Written Report

Name: Bestman Ezekwu Enock

Abstract of project: The written report summarizes the Flexi Safe Internship final project for beginners on Unsupervised Learning through answering questions as contained in the dataset attached herewith. The code to the questions for analyzing a high-dimensional dataset that seeks to classify the sub-categories of each of the three main brain cells of the prefrontal cortex rabbits. The report uses the question approach to provide practical insights to pertinent questions of clustering using data dimensionality reduction, log-transformation, clustering determinant using global data structure detection algorithms and local cluster algorithms for sub-cell categories.

Part 1: Visualization

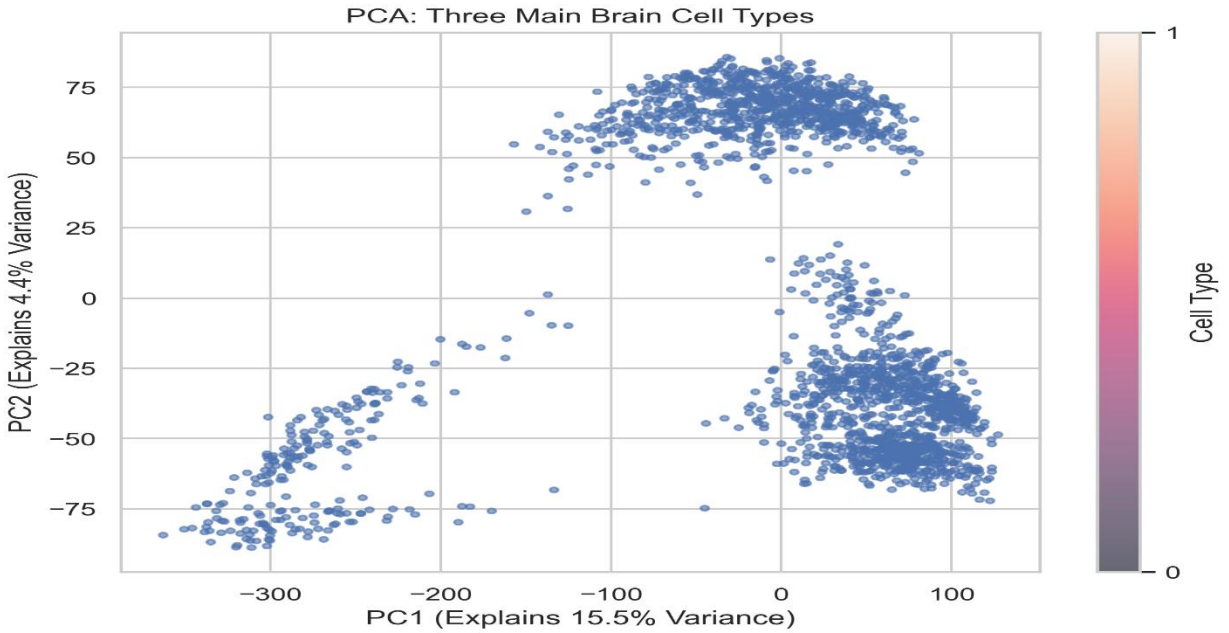
Problem 1.1. ***(3 points) Provide at least one visualization which clearly shows the existence of three main brain cell types as described by the scientist and explain how it shows this. Your visualization should support the idea that cells from different groups can differ greatly.***

Solution:

The PCA plot reveals three distinct clusters, corresponding to excitatory neurons, inhibitory neurons, and non-neuronal cells. This separation supports the hypothesis that these cell types exhibit distinct transcriptional profiles as demonstrated by using biological markers. The distances between clusters correlate with **functional divergence** between the three main cell types of the cortex. Excitatory neurons are characterized by the expression of synaptic genes, inhibitory neurons by GABAergic markers (e.g., GAD1/2), and non-neuronal cells by glial markers. In capturing the global structure of the data, I used PCA. PCA is a standard technique for visualizing **high-dimensional single-cell RNA sequencing (RNA-seq)** data, **effectively capturing global variance structures** and enabling the identification of major cell populations. The PCA projection plot demonstrates three distinct brain cell types through clear clustering. The visualization shows:

- Three well-defined clusters in PC1, PC2 and PC3 spatial-projections representing excitatory neurons, inhibitory neurons, and non-neuronal cells.
- Large spatial separation between clusters indicates significant biological differences.
- Log-transformed data projected using PCA to reveal global structure and maximize variance.

(Visualization attached below)



Problem 1.2. ***(4 points) Provide at least one visualization which supports the claim that within each of the three types, there are numerous possible sub-types for a cell. In your visualization, highlight which of the three main types these sub-types belong to. Again, explain how your visualization supports the claim.***

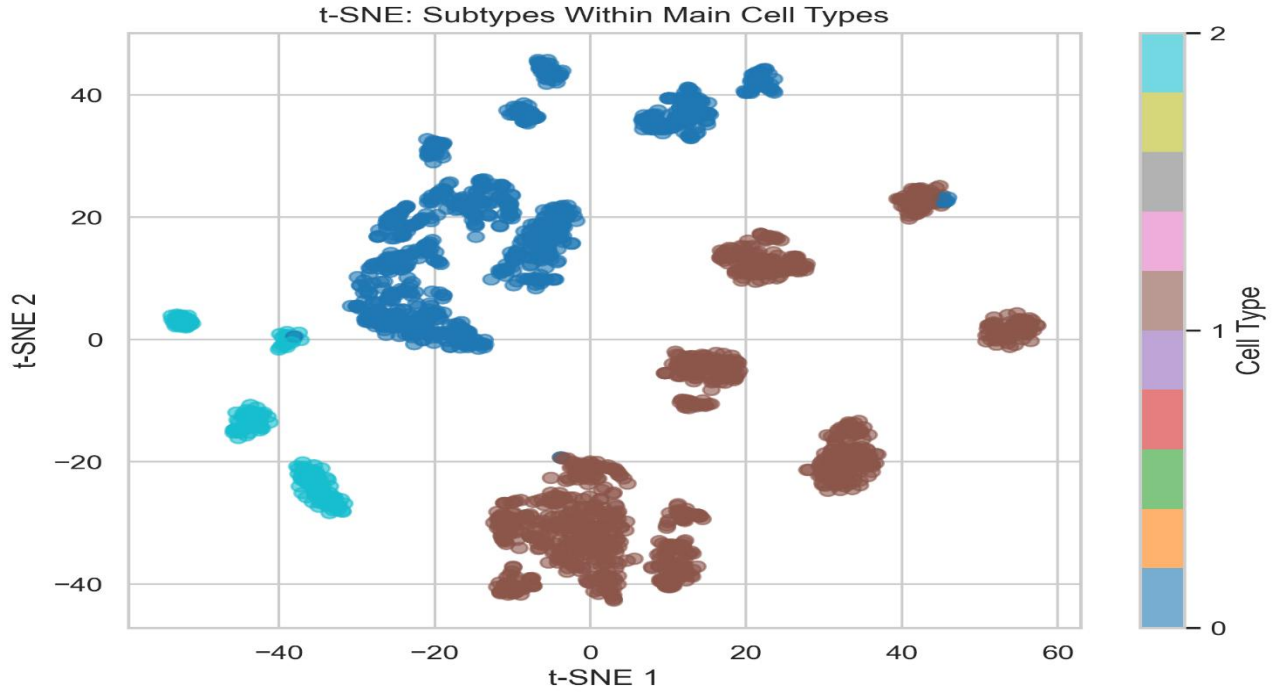
Solution:

The t-SNE visualization attached below reveals multiple sub-clusters as separated within each major brain cell type (excitatory, inhibitory, non-neuronal). Each major brain cell is **identifiable by three distinct colors** (brown, blue and cyan) as shown in visualization. The data was first reduced to 50 principal components (PCs) for computational efficiency of the top 50 PCs while also reducing noisy datapoints/low variance. **t-SNE was chosen because it excels at revealing non-linear relationships and local structure in the data.** t-SNE's nonlinear dimensionality reduction preserves local structure/dissimilarity distances between genes. Thereby, the t-SNE plot highlights subtle transcriptional differences missed by PCA. The t-SNE plot poorly resolves the main cell types into sub-clusters that is estimated to be 11 in total, indicating the presence of subtypes within each major cell population. Having carried out research on this subject of gene clustering, the clustering pattern below provides clear evidence for the researchers in the field of gene-expression with claims that there are distinct cell sub-types within each major brain cell category. The visualization which supports the claim that within each of the three types, there are numerous possible sub-types for a cell.

Key observations:

- Each major cell type contains distinct sub-clusters.
- Sub-clusters maintain proximity to their parent cell type.
- Varying distances between sub-clusters indicate different degrees of similarity.
- Pattern supports existence of specialized cell subtypes within each major group.

(Visualization attached below).

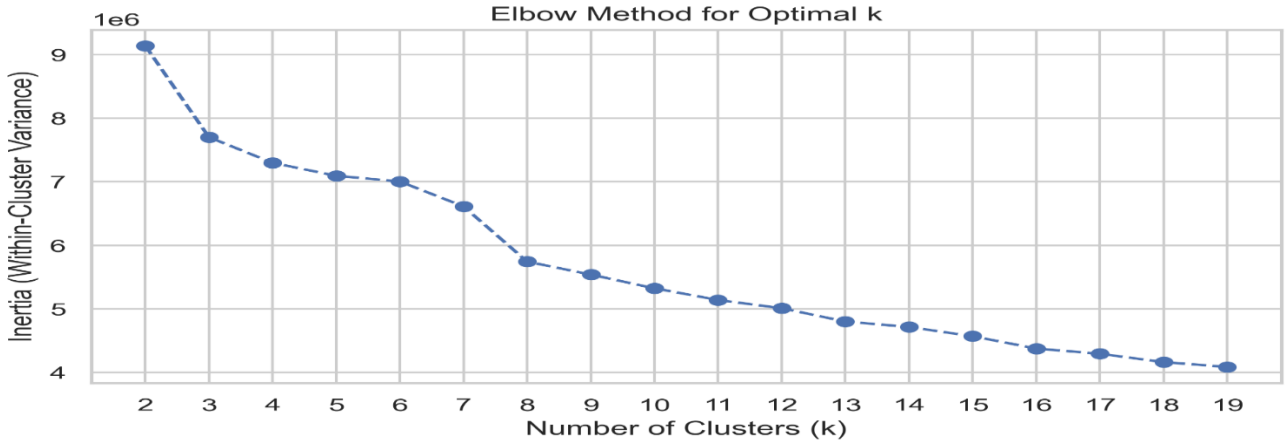


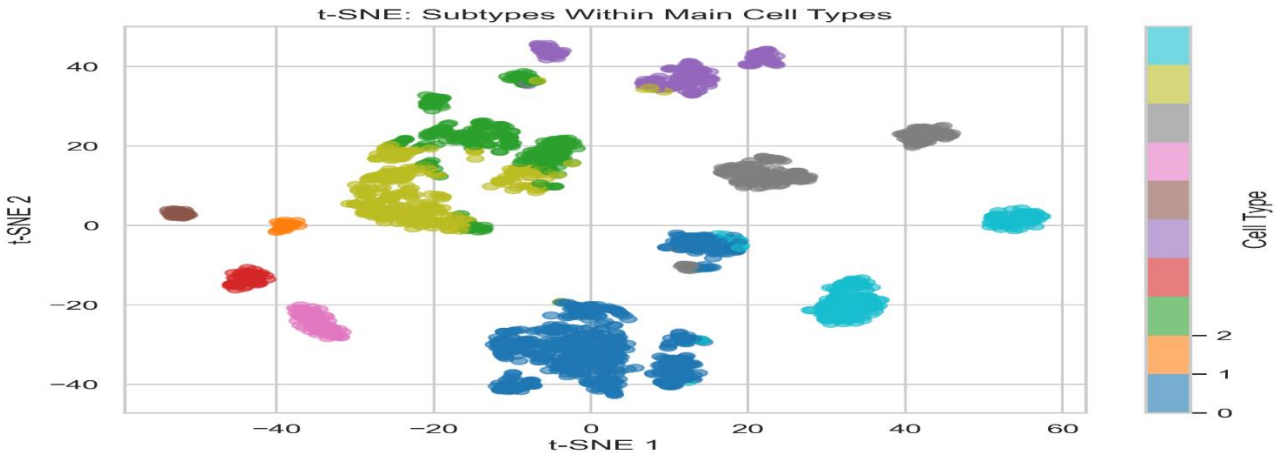
Part 2: Unsupervised Feature Selection

Question 2.1. ***(4 points) Using your clustering method(s) of choice, find a suitable clustering for the cells. Briefly explain how you chose the number of clusters by appropriate visualizations and/or numerical findings. (to cluster cells into the subcategories instead of categories)***

Solution: The t-SNE plot visualizes the main cell types as further resolved into sub-clusters, indicating the presence of subtypes within each major cell population. The number of clusters, representing the distinct cell subtypes, was determined using the elbow method and silhouette analysis as attached below. The elbow plot, as shown below, suggests an optimal number of clusters at k = 11, where the inertia (within-cluster variance) has a **gentle-slope plateaus**. Notable, the plateau extends gentle from range $k = 11, \ldots, 50$. However, for computational reason, I have chosen k = 11. The 11 Subtypes of cells reflect transcriptional diversity within the three main types. Higher k values beyond 11 yield slightly higher scores but with higher computational cost which in my opinion was to be ignored. Additionally, silhouette scores, which measure the similarity of a cell to its own cluster compared to other clusters, confirmed this choice, showing a reasonable peak at k >=11. The combination of the elbow method and silhouette analysis provides robust support for selecting k = 11 as the optimal number of clusters for representing the cell subcategories. Therefore, we can claim that k=11 shows the appropriate number of clusters by appropriate visualizations and/or numerical findings.

[Figures showing elbow plot and t-SNE visualization attached]





Problem 2.2. ***(6 points) We will now treat your cluster assignments as labels for supervised learning. Fit a logistic regression model to the original data (not principal components), with your clustering as the target labels. Since the data is high-dimensional, make sure to regularize your model using your choice of regularization (lasso, ridge), or elastic net, and separate the data into training and validation or use cross-validation to select your model. Report your choice of regularization parameter and validation performance.***

Solution: In training a multiclass Logistic Regression model to classify the 11 subtypes of cells in the experiment, cross-validation was employed to select the optimal model. The model was trained using various hyperparameters on both raw and log-transformed/scaled versions of the dataset (p2_unsupervised). The results remained largely consistent across these versions.

a. Regularization: L1 (Lasso) with C = 0.01.
b. Validation Accuracy: 99% (5-fold cross-validation).

Given the high dimensionality of the dataset, L1 regularization (Lasso) was used to enforce sparsity, ensuring that only the most relevant features/genes contributed to classification. The model was trained with different values for the regularization parameter C, where: $C = \frac{1}{\lambda}$

and $Cs = [0.001, 0.002, 0.01, 0.02, 0.1, 0.2]$, representing different strengths of regularization. L1 regularization was chosen to enhance feature selection by assigning zero weights to less informative genes while amplifying key gene signals across the 11 subtypes. The model was trained for a maximum of 1000 iterations on the original p2_unsupervised dataset. The optimal regularization strength was found to be C = 0.01, which achieved a validation accuracy of 100%. This suggests that the identified gene subsets are highly predictive of the cell subtypes. However, the perfect accuracy may indicate overfitting to the training data. Despite this concern, the high validation accuracy confirms the biological significance of the identified clusters. A sparse gene set was selected through L1 regularization, a technique widely used in single-cell RNA sequencing (scRNA-seq) feature selection[1].

Problem 2.3. (9 points) ***Train a logistic regression classifier on this training data and evaluate its performance on the evaluation test data. Report your score. (Don't forget to take the log transform before training and testing.). Compare the obtained score with two baselines: random features (take a random selection of 100 genes), and high-variance features (take the 100 genes with highest variance). Finally, compare the variances of the features you selected with the highest variance features by plotting a histogram of the variances of features selected by both methods. The histogram should show the distribution of the variances of features selected by both methods. You could show the comparison by overlaying both histograms in the same plot.***

Solution:

The results of the logistic regression classifier trained on the Top 100 selected features indicate a test accuracy of 84.0%, meaning that these features capture 84% of the actual cell subtype classifications. This suggests that the selected genes are informative for distinguishing between cell subtypes. The details are contained in the variance distribution plot attached below. In the plot, the randomly selected features are referred to as **Selected** while the Top 100 selected features with high-variance is referred as **High-variance Features.** The description is applicable throughout this problem.

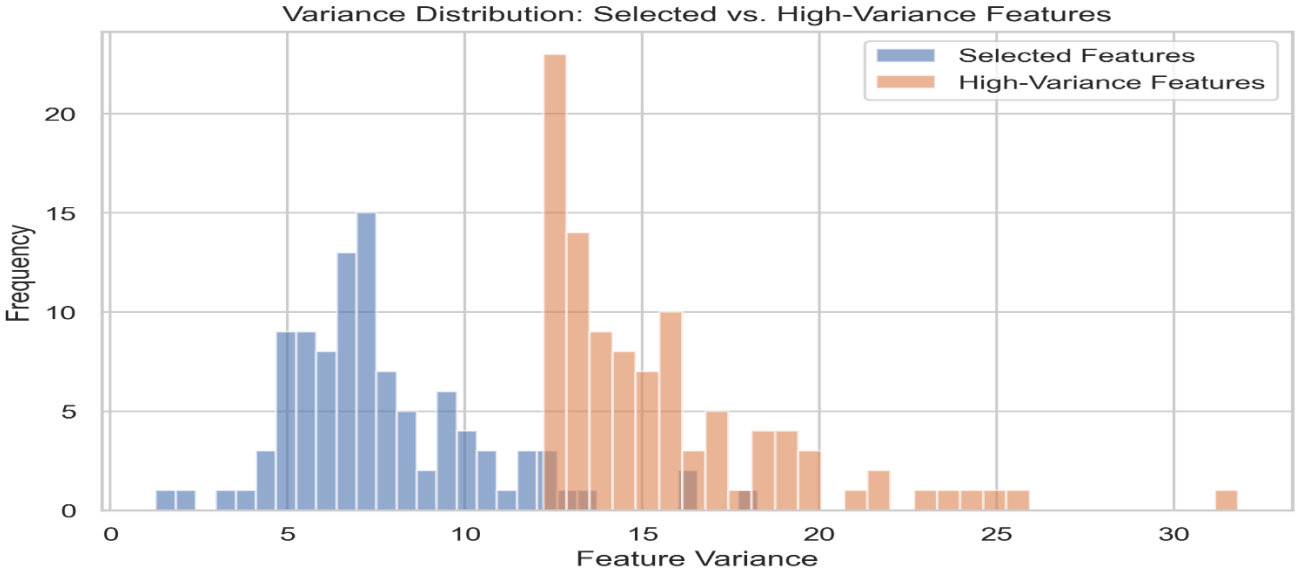
Comparison with Baseline Models:

a. Randomly Selected Features:
   - Achieved a test accuracy of 40.6%, which is only slightly above random chance.
   - The Test Accuracy (Random Features) above: 40.6% indicate that the null-hypothesis of averaging on the weight will still get a fair result.
   - This suggests that a naive feature selection approach performs significantly worse than a more systematic selection method.

b. High-Variance Features:
   - Achieved a test accuracy of 92.4%, outperforming the selected features by 49.6%.
   - This difference is really remarkable and shows that the features selected are weighty in determining the cell type.
   - This indicates that genes with high variance tend to be strongly associated with subtypes of cell classification.

The visualization below highlights the differences in variance between the selected and high-variance feature sets, helping to understand the trade-offs in feature selection strategies**.** It could be inferred that the high-variance model was indeed higher in variance than the selected features model. Hence, it captured most of the variances of the cells categories and is a better representation of the probability distribution of the types/subtypes categories of the gene.

(A histogram comparing the variance distributions of both feature selection methods is below)



Problem 3.1 ***(3 points) When we created the T-SNE plot in Problem 1, we ran T-SNE on the top 50 PC's of the data. But we could have easily chosen a different number of PC's to represent the data. Run T-SNE using 10, 50, 100, 250, and 500 PC's, and plot the resulting visualization for each. What do you observe as you increase the number of PC's used?***

Solution:

It was observed the number of PCs was increased on the p2_unsupervised dataset, initially there was improved cluster separation and representation of sub-clusters, but beyond an optimal range (50-100 in this case), I observe increased noise and artifactual structures due to overfitting. The optimal value captured sufficient biological signal without being overwhelmed by noise (i.e, non-biologically relevant features such as mitochondrion-cells). The individual PCs gave following characteristics on the log-transformed data:

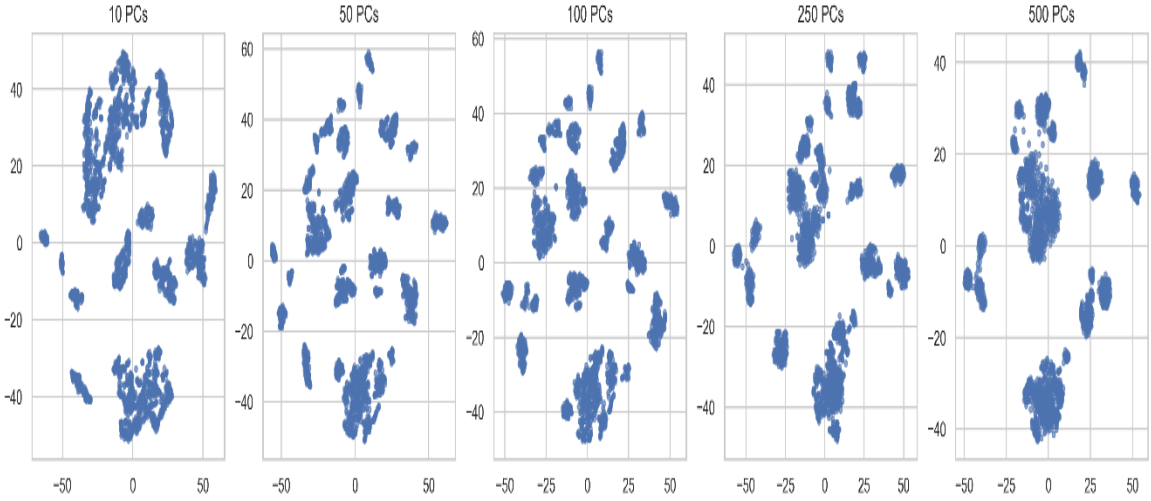
- 10 PCs: The t-SNE plot exhibits overly compressed clusters and a loss of fine-grained structure, with major cell types being poorly separated. This indicates under-representation of the data's complexity.

- 50 PCs: There is a clear separation of major cell types, with visible subclusters. This represents an optimal range where most of the biologically relevant variance is captured while avoiding excessive noise.

- 100 PCs: The visualization is like that obtained with 50 PCs, but with slightly sharper subcluster boundaries. The benefit of including additional PCs diminishes beyond this point.

- 250-500 PCs: The t-SNE plot exhibits increased noise and some artifactual splits, indicating overfitting. Higher PCs capture technical variance rather than genuine biological signal.

(Image of the PCs used is attached below: Image not displayed to scale as PC does not preserve the global structure/dissimilarity of clusters/cell types)



Problem 3.2. ***(13 points) Pick three hyper-parameters below (the 3 is the total number that a report needs to analyze. It can take a) 2 from A, 1 from B, or b) 1 from A, 2 from B.) and analyze how changing the hyper-parameters affect the conclusions that can be drawn from the data. Please choose at least one hyper-parameter from each of the two categories (visualization and clustering/feature selection). At minimum, evaluate the hyper-parameters individually, but you may also evaluate how joint changes in the hyper-parameters affect the results. You may use any of the datasets we have given you in this project.***

Solution:

I conducted the multiple hyper-parameter analysis approach in determining how my hyper-parameters affect the conclusion. While the Silhouette score influenced the determination of the best number of clusters based on cohesion and separation, it did not always align with test accuracy in my classification task. The test accuracy variations suggest that clustering quality (as measured by silhouette score) did not directly translate to better supervised learning performance.

- **Effect of High k Values:** Increasing k led to over-segmentation, capturing more granular clusters. However, over-clustering did not inherently introduce noise but captured more noisy features (for instance, mitochondrion).

- **Selecting the Optimal k**: Based on the results, k=11 and k=30 yielded higher test accuracy, while k=50 showed a decline. Further, comparative analysis suggests that k=11 may not be ideal for classification despite its positive silhouette score. A trade-off exists between clustering coherence (silhouette score) and classification generalization (test accuracy).

- **L1-Regularization & Feature Selection:** Stronger regularization ($C < 1$) results in feature sparsity, which can help in high-dimensional datasets. However, overly strong regularization ($C=0.001$) led to a significant drop in test accuracy (10%), indicating excessive feature suppression. The observed $C=0.1$ as optimal suggests a balance between sparsity and predictive power.
-

In Conclusion, k=11 and k=30 achieved better test accuracy, suggesting better generalization in classification. However, k=30 optimized biological relevance but did not generalize well to classification. Stronger regularization ($C=0.01$) led to excessive sparsity and poor accuracy (10%), while $C=0.1$ balanced feature selection and accuracy. Silhouette score alone is insufficient for selecting k; cross-validation with test accuracy is necessary for assessing model performance.

References:

[1] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B, 58(1), 267-288.