Cell Reports, Volume *31*

Supplemental Information

A Quantitative Framework for Evaluating Single-Cell Data Structure Preservation by Dimensionality Reduction Techniques

Cody N. Heiser and Ken S. Lau

Supplemental Figures

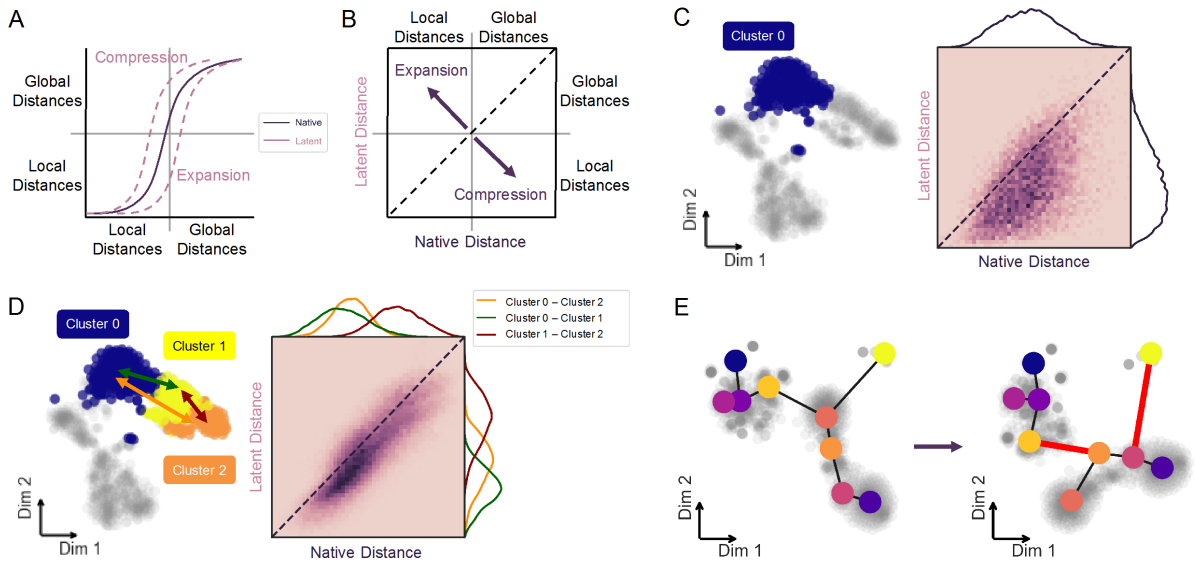

Figure S1, Related to Figure 1. Interpretation of data structure preservation analyses. A) Small distances in cumulative distance distribution represent local cell similarity (within cluster), while large distances represent global relationships and arrangement of data (between clusters). A distribution shift left indicates compression of distances from native to latent space, while a shift right results from expansion or exaggeration of native distances. B) Correlation of latent to native distances; dispersion below identity line (dashed) indicates compression of distances from native to latent space, while dispersion above identity results from expansion of native distances in low-dimensional space. C) Substructure analysis uses same framework as Figure 1 on isolated subset of data to measure intra-cluster distance preservation and determine contribution to global structure. D) Distribution of distances from all cells in one cluster to another define relative substructure. Inter-cluster distances are measured pairwise to interrogate cluster arrangement in latent compared to native space. E) Evaluation of coarse global cluster topology using minimum spanning tree (MST) graph constructed from cluster centroids and their pairwise distances in native and latent dimensions. Black edges between centroids denote MST. Edges not present in native MST graph are highlighted red, indicating relative rearrangement of clusters following dimension reduction.

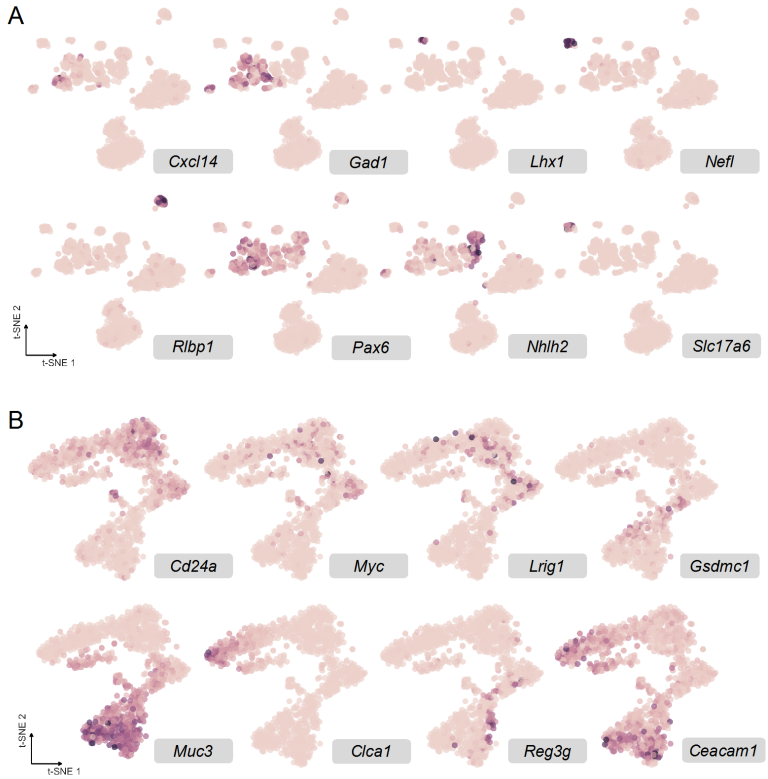

Figure S2, Related to Figure 2. t-SNE visualizations from Figure 2 with overlay of arcsinh-normalized expression of marker genes (Macosko *et al.*, 2015; Herring, Banerjee, *et al.*, 2018) used to assign cell type to Louvain clusters for retina (A) and colon (B) datasets.

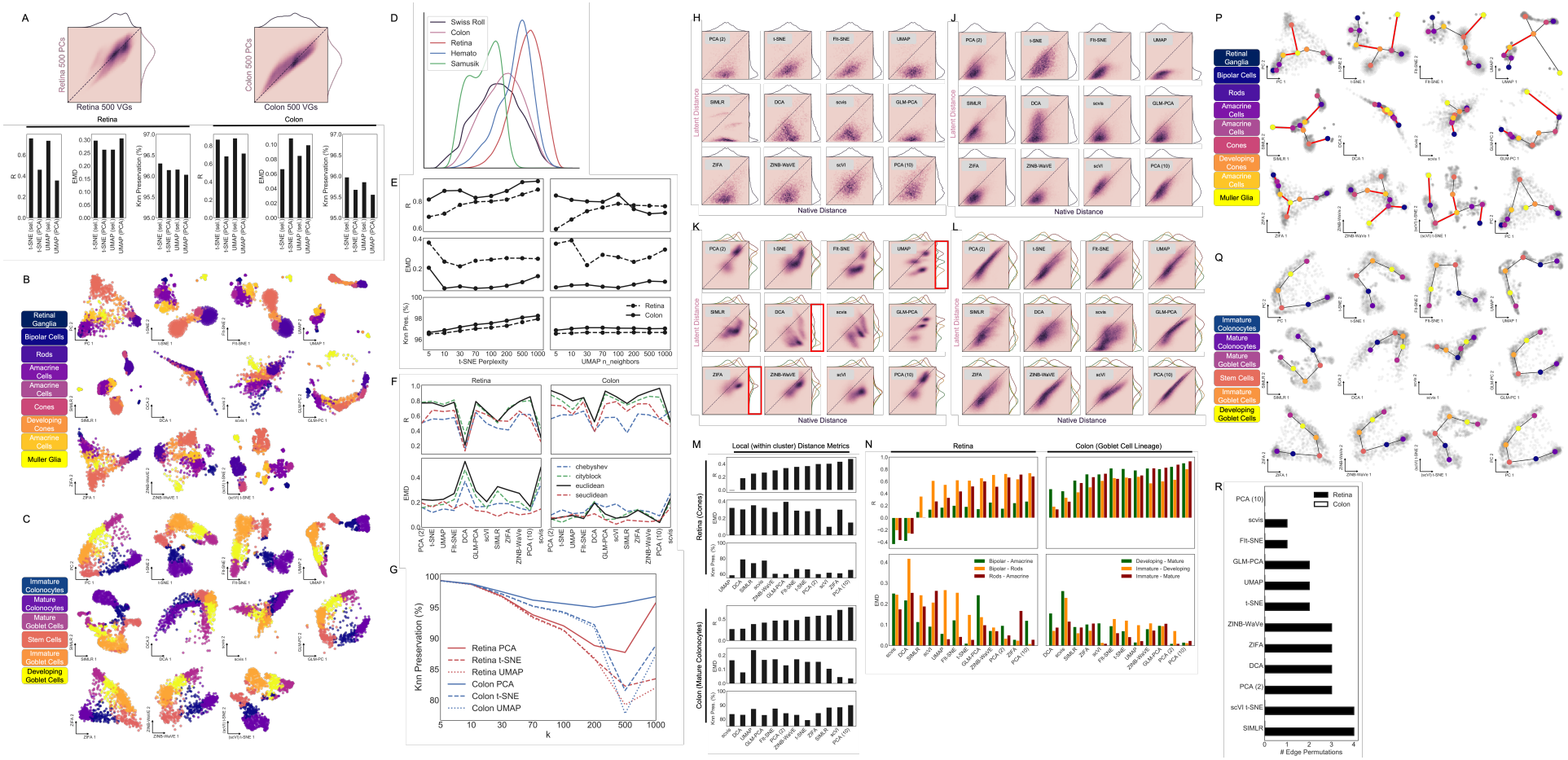

Figure S3, Related to Figure 3. A) Comparison of 500 variable genes (VGs) to 500 principal components (PCs) as latent space for both datasets. 2D histograms of VGs and PCs for retina and colon data (top). Summary of correlation, EMD, and Knn preservation values for t-SNE and UMAP primed with 500 selected VGs ("sel.") and 500 PCs ("PCA") for both datasets. B,C) Low-dimensional projections from 11 evaluated dimensionality reduction methods with overlay of consensus Louvain clusters for retina (B) and colon (C) data. These embeddings were generated using the 500 most variable genes in each dataset. D) Overlay of normalized probability distributions of native spaces to demonstrate varying structure of different datasets. Retina, Colon, and Hemato are scRNA-seq data (normalized counts of all genes shown), Samusik is mass cytometry (CyTOF), and Swiss Roll is a synthetic dataset of 1000 points in 3-dimensional space. E) Resulting distance metrics from titration of perplexity parameter in t-SNE and UMAP (n_neighbors) on retina (discrete) and colon (continuous) datasets. F) Framework is agnostic to chosen distance metric. Comparison of alternative distance metrics and their effect on R and EMD results for both datasets and all 11 methods. Trend is generally conserved relative to default metric (Euclidean) which exhibits the most variability, indicating utility for discriminating method performance. G) Titration of k parameter for construction of Knn graphs and calculation of their preservation following dimension reduction by PCA, t-SNE, and UMAP on both retina and colon data. Optimal window for reliable discrimination of method performance between 3 and 10 % of dataset size (k ≈ 30-100). H) 2D histograms of cell distance correlations within cone cell cluster of retina dataset for evaluated latent spaces. J) Same as in H for distances within mature colonocyte cluster of colon dataset. K) Same as in H for distances between bipolar, amacrine, and rod cells in retina dataset. Methods that rearranged cluster ordering are highlighted in red. L) Same as in H for distances between three goblet cell clusters in colon dataset. M) Local (within cluster) distance preservation metrics for cone cell cluster and mature colonocyte cluster from retina and colon datasets, respectively. N) EMD and distance correlation values for pairwise distance distributions between bipolar cells, rod cells, and amacrine cells in retina dataset, and three clusters along developing goblet cell lineage in colon dataset. P) Minimum spanning tree (MST) graphs constructed from cluster centroids and their pairwise distances in 2D latent

projections for retina dataset. Red edges represent those not present in corresponding native MST graph. Q) Same as in P for colon dataset. R) Summary of number of edge permutations in 2D latent projections relative to native graph (500 VGs). Retina dataset contains 8 total edges; colon dataset contains 5 total edges.

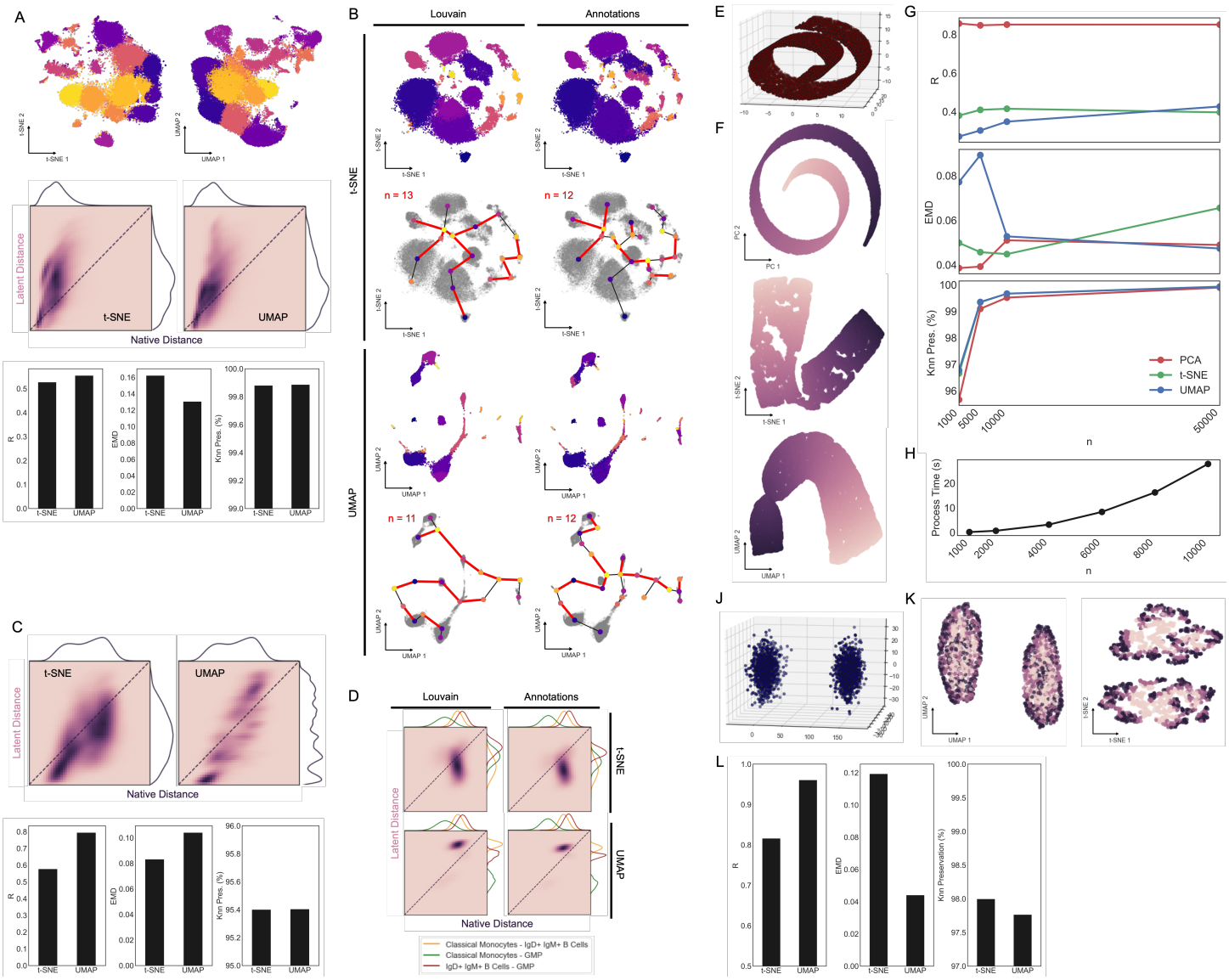

Figure S4, Related to Figure 4. Application of framework to datasets from Becht, *et al.* (2018) and 3D synthetic datasets with intuitive structure. A) Hematopoietic subset of mouse cell atlas (Han *et al.*, 2018); 14 combined scRNA-seq samples from blood and bone marrow totaling 51,252 cells. Data were normalized and preprocessed with 100-component PCA prior to embedding with t-SNE and UMAP (top) as in Han *et al.*, 2018. 2D histograms of cell distance correlation (middle) and summary metrics (bottom). B) Comparison of naïve Louvain clustering and previously published cell annotations via MST topological analysis for Samusik_01 CyTOF data (86,864 cells from mouse bone marrow, 39 features) preprocessed as in Weber and Robinson, 2016. C) 2D histograms of cell distance correlation (top) and summary metrics (bottom) for t-SNE and UMAP embeddings of dataset (both shown in A). D) Comparison of cluster definition as in B, using neighborhood analysis measuring distance distributions between classical monocytes, IgD+ IgM+ B cells, and granulocyte-monocyte progenitors (GMPs). E) Example 3-dimensional swiss roll dataset with 10,000 randomly placed points generated using sklearn.datasets.make_swiss_roll function with 0 noise (vertically away from manifold). F) Example 2D embeddings of data from A using PCA, t-SNE, and UMAP. Points are colored by their position along the manifold to show expected order. G) Correlation, EMD, and Knn preservation metrics (k=30) for swiss roll datasets as in E with increasing number of points. H) Processing time for structural preservation framework (calculating R, EMD, and Knn pres. from distance matrices) for up to 10,000 cells. J) 3D dataset consisting of two Gaussian point clouds (1,000 points each) generated using sklearn.datasets.make_gaussian_quantiles. K) UMAP and t-SNE embeddings of data from J with points colored by their distance from center of their respective Gaussian distribution in 3D space. L) Summary of structural preservation metrics for t-SNE and UMAP of data from J.

Supplemental Tables

Table S1, related to Figure 3. Summary of structural preservation metrics for scRNA-seq data (Figure 3C,K)

	EMD	R	Knn Pres. (%)
Colon			
DCA	0.19544272	0.52133666	96.2199
scvis	0.23221256	0.52698273	95.8854
SIMLR	0.08492918	0.76620863	96.8255
UMAP	0.09363649	0.80019129	97.1171
FIt-SNE	0.06910941	0.84379107	97.148
scVI	0.07250057	0.85703659	96.9871
ZIFA	0.13640594	0.8659447	96.9169
t-SNE	0.08217963	0.88159348	97.1733
ZINB-WaVE	0.08770175	0.92184172	96.7756
GLM-PCA	0.10577689	0.93166378	96.9085
PCA (2)	0.06352942	0.93813773	97.529
PCA (10)	0.03748585	0.96876379	98.0704
Retina			
DCA	0.53030121	0.21048289	96.7021
scvis	0.48463706	0.30398435	96.7131
SIMLR	0.32752819	0.5260237	97.2329
scVI	0.20517063	0.65531778	96.7922
ZIFA	0.2990902	0.68641666	96.5919
UMAP	0.2247211	0.726851	96.6672
PCA (2)	0.22306147	0.77175522	96.7996
t-SNE	0.21654539	0.77393381	96.8218
FIt-SNE	0.27079505	0.78204988	96.8063
ZINB-WaVE	0.27304757	0.79158321	96.6629
GLM-PCA	0.29257587	0.7917963	96.6734
PCA (10)	0.16593524	0.85567081	97.4033

Table S2, related to Figure 4. Summary of structural preservation metrics for simulated data (Figure 4E,K)

	R (Path1-Path2)	R (Path1-Path3)	R (Path2-Path3)	EMD (Path1-Path2)	EMD (Path1-Path3)	EMD (Path2-Path3)
Discrete						
PCA (2)	0.83587583	0.83458192	0.87561784	0.07000233	0.07363921	0.05680564
PCA (10)	0.8030938	0.78904974	0.84910766	0.07586439	0.07702044	0.06146719
t-SNE	-0.1246497	-0.6380908	0.5834859	0.0484876	0.05694764	0.05750208
UMAP	-0.0090184	-0.101758	0.78641275	0.02244971	0.08570284	0.04058559
GLM-PCA	0.80811142	0.75466097	0.81235556	0.07123744	0.07944286	0.08929313
ZINB-WaVE	0.81127499	0.82874357	0.84637534	0.07365733	0.06444041	0.07207731
SIMLR	0.56223239	0.64603738	0.51685226	0.26825249	0.15399134	0.26637604
ZIFA	0.8633688	0.85363458	0.88584082	0.07118298	0.05331366	0.05894853
FIt-SNE	-0.2676533	-0.6328717	0.3503715	0.04275841	0.0556827	0.06112829
DCA	0.16313048	0.15166638	0.09753656	0.13771101	0.20879459	0.13726973
scvis	0.67044307	0.33770459	0.53412756	0.09478033	0.12405217	0.12756443
scVI	0.48149456	0.63819369	0.53262934	0.11928148	0.10449441	0.12406571
Continuous						
PCA (2)	0.93115732	0.93619875	0.94489202	0.06017981	0.07216537	0.04598444
PCA (10)	0.9111296	0.90153843	0.91608509	0.05942224	0.06513866	0.04781115
t-SNE	0.93252508	0.93038337	0.9328201	0.02359744	0.03990939	0.03261977
UMAP	0.95994481	0.95711972	0.95476653	0.02715512	0.03340084	0.02362446
GLM-PCA	0.8722635	0.85968453	0.92948811	0.07441019	0.10451455	0.02761849
ZINB-WaVE	0.92712485	0.92940153	0.93726769	0.04650262	0.06886455	0.03625723
SIMLR	0.83989932	0.78777779	0.83101821	0.04706959	0.03091046	0.03213203
ZIFA	0.94028448	0.94373401	0.95305705	0.06354619	0.05511923	0.05025672
FIt-SNE	0.93951316	0.92713408	0.93855807	0.02646137	0.03889648	0.03341029
DCA	0.77718543	0.74124497	0.78978341	0.09388641	0.05607755	0.07878317
scvis	0.58769625	0.44315141	0.85311606	0.25200176	0.30550852	0.07758574