A data-driven approach for understanding the dynamics of neuronal migration in neocortical development

Ashwin De Silva, Sai Koukuntla, and Ime Essien

Department of Biomedical Engineering, Johns Hopkins University

1 Introduction

Dorsal pallial neurogenesis during mid-gestation is the process where a majority of excitatory neurons in the cortex are formed. There are 6 layers of neurons in the mature neocortex in mammals. The neurons of each of these layers are produced in succession, with neurons of the new layer migrating past previously created layers. As each cortical layer is produced, newborn neurons arise from the progenitor cells in the germinal zone (GZ) and migrate toward the cortical plane (CP). This is where the neurons would be established in the developing cortex.

Given this biological context, we are interested in studying the genetic basis of embryonic neurogenesis. In particular, we wish to determine whether the dynamics of neuronal migration from GZ to CP during mid-gestation are regulated by a non-linear function of gene expression. Identifying the collection of genes characterizing the neuronal activities in the germinal zone and the cortical plate would be important in understanding the mechanisms underlying early brain development.

This migration is a complex process involving the generation and differentiation of neural progenitor cells into neurons and glial cells. Several genes and molecular pathways are known to play crucial roles in this process. SOX2, PAX6, NEUROG1, NEUROG2, ASCL1, and NOTCH1 are some key genes involved in this process. The SOX2 transcription factor helps maintain neural stem cell identity and regulate neural differentiation [3]. PAX6 plays a critical role in the development of the nervous system, including the cerebral cortex and the retina [2]. NEUROG1 and NEUROG2 are proneural basic helix-loop-helix (bHLH) transcription factors that promote neuronal differentiation [6]. MASH1 or ASCL1 (Achaete-scute homolog 1) is another proneural bHLH transcription factor that is crucial for the generation of neurons and oligodendrocytes [1]. Finally, the notch signaling pathway is a highly conserved intercellular signaling mechanism that plays a critical role in cell fate determination, including the regulation of neural stem cell proliferation and differentiation [5].

In this study, we have access to a large dataset single-cell RNA-seq dataset on human embryonic neurogenesis. Assembling such a dataset is extremely difficult as it involves an invasive procedure to collect tissue from a human embryo. Our objective is to study the dynamics of neuronal migration from GZ to CP through a data-driven approach using this dataset. Specifically, we utilize standard feature selection tools to identify the genes whose expression levels are most suited for discriminating between the neurons in GZ and CP, and evaluate whether the potentially non-linear relationship between these gene expressions and the cell location can be captured by several commonly used classification algorithms. We also compare the potency of the aforementioned biologically relevant genes against those selected computationally in the task of discriminating between the two cell locations. Finally, from our data-driven findings, we conclude that neuronal migration from GZ to CP is regulated by a non-linear function of gene expression. We primarily use the `sklearn` package for our experiments and our code is available at `https://github.com/Laknath1996/neurogenesis`.

2 Dataset

The dataset contains $n = 15066$ cells (samples) and expression levels (features) from $d = 35544$ different genes. The cells are categorically labeled according to their location which is either the germinal zone (GZ) or the cortical plane (CP). The samples are well-balanced between the two categories and the dataset does not suffer from missing data. The cells are coming from different donors, gestation weeks, and libraries. While this may lead to undesirable batch effects, for the simplicity of our analysis, we neglect the impact of these meta-variables.

3 Methods

3.1 Feature Selection

Since the number of samples is significantly lower than the feature dimension ($n << d$), to avoid the curse of dimensionality we are compelled to perform a feature selection step before proceeding to the classification task. As the first step, we split the entire dataset into a train set (60%) and a test set (40%), such that the original category proportion is preserved. We hold out the test set and focus on the train set to perform feature selection. To this end, we employ the Wilcoxon rank sum test with Bonferroni correction with $\alpha = 0.05$ to identify the deferentially expressed genes with respect to the cell location. This yields a total 1139 genes whose expression levels differ significantly between the two categories GZ and CP. Interestingly, PAX6 and SOX2 are the only two biologically relevant genes from the literature to be included in the list of genes selected by the Wilcoxon test. After obtaining the set of the differentially expressed genes, we rank them according to the ascending order of the Bonferroni-corrected p-value. We also do a coarse feature selection based on the mutual information between the gene expression level and category label (see Figure 2). The top-10 ranked genes from the Wilcoxon test (denoted as Wilcoxon genes/features) and mutual information-based method, and the 6 biologically relevant genes are reported in Table 1. However, we only use the Wilcoxon features and biologically relevant features for our further analysis in this report. In Figure 1, we plot the histograms of the cells with respect to the expression levels of several hand-picked features from both Wilcoxon genes and biologically relevant genes. We observe that even though these genes are expected to demonstrate a good separation between the GZ and CP categories, it not the case in reality. This may be due to the non-linear and collective nature of the relationship that exists between the gene expression levels and cell location.

Table 1: The 6 biologically relevant genes and the top-10 genes selected by the Wilcoxon rank sum test with Bonferroni correction. Note that none of the biologically relevant features are selected into the top-10 Wilcoxon genes.

Gene/feature pool	Gene names
Biologically relevant genes	`SOX2, PAX6, NEUROG1, NEUROG2, ASCL1, NOTCH1`
Wilcoxon top-10 genes	`CALM1, GAP43, MEF2C, ARPP21, STMN2, SATB2, MAP1B, LIMCH1, CHL1, NCAM1`
Mutual Information top-10 genes	`MEF2C, ARPP21, STMN2, CHL1, SATB2, MAP1B, GAP43, LIMCH1, NEFM, CALM1`

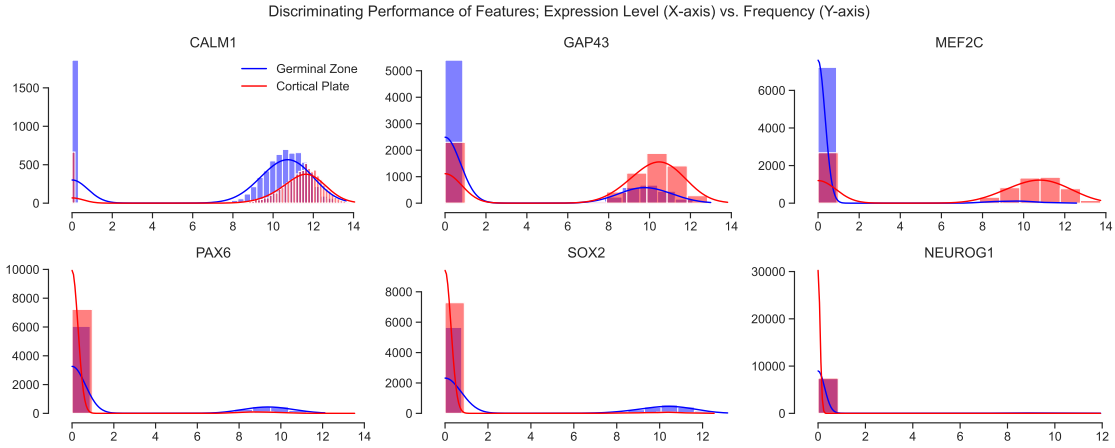

Figure 1: **(top row)** The top three genes ranked according to the lowest p-values obtained using the Wilcoxon rank sum test. (**bottom row)** Three of the biologically relevant genes for the neuronal migration during cortical development. Note that 2 of these genes (PAX6 and SOX2) were also selected by the Wilcoxon test-based feature selection procedure. Based on these histograms, we are not able to conclude the discriminative power of these individual genes with respect to the classification between the two cell locations.

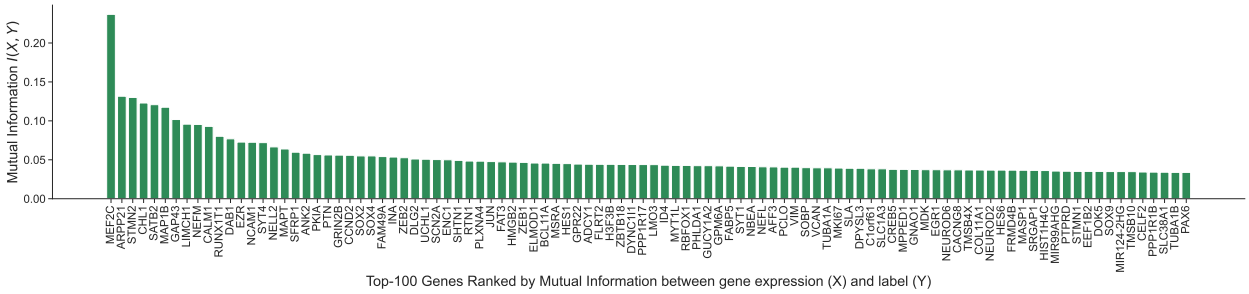

Figure 2: Top-100 genes ranked by mutual information between gene expression level and the label.

3.2 Classification

Based on our findings from the feature selection step, we created 5 sets of features:

1. Top-10 features by Wilcoxon p-value,
2. Top-100 features by Wilcoxon p-value,
3. Top-1000 features by Wilcoxon p-value,
4. 6 biologically relevant features according to literature,
5. Top-10 features by Wilcoxon p-value + 6 biologically relevant features.

For each set of features, we trained 3 classifiers: Linear Discriminant Analysis (LDA), K-nearest neighbors (k-NN), and Random Forest (RF). We chose the two non-linear models (k-NN, RF) since we expect the relationship between gene expression and cell location to be complex and non-linear. The RF can capture more non-linearity than the k-NN model, so we chose to train both to assess how much non-linearity is required to learn the relationship. We also trained a linear model as a baseline, though we expected its performance to be poor. While we did not perform a hyperparameter tuning for our classification algorithms, we adopted the hyperparameter configurations provided in Table 2 across all of our experimental settings.

Table 2: Hyperparameter configurations used in the classifiers

Classifier	Hyperparameter configuration
LDA	None
k-NN	$k = 10$
RF	number of trees = 100

Separately, we employed a recently proposed neural network architecture called LassoNet [4] in our classification step. LassoNet is a state-of-the-art neural network model that automatically selects features using a lasso penalty to regularize the model's coefficients. This penalty encourages some of the coefficients to be exactly equal to zero, effectively removing some of the features from the model. As a result, only the most important features are retained, and these features can skip the intermediate layers and be directly connected to the output layer. Inspired by the success of deep neural networks in problems such as like MRI and X-ray image classification, we evaluated its performance on our gene classification problem. LassoNet was given all the 1139 features that were selected using the Bonferroni corrected p-value threshold, as it should filter out uninformative features among this set on its own.

To obtain good estimates of training performance, we assessed each model on the training set using stratified 10-fold cross-validation. Although we performed cross-validation, these performance metrics may still be biased since the training set was used to select features. Thus, we evaluated each classifier on the held-out testing data to obtain final performance metrics. We report the accuracy, AUC-ROC, sensitivity, and specificity of each classifier. Finally, we analyzed the importance of features for each classifier to understand which genes are most strongly linked to cell location.

4 Results

Figure 3 illustrates the performance of the LDA, k-NN and RF models using the first three sets of features (top-10, top-100, and top-1000). Surprisingly, the LDA classifier performs virtually as well as the RF classifier ($\sim 90\%$ CV accuracy) across all three feature sets. The k-NN classifier performs well but worse than the other two classifiers. This suggests that an intermediate amount of non-linearity is detrimental to model performance. We observe that the performance of k-NN degrades with the number of features used during training. This may be due to the over-fitting caused due to the curse of dimensionality as k-NN requires more training data when the feature dimension increases.

The performance of all three classifiers improved moderately as the feature set size was increased from 10 to 100. However, increasing the feature set size from 10 to 1000 yielded virtually no performance gains. This may imply that the added genes are uninformative for our classification problem. Since the classifier performance is already quite high using just 10 genes, it may be that there are only a few genes highly correlated with cell location.

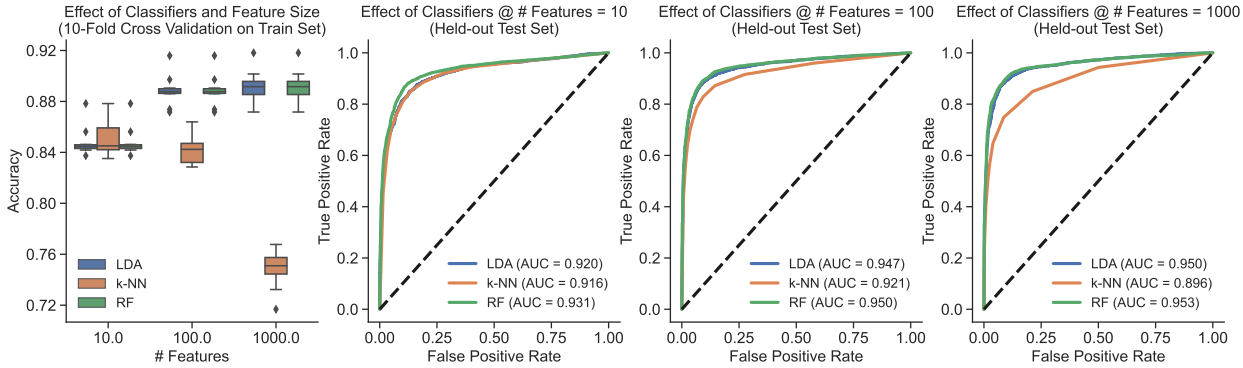

Figure 3: **(left)** Cross-validation performance of LDA, k-NN, and RF models on the top-10, top-100, and top-1000 features with the smallest Wilcoxon p-value. **(right)** Test set ROC curves. Each graph corresponds to one feature set (10, 100, 100 from left to right). See Table 3 for an extensive report of the metrics.

Figure 4 reports the performance of the LDA, k-NN, and RF models using the Wilcoxon features, biologically relevant features, and a combination of both. Notably, all three classifiers perform quite poorly – not far above chance – when it comes to the biological features. Additionally, adding the biologically relevant features to the Wilcoxon features gives virtually no performance gains. These findings suggest that the biologically relevant features from literature may not be informative about cell location on their own. We see a slightly different trend in performance between models here. The LDA and k-NN classifiers show the same performance across all three feature sets. The RF classifier outperforms the other two classifiers on feature sets that contain the Wilcoxon features but performs worse when only the biologically relevant features are used.

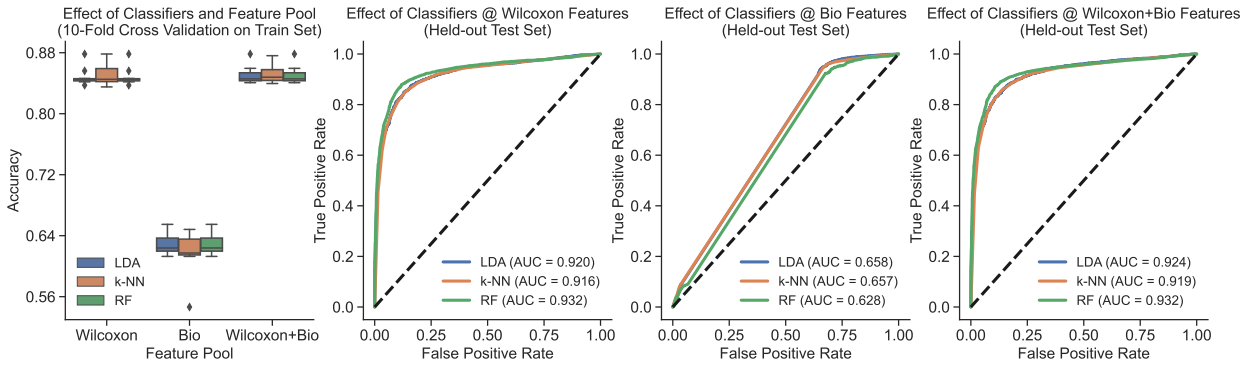

Figure 4: **(left)** Cross-validation performance of LDA, k-NN, and RF models on the top-10 smallest Wilcoxon p-value, biologically relevant, and 10 smallest Wilcoxon p-value + biologically relevant features. **(right)** Test set ROC curves. Each graph corresponds to one feature set (top-10, Bio, top-10 + Bio from left to right). See Table 3 for an extensive report of the metrics.

Table 4 reports the performance metrics of LDA, k-NN, and RF on the held-out test set. These results largely agree with train set cross-validation results stated in Table 3, confirming that the models generalize well over the test set. On the other hand, LassoNet performed on the test set with accuracy 0.89, sensitivity 0.87, and specificity 0.91 when all the Wilcoxon-selected 1139 genes were used as the feature set. Therefore, the performance of the LassoNet on the held-out test set was not far from the traditional models we used.

Table 3: Performance metrics of the 3 classifiers computed using 10-fold cross-validation on the train set across multiple feature pools considered.

Classifier		top-10			top-100			top-1000	
Metric	Accuracy	Specificity	Sensitivity	Accuracy	Specificity	Sensitivity	Accuracy	Specificity	Sensitivity
LDA	0.85 ± 0.01	0.79 ± 0.02	0.91 ± 0.02	0.89 ± 0.01	0.86 ± 0.01	0.92 ± 0.02	0.89 ± 0.01	0.87 ± 0.01	0.92 ± 0.02
k-NN	0.85 ± 0.01	0.80 ± 0.02	0.90 ± 0.01	0.84 ± 0.01	0.73 ± 0.02	0.95 ± 0.01	0.75 ± 0.01	0.51 ± 0.03	0.98 ± 0.01
RF	0.87 ± 0.01	0.86 ± 0.01	0.88 ± 0.02	0.89 ± 0.01	0.88 ± 0.01	0.90 ± 0.02	0.89 ± 0.01	0.89 ± 0.01	0.90 ± 0.02
Classifier		Wilcoxon			Bio			Wilcoxon + Bio	
Metric	Accuracy	Specificity	Sensitivity	Accuracy	Specificity	Sensitivity	Accuracy	Specificity	Sensitivity
LDA	0.85 ± 0.01	0.79 ± 0.02	0.91 ± 0.02	0.63 ± 0.01	0.94 ± 0.01	0.32 ± 0.03	0.85 ± 0.01	0.80 ± 0.02	0.90 ± 0.01
k-NN	0.85 ± 0.01	0.80 ± 0.02	0.90 ± 0.01	0.62 ± 0.03	0.84 ± 0.24	0.40 ± 0.19	0.85 ± 0.01	0.81 ± 0.02	0.90 ± 0.01
RF	0.87 ± 0.01	0.86 ± 0.01	0.88 ± 0.02	0.61 ± 0.01	0.92 ± 0.01	0.32 ± 0.02	0.87 ± 0.01	0.86 ± 0.02	0.88 ± 0.01

Table 4: Performance metrics of the 3 classifiers evaluated on the held-out test set across multiple feature pools considered.

Classifier		top-10			top-100			top-1000	
Metric	Accuracy	Specificity	Sensitivity	Accuracy	Specificity	Sensitivity	Accuracy	Specificity	Sensitivity
LDA	0.85	0.80	0.91	0.90	0.87	0.92	0.90	0.87	0.92
k-NN	0.85	0.81	0.90	0.85	0.74	0.95	0.75	0.51	0.98
RF	0.88	0.88	0.88	0.90	0.90	0.90	0.90	0.90	0.90
Classifier		Wilcoxon			Bio			Wilcoxon + Bio	
Metric	Accuracy	Specificity	Sensitivity	Accuracy	Specificity	Sensitivity	Accuracy	Specificity	Sensitivity
LDA	0.85	0.80	0.91	0.64	0.95	0.33	0.86	0.80	0.91
k-NN	0.85	0.81	0.90	0.53	0.09	0.96	0.86	0.81	0.90
RF	0.88	0.88	0.88	0.62	0.92	0.32	0.88	0.88	0.88

Figure 5 illustrates visualizations of the feature importance for LDA, k-NN, and RF classifiers. Each of the 50 most important genes is visualized as a bubble, where the size of the bubble represents the feature importance and the bubble color represents the rank of the gene's p-value. The LDA and RF classifiers seem to have more diversity in importance among the top 50 features, with only a few very important features. On the other hand, the top 50 features all have roughly similar importance for the k-NN classifier. Interestingly, the gene MEF2C is

the most important for both the LDA and RF classifiers and slightly more important than average for the k-NN classifier. The fact that this gene was identified as important by all 3 classifiers suggests that it is quite informative about cell destination (at least for our dataset).

With these plots, we can also evaluate a prevalent misconception in scientific literature. A common error is equating p-value magnitude with effect size; i.e. the hypothesis (or feature) with the smallest p-value must be correct (or most discriminative). If this were true, we would expect the largest bubbles in Figure 5 to have the smallest p-values and thus be colored the darkest. However, this is obviously not the case. By eye, there seems to be no clear relationship between bubble size and color for all three classifiers.

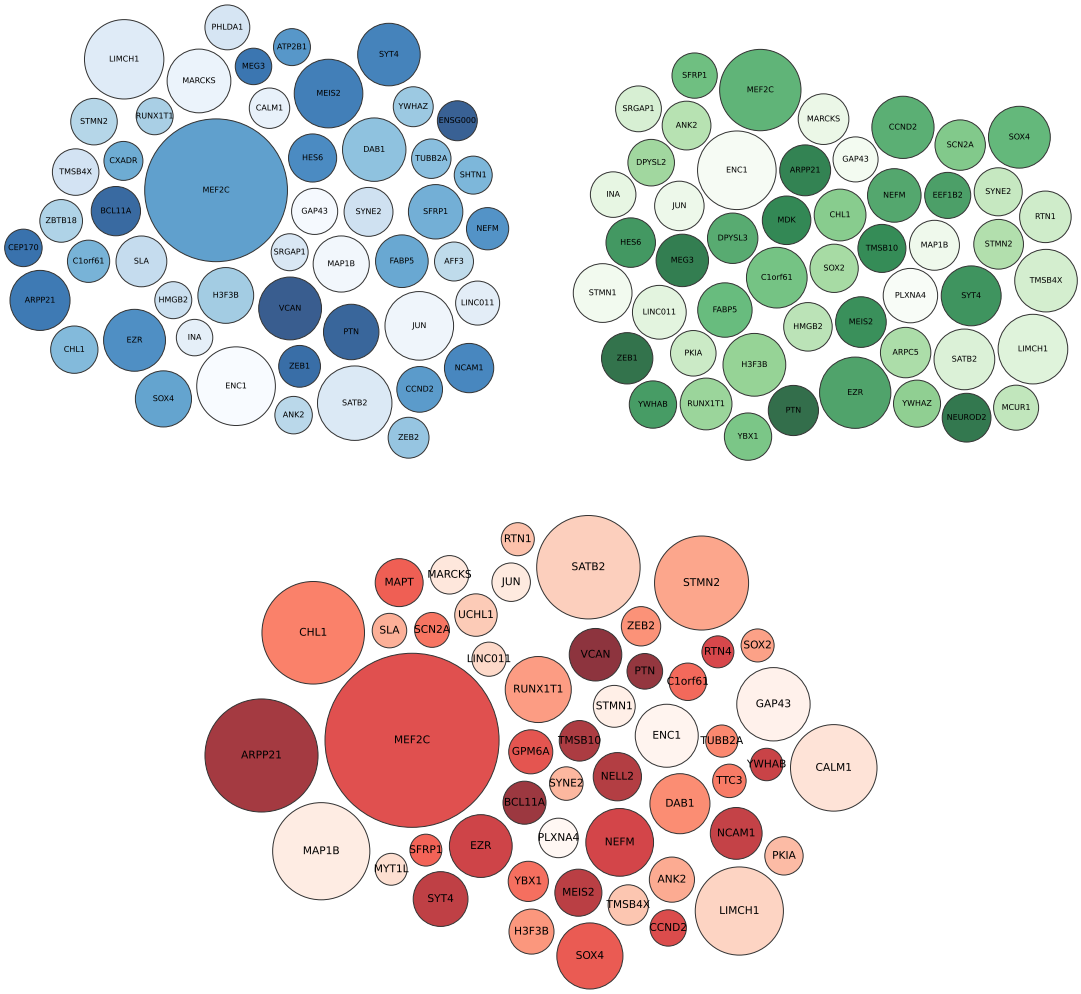

Figure 5: Relative importance of the 50 most discriminative features for the LDA **(red)**, k-NN **(green)**, and RF **(blue)** classifier. Bubble size signifies feature importance, bubble color corresponds to p-value rank (darker color = smaller p-value).

5 Conclusion

Learning the relationship between gene expression and dorsal pallial neurogenesis can potentially further our understanding of embryonic neurogenesis. In pursuit of this goal, we trained 4 classifiers to predict whether cells are located in the germinal zone or cortical plate based on their gene expression. We selected 6 feature sets of various sizes, including from a Wilcoxon test and genes that are biologically relevant according to the literature. We trained our classifiers – Linear Discriminant Analysis, k-Nearest Neighbors, Random Forest, and LassoNet

on these feature nets. Because we the relationship between genes and cell location is complex, we expect the linear method (LDA) to perform very poorly compared to the others. Surprisingly, all 4 models perform similarly, achieving accuracies of around 90%. Although increasing the number of features yielded small performance gains, our models were able to achieve high accuracy (around 85%) with as little as 10 Wilcoxon test features.

We found that classifiers trained on biologically relevant genes alone had poor performance, and adding biologically relevant genes to those identified by a Wilcoxon test made virtually no difference in model performance. This suggests that the genes identified by literature may not be strongly informative about cell location, or they may be linked to cell location in a more subtle way. Finally, we examined the feature importance – amount that each feature contributed to the classification decision – for each of our classifiers and found that a single gene, MEF2C, was a highly important feature in all our models. However, more analysis is needed, particularly using other datasets, to confirm our findings. Issues like batch effects or sampling bias may impair the generalizability of our results. Overall, our analysis suggests that there may be only a few genes that are highly informative in predicting the location of a neural progenitor cell.

References and Notes

[1] Diogo S Castro and François Guillemot. Old and new functions of proneural factors revealed by the genome-wide characterization of their transcriptional targets. *Cell cycle*, 10(23):4026–4031, 2011. 1

[2] Magdalena Götz and Wieland B Huttner. The cell biology of neurogenesis. *Nature reviews Molecular cell biology*, 6(10):777–788, 2005. 1

[3] Victoria Graham, Jane Khudyakov, Pamela Ellis, and Larysa Pevny. Sox2 functions to maintain neural progenitor identity. *Neuron*, 39(5):749–765, 2003. 1

[4] Ismael Lemhadri, Feng Ruan, Louis Abraham, and Robert Tibshirani. Lassonet: A neural network with feature sparsity. *The Journal of Machine Learning Research*, 22(1):5633–5661, 2021. 3

[5] Angeliki Louvi and Spyros Artavanis-Tsakonas. Notch signalling in vertebrate neural development. *Nature Reviews Neuroscience*, 7(2):93–102, 2006. 1

[6] Carol Schuurmans and Francois Guillemot. Molecular mechanisms underlying cell fate specification in the developing telencephalon. *Current opinion in neurobiology*, 12(1):26–34, 2002. 1