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

Isak Bieltvedt Jonsson[1]

Supervisors: Timo Verlaan[2]**, Roy Lardenoije**[2]

[1]**EEMCS, Delft University of Technology, The Netherlands**
[2]**Pattern Recognition And Bioinformatics, Delft University of Technology, The Netherlands**

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 22, 2025

Name of the student: Isak Bieltvedt Jonsson
Final project course: CSE3000 Research Project
Thesis committee: Marcel Reindeers, Timo Verlaan, Roy Lardenoije

Abstract

1 Introduction

Alzheimer's Disease (AD) is an infamous neurodegenerative disorder affecting over 55 million people worldwide [1]. Despite decades of research its exact causes remain unknown.

A critical challenge in AD research is its striking heterogeneity - AD presents on a spectrum with varied clinical manifestations, progression rates, and underlying pathologies. Unravelling this complexity by identifying and analysing AD subgroups could prove a crucial step in understanding, preventing, and treating the varied pathologies of AD.

1.1 Background

Alzheimer's Disease is characterized by the abnormal accumulation of amyloid-beta Aβ plaque and tau protein neurofibrillary tangles in the brain. These proteins interfere with neuron communication, cause inflammation, and contribute to atrophy of brain regions responsible for memory. AD is confirmed posthumously or with neuroimaging by measuring $\alpha\beta$-plaque and tau levels.

Gene expression is the process by which information from a gene is used to produce functional products like proteins. Your cells are genetically identical, but function differently due to variations in gene expression.

Transcriptomics measure messenger RNA (mRNA) levels in tissue samples. mRNA carries instructions from the genome to the ribosome where it is used to make proteins; its abundance directly reflects gene expression levels. Transcriptomics can be applied at either the bulk (large piece of tissue measured and mRNA counts averaged) or single cell level.

Single Cell RNA sequencing (scRNA-seq) measures gene expression across thousands of single cells. It gives us a snapshot of the individual cell states of complex biological systems. scRNA-seq data is very high dimensional and sparse, it is also noisy, with high dropout rates and technical artifacts that can obscure biological signals.

Foundation models learn meaningful representations from complex high dimensional data through pretraining on massive unlabeled datasets with self supervised learning objectives such as masked token prediction.

Differential Gene Expression (DGE), Over Representation Analysis, & GSEA are statistical methods that can be used to find significant transcriptomic differences between groups. DGE finds genes who's expression differs significantly between groups. ORA and GSEA are both used to test whether specific gene sets or biological pathways are significantly enriched between groups. ORA uses a list of significant Differentially Expressed Genes (DEGs) to identify over-represented gene sets while GSEA finds pathway enrichment directly from the full set of genes.

1.2 Related Work

Characterizing AD using scRNA-seq data requires tackling a twofold dimensionality reduction problem. The data matrix X $(n_{cells} \times n_{genes})$ is high dimensional and noisy. Gene level complexity must be reduced to create denser cell features while accounting for noise. Sample level analysis requires cellular dimensions to be compressed or aggregated to form meaningful sample representations.

Seminal work by used scRNA-seq to identify disease associated subgroups of Microglia and Astrocytes [2] [3]. Graph based clustering was applied to find cell-subgroups and DGE & GSEA used to find transcriptional differences. These studies reveal Disease Associated Microglia (DAM) and Disease Associated Astrocytes (DAA) along with key enriched genes and pathways. Further trajectory analysis characterizes the progression of Astrocyte cell states from GFAP low to GFAP high to DAA as Alzheimer's pathology progresses. Later studies expand on their work and identify or refine further AD related cell subgroups through scRNA-seq clustering [4] [5].

Cell subgroups have provided key insights for understanding the development and mechanisms of AD pathology but do not directly address the challenge of subtyping patients at the sample level. Samples are composed of thousands of individual cells and constructing meaningful sample representations from single cell data with minimal loss of information remains an open problem. Various methods to tackle this problem have been proposed in the literature.

Studies address this through the use of Weighted SNP correlation matrix analysis (WSCNA) [6]. Sample-Sample correlation is used to construct a topological overlap matrix and then clustering applied to directly find subgroups of samples. A major drawback of this method is the $O(n_{cells}^2)$ space complexity which hinders analysis of larger datasets.

Another approach clusters scRNA-seq data and uses cell-subgroup proportions as sample representations for downstream trajectory analysis. They model cell state transitions over pseudo-time to distinguish between three distinct AD trajectories [7].

More recently, foundational transformer models, like Geneformer, have emerged as powerful tools for learning meaningful representations from complex high-dimensional single-cell data [8]. Geneformer, pretrained on millions of scRNA-seq transcriptomes, captures intricate gene network dynamics and generates context-aware cell embeddings. Cell representations learned by geneformer demonstrate SOTA performance on various downstream tasks, including cell type annotation, batch integration, in silico perturbation analysis, and disease classification, even in zero-shot or limited fine-tuning scenarios.

A promising new approach explores the use of graph based neural networks and attention pooling to generate sample embeddings directly. They train a Graph Neural Network to generate cell embeddings, a sample pooling mechanism, and a sample level AD classifier simultaneously to learn a latent space with an ordering corresponding to disease severity [9]. Similar methods could be applied with self-supervised objectives but the lack of available GNN foundation model weights would necessitate pretraining from scratch.

1.3 Research Question & Hypotheses

Transcriptomic Foundation models have been shown to learn powerful representations for various downstream tasks but

the investigation of their usage for AD subgroup discovery remains underexplored. To address this knowledge gap we pose our central research question:

"To what extent do the latent representations learned by self supervised scRNA-seq foundation models enable the discrimination and characterization of AD subtypes?"

We split our investigation into two parts and develop concrete hypotheses to test.

Cell subgroup identification with clustering
H1.1 - "Clustering on self-supervised cell embeddings will yield at least one cluster significantly enriched for cells derived from Alzheimer's Disease samples (adjusted p ¡ 0.05, Fisher's exact test with Benjamini-Hochberg correction)"
H1.10 - "Clustering on self-supervised cell embeddings will not yield any cluster significantly enriched for cells derived from Alzheimer's Disease samples (adjusted p ¡ 0.05, Fisher's exact test with Benjamini-Hochberg correction)"
Motivation - Clusters significantly enriched for AD confirm the foundation model's latent space separates disease-specific transcriptional patterns from healthy controls and other sources of cognitive decline.

H1.2 - "Clustering finds Astrocyte subgroups with DEG's corresponding to key markers from DAA, GFAP low, and GFAP high"
H1.20 - "DGE on Astrocyte subgroups does not find significant markers associated with DAA, GFAP low, or GFAP high at any 'resolution' or finds some partial set missing key genes"
Motivation - Revealing known AD associated astrocyte subgroups characterized by DAA, GFAP-low, and GFAP-high markers confirms the clustering procedures ability to identify known disease relevant astrocyte phenotypes from cell embeddings.

Sample level analysis of cell-subgroup proportions
H2.1 - "Clustering on sample proportions of cell subgroups will identify at least one cluster significantly enriched for Alzheimer's Disease samples (adjusted p ¡ 0.05, Fisher's exact test with Benjamini-Hochberg correction)"
H2.10 - "Clustering on sample proportions of cell subgroups will not yield any cluster significantly enriched for Alzheimer's Disease samples (adjusted p ¡ 0.05, Fisher's exact test with Benjamini-Hochberg correction)"
Motivation - Clusters enriched for AD samples based on cell subgroup proportions confirm that proportions discriminate between AD and healthy controls or other sources of cognitive decline.

After finding no AD enriched clusters on the whole proportion matrix we attempt subtype analysis with only AD samples.

H2.2 - "Clustering on only Alzheimer's Disease sample proportions will identify AD subgroups, whose differential gene expression (DGE) and over-representation analysis (ORA) will reveal transcriptional differences and enriched pathways known to be associated with Alzheimer's Disease pathophysiology (adjusted p ¡ 0.05)"
H2.20 - "Clustering on only Alzheimer's Disease sample proportions will not identify distinct AD subgroups whose DGE and ORA reveal transcriptional differences or enriched pathways known to be associated with Alzheimer's Disease pathophysiology (adjusted p ¡ 0.05)"
Motivation - Finding distinct AD subgroups from cell proportions with AD associated transcriptional differences implies that subtypes identified by clustering are biologically meaningfull

2 Results

We leverage self-supervised learning to generate rich embeddings from scRNA-seq data with Geneformer [ref]. ROSMAP data is split by cell type into Astrocytes, Immune Cells, Oligodendrocytes, OPCs, Excitatory neurons, and Inhibitiroy neurons. Clustering is applied to cell embeddings and clusters are tested for AD enrichment. DEG and ORA is applied to astrocyte clusters to test enrichment for a set of known DAA markers. Finally we represent each sample as a distribution of its cell-cluster counts and attempt to identify subtypes of samples.

2.1 Cell subgroup evaluation with clustering

Comparison of silhouette scores and cluster balance for different clustering methods reveals agglomerative ward clustering as the best overall performer. Inspecting other spatial metrics and cluster size balance confirms this further **??**.

Significantly AD enriched clusters for GF embeddings are identified at each tested resolution for every cell type. This indicates some separation between AD and non AD cells but low fold enrichment values of 1.1-1.3 suggest a weak association and require further investigation.

2.2 Identification of known Astrocyte AD subgroups

After clustering on Astrocytes we select a set of configurations with increasing n_clusters. DGE is performed on each configuration to test for enrichment of a set of known DAA marker genes. We especially look for upregulation of CD44, GFAP ,SPP1, LCN2, C3, & CLU and downregulation of SLC1A2, & ALDH1L1 as a strong indication of a DAA candidate cluster.

GF embedding configurations yield clusters enriched for 3 key DAA markers at n_clusters (13, 22, 39, 52) - GFAP, CLU, & CD44 - SLC1A2 & ALDH1l1 are enriched but upregulated . Higher configurations do not yield clusters enriched for more of our target set or with larger subsets of our targets enriched per cluster; higher n_clusters seem to split existing clusters (and keep the same markers) rather than identifying new ones. Clustering GF embeddings does not yield clusters corresponding to DAA.

PCA embedding clusters provide more promising DGE results but still only find a subset of our targets. Clusters are again enriched for key genes GFAP, CLU, & CD44 but none are upregulated for SPP1, LCN2, & C3. Clusters upregulated for GFAP, CLU, & CD44 are also generally enriched for a subset of other DAA supporting genes. ALDH1L1 & SLC1A2 are downregulated for multiple clusters. PCA with 35 clusters has some weak DAA candidates (14, 15, 17, 21,

32) but lacks key targets. PCA with 65 clusters shows even weaker candidates with less enrichment for supporting genes (22, 26, 35, 44, 48). Neither has clusters that can be confidently identified as DAAs.

AD enriched clusters are highlighted in red. Only one enriched clusters was identified as a DAA candidate and they were generally not enriched for more than one of our targets.

Clustering GF and PCA embeddings does not yield clusters strongly corresponding to DAAs based on analysis of marker genes. PCA outperforms GF and identifies clusters enriched for a larger subset of DAA targets.

We construct sample representations by using the proportions of their cells in each subgroup to get a proportion matrix P $(n_{samples} \times n_{cell-subgroups})$. A baseline comparison is established using cell-subgroups already present in the ROSMAP dataset. Clustering is applied to this matrix and clusters tested for AD enrichment. No clusters are found to be AD enriched suggesting low separation between AD and non AD samples in the proportion space. Umap plots confirm this low seperation for GF and baseline.

After finding no enriched clusters we filter P to include only AD samples and apply clustering to find AD subtypes. DGE and ORA are applied vs other AD subtypes and we test for enrichment for a known set of AD related pathways and marker genes. We especially look for enrichment for pathways from 'KEGG_ALZHEIMERS', 'Alzheimer's disease', 'GO_NEUROGENESIS', and 'Alzheimer's disease (WP5124)'.

DGE results show significant transcriptional differences between groups. GF subtypes are enriched for a subset of our targets with high log2FC scores and distinct seperation between enriched targets for each cluster. GF subgroup proportions outperform our baseline and show better transcriptional distinction between groups.

Enriched pathways do not indicate strong AD related pathway differences between subgroups **??**.

3 Discussion

3.1 Cell clustering

Clustering on geneformer embeddings yielded enriched clusters but enrichment levels around 1.1-1.3 mean these clusters are only 10-30% more likely to contain AD samples. GF clustering does not seperate AD and non AD samples, this suggests the latent space learned by geneformer is not well seperated between AD and non AD clusters.

Similar results for clustering on the pca baseline suggests that while geneformer embeddings are not well seperated, they are not worse than the baseline.

These results indicate that a simple clustering procedure is not sufficient to distinguish between healthy and AD cells.

3.2 Known AD associated Astrocytes subgroups

Astrocyte GF clusters do not identify clusters that are enriched for key DAA markers. Partial enrichment for some targets but a lack of key genes (SPP1, LCN2, C3) and weak seperation between cluster markers suggests our clusters are not associated with DAAs markers. This implies transcriptional differences between Astrocyte clusters are largely not associated with AD.

While PCA results are better, and may present some DAA candidates, they can not be said to identify DAAs.

These results indicate that the latent space learned by geneformer does not distinguish between AD associated astrocyte subgroups and healthy controls. In fact they seem to separate healthy and DA Astrocytes worse than a simple pca procedure.

3.3 Sample Proportion Subgroups

Raw sample proportions do not show any separation between AD and non AD samples and the lack of any enriched clusters for the baseline and GF shows they can not be used to distinguish between AD subtypes.

Filtering on AD samples only we see a significant subset of our target genes are enriched and that subgroups have good seperation between enriched sets of genes (i.e. two subgroups are generally not enriched for the same gene). Geneformer subgroups outperform the ROSMAP subgroup baseline despite weak performance in the identification of DAAs.

Pathway enrichement results do not show major enrichment for AD related pathways between subgroups. Since we only perform DGE on a subset of cell types this could be explained.

These results are conflicting but show some promise. Clustering on AD sample cell subgroup proportions does yield clusters with DEGs associated with AD and indicates a discovery of meaningful transcriptomic differences but further research and analysis is requried.

4 Conclusions, Limitations, and Future Work

4.1 Conclusion

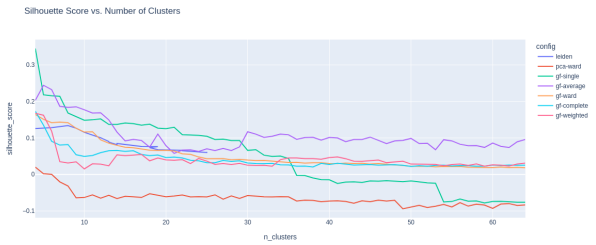
This study investigated the extent to which latent representation learned by scRNA-seq Foundation Models could be used to identify AD subtypes. Our analysis of geneformer embeddings on ROSMAP data reveals mixed results that challenge assumptions on foundation model superiority.

Geneformer embeddings show superior cell type seperation than PCA, but this advantages does not translate to better identification of AD enriched clusters with both methods acheiving similar numbers of enriched clusters weak enrichment levels.

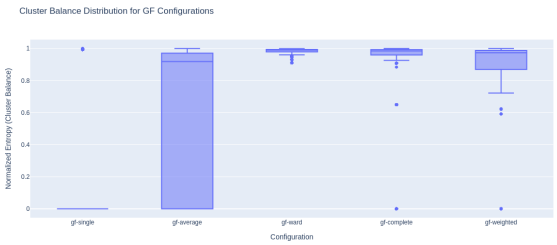
Neither geneformer nor PCA identifies clusters with markers corresponding to a known Disease Associated Astrocytes profile.

Analysis using cell subgroup proportions fails to identify AD-enriched sample clusters, indicating that this approach may not be sufficient for patient-level classification. When clustering only AD samples, both methods revealed AD associated transcriptional differences between subgroups and this method may be promising for further subtype analysis.

In conclusion, our research demonstrates that extracting AD subtype information from the latent space learned from foundation models requires careful consideration and potentially disease specific adaptations. The mixed results observed suggest that the path from general-purpose models to



(a) 1.A Spatial metrics for Immune Cells



(b) 1.B Cluster balance

Figure 1: Cluster balance and Silhouette scores for Immune Cells

disease-specific insights may be more complex than initially anticipated and that the latent space learned by geneformer cannot be easily used for AD subtyping.

4.2 Limitations

Analysis was limited to the ROSMAP cohort, which primarily consists of white female participants, potentially limiting generalizability. DGE for sample proportions was only performed for a subset of cell types and this may explain weak AD pathway enrichment results.

Our choice of clustering parameters and specific resolution parameters may have influenced subgroup identification

The DAA marker set used may not be universally applicable across all AD samples or stages. The full set of AD target genes and pathways is not exhaustive.

Our analysis cannot capture disease progression dynamics.

Practical DAIC jobs are limited to 1TB of RAM, this made the analysis of larger datasets (i.e. entire ROSMAP cohort) infeasible.

4.3 Future Work

Disease-specific fine-tuning: Investigate whether fine-tuning foundation models on AD-specific datasets improves disease-relevant pattern recognition

Sample level clustering approaches: Use sample level clustering to find subgroups of samples directly.

Multi-omic integration: Combine transcriptomic with other omics modalities for more comprehensive analysis

Further embedding space exploration: Embeddings for other cell types could be clustered to analyze discovery of other known AD related cell subgroups

Direct Sample Embeddings: The procedure from [9] could be combined a contrastive attention pooling goal to learn sample embeddings directly.

References

[1] World Health Organization. (2023). Accessed: 2025-04-24.

[2] Keren-Shaul, H., Spinrad, A., Weiner, A., Matcovitch-Natan, O., Dvir-Szternfeld, R., Ulland, T. K., David, E., Baruch, K., Lara-Astaiso, D., Toth, B., et al. *Cell* **169**(7), 1276–1290 (2017).

[3] Deczkowska, A., Keren-Shaul, H., Weiner, A., Colonna, M., Schwartz, M., and Amit, I. *Cell* **173**(5), 1073–1081 (2018).

[4] Mathys, H., Davila-Velderrain, J., Peng, Z., Gao, F., Mohammadi, S., Young, J. Z., Menon, M., He, L., Abdurrob, F., Jiang, X., Martorell, A. J., Ransohoff, R. M., Hafler, D. A., Bennett, D. A., Kellis, M., and Tsai, L.-H. *Nature* **570**(7761), 332–337 (2019).

[5] Zhou, Y., Song, W. M., Andhey, P. S., Swain, A., Levy, T., Miller, K. R., Poliani, P. L., Cominelli, M., Grover, S., Gilfillan, S., Cella, M., Uwanogho, D., Yan, Q., Petrus, S. G., Phung, A., Norwood, B., Zhang, A., Sindiri, S., Ficarra, E., Barta, K., Calabresi, P. A., Bennett, D. A., Goate, A. M., Artyomov, M. N., and Colonna, M. *Nature Medicine* **26**(1), 131–142 (2020).

[6] Neff, R. A., Wang, M., Vatansever, S., Guo, L., Ming, C., Wang, Q., Wang, E., Horgusluoglu-Moloch, E., Song, W.-m., Li, A., Castranio, E. L., Tcw, J., Ho, L., Goate, A., Fossati, V., Noggle, S., Gandy, S., Ehrlich, M. E., Katsel, P., Schadt, E., Cai, D., Brennand, K. J., Haroutunian, V., and Zhang, B. *Science Advances* **7**(2), eabb5398 (2021).

[7] Ferreira, D., Nordberg, A., and Westman, E. *Scientific Reports* **10**(1), 8731 (2020).

[8] Chen, H., Venkatesh, M. S., Gomez Ortega, J., Mahesh, S. V., Nandi, T., Madduri, R., Pelka, K., and Theodoris, C. V. *Nature* **617**(7961), 616–622 (2023).

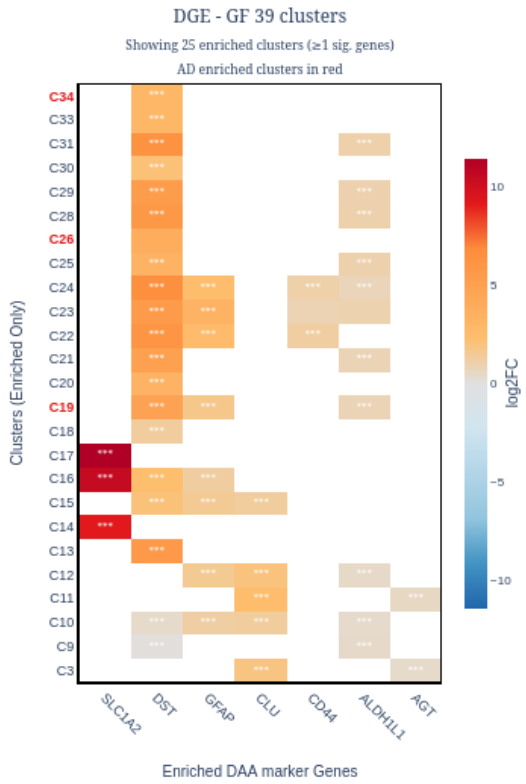
[9] Verlaan, T., Bouland, G., Mahfouz, A., and Reinders, M. J. T. *bioRxiv* (2025). Preprint.

[10] Kunkle, B. W., Grenier-Boley, B., Sims, R., Bis, J. C., Damotte, V., Naj, A. C., Boland, A., Vronskaya, M., van der Lee, S. J., Amlie-Wolf, A., et al. *Nature Genetics* **51**(3), 414–430 (2019).

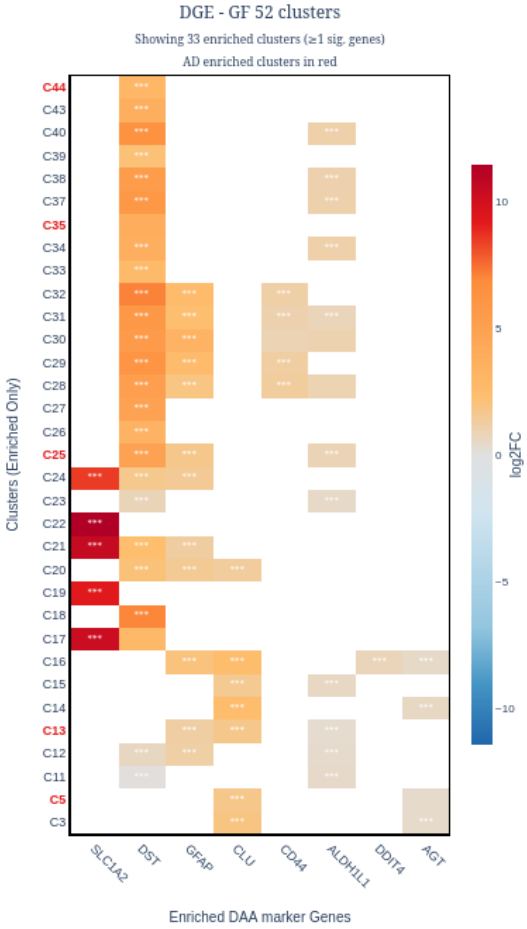
[11] Cui, C. V., Zhang, J., Blanco-Míguez, A., Bhatti, A., Zou, J., Gehman, S., and Theodoris, C. V. *Nature* **626**(7999), 620–628 (2024).

[12] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. *arXiv preprint arXiv:1810.04805* (2018).

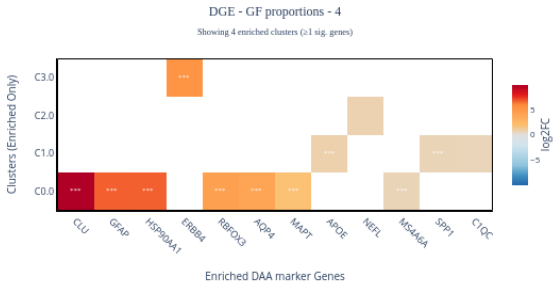
[13] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma,



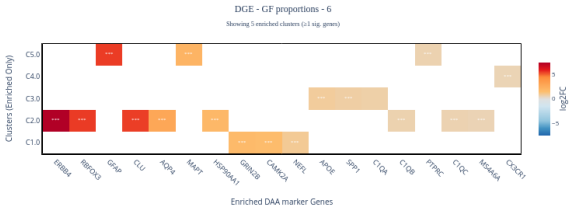
(a) GF Astrocyte DGE results 39 clusters



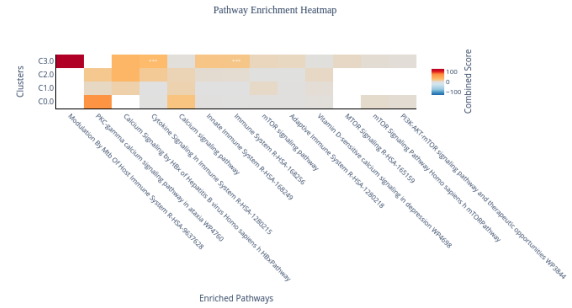
(b) GF Astrocyte DGE results 52 clusters



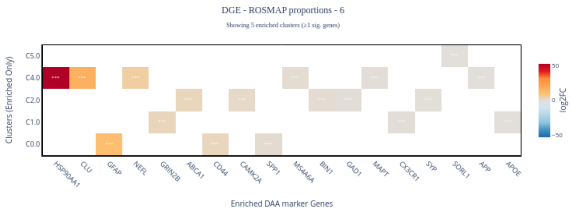
(a) DEGs GF AD subtypes - 4



(b) DEGs GF AD subtypes - 4



(c) DEGs ROSMAP state AD subtypes - 4



(d) DEGs ROSMAP state AD subtypes - 6

Figure 3: DGE results for AD subtypes

C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* , 38–45 (2020).

[14] Wolf, F. A., Angerer, P., and Theis, F. J. *Genome Biology* **19**(1), 15 (2018).

[15] Rousseeuw, P. J. *Journal of Computational and Applied Mathematics* **20**, 53–65 (1987).

[16] Caliński, T. and Harabasz, J. *Communications in Statistics* **3**(1), 1–27 (1974).

[17] Davies, D. L. and Bouldin, D. W. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **PAMI-1**(2), 224–227 (1979).

[18] Bennett, D. A., Schneider, J. A., Arvanitakis, Z., and Wilson, R. S. *Current Alzheimer Research* **9**(6), 628–645 (2012).

5 Materials And Methods

5.1 Data Sources and Preprocessing

We utilized data from the Religious Orders Study and Memory and Aging Project (ROSMAP) [**?**; **?**], accessed through the AD Knowledge Portal on Synapse [10]. ROSMAP represents one of the largest longitudinal cohort studies of aging and dementia, encompassing clinical assessments, cognitive evaluations, and comprehensive neuropathological characterization of participants. The multi-omic datasets include genomics, transcriptomics, and epigenomics data generated primarily from postmortem dorsolateral prefrontal cortex tissue samples.

Single-cell RNA sequencing data were obtained from *426* individuals, including both cognitively normal controls and individuals with varying degrees of cognitive impairment. All data access and usage adhered to the appropriate data use agreements established through the AD Knowledge Portal (https://adknowledgeportal.synapse.org).

5.2 Cell Embedding Generation

To generate our cell embeddings we employed two methods, a transformer based self-supervised approach and a principal component analysis (PCA) baseline. scRNA-seq data is embedded in a 512 dimensional space and SS and PCA embeddings are then compared in later steps of our analysis pipeline. The PCA baseline comparison allows us to assess whether self-supervised embeddings improve upon capture of latent AD subtypes compared to traditional linear dimensionality reduction.

Self-Supervised Learning Approach

Cell embeddings were generated using Geneformer [11], a transformer-based foundation model pre-trained on a diverse atlas of 95 million human cells. Geneformer employs a BERT-like architecture [12] with a masked token prediction self-supervised learning objective, allowing it to learn rich cellular representations from unlabeled scRNA-seq data.

We utilized the 12-layer Geneformer model (95M parameters) available through Hugging Face [13]. Input preprocessing followed the recommended protocol: expressed genes were converted to ensembl ID's and raw RNA expression counts fed into a tokenizer that generated rank value encodings by ranking genes by their relative expression and normalizing by their expression across the entire pre-training atlas.

PCA Baseline Comparison

Alternative cell embeddings were generated using PCA. scRNA-seq data was preprocessed by filtering out cells with less than 200 genes and genes expressed in less than 3 cells. It was then normalized, log1p transformed, and the 5000 most highly variable genes selected. Finally we reduce dimensionality to 512 with PCA.

All preprocessing and PCA was performed using `scanpy` [14].

5.3 Hierarchical Cell Embedding Clustering

Cell embeddings were Hierarchically clustered using the agglomerative clustering implementation from `scipy`. Various configurations were tested:

- Linkage Methods - Ward, Weighted, Average, Single
- Distance Metrics - Euclidean, Cosine

Clustering was performed with n_clusters 4 to 64.

When this was infeasible due to the size of the dataset, we applied leiden clustering with n_neighbours=15 and resolutions [0.1-5] with 40 steps.

Spatial Metrics

Clustering quality was evaluated using 3 spatial metrics:

- **Silhouette Score** [15]: Measures the similarity of cells to their assigned cluster relative to other clusters, with values ranging from -1 to 1.
- **Calinski-Harabasz Index** [16]: Calculates the ratio of between-cluster to within-cluster dispersion, with higher values indicating better-separated clusters.
- **Davies-Bouldin Index** [17]: Quantifies clustering quality using the average ratio of within-cluster scatter to between-cluster separation, where lower values (closer to 0) indicate better cluster distinctiveness

Cell Type Stratification

Clustering was performed separately for each major cell type (astrocytes, OPCs, oligodendrocytes, immune cells, inhibitory neurons and excitatory neurons). This was done as the embedding space is largely dominated by cell type and would start by clustering by cell type. It was also done for memory and time complexity constraints as agglomerative clustering is $O(n^2)$ for both.

AD enrichment testing

Throughout the clustering procedure, we tracked cognitive diagnostic status for each cell based on the individual donor's clinical assessment. Cognitive status was categorized according to the ROSMAP "cogdx" variable: cognitively normal controls (CT), other cognitive decline (OCD), and Alzheimer's disease (AD) [18].

For each identified cluster, we computed the distribution of cells across diagnostic categories and assessed statistical significance using Fisher's exact test, with false discovery rate (FDR) correction for multiple testing. Clusters showing significant enrichment (adjusted p-value ¡ 0.05) for AD diagnosis were designated as AD-enriched clusters. Enrichment level was quantified by taking the cluster AD proportion over the background AD proportion.

Selecting n_clusters

In order to select n_clusters for each cell type we analyze spatial metrics and enrichment results ordered by silhouette score.

Scores will generally be decreasing as n_clusters increases.

We look for a "corner" value where silhouette score either increases on a larger n or where scores decrease sharply after staying relatively stable. Other metrics are also taken into account as tiebreakers.

Choice was then verified by visually inspecting UMAP plots of cell embeddings.

5.4 Differential Gene Expression & ORA

Differential gene expression was performed using scanpy. Comparisons are vs every other cluster unless otherwise specified. ORA was performed using enrichr to test for enriched pathways.

5.5 Sample proportion analysis

Sample proportion matricies were constructed from cell subgroup assignements and clustered with k means clustering. Fishers exact test is applied like above to test for enriched clusters

5.6 Computational Resources

All analyses were performed on a the Delft AI Cluster using NVIDIA A100 GPUs. Geneformer inference required approximately 1 hour per 20,000 cells. Hierarchical clustering and differential expression analyses were parallelized across multiple CPU cores to optimize computational efficiency.

Statistical analyses were conducted using Python 3.9 with the `scanpy`, `pandas`, and `scipy` libraries. Visualization utilized `matplotlib`, `seaborn`, and `plotly` packages. All random number generators were seeded to ensure reproducible results.

6 Responsible Research

6.1 Code and Data Availibility

Our analysis code, pre-processing pipelines, and documentation are publically available at our GitHub repository. ROSMAP data can be accessed publicly through the AD Knowledge Portal on Synapse, with access being granted after approval via standard data use agreements. Data preprocessing scripts, which are specific to our computational environement, can be found at our processing repository. Computational Requirements: Geneformer embedding generation needs GPU support with CUDA compatability and roughly 12GB of GPU memory. We provide detailed hardware specs and software dependencies in our repo.

6.2 Reproducibility and Replicability

Our pipeline was designed with reproducibility in mind. Our GitHub repository contains extensive documentation, step by step instructions, bash scripts for every step of the analysis, and a containerized Apptainer environment to ensure computational reproducibility. Data pre-processing, tokenization, embedding generation, and clustering analyses can all be executed following our documented protocols. Replication on other datasets should be feasable given Geneformer's training on diverse human cell atlases. We've included recommendations for adapting our preprocessing pipeline to other single-cell datasets, although validation on independent cohorts will be necessary to establish generalizability.

6.3 Ethical Considerations

Ethics remains an open problem in computer science and there are several implications that need careful consideration in this high-stakes medical research context.

Clinical Translation: Our computational findings represent research hypotheses at an early stage and require rigorous validation prior to any clinical translation. Premature translation of subtyping results could potentially mislead therapeutic development efforts, so responsible reporting and clear communication of limitations is essential.

Representation and Bias: The ROSMAP data consists primarily of white female participants, which may limit how generalizable our identified subtypes are to diverse populations. AD pathology and disease progression can vary across demographic groups, and our findings might not fully capture this important heterogeneity. Validation in diverse populations will be needed to ensure equitable medical advancement.

Data Privacy: Although ROSMAP contains de-identified data from deceased participants, we recognize the sensitive nature of genomic and health data. Our aggregate analysis approach helps minimize re-identification risks, but we strictly follow data use agreements and privacy protocols. Possible Misuse: Outside of research applications, our subtyping method could potentially be misapplied for healthcare rationing or insurance discrimination purposes. While we can't control all downstream applications, we want to emphasize that our methods are intended solely for advancing disease understanding and therapeutic development. We're committed to openly disclosing uncertainties, limitations, and the exploratory nature of our findings, stressing that results should be treated as research hypotheses rather than clinical tools.

6.4 Use of Large Language Models

Large language models were used when writing this report, specifically we employed claude 4 sonnet and opus.

LLMs were used to:

- Generate summaries of relevant topics for reference
- Generating rough drafts and ideas referenced while writing.
- Explaining/finding biological and ML terminology and concepts

Table 1: Top 15 Pathway Enrichment Results for ROSMAP proportions - 4 clusters

Gene Set	Term	Overlap	P-value	Adj. P-value	Odds Ratio	Cluster
Reactome_2022	HSF1 Activation R-HSA-3371511	13/29	2.33×10^{-16}	4.61×10^{-14}	51.20	3
Reactome_2022	Attenuation Phase R-HSA-3371568	12/26	2.18×10^{-15}	3.60×10^{-13}	53.84	3
GO_Biological_Process_2021	positive regulation of tau-protein kinase activity	4/6	1.00×10^{-6}	1.30×10^{-4}	122.57	3
Reactome_2022	Formation of the cornified envelope R-HSA-6809371	8/14	3.11×10^{-10}	1.53×10^{-8}	65.14	3
Reactome_2022	Keratinization R-HSA-6805567	8/14	3.11×10^{-10}	1.53×10^{-8}	65.14	3
GO_Biological_Process_2021	keratinocyte differentiation	8/22	1.78×10^{-8}	1.54×10^{-6}	41.51	3
GO_Biological_Process_2021	scroll development	9/29	3.64×10^{-8}	2.84×10^{-6}	35.40	3
GO_Biological_Process_2021	epidermis development	9/30	4.51×10^{-8}	3.26×10^{-6}	34.23	3
Reactome_2022	HSF1-dependent transactivation R-HSA-3371497	6/11	5.59×10^{-8}	3.72×10^{-6}	62.21	3
GO_Cellular_Component_2021	cornified envelope	6/11	5.59×10^{-8}	1.67×10^{-6}	62.21	3
GO_Biological_Process_2021	cornification	6/11	5.59×10^{-8}	3.72×10^{-6}	62.21	3
GO_Biological_Process_2021	epidermal cell differentiation	8/25	5.70×10^{-8}	3.72×10^{-6}	36.51	3
Reactome_2022	Cellular response to heat stress R-HSA-3371556	11/47	1.58×10^{-7}	9.14×10^{-6}	26.70	3
GO_Biological_Process_2021	cellular response to unfolded protein	7/20	1.86×10^{-7}	1.03×10^{-5}	39.93	3
GO_Biological_Process_2021	cellular response to topologically incorrect protein	7/20	1.86×10^{-7}	1.03×10^{-5}	39.93	3

Table 2: Top 15 Pathway Enrichment Results for GF proportions - 4 clusters

Gene Set	Term	Overlap	P-value	Adj. P-value	Odds Ratio	Cluster
WikiPathway_2021_Human	Wnt Signaling WP428	14/35	5.09×10^{-14}	1.27×10^{-11}	30.40	1
Reactome_2022	Signaling by WNT R-HSA-195721	17/61	6.82×10^{-12}	8.47×10^{-10}	21.17	1
GO_Biological_Process_2021	canonical Wnt signaling pathway	12/32	2.25×10^{-11}	1.87×10^{-9}	28.50	1
Reactome_2022	TCF dependent signaling R-HSA-201681	11/26	3.04×10^{-11}	1.89×10^{-9}	32.13	1
GO_Molecular_Function_2021	transcriptional activator activity	15/54	4.44×10^{-11}	2.77×10^{-9}	21.10	1
Reactome_2022	Beta-catenin independent WNT signaling R-HSA-3858494	9/18	4.82×10^{-11}	2.00×10^{-9}	38.00	1
GO_Biological_Process_2021	Wnt signaling pathway	14/45	8.44×10^{-11}	3.52×10^{-9}	23.62	1
GO_Biological_Process_2021	positive regulation of transcription	22/125	1.19×10^{-10}	3.71×10^{-9}	13.37	1
Reactome_2022	PCP/CE pathway R-HSA-4086398	8/15	1.56×10^{-10}	4.33×10^{-9}	40.53	1
Reactome_2022	Noncanonical activation of NOTCH3 R-HSA-9013695	8/15	1.56×10^{-10}	4.33×10^{-9}	40.53	1
GO_Biological_Process_2021	positive regulation of gene expression	20/109	2.05×10^{-10}	4.64×10^{-9}	13.94	1
GO_Biological_Process_2021	planar cell polarity pathway	8/16	2.64×10^{-10}	5.48×10^{-9}	38.00	1
GO_Molecular_Function_2021	DNA-binding transcription activator activity	14/48	2.92×10^{-10}	5.59×10^{-9}	22.17	1
Reactome_2022	Signaling by NOTCH3 R-HSA-9013507	8/16	2.64×10^{-10}	6.18×10^{-9}	38.00	1
GO_Biological_Process_2021	positive regulation of RNA polymerase II transcription	19/104	3.51×10^{-10}	6.32×10^{-9}	13.88	1

Table 3: Top 15 Pathway enrichment Results for ROSMAP subgroups - 6 clusters

Gene Set	Term	Overlap	P-value	Adj. P-value	Odds Ratio	Cluster
Reactome_2022	HSF1 Activation R-HSA-3371511	12/29	8.99×10^{-14}	1.78×10^{-11}	35.73	3
Reactome_2022	Attenuation Phase R-HSA-3371568	11/26	1.33×10^{-12}	1.31×10^{-10}	36.54	3
GO_Biological_Process_2021	positive regulation of tau-protein kinase activity	4/6	8.85×10^{-7}	1.15×10^{-4}	100.80	3
Reactome_2022	Formation of the cornified envelope R-HSA-6809371	7/14	1.70×10^{-8}	8.34×10^{-7}	43.20	3
Reactome_2022	Keratinization R-HSA-6805567	7/14	1.70×10^{-8}	8.34×10^{-7}	43.20	3
GO_Biological_Process_2021	keratinocyte differentiation	7/22	1.12×10^{-7}	4.88×10^{-6}	27.50	3
GO_Biological_Process_2021	scaldevelopment	8/29	1.47×10^{-7}	5.75×10^{-6}	23.86	3
GO_Biological_Process_2021	epidermis development	8/30	1.81×10^{-7}	6.58×10^{-6}	23.04	3
Reactome_2022	HSF1-dependent transactivation R-HSA-3371497	5/11	3.66×10^{-7}	1.22×10^{-5}	39.27	3
GO_Cellular_Component_2021	cornified envelope	5/11	3.66×10^{-7}	5.49×10^{-6}	39.27	3
GO_Biological_Process_2021	cornification	5/11	3.66×10^{-7}	1.22×10^{-5}	39.27	3
GO_Biological_Process_2021	epidermal cell differentiation	7/25	3.96×10^{-7}	1.30×10^{-5}	24.19	3
Reactome_2022	Cellular response to heat stress R-HSA-3371556	10/47	5.71×10^{-7}	1.66×10^{-5}	18.39	3
GO_Biological_Process_2021	cellular response to unfolded protein	6/20	1.11×10^{-6}	3.06×10^{-5}	25.92	3
GO_Biological_Process_2021	cellular response to topologically incorrect protein	6/20	1.11×10^{-6}	3.06×10^{-5}	25.92	3

Table 4: Top 15 Gene Enrichment Results for GF proportions - 6 clusters

Gene Set	Term	Overlap	P-value	Adj. P-value	Odds Ratio	Cluster
WikiPathway_2021_Human	Wnt Signaling WP428	14/35	5.09×10^{-14}	1.27×10^{-11}	30.40	1
Reactome_2022	Signaling by WNT R-HSA-195721	17/61	6.82×10^{-12}	8.47×10^{-10}	21.17	1
GO_Biological_Process_2021	canonical Wnt signaling pathway	12/32	2.25×10^{-11}	1.87×10^{-9}	28.50	1
Reactome_2022	TCF dependent signaling R-HSA-201681	11/26	3.04×10^{-11}	1.89×10^{-9}	32.13	1
GO_Molecular_Function_2021	transcriptional activator activity	15/54	4.44×10^{-11}	2.77×10^{-9}	21.10	1
Reactome_2022	Beta-catenin independent WNT signaling R-HSA-3858494	9/18	4.82×10^{-11}	2.00×10^{-9}	38.00	1
GO_Biological_Process_2021	Wnt signaling pathway	14/45	8.44×10^{-11}	3.52×10^{-9}	23.62	1
GO_Biological_Process_2021	positive regulation of transcription	22/125	1.19×10^{-10}	3.71×10^{-9}	13.37	1
Reactome_2022	PCP/CE pathway R-HSA-4086398	8/15	1.56×10^{-10}	4.33×10^{-9}	40.53	1
Reactome_2022	Noncanonical activation of NOTCH3 R-HSA-9013695	8/15	1.56×10^{-10}	4.33×10^{-9}	40.53	1
GO_Biological_Process_2021	positive regulation of gene expression	20/109	2.05×10^{-10}	4.64×10^{-9}	13.94	1
GO_Biological_Process_2021	planar cell polarity pathway	8/16	2.64×10^{-10}	5.48×10^{-9}	38.00	1
GO_Molecular_Function_2021	DNA-binding transcription activator activity	14/48	2.92×10^{-10}	5.59×10^{-9}	22.17	1
Reactome_2022	Signaling by NOTCH3 R-HSA-9013507	8/16	2.64×10^{-10}	6.18×10^{-9}	38.00	1
GO_Biological_Process_2021	positive regulation of RNA polymerase II transcription	19/104	3.51×10^{-10}	6.32×10^{-9}	13.88	1