Predicting Gene Transcriptional Activity Using 3D Spatial Genomic Data

Kristine Yang
DATA1030 Final Report
https://github.com/kristineyang/data1030_project

I. Introduction

The human genome is vast, containing over 3 billion base pairs of DNA. However, only a small fraction of this DNA constitutes transcriptionally active genes – those that code for functional protein products. While recent breakthroughs in genomics techniques have significantly increased our ability to gather biological data, probing for active genes on a single-cell basis remains a time-consuming and expensive task with limited scalability. Moreover, current techniques for detecting gene activity require separate samples and experimental workflows, which cannot be easily integrated into other experimental procedures.

3D genomics techniques and datasets are growing in both popularity and availability. By capturing the spatial coordinates and locational information of individual genes, this data offers the potential to provide insights into gene activity, especially as studies have shown that genes have a non-random pattern of organization within the cell's nucleus and genes with similar expression patterns localize together in "gene-expression neighborhoods" (Oliver et al., 2005). While machine learning has been increasingly incorporated into genomics studies, research on the classification of a gene's activity based on its spatial positioning has yet to be published.

The goal of this project is to predict a gene's activity using information about its spatial position, distance to key nuclear landmarks, and chromosomal coordinates. The data comes from a genome-scale chromatin tracing experiment on human female cells published in a 2020 Cell Press paper, "Genome-Scale Imaging of the 3D Organization and Transcriptional Activity of Chromatin" by Su et al. The target variable is 'transcription' with classes 'on' or 'off', referring to the activity of the individual gene.

II. Exploratory Data Analysis

Due to the experimental nature of the data collected, the original dataset — with over eight million data points — contained a significant proportion of missing values and, in some cases, indistinguishable gene identification due to ambiguous signal. After filtering out these instances, the remaining dataset still contained approximately 2.6 million data points. Of this data, 86.3% consisted of transcriptionally inactive genes and 13.7% consisted of transcriptionally active genes (Figure 1). This imbalance is biologically representative as only a small proportion of the genome is actively transcribed.

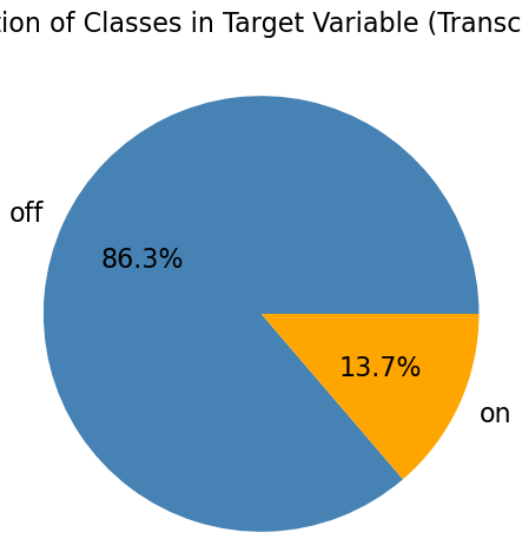

Figure 1: Proportion of data belonging to classes 'on' (orange) and 'off' (blue) for the target variable, transcription.

Each cell in the human body has 23 pairs of chromosomes, and these chromosomes vary in size, gene density, and transcriptional activity. Consistent with this knowledge, Figure 2 shows a non-uniform distribution of active genes across chromosomes.

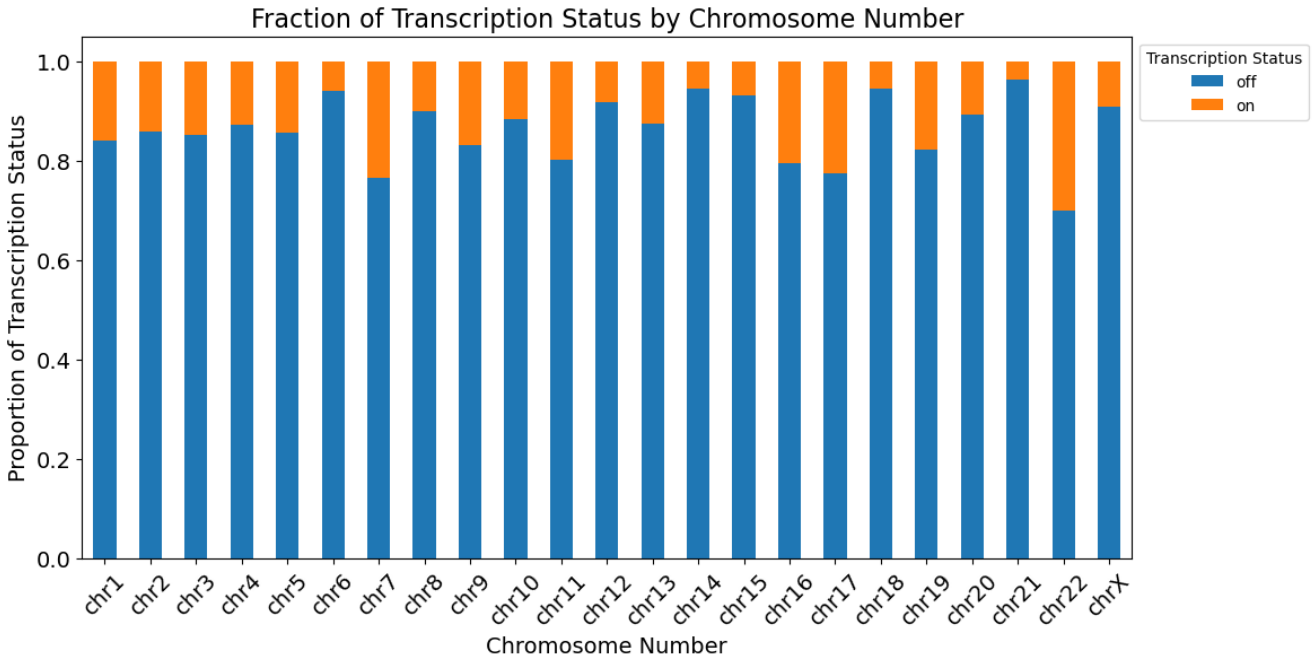

Figure 2: Non-uniform transcription activity status across all 23 chromosomes.

Nuclear organization studies have shown that transcriptionally silent genes tend to be associated closer to the nuclear lamina – a peripheral protein meshwork lining the inner nuclear

membrane – whereas active genes tend to localize closer to the interior of the nucleus (Geyer et al., 2011). Figure 3 shows that the observations in this dataset are consistent with these findings as transcriptionally 'on' genes tend to be more centrally located, or in other words, further from the nuclear lamina.

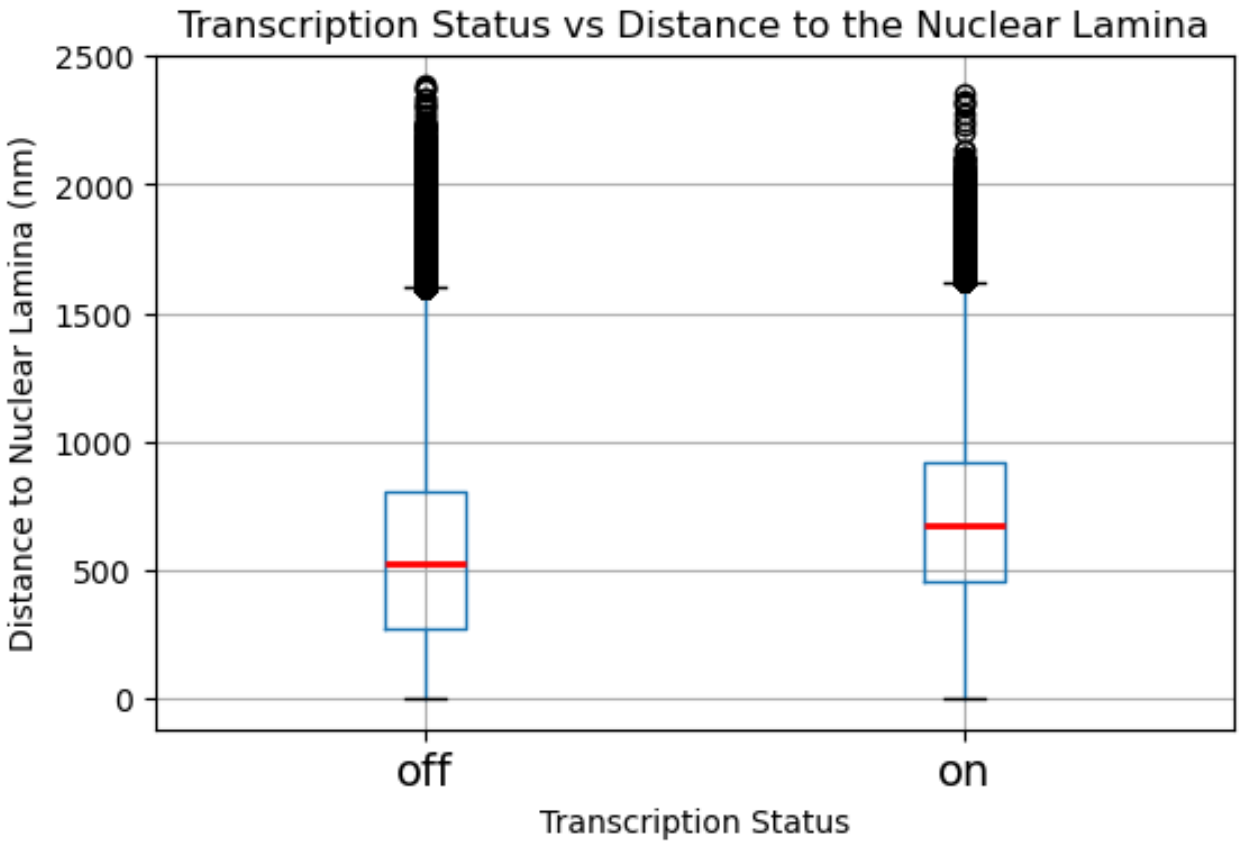

Figure 3: Transcriptionally active genes tend to be located further away from the nuclear lamina than inactive genes.

While the x and y positions of the genes (measured relative to the microscopic imaging plane) have a similar distribution between transcriptionally inactive and active genes, Figure 4 shows that the bimodal distribution of the z position is more pronounced in transcriptionally 'on' genes.

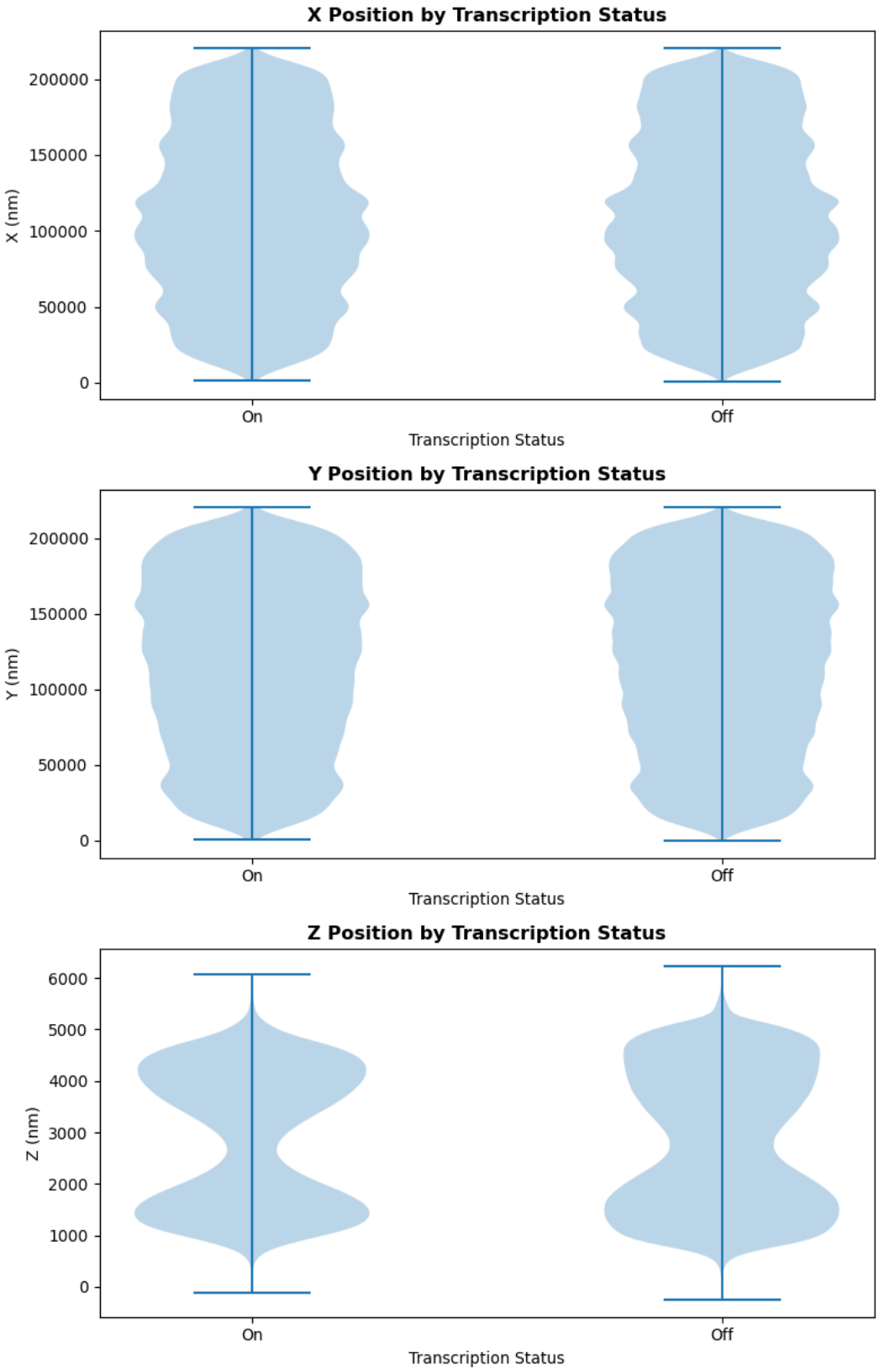

Figure 4: The distribution of X, Y, and Z positions of genes by transcription status.

III. Methods

Due to the large size of the dataset, a stratified random subsample of 100,000 data points was taken, and this subset was used as input for the models. Prior to splitting and preprocessing, the original 'genomic coordinate' feature (ex: chr1:2950000-3050000) was divided into 'chromosome', 'coordinate start', and 'coordinate end' features, for a total of 11 features and one target variable.

For the Logistic Regression, k-Nearest Neighbors, and Random Forest models, the GroupShuffleSplit function from sklearn was used to split the data into a 95:5 ratio. A 3-fold GroupKFold split was subsequently applied to the 95% subset for cross validation. A group-based splitting strategy was used to preserve the integrity of the single-cell data collection method. For the XGBoost model, GroupShuffleSplit was used to partition the data into training, validation, and testing sets with a 90:5:5 ratio. A 98/1/1 ratio for training, validation, and testing was originally implemented but was adjusted to 90/5/5 after the observation of large variances in the test score.

The categorical features, 'chromosome', 'gene names', 'homolog number', were preprocessed using OneHotEncoder. The x, y, and z features were bounded by the frame of the microscope and thus were preprocessed using MinMaxScaler. The remaining continuous features, 'distance to lamina (nm)','distance to nucleoli (nm)', 'distance to speckles (nm)','coordinate start', and 'coordinate end', were preprocessed with StandardScaler.

Four supervised machine learning algorithms were implemented: Logistic Regression, k-Nearest Neighbors Classifier, Random Forest Classifier, and XGBoost Classifier with early stopping. Table 1 shows the models and the hyperparameters that were tuned for each algorithm.

Table 1: Machine learning models and their corresponding hyperparameters.

Model	Hyperparameters	Values
Logistic Regression	C	[1e-2, 1e-1, 1e0, 1e1, 1e2]
	penalty	['l2', 'l1']
k-Nearest Neighbors	n_neighbors	[3, 5, 7, 9, 10]
	weights	['uniform', 'distance']
Random Forest	max_features	[0.25, 0.5, 0.75, 1.0]
	max_depth	[1, 10, 50, 70, 100]
XGBoost (with early stopping)	reg_alpha	[0e0, 1e-2, 1e-1, 1e0, 1e1, 1e2]
	reg_lambda	[0e0, 1e-2, 1e-1, 1e0, 1e1, 1e2]
	max_depth	[5, 30, 70, 100]

Each model was trained over three random states to address uncertainties from splitting and non-deterministic models. The F1 score was chosen as the evaluation metric because active and inactive genes are of equal interest, and it provides a more reliable measure in the presence of class imbalance.

IV. Results

Based on the mean F1 scores of the models, the logistic regression (with L2 penalty) model performed the best with a mean test score of 0.493 (Table 2). All models performed better than the baseline model, which was generated using the DummyClassifier function from sklearn with a 'stratified' random split. Compared to the baseline performance of the DummyClassifier, all the models significantly outperform it by about 20 standard deviations above the baseline.

Table 2: Performance of models across three random states evaluated using mean F1 score.

Model	Mean F1 Score
Baseline (DummyClassifier)	0.166
Logistic Regression	0.493 ± 0.040
k-Nearest Neighbors	0.492 ± 0.045
Random Forest	0.492 ± 0.031
XGBoost (with early stopping)	0.487 ± 0.007

The confusion matrix for the best performing logistic regression model is shown in Figure 5. The model exhibits a high true negative rate, indicating it correctly identified transcriptionally inactive genes in most cases. The model has a false negative rate of 0.55 and a true positive rate of 0.45, meaning it incorrectly classifies 55% of active genes as inactive, while correctly identifying 45% of active genes.

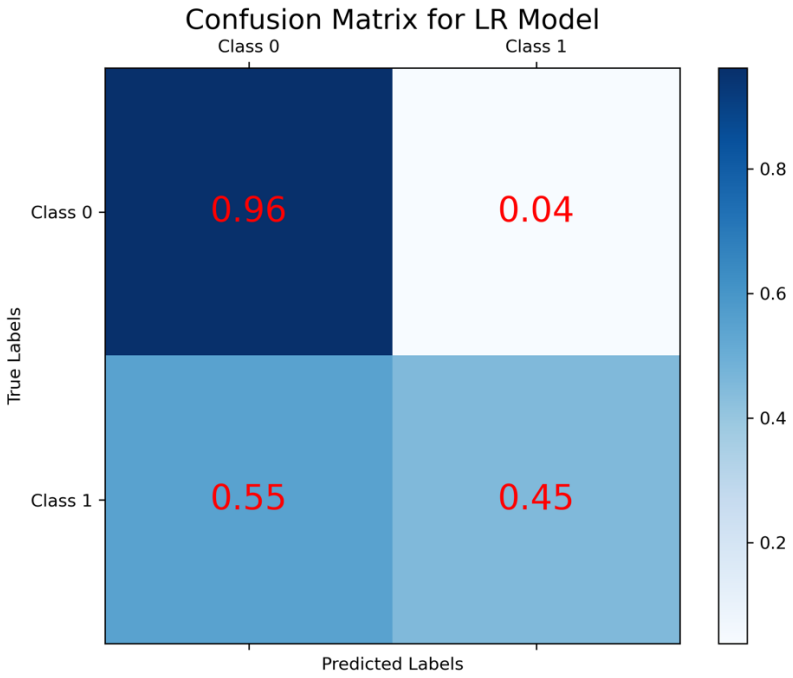

Figure 5: The confusion matrix for the best performing logistic regression model.

To understand which features contributed most to the predictions of this logistic regression model, permutation importance, coefficient importance, and SHAP were used. Figure 6 shows that gene names, chromosome number, and distance to speckles were among the most important features when calculated using permutation importance, as dropping these features resulted in the largest decreases in the test score.

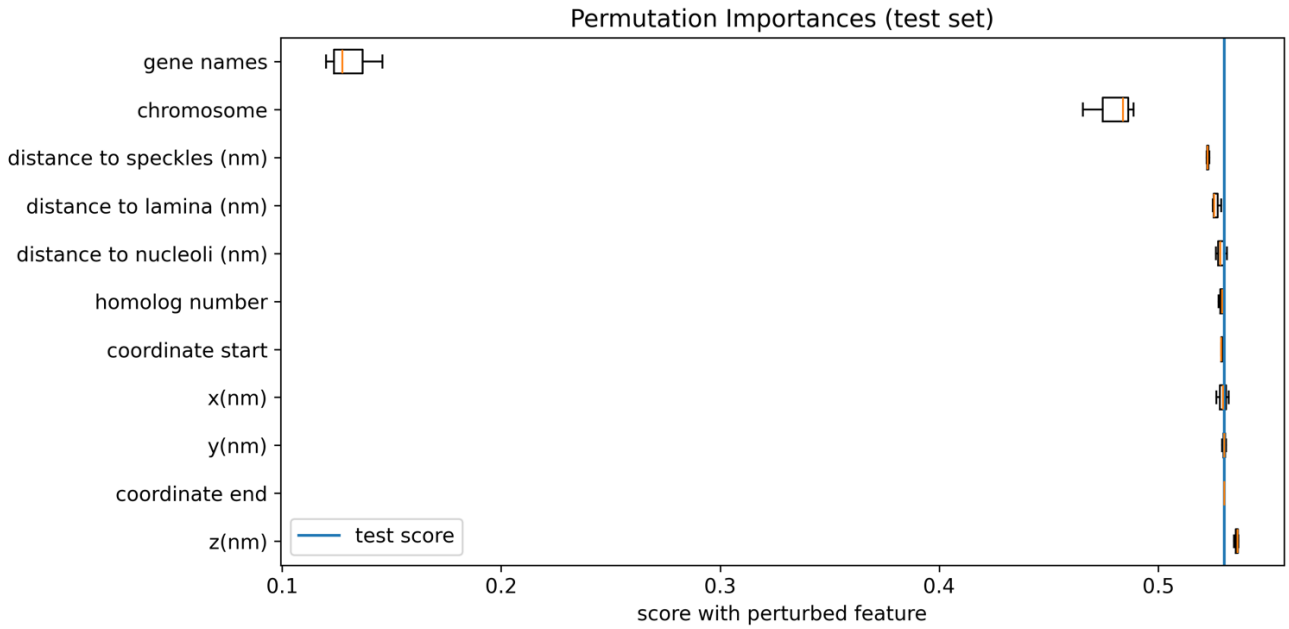

Figure 6: Permutation importances calculated using the best performing logistic regression model.

Figure 7 shows the top 20 most important features in the logistic regression model, ranked by the relative importance of each feature's coefficient. All the most important features in this plot correspond to specific genes, consistent with the findings in Figure 6, where gene identification plays a crucial role in determining the probability of transcriptional activity.

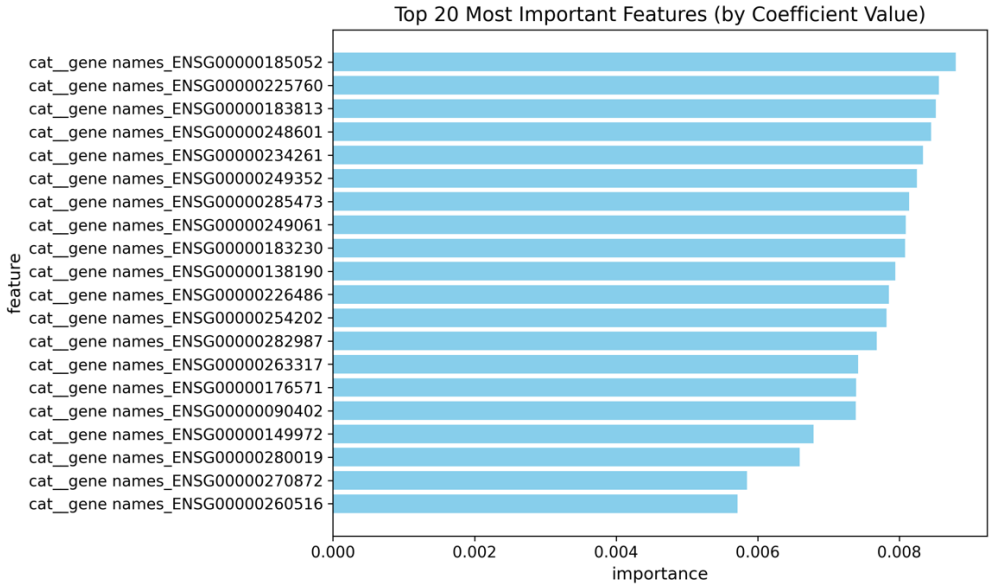

Figure 7: Top 20 most important features in the logistic regression model, ranked by coefficient.

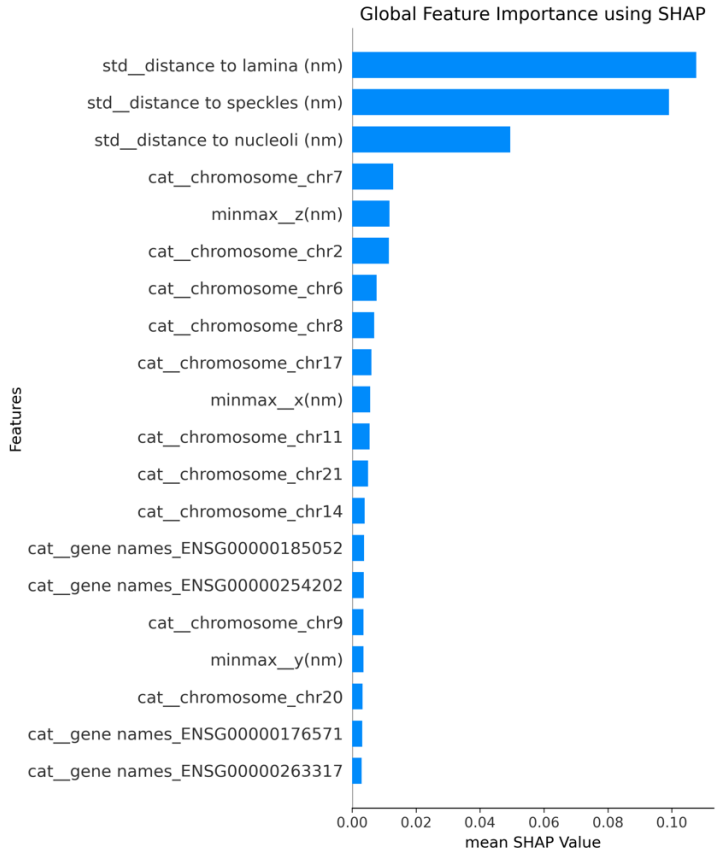

Figure 8: SHAP summary plot showing global feature importance. Distances to nuclear landmarks are the top three features.

Figure 8 shows the SHAP summary plot displaying the global feature importance. Distances to the nuclear lamina, nuclear speckles, and nucleoli are the three features that influence the prediction the most, showing the importance of gene positioning relative to the rest of the nuclear architecture.

Local feature importance was also calculated at for three independent instances to show which features contribute the most for a single observation. Specific gene names and distance to the nuclear lamina pushed the prediction to the left in all cases (Figure 9). Chromosome identification, where chromosome 2 pushed the prediction in Figure 9a to the right and chromosome 6 pushed the predictions to the left (Figure 9b and 9c), highlights its importance in influencing the model's prediction. Interestingly, the x, y, and z coordinates of the genes did not rank among the most important features in any of the assessments. This could perhaps be explained by the lack of a universal coordinate system in the experimental setup, as these coordinates were calculated relative to the microscope frame, with the possibility of cells being oriented in various ways.

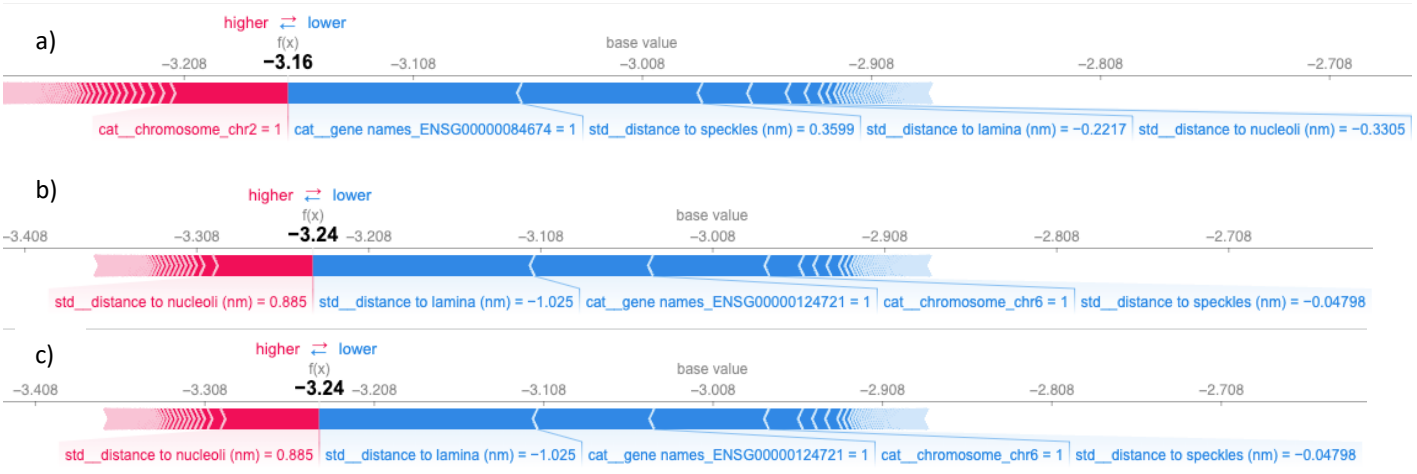

Figure 9: SHAP local feature importance for three independent single observations.

V. Outlook

Given the large size of the original dataset, a random stratified subsample of 100,000 points was selected. Due to biological variability, this subsample may not have fully captured all patterns in the data, and therefore, further iterations of these models using a larger proportion of the dataset could help improve model performance. More extensive hyperparameter tuning could also be performed, including the addition of the 'elastic net' penalty to the logistic regression model. Additionally, incorporating feature engineering for the x, y, and z spatial

coordinates, as well as more data collection of key biological factors like cell state and size, could further improve the model's predictive power.

VI. References

1. Geyer, P. K., Vitalini, M. W., & Wallrath, L. L. (2011). Nuclear organization: Taking a position on gene expression. *Current Opinion in Cell Biology, 23*(3), 354-359. https://doi.org/10.1016/j.ceb.2011.03.002
2. Oliver, B., & Misteli, T. (2005). A non-random walk through the genome. *Genome Biology, 6*(4), 214. https://doi.org/10.1186/gb-2005-6-4-214
3. Su, J.-H., Zheng, P., Kinrot, S. S., Bintu, B., & Zhuang, X. (2020). Genome-scale imaging of the 3D organization and transcriptional activity of chromatin. *Cell*.