A Foundation Model for Single-Cell Transcriptomics in Alzheimer's Disease

Athanasios Balomenos
Libra AI Technologies
Athens, Greece
athanasios.balomenos@libramli.ai

Christos Petrou
Libra AI Technologies
Athens, Greece
christos.petrou@libramli.ai

Theodoros Siozos
Libra AI Technologies
Athens, Greece
theodoros.siozos@libramli.ai

Ioannis Charalampopoulos
Pharmacology Department, University of Crete
IMBB, FORTH
Heraklion, Greece
charalampn@uoc.gr

Calogeropoulou Theodora
Institute of Chemical Biology, NHRF
Athens, Greece
tcalog@eie.gr

Yannis Kopsinis
Libra AI Technologies
Athens, Greece
yannis.kopsinis@libramli.ai

***Abstract*—Alzheimer's Disease (AD) presents a complex, heterogeneous pathology that is difficult to decode with traditional bulk or static transcriptomic approaches. Single-cell RNA sequencing (scRNA-seq) offers unprecedented resolution, but extracting consistent signals across large, diverse datasets remains a major challenge. In this work, we introduce a foundation model trained on a harmonized compendium of 1.7 million single cells from AD and control brains. Built on a transformer-based generative architecture and extended with a curated gene vocabulary, our model achieves 83% accuracy in distinguishing AD from control cells and 83% accuracy in predicting Braak stage, substantially outperforming non-pretrained baselines. Beyond classification, our model supports unsupervised extraction of biologically meaningful gene clusters, that may reveal transcriptional modules related to AD. These results demonstrate that foundation models can unlock scalable, interpretable frameworks for modelling neurodegeneration and support systematic hypothesis generation in AD research.**

***Index Terms*—Alzheimer's Disease, single-cell RNA sequencing, foundation models, transformer, generative pretraining, cell embeddings, gene clustering, Braak stage classification**

I. Introduction

Single-cell RNA sequencing (scRNA-seq) enables profiling of gene expression at cellular resolution, offering a powerful means to dissect the molecular and cellular heterogeneity of complex tissues. By capturing transcriptomic variation across individual cells, scRNA-seq has transformed our understanding of development, aging, and disease mechanisms that are often obscured in bulk RNA sequencing analyses. Alzheimer's disease (AD), the leading cause of dementia worldwide, affects over 50 million individuals and is characterized by progressive neurodegeneration and pronounced cellular diversity [1]. Understanding how distinct cell populations contribute to or respond to AD pathology remains a major challenge. Recent applications of scRNA-seq to AD research have revealed disease-associated activation states in microglia, reactive astrocyte phenotypes, and selective neuronal vulnerability [2]–[4]. However, these findings are often dataset-dependent, as technical variability, cohort differences, and protocol heterogeneity complicate cross-study integration [2], [5], [6]. For instance, Lau et al. [5] reported that batch effects across AD cohorts hinder direct comparison of cell states, while Morabito et al. [6] emphasized the need for advanced harmonization approaches such as Seurat and Harmony to achieve consistent integration across single-cell datasets. Consequently, building unified analytical frameworks that generalize across datasets remains a key bottleneck in AD transcriptomics. Foundation models, originally developed for natural language processing, have recently been adapted to biological sequence and single-cell omics data, capturing high-dimensional relationships through generative pretraining [7], [8]. The scGPT framework, for example, jointly learns gene and cell embeddings to support downstream biological inference [8]. Yet, existing foundation models are general-purpose and not optimized to capture disease-specific transcriptional programs. In AD, where gene–gene and cell–cell relationships deviate markedly from healthy states due to neuroinflammatory and degenerative processes, disease-focused pretraining can enhance both the sensitivity and interpretability of single-cell representations [9]. This specialization enables the model to uncover subtle AD-related transcriptional shifts that may be diluted in general-purpose embeddings. Here, we introduce a disease-tuned foundation model trained on 96 curated scRNA-seq datasets from AD postmortem brain samples. Our approach integrates harmonized gene vocabularies, AD-specific pretraining, and systematic validation through classification and gene network analyses. We demonstrate that the model improves disease-context prediction tasks and reveals biologically interpretable gene clusters linked to AD pathology, providing a scalable computational framework for modeling disease complexity and supporting hypothesis generation in neurodegenerative research.

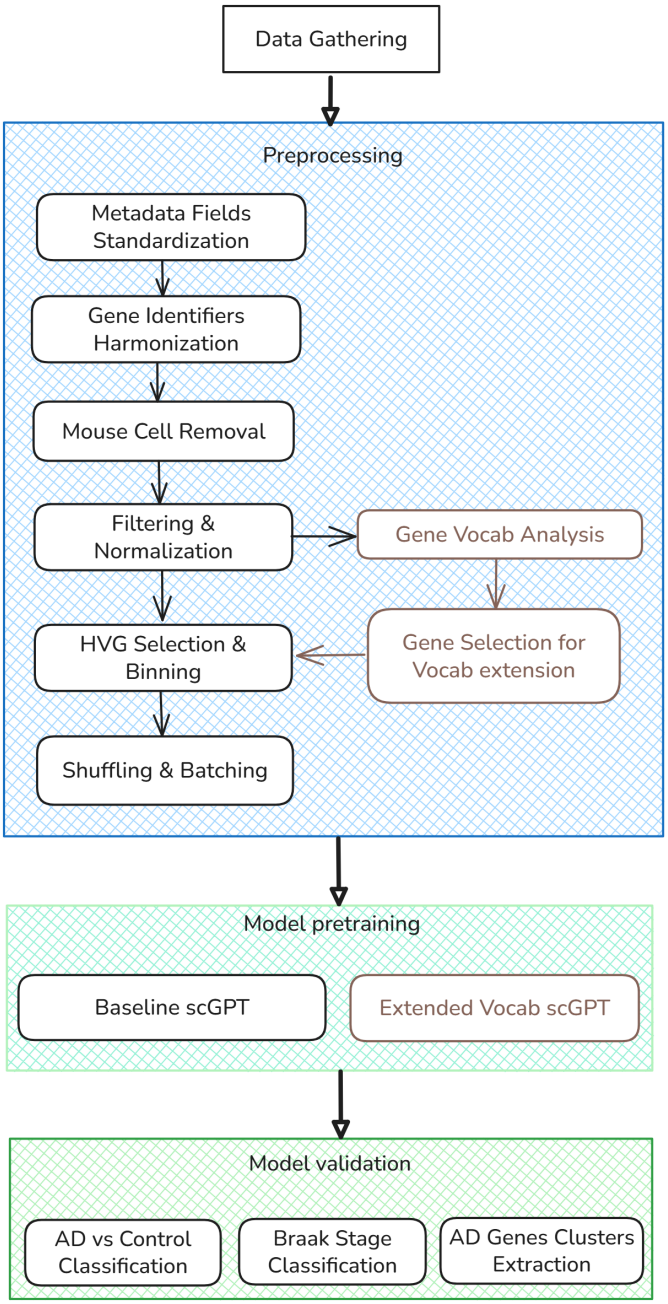

Fig. 1. Overview of the proposed methodology.

II. Methodology

A. Data gathering

The dataset integrates data from four primary sources: ssREAD [9], Dharsini et al. [10], Rexach et al. [11] and Pan et al. [12]. ssREAD serves as a comprehensive AD data portal aggregating multiple studies, while the rest represent recent studies accessed via cellXgene [13]. We focused exclusively on human scRNA-sequencing data, accompanied by metadata including brain region, cell type, donor age, sex, and Braak stage. ssREAD contributed 23 projects with 93 donors, while cellXgene provided six collections from 73 donors, prioritizing the latest studies not yet included in ssREAD.

B. Preprocessing:

To ensure high-quality inputs for training and evaluation, we performed a series of preprocessing steps on the single-cell RNA-seq dataset. These steps addressed both expression matrix consistency and metadata harmonization across sources.

Metadata Fields Standardization: The dataset was sourced from multiple studies with varying annotation formats and conventions. We standardized key metadata fields, including cell type labels, brain region identifiers, donor sex, and Braak stage annotations, by resolving naming discrepancies (e.g., "microglial cell" vs. "microglia"), unifying categorical fields, and removing ambiguous entries.

Gene Identifier Harmonization: We used MyGene.info [14], BioMart [15], and the HGNC dataset [16] to standardize gene symbols and Ensembl IDs across datasets, resolving synonyms, deprecated identifiers, and inconsistencies to ensure a unified gene vocabulary.

Mouse Cell Identification and Removal: Mouse-derived cells were identified and removed using species-specific gene annotations to retain only human cells for model training.

Following these preprocessing steps, we compiled a harmonized single-cell RNA-seq dataset suitable for model training and downstream analysis. The finalized dataset, which is publicly available via Zenodo [17], integrates 1.7 million annotated cells from AD and control donors, with metadata including cell type, Braak stage and brain regions. A summary of key dataset statistics is provided in Table I.

TABLE I
Summary of the created single-cell RNA-seq Dataset

Total Cells (with cell type)	1.7 M
Cells with Braak stage	761 k
Unique Genes	60 k
Donors	166

Filtering & Normalization: To reduce technical noise and remove uninformative features, we filtered out cells expressing fewer than three genes and genes in at least one cell. Gene expression values were normalized on a per-cell basis using total-count normalization and logarithmic transformation to ensure uniform scaling and variance stabilization across cells, with strictly positive expression values.

HVG Selection: Highly variable genes (HVG) were identified using Cell Ranger's dispersion-based method, which detects genes with significant expression variability across cells, a standard approach in single-cell RNA-seq preprocessing pipelines [18]. For each dataset, 10,000 HVGs were retained, while for larger datasets, 6,000 HVGs were selected.

Binning: For each cell, the raw expression value of each gene is placed into one of 50 equidistant bins. The lowest bin corresponds to the smallest (often zero or near-zero) expression values, and the highest bin corresponds to the largest observed expression values of genes across the cell.

This step both preserves biological signals and mitigates batch effects, preparing the data for transformer-based modeling.

Shuffling and Batching: To efficiently process large datasets, we developed data handling pipelines based on scBank data format. Training was performed using a batch size of 128.

C. Gene Vocabulary Analysis & Extension

To evaluate the alignment between our processed datasets and the original scGPT vocabulary of 60,697 genes, we conducted an extensive gene coverage analysis, revealing that only 37,215 genes (61.3%) were found in our data. Among the detected genes, file-level presence varied widely, with a median of 74.7%, while many genes appeared in only a small subset of the 96 curated datasets, often concentrated within specific cohorts such as Pan et al., Dharsini et al., and Rexach et al.

This bimodal distribution, with a large fraction of genes sparsely represented and a smaller core set appearing consistently across nearly all files, highlighted the limitations of relying on raw vocabulary presence alone for foundation model training. To ensure that our extended vocabulary captures genes that are both biologically relevant and broadly represented across datasets, we identified the need for a systematic, quantitative scoring analysis. This approach enables prioritization of genes that contribute meaningfully to generalizable representation learning, while filtering out low-coverage genes that add noise and do not support robust downstream inference in the context of AD modeling.

To address missing or sparsely covered genes, we developed a quantitative scoring system to assess biological relevance and prioritize genes for vocabulary inclusion. Each gene was evaluated using the following expression-based metrics:

Study Coverage (SC) [20%]. This metric quantifies the proportion of studies in which the ith gene is present, reflecting its cross-cohort consistency.

$$SC_i = \frac{\text{\# of studies including } i^{\text{th}} \text{ gene}}{\text{Total \# of studies}} \times 100 \tag{1}$$

Cell Expression Coverage (CEC) [25%]. This metric measures the proportion of cells in which the ith gene is expressed (non-zero), indicating its prevalence across the dataset.

$$CEC_i = \frac{\text{\# of cells where } g_{\text{expr}} > 0}{\text{Total \# of cells}} \times 100 \tag{2}$$

Scaled Mean Expression (SME) [20%]. This metric captures the mean expression of the gene, scaled relative to the 95th percentile across all genes to ensure comparability.

$$SME_i = \min\left(\frac{\text{Mean}(g_{\text{expr}})}{\text{95th percentile of } (g_{\text{expr}})}, 1\right) \tag{3}$$

Coefficient of Variation (CV) [20%]. This metric reflects the variability of the gene's expression relative to its mean, providing insight into its dynamic range across cells.

$$CV_i = \frac{\text{std}(g_{\text{expr}})}{\text{mean}(g_{\text{expr}})} \tag{4}$$

Expression Heterogeneity (EH) [15%]. This metric uses the standard deviation of the gene's expression across all cells as a direct measure of heterogeneity.

$$EH_i = \text{std}(g_{\text{expr}}) \tag{5}$$

where g_{expr} is the expression level of the ith gene.

Using these metrics, we compute the final gene prioritization score as:

$$\begin{aligned} S_i = 0.20 \cdot SC_i + 0.25 \cdot CEC_i + 0.20 \cdot SME_i \\ +0.20 \cdot CV_i + 0.15 \cdot EH_i \end{aligned} \tag{6}$$

This systematic, biologically informed scoring ensures that genes prioritized for vocabulary extension in the foundation model are both representative across datasets and relevant to Alzheimer's disease biology.

The top 2,000 missing genes by score were selected for inclusion in the extended vocabulary, ensuring both biological relevance and cross-dataset consistency. This detailed gene coverage analysis allowed us to move beyond binary presence/absence checks. By incorporating both coverage frequency and expression quality, we ensured that the extended vocabulary better represents AD-relevant biology, improves model generalizability, and supports downstream analyses such as gene regulatory network inference.

D. Foundation model development

Each file was split into training and validation datasets using a 90/10 ratio. In total, we used 1,729,662 cells with 1,556,695 in training and 172,967 in validation. A custom PyTorch dataloader was implemented to handle tokenization, batching, and file management.

To establish a robust representation of single-cell transcriptomic data in the context of AD, we trained two configurations of our foundation model. The first model was trained using the original scGPT vocabulary, serving as a baseline to evaluate generalization across diverse cell populations and experimental conditions. The second configuration utilized an extended vocabulary, which included an additional 2,000 genes selected based on coverage, expression quality, and biological relevance to AD. This extended model was designed to enhance the representation of disease-relevant molecular features, with the goal of improving downstream tasks such as GRN inference and microglial subtype identification.

The training dynamics of the foundation model, visualized in Fig 2, indicate a healthy and stable optimization process. Both foundation model variants, one trained on the original scGPT vocabulary and the other on the extended vocabulary, exhibited stable training dynamics and converged smoothly without signs of overfitting. The second model, which incorporated an additional 2,000 biologically prioritized genes, maintained comparable loss profiles while expanding its representational capacity. This confirms that the extended vocabulary did not compromise training stability and provides a stronger foundation for downstream biological interpretation, particularly in AD-related contexts.

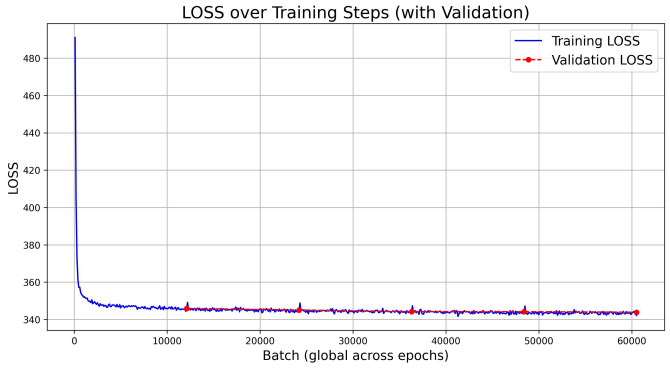

Fig. 2. Training and validation loss of foundation model training with extended scGPT vocabulary. This line plot shows the training loss (blue) and validation loss (red dots) across 60,000 batches. The training loss decreases rapidly in the early steps, then gradually converges toward a stable minimum. The close alignment between training and validation loss throughout the training process indicates strong generalization performance and an absence of overfitting, confirming the model's robustness on unseen data.

In conclusion, we laid the foundation for robust modeling of single-cell RNA-seq data in AD. By harmonizing datasets, curating a biologically informed gene vocabulary, and implementing scalable training pipeline, we are creating a model capable of empowering multiple downstream tasks related to AD.

III. Results & Discussion

A. Control vs AD Cell Classification

The first downstream task used to validate our foundation models is binary classification, predicting whether a cell belongs to an AD or Control condition. For this task, we train a classification head connected to the frozen foundational transformer model.

We report results for three models:

- **scGPT Brain:** The baseline model pre-trained on brain cells.
- **ADPT:** The pretrained model on AD with the same vocabulary as scGPT Brain.
- **ADPT extended vocab:** The pretrained model on AD with 2,000 additional genes extending the base vocabulary.

All models use the same classification head, consisting of 34k trainable parameters. The goal is to test whether pretraining enables the foundational model to learn AD-specific features, generating more meaningful embeddings that improve classification accuracy with a simple head.

In the classification task we uses the whole dataset, 1.7M cells, split as follows:

- Train set: 1M (60%)
- Validation set: 350k (20%)
- Test set: 350k (20%)

Each input consists of gene expression values, with labels indicating the condition ("alzheimer" or "control"). The mini-batch size is 128, and models are finetuned for 8 epochs using Binary Cross Entropy Loss.

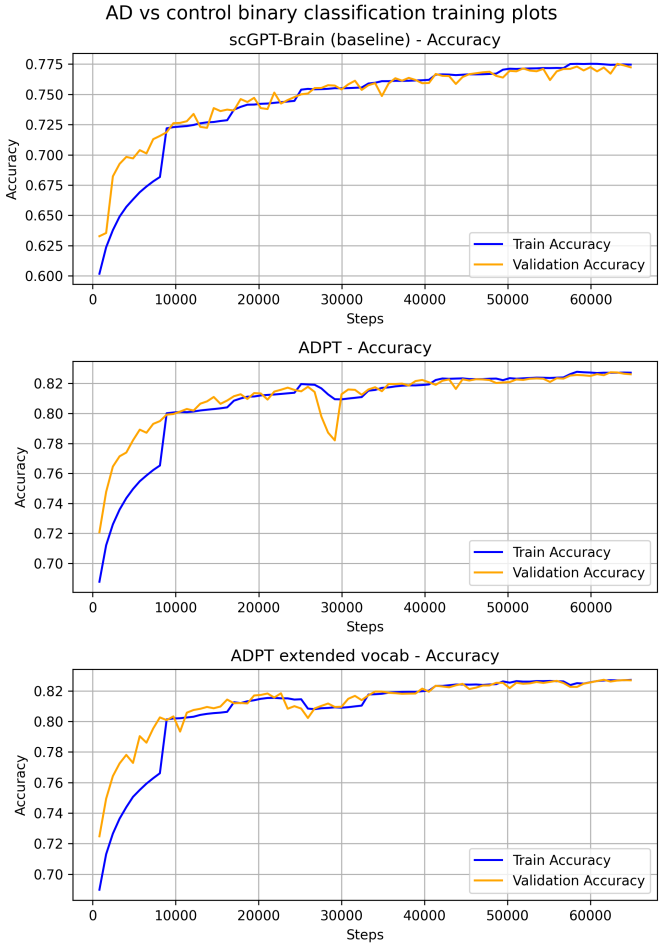

Fig. 3. Accuracy plots for AD vs Control binary classification across distinct model categories. The top panel shows scGPT-Brain (baseline), the middle panel shows ADPT, and the bottom panel shows ADPT with extended vocabulary. Validation and training accuracy improvements illustrate the benefits of pretraining on AD-specific data and extended gene vocabulary for capturing disease-relevant signals during fine-tuning.

Figure 3 shows the training and validation loss curves. We observe a clear performance improvement in both ADPT models compared to the baseline model. However, the extended vocabulary variant performs similarly to the standard ADPT model.

Figures 4(a) and (b) present confusion matrices for the ADPT models on the test set. Both models achieve an accuracy of 0.83 but exhibit a higher number of false negatives, misclassifying AD cells as Control, a common challenge in AD classification, as resilient subpopulations can mask disease signatures [2], [19].

Table II summarizes validation loss and accuracy for each model, showing clear improvements from pretraining.

B. Braak Stage Classification

The second downstream task used to validate our scGPT-AD finetuned models is multi-class classification, predicting the Braak stage (0–6) for each cell. This task assesses whether the

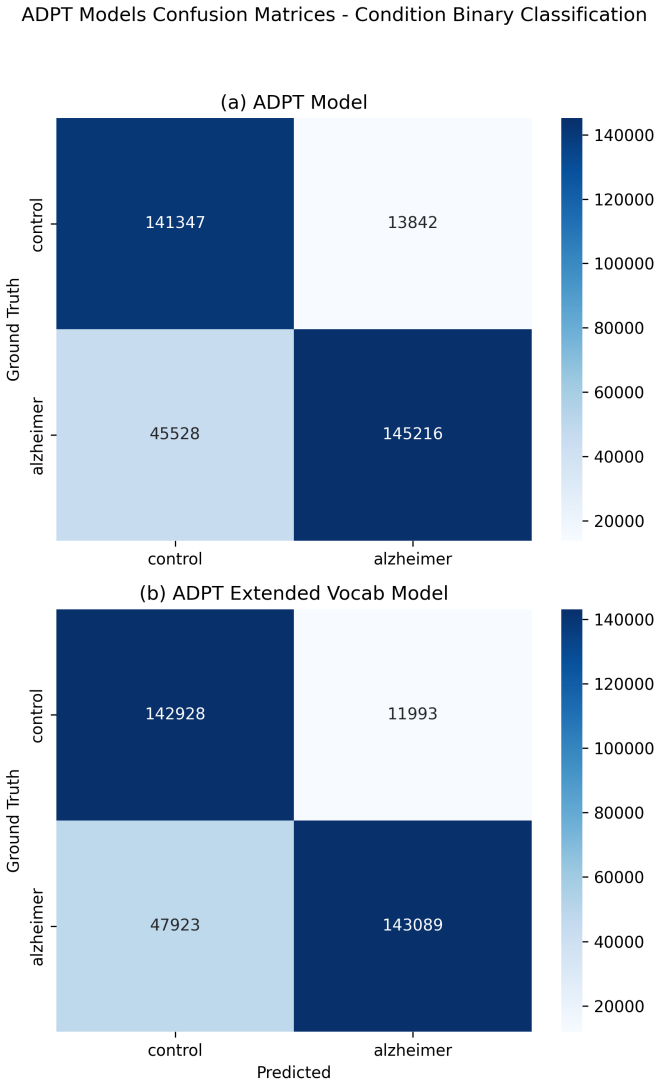

Fig. 4. Condition binary classification confusion matrices on the test set for finetuned ADPT models: (a) standard vocabulary (Accuracy: 0.83), (b) extended vocabulary (Accuracy: 0.83).

TABLE II
MODEL PERFORMANCE OVERVIEW FOR THE AD VS CONTROL CLASSIFICATION TASK.

Model	Validation Loss	Validation Accuracy (%)
scGPT Brain	0.43	77
ADPT	0.32	**83**
ADPT extended vocab	0.32	**83**

pretrained scGPT-AD models capture granular AD progression signals within single-cell transcriptomic profiles.

As in the previous task, we evaluate three models, **scGPT Brain**, **ADPT** and **ADPT extended vocab**. The models use the same classification head with 34k trainable parameters, adapted for 7 output classes corresponding to Braak stages 0–6. This setup tests whether pre-training helps the foundational model learn stage-specific features, enabling it to generate

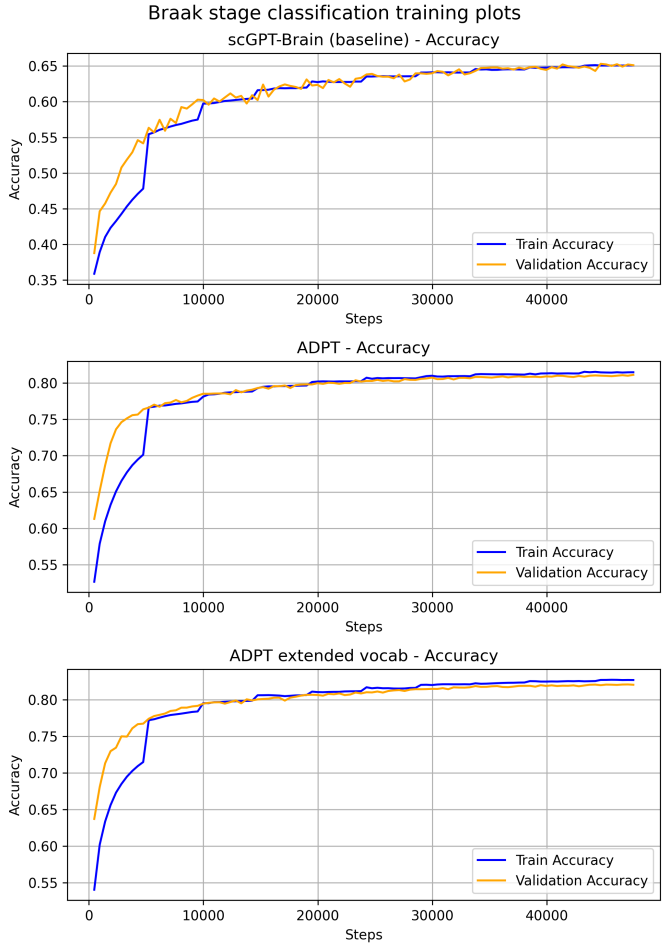

Fig. 5. Accuracy plots for Braak stage classification across distinct model categories. The top panel shows scGPT-Brain (baseline), the middle panel shows ADPT, and the bottom panel shows ADPT with extended vocabulary. The plots illustrate improved validation and training accuracy with AD-specific pretraining and extended vocabulary, highlighting the models' ability to capture stage-specific signals in single-cell transcriptomic data.

meaningful embeddings that improve Braak stage classification accuracy when paired with a simple head.

Due to missing or vague Braak stage annotations in parts of the dataset, we exclude these cells, resulting in 761k cells labeled with Braak stages (0–6). The dataset is split as:

- Train set: 609k cells (80%)
- Validation set: 76k cells (10%)
- Test set: 76k cells (10%)

Each input consists of gene expression values, with the label indicating the Braak stage (integer 0–6). The mini-batch size is 128, and models are finetuned for 10 epochs using CrossEntropyLoss for this multiclass classification task.

Figure 5 shows the training and validation accuracy curves in all models. We observe a clear improvement in both ADPT models compared to the baseline, with the extended vocabulary variant outperforming the standard ADPT model.

Table III summarizes validation loss and accuracy for each model, demonstrating the improvements from disease-specific

TABLE III
MODEL PERFORMANCE OVERVIEW FOR THE BRAAK STAGE CLASSIFICATION TASK.

Model	Validation Loss	Validation Accuracy (%)
scGPT Brain	1.51	65
ADPT	1.36	82
ADPT extended vocab	1.35	**83**

pretraining and vocabulary extension.

Figures 6(a) and (b) present confusion matrices for our models on the test set, confirming that test set results align with training distributions. Notably, the ADPT extended vocab model demonstrates strong diagonal dominance, particularly for stages 0, 2, and 6, indicating reliable classification for these stages. Higher stages, particularly stages 4 and 5, often show confusion with stage 6, reflecting known challenges in distinguishing intermediate stages within continuous neurodegenerative progression while confirming that the model captures key stage-dependent patterns in single-cell profiles.

C. Brain region-Specific Gene Clusters related to AD

To explore the biological relevance of the learned representations, we leveraged the ADPT gene embeddings to construct gene clusters in a fully unsupervised manner. Gene modules were identified by clustering the model-derived embeddings, then thresholding to retain clusters with at least five genes, following the strategy of [20]. Using STRING v12 [21] and KEGG [22], we identified 141 gene clusters and 550 genes associated with AD. These results provide a foundation for the extension of the molecular mechanisms and biological pathways involved in the pathology of AD.

Figure 7 presents a hierarchically clustered heatmap of gene module activity across major brain regions in the context of AD. Rows correspond to gene clusters, and columns represent brain regions. The color intensity indicates normalized cluster activity, with lighter tones reflecting higher expression. Dendrograms reveal structure across both modules and cell types, suggesting patterns of co-regulation and expression specificity. Notably, several gene clusters display strong activation in the entorhinal cortex [23], [24] and hippocampus [24], [25] to play central roles in AD pathogenesis, while others are more broadly expressed in brain regions. These patterns underline the model's ability to uncover brain region-specific transcriptional signatures directly from AD single-cell data, providing candidates for further regulatory network analysis and functional validation.

IV. CONCLUSIONS

We present ADPT, a foundation model that leverages generative pretraining on large-scale single-cell transcriptomic data to model gene and cell relationships in the context of AD. Built on the scGPT architecture, our model jointly learns both gene and cell embeddings, analogous to word and sentence embeddings in natural language processing, providing a unified framework for capturing the molecular complexity of AD

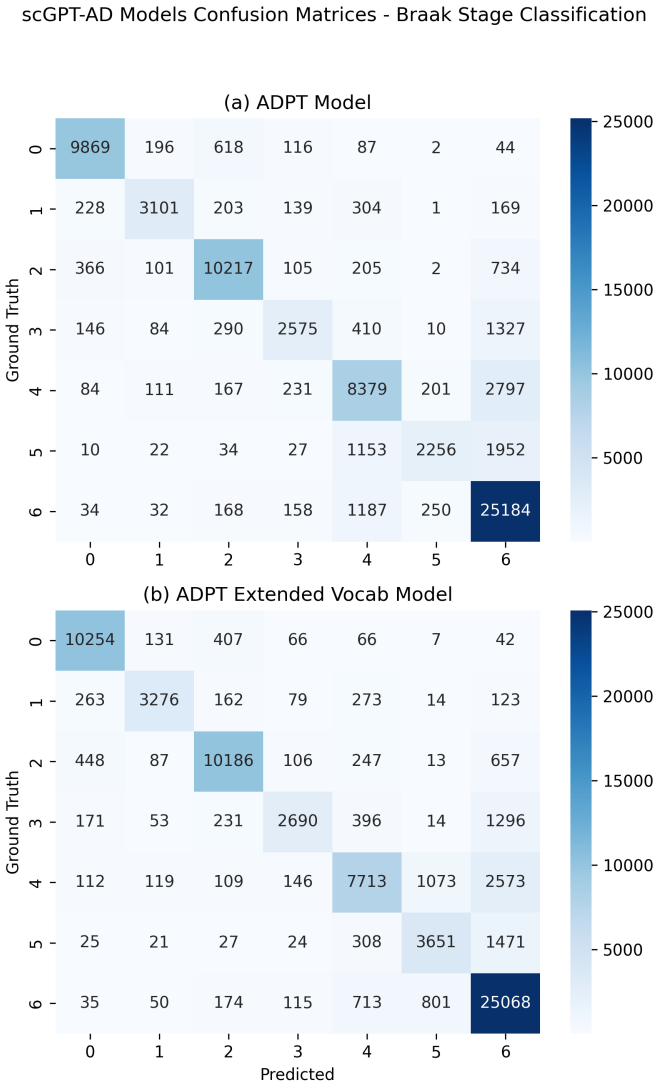

Fig. 6. Braak stage classification confusion matrices on the test set for fine-tuned ADPT models: (a) standard vocabulary (Accuracy: 0.81), (b) extended vocabulary (Accuracy: 0.83).

at cellular resolution. We demonstrate the utility of ADPT across multiple downstream tasks. These results highlight the model's ability to generalize across tasks while maintaining interpretability, a crucial feature for biological applications.

Moving forward, ADPT opens new avenues for data-driven discovery in neurodegeneration research. Its ability to infer cell-type gene programs from large, heterogeneous datasets positions it as a powerful tool for hypothesis generation and targeted experimental design. By bridging large-scale computational modelling with biological insight, ADPT sets the stage for a new class of AI-guided investigation in AD.

ACKNOWLEDGMENT

The Generative LLM for Single-Cell Transcriptomics of Alzheimer's Disease has received funding from the European Union, via the oc1-2024-TIS-01 issued and implemented

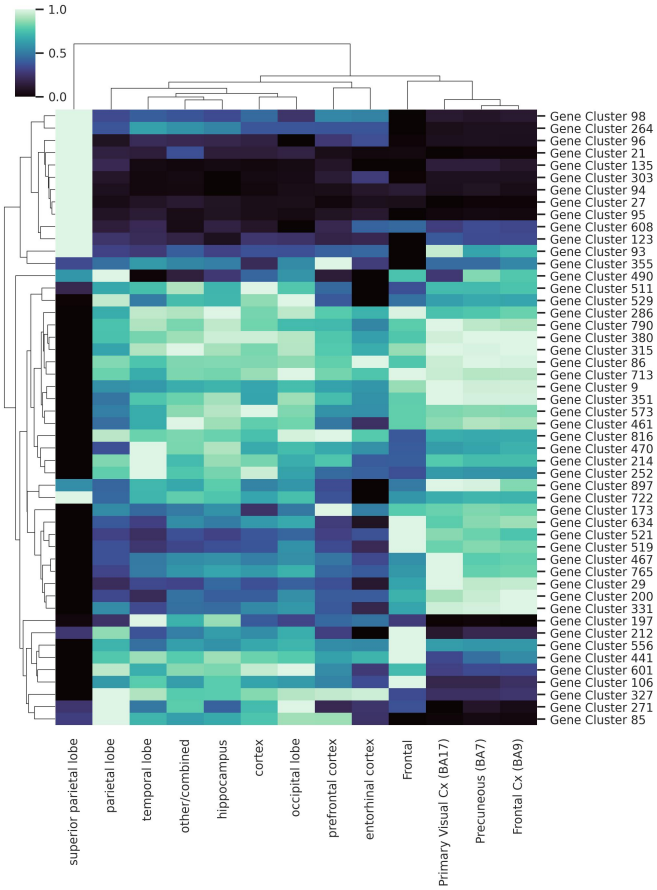

Fig. 7. Brain region correlation with gene clusters derived from unsupervised clustering. The heatmap shows the differential activity of gene clusters across annotated brain regions. Enrichment scores reveal biologically coherent AD-related genes per per brain region. Only AD-labeled cells were included for clustering.

by the ENFIELD project, under the grant agreement No 101120657.

References

[1] World Health Organization, "Dementia," Fact Sheets, 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/dementia

[2] H. Mathys, J. Davila-Velderrain, Z. Peng, et al., "Single-cell transcriptomic analysis of Alzheimer's disease," *Nature*, vol. 570, no. 7761, pp. 332–337, 2019.

[3] A. Grubman, G. Chew, J. F. Ouyang, et al., "A single-cell atlas of entorhinal cortex from individuals with Alzheimer's disease reveals cell-type-specific gene expression regulation," *Nat. Neurosci.*, vol. 22, no. 12, pp. 2087–2097, 2019.

[4] B. De Strooper and E. Karran, "The Cellular Phase of Alzheimer's Disease," *Cell*, vol. 164, no. 4, pp. 603–615, 2016.

[5] Lau, S. F., Cao, H., Fu, A. K. Y., Ip, N. Y. (2020). Single-nucleus transcriptome analysis reveals dysregulation of angiogenic endothelial cells and neuroprotective glia in Alzheimer's disease. *Nature Neuroscience*, 23(6), 701–713. https://doi.org/10.1038/s41593-020-0631-7

[6] Morabito, S., Miyoshi, E., Michael, N., Shahin, M., Martini, A. C., Head, E., Silva, J., Lehallier, B., Wyss-Coray, T., Bennett, D. A., et al. (2021). Single-nucleus chromatin accessibility and transcriptomic characterization of Alzheimer's disease. *Nature Genetics*, 53(8), 1143–1155. https://doi.org/10.1038/s41588-021-00894-z

[7] M. Hao, J. Gong, X. Zeng, C. Liu, Y. Guo, X. Cheng, T. Wang, J. Ma, X. Zhang, and L. Song, "Large-scale foundation model on single-cell transcriptomics," *Nat. Methods*, vol. 21, no. 8, pp. 1481–1491, Aug. 2024, doi: 10.1038/s41592-024-02305-7.

[8] H. Cui, C. Wang, H. Maan, K. Pang, F. Luo, N. Duan, and B. Wang, "scGPT: toward building a foundation model for single-cell multi-omics using generative AI," *Nat. Methods*, vol. 21, no. 8, pp. 1470–1480, Aug. 2024, doi: 10.1038/s41592-024-02201-0.

[9] Wang, C., Acosta, D., McNutt, M. et al. A single-cell and spatial RNA-seq database for Alzheimer's disease (ssREAD). *Nat Commun* 15, 4710 (2024). https://doi.org/10.1038/s41467-024-49133-z

[10] S. A. P. Dharshini, A. P. Rajkumar, K. Rajamani, and A. Mani, "Molecular signatures of resilience to Alzheimer's disease in neocortical layer 4 neurons," *bioRxiv*, 2024.

[11] J. E. Rexach, D. S. Lee, S. A. Sloan, B. T. Staahl, and M. Kampmann, "Cross disorder and disease specific pathways in dementia revealed by single cell genomics," *Cell*, vol. 187, no. 20, pp. 5753–5774, 2024.

[12] J. Pan, J. Kim, T. M. Le, Y. Wang, and R. D. Salinas, "Deciphering glial contributions to CSF1R-related disorder via single nuclear transcriptomic profiling: a case study," *Acta Neuropathol. Commun.*, vol. 12, no. 1, p. 139, 2024.

[13] CZ CELLxGENE Discover: A single-cell data platform for scalable exploration, analysis and modeling of aggregated data CZI Single-Cell Biology, et al. *bioRxiv* 2023.10.30; doi: https://doi.org/10.1101/2023.10.30.563174

[14] J. Xin, A. Mark, C. Afrasiabi, G. Tsueng, M. Juchler, N. Gopal, ... and C. Wu, "High-performance web services for querying gene and variant annotation," *Genome Biol.*, vol. 17, p. 91, 2016, doi: 10.1186/s13059-016-0953-9.

[15] D. Smedley, S. Haider, B. Ballester, R. Holland, D. London, G. Thorisson, and A. Kasprzyk, "BioMart–biological queries made easy," *BMC Genomics*, vol. 10, p. 22, 2009, doi: 10.1186/1471-2164-10-22.

[16] S. Tweedie, B. Braschi, K. Gray, T. E. M. Jones, R. L. Seal, B. Yates, and E. A. Bruford, "Genenames.org: the HGNC and VGNC resources in 2021," *Nucleic Acids Res.*, vol. 49, no. D1, pp. D939–D946, 2021, doi: 10.1093/nar/gkaa980.

[17] T. Siozos, C. Petrou, A. Balomenos, and Y. Kopsinis, "Single-Cell RNA Data Portal for Alzheimer's Disease (Version 3)," Zenodo, dataset, 2025. [Online]. Available: https://zenodo.org/records/15526235

[18] G. X. Y. Zheng, J. M. Terry, P. Belgrader, et al., "Massively parallel digital transcriptional profiling of single cells," *Nature Communications*, vol. 8, no. 1, pp. 1–12, 2017. doi:10.1038/ncomms14049.

[19] Tandon, R., Levey, A.I., Lah, J.J., Seyfried, N.T. and Mitchell, C.S., 2023. Machine learning selection of most predictive brain proteins suggests role of sugar metabolism in Alzheimer's disease. Journal of Alzheimer's Disease, 92(2), pp.411-424.

[20] Nicholas Ceglia, et al. "GeneVector: Identification of transcriptional programs using dense vector representations defined by mutual information". *bioRxiv* (2022), pp. 2022–04.

[21] Szklarczyk, D., Kirsch, R., Koutrouli, M., Nastou, K. C., Mehryary, F., Hachilif, R., et al. (2023). The STRING database in 2024: Protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Research*, 52(D1), D676–D685. https://doi.org/10.1093/nar/gkad1020

[22] Kanehisa, M., Furumichi, M., Sato, Y., Kawashima, M., Ishiguro-Watanabe, M. (2023). KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Research*, 51(D1), D587–D592. https://doi.org/10.1093/nar/gkac963

[23] K. M. Igarashi, "Entorhinal cortex dysfunction in Alzheimer's disease," *Trends Neurosci.*, vol. 46, no. 2, pp. 124–136, Feb. 2023.

[24] A. P. Mecca, M. K. Chen, R. S. O'Dell, et al., "Association of entorhinal cortical tau deposition and hippocampal synaptic density in older individuals with normal cognition and early Alzheimer's disease," *Neurobiol. Aging*, vol. 111, pp. 44–53, Mar. 2022.

[25] Y. L. Rao, B. Ganaraja, B. V. Murlimanju, T. Joy, A. Krishnamurthy, and A. Agrawal, "Hippocampus and its involvement in Alzheimer's disease: a review," *3 Biotech*, vol. 12, no. 2, p. 55, Feb. 2022.