Single-cell Epigenetic Analysis of Bipolar Disorder

Stephen Wu
Carnegie Mellon University
shuangw2@andrew.cmu.edu

April 2026

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

COMMITTEE

Andreas Pfenning

Submitted in partial fulfillment of the requirements
for the Senior Honors Thesis

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Keywords: single-cell genomics, machine learning, epigenetics, bipolar disorder.

1 Abstract

Bipolar disorder, a mental illness marked by drastic shifts in the patients' energy levels, severely interferes with patients' daily lives. While treatment exists, they are non-specific due to the lack of understanding of bipolar disorder's genetic regulatory pathways. This project aims to explore the epigenetic cause of bipolar disorder, especially its connection with Ca^{2+} ion level oscillations in cells, to eventually aid the development of potential RNA therapies specific to bipolar disorder. The project is carried out with two aims. The first aim attempts to associate cellular Ca^{2+} level regulation of specific cell populations with bipolar disorder through analyzing mouse brain single-cell RNA-sequencing (scRNA-seq) and human GWAS data using computational tools. The second aim is training TACIT (Kaplow et al. [2023]), a machine learning pipeline, to predict the open chromatin regions of the genome across multiple species. The pipeline would then be used to predict chromatin accessibility of the list of genes identified in the first phase and associate gene regulatory patterns with bipolar disorder traits. Our initial results show that there is a specific cell population in the hypothalamus that is enriched in Ca^{2+} related genes and statistically associated with bipolar disorder. And through training on mouse hypothalamus single-nucleus ATAC-sequencing (snATAC-seq) data, we produced cell-type-specific models that perform well on predicting chromatin accessibility in hypothalamus cell types.

2 Acknowledgments

This project is primarily supported by Dr. Andreas Pfenning and the Pfenning Lab. I would like to thank Dr. Andreas Pfenning for his support and guidance throughout the course of this project. I would like to thank Rajee Ganesan for her thorough mentorship on various aspects of the project and Heather Sestili for her exhaustive support on TACIT model training. And I would like to thank members of the Neurogenomics Lab for a positive and supportive environment where amazing projects sprouts.

I would also like to thank our collaborators: Dr. Sarah Ross for her insight on the relationship between Ca^{2+} level regulation in the brain and bipolar disorder, and Dr. Christopher Gregg, Elliott Ferris, and the Gregg Lab for providing the data used in this project.

Last but not the least, I would like to thank my family and friends for their continuous support, which gives me the strength to overcome all obstacles, as well as Carnegie Mellon Fencing Club–for keeping me physically and mentally healthy throughout my undergraduate years.

3 Introduction

3.1 Epigenetics and Enhancers

All biological entities are composed of cells, each of which carries a copy of DNA genome, serving as the blueprint for how each cell functions. Parts of the genome encodes functional units called genes, but there are also other parts in the genome that regulates the expression of these genes. It is hypothesized that the difference in regulation and expression of genes gives rise to the development of various different cell types. The study of genomic regulations that alters the expressions of these genes is called epigenomics.

Gene expression in eukaryotes is orchestrated not only by genomic regions called promoters directly adjacent to coding sequences, but also by distal cis-regulatory elements known as enhancers. Enhancers are noncoding sequences that are believed to contact the target promoters by chromatin looping during transcriptional activation, regulating the expression of target genes transcribed by RNA polymerase II (Panigrahi and O'Malley [2021]). Enhancers are found primarily in intergenic and intronic regions, consisting of dense clusters of the transcription factor binding sites and later bound by cell type-specific transcription factors, coregulators, chromatin modifiers and other proteins and enzymes (Panigrahi and O'Malley [2021]). Given that enhancers are key effectors of gene expression regulation during differentiation, we aim to study how activity of specific enhancers can influence complex traits, such as the onset of psychiatric diseases like bipolar disorder.

3.2 Cell Type Heterogeneity of the Brain and Single-Cell Genomics Technologies

The human brain is one of the most complex organs in the human body. It is composed of various different types of cells that can be broadly categorized into neurons and glial cells (Soorajkumar et al. [2025]). Neurons, widely regarded as the primary functional units of the brain, are responsible for the transmission of electrical signals that controls our thoughts and behaviors. But glial cells, such as astrocytes, oligodendrocytes, and microglia, are equally important in maintaining the homeostasis of the brain. Previous studies have shown that abnormalities in glial cells like astrocyte and microglial is associated with psychiatric disorders like major depressive disorder (Carrier et al. [2022]). Thus, pinpointing the epigentic origin of diseases takes careful examination the cell-type-level contribution of different cells composing the brain.

To profile the gene regulations of cell populations at single-cell resolution, we used single-cell genomics technologies. Traditional bulk-sequencing technologies involves breaking all cells in a tissue and profiling the aggregated

regulatory signals of genes across multiple cell populations. However, many genetic conditions and disease results from the mutations of gene regulatory elements of a small population of cells. Bulk-sequencing technologies, in this case, would diminish the mutation signal in the data, while single-cell technologies, with their abilities measure gene regulation of each single cells, allows genes and cell populations associated with diseases to be more clearly identified (Li and Wang [2021]). The main single-cell genomics technologies we are going to use are single-nucleus RNA-sequencing (snRNA-seq) and the Assay for Transposase-Accessible chromatin with sequencing, at a single nuclei or cell level (snATAC-seq).

snRNA-seq creates expression profiles of individual cells, and patterns of subsequent gene expression can be identified computationally through clustering analyses. This allows us to analyze the expression of RNA from large, mixed cell populations, and identify critical differences between cell types within a given microenvironment. Alternatively, snATAC-seq can be used to identify cis-regulatory elements controlling cell-type specific gene expression patterns. This allows us to understand cell types and states in samples with large amounts of heterogeneity, as well as subsequent gene regulatory mechanisms. By completing an integrative analysis using both snRNA-seq and snATAC-seq data, we are able to understand the cellular states at which gene regulation and expression occurs in complex traits, such as bipolar disorder.

3.3 Bipolar Disorder and Project Motivation

Bipolar Disorder (BD) is a mental illness that is marked by dramatic shifts in the patient's energy levels, affecting mood, activity levels, and concentration (Nierenberg et al. [2023]). Patients experience sudden changes from a highly energetic and excited state to a depressed and low-energy state, which significantly interferes with daily tasks. Current treatment for BD, though present, mainly targets symptoms in BD that are also prevalent among many psychiatric disorders. For example, mood stabilizers (Li^+) as a common prescription mainly serves to alleviate extreme mood shifts, but only around 30% of the patients are considered "good responders" to the treatment (Zafrilla-López et al. [2024]); and antidepressants are suspected to have the risk of inducing mania episodes and triggering affective switches between mood states more frequently (Oliva et al. [2025]). The absence of clinically approved treatments that targets BD-specific genetic pathways indicates that the underlying epigenetic cause for BD has not been fully understood. An exploration into the genetic origin of BD could provide insights for development of more effective treatments against BD, such as gene therapies, by targeting cell-type specific expression.

This project aims to find potential epigenetic regulators in the brain that is associated with BD through analyzing single-cell genomics data using bioinformatics and machine learning. The main motivation of the project came from a potential connection between Ca^{2+} ion level oscillation in cells and symptoms of BD suggested by our collaborator (Dr. Sarah Ross from University of Pittsburgh)'s preliminary data. Upon investigating the role of our candidate gene, PlcZ1, in regulation of brain Ca2+ levels and finding no significant connections, we searched for new candidates in two directions: 1) Find cell populations that are enriched in both Ca^{2+} regulating gene sets and are significantly associated with BD; 2) Train cell-type-specific machine learning models to identify enhancer regulation differences between BD and healthy samples through predicting enhancer chromatin-openness.

4 Results

4.1 PlcZ1 showed little association with BD and Ca^{2+} regulaion

We first explored the association between PlcZ1 and Ca^{2+} regulation of the brain. PlcZ1 is selected as our first candidate gene, as it regulates Ca^{2+} oscillation in non-neuronal cells (Igarashi et al. [2007])–we would like to investigate whether it carries the same regulatory function in the brain. A list of genes associated with bipolar disorder is provided by a GWAS study (O'Connell et al. [2025]), and two methods are applied to find genes associated with Ca2+ regulation using single cell RNA sequencing (scRNA-seq) data from human and mouse brains (which reflects the expression of genes in cells of the brain).

First, mouse and human whole brain scRNA-seq data is analyzed to find genes co-expressed with PlcZ1. By fitting a linear model to the gene expression data using LimmaVoom, a list of genes significantly coexpressed with PlcZ1 is found (see **Figure 1.c**). However, none of them overlaps with the BD-associated genes identified by the GWAS study. Next, we aimed to identify specific brain regions or cell types that differentially express PlcZ1, and we planned to perform differential expression (DE) analysis to identify genes that have significantly different expression levels in these cell types or brain regions. After analyzing mouse and human whole brain scRNA-seq datasets with Seurat, although we found in some samples that PlcZ1 has higher expression in the hypothalamus, the overall expression level of PlcZ1 in the brain is low. Further analysis of the human and mouse hypothalamus scRNA-seq datasets proved this by confirming that overall expression of PlcZ1 in the hypothalamus is relatively low as well (See **Figure 1.a and 1.b**). And although PlcZ1 expression is relatively higher in astrocytes and splatter neurons, they cannot be seen as significant cell types due to the overall low expression of PlcZ1 and the heterogeneous nature of the splatter neurons.

4.2 Hypothalamus cell population enriched in Ca^{2+} regulation and associated with BD

As our preliminary results showed that PlcZ1 is not associated with BD or Ca^{2+} regulation in the brain, we searched for other potential cell populations or genes connecting BD with brain Ca^{2+} level regulation. And we focused our search on the hypothalamus due to its strong association with BD shown in previous studies: hypothalamus is a part of the HPA-axis, which has strong association with BD (Belvederi Murri et al. [2016]).

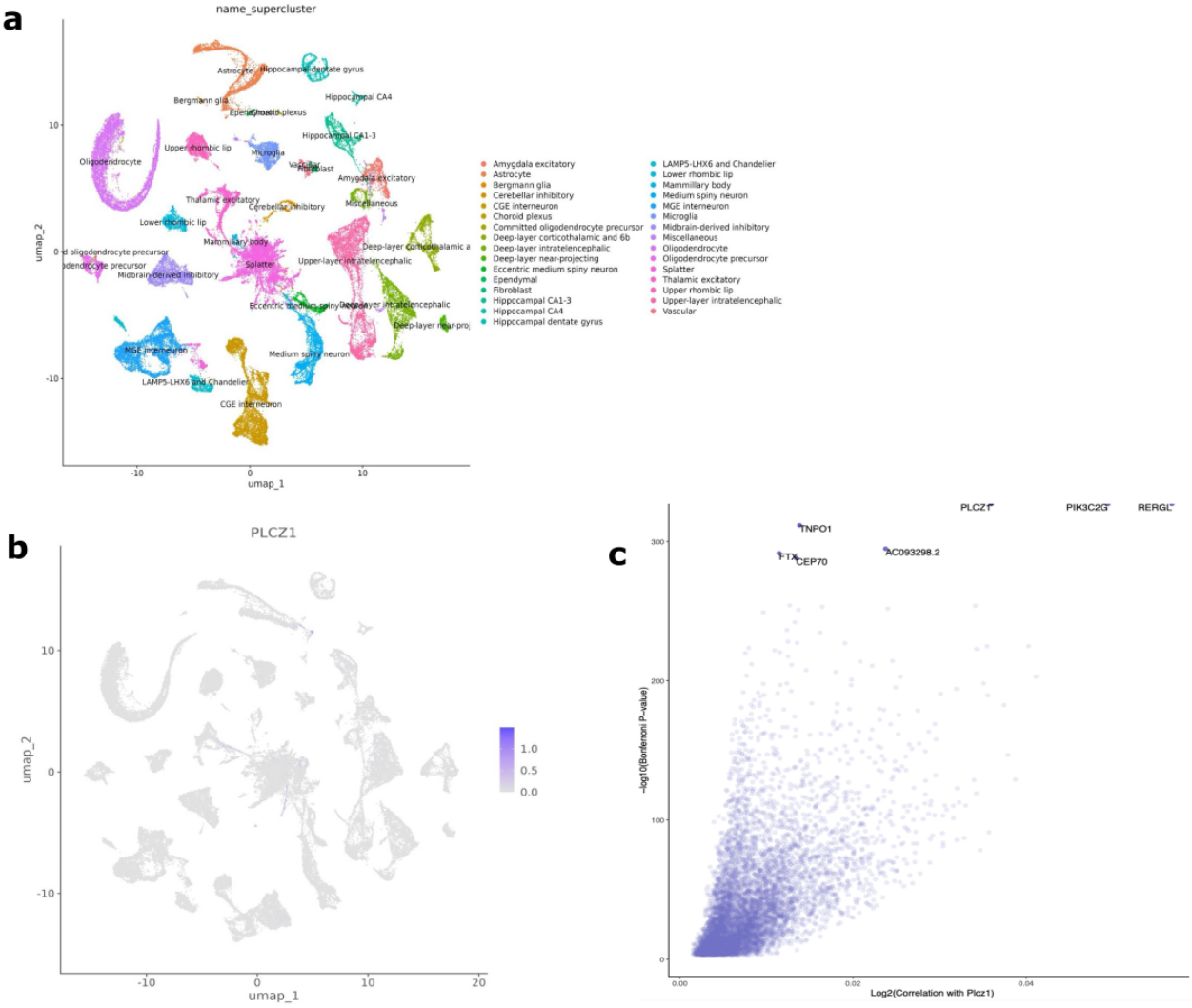

Figure 1 | PlcZ1 BD association and Calcium ion regulation results. a) UMAP of cell types in the mouse whole brain from Allen Institute (each dot is a single cell, and differently colored clusters on the plot represent different cell types in the mouse brain); **b)** Feature plot of PlcZ1 expression in the mouse whole brain (data from Allen Institute)--the more purple the color of a dot, the higher the expression of PlcZ1 in that cell; **c)** Co-expression analysis of PlcZ1 with other genes using Limma-Voom (each dot on the plot is a gene, and genes on the top right are genes more highly co-expressed with PlcZ1).

4.2.1 Methods Overview

To search for cell population candidates that could connect BD with Ca^{2+} regulation in the hypothalamus, we used a snRNA-seq data of the mouse hypothalamus (Ferris et al. [2025]) and applied two measures on to each cell: statistical association with BD and enrichment of Ca^{2+} regulatory gene sets.

To calculate the statistical disease association with BD, we applied Single-Cell Disease Relevance Score (scDRS) (Zhang et al. [2022]) on our dataset. scDRS is a bioinformatics pipeline that quantifies the statistical association of genes and cells to a disease trait in the context of each cell. It first converts Genome Wide Association Studies (GWAS) summary data, which is a method that associates specific gene variants to a disease trait, to gene-level z-score using MAGMA (de Leeuw et al. [2015]). Then, the genes with top 1000 z-scores are selected as putative disease genes. Next, the expression of putative disease genes are quantified in the context of each cell using a score weighted by the z-scores and inversely weighted by gene-specific technical noise values in the snRNA-seq dataset. After computing these "disease scores", the algorithm generates 1000 sets of cell-specific control scores by generating Monte Carlo samples of control gene sets with matching gene set size, mean expression, and the expression of putative disease genes to those of the original cells. Finally, a cell-level p-score is computed for each cell by normalizing the disease scores and control scores (Zhang et al. [2022]).

To measure the gene set enrichment of Ca^{2+} regulatory gene sets in each cell, AUCell is used. AUCell is a bioinformatics package that takes in a snRNA-seq dataset and a set of genes and outputs a score that reflects the degree that the gene set shows significant expression in each cell in the dataset. The algorithm behind this package works by raking the genes by their expression levels in each cell and define a set of top-ranked genes (which is where the enrichment will be covered). It then builds a recovery curve with the ranked genes on the x-axis and cumulative number of genes in the input gene set covered so far on the y-axis. The area-under-curve of this recovery curve for each cell will be its AUCell gene set enrichment score.

After calulating the scDRS and AUCell score for each cell in the mouse hypothalamus dataset, PCA, clustering, and UMAP algorithms are applied to group cells with similar gene expression profiles into distinct cell populations. And the scDRS score and AUCell scores are visualized on the UMAP plot and compared to search for any cell population that are high in both scores.

The dataset we used is a snRNA-seq result of mouse hypothalamus (Ferris et al. [2025]). After quality control and filtering, the gene expression profiles

of 89036 cells are preserved. For AUCell analysis, we used four gene sets regulating Ca2+ levels: calcium-mediated signaling, positive regulation of cytosolic calcium ion concentration, store-operated calcium channel activity, and intracellular calcium ion homeostasis.

4.2.2 Results

Through this analysis pipeline, we found a specific cell population that showed high scDRS score and AUCell score (see **Figure 2a, 2b, 2c**). In the UMAP corlored according to scDRS score, a small population of cells showed significantly higher scores than others. And the same population of cells showed high AUCell scores for gene sets responsible for calcium-mediated signaling and positive regulation of cytosolic calcium ion concentration. And for gene sets regulating store-operated calcium channel activity and intracellular calcium ion homeostasis, we found that they are not differentially expressed in this dataset (see **Figure 3**). In the histograms produced by AUCell (with AUCell score on the x-axis and number of cells within that range AUCell scores on the y-axis), a differentially expressed gene set would produce a bi-modal distribution, reflecting that there exist a population of cells that high express this gene set and other cells have low expression of this gene set. We can see that the gene sets responsible for calcium-mediated signaling and positive regulation of cytosolic calcium ion concentration (**Figure 3a, 3b**) shows a bi-modal distribution in their histograms, while the other two gene sets (**Figure 3c, 3d**) do not. This result suggests that this cell population (we will refer to it as "cell population of interest") in the hypothalamus is both correlated with BD and differentially expressing some gene sets regulating neuron Ca2+ levels, which provides us with a promising candidate for further investigation into the connection between Ca2+ regulation in the hypothalamus and BD.

Interested in knowing the neuron subtype of that cell population, a label transfer was performed using Seurat's MapQuery function, mapping cell-subtype labels from a finely annotated mouse hypothalamus dataset (Tadross et al. [2025]) to our current snRNA-seq dataset. The result of the label transfer is presented in **Figure 4a**. By isolating individual clusters and grouping others together, we found that our cell population of interest is of type "C19.Sst.Epha3"–a group of Somatostatin-expressing (SST) neurons marked by expression of Epha3 gene, and this label transfer result has high confidence score (see **Figure 4b**). Since SST neurons are GABAergic neurons, this contradicts with the original cell type annotation from Ferris et al. [2025], which suggests that this cell population are Glutamatergic neurons. To further elucidate the identity of this cell population, we 1) clustered our current snRNA-seq dataset into higher-level cell type clusters and perform marker

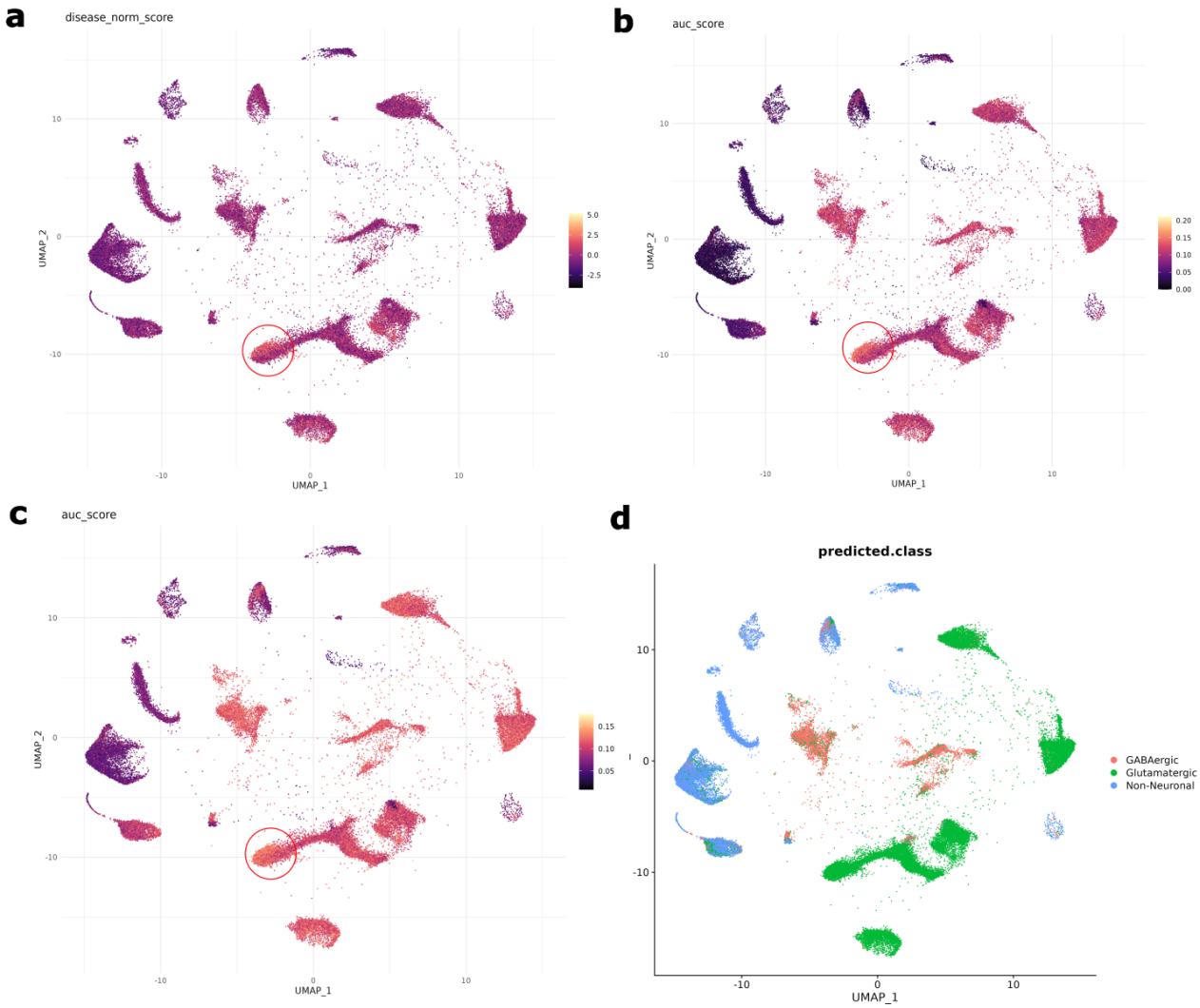

Figure 2 | scDRS and AUCell analysis of mouse hypothalamus snRNA-seq data (Ferris et al. [2025]). **a)** UMAP of scDRS analysis of the snRNA-seq data--the lighter the color of the dots, the greater the cells' statistical disease relevance with BD; **b) & c)** UMAP of AUCell analysis on the snRNA-seq data on positive regulation of cytosolic Calcium ion concentration gene sets and Calcium-mediated signaling gene sets, respectively--the lighter the color of the dots, the greater the gene set enrichment of the cells; **d)** The reference UMAP produced by Ferris et al. including the neuronal cell type labels, indicating the cell population that menifests high scores in both scDRS and AUCell analysis are glutamatergic neurons.

gene analysis on the cell population of interest to identify potential marker genes, and 2) used MapMyCell from Allen Institute to perform label transfer onto our current snRNA-seq dataset and performed marker gene analysis for the cell population of interest.

The clustering results are shown in **Figure 5**. We used Seurat to perform the clustering, and we identified the clusters that reflects our cell population of interest by transferring the scDRS and AUCell scores onto the same cells.

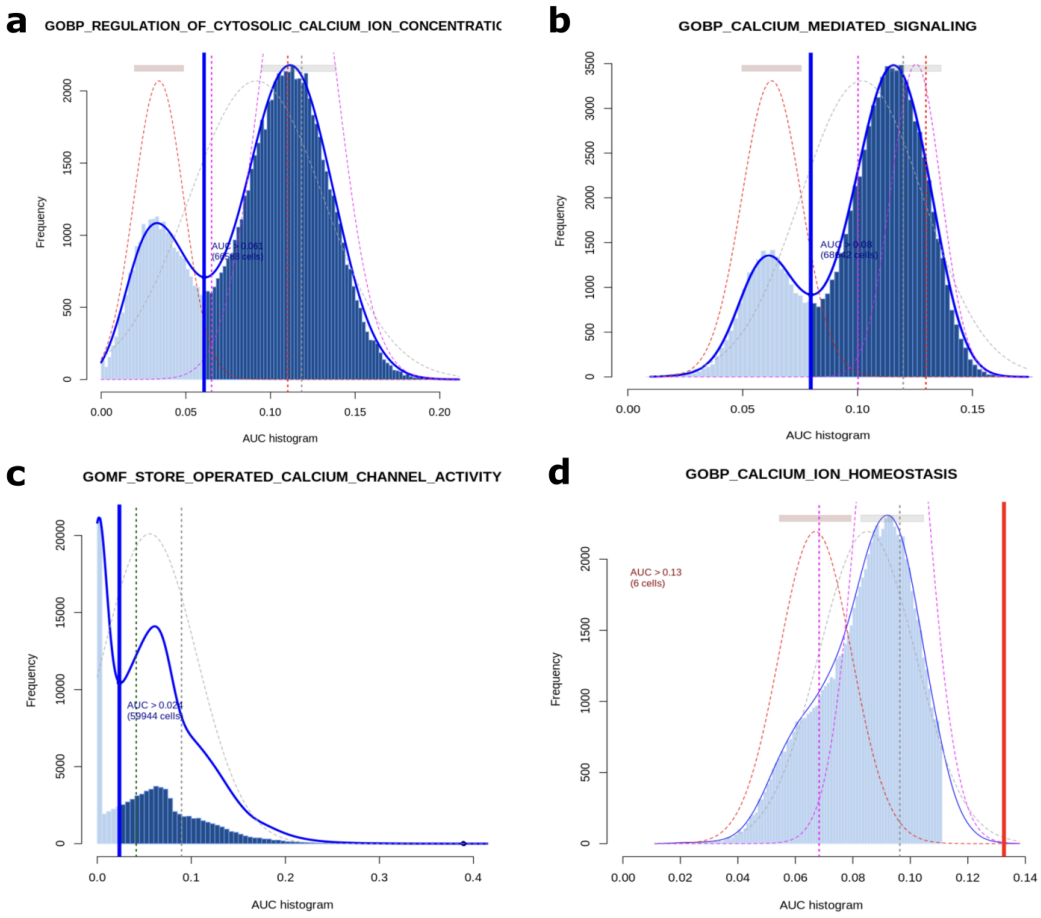

Figure 3 | AUCell histogram of Calcium ion regulation gene sets. The AUCell histograms reflects the AUCell scores across all cells in a dataset. The ideal histogram (one where the gene set is differentially expressed in a cell population) should be a bi-modal distribution. **a)** Histogram of genes regulating cytosolic Calcium ion concentration; **b)** Histogram of genes regulating Calcium-mediated signaling; **c)** Histogram of genes regulating storage-operated Calcium channel activity; **d)** Histogram of genes regulating Calcium ion homeostasis.

From the resulting UMAP reflecting scDRS and AUCell scores, clusters 2 and 11 has high scores in both measures, which represents our cell population of interest (see **Figure 5a, 5b, 5c**). Marker gene analysis was performed to identify genes representative of these two clusters against all other clusters using the Seurat FindMarkers function. Expression plots of these 5 marker genes are then examined to identify which genes more uniquely represent clusters 2 and/or 11. We found that wnt2 and sh3rf2 are more differentially expressed in clusters 2 and/or 11 (see **Figure 5d, 5e**).

To further validate our label transfer results, we used MapMyCell to perform label transfer onto our current snRNA-seq data. MapMyCell is a web tool

developed by the Allen Institute for Brain Science that uses the dataset it curated to annotate cell types for human or mouse brain single cell data. The label transfer results are shown in **Figure 4c**. Our cell population of interest is of type "CNU-LGE GABA"–GABAergic neuron at Lateral Ganglionic Eminence (LGE). This result supported the label transfer result from HypoMap that our cell population of interest are GABAergic neurons. However, since they are mapped to a cell identified from a brain region that does not belong to the hypothalamus, this result needs to be further examined. Marker gene analysis of the cell population of interest was then performed against all GABAergic neuron clusters and all other clusters. Sh3rf2 showed up again as a gene that is differentially expressed in our cell population of interest when comparing to all other GABAergic neuron clusters, as well as against all other clusters (see **Figure 4f**). Wnt2 also showed up as a gene that is differentially expressed in our cell population of interest when comparing to all other clusters (see **Figure 4e**). These two genes could be potential targets for further analysis of their connection with Ca^{2+} and BD.

4.3 Hypothalamus TACIT Model Training

Chromatin openness at enhancer regions is strongly correlated with the up-regulation of these enhancers. We aimed to train cell-type specific models for predicting chromatin openness, so that a comprehensive scan of enhancer region chromatin openness can be performed on BD and healthy samples to reveal the enhancer regulation differences at cell-type resolution. The enhancers that are regulated differently would be candidates for further investigation, as they could be the underlying cause for BD. For this project, a cell-type specific classification model is trained for each hypothalamus cell type.

4.3.1 Data preparation

The dataset used is a snATAC-seq dataset of mouse hypothalamus (Ferris et al. [2025]). We used positive sets from the reproducible Open Chromatin Regions (OCR) set detected in hypothalamus cell types within this dataset, where we removed OCRs overlapping exons and within 20,000 base pairs from a Transcription Start Site (TSS) to thoroughly exclude promoter elements, because promoters and enhancers within the same cell type have been shown to be bound by only partially overlapping groups of transcription factors. Then, we filtered this set of OCR again for enhancer regions only and scaled them to 500 base pairs each based on peak location of every OCR. Next, GC-matched negative targets to the positive set are generated by taking the parts of the mouse genome that have similar GC-content as the positive targets and are not detected by snATAC-seq. To do so, we used the biasaway tool with the following parameters: biasaway c -n 10. While 10 times GC-matched negatives are generated, we further filter them to exclude exonic and promoter

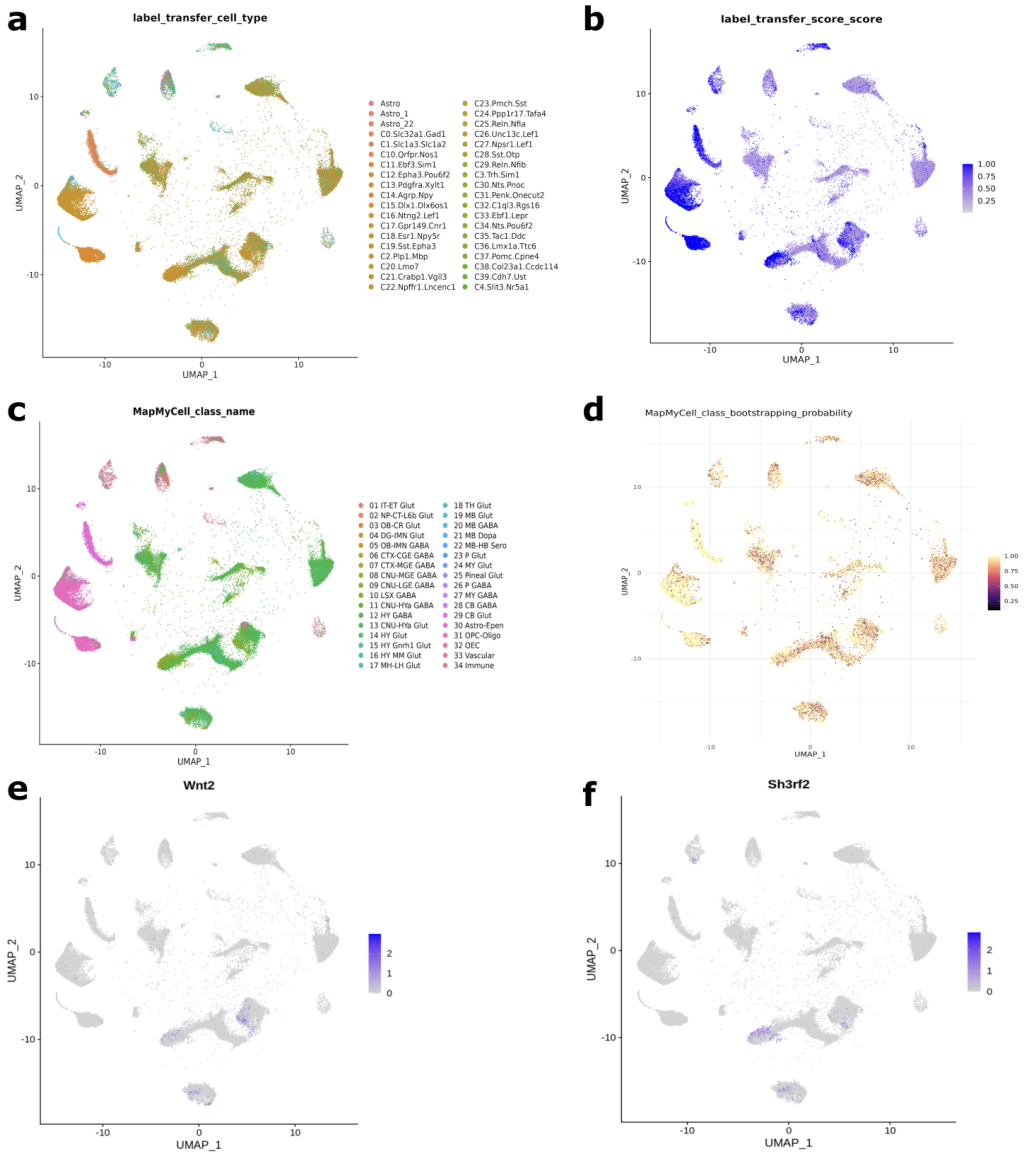

Figure 4 | Label transfer results and corresponding marker gene analysis results. a) Label transfer results from HypoMap (the cell population of interest is annotated as C19.Sst.Epha3, more cell type coloring labels are hidden to prevent over-crowding of the figure); **b)** Confidence score of the label transfer from HypoMap (the more purple the color of a dot, the greater the confidence of its label transferred cell type labeling); **c)** Label transfer results from MapMyCell (the cell population of interest is annotated as CNU-LGE GABA); **d)** MapMyCell label transfer result confidence score UMAP (the lighter the color of a dot, the greater the confidence of its label transferred cell type labeling); **e) & f)** Expression plots of Wnt2 and Sh3rf2 on the UMAP clustered by Ferris et al.--these two genes are differentially expressed in the cell type of interest.

regions and regions overlapping any reproducible or non-reproducible OCR of the target cell type, resulting in 4-5 times as many GC-matched negatives as positives. This negative set aims to push models to learn to identify random non-coding regions in the genome that are inactive, but have a similar GC-content to our training set. We want to make sure the GC-content of the

positive and negative targets are similar, so that the models are not learning GC-content as a measure of enhancer openness.

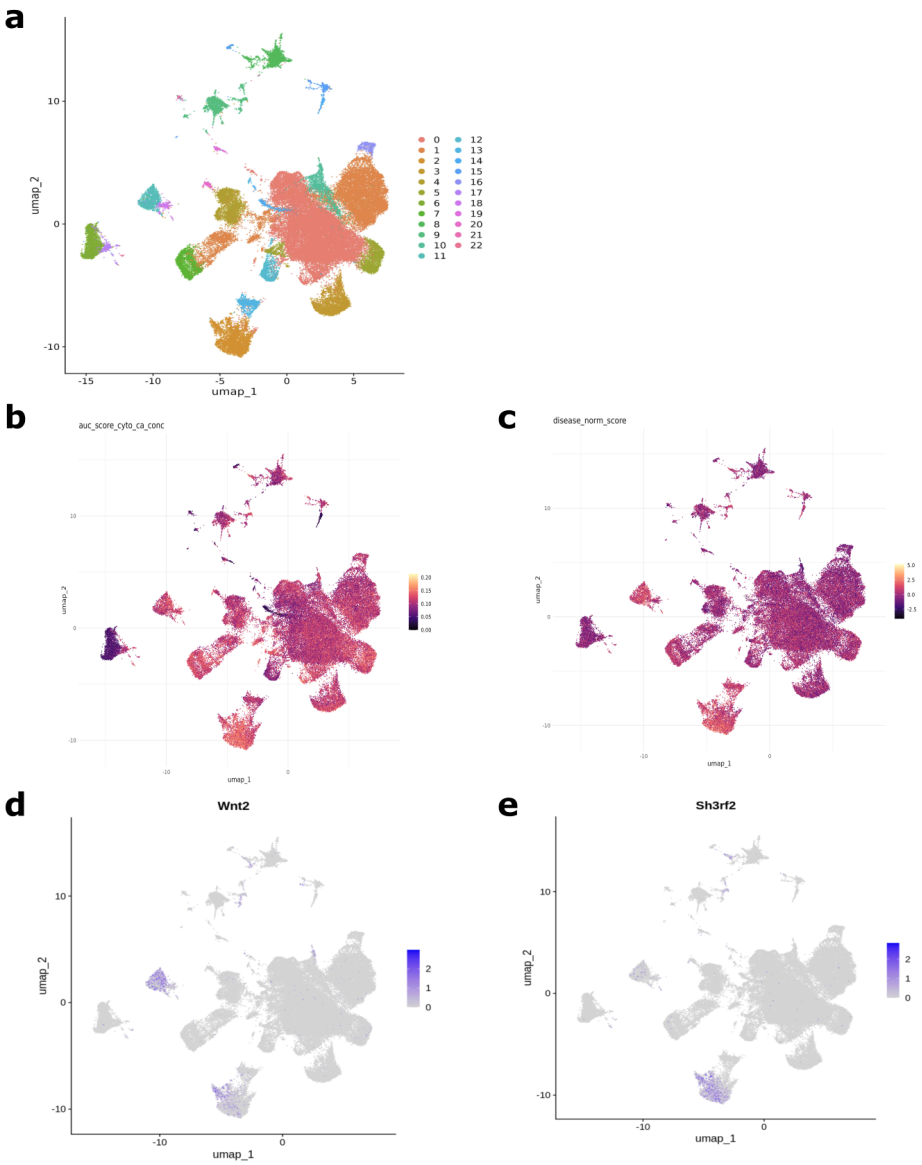

Figure 5 | Clustering and marker gene analysis results. a) Clustering result of the Ferris et al. [2025] snRNA-seq data using Seurat (clusters 2 and 11 represents the cell population of interest identified in scDRS and AUCell score mapping); **b)** AUCell score (regulation of cytosolic Calcium ion concentration) mapping onto corresponding cells of the clustering result; **c)** scDRS score mapping onto the corresponding cells of the clustering result; **d) & e)** Expression plot of Wnt2 and Sh3rf2 genes on Seurat-clustered UMAP--these two genes are differentially expressed in clusters 2 and 11.

4.3.2 Model architecture

The first set of models trained are TACIT (Tissue-Aware Conservation Inference Toolkit, Kaplow et al. [2023]) models. TACIT is a CNN pipeline developed in the Pfenning Lab that takes in a fixed length of genomic se-

quences and outputs the openness of that genomic region. The inputs were one-hot encoded, 500 base pair DNA sequences, and the outputs varied based on whether we trained a classification or regression model: classification model outputs were a probability within [0,1] that the input OCR was accessible, and regression model outputs were a continuous-valued number greater than or equal to zero, representing the predicted strength of the ATAC-seq peak signal. TACIT uses a standard CNN architecture: The input DNA sequence is first converted to a one-hot encoding that is then passed to a number of 1D convolutional layers for capturing local regulatory motifs; after that, a max-pooling layer and a number of dense layers are applied for feature selection and further transformations, and lastly an output layer is added (See **Figure 6a**). The pipeline was originally developed for predicting chromatin openness across species, and this project adapts the pipeline in predicting chromatin openness across different cell types in the hypothalamus. In this part of the project, we trained TACIT models on classification targets.

4.3.3 Results

The models are trained in default configuration (2 convolutional layers and 1 dense layer), and cyclic learning rate (LR) and momentum are used to accelerate training time. We trained one model each for the eight major cell types in our snATAC-seq dataset: Oligodendrocyte, OPC, GABAergic Neuron, Glutamatergic Neuron, General Neuron, Microglia, Endothelial Cell, and Astrocyte. These models are evaluated across folds for overall performance on the validation set using the area under the receiver-operator curve (AUROC) and area under the Precision-Recall curve (AUPRC). When evaluating the models on the validation set, we observed that the models' AUROC ranged from 0.81 to closer to 1.0 over the 8 cell types, and the validation AUPRC ranged from 0.50-0.83. The Glutamatergic neuron model and GABAergic neuron model performed exceptionally well, obtaining validation AUROC of above 0.95 and validation AUPRC of above 0.8 (see **Figure 6b**). The overall positive training result lays the foundation for our next step—training TACIT models on finer cell types of the hypothalamus.

4.4 Contrastive-TACIT Model Training

Besides training the traditional TACIT model, we also designed a "contrastive-TACIT" architecture—modifying the architecture of TACIT so that it can take in two sequences and predict the relative openness between the two sequences. The aim of the new design is to allow the model to more accurately reflect the relative openness differences in comparative tasks—in the context of this project, it could be used to compare the openness of the same enhancer regions across BD and healthy samples.

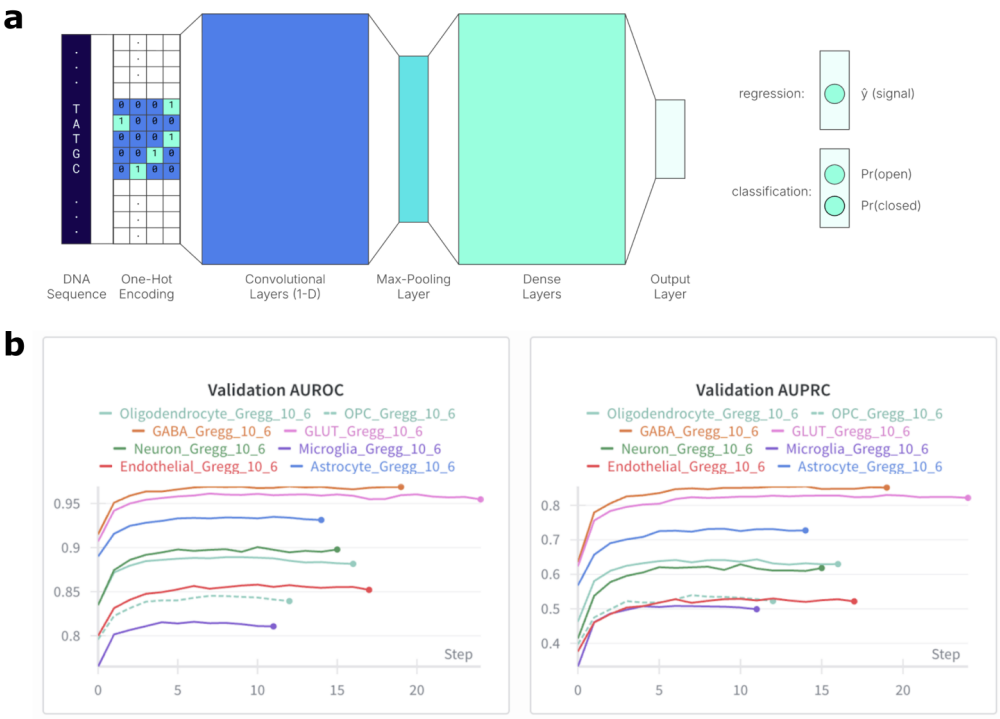

Figure 6 | Hypothalamus cell-type-specific TACIT model training. a) Architecture of TACIT (Figure adapted from Kaplow et al. [2023]); **b)** Hypothalamus cell-type-specific classification model training validation AUROC and AUPRC--models trained on GABAergic neuron and Glutamatergic neuron data performed exceptionally well.

4.4.1 Model architecture

The architecture of the original TACIT model (Kaplow et al. [2023]) was modified by stacking the one-hot encodings of two input sequences on a new dimension and applying a 2D convolutional layer right after that, leaving other parts of the model architecture unchanged (**Figure 7a**). With this design, our intention was to allow the model to capture local motif differences between the two sequences and examine if that can help predict the relative openness between two genomic regions.

4.4.2 Dataset Preparation

The modified architecture was trained as a regression model on mouse hypothalamus GABAergic neuron snATAC-seq data—two sequences are randomly selected from a pool of "open" sequences and "close" sequences, and the regression target is the log-fold change of the snATAC-seq openness measure between the first and second sequence (the openness measures for the negative sequences are 0). The loss function for model training is Mean Squared Error (MSE).

Besides this dataset, we also prepared an ortholog-matching dataset, as our lab member's preliminary results showed that training TACIT models on ortholog matching data significantly improved their performance. This dataset was generated by a 2-step filtering process: 1) take Open Chromatin Regions (OCR) in Human single-cell ATAC-sequencing (scATAC-seq) data and filter for the OCR peaks that overlaps with Human orthologs; 2) for every Human OCR peak, find Macaque OCR peaks from Macaque scATAC-seq data that maps to the Human ortholog that overlaps with the human OCR peak. These pairs of Macaque-Human peaks will be the input to the regression model, and the output will be the log-fold change of the snATAC-seq openness measure between the Human and Macaque sequence.

4.4.3 Results

Training results on the mouse hypothalamus GABAergic neuron dataset is shown in **Figure 7b**. Although training loss gradually decreases, validation losses over training epochs shows instability. To examine whether this instability was the result of inappropriate learning rate, a hyperparameter sweep across initial learning rates was performed(see **Figure 7c**). Initial learning rates are sampled uniformly between 6×10^{-5} and 5×10^{-3} and the model was trained on the GABAergic neuron data. Final validation loss was recorded. From the result, higher learning rates seemed to give lower validation loss, but the validation losses across epochs still changes unstably, suggesting that this instability in validation loss change was not a result of inappropriate learning rate selection. The immediate next step would be to train the model on the ortholog matching dataset and observe if validation loss still exhibits instability.

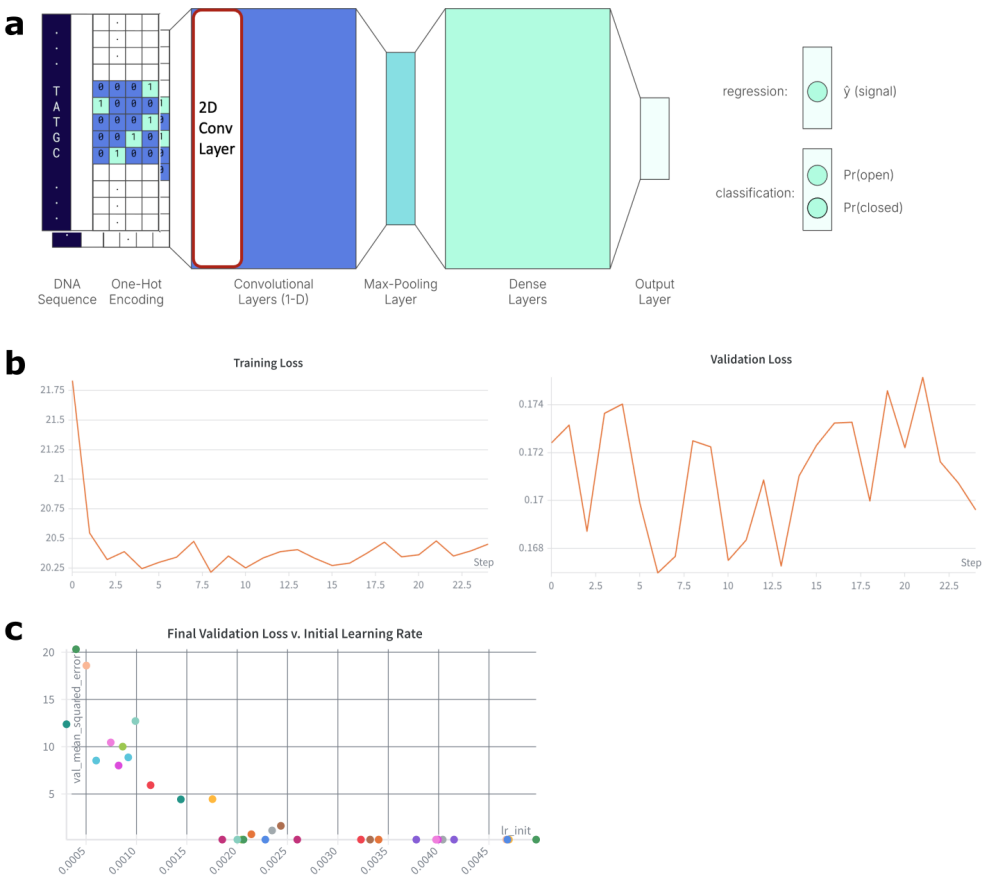

Figure 7 | Contrastive-TACIT regression model training. a) The architecture of contrastive-TACIT model--it was modified from the original Kaplow et al. [2023] design by stacking two input sequence embeddings and replacing the first 1D convolutional layer with a 2D convolutional layer; **b)** Training and validation loss of contrastive-TACIT model training on GABAnergic neuron snATAC-seq data (initial learning rate = 0.0039), training loss shows gradual decrease but validation loss shows instability; **c)** Hyperparameter sweep result on initial learning rate--higher learning rate seems to give lower validation loss.

5 Discussion and Limitation Analysis

The preliminary investigation into PlcZ1 revealed that, despite its established role in regulating Ca^{2+} oscillations in non-neuronal cells, it does not appear to carry out a parallel function in the brain. Its overall low expression across both the whole-brain and hypothalamus-specific scRNA-seq datasets, and the absence of overlap between its co-expressed genes and BD-associated GWAS genes, collectively suggest that PlcZ1 is unlikely to have a significant correlation with BD or Ca^{2+} regulation in the brain. These results justify the pivot toward a broader, unbiased search for Ca^{2+}-regulating cell populations associated with BD.

Through disease relevance calculation and gene set enrichment analysis, we found a hypothalamic cell population simultaneously exhibiting elevated scDRS scores for BD and enriched AUCell scores for Ca^{2+}-regulating gene sets–specifically calcium-mediated signaling and positive regulation of cytosolic calcium ion concentration. This double enrichment provides preliminary evidence that this cell population might reveal a cell-type specific loci where Ca^{2+} dysregulation intersects with BD risk. However, the cell type identity of this population remains ambiguous. The label transfer using Seurat against HypoMap yielded a GABAergic identity (C19.Sst.Epha3), while the original Ferris et al. [2025] annotation classified the region as Glutamatergic. The subsequent MapMyCell analysis supported a GABAergic identity (CNU-LGE GABA), but mapped the population to the LGE rather than the hypothalamus, which raises questions about either the accuracy of the mapping or a genuine transcriptomic similarity between this hypothalamic subpopulation and LGE GABAergic neurons. Combined, the evidence across two independent label transfer methods do lend some confidence to the GABAergic characterization, but the discordance with the original annotation suggests further validation needs to be done.

Marker gene analysis offered additional insights. Wnt2 and Sh3rf2 are consistently differentially expressed in this cell population of interest, both in comparisons against all other clusters and all other GABAergic neuron clusters. Wnt2 has established roles in dendritic development (Wayman et al. [2006]), and Sh3rf2 has been linked to hippocampal dendrite development and synaptogenesis (Wang et al. [2018]). Whether either gene has a direct functionality in Ca^{2+} regulation or BD-associated pathways remains unclear, but this result positions them as potential candidates for follow-up studies.

The cell-type-specific hypothalamus TACIT models trained on eight broad hypothalamic cell types produced encouraging preliminary results, demonstrating that the pipeline can generalize from its original cross-species design to a cross-cell-type prediction task within the hypothalamus. This result sets up the stage for training the models on finer cell types, taking us one step closer to the longer-term goal of comparing enhancer openness between BD and healthy samples at detailed cell-type resolution.

While conceptually motivating, the contrastive-TACIT architecture showed instability in validation loss across training epochs that was not resolved by hyperparameter sweeps over learning rate. This suggests the instability could be architectural rather than an optimization artifact. A likely contributing factor is that randomly pairing sequences to form training examples introduces substantial noise into the regression targets: the model was tasked with

predicting relative openness between pairs (most likely non-orthologous) that may differ in ways unrelated to relative regulatory sequence content. Training the model on orthologous sequence pairs could potentially help with model learning. The 2D convolutional design for capturing local motif differences between two sequences might also not be optimal for capturing the most informative signal for relative openness prediction. Alternative formulations, such as attention-based sequence comparison, may be more appropriate.

6 Future Directions

Following the identification of the cell population in hypothalamus that has high significance in BD disease relevance and Ca^{2+} regulatory gene set enrichment, further label transfer and marker gene analysis from other more finely annotated mouse hypothalamus datasets is going to be performed to validate its identity. From there, pathway analysis and further gene set analysis on this cell population would be performed to link specific cell types with BD, allowing us to gain insight into the regulatory mechanisms influencing BD.

On the machine learning side, TACIT models would be trained on finer cell types to detect enhancer regulation differences at more detailed resolution. And besides training the models on “negative" input sequences (genomic regions where chromatin is closed) with matching GC-contents, additional training on “cell-type negatives”, which are orthologous regions that are opened in another cell type but closed in the target cell type, would also be conducted. This would push the model to capture finer regulatory differences across cell types, achieving greater cell-type-specificity. Contrastive-TACIT would also be further improved on several aspects. Designs that confer non-local difference detection would be introduced into the convolution module. We are also planning on training the models on pairs of sequences with more significant openness differences, as preliminary results from lab members suggest that small differences in openness are often not reflected on the sequence level. And after successful TACIT and contrastive-TACIT model training, a comprehensive scan of all enhancers in BD and healthy samples would be performed to identify enhancer regulation differences at fine cell-type resolution, revealing candidate epigenetic drivers for BD.

References

Martino Belvederi Murri, Davide Prestia, Valeria Mondelli, Carmine Pariante, Sara Patti, Benedetta Olivieri, Costanza Arzani, Mattia Masotti, Matteo Respino, Marco Antonioli, Linda Vassallo, Gianluca Serafini, Giampaolo Perna, Maurizio Pompili, and Mario Amore. The hpa axis in bipolar disorder: Systematic review and meta-analysis. *Psychoneuroendocrinology*, 63:327–342, 2016. ISSN 0306-4530. doi: https://doi.org/10.1016/j.psyneuen.2015.10.014. URL https://www.sciencedirect.com/science/article/pii/S0306453015009622.

Micaël Carrier, Kira Dolhan, Bianca Caroline Bobotis, Michèle Desjardins, and Marie-Ève Tremblay. The implication of a diversity of non-neuronal cells in disorders affecting brain networks. *Frontiers in Cellular Neuroscience*, Volume 16 - 2022, 2022. ISSN 1662-5102. doi: 10.3389/fncel.2022.1015556. URL https://www.frontiersin.org/journals/cellular-neuroscience/articles/10.3389/fncel.2022.1015556.

Christiaan A. de Leeuw, Joris M. Mooij, Tom Heskes, and Danielle Posthuma. Magma: Generalized gene-set analysis of gwas data. *PLOS Computational Biology*, 11(4):e1004219, 2015. doi: 10.1371/journal.pcbi.1004219. URL https://doi.org/10.1371/journal.pcbi.1004219.

Elliott Ferris, Josue D. Gonzalez Murcia, Adriana Cristina Rodriguez, Susan Steinwand, Cornelia Stacher Hörndli, Dimitri Traenkner, Pablo J. Maldonado-Catala, and Christopher Gregg. Genomic convergence in hibernating mammals elucidates the genetics of metabolic regulation in the hypothalamus. *Science*, 389(6759):494–500, 2025. doi: 10.1126/science.adp4025. URL https://www.science.org/doi/abs/10.1126/science.adp4025.

Hideki Igarashi, Jason G. Knott, Richard M. Schultz, and Carmen J. Williams. Alterations of plc1 in mouse eggs change calcium oscillatory behavior following fertilization. *Developmental Biology*, 312(1):321–330, 2007. ISSN 0012-1606. doi: https://doi.org/10.1016/j.ydbio.2007.09.028. URL https://www.sciencedirect.com/science/article/pii/S001216060701384X.

Irene M. Kaplow, Alyssa J. Lawler, Daniel E. Schäffer, Chaitanya Srinivasan, Heather H. Sestili, Morgan E. Wirthlin, BaDoi N. Phan, Kavya Prasad, Ashley R. Brown, Xiaomeng Zhang, Kathleen Foley, Diane P. Genereux, Zoonomia Consortium**, Elinor K. Karlsson, Kerstin Lindblad-Toh, Wynn K. Meyer, Andreas R. Pfenning, Gregory Andrews, Joel C. Armstrong, Matteo Bianchi, Bruce W. Birren, Kevin R. Bredemeyer, Ana M. Breit, Matthew J. Christmas, Hiram Clawson, Joana Damas, Federica Di Palma, Mark Diekhans, Michael X. Dong, Eduardo Eizirik, Kaili Fan, Cornelia Fanter, Nicole M. Foley, Karin Forsberg-Nilsson, Carlos J. Garcia,

John Gatesy, Steven Gazal, Diane P. Genereux, Linda Goodman, Jenna Grimshaw, Michaela K. Halsey, Andrew J. Harris, Glenn Hickey, Michael Hiller, Allyson G. Hindle, Robert M. Hubley, Graham M. Hughes, Jeremy Johnson, David Juan, Irene M. Kaplow, Elinor K. Karlsson, Kathleen C. Keough, Bogdan Kirilenko, Klaus-Peter Koepfli, Jennifer M. Korstian, Amanda Kowalczyk, Sergey V. Kozyrev, Alyssa J. Lawler, Colleen Lawless, Thomas Lehmann, Danielle L. Levesque, Harris A. Lewin, Xue Li, Abigail Lind, Kerstin Lindblad-Toh, Ava Mackay-Smith, Voichita D. Marinescu, Tomas Marques-Bonet, Victor C. Mason, Jennifer R. S. Meadows, Wynn K. Meyer, Jill E. Moore, Lucas R. Moreira, Diana D. Moreno-Santillan, Kathleen M. Morrill, Gerard Muntané, William J. Murphy, Arcadi Navarro, Martin Nweeia, Sylvia Ortmann, Austin Osmanski, Benedict Paten, Nicole S. Paulat, Andreas R. Pfenning, BaDoi N. Phan, Katherine S. Pollard, Henry E. Pratt, David A. Ray, Steven K. Reilly, Jeb R. Rosen, Irina Ruf, Louise Ryan, Oliver A. Ryder, Pardis C. Sabeti, Daniel E. Schäffer, Aitor Serres, Beth Shapiro, Arian F. A. Smit, Mark Springer, Chaitanya Srinivasan, Cynthia Steiner, Jessica M. Storer, Kevin A. M. Sullivan, Patrick F. Sullivan, Elisabeth Sundström, Megan A. Supple, Ross Swofford, Joy-El Talbot, Emma Teeling, Jason Turner-Maier, Alejandro Valenzuela, Franziska Wagner, Ola Wallerman, Chao Wang, Juehan Wang, Zhiping Weng, Aryn P. Wilder, Morgan E. Wirthlin, James R. Xue, and Xiaomeng Zhang. Relating enhancer genetic variation across mammals to complex phenotypes using machine learning. *Science*, 380(6643):eabm7993, 2023. doi: 10.1126/science.abm7993. URL https://www.science.org/doi/abs/10.1126/science.abm7993.

X. Li and C. Y. Wang. From bulk, single-cell to spatial rna sequencing. *International Journal of Oral Science*, 13:36, 2021. doi: 10.1038/s41368-021-00146-0. URL https://doi.org/10.1038/s41368-021-00146-0.

Andrew A. Nierenberg, Bruno Agustini, Ole Köhler-Forsberg, Cristina Cusin, Douglas Katz, Louisa G. Sylvia, Amy Peters, and Michael Berk. Diagnosis and treatment of bipolar disorder: A review. *JAMA*, 330(14):1370–1380, 10 2023. ISSN 0098-7484. doi: 10.1001/jama.2023.18588. URL https://doi.org/10.1001/jama.2023.18588.

K. S. O'Connell, M. Koromina, T. van der Veen, et al. Genomics yields biological and phenotypic insights into bipolar disorder. *Nature*, 639:968–975, 2025. doi: 10.1038/s41586-024-08468-9. URL https://doi.org/10.1038/s41586-024-08468-9.

Vincenzo Oliva, Giovanna Fico, Michele De Prisco, Xenia Gonda, Adriane R. Rosa, and Eduard Vieta. Bipolar disorders: an update on critical aspects. *The Lancet regional health. Europe*, 48:101135–101135, 2025. ISSN 2666-7762.

A. Panigrahi and B. W. O'Malley. Mechanisms of enhancer action: the known and the unknown. *Genome Biology*, 22:108, 2021. doi: 10.1186/s13059-021-02322-1. URL https://doi.org/10.1186/s13059-021-02322-1.

A. Soorajkumar, B. Balan, N. Nassir, et al. Mapping human brain cell type origin and diseases through single-cell transcriptomics. *Translational Psychiatry*, 15:349, 2025. doi: 10.1038/s41398-025-03562-6. URL https://doi.org/10.1038/s41398-025-03562-6.

J. A. Tadross, L. Steuernagel, G. K. C. Dowsett, et al. A comprehensive spatio-cellular map of the human hypothalamus. *Nature*, 639:708–716, 2025. doi: 10.1038/s41586-024-08504-8. URL https://doi.org/10.1038/s41586-024-08504-8.

Shuo Wang, Ningdong Tan, Xingliang Zhu, Minghui Yao, Yaqing Wang, Xiaohui Zhang, and Zhiheng Xu. Sh3rf2 haploinsufficiency leads to unilateral neuronal development deficits and autistic-like behaviors in mice. *Cell Reports*, 25(11):2963–2971.e6, 2018. ISSN 2211-1247. doi: https://doi.org/10.1016/j.celrep.2018.11.044. URL https://www.sciencedirect.com/science/article/pii/S2211124718318072.

Gary A. Wayman, Soren Impey, Daniel Marks, Takeo Saneyoshi, Wilmon F. Grant, Victor Derkach, and Thomas R. Soderling. Activity-dependent dendritic arborization mediated by cam-kinase i activation and enhanced creb-dependent transcription of wnt-2. *Neuron*, 50(6):897–909, 2006. ISSN 0896-6273. doi: https://doi.org/10.1016/j.neuron.2006.05.008. URL https://www.sciencedirect.com/science/article/pii/S0896627306003746.

Marina Zafrilla-López, Miriam Acosta-Díez, Marina Mitjans, Anna Giménez-Palomo, Pilar A Saiz, Carme Barrot-Feixat, Ester Jiménez, Sergi Papiol, Victoria Ruiz, Patrícia Gavín, María Paz García-Portilla, Leticia González-Blanco, Julio Bobes, Thomas G Schulze, Eduard Vieta, Antoni Benabarre, and Bárbara Arias. Lithium response in bipolar disorder: Epigenome-wide dna methylation signatures and epigenetic aging. *European Neuropsychopharmacology*, 85:23–31, 2024. ISSN 0924-977X. doi: https://doi.org/10.1016/j.euroneuro.2024.03.010. URL https://www.sciencedirect.com/science/article/pii/S0924977X24000683.

Martin Jinye Zhang, Kangcheng Hou, Kushal K. Dey, Saori Sakaue, Karthik A. Jagadeesh, Kathryn Weinand, Aris Taychameekiatchai, Poorvi Rao, Angela Oliveira Pisco, James Zou, Bruce Wang, Michael Gandal, Soumya Raychaudhuri, Bogdan Pasaniuc, and Alkes L. Price. Polygenic enrichment distinguishes disease associations of individual cells in single-cell rna-seq data. *Nature Genetics*, 54(10):1572–1580, Oct 2022. doi: 10.1038/s41588-022-01167-z. URL https://doi.org/10.1038/s41588-022-01167-z.