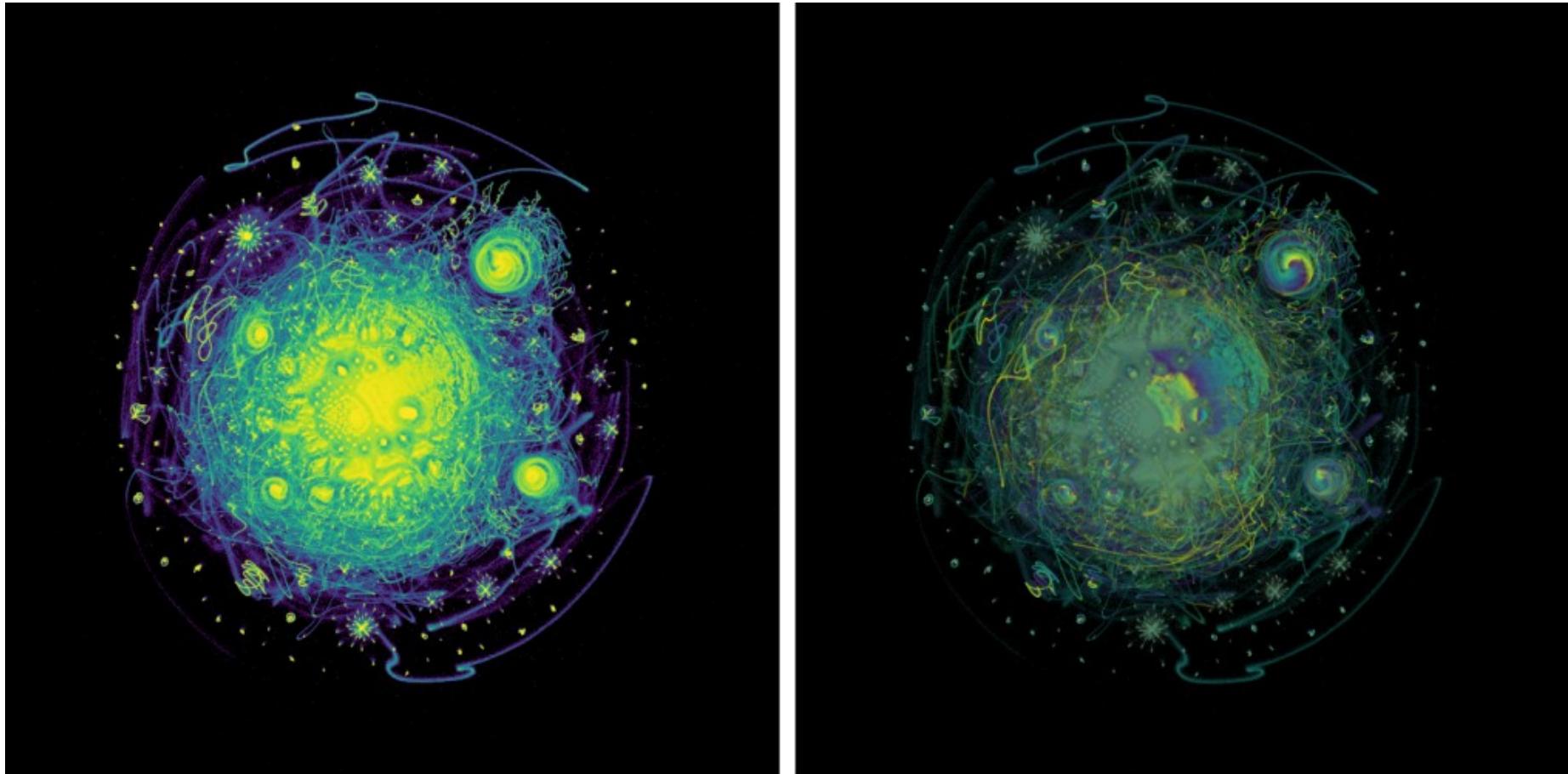


UMAP for OMICs Integration

OMICs Integration and Systems Biology course

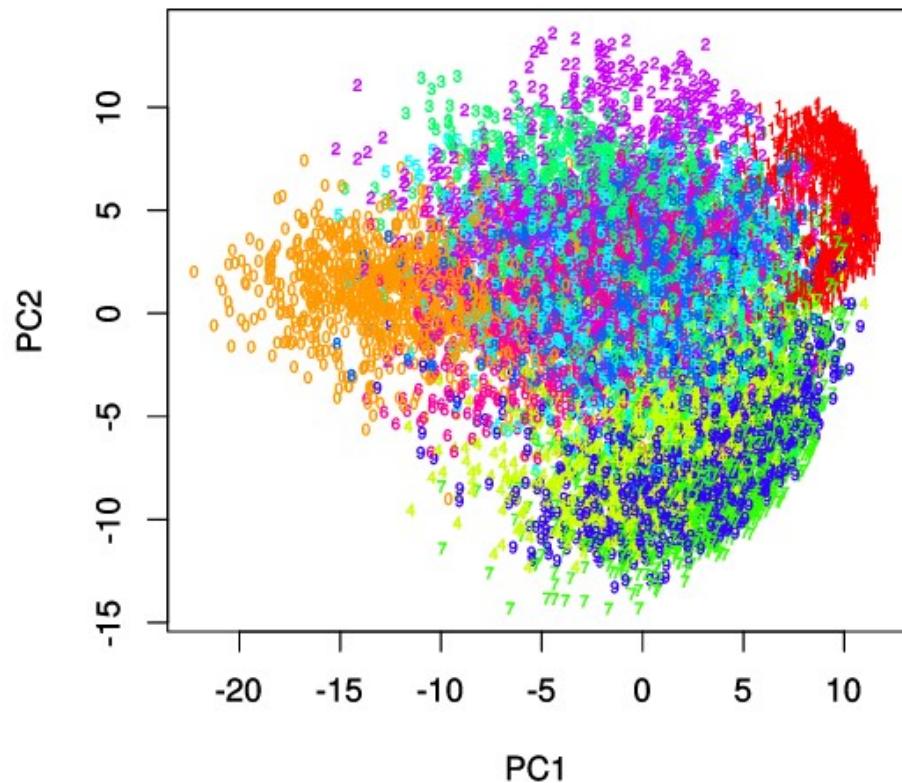
Nikolay Oskolkov, NBIS SciLifeLab

Lund, 7.09.2021

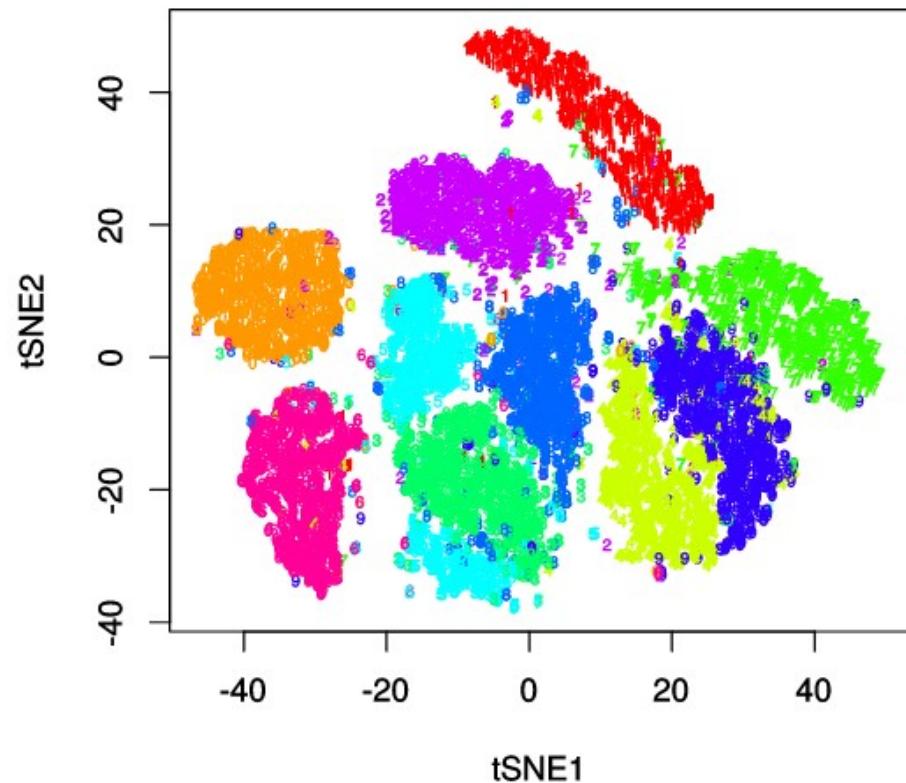


0 0 0 0 0 0 0 0 0 0 0 0 0 0
1 1 1 1 1 1 1 1 1 1 1 1 1 1
2 2 2 2 2 2 2 2 2 2 2 2 2 2
3 3 3 3 3 3 3 3 3 3 3 3 3 3
4 4 4 4 4 4 4 4 4 4 4 4 4 4
5 5 5 5 5 5 5 5 5 5 5 5 5 5
6 6 6 6 6 6 6 6 6 6 6 6 6 6
7 7 7 7 7 7 7 7 7 7 7 7 7 7
8 8 8 8 8 8 8 8 8 8 8 8 8 8
9 9 9 9 9 9 9 9 9 9 9 9 9 9

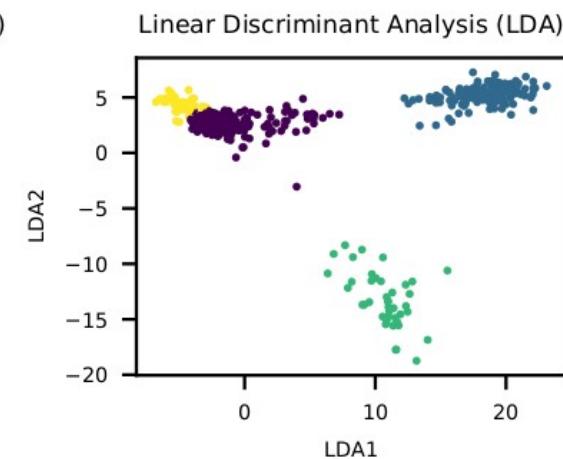
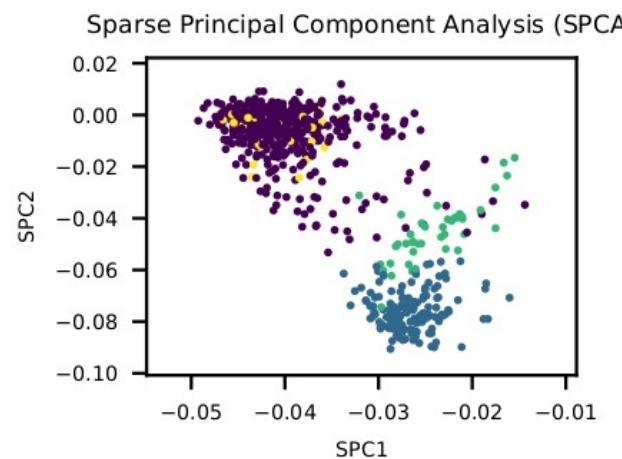
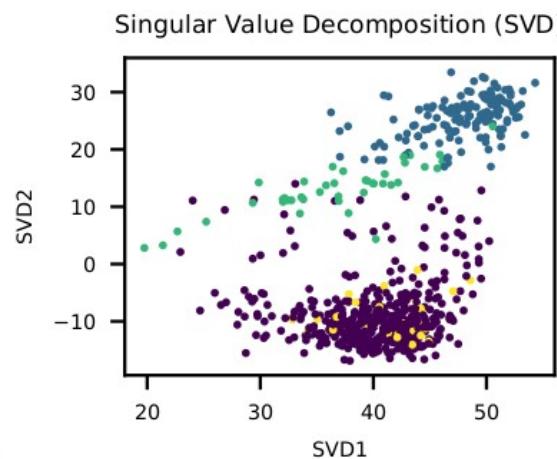
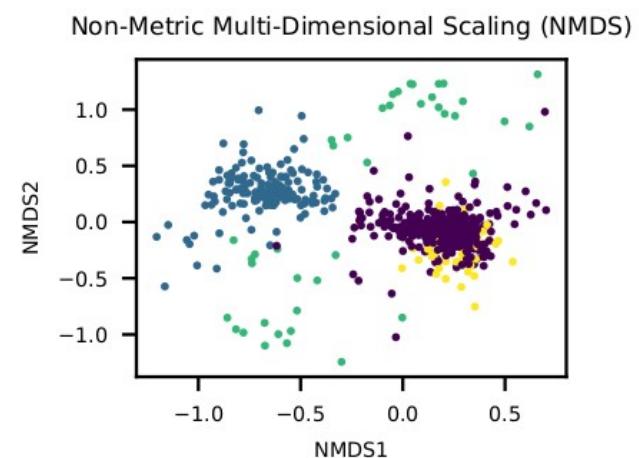
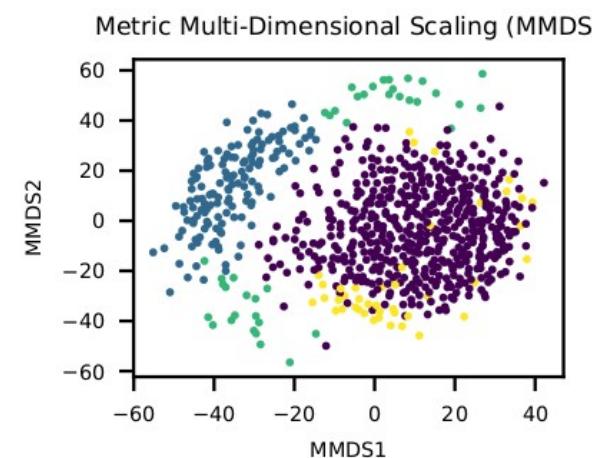
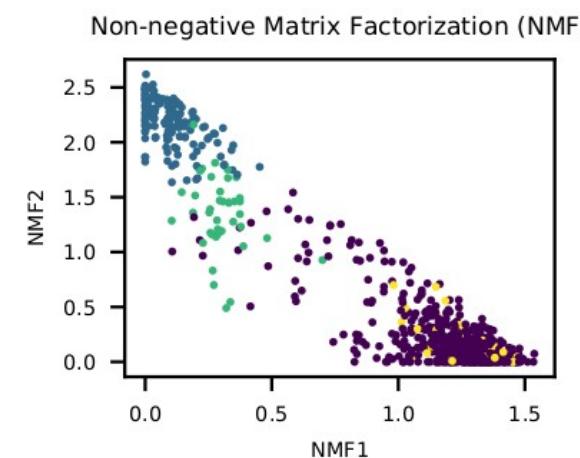
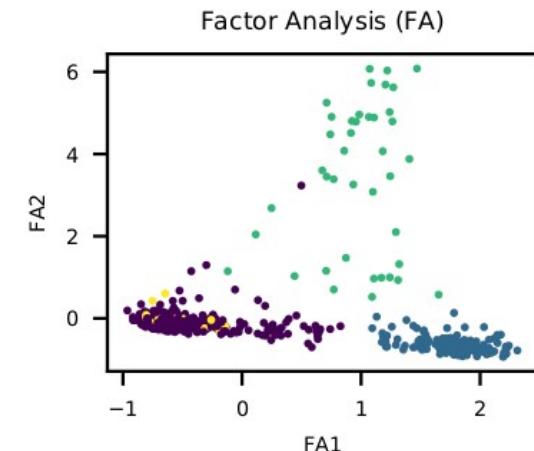
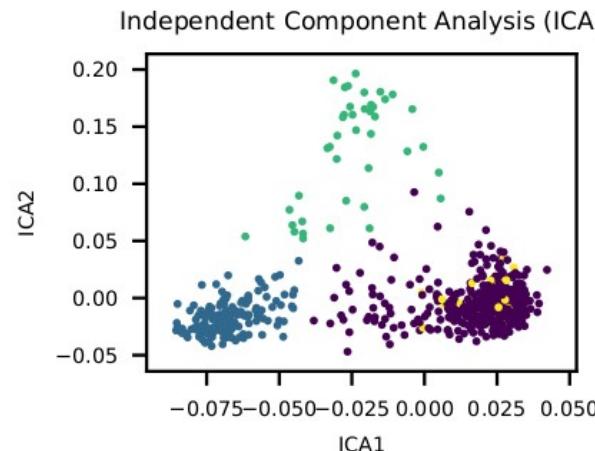
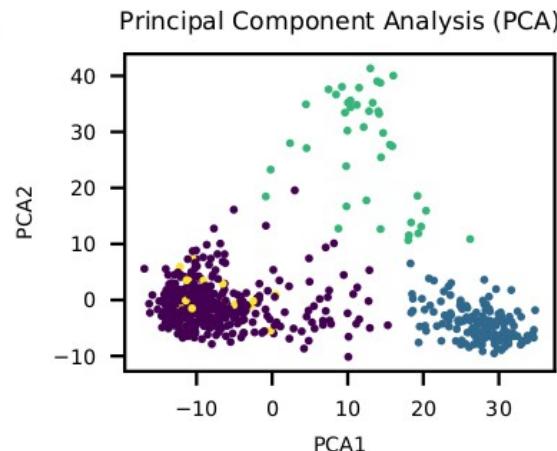
PCA PLOT WITH PRCOMP

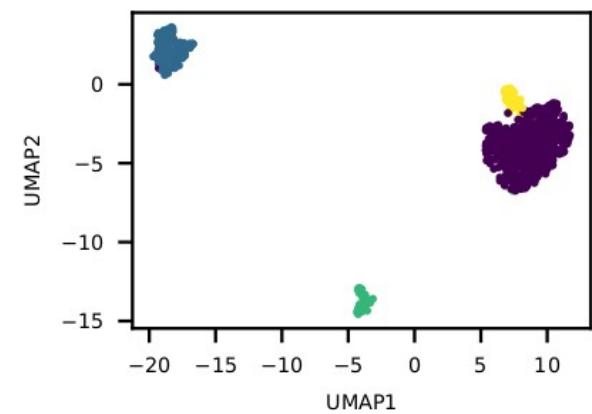
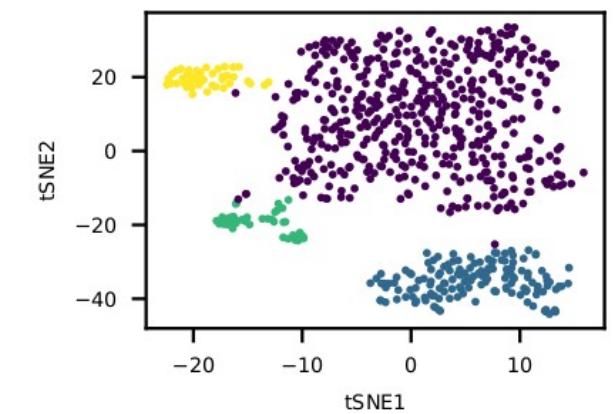
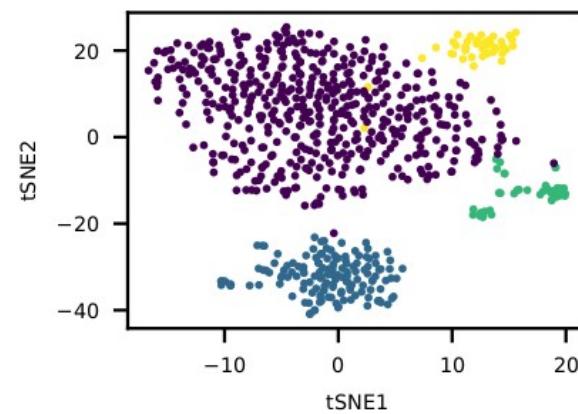
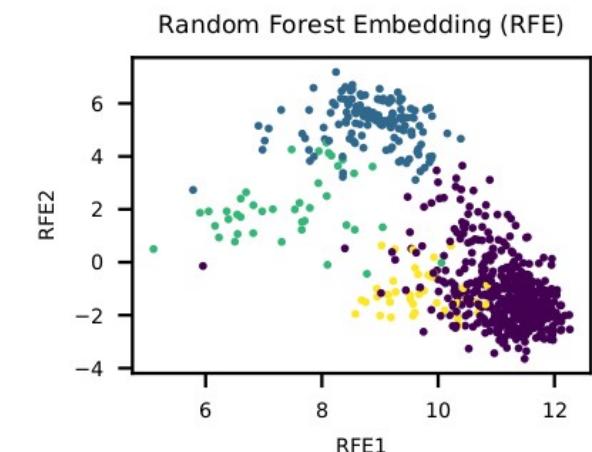
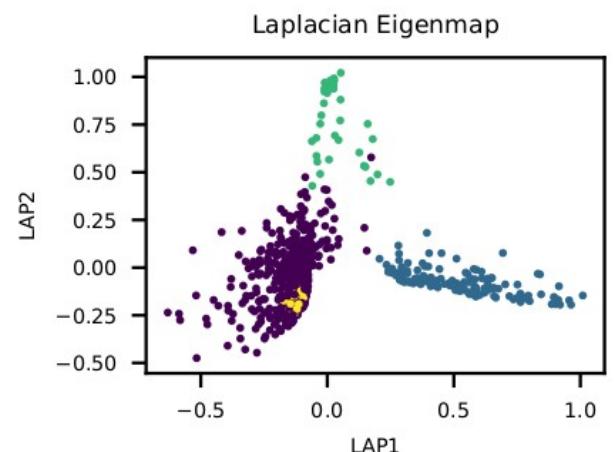
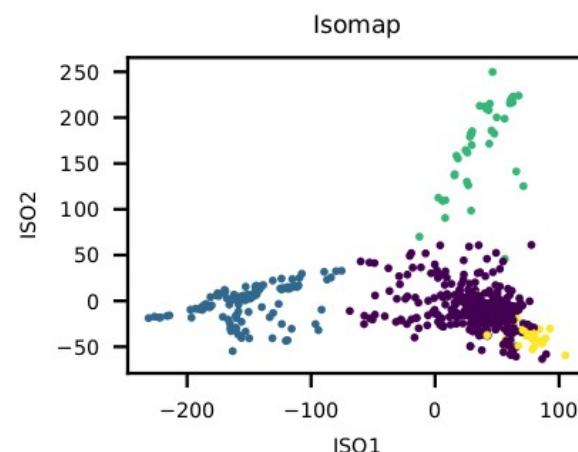
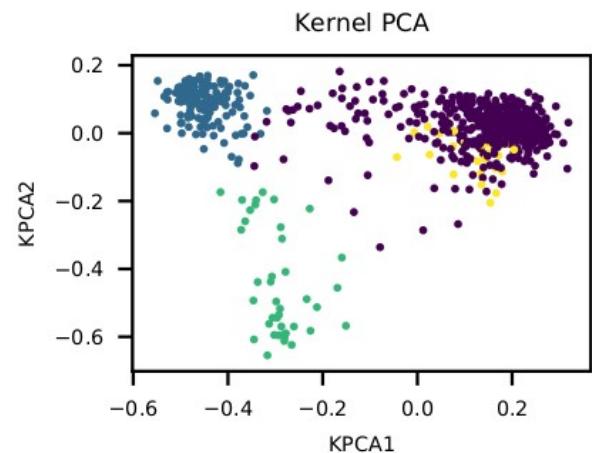
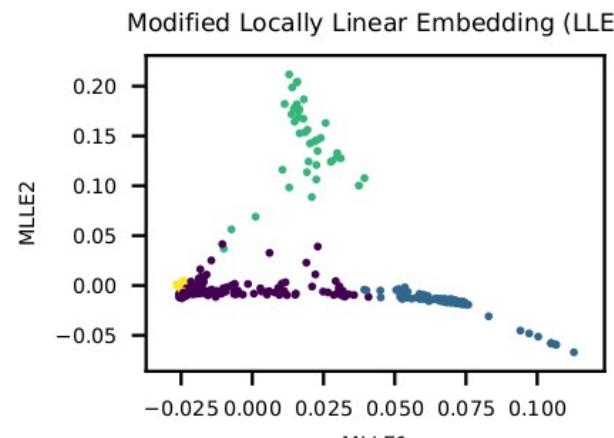
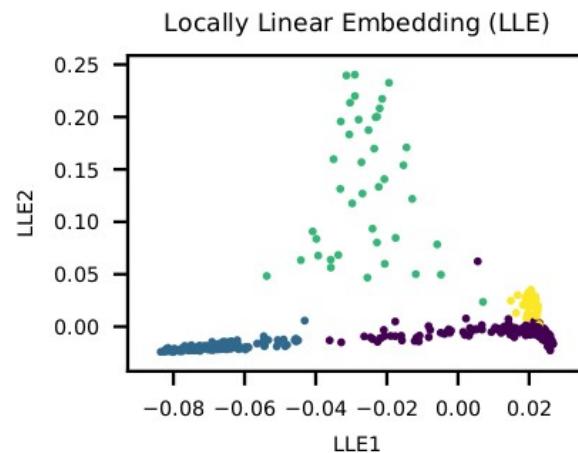


tSNE MNIST

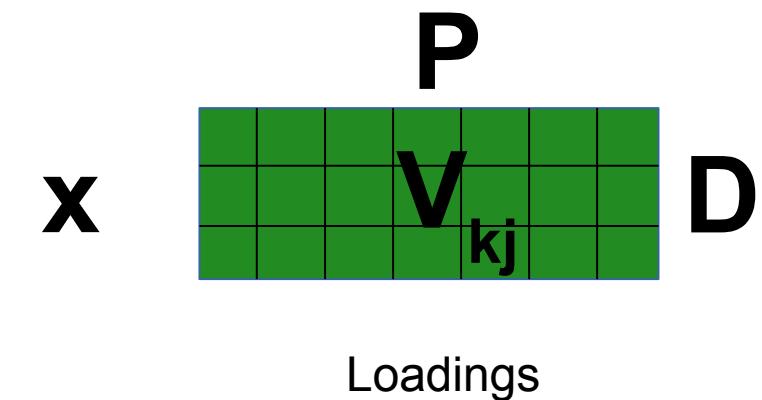
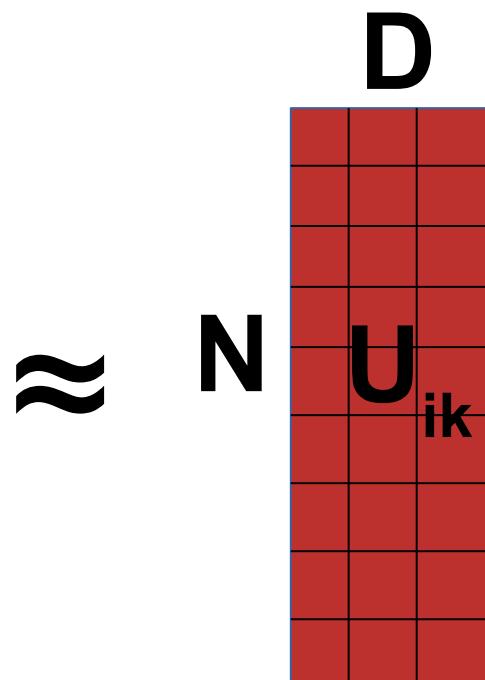
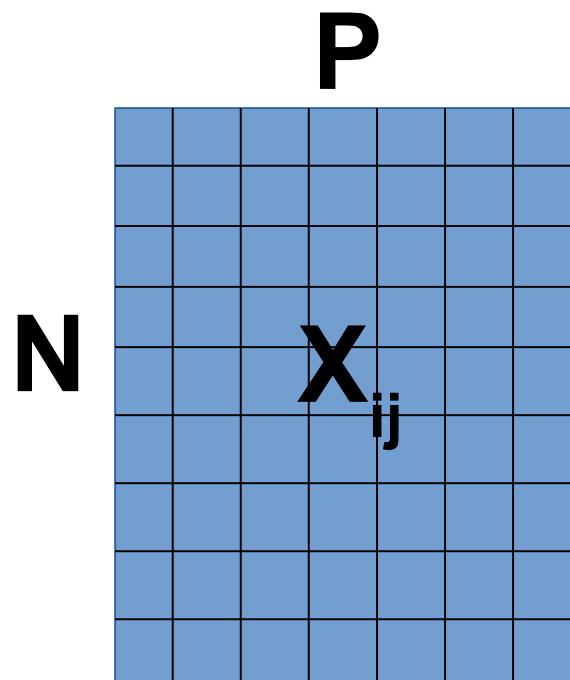


Dimension reduction is not only for visualization but overcoming the Curse of Dimensionality



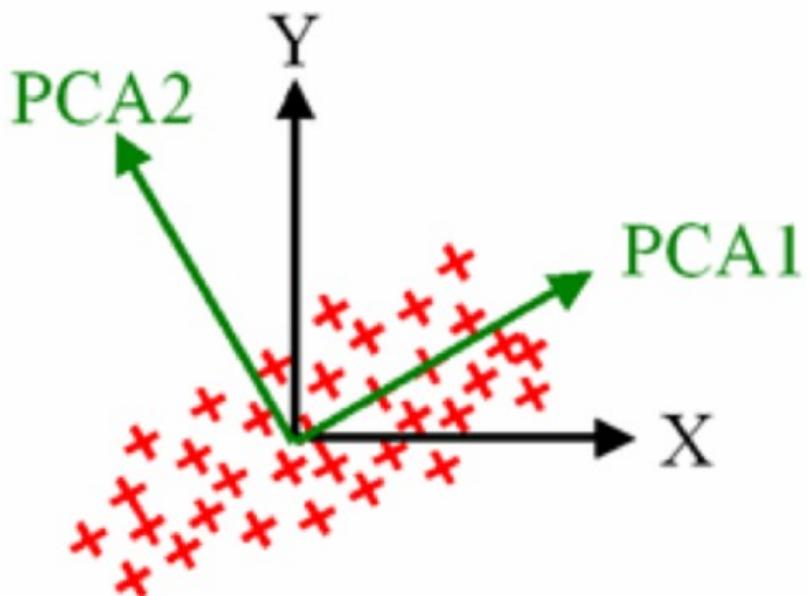


$$\mathbf{X}_{ij} \approx \mathbf{U}_{ik} \mathbf{V}_{kj}$$



$$\text{Loss} = \sum_{i=1}^N \sum_{j=1}^P (\mathbf{X}_{ij} - \mathbf{U}_{ik} \mathbf{V}_{kj})^2$$

- Collapse p features ($p >> n$) to few latent features and keep variation
- Rotation and shift of coordinate system toward maximal variance
- PCA is an **eigen matrix decomposition** problem



$$PC = u^T X = X^T u$$

X is mean centered $\implies \langle PC \rangle = 0$

$$\langle (PC - \langle PC \rangle)^2 \rangle = \langle PC^2 \rangle = u^T X X^T u$$

$$X X^T = A$$

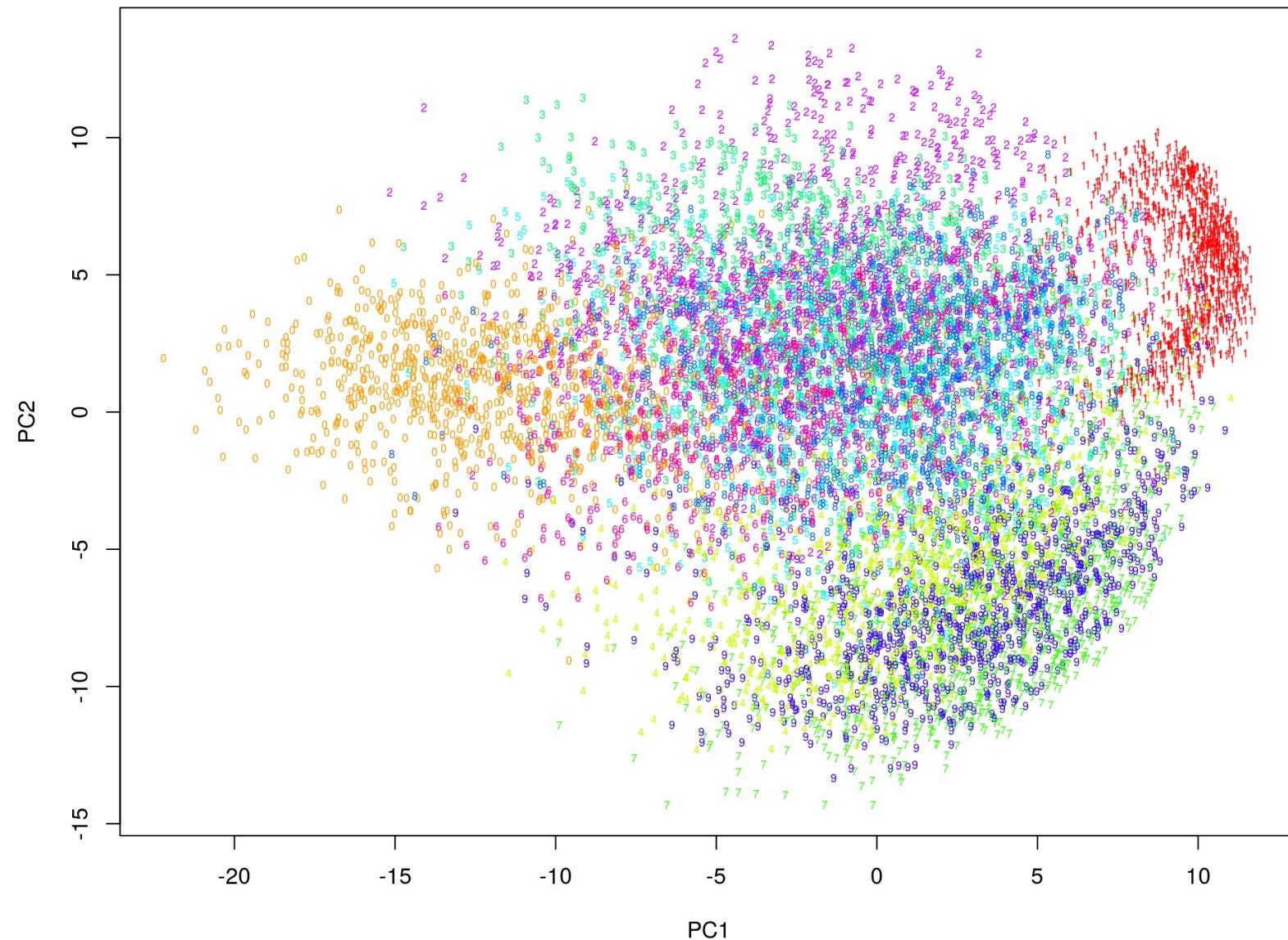
$$\langle PC^2 \rangle = u^T A u$$

A is **variance-covariance** of X

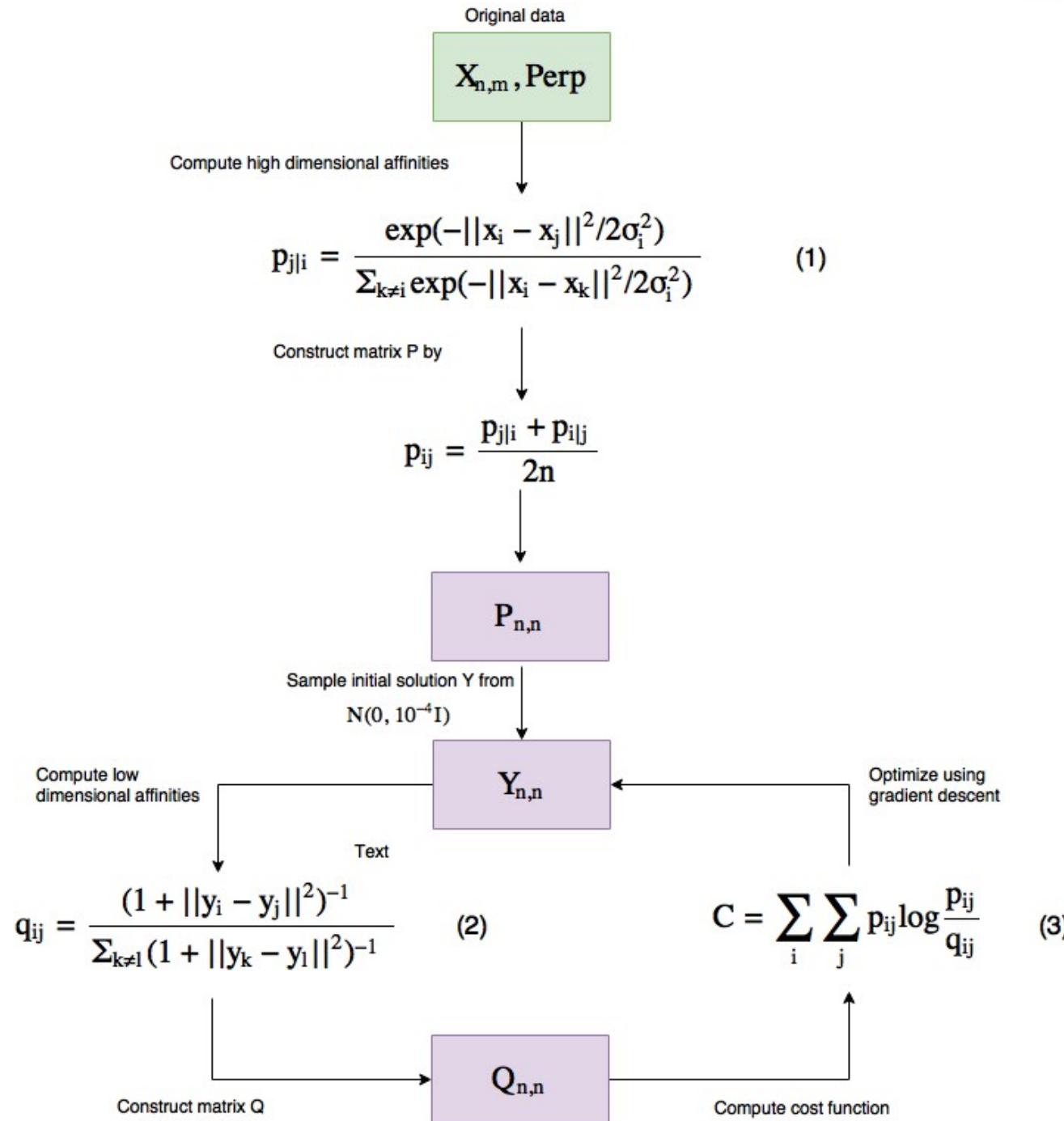
$$\max(u^T A u + \lambda(1 - u^T u)) = 0$$

$$A u = \lambda u$$

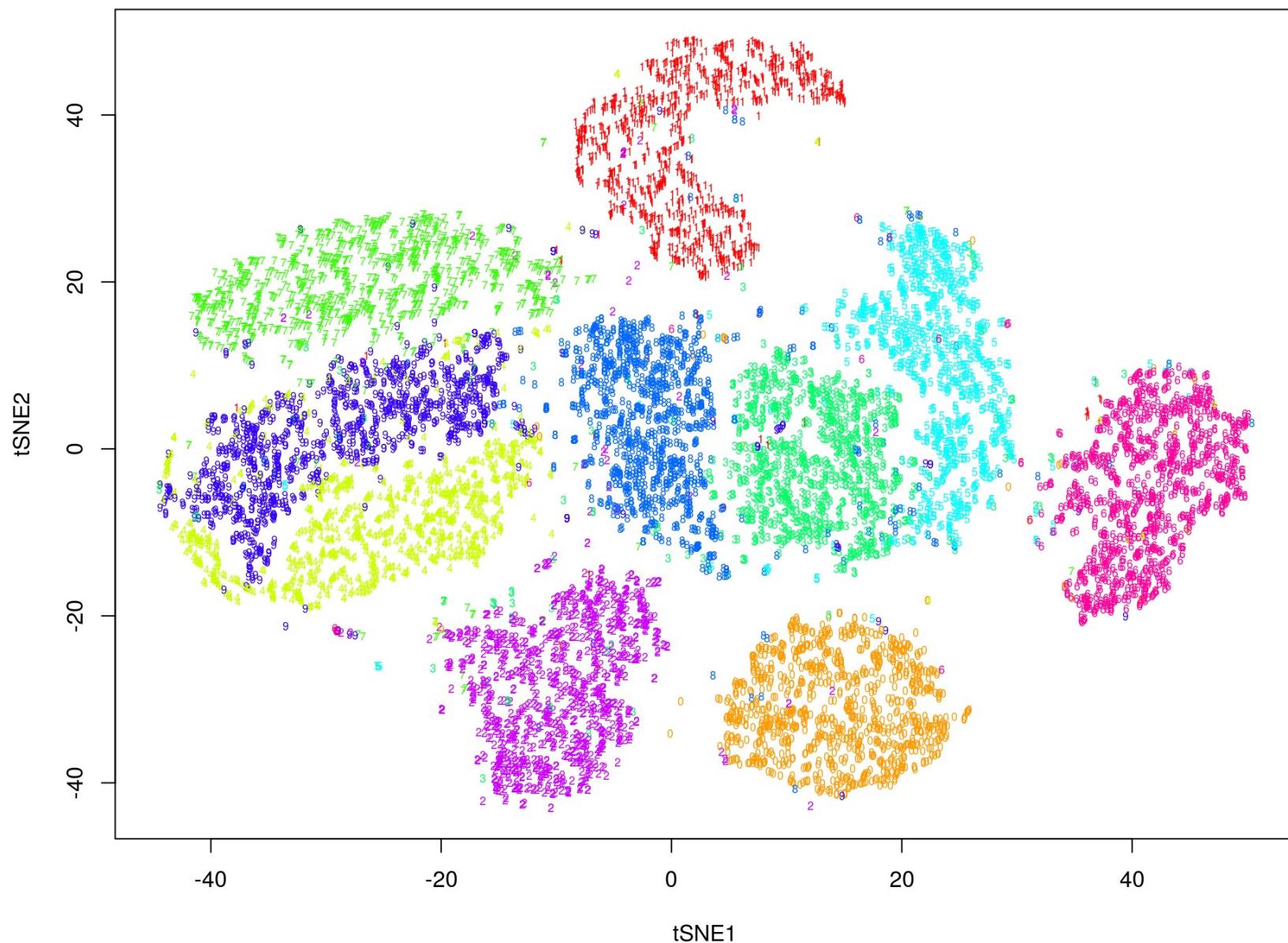
PCA PLOT WITH PRCOMP



t-distributed Stochastic Neighbor Embedding (tSNE)



tSNE MNIST



tSNE does not scale for large data sets?

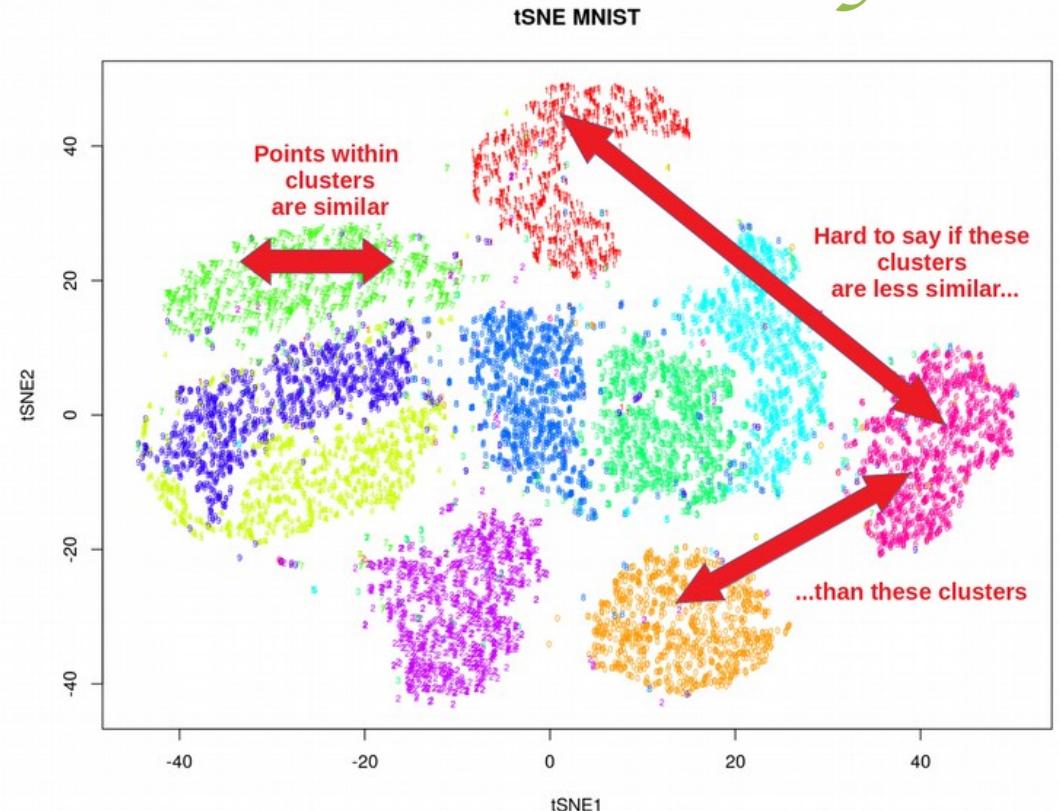
tSNE does not preserve global structure?

tSNE can only embed into 2-3 dims?

tSNE performs non-parametric mapping?

tSNE can not work with high-dimensional data directly (PCA needed)?

tSNE uses too much memory at large perplexities (FitSNE does not solve it)?

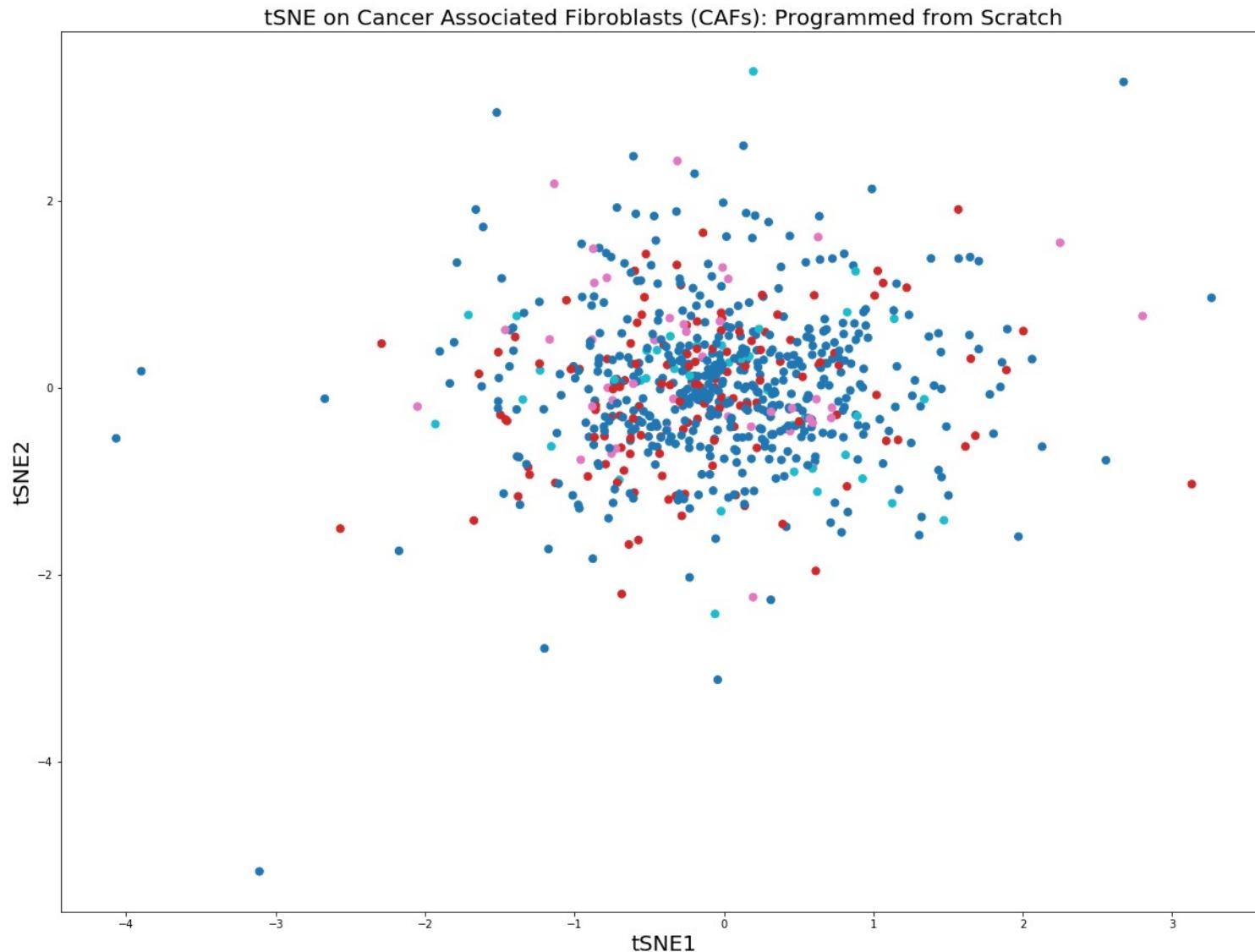


$$p_{j|i} = \frac{\exp(-\|x_i - x_j\|^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|x_i - x_k\|^2 / 2\sigma_i^2)}, \quad p_{ij} = \frac{p_{i|j} + p_{j|i}}{2N} \quad (1)$$

$$\text{Perplexity} = 2^{-\sum_j p_{j|i} \log_2 p_{j|i}} \quad (2)$$

$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq l} (1 + \|y_k - y_l\|^2)^{-1}} \quad (3)$$

$$KL(P_i || Q_i) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}}, \quad \frac{\partial KL}{\partial y_i} = 4 \sum_j (p_{ij} - q_{ij})(y_i - y_j) (1 + \|y_i - y_j\|^2)^{-1} \quad (4)$$



UMAP uses local connectivity for high-dim probabilities

$$p_{i|j} = e^{-\frac{d(x_i, x_j) - \rho_i}{\sigma_i}}$$

UMAP does not normalize probabilities (speed-up)

$$k = 2 \sum_i p_{ij}$$

UMAP uses slightly different expression for nearest neighbors

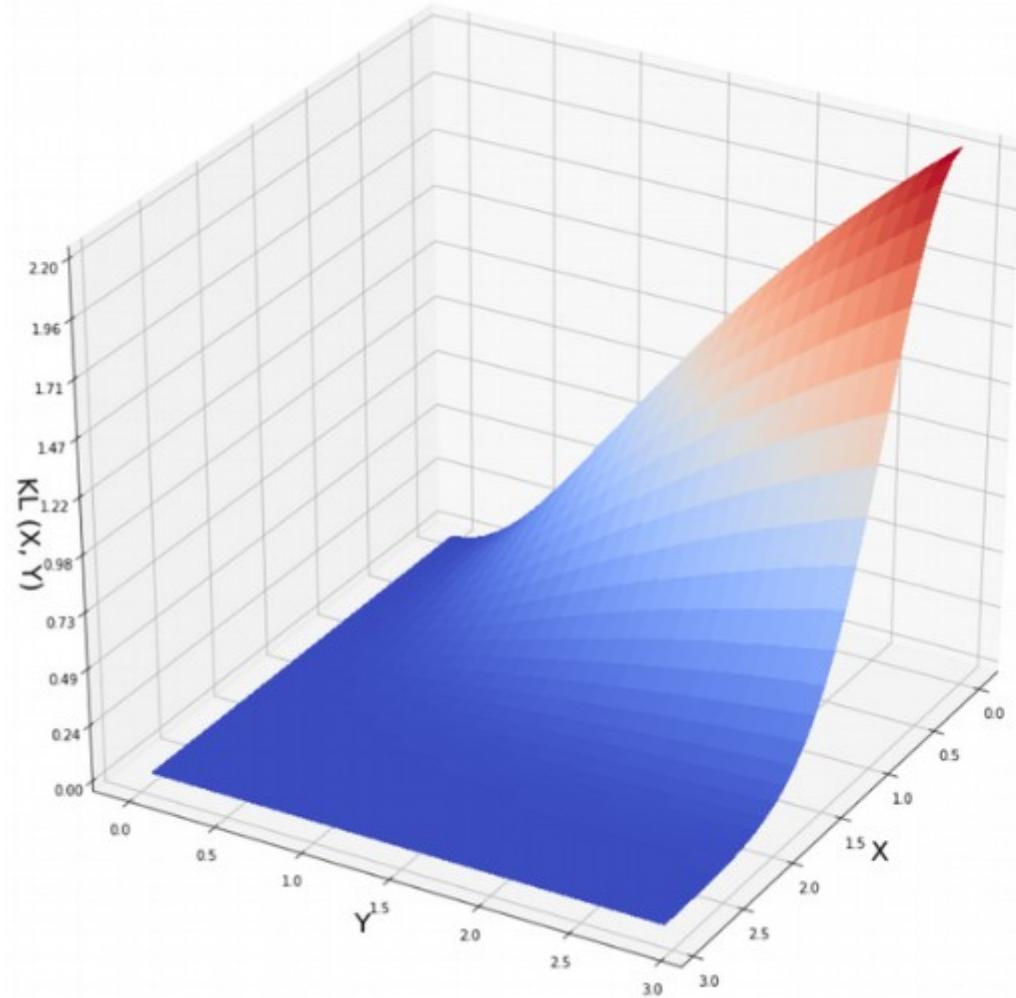
UMAP uses Laplacian Eigenmap for initialization

UMAP uses Cross-Entropy (not KL) as cost function

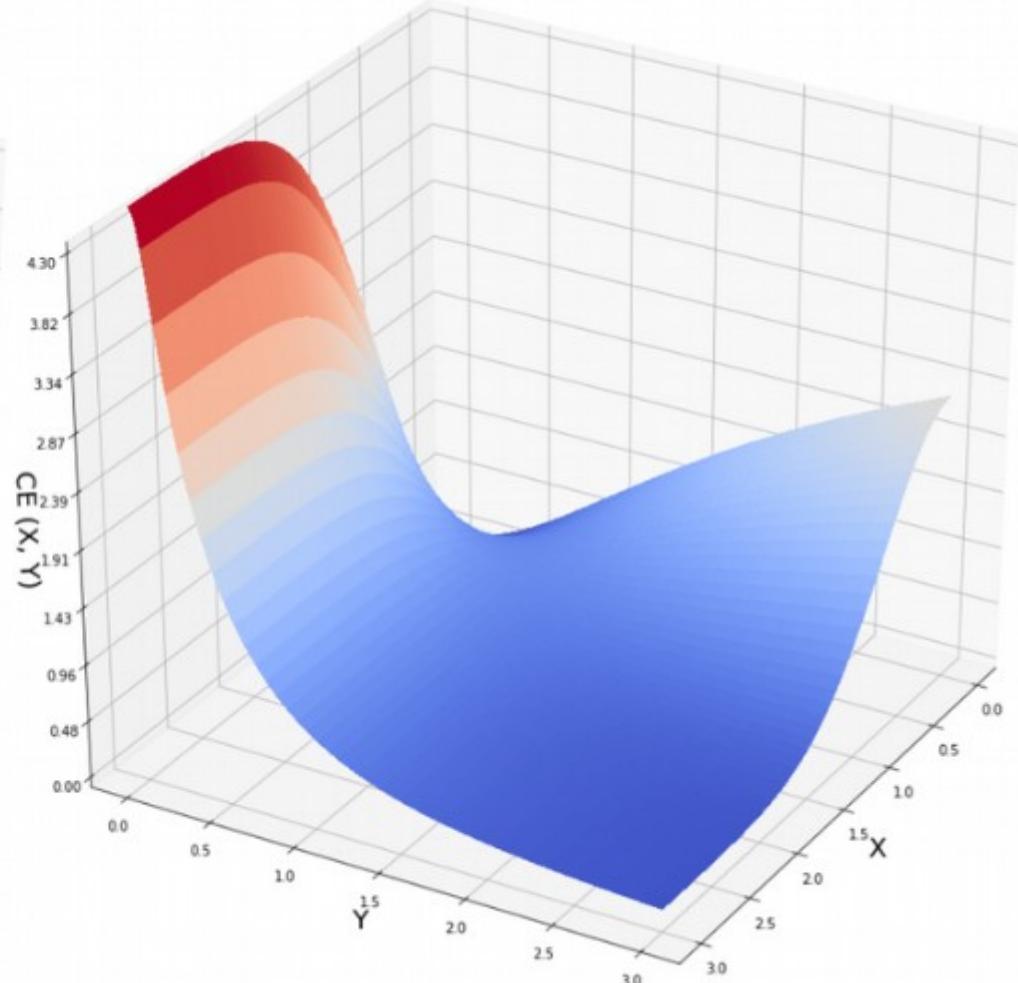
$$CE(X, Y) = \sum_i \sum_j \left[p_{ij}(X) \log \left(\frac{p_{ij}(X)}{q_{ij}(Y)} \right) + (1 - p_{ij}(X)) \log \left(\frac{1 - p_{ij}(X)}{1 - q_{ij}(Y)} \right) \right]$$

This is similar to tSNE cost function

This term is UMAP specific

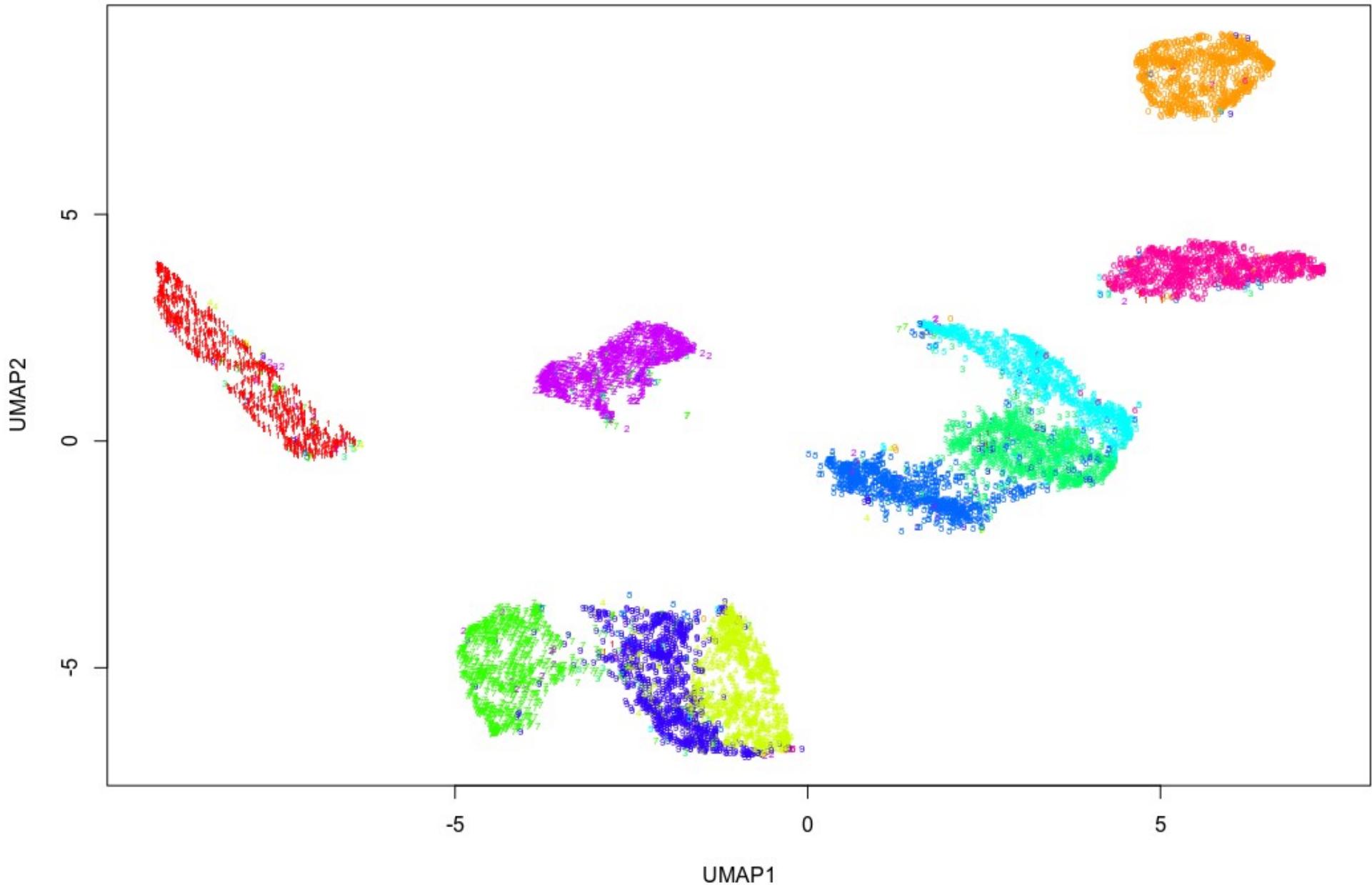


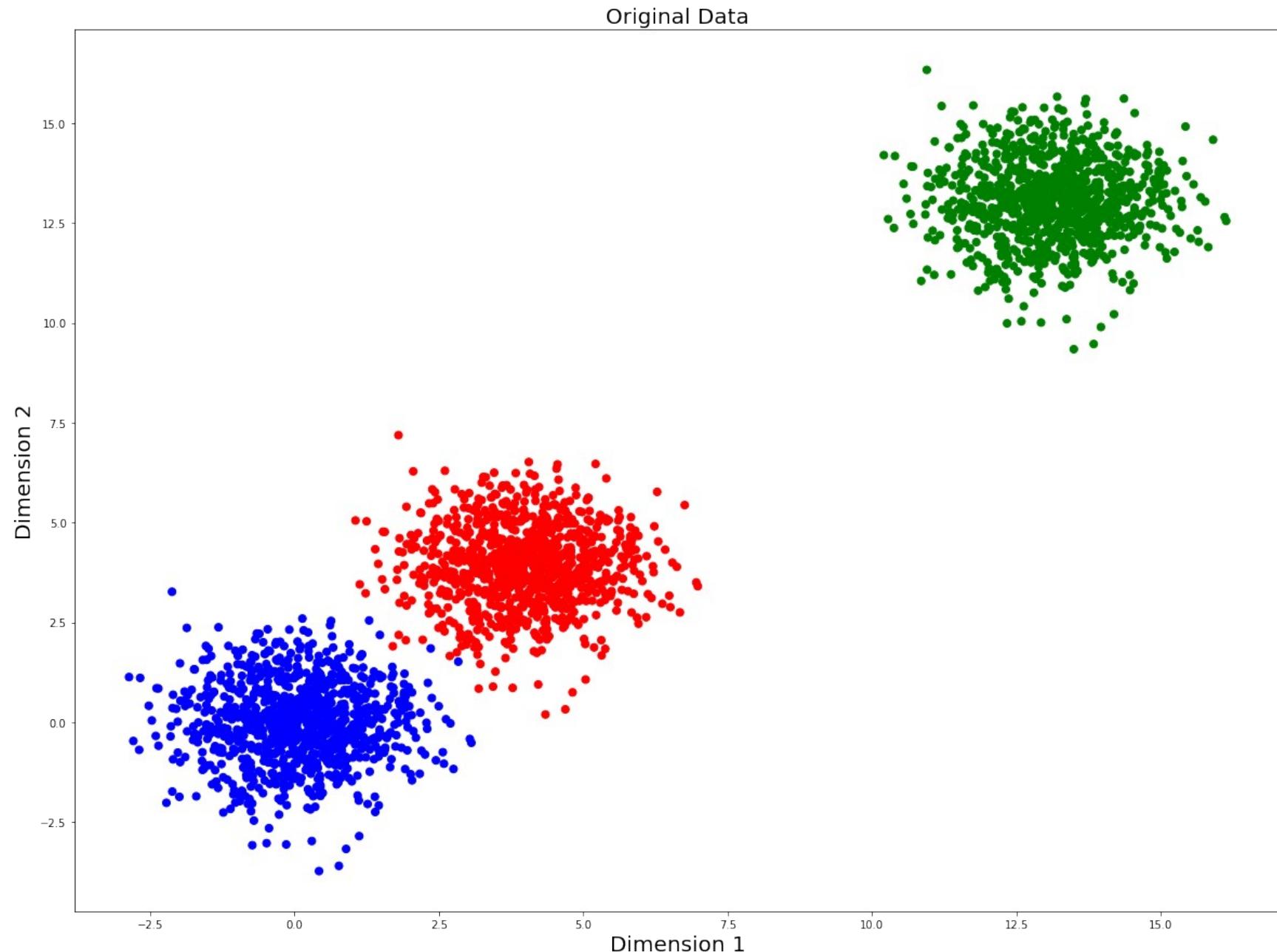
$X \rightarrow \infty$, Y can be any

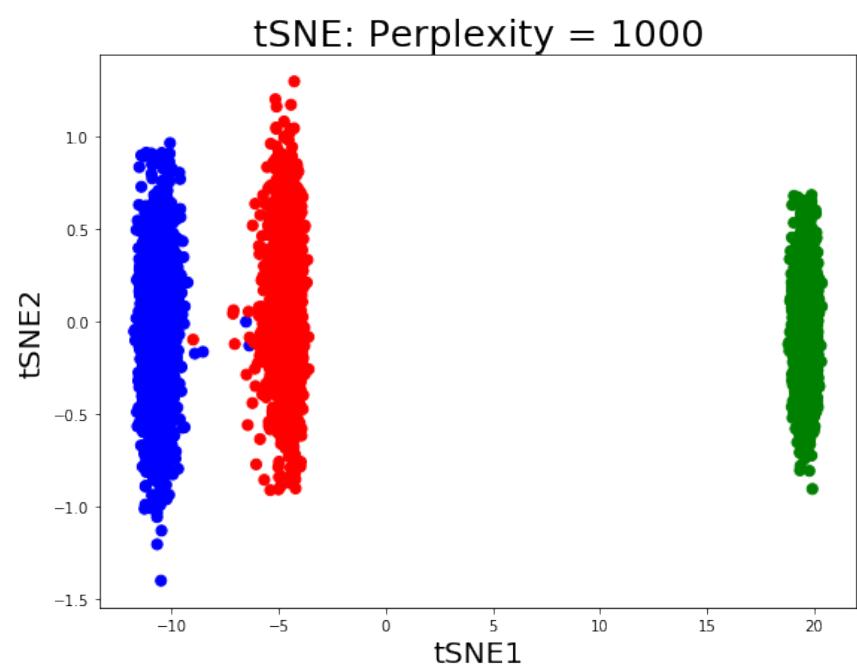
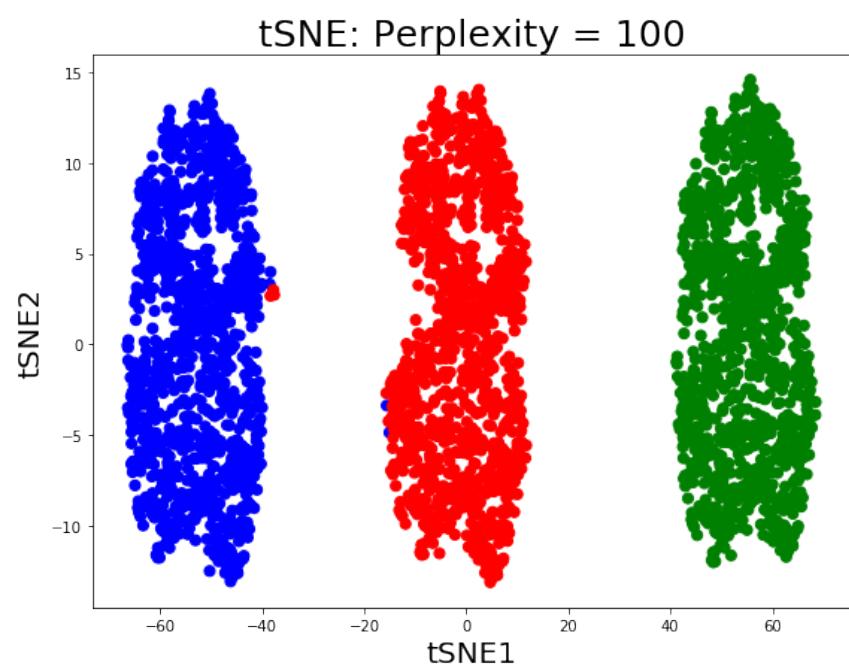
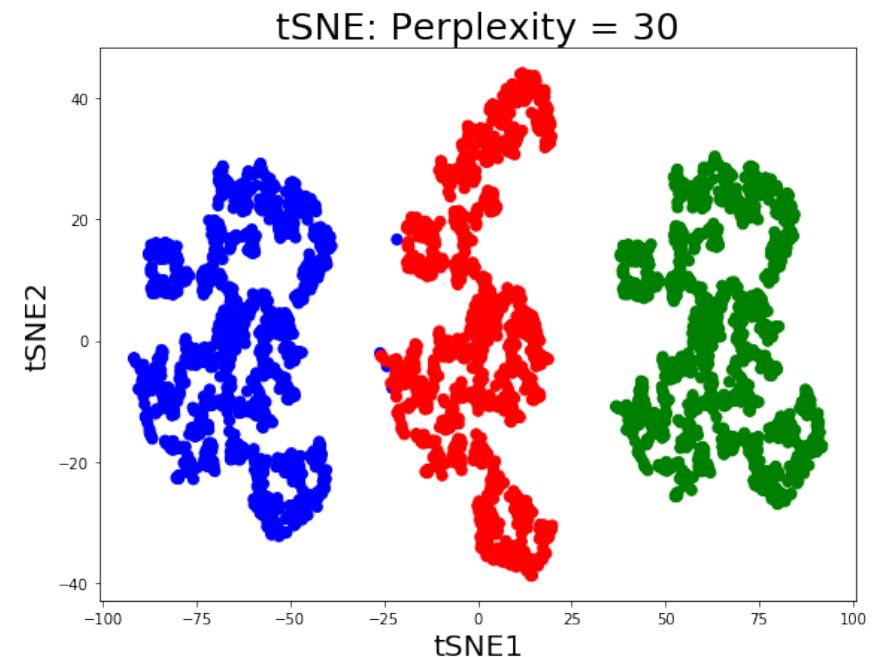
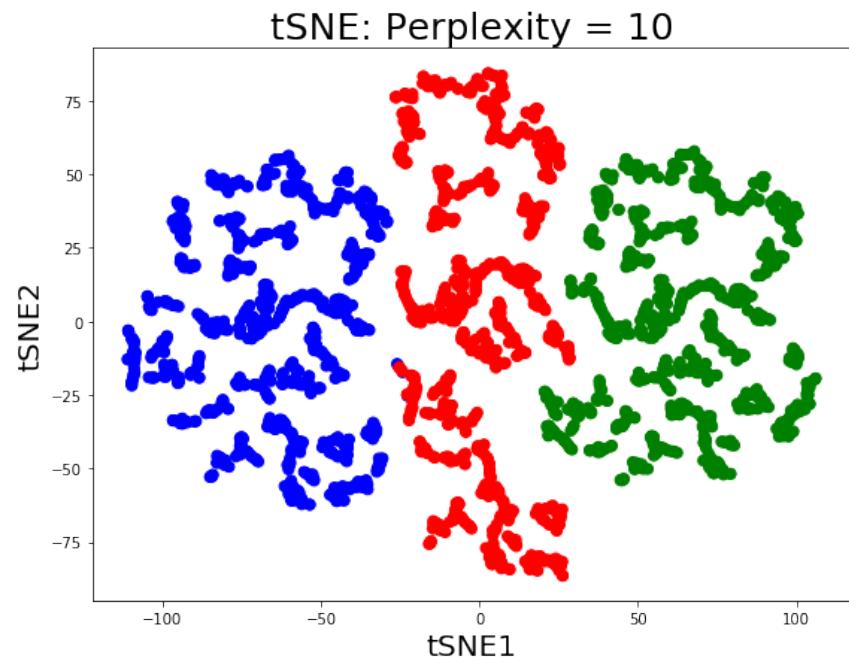


$X \rightarrow \infty$, $Y \rightarrow \infty$

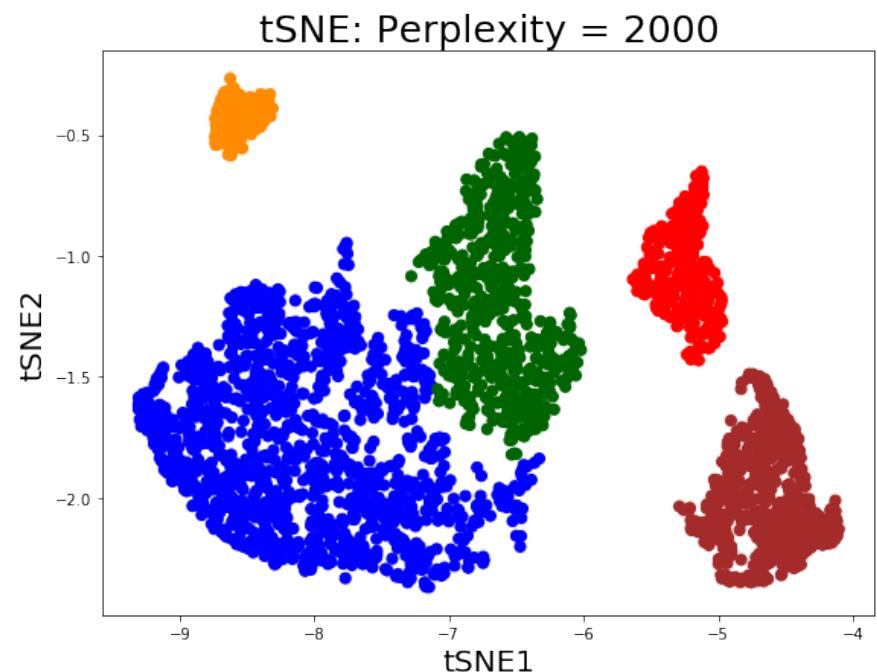
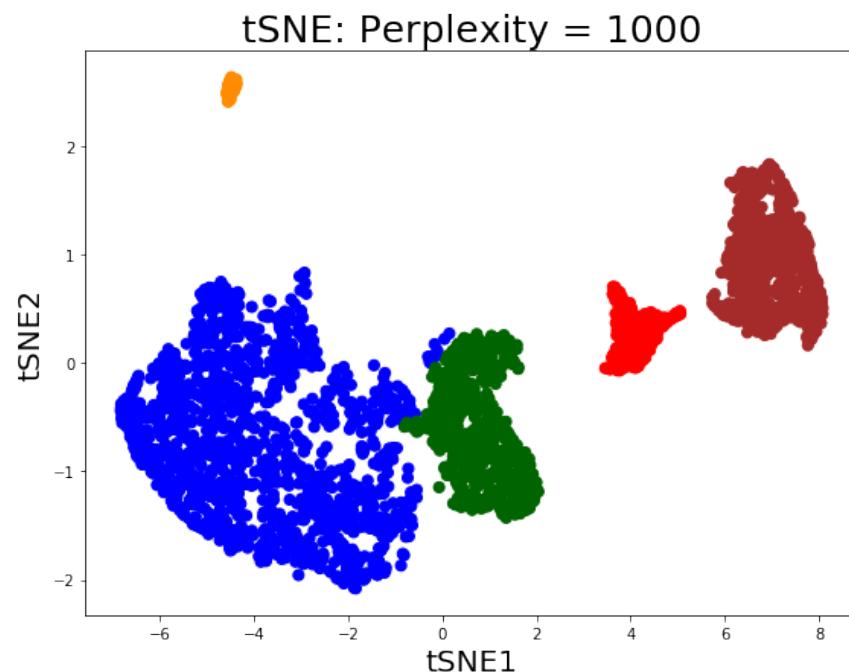
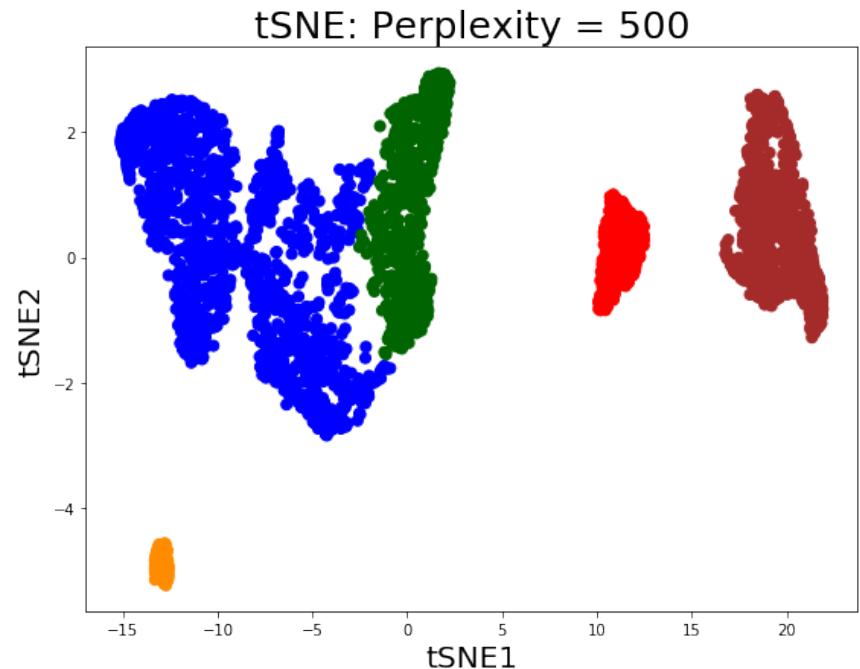
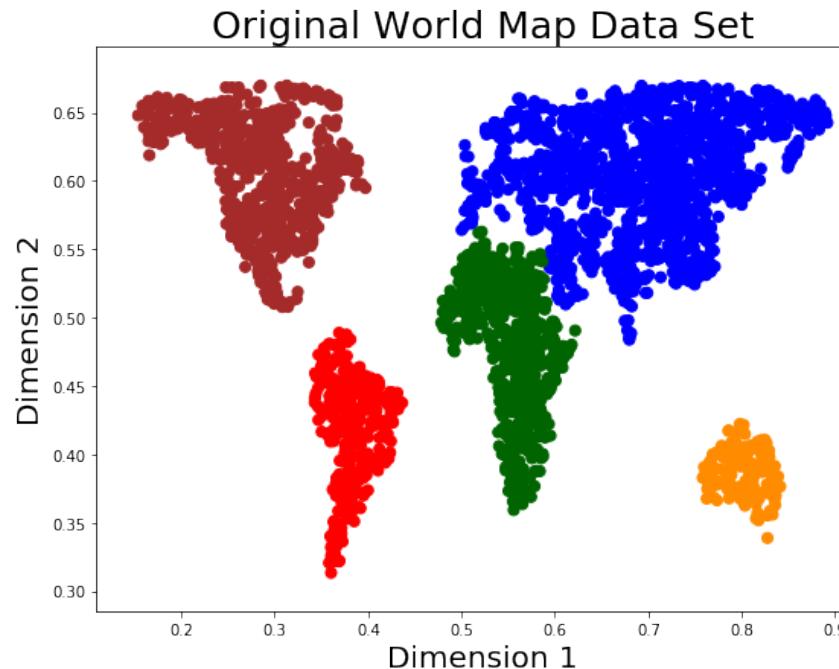
UMAP MNIST





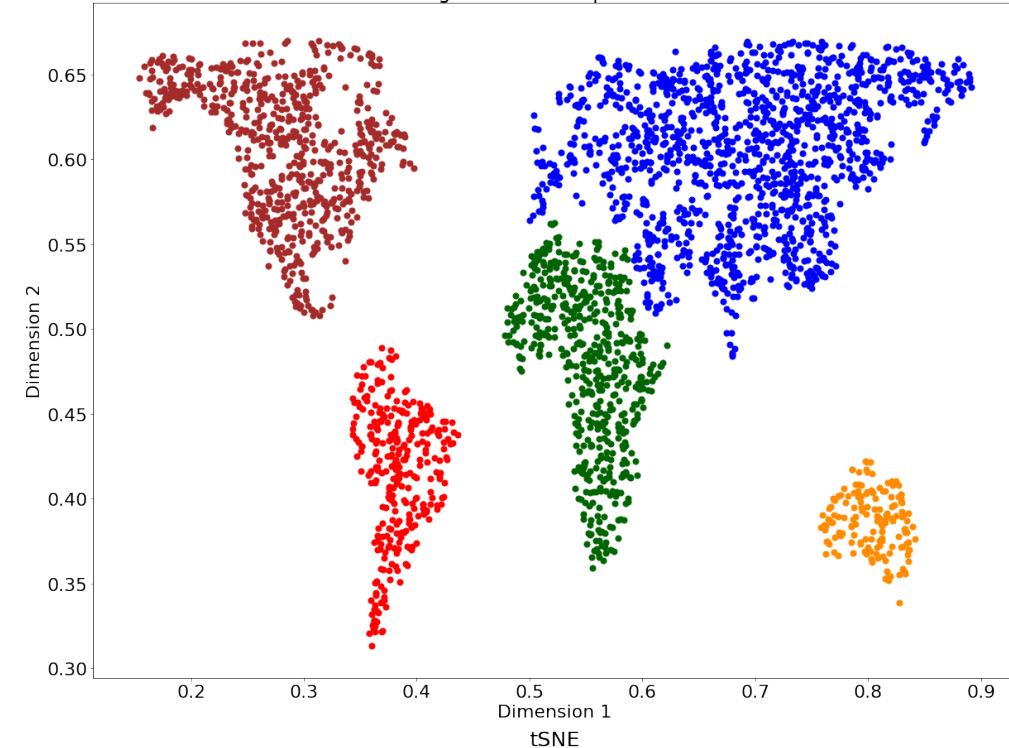


Why Global Structure Preservation is Important: Example #2

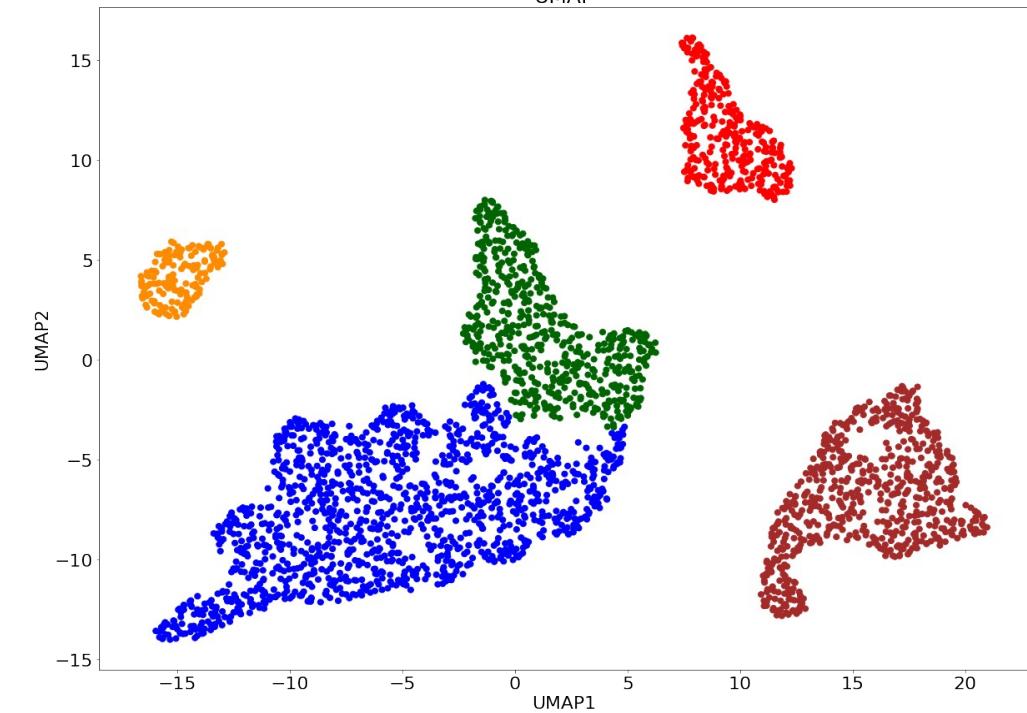
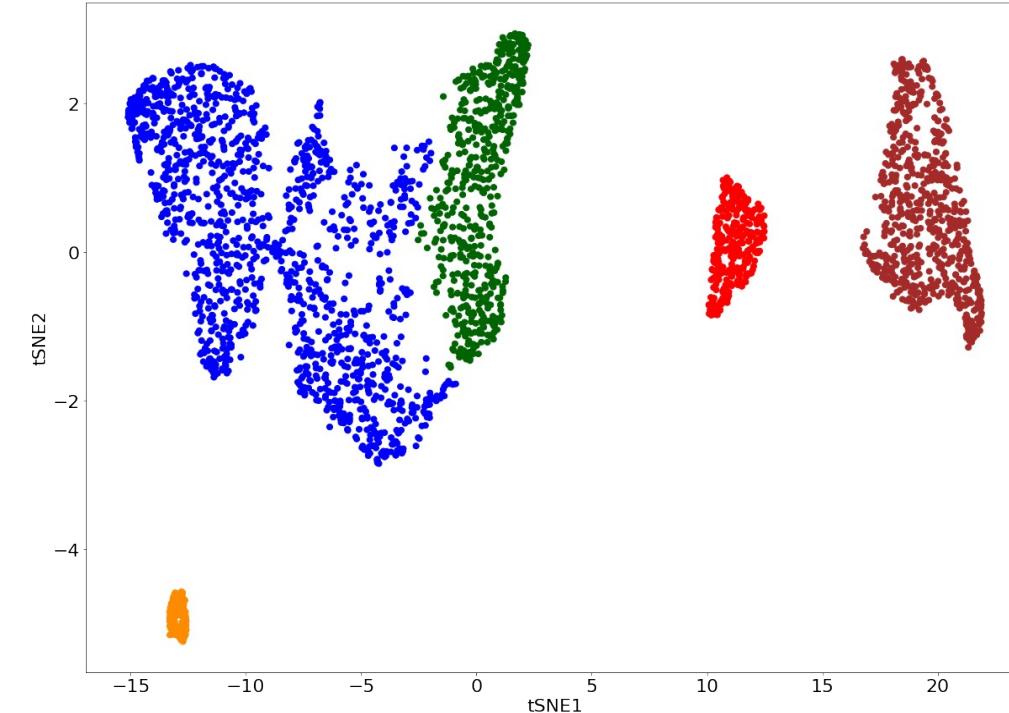
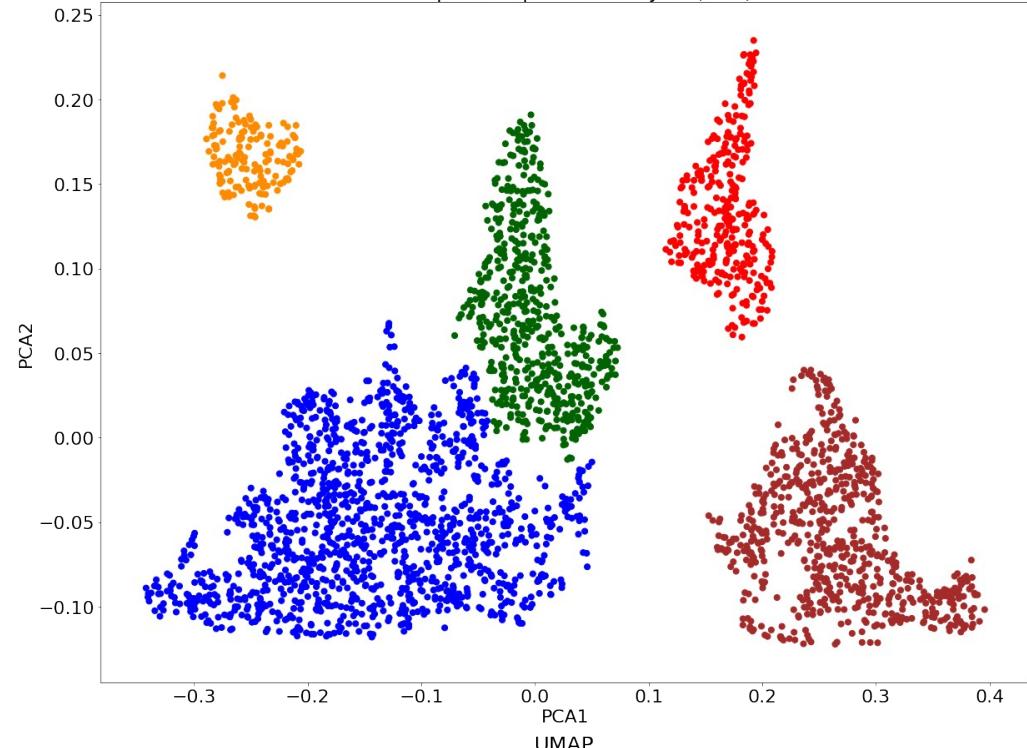


Global Structure Preservation for Linear Manifold

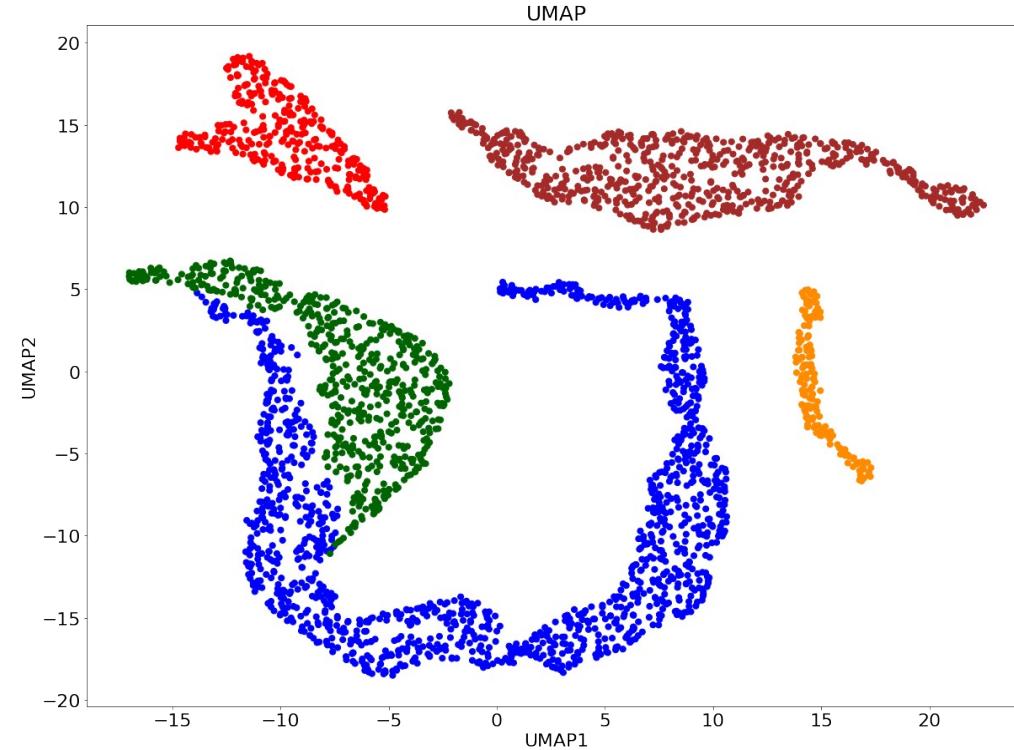
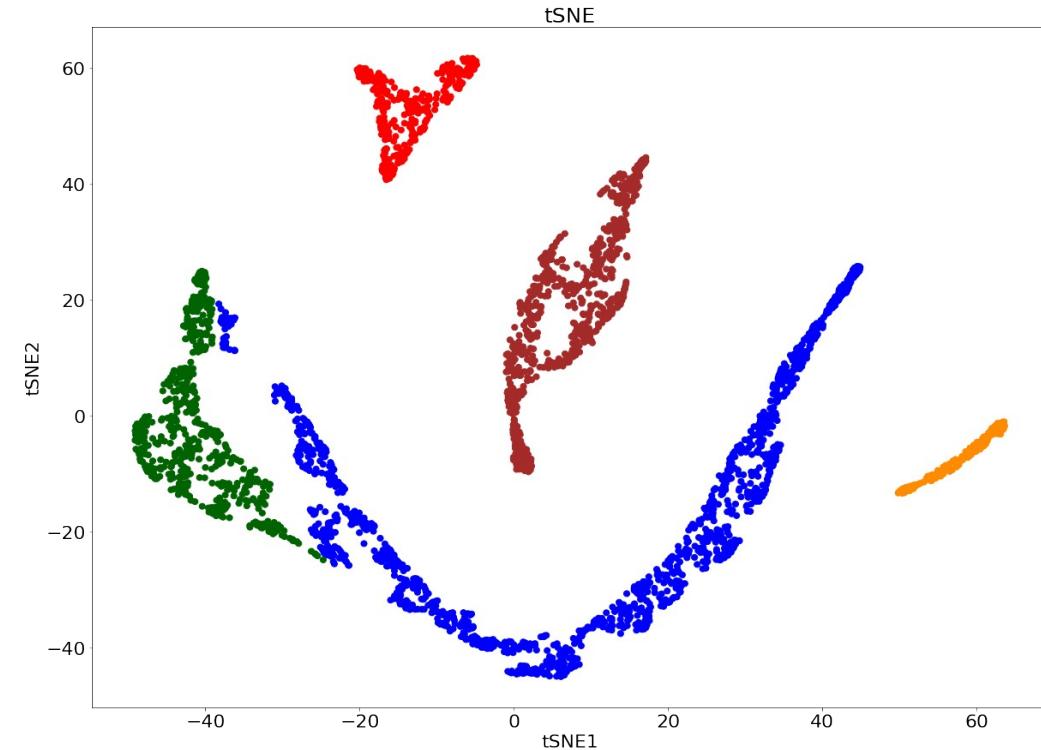
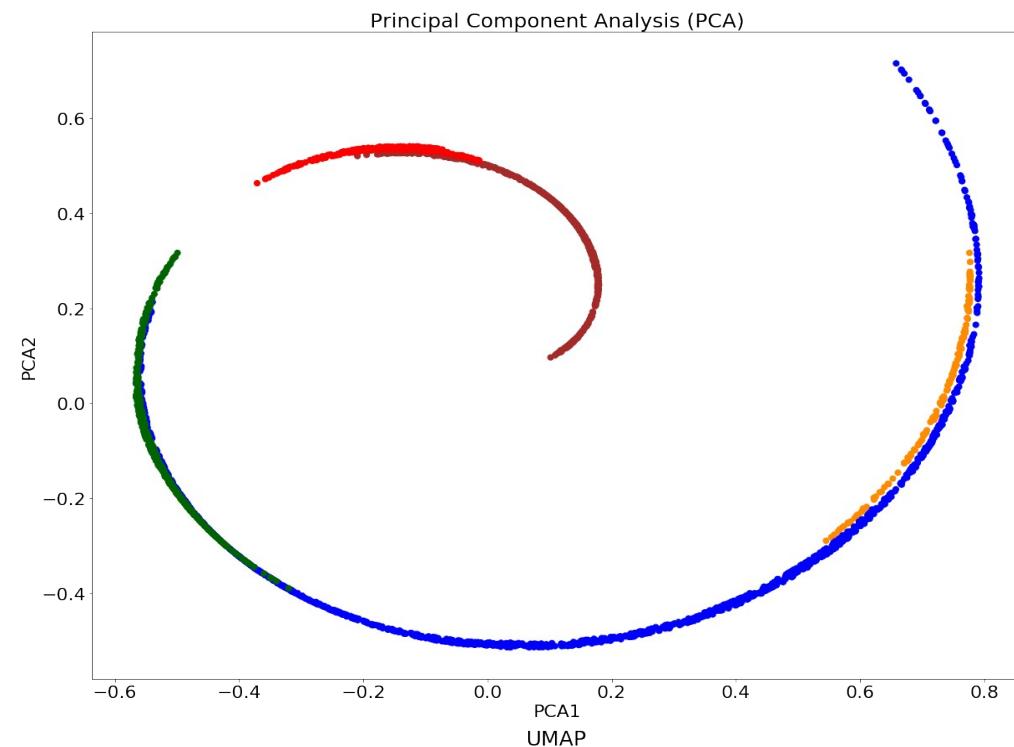
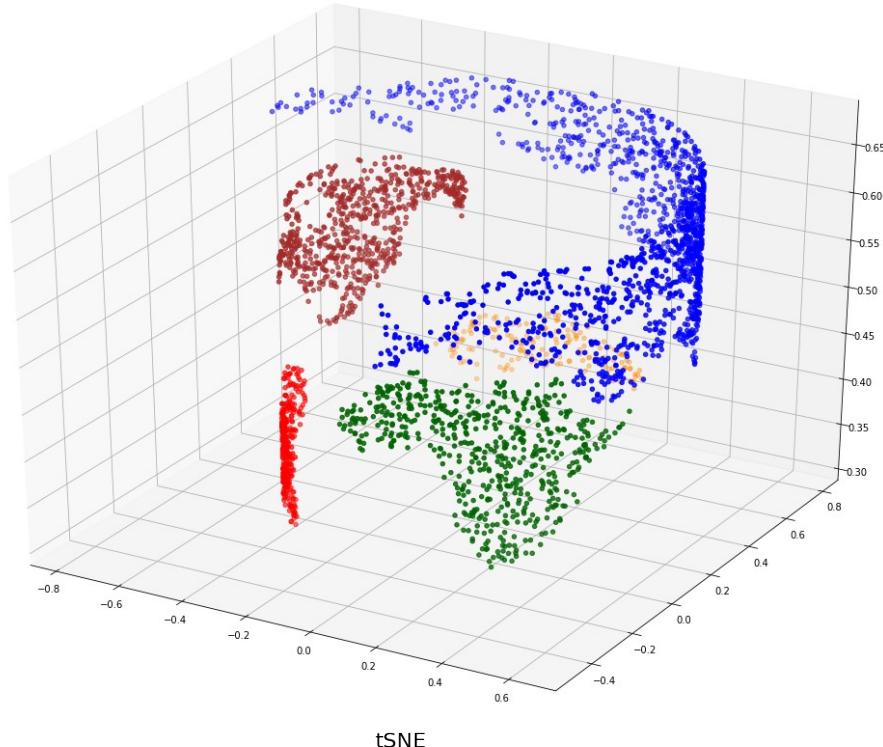
Original World Map Data Set



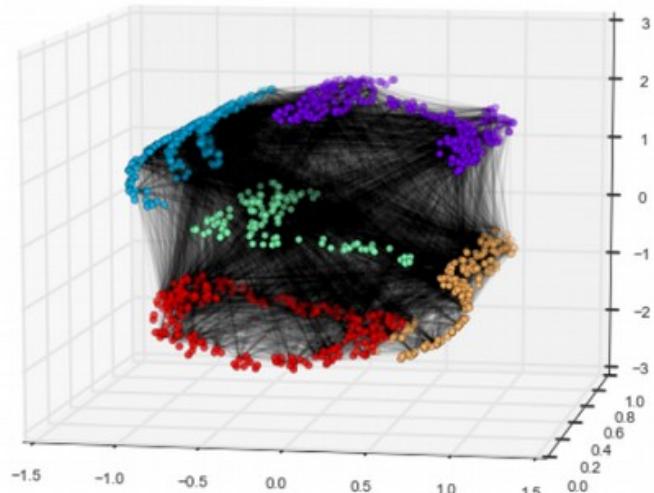
Principal Component Analysis (PCA)



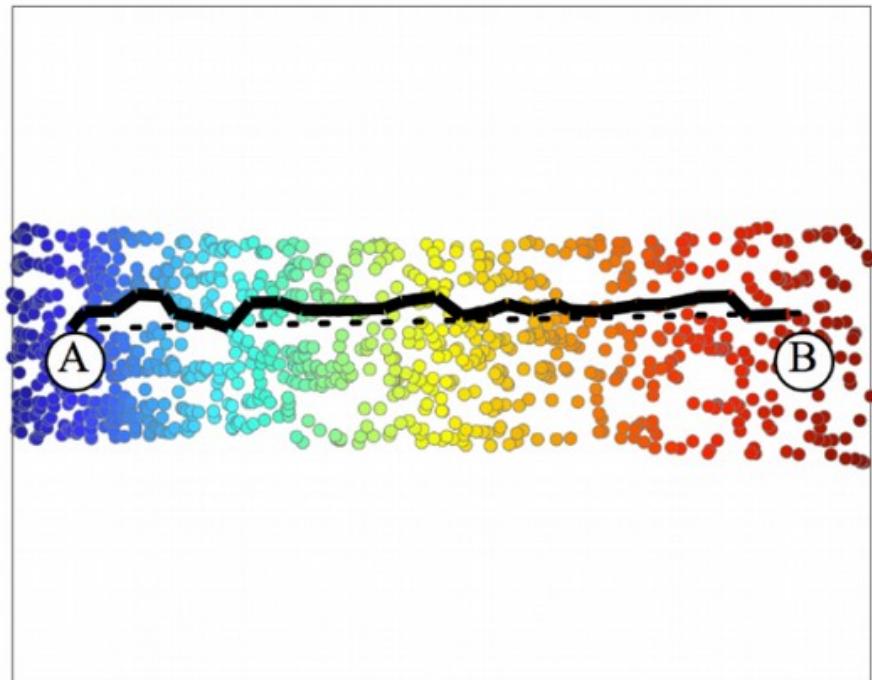
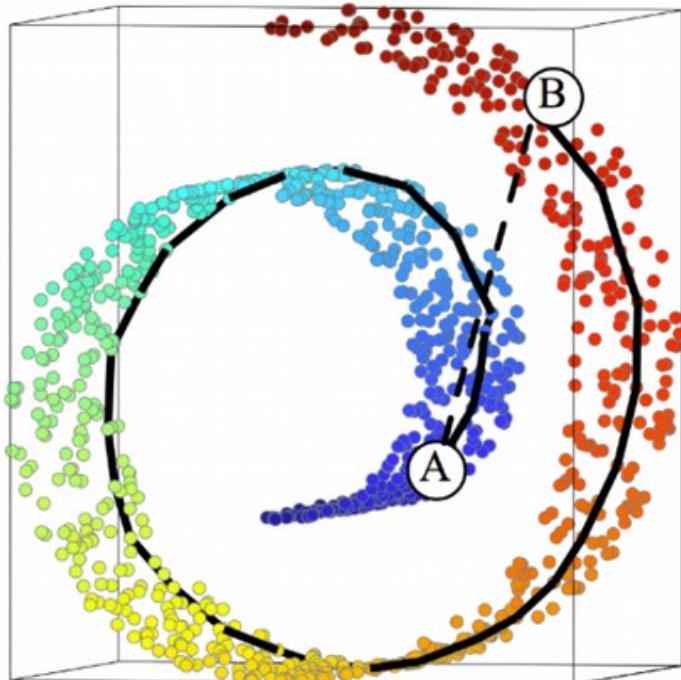
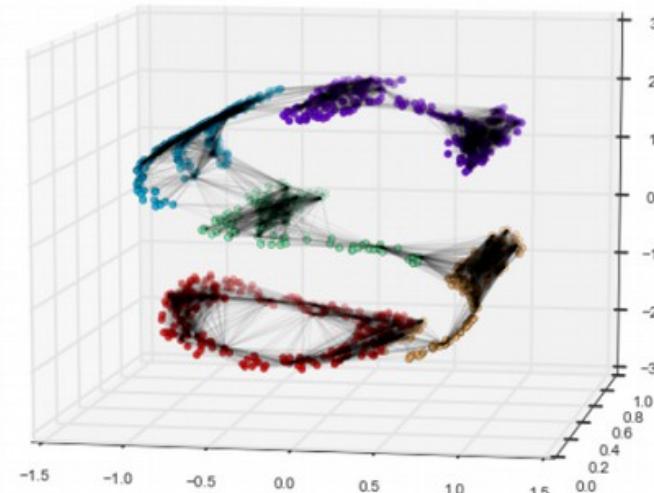
Global Structure Preservation for Non-Linear Manifold



MDS Linkages

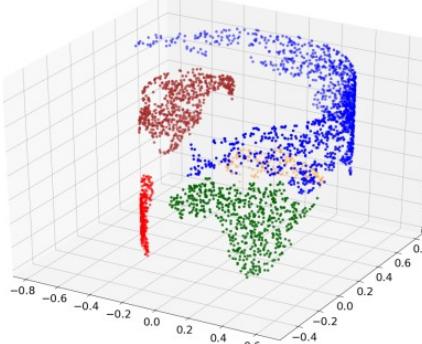


LLE Linkages (100 NN)

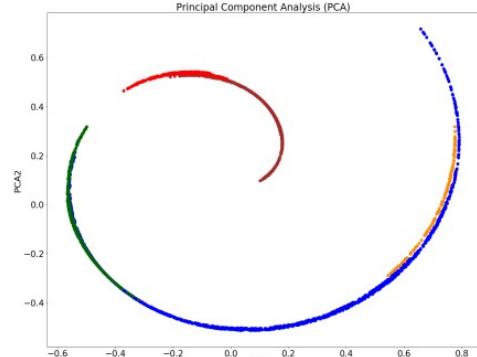


Non-Linear Manifold: PCA vs. tSNE vs. UMAP

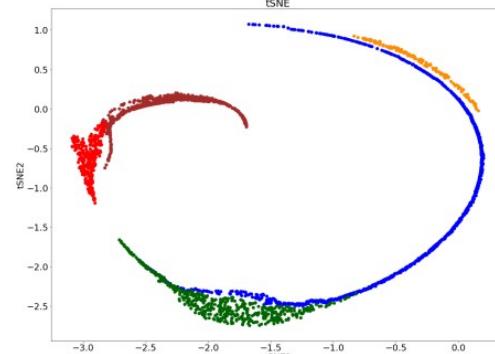
Swiss Roll: 3023 points



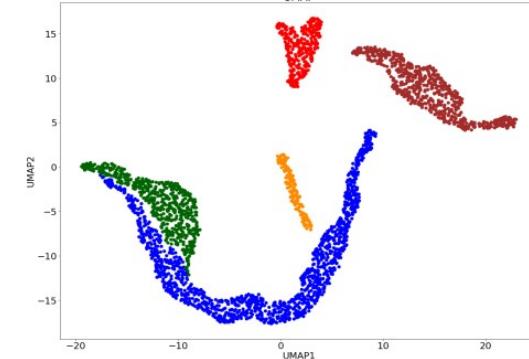
Principal Component Analysis



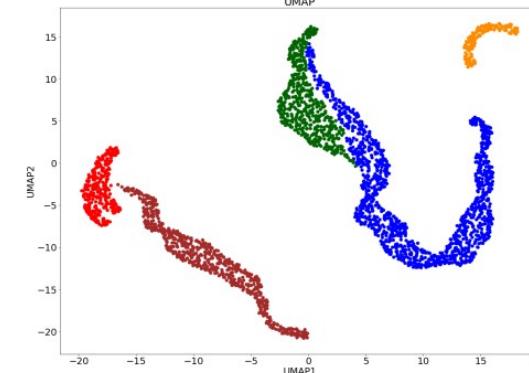
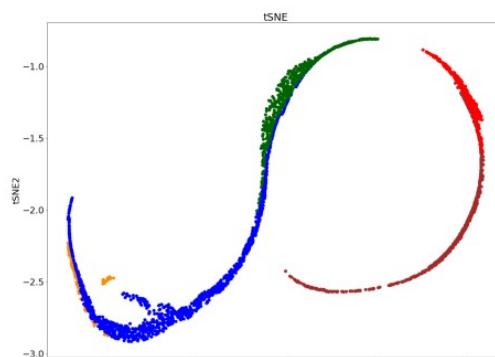
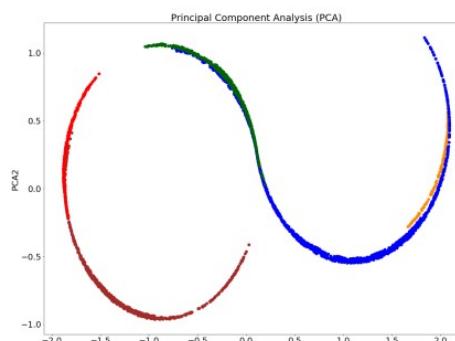
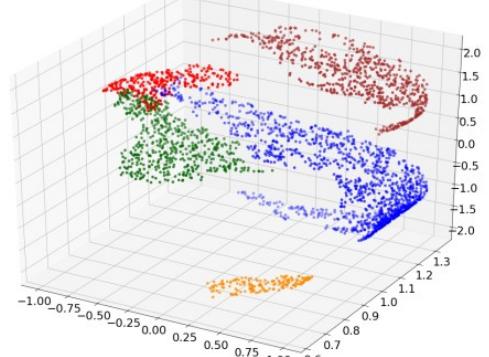
tSNE: perplexity = 2000



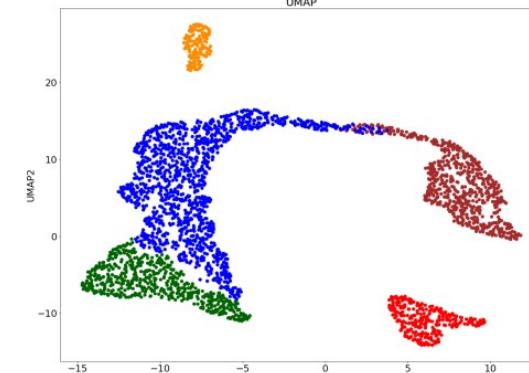
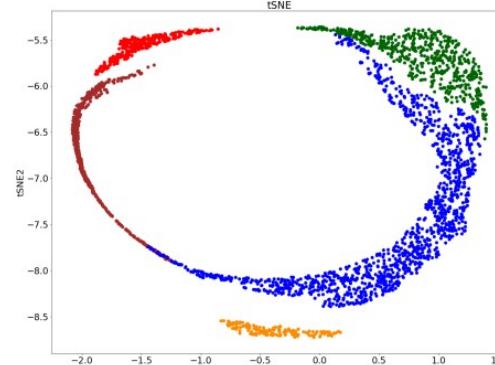
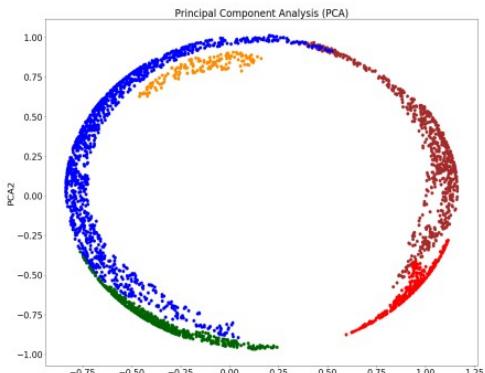
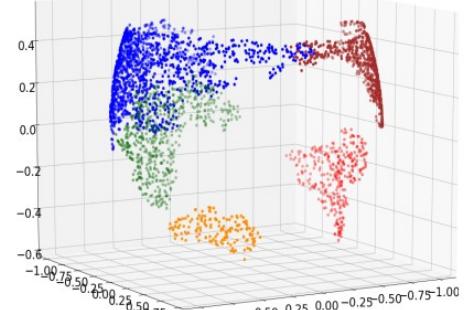
UMAP: n_neighbor = 2000

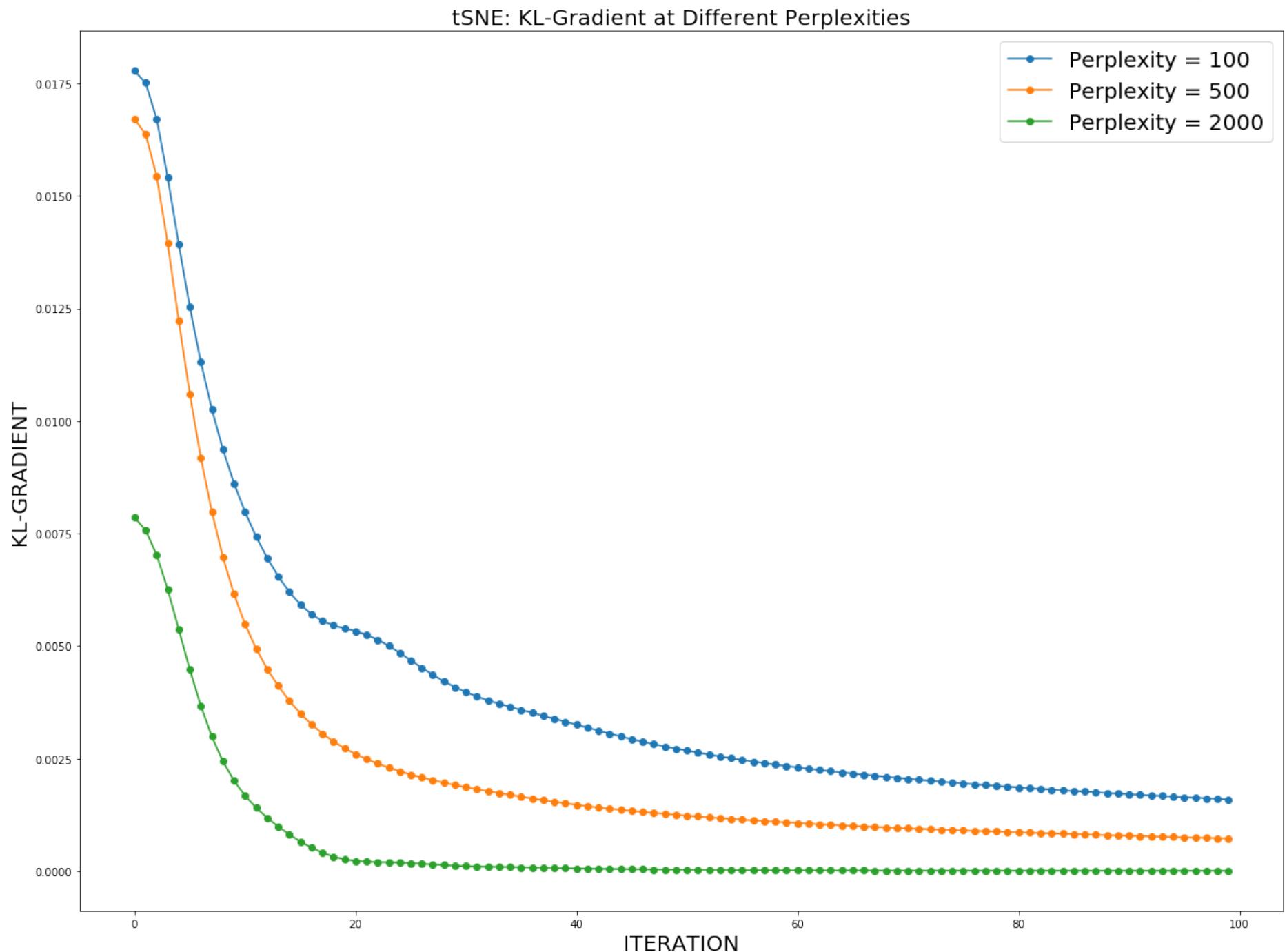


S-shape: 3023 points

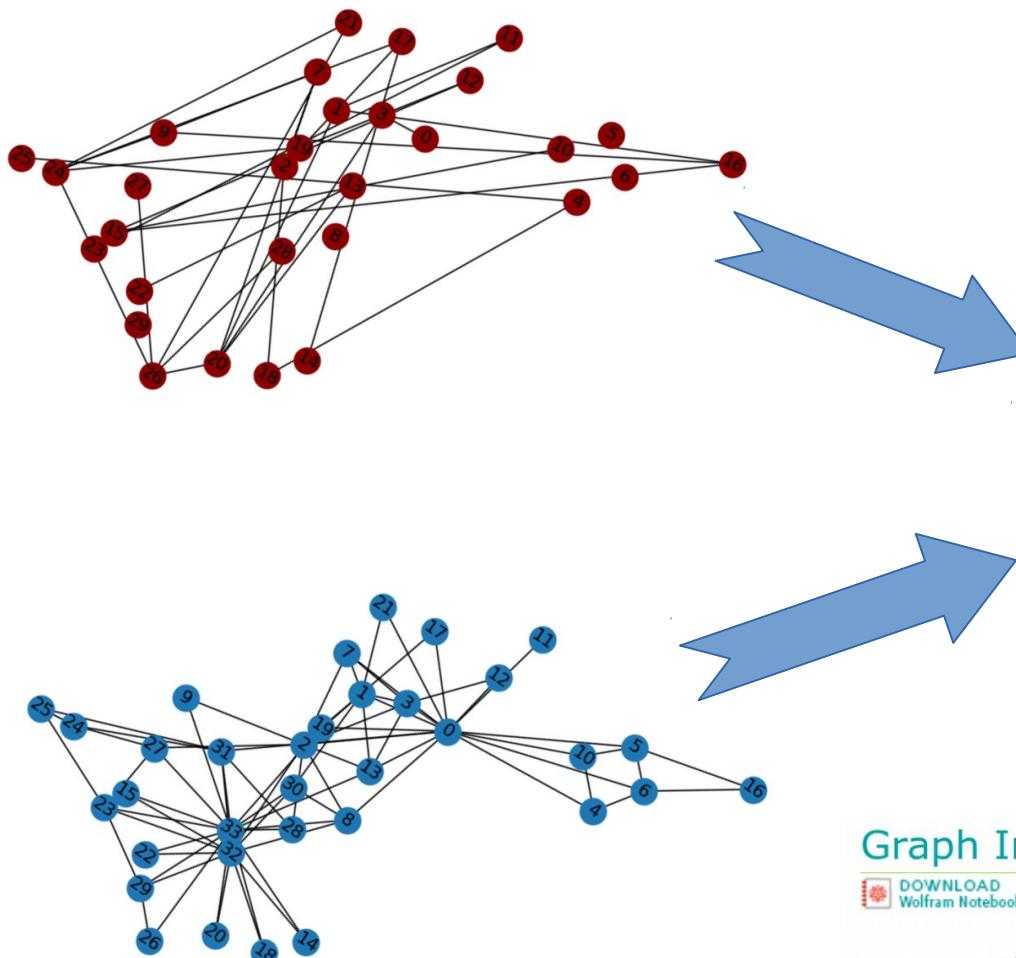


Sphere: 3023 points





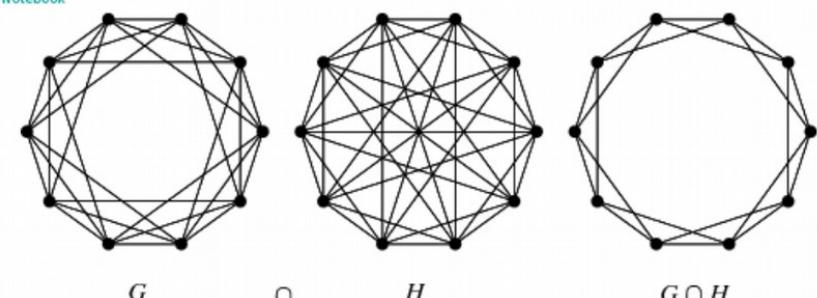
UMAP for OMICs Integration



Integration: keep edges consistently present across individual OMICs graphs

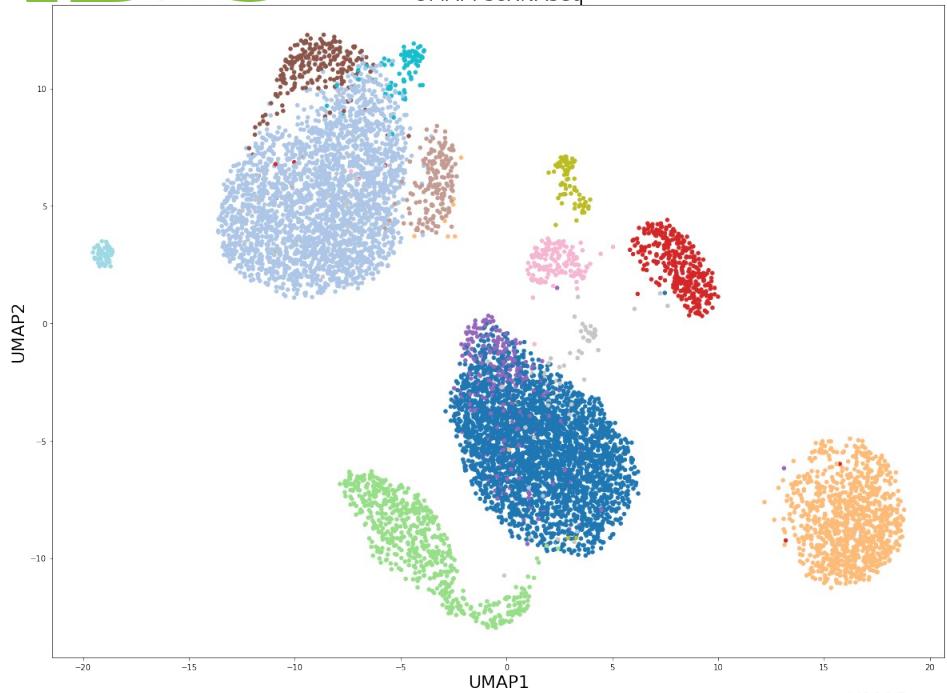
Graph Intersection

[DOWNLOAD](#)
Wolfram Notebook

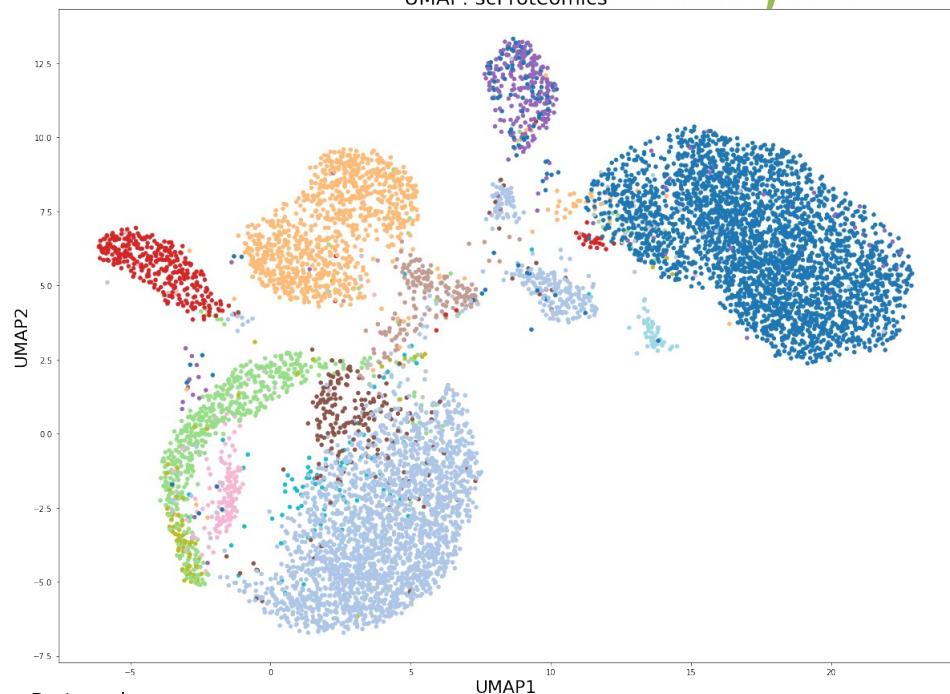


Let S be a set and $F = \{S_1, \dots, S_p\}$ a nonempty family of distinct nonempty subsets of S whose union is $\bigcup_{i=1}^p S_i = S$. The intersection graph of F is denoted $\Omega(F)$ and defined by $V(\Omega(F)) = F$, with S_i and S_j adjacent whenever $i \neq j$ and $S_i \cap S_j \neq \emptyset$. Then a graph G is an intersection graph on S if there exists a family F of subsets for which G and $\Omega(F)$ are isomorphic graphs (Harary 1994, p. 19). Graph intersections can be computed in the Wolfram Language using `GraphIntersection[g, h]`.

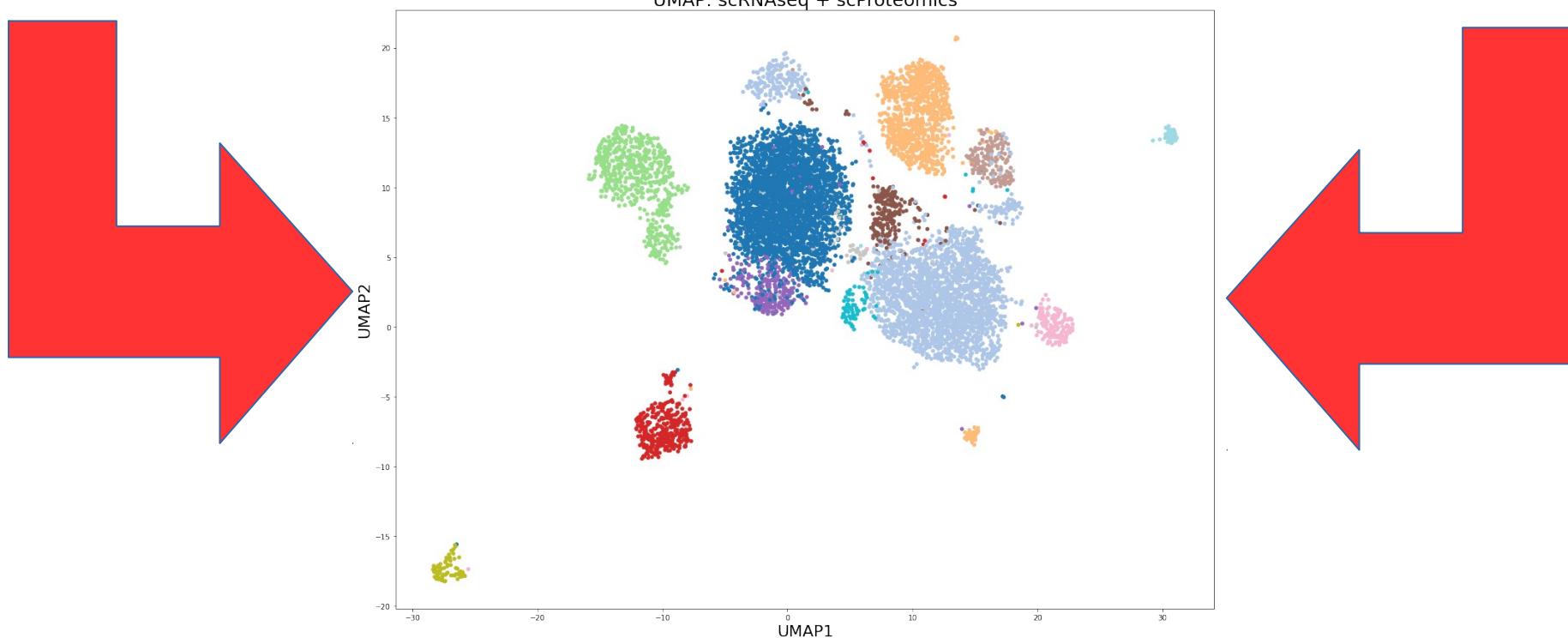
UMAP: scRNAseq



UMAP: scProteomics



UMAP: scRNAseq + scProteomics





National Bioinformatics Infrastructure Sweden (NBIS)

SciLifeLab



*Knut och Alice
Wallenbergs
Stiftelse*



LUNDS
UNIVERSITET