MTH312 Project 4 Group 8

Team Members - Devansh Gupta (220345) Jiyanshu Dhaka (220481)

Arnab Das (231080020) Shivam Chaurasia (231080082)

1 Introduction

Studying individual cells using multiple types of data, like RNA, proteins, and chromatin accessibility, gives us a more complete understanding of cellular diversity. Instead of analyzing each type of data separately, combining them helps reveal deeper insights. However, since different types of data are available for different cells, it becomes challenging to study them together. Our method addresses this by mapping these different inputs into the same hidden space and grouping the cells into eight clusters.

2 Data Description

- RNA expression data (`rna_data`): 10,000 cells $\times$ 2,000 features
- Protein expression data (`protein_data`): 10,000 cells $\times$ 2,000 features
- Cell type labels for validation purposes

3 Methodology

We drew analogy from a different problem. Let's say we are given a dataset of English and French sentences, these sentences may not be the same, and we need to find words that have similar semantics. Then, we have the problem of mapping words from different languages into the same latent space. This gives us an approach to embed words that have different inputs, which is usually done using autoencoders. Hence, we use dual-branch autoencoder where the RNA and protein data are encoded separately but forced to align in a shared latent space. This is similar to a dual-encoder model for cross-lingual embeddings.

3.1 Autoencoder Architecture

- Two encoders (one for RNA, one for protein) compress their respective inputs into a shared latent space.
- A shared latent representation ensures that correlated RNA and protein data points align in the same space.
- Two decoders reconstruct RNA and protein from the shared representation.
- Reconstruction Loss: Ensures encoders learn meaningful representations.
- Alignment Loss: Ensures correlated RNA and protein embeddings are close.

Once we get the final embeddings, we use spectral clustering with number of clusters as 8 to obtain the clusters labels of each cell.

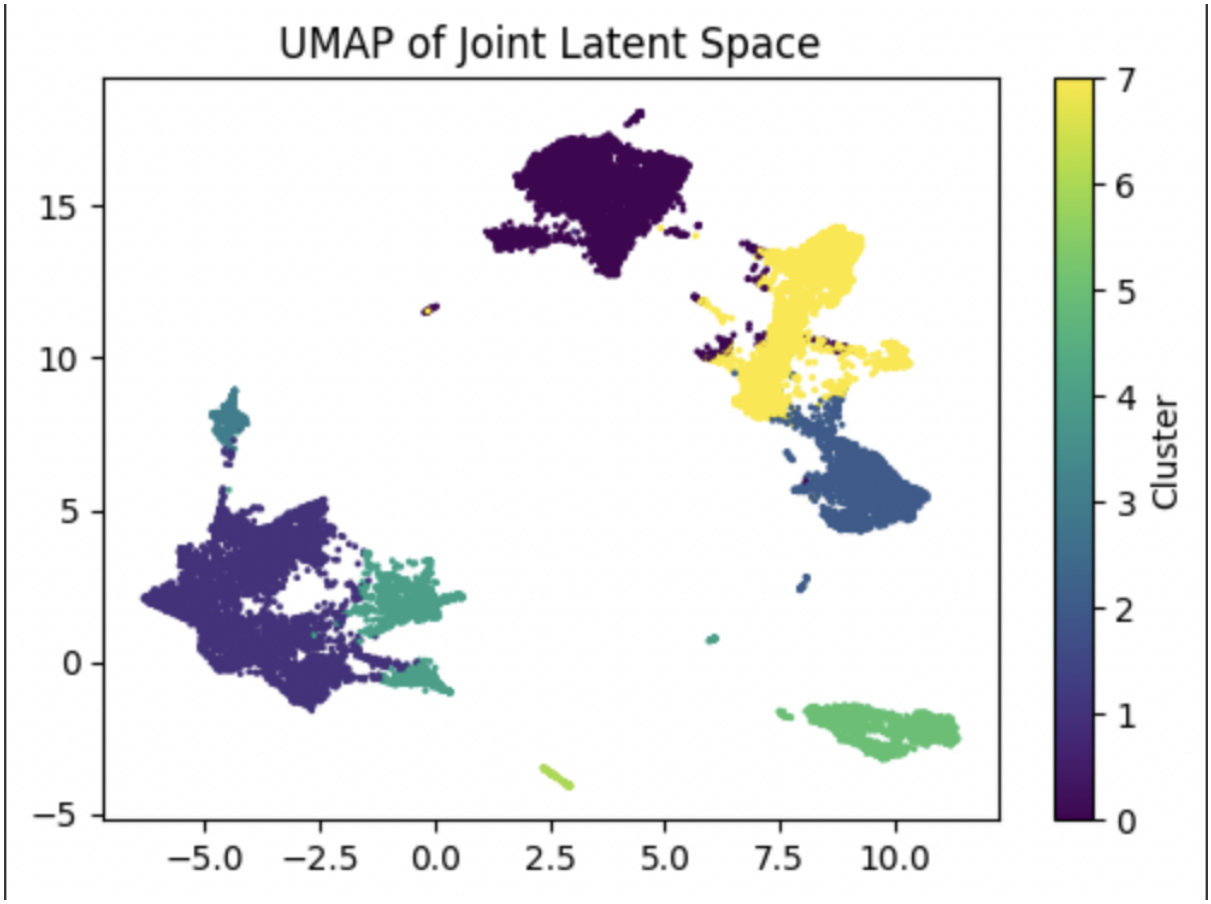

Figure 1: UMAP of Joint Latent Space

3.2 UMAP Visualization

UMAP revealed distinct clusters with overlapping RNA-protein regions, indicating successful integration while preserving modality-specific variations.

The UMAP visualization of the integrated data revealed:

- Clear separation between certain biological groups
- Regions of overlap between RNA and protein data points from the same biological origin
- Distinct cluster formations suggesting meaningful biological groupings

4 Experimentation

4.1 PCA and CCA with Leiden Clustering (Score: 0.49)

We performed dimensionality reduction on both datasets independently using PCA. Then, we applied CCA to maximize correlation between PCA components of both datasets.
Let Z_X and Z_Y be PCA-transformed representations of both datasets. We computed projection matrices A and B such that

$$U = Z_X A, \quad V = Z_Y B$$

where correlation between U and V is maximized.

We concatenated U and V to form joint latent space. We constructed k-nearest neighbor graph from joint data and applied Leiden clustering to assign clusters. We performed CCA

on PCA-transformed data to align both datasets in common latent space. We concatenated latent representations from CCA to form unified dataset. We constructed neighborhood graph using joint latent space and applied Leiden clustering to identify groups of cells with similar expression profiles.

4.2 PCA with Fuzzy Nearest-Neighbor Smoothing (Score 0.58)

We converted RNA and protein datasets into AnnData objects, then applied normalization and log transformation to standardize values. Similar to the previous method, we performed PCA separately on both datasets.Then to reduce noise, Fuzzy nearest-neighbor smoothing was applied by computing pairwise distances and averaging features of nearest neighbors. For Initial cell matching we used cost matrix based on Euclidean distances. then rest we did similar to previous method like CCA for iterative refinement, etc. After obtaining joint embedding by concatenation we applied KMeans clustering. this method gave score of 0.58.

4.3 MaxFuse

MaxFuse integrates multi-modal data using optimal transport, feature alignment, and graph-based techniques. We use the maxfuse library to pre-process and then extract the final embeddings as shown in the paper [1]. Once we obtained the final embeddings we used K-Means clustering, h-clustering, Spectral-clustering and Gaussian Mixture Models to obtain clusters.

Type of clustering	Accuracy
K-Means	0.583
h-Clustering	0.744
GMM	0.734
Spectral Clustering	0.830

Table 1: Accuracy of different clustering methods with MaxFuse embeddings

Implementation Steps

- **Pre-processing**: Normalize, log-transform, select variable features, and scale data.
- **Feature Processing**: Extract shared features via fuzzy matching, select active features by variability.
- **Graph Construction**: Compute neighborhood graphs, construct cost matrices, apply SVD.
- **Fusion Process**: Identify initial pivots via CCA, refine iteratively, generate integrated embeddings.

5 Conclusion

In this work, we experimented with different embedding and clustering methods and finally developed a dual-autoencoder model to map RNA and protein data into a shared latent space, enabling effective clustering. Our approach was inspired by cross-lingual word similarity, where different representations (RNA and protein) are aligned in a common space despite differences in structure and dimensionality.

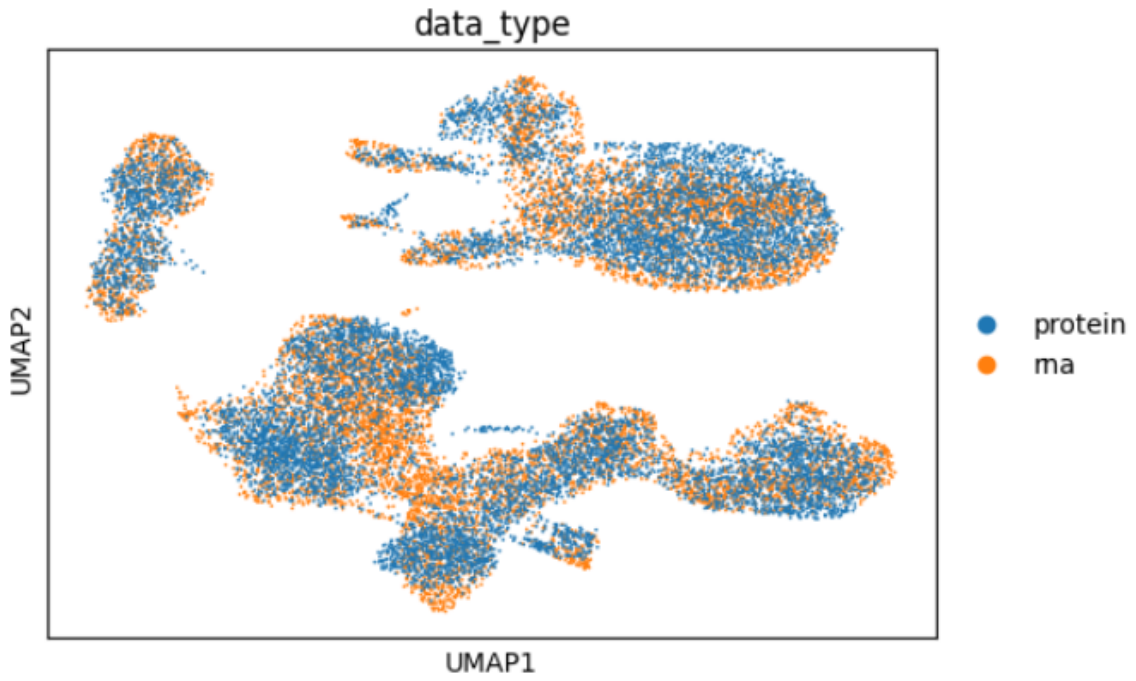

Figure 2: UMAP Components Plot

References

[1] Shuxiao Chen, Bokai Zhu, Sijia Huang, John W. Hickey, Kevin Z. Lin, Michael Snyder, William J. Greenleaf, Garry P. Nolan, Nancy R. Zhang, and Zongming Ma. Integration of spatial and single-cell data across modalities with weakly linked features. *Nature Biotechnology*, 42:1096–1106, 2024.

[2] Hongru Zong and Jianjun Liu. Dual-branch autoencoder with clustering information for hyperspectral blind unmixing. In *IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium*, pages 9139–9142, 2024.