



Enhancing Engineered Cell Therapies for CNS Conditions (Stage 1 Registered Report)

Sean Simonini^{1*}

Published: January 17, 2026

¹ University of Massachusetts Lowell, Lowell, MA, US

*Corresponding author: sean_simonini@student.uml.edu

ABSTRACT

Cell replacement and supportive cellular therapies for central nervous system (CNS) conditions, including neurodegenerative diseases, traumatic brain injuries, stroke, spinal cord injury, and brain aging, are limited by challenges in transplanted-cell differentiation, survival, integration, and maturation. Genetically engineering therapeutic cells has been proposed to overcome these hurdles, but genome-wide experimental screening is often logistically infeasible. This Stage 1 Registered Report proposes a computational approach to narrow a candidate dataset of genetic perturbations by leveraging publicly available large-scale perturbation screens and transcriptomic state prediction. Using CRISPRbrain survival and stress-response screens in iPSC-derived human glutamatergic neurons, iPSCs, and induced transcription factor–microglia (iTF-microglia), perturbations with beneficial phenotypes (manually annotated; $p\text{-adj} \leq 0.05$) will be selected. Perturbation-induced single-cell transcriptomic profiles will be generated in silico using a pretrained transformer-based state model (scGenePT or STATE), referencing baseline single-cell transcriptomes. Integrated profiles will be clustered, followed by pathway enrichment and LASSO-regularized logistic regression to identify enrichment patterns associated with resilient states. The output will be a ranked set of approximately 250 genes/pathways intended to inform subsequent in vitro and in vivo genetic screening for improved CNS cell therapy engraftment.

INTRODUCTION

Research Question: Background, Importance, and Relevance

Neurodegenerative diseases (NDDs), traumatic brain injuries (TBIs), stroke, spinal cord injuries (SCIs), and broader CNS insults such as brain aging often involve substantial neuronal loss. Because the brain has limited capacity to compensate for large-scale neuron death, these insults can lead to persistent impairment in basic functions, cognition, and personality. A promising approach to reverse or mitigate such deficits is exogenous transplantation of cells to replace lost neurons and/or provide neurotrophic support. However, many preclinical and clinical efforts have been hindered by limitations in transplanted cells' ability to differentiate appropriately, survive post-transplant stress, integrate into circuits, and mature.^{1–3}

Genetic engineering has been proposed, and in some contexts used, to enhance transplanted cell performance. Yet comprehensive experimental exploration of the genome is logistically infeasible for many programs, especially when the goal is to prioritize candidates for subsequent pooled screening efforts. Computational approaches that synthesize existing perturbation

screening data and predict transcriptomic effects of candidate perturbations could therefore provide a tractable path to shortlist genes/pathways likely to enhance cellular resilience and integration.

Large-scale CRISPR perturbation screening resources in neuronal and glial contexts, such as the Kampmann Lab's CRISPRbrain platform,^{4,5} provide phenotypic readouts across survival and stress paradigms. In parallel, transformer-based models that predict transcriptomic state changes after perturbation^{6–8} enable in silico generation of perturbation-induced single-cell profiles. Combining these resources may allow identification of shared transcriptomic features and pathways associated with resilience under stress conditions relevant to intracranial transplantation.

Central research question: Can machine-learning analysis of large-scale genetic perturbation datasets, including in silico transcriptomic state prediction models, be leveraged to identify perturbations that enhance neuronal cell survival, integration, and resilience in stress conditions relevant to CNS transplantation?

Hypotheses

H1 – Mechanistic/Common-Feature Hypothesis

Perturbation-derived transcriptomic profiles (via a transformer state model such as STATE) will reveal common beneficial features and resilience-associated pathways (e.g., stress response, synaptic plasticity, metabolic regulation) that are predictive of long-term resilience, functional integration, and maturation of transplanted neurons.

H2 – Predictive/Generalization Hypothesis

Machine learning models (LASSO-regularized logistic regression) trained on pathway enrichment features derived from perturbed in silico single-cell transcriptomic profiles can generalize to predict novel perturbations and combinations of perturbations that improve cellular phenotypes relevant to CNS transplantation.

Primary Outcome (Project Deliverable)

A ranked, curated list of approximately 250 genes and/or pathways prioritized for downstream pooled in vitro/in vivo screening to enhance engraftment efficacy of CNS cell therapies.

Timeline

Estimated project duration: approximately 5 months, contingent on access to required datasets and funding for computational credits.

Months 1–2: Build and validate the extraction and preprocessing pipeline; finalize screen selection; generate initial in silico single-cell transcriptomic profiles (scTPROs).

Months 2–4: Integration, batch correction, clustering, enrichment, ML feature extraction, contingency iterations.

Months 4–5: Gene ontology filtering; finalize ranked ~250 candidate list; package code and outputs for public release.

METHODS

Experimental Design

This is a computational, secondary-data analysis study using publicly available (or available-by-request) datasets; no new human or animal data will be collected. The workflow proceeds as follows: (1) screen selection and data extraction from CRISPRbrain; (2) selection of beneficial perturbations (manual annotation of sign; $p\text{-adj} \leq 0.05$); (3) generation of perturbation-induced in silico single-cell transcriptomic profiles (scTPROs) using a pretrained state model; (4) integration, dimensionality reduction, and batch correction; (5) unsupervised clustering (multi-group Leiden) plus additional binary clustering driven by entropy criteria; (6) pathway enrichment analysis; (7) ML feature extraction via LASSO-regularized logistic regression and likelihood ratio test (LRT) ranking via in silico mutagenesis; (8) gene ontology filtering and output of a curated ranked list (~250 genes/pathways).

Data Sources and Materials

CRISPRbrain Screens

Data will be obtained via the CRISPRbrain platform (Kampmann Lab, UCSF) from survival and stress-response screens of iPSC-derived human glutamatergic neurons, iPSCs, and iTF-microglia. Extracted fields include perturbation names, barcodes, perturbation type (CRISPRi/a), phenoscore, adjusted p-value, screened cell type, and assay context. Screens planned for inclusion (sign indicates whether positive phenoscores represent beneficial phenotypes): Glutamatergic Neuron-CellRox-CRISPRi (-); Glutamatergic Neuron-Day-14-Survival-CRISPRi (+); Glutamatergic Neuron-Day-21-Survival-CRISPRi (+); Glutamatergic Neuron-Day-28-Survival-CRISPRi (+); Glutamatergic Neuron-FerRhoNox-CRISPRi (-); Glutamatergic Neuron-Liperfluo-CRISPRi (-); Glutamatergic Neuron-No Antioxidants-Survival-CRISPRa (+); Glutamatergic Neuron-No Antioxidants-Survival-CRISPRi (+); Glutamatergic Neuron-PSAP KO-No Antioxidants-Survival-CRISPRa (+); Glutamatergic Neuron-PSAP KO-No Antioxidants-Survival-CRISPRi (+); Glutamatergic Neuron-PSAP KO-Survival-CRISPRa (+); Glutamatergic Neuron-PSAP KO-Survival-CRISPRi (+); Glutamatergic Neuron-Survival-CRISPRa (+); Glutamatergic Neuron-Survival-CRISPRi (+); iPSC-Survival-CRISPRi (+); iTF-Microglia-Survival-Proliferation-CRISPRi (+).

Reference Transcriptomes

Baseline single-cell expression profiles from GEO accession GSE124703 (Tian et al., 2019), specifically GSM3543622/SRR8400813 iNeuron CROP-seq data, used as a reference for STATE/scTPRO transformations.

Pretrained Models and Compute

Arc Institute's STATE framework (ST-Tahoe pretrained checkpoint) or CZI's scGenePT model (both via HuggingFace implementation). Analyses implemented in Python/R using Scanpy, Seurat, Harmony, scikit-learn, and statsmodels.

Inclusion and Exclusion Criteria

Inclusion: CRISPRbrain screens listed above will be included. Within each screen, perturbations will be included if they meet (i) beneficial phenotype direction based on manual assay-specific annotation; and (ii) adjusted p-value ≤ 0.05 . Exclusion: Perturbations failing the above criteria will

be excluded. scTPRO generation runs that fail due to model/inference errors, missing metadata, or unusable reference mapping will be excluded and rerun where feasible. Data replacement: If a perturbation is excluded due to technical failure (not biological criteria), it will be rerun; if rerun fails, it will be excluded and documented in the repository changelog.

Sample and Statistical Power

This study does not recruit participants. The effective sample size is defined by the number of in silico single-cell transcriptomic profiles (scTPROs) generated across genetic perturbations. Approximately 200 scTPROs per perturbation will be generated to support downstream clustering and statistical testing, consistent with prior machine learning-based single-cell analyses.

Statistical power is justified using an events-per-variable (EPV) framework. For the primary predictive analysis (LASSO-regularized logistic regression), an EPV threshold of ≥ 20 will be maintained to reduce overfitting and ensure stable coefficient estimation. For unsupervised clustering analyses, the chosen scTPRO counts are intended to support robust Leiden clustering and enable assessment of cluster stability using metrics such as the Adjusted Rand Index.

Data Collection, Processing, and Planned Analysis

Step 1: Screen Selection and Extraction

Screens will be manually annotated for beneficial phenotype direction. Data will be extracted and filtered to perturbations that match the beneficial direction and meet $p\text{-adj} \leq 0.05$. Control cells from the baseline transcriptome will be cleaned and loaded.

Step 2: scTPRO Transformation

Using STATE (ST-Tahoe checkpoint) or scGenePT, baseline single-cell expression matrices will be transformed to generate predicted perturbation-induced scTPROs for each selected perturbation. Outputs will be annotated with CRISPRbrain metadata from the assay context.

Step 3: Integration and Batch Correction

Expression profiles will be normalized (including dropping zero-variance genes) and concatenated while retaining annotations. PCA will be run using Scanpy, and Harmony will be used for batch correction.

Step 4: Clustering

Unsupervised multi-group Leiden clustering will be conducted in Seurat. In addition, binary clustering will be conducted using Shannon's entropy to maximize the diversity of assays represented in clusters and the overall number of cells. The best cluster (highest entropy and number of cells) will be retained; all other cells will be aggregated into an 'other' cluster.

Step 5: Pathway Enrichment

Gene Set Enrichment Analysis (GSEA) will be run comparing the best binary cluster versus 'other.' Enrichment rankings will be saved.



Step 6: ML Feature Extraction and Ranking

A LASSO-regularized logistic regression model will be trained to predict membership in the best cluster using pathway enrichment features. A likelihood ratio test (LRT) will be run as an in silico mutagenesis screen to rank pathway enrichment patterns identified as predictive.

Step 7: Gene Ontology Selection

Top 250 genes in the enriched pathways (ranked by the ML model and LRT) will be extracted and undergo Gene Ontology (GO) analysis. Genes with GO terms relevant to neurological conditions, disease-state microenvironments, stress response, apoptosis, neuronal development, axonal guidance, synaptic integration, and related processes will be prioritized, retaining metadata including the perturbation direction in the origin screen.

Variations from Intended Sample Size

Contingent Analyses (Pre-Specified)

For H1: If Adjusted Rand Index (ARI) < 0.75 (less stable clusters), then the number of scTPROs generated per perturbation will be increased and additional control cells included; PCA parameters will be retuned. If pathway discovery returns fewer than 5 significant pathways per assay at FDR ≤ 0.10 , then rank-based enrichment (AUCell/GSVA) will be substituted.

For H2: If Events per Variable (EPV) < 20, indicating potential overfitting, then the number of features will be reduced, assays collapsed or removed, and/or gene sets restricted to top-percent high-variance genes.

Pilot Data

At the time of this Stage 1 submission, the study is in early pipeline-layout and preparation stages. No pilot data have been generated. The computational approach is novel and the study is designed as a first-in-kind application of in silico perturbation modeling to CNS cell therapy target prioritization.

Statistical Methods

Binary cluster membership will be modeled using regularized logistic regression (LASSO). Cells will be classified according to whether they belong to the entropy-optimized best cluster or to the aggregated comparison cluster, using pathway enrichment scores as predictors. Regularization will be applied to reduce overfitting, and model fitting will follow standard cross-validation procedures. To rank the contribution of individual pathway features, likelihood ratio testing will be performed by comparing models with and without each feature of interest.

Software: Scanpy, Seurat, Harmony, scikit-learn, scGenePT/STATE (model and checkpoint version identifiers to be logged and reported in the Stage 2 report).

Interpreting Results

If the analyses identify consistent pathways and gene sets associated with resilient-state clusters (supporting H1), these findings will be interpreted as candidate molecular programs that may enhance transplanted-cell stress resilience and downstream integration/maturation. If the LASSO model generalizes to predict additional perturbations or combinations (supporting H2), it

will suggest that pathway-level signatures can be used to accelerate prioritization of edits for downstream pooled screening.

If hypotheses are not supported (e.g., unstable clusters, weak enrichment, poor predictive performance even after contingency steps), results will be interpreted as evidence that current perturbation-to-state predictions and/or the selected screen contexts are insufficient to robustly infer transferable resilience programs, motivating revision of reference datasets, model choice, or assay selection in future work.

Dissemination of Information

Updates on progress (code changes, protocol changes, intermediate results) will be posted monthly on ResearchHub. All raw code, requirements, data pulls (where permitted), and final outputs will be made publicly available for sharing with the scientific community.

Study Status

At the time of writing, the study has not started. The project is in early pipeline-layout and preparation stages.

OTHER REQUIRED INFORMATION

Acknowledgements

The author thanks Dr. Sarah Rozelle of UMass Lowell for her advice on the formation of this project.

Funding

This work is supported by the ResearchHub Foundation through the 2025 “Bioinformatics Request for Proposals” awarded to S.S.

Author Contributions

S.S. conceived and designed the study, will perform the analyses, and write the manuscript. All authors read and approved the final manuscript.

Corresponding author: Sean Simonini, University of Massachusetts, Lowell, Lowell, MA. Email: sean_simonini@student.uml.edu.

Ethical Approval

No human or animal subjects will be utilized. All data are de-identified and derived from in vitro experimental assays; ethics approval is not required.

Registration

This study is not a clinical trial and does not require trial registration. A preregistration of this protocol will be posted on ResearchHub prior to commencement of data collection.

Conflicts of Interest

The author declares no conflicts of interest.



REFERENCES

1. Kim TW, et al. TNF-NF- κ B-p53 axis restricts in vivo survival of hPSC-derived dopamine neurons. *Cell*. 2024;187:3671–3689.e23.
2. Bernstock JD, et al. Neural stem cell transplantation in ischemic stroke: A role for preconditioning and cellular engineering. *J Cereb Blood Flow Metab*. 2017;37:2314–2319.
3. Gowing G, Svendsen S, Svendsen CN. Ex vivo gene therapy for the treatment of neurological disorders. In: *Progress in Brain Research*. 2017;230:99–132.
4. Tian R, et al. CRISPR Interference-Based Platform for Multimodal Genetic Screens in Human iPSC-Derived Neurons. *Neuron*. 2019;104:239–255.e12.
5. CRISPRbrain | Compare Simple Screens. <https://crisprbrain.org/simple-screen-compare/>.
6. Tian R, et al. Genome-wide CRISPRi/a screens in human neurons link lysosomal failure to ferroptosis. *Nat Neurosci*. 2021;24:1020–1034.
7. Adduri AK, et al. Predicting cellular responses to perturbation across diverse contexts with State. *bioRxiv*. 2025. doi:10.1101/2025.06.26.661135.
8. Appleton E, et al. Machine-guided cell-fate engineering. *Cell Rep*. 2025;44.