Understanding Dimension Reduction Algorithms

Yingfan Wang, Haiyang Huang

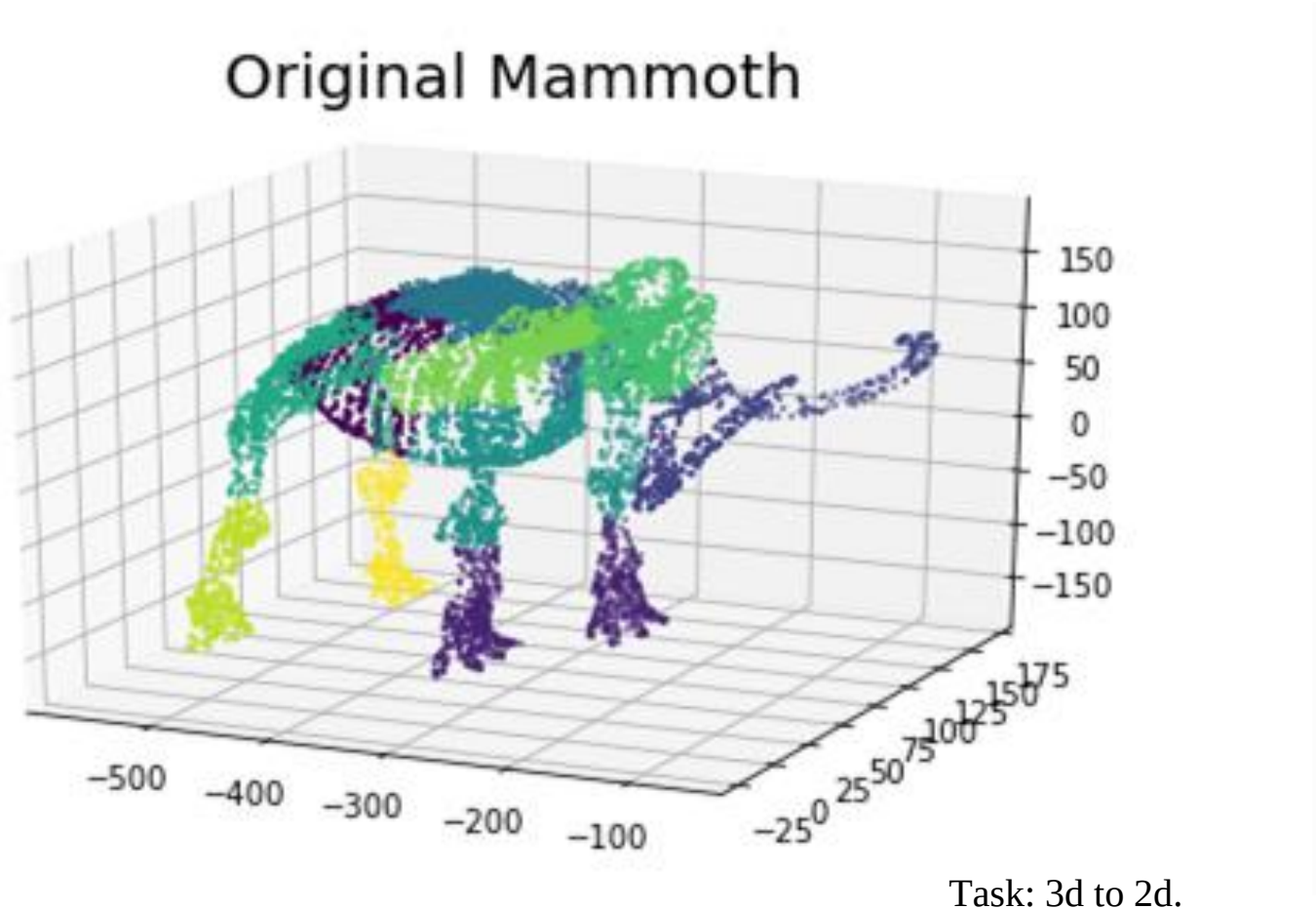
Duke DEPARTMENT of COMPUTER SCIENCE

Dimension reduction (DR) algorithms

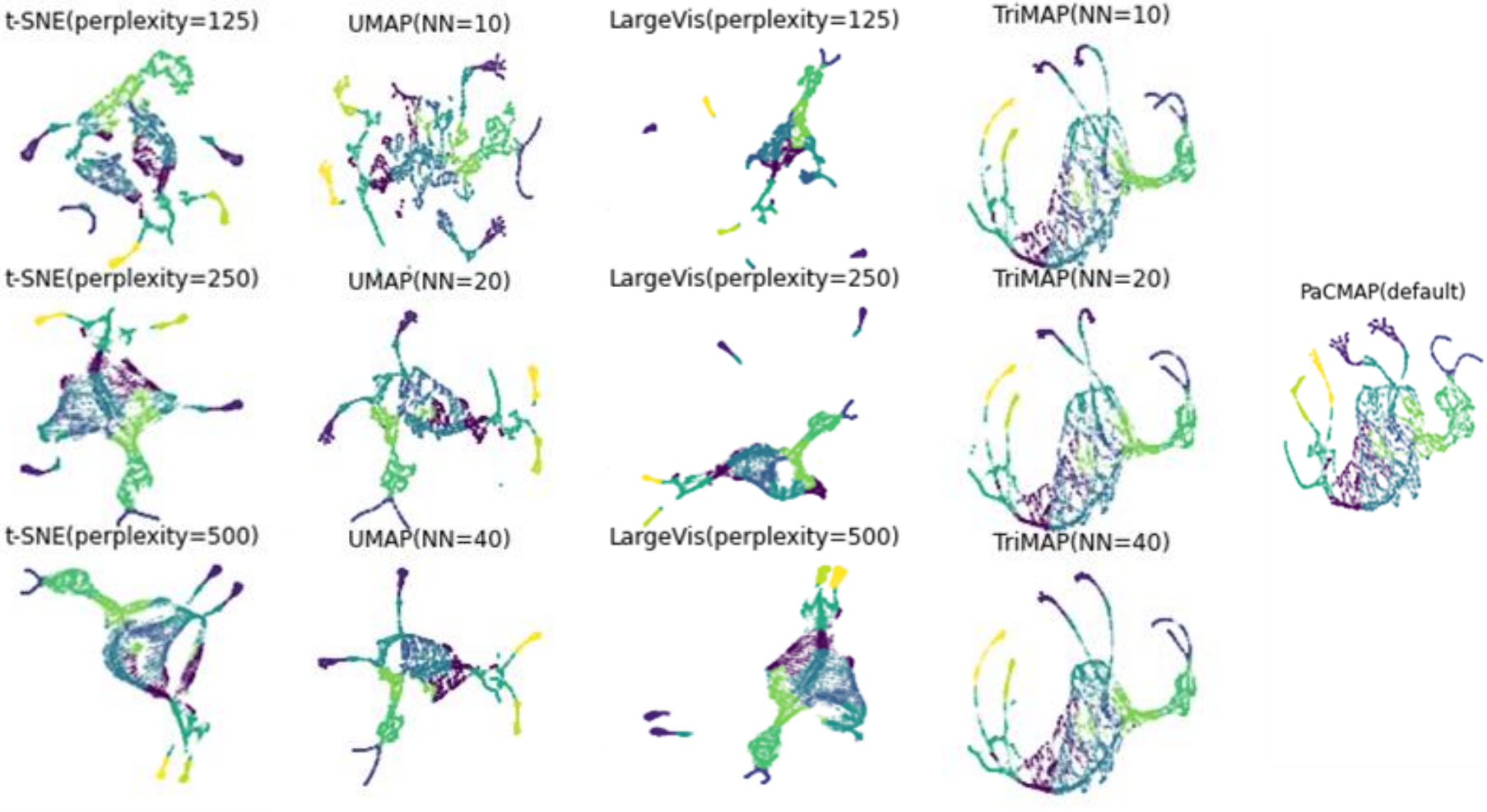
Input: high-dimensional data
Output: low-dimensional data that preserves…

- the graph structure?
- local neighborhoods?
- global structure?

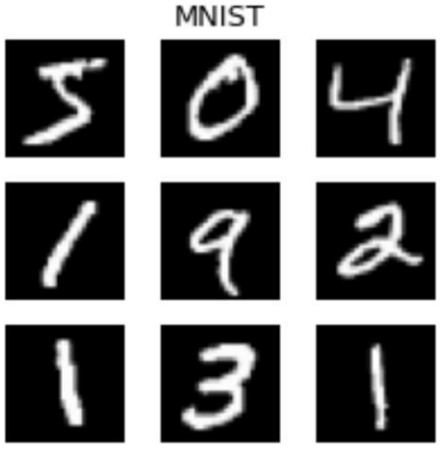
Previous successful DR algorithms: t-SNE, UMAP, Largevis, TriMAP, …
Our new algorithm: PaCMAP

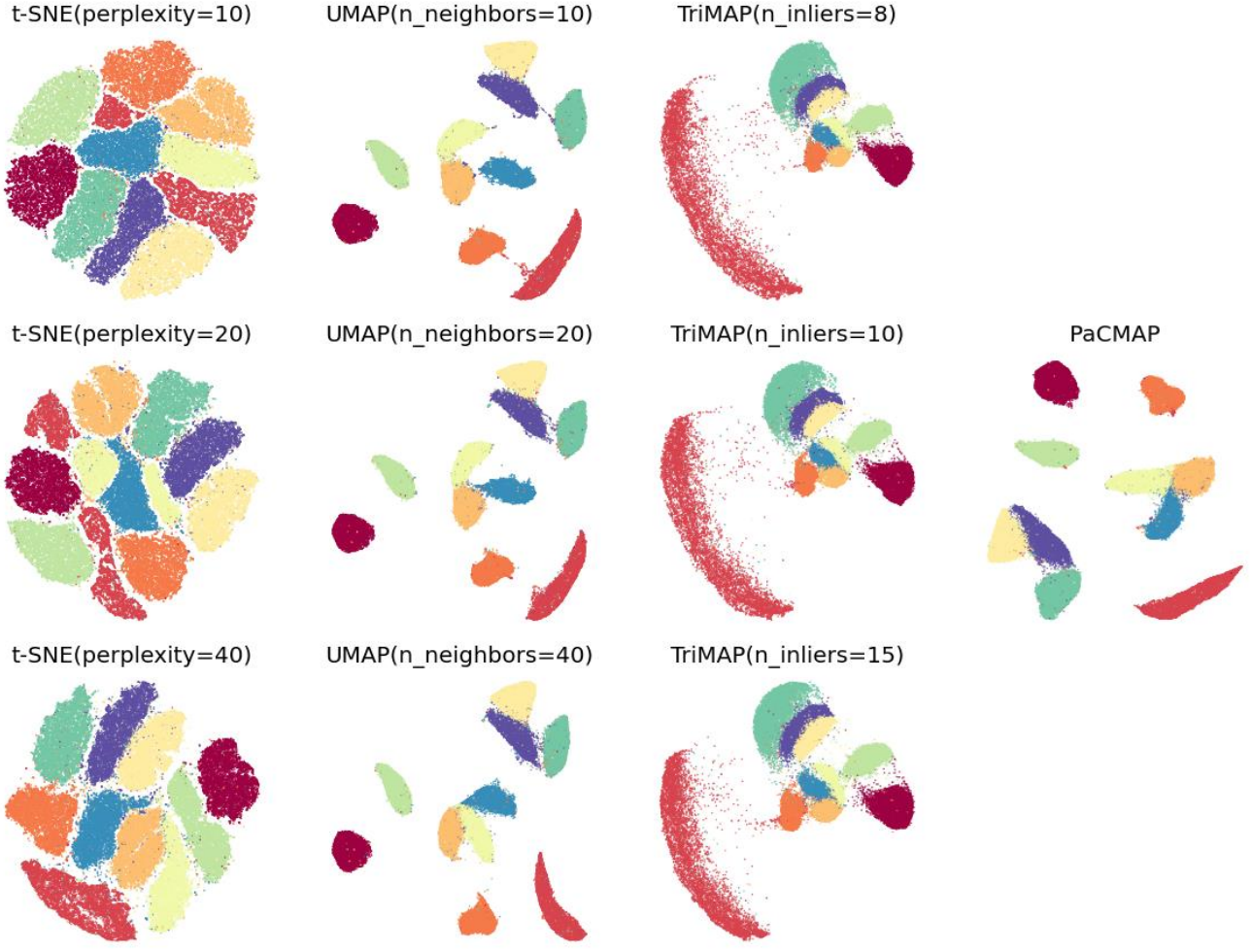


Task: 3d to 2d.
Global structure is important here!


t-SNE(perplexity=125)
UMAP(NN=10)
LargeVis(perplexity=125)
TriMAP(NN=10)
t-SNE(perplexity=250)
UMAP(NN=20)
LargeVis(perplexity=250)
TriMAP(NN=20)
PaCMAP(default)
t-SNE(perplexity=500)
UMAP(NN=40)
LargeVis(perplexity=500)
TriMAP(NN=40)

MNIST dataset (handwritten digit image)





Algorithm	Graph component	Loss function
t-SNE	Edges (i, j)	$\text{Loss}_{i,j}^{\text{t-SNE}} = p_{ij} \log \frac{p_{ij}}{q_{ij}}$, where $q_{ij} = \frac{\left(1+\|\mathbf{y}_i - \mathbf{y}_j\|^2\right)^{-1}}{\sum_{k \neq l}(1+\|\mathbf{y}_k - y_l\|^2)^{-1}}$
UMAP	Edges (i, j)	$\text{Loss}_{i,j}^{\text{UMAP}} = \begin{cases} \bar{w}_{i,j} \log\left(1 + a\left(\|\mathbf{y}_i - \mathbf{y}_j\|_2^2\right)^b\right)^{-1} & i, j \text{ neighbors} \\ (1 - \bar{w}_{i,j}) \log\left(1 - \left(1 + a\left(\|\mathbf{y}_i - \mathbf{y}_j\|_2^2\right)^b\right)^{-1}\right) & \text{Otherwise} \end{cases}$
TriMAP	Triplets (i, j, k) where $\text{Distance}_{i,j} \leq \text{Distance}_{i,k}$	$\text{Loss}_{i,j,k}^{\text{TM}} = \omega_{i,j,k} \frac{s(\mathbf{y}_i, \mathbf{y}_k)}{s(\mathbf{y}_i, \mathbf{y}_j) + s(\mathbf{y}_i, \mathbf{y}_k)}$, where $s(\mathbf{y}_i, \mathbf{y}_j) = \left(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2\right)^{-1}$

t-SNE (van der Maaten and Hinton, 2008), UMAP (McInnes et al., 2018), TriMAP (Amid & Warmuth, 2019)

What elements of these algorithms are important?

What we knew before:

Certain properties of the loss function are important:

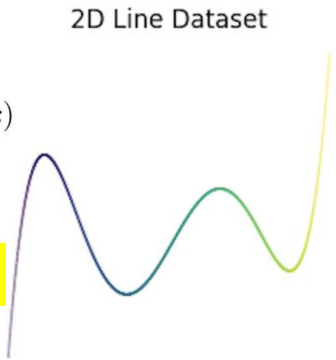
- Attraction: neighbors should be attracted. But not too close! (Crowding)
- Repulsion: farther points in original space should be far in low-dim space.

【Local Structure!】

But that's not enough

$$\sum_{(i,j)\in\mathcal{T}_{\text{neighbors}}} l^{\text{attract}}(i,j) + \sum_{(i,k)\in\mathcal{T}_{\text{further}}} l^{\text{repulse}}(i,k)$$

【Global Structure!】



After a huge amount of experimentation, we found that:

For local structure:

- Certain specific properties of the loss function are important.

For global structure:

- We must have forces on non-neighbors.
- The choice of which graph components to preserve is important.

Algorithm	Graph component	Loss function
t-SNE	Edges (i, j)	$\text{Loss}_{i,j}^{\text{t-SNE}} = p_{ij} \log \frac{p_{ij}}{q_{ij}}$, where $q_{ij} = \frac{\left(1+\|\mathbf{y}_i - \mathbf{y}_j\|^2\right)^{-1}}{\sum_{k \neq l}(1+\|\mathbf{y}_k - y_l\|^2)^{-1}}$
UMAP	Edges (i, j)	$\text{Loss}_{i,j}^{\text{UMAP}} = \begin{cases} \bar{w}_{i,j} \log\left(1 + a\left(\|\mathbf{y}_i - \mathbf{y}_j\|_2^2\right)^b\right)^{-1} & i, j \text{ neighbors} \\ (1 - \bar{w}_{i,j}) \log\left(1 - \left(1 + a\left(\|\mathbf{y}_i - \mathbf{y}_j\|_2^2\right)^b\right)^{-1}\right) & \text{Otherwise} \end{cases}$
TriMAP	Triplets (i, j, k) where $\text{Distance}_{i,j} \leq \text{Distance}_{i,k}$	$\text{Loss}_{i,j,k}^{\text{TM}} = \omega_{i,j,k} \frac{s(\mathbf{y}_i, \mathbf{y}_k)}{s(\mathbf{y}_i, \mathbf{y}_j) + s(\mathbf{y}_i, \mathbf{y}_k)}$, where $s(\mathbf{y}_i, \mathbf{y}_j) = \left(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2\right)^{-1}$

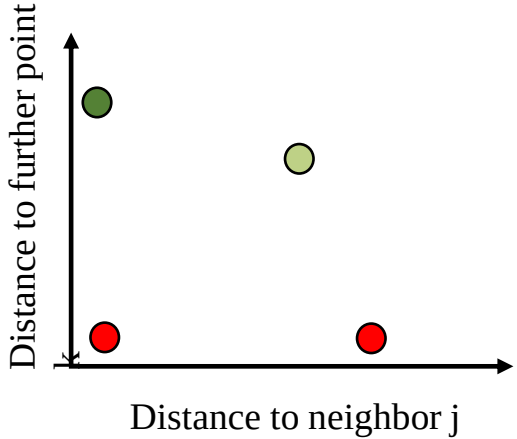
t-SNE (van der Maaten and Hinton, 2008), UMAP (McInnes et al., 2018), TriMAP (Amid & Warmuth, 2019)

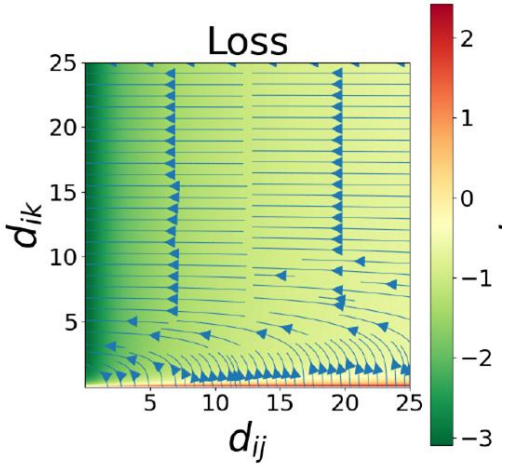
For local structure:

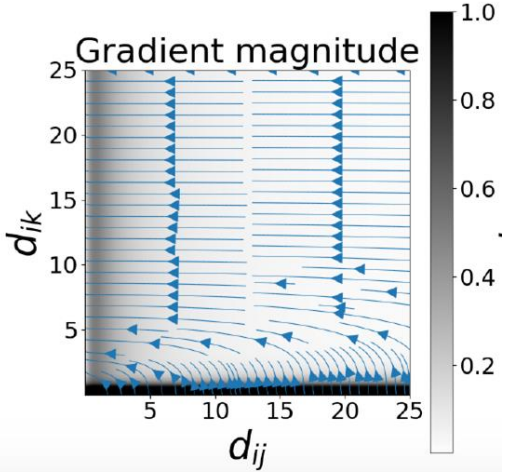
Certain specific properties of the loss function are important.

The “rainbow” plot

Triplet i, j (neighbor), k (further)



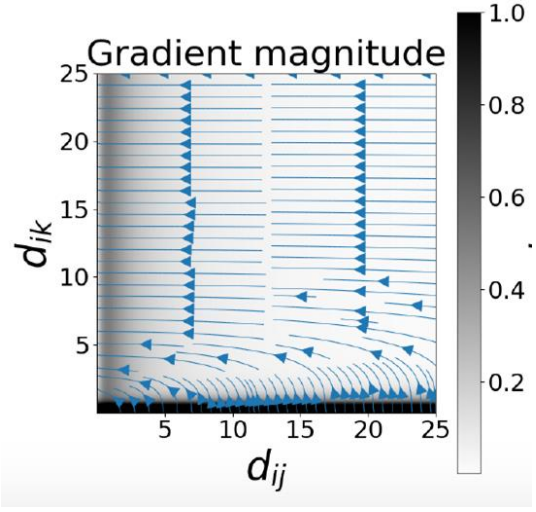


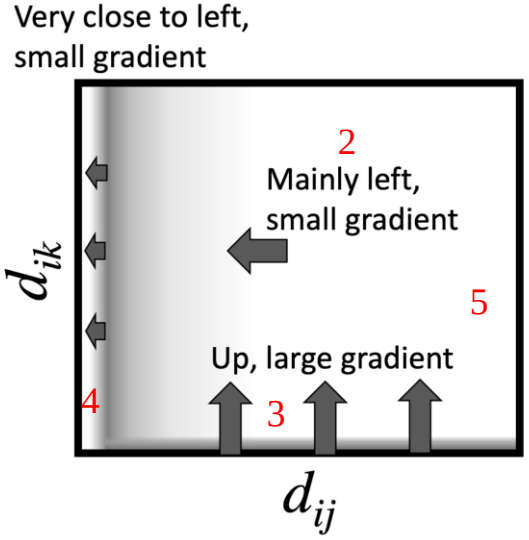


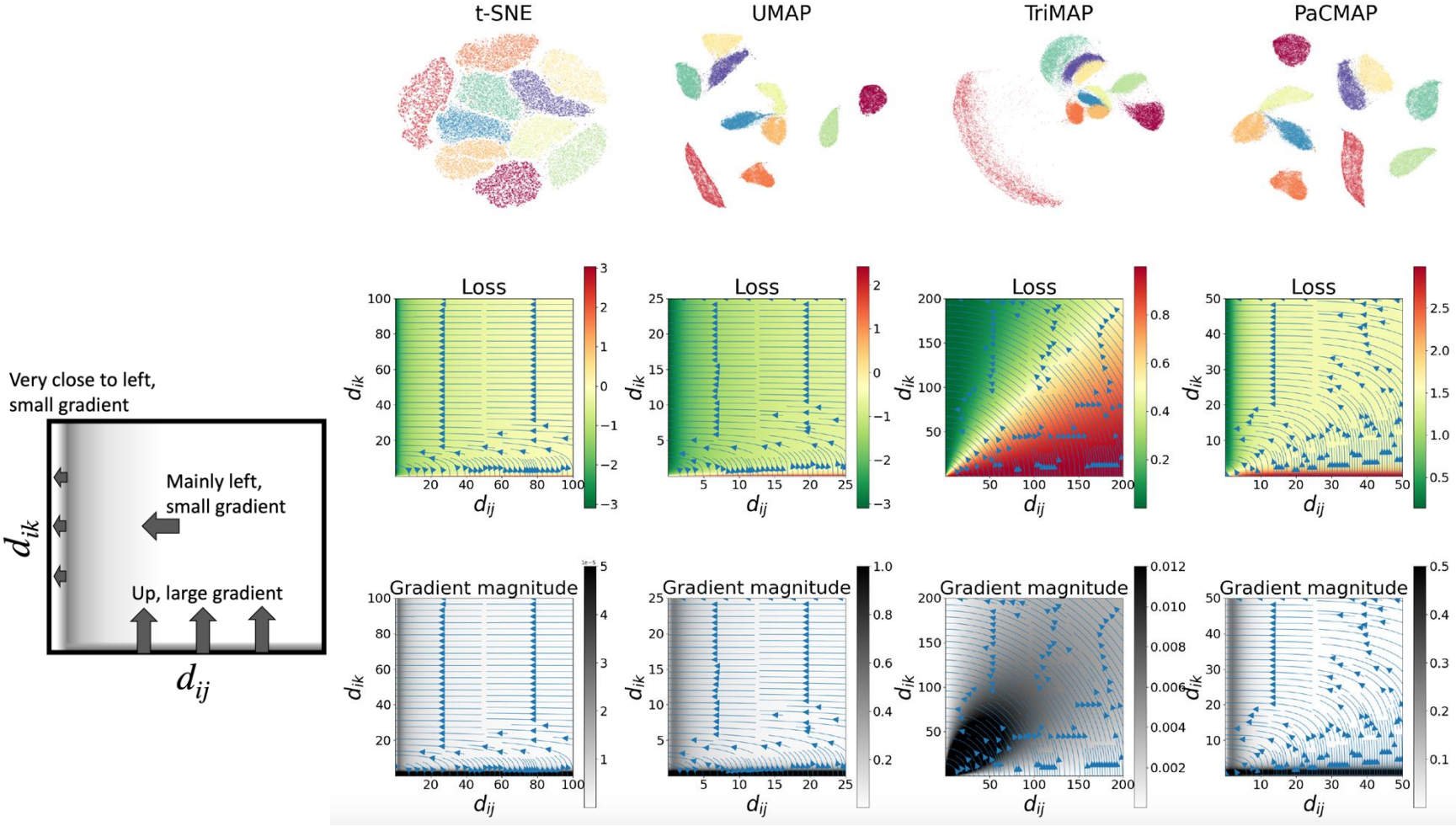

t-SNE
UMAP
TriMAP
PaCMAP

Principles for a good loss for DR

1) Monotonicity: pull neighbors closer, push farther points away (go left, go up)
2) Except along the bottom, gradient should go mainly to the left (broadly attract neighbors, further points are far enough), sufficient attraction
3) Along bottom, gradient goes mainly up (further point is too close) with large gradient
4) Along vertical axis, small magnitude (neighbor is close enough)
5) Weak pull on far neighbors: gradients should become small as distance to neighbor j becomes large





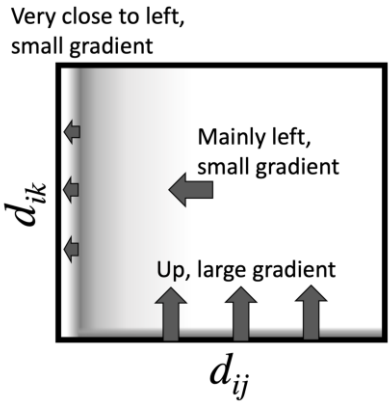

t-SNE
UMAP
TriMAP
PaCMAP
Loss
Gradient magnitude
d_{ik}
d_{ij}
Very close to left,
small gradient
Mainly left,
small gradient
Up, large gradient

$Loss = log(1 + exp(\frac{d_{ij}^2 - d_{ik}^2}{10}))$

$Loss = \frac{d_{ij}^2 + 1}{d_{ik}^2 + 1}$

$Loss = -\frac{d_{ik}^2 + 1}{d_{ij}^2 + 1}$

$Loss = log(1 + exp(d_{ij}^2) + exp(-d_{ik}^2))$



Too much repulsion

Insufficient attraction

No gradient on repulsion

Insufficient local attraction

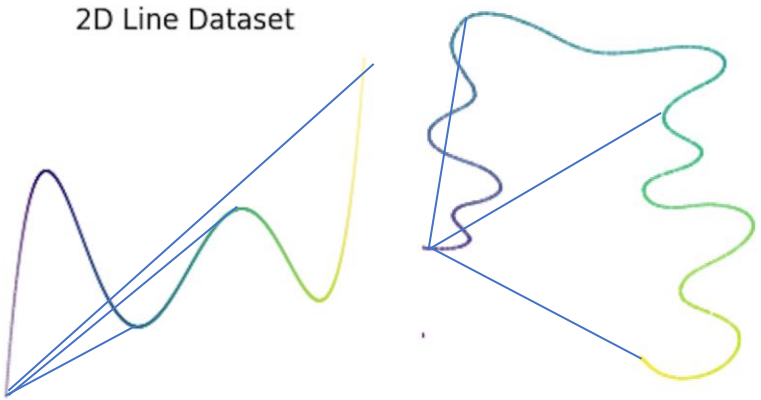
After a huge amount of experimentation, we found that:

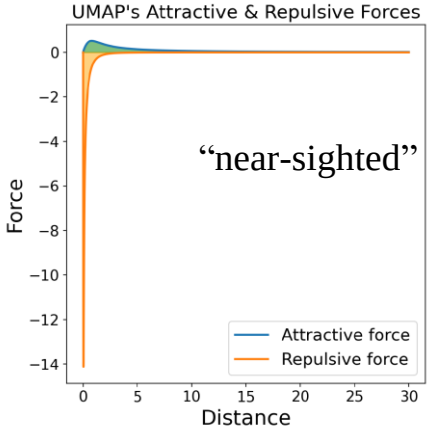
For local structure:

- Certain specific properties of the loss function are important.

For global structure:

- We must have forces on non-neighbors.
- The choice of which graph components to preserve is important.





"near-sighted"

$$\sum_{(i,j)\in\mathcal{T}_{\text{neighbors}}} l^{\text{attract}}(i,j) + \sum_{(i,k)\in\mathcal{T}_{\text{further}}} l^{\text{repulse}}(i,k)$$

For global structure:

- We must have forces on non-neighbors.
- The choice of which graph components to preserve is important.

MN: mid-near FP: further point

$$\text{Loss}^{\text{PaCMAP}} = w_{\text{neighbors}}\text{Loss}_{\text{neighbors}} + w_{MN}\text{Loss}_{MN} + w_{FP}\text{Loss}_{FP}$$

$$\text{Loss}_{\text{neighbors}} = \frac{\tilde{d}_{ij}}{10 + \tilde{d}_{ij}}, \quad \text{Loss}_{MN} = \frac{\tilde{d}_{ik}}{10000 + \tilde{d}_{ik}}, \quad \text{Loss}_{FP} = \frac{1}{1 + \tilde{d}_{il}}$$

Neighbors: attractive

Mid-near pairs: mild attractive

Further points: repulsive

For global structure:

- We must have forces on non-neighbors.
- The choice of which graph components to preserve is important.

$$\text{Loss}^{\text{PaCMAP}} = w_{\text{neighbors}}\text{Loss}_{\text{neighbors}} + w_{MN}\text{Loss}_{MN} + w_{FP}\text{Loss}_{FP}$$

$$\text{Loss}_{\text{neighbors}} = \frac{\tilde{d}_{ij}}{10 + \tilde{d}_{ij}}, \quad \text{Loss}_{MN} = \frac{\tilde{d}_{ik}}{10000 + \tilde{d}_{ik}}, \quad \text{Loss}_{FP} = \frac{1}{1 + \tilde{d}_{il}}$$

Neighbors: attractive

Mid-near pairs: mild attractive

Further points: repulsive

The weights change on a schedule:

Period 1: $w_{\text{neighbors}}$ is medium, w_{MN} is huge, w_{FP} is medium
Period 2: $w_{\text{neighbors}}$ is large, w_{MN} is small, w_{FP} is medium
Period 3: $w_{\text{neighbors}}$ is medium, w_{MN} is 0, w_{FP} is medium

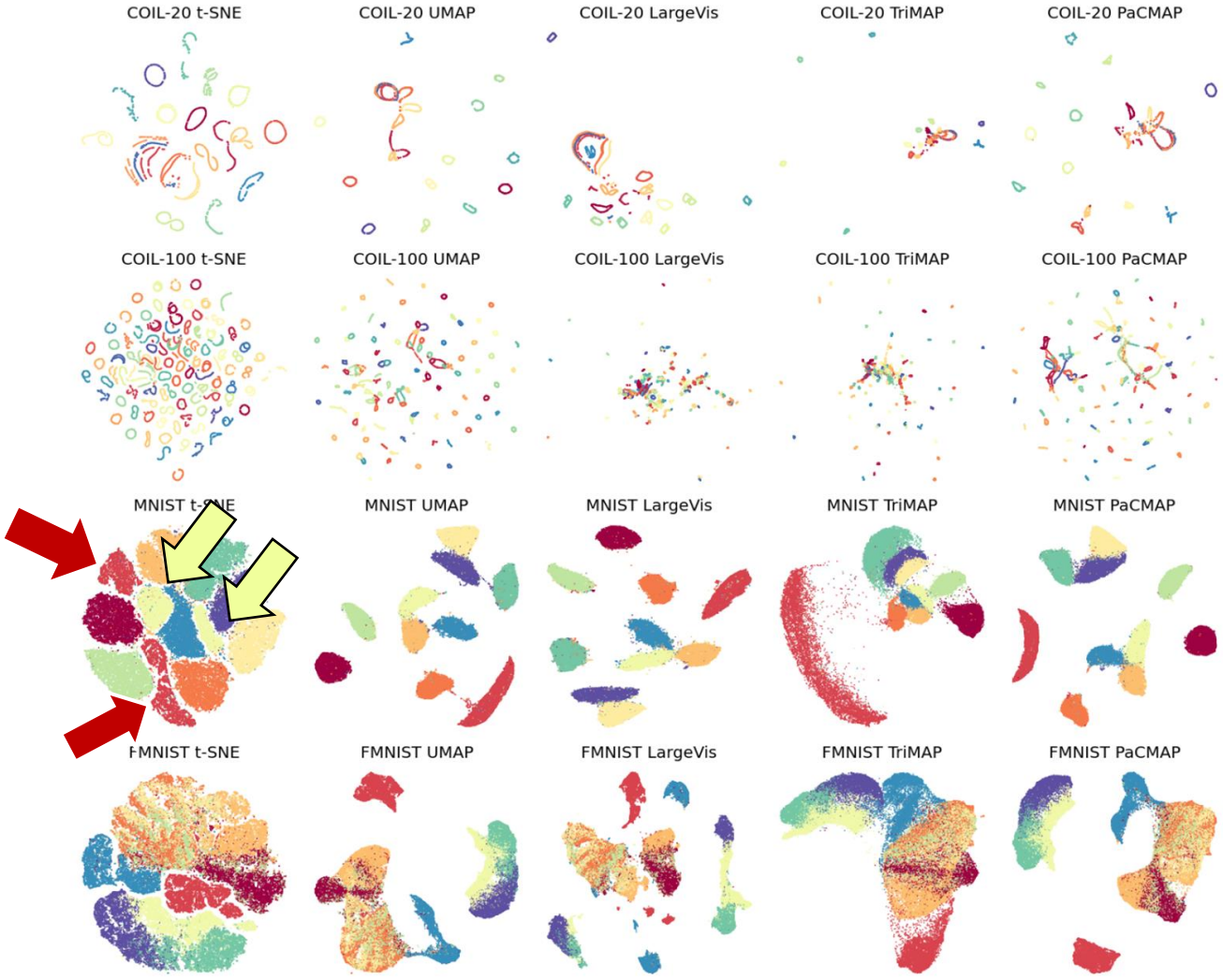
SVM accuracy (measures local structure preservation)

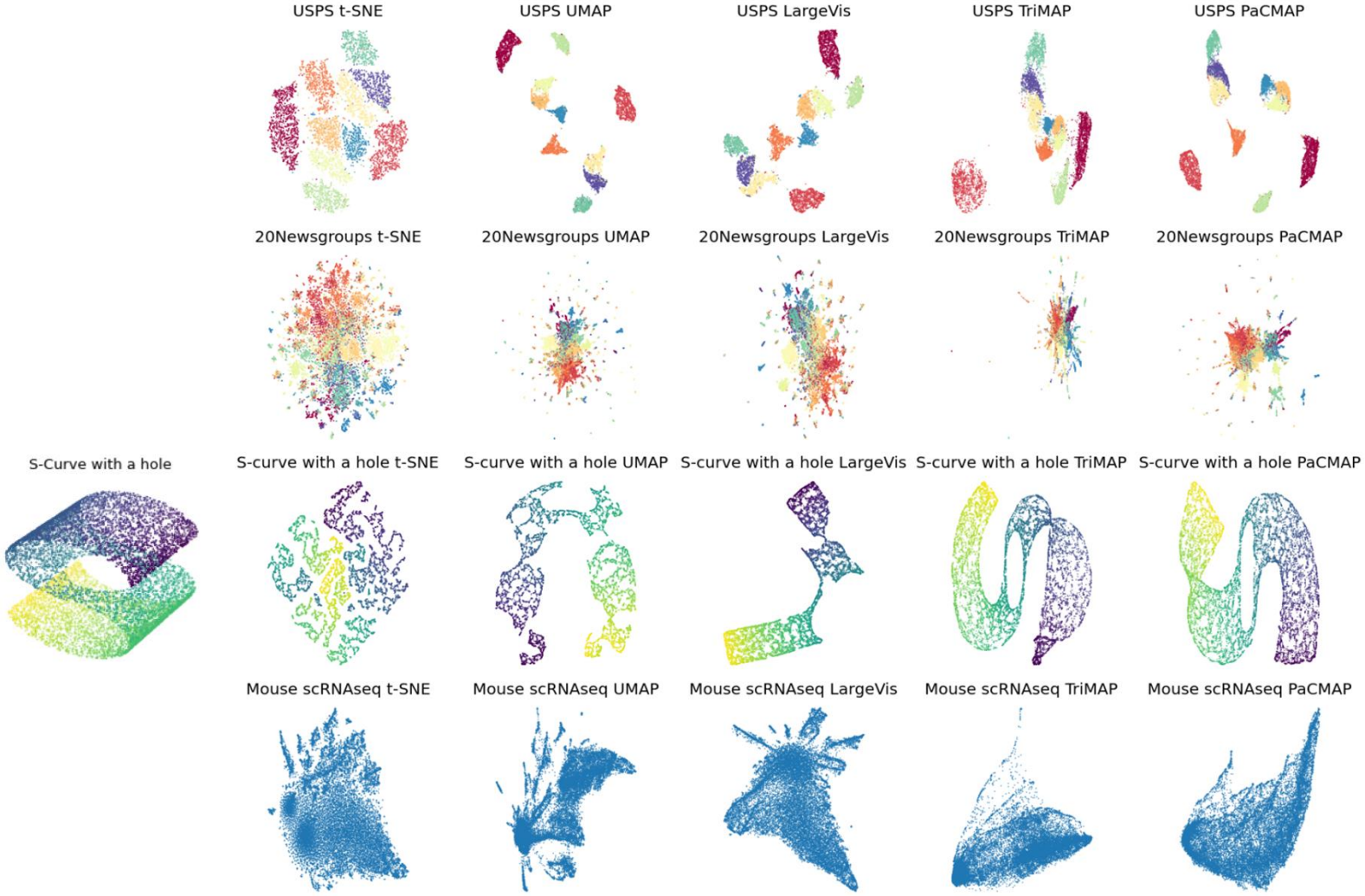
Dataset (Size)	Baseline	t-SNE	LargeVis	UMAP	TriMAP	PaCMAP
COIL-20 (1.4K)	0.972	0.909 ± 0.015	0.799 ± 0.020	0.844 ± 0.004	0.778 ± 0.010	**0.942 ± 0.009**
COIL-100 (7.2K)	0.989	0.911 ± 0.004	0.707 ± 0.014	0.879 ± 0.007	0.737 ± 0.019	**0.933 ± 0.009**
USPS (9K)	0.949	***0.959 ± 0.002***	*0.957 ± 0.001*	*0.956 ± 0.002*	0.946 ± 0.001	*0.958 ± 0.001*
Mammoth (10K)	0.961	0.927 ± 0.009	0.923 ± 0.011	***0.941 ± 0.003***	0.900 ± 0.004	*0.933 ± 0.004*
20Newsgroups (18K)	0.792	0.435 ± 0.014	*0.444 ± 0.012*	0.431 ± 0.013	0.410 ± 0.007	***0.447 ± 0.006***
MNIST (70K)	0.926	0.967 ± 0.002	0.965 ± 0.004	0.970 ± 0.001	0.960 ± 0.001	**0.974 ± 0.001**
F-MNIST (70K)	0.854	***0.754 ± 0.003***	*0.748 ± 0.003*	0.742 ± 0.003	0.729 ± 0.001	*0.752 ± 0.004*

Random triplet accuracy (measures global structure preservation)

Dataset (Size)	t-SNE	LargeVis	UMAP	TriMAP	PaCMAP
COIL-20 (1.4K)	0.698 ± 0.016	**0.735 ± 0.011**	0.649 ± 0.014	0.659 ± 0.006	0.699 ± 0.007
COIL-100 (7.2K)	0.577 ± 0.012	0.630 ± 0.021	0.568 ± 0.011	0.633 ± 0.002	**0.718 ± 0.005**
USPS (9K)	0.654 ± 0.013	*0.668 ± 0.011*	***0.669 ± 0.002***	0.640 ± 0.002	*0.665 ± 0.002*
S-Curve with Hole (9.5K)	0.722 ± 0.045	0.834 ± 0.041	0.800 ± 0.013	0.838 ± 0.004	**0.866 ± 0.010**
Mammoth (10K)	0.701 ± 0.038	0.766 ± 0.024	0.816 ± 0.001	***0.874 ± 0.001***	*0.872 ± 0.003*
20Newsgroups (18K)	0.645 ± 0.002	0.632 ± 0.001	0.664 ± 0.002	**0.704 ± 0.002**	0.666 ± 0.003
Mouse scRNA-seq (20K)	0.715 ± 0.002	0.719 ± 0.003	*0.727 ± 0.002*	***0.728 ± 0.001***	*0.727 ± 0.001*
MNIST (70K)	0.600 ± 0.007	0.601 ± 0.007	0.614 ± 0.001	0.600 ± 0.001	**0.619 ± 0.001**
F-MNIST (70K)	0.679 ± 0.019	0.657 ± 0.011	0.740 ±0.001	**0.777 ± 0.001**	0.741 ± 0.002
Flow Cytometry (3M)	–	–	–	0.857 ± 0.001	**0.894 ± 0.005**
KDD Cup99 (4M)	(Ran out of memory or time, >24 hrs)		–	0.660 ± 0.007	**0.752 ± 0.002**

t-SNE version: opt-SNE (Belkina et al., 2019) built on Multi-core t-SNE (Ulyanov et al., 2016)


COIL-20 t-SNE
COIL-20 UMAP
COIL-20 LargeVis
COIL-20 TriMAP
COIL-20 PaCMAP
COIL-100 t-SNE
COIL-100 UMAP
COIL-100 LargeVis
COIL-100 TriMAP
COIL-100 PaCMAP
MNIST t-SNE
MNIST UMAP
MNIST LargeVis
MNIST TriMAP
MNIST PaCMAP
FMNIST t-SNE
FMNIST UMAP
FMNIST LargeVis
FMNIST TriMAP
FMNIST PaCMAP


USPS t-SNE
USPS UMAP
USPS LargeVis
USPS TriMAP
USPS PaCMAP
20Newsgroups t-SNE
20Newsgroups UMAP
20Newsgroups LargeVis
20Newsgroups TriMAP
20Newsgroups PaCMAP
S-Curve with a hole
S-curve with a hole t-SNE
S-curve with a hole UMAP
S-curve with a hole LargeVis
S-curve with a hole TriMAP
S-curve with a hole PaCMAP
Mouse scRNAseq t-SNE
Mouse scRNAseq UMAP
Mouse scRNAseq LargeVis
Mouse scRNAseq TriMAP
Mouse scRNAseq PaCMAP

Thanks!