Title

NeuroSC: Leveraging Pretrained Single-Cell Models for Brain Cell Classification (Stage 1 Registered Report)

Authors

Hien Trinh[1], Minh Truong[2], Duy Pham[3]

Affiliations

1 MindX Technology School, Hanoi Capital Region, Vietnam
2 Pythia Bioscience, Carlsbad, California 90033, USA
3 Wellcome Sanger Institute, Hinxton, Saffron Walden CB10 1RQ, United Kingdom

Correspondence

Duy Pham, Ph.D.
Senior Computational Biologist,
Wellcome Sanger Institute,
Parasites and Microbes Program,
Wellcome Genome Campus,
Hinxton, Saffron Walden
CB10 1RQ
email: duy.pham@sanger.ac.uk

Abstract

The immense cellular diversity of the mammalian brain presents a significant challenge for precise cell type classification, particularly when dealing with the noise and sparsity inherent in single-cell RNA sequencing (scRNA-seq) data. We introduce NeuroSC, a computational framework designed to enhance brain cell identification by leveraging transfer learning from large-scale, pretrained single-cell foundation models. While traditional classification methods rely on supervised learning from scratch, NeuroSC utilizes models already trained on millions of diverse transcriptomes to capture complex gene-gene dependencies. In this study, we will fine-tune these models on specialized brain atlases, NeuroSC aims to achieve superior accuracy in identifying cell subtypes and transitional states that are often missed by standard clustering algorithms or prediction models. This transfer learning approach is potentially resilient to batch effects and technical variability across different sequencing platforms. Ultimately, NeuroSC provides a robust and scalable solution for automated cell annotation, supporting the development of high-esolution brain maps and the discovery of cellular drivers in neurological diseases.

1. Introduction

1.1 Research question, background, and relevance

The transcriptomic mapping of the mammalian brain has historically relied on unsupervised clustering and manual annotation, a process that is increasingly becoming a bottleneck as datasets scale to millions of cells[1], [2], [3]. While traditional tools like Seurat[4] or Scanpy[5] have provided a foundation for cell type discovery, they often struggle with the inherent noise, high dimensionality, and extreme sparsity of single-cell RNA sequencing (scRNA-seq) data. Furthermore, conventional supervised learning models frequently fail to generalize across different laboratories or sequencing platforms due to significant batch effects[6], [7]. Recent literature has seen a shift toward task specific deep learning, yet these models are typically trained de novo on narrow datasets, failing to capture the broader biological syntax of gene-gene interactions that govern cellular identity across the entire organism[8], [9], [10], [11].

NeuroSC addresses these limitations by applying transfer learning from large scale single-cell foundation models, such as scGPT[10] or Geneformer[11], specifically to the complex architecture of the brain. The importance of this study lies in its ability to automate high resolution annotation, allowing researchers to identify cell subtypes and subtle transitional states that are often not perform well in standard pipelines. By leveraging models pretrained on massive, diverse datasets, NeuroSC aims to achieve a level of robustness and cross-platform accuracy that is unattainable by training from scratch. This relevance extends directly to clinical research, as many neurodegenerative and neuropsychiatric disorders are cell type specific, the ability to precisely and scalable map these cells is critical for identifying disease biomarkers and developing targeted therapeutic interventions. Ultimately, NeuroSC serves as a vital bridge between general purpose artificial intelligence and the specialized needs of high-resolution neurobiology.

1.2 Hypotheses

The primary outcome of interest for NeuroSC is classification model with high performance, specifically the Macro-F1 score, which ensures that cell subtypes are given equal weight. This is supplemented by secondary outcomes focusing on cross-platform robustness, the model's ability to maintain accuracy when applied to datasets from different laboratories or sequencing technologies. These outcomes are defined by the degree of overlap between NeuroSC's automated predictions and the high-confidence, expert curated labels provided by reference atlases. By aggregating performance metrics across diverse cortical and subcortical regions, the study will determine if the framework provides a consistent advantage over traditional, non-transfer-learning-based classification tools.

We hypothesize that by leveraging biological syntax learned during large scale pretraining, NeuroSC will achieve a 5-10% improvement in F1-score for cell subtype populations compared to models trained de novo. Furthermore, we test the hypothesis that this transfer learning approach will demonstrate superior resilience to batch effects, maintaining high classification accuracy on held out datasets without the need for manual batch correction. These hypotheses will be measured through a direct head-to-head comparison with baseline supervised models (e.g., celltypist[12]) linking the model latent representations directly to its precision in identifying the complex cellular landscape of the mammalian brain.

1.3 Timeline

The project is structured into a rigorous six-month timeline, beginning with the first three months dedicated to data acquisition, preprocessing, and the fine-tuning of pretrained foundation models on specialized brain atlases. The subsequent three months will focus on benchmarking the model against traditional supervised methods, performing cross platform validation, and analyzing its performance on cell subtypes. This phased approach ensures that all computational experiments, result synthesis, and manuscript preparation are finalized for submission by the end of the six-month period.

2. Methods

2.1 Experimental design

The experimental design centers on a computational intervention wherein a pretrained single-cell foundation model (such as scGPT or Geneformer) is fine-tuned on standardized brain-specific scRNA-seq datasets, specifically the Allen Brain Atlas, to evaluate improvements in classification accuracy compared to traditional de novo supervised models. This intervention will be implemented by the research team over a six-month period using high-performance computing environments, with all model architecture, training hyperparameters, and preprocessing scripts documented to allow for full scientific replication. To minimize performance and expectancy biases, a blinded evaluation strategy will be employed where labels for the validation and test datasets are entirely withheld from the models during the inference phase, and all final performance metrics (such as Macro-F1 and accuracy) will be generated through automated, non-human-intervened scripts. Exclusion criteria for the dataset will follow standard scRNA-seq quality control protocols, specifically removing low-quality cells characterized by high mitochondrial content (e.g., >15%) or insufficient gene counts (<200) to ensure data integrity; any data failing these criteria will be excluded from the analysis rather than replaced to preserve the biological authenticity of the reference atlas.

2.2 Sample size, randomization and blinding

The study will utilize an expected sample size of approximately 500,000 to 1,000,000 single cells curated from the Allen Brain Atlas[3]. Statistical power is determined by the number of biological replicates and cell counts per cluster rather than traditional human subject counts; based on power analysis using scPower[13], a sample size of this magnitude provides >95% power to detect differentially expressed marker genes and classification improvements in rare populations representing as little as 0.5% of the total dataset. The study is powered to detect a minimum effect size of 5% improvement in the Macro-F1 score with a significant level.

2.3 Data collection, processing, and planned analysis

Data for this study will be primarily sourced from high-resolution, publicly available repositories including the Allen Brain Atlas and the Chan Zuckerberg Biohub, focusing on mammalian cortical scRNA-seq datasets[3]. Preprocessing will involve standard quality control (QC) to filter out low-quality cells, specifically those with high mitochondrial DNA content or low gene counts, followed by the specialized tokenization and gene-vocabulary mapping required for input into foundation models like scGPT or Geneformer. While the initial data acquisition will be completed within the first month, the total curation and normalization process is slated for the first three months of the project to ensure dataset harmonization across different platforms. Analysis decisions are contingent upon model convergence; if the initial fine-tuning phase yields suboptimal classification stability, the study will pivot an ensemble approach or adjust the weighting of rare cell populations to enhance model sensitivity. All analysis decisions will be documented sequentially, ensuring that adjustments to hyperparameters are made only after baseline performance benchmarks have been established and verified.

2.4 Variations from the intended sample size

While the study relies on high-quality, publicly available repositories, we anticipate potential challenges regarding data heterogeneity and missing feature metadata across different sequencing platforms. To address these variations, we will implement a robust data-harmonization pipeline that utilizes padding for missing genes and integrates anchor-based batch correction to ensure compatibility with the foundation model's fixed vocabulary. If the intended sample size of 500,000 cells is reduced due to stringent quality control (QC) filtering, we will prioritize maintaining the proportional representation of rare subtypes over raw cell counts to preserve statistical power. Furthermore, if certain datasets prove too noisy for integration, we will substitute them with alternative high-depth datasets from the Human Cell Atlas (HCA)[14] to ensure the final analysis remains sufficiently powered to detect a 5% improvement in classification accuracy.

2.5 Statistical methods

The study will primarily employ Macro-F1 scoring and Precision-Recall (PR) curves to evaluate classification accuracy, as these metrics are robust to the class imbalance inherent in brain tissue. The underlying assumption is that the gene expression distributions follow a zero-inflated negative binomial (ZINB) distribution[15], which is standard for scRNA-seq. To compare NeuroSC against baseline models, we will use the Wilcoxon rank-sum test to determine if improvements in classification accuracy are statistically significant across multiple cross-validation folds.

Software Stack:

- **Model Implementation: PyTorch** and the **Hugging Face Transformers** library for fine-tuning and hosting.
- **Data Processing: Scanpy** for data storage and QC.
- **Statistical Validation: SciPy** and **scikit-learn** for calculating F1-scores and performing significance testing.
- **Visualization: Matplotlib**

2.6 Interpreting results

The interpretation of results will focus on whether the NeuroSC framework confirms the biological syntax hypothesis that pretrained models capture universal gene-regulatory rules or if brain-specific cellular identity is too niche for general foundation models. If NeuroSC achieves superior accuracy, particularly in rare interneuron populations, it will challenge traditional model and support the theory that transfer learning is the most efficient path for high resolution cell mapping. This contribution to literature would shift the focus of computational neurobiology from manual, lab-specific annotation toward a unified, automated global atlas. Furthermore, the success of this model has significant policy implications for precision medicine and drug development; by providing a more accurate baseline for normal brain cell states, it enables regulatory bodies and pharmaceutical researchers to better validate the cell-type-specific targets of new therapeutics for neurodegenerative diseases.

3. Other required information

3.1 Acknowledgements

n/a

3.2 Funding

This work is supported by ResearchHub Foundation awarded to D.P. No other specific external funding was received for this study.

3.3 Author contributions

D.P. conceived and designed the study. H.T. and M.N. contributed to study design, data collection, method development and manuscript preparation. D.P. wrote the registered report. All authors reviewed and approved the final manuscript.

3.4 Ethical Approval

n/a

3.5 Conflicts of Interest

Hien Nguyen is working for MindX Technology School and Minh Truong is working for Pythia Biosciences

4. References

[1] Z. Yao *et al.*, “A high-resolution transcriptomic and spatial atlas of cell types in the whole mouse brain,” *Nature*, vol. 624, no. 7991, pp. 317–332, Dec. 2023, doi: 10.1038/s41586-023-06812-z.

[2] D. Li, J. Ding, and Z. Bar-Joseph, “Unsupervised cell functional annotation for single-cell RNA-seq,” *Genome Res.*, vol. 32, no. 9, pp. 1765–1775, Sep. 2022, doi: 10.1101/gr.276609.122.

[3] K. Siletti *et al.*, “Transcriptomic diversity of cell types across the adult human brain,” *Science*, vol. 382, no. 6667, p. eadd7046, Oct. 2023, doi: 10.1126/science.add7046.

[4] Y. Hao *et al.*, “Dictionary learning for integrative, multimodal and scalable single-cell analysis,” *Nat Biotechnol*, vol. 42, no. 2, pp. 293–304, Feb. 2024, doi: 10.1038/s41587-023-01767-y.

[5] F. A. Wolf, P. Angerer, and F. J. Theis, “SCANPY: large-scale single-cell gene expression data analysis,” *Genome Biol*, vol. 19, no. 1, p. 15, Dec. 2018, doi: 10.1186/s13059-017-1382-0.

[6] X. Sun, X. Lin, Z. Li, and H. Wu, “A comprehensive comparison of supervised and unsupervised methods for cell type identification in single-cell RNA-seq,” *Briefings in Bioinformatics*, vol. 23, no. 2, p. bbab567, Mar. 2022, doi: 10.1093/bib/bbab567.

[7] I. Korsunsky *et al.*, “Fast, sensitive and accurate integration of single-cell data with Harmony,” *Nat Methods*, vol. 16, no. 12, pp. 1289–1296, Dec. 2019, doi: 10.1038/s41592-019-0619-0.

[8] J. Wu *et al.*, “Biology-driven insights into the power of single-cell foundation models,” *Genome Biol*, vol. 26, no. 1, p. 334, Oct. 2025, doi: 10.1186/s13059-025-03781-6.

[9] S. Baek, K. Song, and I. Lee, “Single-cell foundation models: bringing artificial intelligence into cell biology,” *Exp Mol Med*, vol. 57, no. 10, pp. 2169–2181, Oct. 2025, doi: 10.1038/s12276-025-01547-5.

[10] H. Cui *et al.*, “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.

[11] C. V. Theodoris *et al.*, “Transfer learning enables predictions in network biology,” *Nature*, vol. 618, no. 7965, pp. 616–624, Jun. 2023, doi: 10.1038/s41586-023-06139-9.

[12] C. Domínguez Conde *et al.*, “Cross-tissue immune cell analysis reveals tissue-specific features in humans,” *Science*, vol. 376, no. 6594, p. eabl5197, May 2022, doi: 10.1126/science.abl5197.

[13] K. T. Schmid *et al.*, “scPower accelerates and optimizes the design of multi-sample single cell transcriptomic studies,” *Nat Commun*, vol. 12, no. 1, p. 6625, Nov. 2021, doi: 10.1038/s41467-021-26779-7.

[14] A. Regev *et al.*, “The Human Cell Atlas,” *eLife*, vol. 6, p. e27041, Dec. 2017, doi: 10.7554/eLife.27041.

[15] D. Risso, F. Perraudeau, S. Gribkova, S. Dudoit, and J.-P. Vert, “A general and flexible method for signal extraction from single-cell RNA-seq data,” *Nat Commun*, vol. 9, no. 1, p. 284, Jan. 2018, doi: 10.1038/s41467-017-02554-5.

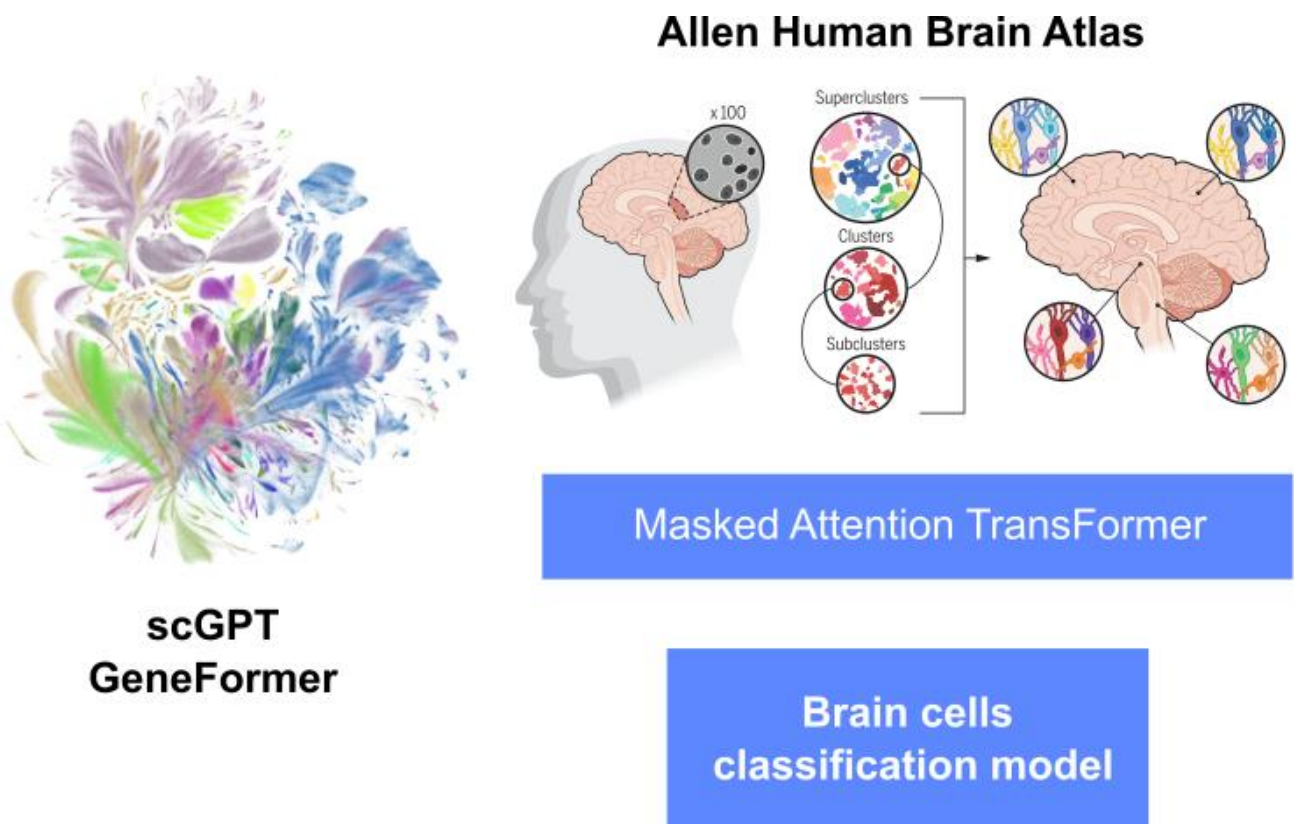

Fig:1: Graphical abstract for proposed project: 1) Single-cell foundation models for fine-tuning 2) Allen Human Brain Atlas that generated diversity of cell types across human brain, 3) Proposed masked architecture to build brain cells classification model.