unifying Perspective on Neighbor Embeddings along the Attraction-Repulsion Spectrum. *arXiv:2007.08902*.

Böhm, J. N.; Berens, P.; and Kobak, D. 2023. Unsupervised visualization of image datasets using contrastive learning. In *International Conference on Learning Representations*.

Cao, J.; Spielmann, M.; Qiu, X.; Huang, X.; Ibrahim, D. M.; Hill, A. J.; Zhang, F.; Mundlos, S.; Christiansen, L.; Steemers, F. J.; et al. 2019. The single-cell transcriptional landscape of mammalian organogenesis. *Nature*, 566(7745): 496–502.

Cao, Y.; and Wang, L. 2017. Automatic selection of t-SNE Perplexity. *arXiv:1708.03229*.

Chen, P.; Liao, B.; Chen, G.; and Zhang, S. 2019. Understanding and Utilizing Deep Neural Networks Trained with Noisy Labels. In *International Conference on Machine Learning*.

Damrich, S.; Böhm, J. N.; Hamprecht, F. A.; and Kobak, D. 2023. From t-SNE to UMAP with contrastive learning. In *International Conference on Learning Representations*.

Dries, R.; Zhu, Q.; Dong, R.; Eng, C.-H. L.; Li, H.; Liu, K.; Fu, Y.; Zhao, T.; Sarkar, A.; Bao, F.; et al. 2021. Giotto: a toolbox for integrative analysis and visualization of spatial expression data. *Genome Biology*, 22: 1–31.

Falcinelli, S. D.; Cooper-Volkheimer, A. D.; Semenova, L.; Wu, E.; Richardson, A.; Ashokkumar, M.; Margolis, D. M.; Archin, N. M.; Rudin, C. D.; Murdoch, D.; and Browne, E. P. 2023. Impact of Cannabis Use on Immune Cell Populations and the Viral Reservoir in People With HIV on Suppressive Antiretroviral Therapy. *The Journal of Infectious Diseases*, jiad364.

Han, B.; Yao, Q.; Yu, X.; Niu, G.; Xu, M.; Hu, W.; Tsang, I. W.; and Sugiyama, M. 2018. Co-teaching: robust training of deep neural networks with extremely noisy labels. In *Neural Information Processing Systems*, 8536–8546.

Huang, H.; Wang, Y.; and Rudin, C. 2024. Navigating the Effect of Parametrization for Dimensionality Reduction. In *Neural Information Processing Systems*.

Huang, H.; Wang, Y.; Rudin, C.; and Browne, E. P. 2022. Towards a comprehensive evaluation of dimension reduction methods for transcriptomic data visualization. *Communications Biology*, 5(1): 719.

Hull, J. J. 1994. A database for handwritten text recognition research. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(5): 550–554.

Johnson, W. E.; Li, C.; and Rabinovic, A. 2007. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, 8(1): 118–127.

Kang, H. M.; Subramaniam, M.; Targ, S.; Nguyen, M.; Maliskova, L.; McCarthy, E.; Wan, E.; Wong, S.; Byrnes, L.; Lanata, C. M.; et al. 2018. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nature Biotechnology*, 36(1): 89.

Kobak, D.; and Linderman, G. C. 2021. Initialization is critical for preserving global data structure in both t-SNE and UMAP. *Nature Biotechnology*, 39(2): 156–157.

Lang, K. 1995. Newsweeder: Learning to filter netnews. In *International Conference on Machine Learning*, 331–339.

LeCun, Y.; Cortes, C.; and Burges, C. 2010. MNIST handwritten digit database. *ATT Labs [Online]. Available: http://yann. lecun. com/exdb/mnist*, 2.

Lee, D. D.; and Seung, H. S. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755): 788–791.

Li, J.; Socher, R.; and Hoi, S. C. 2020. DivideMix: Learning with Noisy Labels as Semi-supervised Learning. In *International Conference on Learning Representations*.

Martin, R. D.; Yohai, V. J.; Maronna, R. A.; and Salibián-Barrera, M. 2018. *Robust Statistics: Theory and Methods (with R), 2nd Edition*. Wiley.

McInnes, L.; Healy, J.; and Melville, J. 2018. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv:1802.03426*.

Moon, K. R.; van Dijk, D.; Wang, Z.; Chen, W.; Hirn, M. J.; Coifman, R. R.; Ivanova, N. B.; Wolf, G.; and Krishnaswamy, S. 2017. PHATE: a dimensionality reduction method for visualizing trajectory structures in high-dimensional biological data. *BioRxiv*, 120378.

Mu, J.; Bhat, S.; and Viswanath, P. 2017. All-but-the-Top: Simple and Effective Postprocessing for Word Representations. *arXiv:1702.01417*.

Nene, S. A.; Nayar, S. K.; and Murase, H. 1996. Columbia Object Image Library (COIL-20). Technical report, Technical Report CUCS-005-96.

Nguyen, L. H.; and Holmes, S. 2019. Ten quick tips for effective dimensionality reduction. *PLoS Computational Biology*, 15(6).

Northcutt, C. G.; Athalye, A.; and Mueller, J. 2024. Pervasive label errors in test sets destabilize machine learning benchmarks. *Neural Information Processing Systems*.

Pearson, K. 1901. On Lines and Planes of Closest Fit to Systems of Points in Space. *Philosophical Magazine*, 2(11): 559–572.

Poličar, P. G.; Stražar, M.; and Zupan, B. 2023. Embedding to reference t-SNE space addresses batch effects in single-cell classification. *Machine Learning*, 112(2): 721–740.

Raunak, V.; Gupta, V.; and Metze, F. 2019. Effective Dimensionality Reduction for Word Embeddings. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, 235–243.

Rousseeuw, P. J. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20: 53–65.

Roweis, S. T.; and Saul, L. K. 2000. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science*, 290(5500): 2323–2326.

Sarfraz, S.; Koulakis, M.; Seibold, C.; and Stiefelhagen, R. 2022. Hierarchical nearest neighbor graph embedding for efficient dimensionality reduction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 336–345.

Semenova, L.; Wang, Y.; Falcinelli, S.; Archin, N.; Cooper-Volkheimer, A. D.; Margolis, D. M.; Goonetilleke, N.; Murdoch, D. M.; Rudin, C. D.; and Browne, E. P. 2024. Machine learning approaches identify immunologic signatures of total and intact HIV DNA during long-term antiretroviral therapy. *eLife*, 13: RP94899.

Stoeckius, M.; Hafemeister, C.; Stephenson, W.; Houck-Loomis, B.; Chattopadhyay, P. K.; Swerdlow, H.; Satija, R.; and Smibert, P. 2017. Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods*, 14(9): 865–868.

Stuart, T.; Butler, A.; Hoffman, P.; Hafemeister, C.; Papalexi, E.; Mauck III, W. M.; Hao, Y.; Stoeckius, M.; Smibert, P.; and Satija, R. 2019. Comprehensive integration of single-cell data. *Cell*, 177(7): 1888–1902.

Tang, J.; Liu, J.; Zhang, M.; and Mei, Q. 2016. Visualizing large-scale and high-dimensional data. In *Proceedings of the 25th International Conference on the World Wide Web*, 287–297.

Tenenbaum, J. B.; de Silva, V.; and Langford, J. C. 2000. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 290(5500): 2319–2323.

Torgerson, W. 1952. Multidimensional scaling: I Theory and Method. *Psychometrika*, 17(4): 401–419.

Van Assel, H.; Espinasse, T.; Chiquet, J.; and Picard, F. 2022. A probabilistic graph coupling view of dimension reduction. *Advances in Neural Information Processing Systems*, 35: 10696–10708.

van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9: 2579–2605.

Wang, Y.; Huang, H.; Rudin, C.; and Shaposhnik, Y. 2021. Understanding How Dimension Reduction Tools Work: An Empirical Approach to Deciphering t-SNE, UMAP, TriMAP, and PaCMAP for Data Visualization. *Journal of Machine Learning Research*, 22.

Wattenberg, M.; Viégas, F.; and Johnson, I. 2016. How to use t-SNE effectively. *Distill*, 1(10): e2.

Wolf, F. A.; Angerer, P.; and Theis, F. J. 2018. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology*, 19: 1–5.

Xiao, H.; Rasul, K.; and Vollgraf, R. 2017. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. *arXiv:cs.LG/1708.07747*.

Zhu, K.; Bendl, J.; Rahman, S.; Vicari, J. M.; Coleman, C.; Clarence, T.; Latouche, O.; Tsankova, N. M.; Li, A.; Brennand, K. J.; et al. 2023. Multi-omic profiling of the developing human cerebral cortex at the single-cell level. *Science Advances*, 9(41): eadg3754.

Zu, X.; and Tao, Q. 2022. SpaceMAP: Visualizing High-Dimensional Data by Space Expansion. In *International Conference on Machine Learning*, 27707–27723.