**Project Report**

**BF550: Programming in Python, Fall 2022**

**Pooja Paresh Savla**

Research Paper: A generalization of t-SNE and UMAP to single-cell multimodal omics DOI: <https://doi.org/10.1186/s13059-021-02356-5>

**Introduction:**

The paper that I selected to replicate images is titled: “A generalization of t-SNE and UMAP to single-cell multimodal omics”. Emerging single-cell technologies profile multiple types of molecules within individual cells. A fundamental step in the analysis of the produced high dimensional data is their visualization using dimensionality reduction techniques such as tSNE and UMAP. The motivation behind this paper is to introduce j-SNE and j-UMAP as their natural generalizations to the joint visualization of multimodal omics data. However, in our figures we will just use the UMAP technique to model our visualizations.

**Data Analysis:**

UMAP Principle:

UMAP (Uniform Manifold Approximation and Projection) is a novel manifold learning technique for dimension reduction. The algorithm is founded on three assumptions about the data:

1. The data is uniformly distributed on a Riemannian manifold;
2. The Riemannian metric is locally constant (or can be approximated as such);
3. The manifold is locally connected

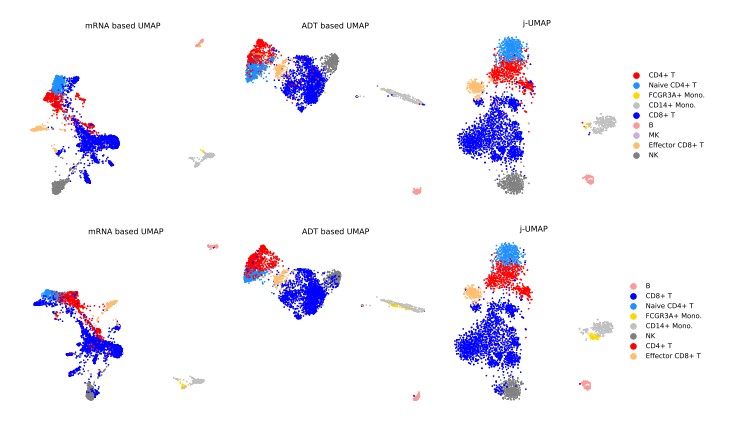
In this paper they have also introduced a modified version of UMAP called j-UMAP available in the JVis Python package as their natural generalizations to the joint visualization of multimodal omics data. This approach automatically learns the relative contribution of each modality to a concise representation of cellular identity that promotes discriminative features but suppresses noise.

The result is a practical scalable algorithm that applies to real world data. The UMAP and JUMAP algorithms are competitive with t-SNE and J-tSNE for visualization quality, and arguably preserve more of the global structure with superior run time performance. Furthermore, UMAP/ J-UMAP have no computational restrictions on embedding dimension, making it viable as a general-purpose dimension reduction technique for machine learning.

I recreated the first four clusters from the supplementary images from Figure 14 using UMAP clustering algorithm. In the figure S14, the UMAP/j-UMAP visualization was done of the Human Peripheral Blood Mononuclear cells (PBM cells). There were two cluster labels identified by Specter shown on the top rows of this figure and CiteFuse shown on the bottom row of this figure. The embeddings were computed in three different ways.

* RNA measurements alone (left side of S14)
* Protein expression (ADT) alone (middle side of S14)

Chart

Description automatically generated with medium confidence

**Methodology**

The thought process for clustering used in the project was as follows:

Specter: adopts and extends recent algorithmic advances in (fast) spectral clustering. In contrast to methods that cluster a (random) subsample of the data, Specter adopts the idea of landmarks that are used to create a sparse representation of the full data from which a spectral embedding can then be computed in linear time. CiteFuse: is a streamlined package consisting of a suite of tools for doublet detection, modality integration, clustering, differential RNA and protein expression analysis, antibody-derived tag evaluation, ligand– receptor interaction analysis and interactive web-based visualization of CITE-seq data. Both Specter and CiteFuse are clustering algorithms; they have provided pre-processed data to use where the author has already used these algorithms.

1. We used JVis package to compute a joint embedding of accessible chromatin and gene expression measured simultaneously by SNARE-seq in 1047 single cells from cultured human cell lines BJ, H1, K562, and GM12878.(The SNARE-seq and CBMC CITE-seq data sets were downloaded from Gene Expression Omnibus with accession codes GSE126074 and GSE126310, respectively).
2. Part A) For the figures A and D, we used an rna\_matrix that we got from the data/pbmc\_rna\_pca.csv path. Part B) For the figures B and E, we used an adt\_matrix that we got from the data/pbmc\_adt\_pca.csv path.
3. I found that clustering method with the Euclidean distance measurement produced the most similar UMAP to the one produced in the paper.
4. The neighborhood graph was contructed with n=30 neighbors, using all the principal components using the nearest neighbors (n\_neighbors = 30). This is the data upon which the UMAP is generated.
5. As you can see in the code I incorporated a color map to identify different cell types and visually color them to understand the UMAP clearly. I plotted a scatter plot of the dataframe created after dimensionality reduction (UMAP) in the notebook for each Specter and Citefuse
6. For both Part A, Part B we used the methods of Specter and Citefuse found in the path of 'data/pbmc\_jointspecter\_labels.csv' and 'data/pbmc\_citefuse\_labels.csv' respectively. This was done because the cell type associated with Specter and Citefuse are different.
7. The UMAP was generated and stored in a dataframe. Using the columns in this dataframe, a scatter plot was created visually displaying the clusters of Human Peripheral Blood Mononuclear cells (PBM cells).

**Figures**

**Part A) S14: Figure A - D**

Map

Description automatically generated

Map

Description automatically generated

**Results and Conclusions:**

In these figures it can be concluded that there is a high density of clusters of CD8+ T. The FCGR3A+ Mono and CD14+ Mono are an island and are away from the other cell\_types. Perhaps, we can do a dendogram and find more significance because they are not related to the other cell\_types by method of Specter or Citefuse.

Higher expected levels of noise in the measurements can be counteracted by smaller regularization coefficients λ that allow to downweight noisy modalities. Not surprisingly, projecting RNA and protein velocities into the joint embedding of both modalities yielded less noisy acceleration landscapes compared to embeddings of mRNA measurements alone. This projection serves as a useful analogy to understand how UMAP prioritizes global vs local structure depending on its parameters because I used both n\_NN as 5,10, 15 and then 30 to get the final plot as shown in paper. As n\_neighbors increases, UMAP connects more and more neighboring points when constructing the graph representation of the high-dimensional data, which leads to a projection that more accurately reflects the global structure of the data. At very low values, any notion of global structure is almost completely lost. As the min\_dist parameter increases, UMAP tends to "spread out" the projected points, leading to decreased clustering of the data and less emphasis on global structure. After using n\_NN I was able to get a graph that was represented exactly like the one shown in the research paper.

References:

1. PMID: 33941244, Do et al, A generalization of t-SNE and UMAP to single-cell multimodal omics (2021) DOI: [10.1186/s13059-021-02356-5](https://doi.org/10.1186/s13059-021-02356-5)
2. <https://umap-learn.readthedocs.io/en/latest/parameters.html>
3. <https://pubmed.ncbi.nlm.nih.gov/34963591/>(The application of Uniform Manifold Approximation and Projection (UMAP) for unconstrained ordination and

classification of biological indicators in aquatic ecology)

1. What is SNARE sequencing: single-nucleus chromatin accessibility and mRNA expression sequencing: a method that can link a cell’s transcriptome with its accessible chromatin for sequencing at scale [(High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell | Nature Biotechnology)](https://www.nature.com/articles/s41587-019-0290-0#:~:text=Here%2C%20we%20describe%20droplet-based%20single-nucleus%20chromatin%20accessibility%20and,with%20its%20accessible%20chromatin%20for%20sequencing%20at%20scale.)
2. Zenodo Link to the Data: [A generalization of t-SNE and UMAP to single-cell multimodal omics | Zenodo](https://zenodo.org/record/4682805#.Y5fhMFHMJEY)