Exploring the latent space learned by scRNA-seq foundation models to identify AD subtypes

Isak Bieltvedt Jonsson
ijonsson@student.tudelft.nl
Supervisors: T. Verlaan[1] & R. Lardenoije[1]
[1] Pattern Recognition and Bioinformatics, TU Delft
June 23, 2025

1 Background

Alzheimers Disease (AD)

- Heterogeneous - manifestations, progressions, and underlying pathologies can vary between patients [2]
- Understanding of causes and effects of these variations is limited

Techniques

Single Cell RNA Sequencing (scRNA-seq or transcriptomics) measures messenger RNA levels in tissue at the cell level to quantify gene activity and provide a snapshot of cell state. Data is high dimensional, sparse, and noisy.
Differential Gene Expression (DGE) is a statistical method to find genes who's expression differ significantly between groups (DEGs). These may also be referred to as "markers" or "marker genes".

Related Work

Clustering scRNA-seq data has revealed multiple disease associated subgroups of cells with distinct transcriptional signatures [3].
Attempts to discover sample (individual) level subtypes have clustered samples directly [4] or used cell-subgroup proportions as sample representations for disease trajectory analysis [2].
Foundation Models like Geneformer [1] (GF) learn dense cell representations through self-supervised pretraining on massive unlabeled scRNA-seq data. GF cell embeddings demonstrate SOTA performance for various downstream tasks but their applications to disease subtype identification are underexplored.

Dataset - ROSMAP

We use transcriptomic data from the Religious Orders Study and Memory and Aging Project. Contains samples from prefrontal cortex of 426 individuals with varying levels of cognitive decline. 2.5M cells x 30k genes.

2 Research Question

"To what extent does the latent space learned by self-supervised scRNA-seq foundation models enable the discrimination and charecterization of AD subtypes?"

References

[1] H. Chen et al. "Geneformer: Transfer learning enables predictions in network biology". In: *Nature* 617.7961 (2023), pp. 616–622.
[2] D. Ferreira et al. "Distinct subtypes of Alzheimer's disease based on patterns of brain atrophy: longitudinal trajectories and clinical applications". In: *Scientific Reports* 10.1 (2020), p. 8731.
[3] H. Mathys et al. "Single-cell transcriptomic analysis of Alzheimer's disease". In: *Nature* 570.7761 (2019), pp. 332–337.
[4] R. A. Neff et al. "Molecular subtyping of Alzheimer's disease using RNA sequencing data reveals novel mechanisms and targets". In: *Science Translational Medicine* 13.600 (2021), eabb5398.
[5] T. Verlaan et al. "scAGG: Sample-level embedding and classification of Alzheimer's disease from single-nucleus data". In: *bioRxiv* (2025). Preprint.

3 Results

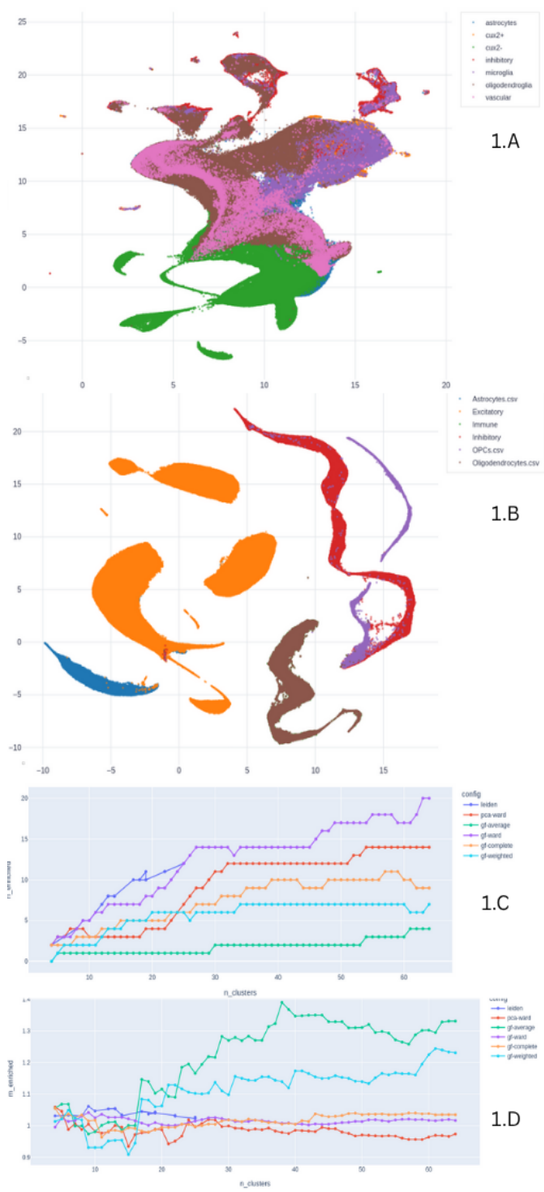

Figure 1: Embedding spaces and clustering on Immune Cells. Embeddings are generated with Geneformer [1] and by applying PCA. Enrichment for AD is tested using fisher's exact test and benjamini-hochberg correction.
1.A: UMAP of PCA embeddings **1.B:** UMAP of GF embeddings.**1.C:** Nr. of AD enriched clusters vs method **1.D:** Mean fold enrichment vs method

Figure 2: Astrocyte cluster DEGs filtered for known AD markers. DGE on clusters should yield markers characteristic of known AD associated astrocyte subgroups
2.A, 2.C: GF - 39 & 52 clusters **2.B, 2.D:** PCA - 35 & 62 clusters

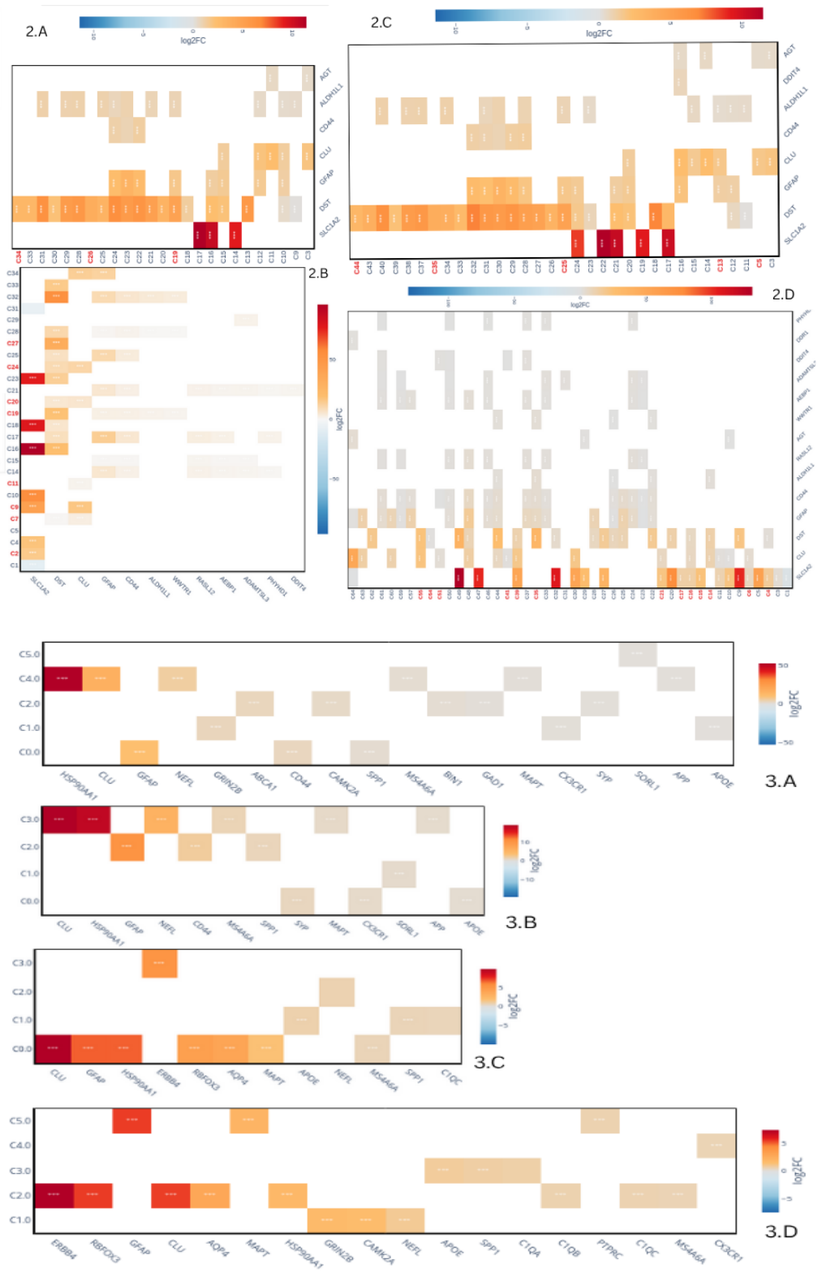

Figure 3: AD subgroup DEGs filtered for known AD markers. Sample representations are constructed from GF cluster proportions and ROSMAP cell-subtype proprotions (Baseline). Initial exploration of all samples revealed low seperation between AD and non AD proportions. Proportions are filtered to only include AD samples, clustering applied, and DGE performed vs other AD clusters. Meaningful subtypes should yield markers known to be associated with AD.
3.A, 3.B: ROSMAP cell-subtype proportions. 4 & 6 subgroups.
3.C, 3.D: GF cluster proportions. 6 & 4 subgroups.

4 Conclusions

GF embeddings show stronger cell type seperation than PCA after applying UMAP (F:1.A,1.B)
Clustering GF embeddings yields more AD enriched clusters and slightly higher mean fold enrichment levels than PCA (F:1.C,1.D). Fold enrichment levels of 1.0-1.3 suggests low separation between AD and non AD cells.
DGE on astrocyte clusters does not identify clusters corresponding to known AD associated subgroups for GF or PCA embeddings. We especially look for upregulation of CD44, GFAP, SPP1, LCN2, C3, & CLU and downregulation of SLC1A2 & ALDH1L1. GF clusters are enriched for a partial set but lack presense of SPP1, LCN2, & C3. SLC1A2 & ALDH1L1 are enriched but upregulated (F:2.A,2.C). PCA yields stronger results with enrichment of more 'secondary' genes and downregulation of SLC1A2 & ALDH1L1 but key genes are still unenriched (F:2.B,2.D).
DGE on AD sample proportion clusters reveals transcriptional differences between clusters known to be associated with AD. Clusters are well separated with distinct DEGs. GF cluster proportions seem to outperform ROSMAP cell subtype proportions with higher logFC scores but further investigation is required.

In conclusion, we find that the latent space learned by Geneformer does not seperate AD cells from controls, that applying clustering directly to cell embeddings does not yield known AD subgroups, but that GF cluster proportions may have potential for downstream applications. Identifying AD subtypes from the latent space learned by Foundation Models could not be accomplished through straightforward methods and further research is required.

5 Limitations & Future Work

Further analysis of the identification of known AD associated subtypes of cells is required. More cell types should be considered.
DGE at sample level could only be performed for a subset of cells. Our results are indicative but not exhaustive

Direct Sample Embeddings could be generated using graph based sample embedding methods [5] adapted with contrastive graph pooling objectives.
Multi-Omics Approaches could integrate other omic subgroup proportions into sample level analysis.
Sample Level Clustering implementations could be developed, published, optimized for large datasets, and applied.
Disease Specific Fine Tuning may yield a latent space better associated with disease subtypes.
Disease Trajectory Analysis (or other advanced methods) could be applied. The latent space could lend itself to such a spatial analysis.