Developing a framework to study the effects of enhancer-like promoters on gene regulation

Arpit Omprakash

A dissertation submitted for the partial fulfilment of BS-MS dual degree in Science

Indian Institute of Science Education and Research Mohali

April 2021

Certificate of Examination

This is to certify that the dissertation titled "**Developing a framework to study the effects of enhancer-like promoters on gene regulation**" submitted by Mr. **Arpit Omprakash** (Reg. No. MS16124) for the partial ful-filment of BS-MS dual degree programme of the Institute, has been examined by the thesis committee duly appointed by the Institute. The committee finds the work done by the candidate satisfactory and recommends that the report be accepted.



Dr. Shashi Bhushan Pandit Dr. Rajesh Ramachandran Dr. Kuljeet Singh Sandhu

(Supervisor)

Dated: April 30, 2021

Declaration

The work presented in this dissertation has been carried out by me under the guidance of Dr. Kuljeet Singh Sandhu at the Indian Institute of Science Education and Research Mohali.

This work has not been submitted in part or in full for a degree, a diploma, or a fellowship to any other university or institute. Whenever contributions of others are involved, every effort is made to indicate this clearly, with due acknowledgement of collaborative research and discussions. This thesis is a bonafide record of original work done by me and all sources listed within have been detailed in the bibliography.

Arpit Omprakash
(Candidate)
Dated: April 30, 2021

In my capacity as the supervisor of the candidate's project work, I certify that the above statements by the candidate are true to the best of my knowledge.

Dr. Kuljeet Singh Sandhu
(Supervisor)

Acknowledgements

I would like to thank Dr. Kuljeet Singh Sandhu for giving me tremendous support, guidance, and freedom to work on this thesis. I would like to extend my gratitude to the committee members Dr. Shashi Bhushan Pandit and Dr. Rajesh Ramachandran for suggestions on my work.

I would like to thank the members of Genome biology lab, especially Shagun, Sunandini, Lipika, Yachna, and Mohan for always being ready to help and for their insightful advice.

To my dearest friends: Subhashree for providing emotional support and timely advice throughout my thesis. Srichandan, Jeevanjeet, Kundu, and Monty for all the wonderful meetings and helping me take a breather when I needed the most. Utkarsh for keeping me focused on my goals. Bhavish, Antriksh, Tejaswar, and Saswat for being there when I needed them the most and all the sweet memories of IISER life.

I would like to express my sincere gratitude to Dr. Manjari Jain and Dr. Anand Krishnan for instilling in me a love for curiosity and discovery and their support during difficult times. I will always be grateful to the members of BEL: Richa, Nakul, Sonam, Soniya, Esha, Prathibha, and Mahi for their continual support and all the exposure to different model organisms and field work that helped me grow as a student. I would also like to extend my gratitude to Ram, Abhinav, Nafiza, Varun, and Vaibhav for all the scientific expeditions and discussions I had during my summer that helped me grow as an independent thinker.

I would like to thank my parents for being incredibly supportive. They have always encouraged me to make my way and made sure that I grow up with all the comforts of life. I would like to sincerely thank them for giving me the freedom to choose my path and instil in me the confidence to overcome obstacles. I would like to thank my elder sister for her faith in me and her support throughout the time of the thesis. I would also like to thank my family for being supportive and helping me stay calm during stressful times.

And lastly, my biggest thanks to KVPY for the monetary support, the IISER Mohali library for all the resources that helped me get through my 5 years at IISER, and public databases – UCSC Genome Browser, ENA, NCBI, ENCODE – for without these, my project would not have been possible.

— Arpit Omprakash

List of Figures

List of Tables

Databases used

1. UCSC Genome Browser: *https://hgdownload.soe.ucsc.edu/downloads.html*
2. European Nucleotide Association (ENA) Browser: *https://www.ebi.ac.uk/ena/browser/*
3. NCBI GEO: *https://www.ncbi.nlm.nih.gov/geo/*
4. The Encyclopedia of DNA Elements (ENCODE): *https://www.encodeproject.org/*

Contents

Abstract

GWAS (Genome Wide Association Studies) have been crucial to identifying genetic loci associated with diseased phenotype. The hypothesis-free nature of GWAS studies have been a success at predicting specific cancer markers. However, this hypothesis-free nature has also led to one of its main issues, i.e., the large number of distant SNPs discovered with no biological link to the known genetic pathways of the diseased phenotype. Recent advancements in chromatin interaction mapping techniques have identified long-ranged promoter-promoter interactions that regulate gene expression pathways in eukaryotes. The presence of regulatory enhancer-like activity in some promoters and differences in the epigenetic features associated with the promoters and enhancer-like promoters (ELPs) have also been described. It led us to hypothesize that studying such long-range promoter-promoter contacts using ELPs may provide insights into biological links between distant SNPs and genetic pathways of disease in a population.

Here we explore possible histone markers and transcription factor bindings (epigenetic factors) that can distinguish between promoters and ELPs. We also build machine learning models that can predict the magnitude of enhancer-like activity (enhancer potential) of a promoter given its epigenetic factors. Regression models to predict the enhancer potential values were made, but the models' accuracy was not up to the mark. Improvements have been suggested for the models, including better feature extraction methods using machine learning classifiers. In the case of HeLa cancer cells, biologically significant epigenetic factors are identified via the classifiers that distinguish between promoters and ELPs. However, the models did not exhibit sufficient accuracy to get relevant features in K562 cancer cells.

In the later part of the thesis, spatial interactions between distant promoters have been characterised using Hi-C data. A mathematical framework incorporating the enhancer potentials and spatial interactions between promoters has been proposed to study the propagation of gene regulation in promoter-promoter networks. Initial results from the framework indicate that it can be used to identify distant upstream interacting promoters of a given promoter of interest and model time-course gene expression data to identify novel pathways of gene regulation.

Keywords – GWAS, Hi-C, enhancer-like promoters (ELPs), gene regulation, machine learning

Chapter 1

Introduction

Phenotypic variation in individuals is underlaid by genetic variation. Some of these genetic variations are associated with diseased traits in human populations and are termed causal variants (Bush & Moore, 2012). Studies linking causal variants and diseases pave the path for an era of personalized medicine, where all individual variations in the genetic code of an individual are used to guide clinical practice (Schierding, Cutfield, & O'Sullivan, 2014). Cataloguing of common genetic variants that are associated with complex traits and diseases has been made possible in recent years due to developments in high throughput genome sequencing and Genome Wide Association Studies (GWAS), among other advances (Haines et al., 2005; Reuter, Spacek, & Snyder, 2015).

In the late 1990s, the primary method of genetic investigation for diseased phenotypes was through pedigree analysis and inheritance studies of genetic linkage in families. Linkage disequilibrium (LD) has been defined as the "*non-random association between alleles at different loci*" (Visscher, Brown, McCarthy, & Yang, 2012). Physically proximal loci exhibit stronger LD than loci that are farther apart on the chromosome. Genetic linkage studies exploit the large LD within pedigrees to identify the chromosomal location of a single disease-causing allele (Pulst, 1999). Thus, single-gene disorders were often easily investigated using a pedigree analysis approach (Altmüller, Palmer, Fischer, Scherb, & Wjst, 2001). However, most common diseases have complex genetic and epigenetic mechanisms at work. Familial genetic linkage studies do not seem to be successful at studying such diseases. (Altmüller et al., 2001).

A higher (effective) population size leads to weaker LD for a given distance (Hill & Robertson, 1968). The genomic distance at which LD decays determines the number of genetic markers needed to "tag" a genomic haplotype (Visscher et al., 2012). Common

SNPs (Single Nucleotide Polymorphisms) are used as genetic markers to tag genomic regions exhibiting variation. It is estimated that a selection of approximately 500,000 common SNPs would be sufficient to tag variation in non-African human populations, even though the total number of common SNPs exceeds 10 million (Belmont et al., 2005; Visscher et al., 2012). In a landmark paper, Risch and Merikangas (Risch & Merikangas, 1996) showed that the statistical power of an association study consisting of 1 million variants in the genome and a sample of unrelated individuals would be higher when compared to linkage analysis using a few hundred markers (Visscher et al., 2012). Along with the theoretical foundations laid down by Risch and Merikangas (Risch & Merikangas, 1996), the international HapMap project (Belmont et al., 2003) provided experimental foundations for carrying out GWAS by identifying a list of SNP tags that captured most of the genomic variation in different human populations.

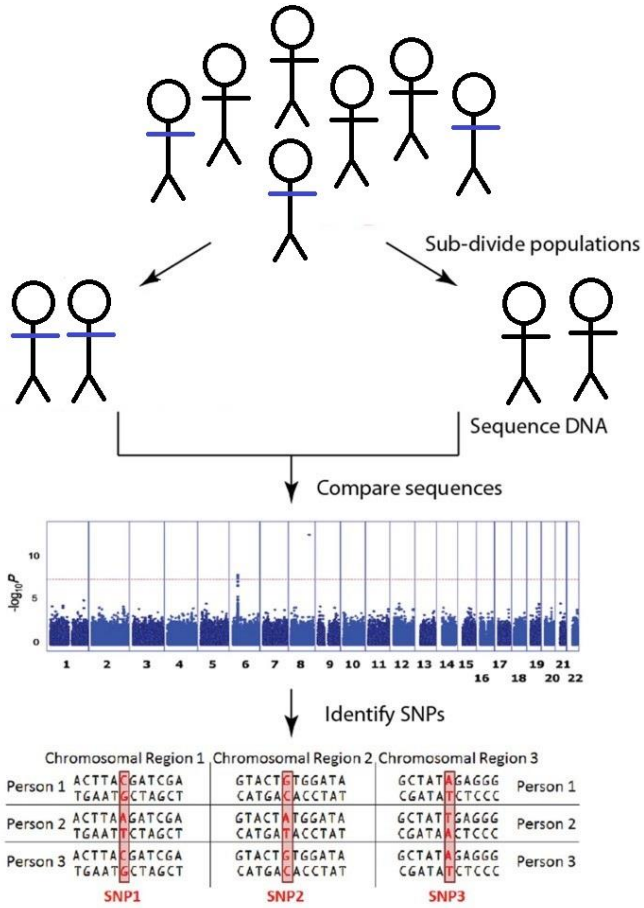


Figure 1: GWAS to identify genetic associations by comparing SNPs across the human genome. Adapted from (Schierding et al., 2014)

GWAS provides a biologically unbiased, hypothesis-free method to detect associations between genetic loci and phenotypic traits. GWA studies are primarily conducted to

identify various SNPs related to phenotypic traits that are involved in a given disease (Schierding et al., 2014). A GWA study's typical approach is the case-control setup, where a comparison between a control group (healthy population) and a case group (diseased population) is carried out. Individuals in both groups are genotyped for common, known SNPs (typically a million or more SNPs per study). For each SNP, the allele frequency is investigated. Differences in the allele frequencies between the case and control group are reported as an odds ratio. In the context of GWA studies, the odds ratio is defined as "the ratio of the odds of the case for individuals having a specific allele and the odds of the case for individuals who do not have the same allele" ("Genome-wide association study - Wikipedia," n.d.). If an observed odds ratio is relatively high or quite low, a p-value for the significance is calculated using a chi-squared test. An odds ratio significantly higher than one indicates that the given SNP is associated with the disease case.

In the case of cancer genomics, GWAS has played a significant role in detecting SNPs that have further led to clinically relevant predictions (Jostins & Barrett, 2011; Klein, Xu, Mukherjee, Willis, & Hayes, 2010). However, several researchers have pointed out that nearly half of the disease-associated SNPs from GWA studies are not located in or near recognized disease-causing genes (McClellan & King, 2010; Visel, Rubin, & Pennacchio, 2009). In the words of McClellan and King (McClellan & King, 2010),

"To date, genome-wide association studies (GWAS) have published hundreds of common variants whose allele frequencies are statistically correlated with various illnesses and traits. However, the vast majority of such variants have no established biological relevance to disease or clinical utility for prognosis or treatment."

Most of the SNPs lie in gene deserts, i.e., genomic regions larger than 500kb that lack identified genes or annotated protein-coding sequences (Craig Venter et al., 2001; Libioulle et al., 2007). Such SNPs' location in gene deserts means either the biological pathway of action of such regions is unknown or it is a false positive with no biological relevance. There have been numerous debates on the implications of SNPs' occurrence in the gene deserts with no unanimous decision from the community (Klein et al., 2010; McClellan & King, 2010). However, recent studies on genome architecture have revealed the presence of long-range interacting gene regions that may be the key to determining the role such gene deserts play in gene regulation (Liu et al., 2020; Ritchie & Van Steen, 2018; Singh Sandhu et al., 2012).

Traditionally, gene regulation was viewed as involving single-gene interactions, i.e., an enhancer activating a single promoter that leads to the production of a specific protein. Recently the traditional view has been supplemented with a multi-gene interaction model for eukaryotic genomes (Wei, Nicolis, Zhu, & Pagin, 2019; Yang, Lin, Wu, Chuang, & Chang, 2015). The multi-gene interaction model proposes a multivalent spatial interaction model, i.e., where more than two promoters and enhancers aggregate to regulate gene expression. Functional understanding of higher-order (multivalent) chromosomal interactions and organization of the eukaryotic genome is minimal. However, comparative analyses of different cell lines show that cell-specific chromatin interactions provide structural frameworks to study gene regulation and suggest significant enrichment of enhancer-promoter interactions for cell-specific functions (Li et al., 2012). Promoters with enhancer-like regulatory activity have been described recently by multiple studies (Dao & Spicuglia, 2018; Nguyen et al., 2016). Promoters frequently form long-range contacts with other promoters, and some promoter elements are shown to function as enhancers in enhancer reporter assays (Andersson, 2015; Dao & Spicuglia, 2018; Li et al., 2012; Schaffner, 2015). This points to the existence of promoter-promoter networks that may regulate gene expression in eukaryotes. The study of gene regulation via such long-range promoter-promoter networks may provide insights into the interaction pathways between diseased gene-associated regions and regions marked by the distant SNPs.

This thesis has characterized differences between regular promoters and enhancer-like promoters and used the differences to build machine learning models that can predict the enhancer-like activity for a promoter given its histone markers and transcription factor bindings. A mathematical framework to study the propagation of gene regulation by building a promoter-promoter network is also established that can be used to model interactions and explain the unexplained association of distant promoter (SNPs) with a given downstream gene.

Chapter 2

Predicting Enhancer-like activity in promoters using epigenetic data

Epigenetic markers are widely used to differentiate between gene regions containing typical promoters, enhancers, and super-enhancers (Creyghton et al., 2010; Heintzman & Ren, 2009; Hnisz et al., 2013). Typical enhancers and promoters have characteristic histone markers and transcription factor pairings. Regression analysis was carried out using epigenetic datasets to find differences between the histone markers and transcriptional factor bindings in promoter-like promoters (typical non-regulatory promoters) and enhancer-like promoters (regulatory promoters). The analysis aims to build models to predict different promoters' capacities to regulate gene expression (termed enhancer potential).

2.1 Materials and Methods

2.1.1 Data Collection and Clean-Up

Enhancer potential data for the different promoters were collected from the CapStarr-seq experimental data (Dao et al., 2017) for HeLa S3 cells. The data consisted of genomic locations of promoters and their associated enhancer activity values determined experimentally. Genomic locations of the promoters were extracted from the matrix. The genomic location data was used to collect histone modification and transcription factor enrichment values for the relevant genomic regions from ChIP-seq experimental data at ENCODE.

Three types of matrices for each cell line were created using the above data:

- Matrix of Histone modifications (Histone data): The matrix had histone modifications as columns and the promoter locations as rows.
- Matrix of Transcription Factors (TF data): The matrix had transcription factors as columns and promoter locations as rows.
- Enhancer Potential matrix: The matrix has a single column (enhancer potential values) and the promoter locations as rows.

After collecting and organizing the data into the above matrices, rows (promoters) with at least one NaN value were removed from the dataset. The final dataset consisted of 20719 promoters.

2.1.2 Regression Analysis

Regression analysis was carried out using the sci-kit learn library on the HeLa Histone dataset to determine the relationship between a promoter's epigenetic factors and enhancer potential. The histone matrix was used as the X matrix, and the enhancer potential matrix was used as the Y matrix.

One-third of the dataset (6837 promoters) was used as the validation dataset, and two-thirds (13881 promoters) were used for training. The data was scaled using the standard scaler class in the sci-kit learn library.

Regression analysis was carried out using MLP (Multi-Layer Perceptron) Regressor, SVR (Support Vector Regressor), and Decision Tree Regressor models.

2.1.2.1 Multi-Layer Perceptron Regressor

An artificial neuron is modelled as a perceptron. The perceptron takes in a few inputs and produces an output based on weighted linear combination of all the inputs and a bias variable.

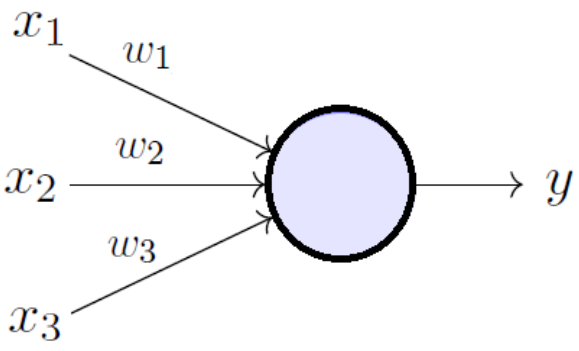


Figure 2: Perceptron model (Minsky & Papert, 2019)

In the case of a multi-layer perceptron model, several perceptrons are connected in a structure, as shown in Figure 3. Each vertical stack of perceptrons is called a layer.

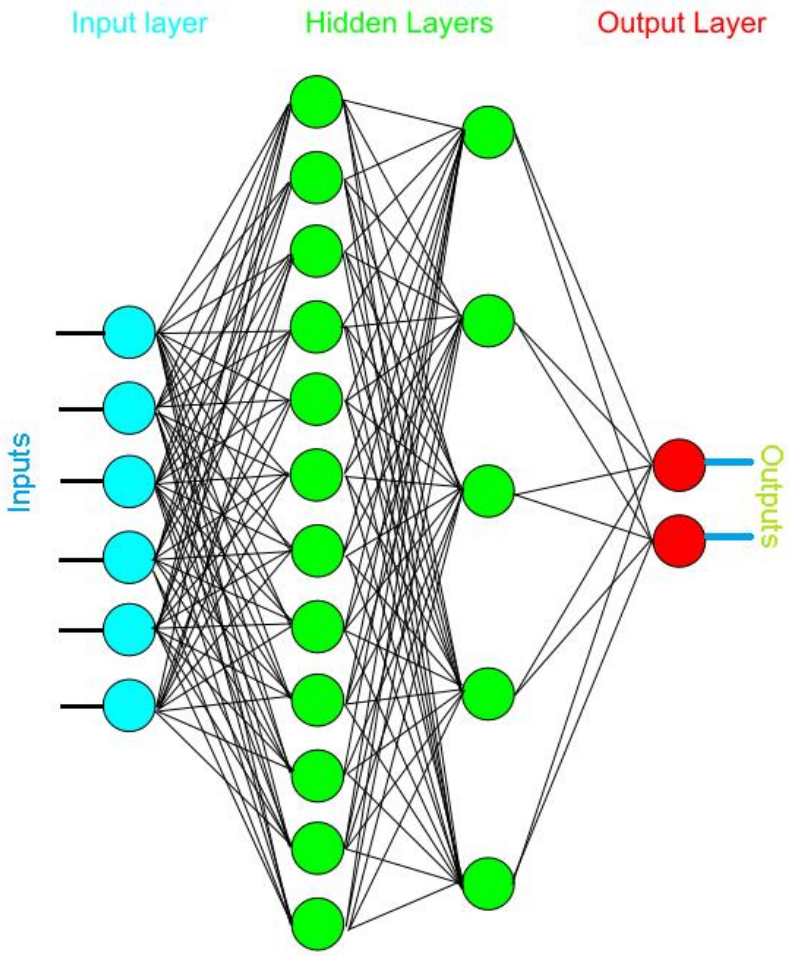


Figure 3: Structure of a Multi-layer perceptron model

The first layer (that takes in input) is called the input layer, followed by hidden layers that finally lead to the output layer. Output from a layer is processed using an activation function such as relu or sigmoid activation before passing on to the next layer. The weights and biases of individual perceptrons are the tuneable parameters in the model. The model is trained using backpropagation via stochastic gradient descent (Amari, 1993).

2.1.2.2 SVR (Support Vector Regressor)

Support Vector Regressor (SVR) constructs a hyperplane in a high-dimensional space containing the data points. An ε-insensitive region is introduced around the hyperplane called the ε-tube. The model then optimizes the location of the ε-tube (and also the hyperplane) such that most of the training data points fall inside the ε-tube.

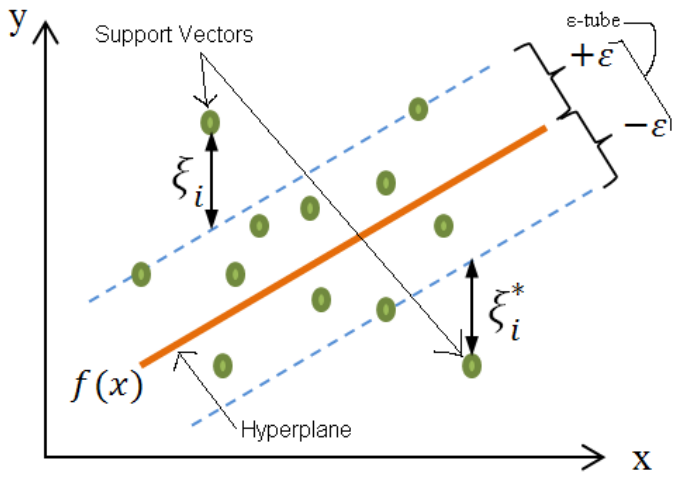


Figure 4: Working of Support Vector Regressor. Adapted from (Chanklan, Kaoungku, Suksut, Kerdprasop, & Kerdprasop, 2018)

The hyperplane is represented in terms of support vectors, which are training samples that lie outside the boundary of the tube. The support vectors are the most influential instances that affect the shape of the tube and the training and test data are assumed to be drawn independently from the same fixed but unknown probability distribution function (Awad, Khanna, Awad, & Khanna, 2015).

2.1.2.3 Decision Tree Regressor

A decision tree regressor arrives at an estimate for a given set of input variables, by asking a series of true-false questions and using if-else conditioned responses to narrow the possible values till the model is confident enough to make a single prediction.

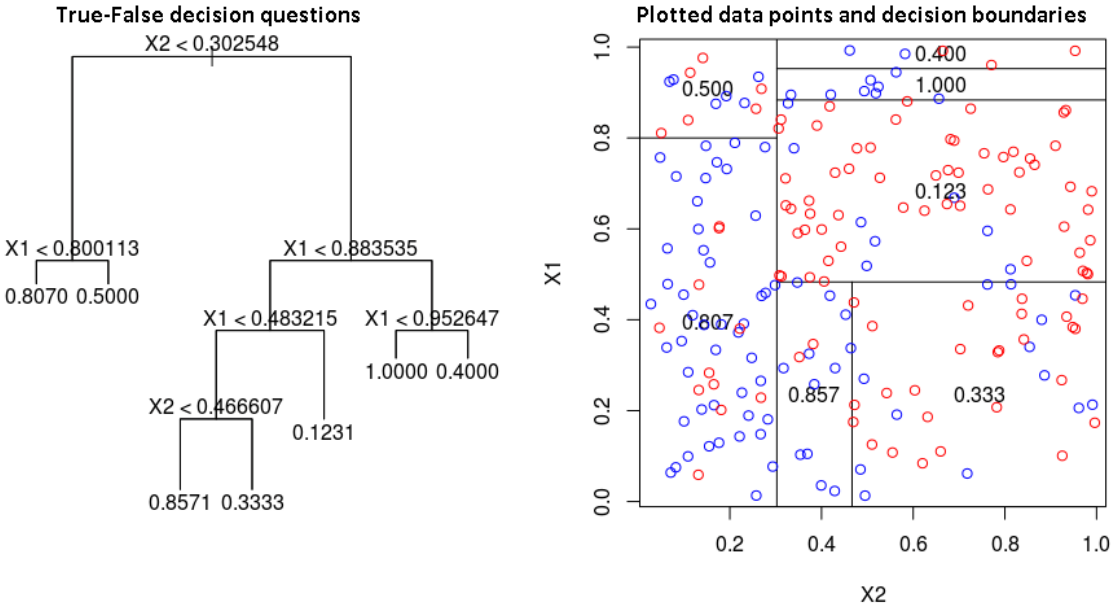


Figure 5: Example working of a decision tree regressor. Adapted from (Drakos, 2019)

In the diagram above, there are two features x_1 and x_2. At each branch, the model asks and answers a true-false question. Each branch leads to another branch till we reach a leaf node (end node). The leaf node represents the predicted y value for a given set of x_1 and x_2 values. Given training data, the model determines the best questions as well as the order in which to ask the questions to make the most accurate estimate of the underlying data distribution (Drakos, 2019).

2.2 Results

2.2.1 Model Evaluation

The regression models were evaluated using the RMSE (Root Mean Squared Error), MSE (Mean Squared Error), MAE (Mean Absolute Error) rates, and Overall Score.

The MAE represents the difference between the true and predicted values extracted by averaging the absolute difference in the predicted and actual values. The MSE represents the difference between the true and predicted values extracted by squared average difference over the dataset. RMSE is the square root of the MSE (DTN, 2019). The overall score metric represents the coefficient, R^2 and is defined as $(1 - {}^{u}/_{v})$, where u is the residual sum of squares and v is the total sum of squares. The best possible score is 1.0, and it can be negative (because the model can be arbitrarily worse). A constant model that always predicts the expected value of y, disregarding the input features, would get a score of 0.0 ("sklearn.neural_network.MLPRegressor — scikit-learn 0.24.1 documentation," n.d.).

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}|$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y})^2$$

$$RMSE = \sqrt{MSE}$$

$$R^2 = 1 - \frac{u}{v} = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}$$

Where,

y_i = true y values

$\hat{y}$ = predicted y values

$\bar{y}$ = mean y value

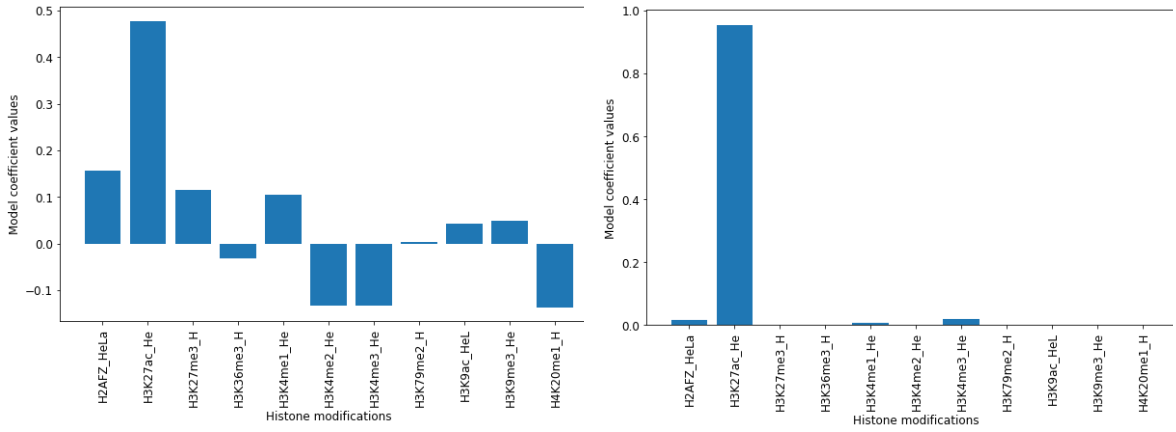
Table 1: Regression model error rates for HeLa Histone data

Model	**Overall Score**	**Training statistics**			**Testing statistics**		
		RMSE	**MSE**	**MAE**	**RMSE**	**MSE**	**MAE**
MLP Regressor	0.12	0.39	0.15	0.27	0.43	0.18	0.29
SVR	0.13	0.39	0.15	0.26	0.43	0.18	0.28
Decision Tree Regressor	0.12	0.40	0.16	0.28	0.43	0.18	0.29

All the models performed similarly on the dataset. The coefficient of R^2 for the models was closer to 0.0 rather than 1.0. This, combined with the high and consistent error values, indicates that the models were not working with sufficient accuracy.

2.2.2 Diagnosing the regression models

To understand the regression algorithms' inner workings and decision processes, we looked at the decision tree and MLP models' feature importance.



(a) MLP Regressor (b) Decision Tree regressor

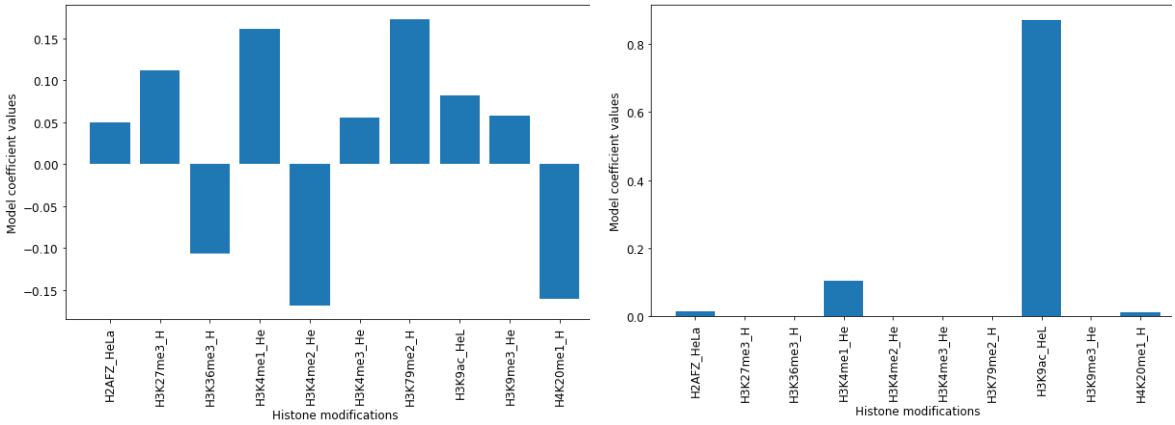
Figure 6: Feature importance of the regression models

For both the Decision Tree regressor and MLP model, almost all values were highly influenced by just a single histone modification (H3K27ac). It can point to a skewed feature dependence leading to erroneous predicted values. The single feature dependence may result from poorly chosen feature columns that interfere with other features or a very high dependence of enhancer potential values on the H3K27ac histone modification. The H3K27ac column was removed from the dataset, and the models were trained to test for the same. The results of the run are shown below.

Table 2: Regression Model error rates after removing H3K27ac

Model	**Overall Score**	**Training statistics**			**Testing statistics**		
		RMSE	**MSE**	**MAE**	**RMSE**	**MSE**	**MAE**
MLP Regressor	0.12	0.39	0.15	0.27	4.57	20.86	4.48
SVR	0.12	0.39	0.16	0.26	0.43	0.19	0.28
Decision Tree Regressor	0.11	0.41	0.17	0.28	0.43	0.19	0.29

The model performance slightly deteriorated after the removal of the H3K27ac column. The model fits well on the training dataset, but it cannot generalize the learning and, thus, not predict values well for the testing dataset. It indicates that the H3K27ac data was necessary for the generalization of the model learning. The models' feature importance after removing the H3K27ac column was expected to remain the same if there were no interactions between the columns. The model importance values were plotted to check if the model importance pattern changed.



(a) MLP Regressor (b) Decision Tree regressor

Figure 7: Feature importance of regression models after removal of H3K27ac data

It can be seen that the pattern of the importance of columns has changed after the removal of H3K27ac data. This can be attributed to eliminating interference between H3K27ac and other features. Thus, the removal of H3K27ac data is not the only factor leading to the models' deterioration.

2.3 Conclusion

Diagnosing the models by plotting their feature coefficients led to the realization that most of the predictions were based on only one feature (H3K27ac). The removal of the H3K27ac data led to the deterioration of the model performance. However, the removal of H3K27ac also led to drastic changes in the feature importance plots. The changes can be attributed to removing interference between the columns and removing a significant feature. Thus, better feature extraction methods need to be developed to improve the performance of the regression models.

Chapter 3

Extracting relevant features from the epigenetic datasets

Classification models are generally run to help in predicting the class types in categorical classification tasks. However, we can look into the feature coefficients of successful classification models and determine which features are essential to determine differences between the promoters and epromoters. This, in turn, can help us extract features that might perform better with regression models.
Data from the Histone, TF, and Enhancer Potential matrices were used to train classification models that can then be utilized as feature extraction pipelines to deduce essential features that may help build highly accurate regression models.

3.1 Materials and Methods

3.1.1 Dichotomizing the Enhancer Potential matrix

The histone modification, transcription factor, and enhancer potential data collected were used to train classification models and extract essential features from such models. Classification requires data to be in the form of discrete categorical classes. Thus, the enhancer potential values were used to divide promoters into two classes – promoters (promoter-like promoters) and epromoters (enhancer-like-promoters).
A threshold to differentiate between the two classes was determined to be 2.152 as follows (Vanhille et al., 2015):

- The enhancer potential data was first arranged in ascending order and plotted.
- The line formed by connecting the greatest and the smallest value in the curve was then slid across the curve till it formed a tangent to the curve.

- The data point where the line formed a tangent was then taken as the threshold value for distinguishing promoters from epromoters.

An R script was used to calculate the threshold for the datasets using only the enhancer potential matrix.

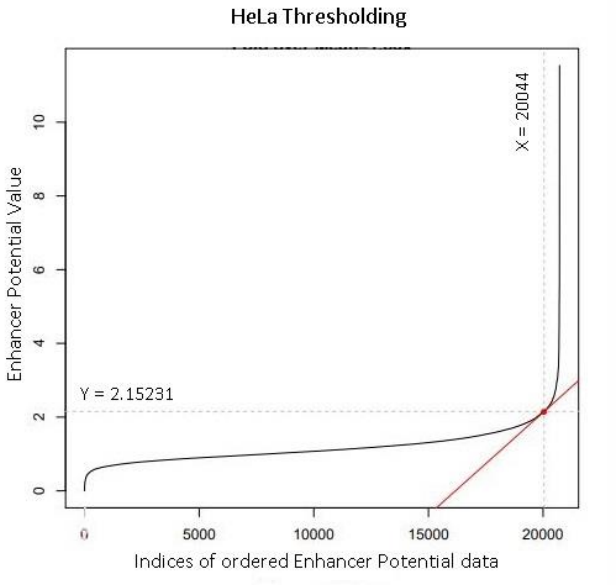


Figure 8: Enhancer potential thresholds for HeLa cells

3.1.2 Sampling to balance the dataset

The above dichotomizing the enhancer potential values created a high imbalance in the dataset (promoters ~ 96% & epromoters ~ 4%). SMOTE (Synthetic Minority Oversampling Technique) algorithm (Chawla, Bowyer, Hall, & Kegelmeyer, 2002) was chosen as the best method to oversample datapoints and balance the training dataset (see Appendix B).

3.1.2.1 SMOTE

SMOTE is perhaps the most widely used approach to balance datasets using oversampling of the minority class samples. First n-nearest neighbours of the minority class are calculated for all elements of the minority class. Lines are drawn, joining the n-nearest samples to the focal sample. Random points on the lines are taken as samples for the minority class. The figure below depicts the generation of synthetic minority samples using SMOTE.

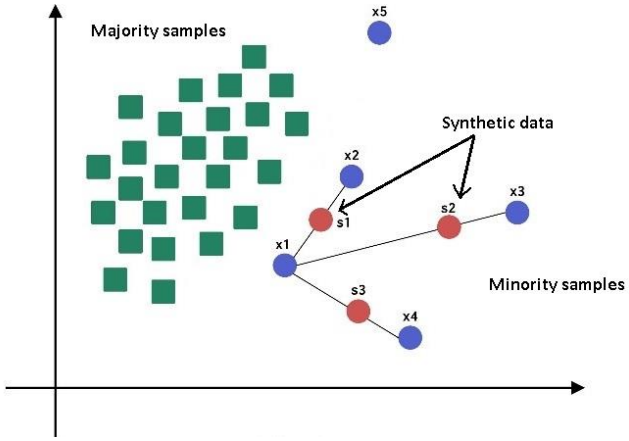


Figure 9: Graphical representation of SMOTE algorithm. Adapted from (Lopez, 2021)

3.1.3 Classification Models

Four classifiers were chosen for the task - Logistic Regression, Decision Tree Classifier, Support Vector Machine (SVM), and Multi-Layered Perceptron (MLP).

3.1.3.1 Logistic Regression

The logistic regression model is widely preferred among researchers because of its simplicity. The model is given by,

$$p(Y_i|X_i, \dots, X_p) = \frac{e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}}$$

Where, $X = (X_1, \dots, X_p)$ are the known variables used to predict the response variable Y. An observation with variables $x_1, \dots x_p$ should be assigned to the class j for which $p(Y = j|X_1 = x_1, \dots x_p)$ is largest. In binary classification, this corresponds to assigning an observation to class 1 if

$$p(Y = 1|X_1 = x_1, \dots, X_p = x_p) > 0.5$$

and to class -1 otherwise.

Training data is used to determine the coefficients $\beta_0, \dots, \beta_p$ of the model by minimizing the error rate by assigning each observation to its most likely class, conditioned on the variables' values (Brandt & Lanzén, 2021).

3.1.3.2 Decision Tree Classifier

Similar to a decision tree regressor (section 2.1.2.3), a decision tree classifier also builds a model that asks true-false questions and depending on the output narrows down a given data point to get its class. The model building and training is similar to that of a decision tree regressor, however, the leaf nodes in the case of a decision tree classifier represent the available class variables (y values).

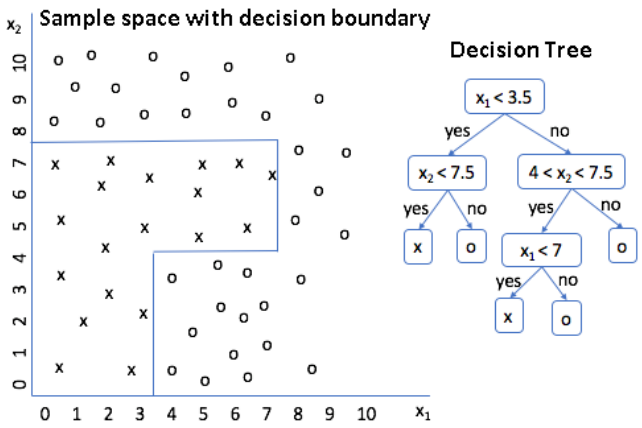


Figure 10: Example working of a decision tree classifier. Adapted from (Berry, Browne, & Omitaomu, 2006)

Based on the values of different input variables, x_1 and x_2 , and a trained decision tree, the model can arrive at a node and assign a certain class to a data point.

3.1.3.3 Multi-Layered Perceptron

A multi-layer perceptron classifier works very similarly to a multi-layer perceptron regressor. The only point of difference is the application of a final softmax activation layer that converts the numerical outputs of different nodes to probability values for a given class. The function is given by,

$$Softmax(x_i) = \frac{e^{x_i}}{\sum_j e^{x_j}}$$

As softmax converts a given output vector into a probabilistic representation, the sum of the softmax outputs for a given vector equal 1. In the context of a multi-layer perceptron classifier, suppose the two output nodes (in Figure 3) have values 2 and -3 for classes 'promoter' and 'epromoter'. Passing the values though a softmax activation function would provide values 0.993 and 0.007 for the respective classes, indicating that the chance that the data point is a promoter is 99.3%.

3.1.3.4 Support Vector Machine

A Support Vector Machine (SVM) constructs a hyperplane or a set of hyperplanes in a high-dimensional data space that can differentiate between observations of different class. A SVM classifier consists of a hyperplane and a functional margin instead of a ε-tube (SVR, Figure 4). The functional margin represents the largest distance from the training data points in any class on either side of the hyperplane. The data points lying on the functional margin are called the support vectors and the task of classification is posed as an optimization problem.

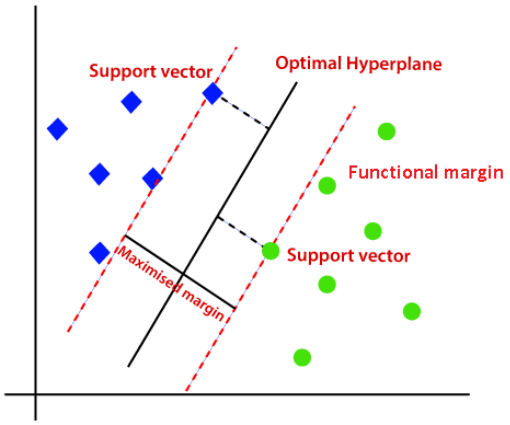


Figure 11: Graphical representation of SVM hyperplane. Adapted from (Javatpoint, 2018)

The aim of the optimization is to maximize the distance between the hyperplane and either of the functional margins. To make the optimization computationally efficient, kernels are used that can map data points to higher dimensional spaces where discovering a hyperplane is computationally easier.

3.1.3.5 Training the classification models

Models were built for each of the datasets (Histone and TF) for HeLa cells. Similar to the regression models, the data was first scaled using the standard scaler in the sci-kit learn library. The classification models were trained on two-thirds of the data, and one-third was used for validation. After splitting the dataset into testing and training datasets, the training dataset was oversampled using the SMOTE algorithm. The total number of samples in the testing and validation datasets for both cell lines and kinds of epigenetic factors are outlined in Table 3.

Table 3: Number of samples in cleaned HeLa datasets

Cell Line	**Dataset**	**No. of Training Samples**		**No. of Validation Samples**		**No. of Features**
		No. of promoters	**No. of epromoters**	**No. of promoters**	**No. of epromoters**	
HeLa S3	Histone	14419	14419	6181	35	11
	TF	14416	14416	6182	33	17

After optimizing the different models' performance, the coefficients for feature importance were extracted from the models.

3.2 Results

3.2.1 Model Evaluation

Various metrics were used for the evaluation of the different classification models. These include confusion matrices, precision, recall values, and ROC-AUC curves.

3.2.1.1 Confusion Matrices

Confusion matrices are tables used to describe a classification model's performance on a set of data for which the actual values are known. The basic structure of a confusion matrix is shown below.

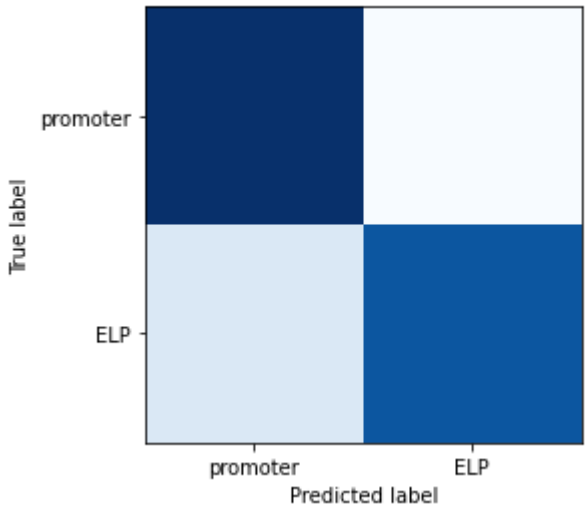
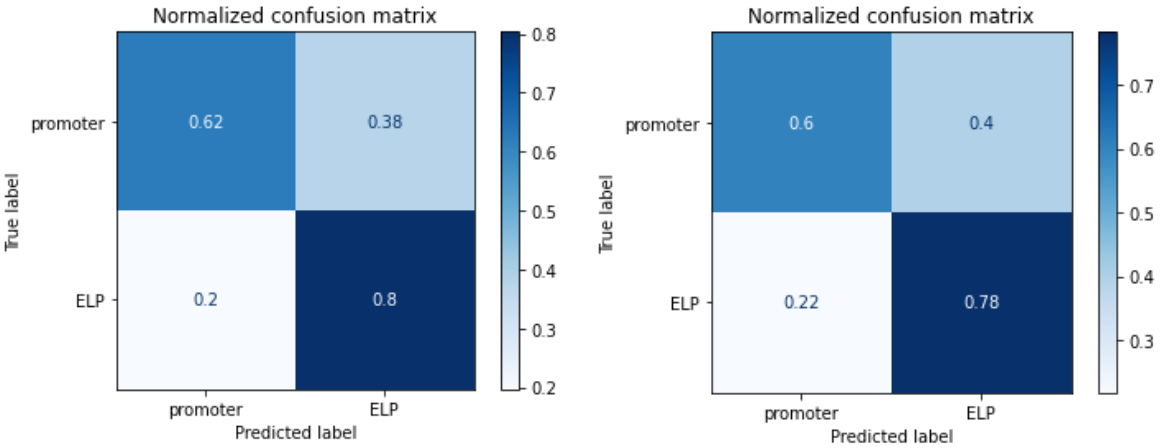


Figure 12: Structure of a confusion matrix

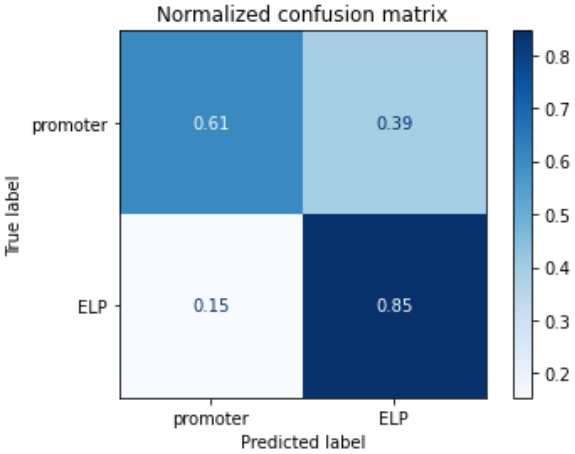
A binary confusion matrix has four classes (quadrants). A true label represents the known class of a data point, whereas a predicted label represents the class the model predicts the data point. In our binary classification problem, the epromoter (ELP) class is a positive class, and the promoter class is negative. The upper left quadrant in Figure 12 represents the number of real promoters predicted as promoters (True Negatives or TN). The lower right quadrant represents the number of real epromoters classified as epromoters (True Positives or TN). The upper right quadrant represents the number of real promoters misclassified as epromoters (False Positives or FP). The lower left quadrant represents the number of real epromoters misclassified as promoters (False Negatives or FN).
As the validation datasets were imbalanced, normalized confusion matrices were calculated and plotted. The "normalized" term means that each of the true labels is represented as having 1.00 samples. Thus, the sum of each row in a normalized confusion

matrix is 1.00 as each row represents the total number of elements in a particular class (Simske, 2019).

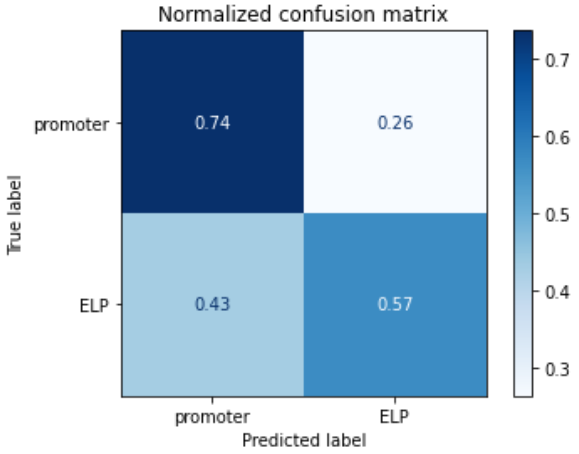


(a) Logistic Regression

(b) Decision Tree Classifier

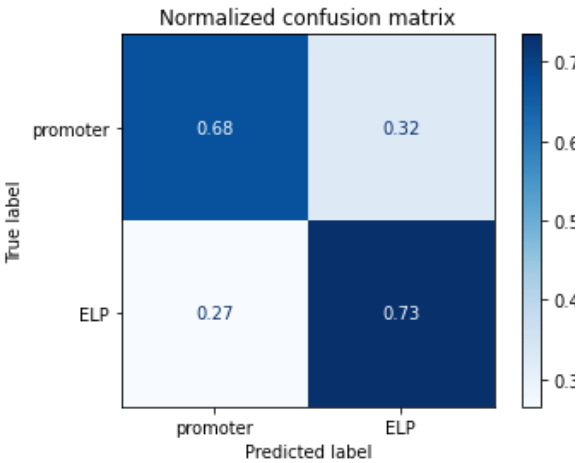


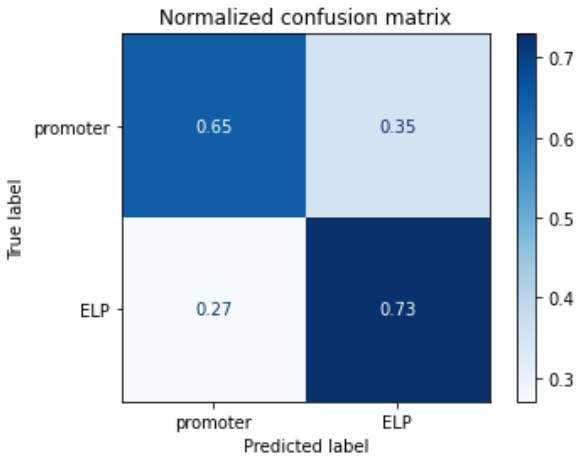
(c) MLP classifier



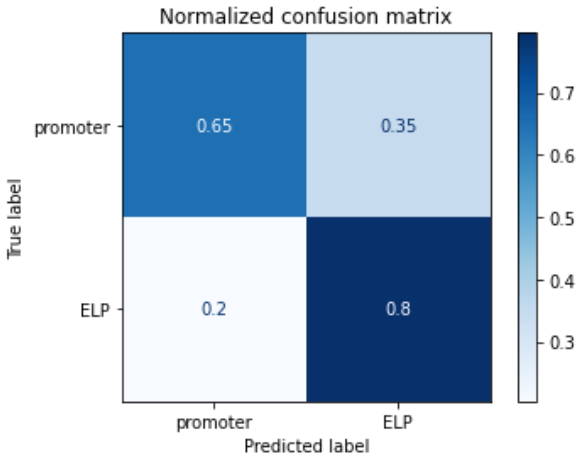
(d) SVM classifier

Figure 13: Confusion matrices for HeLa S3 histone dataset

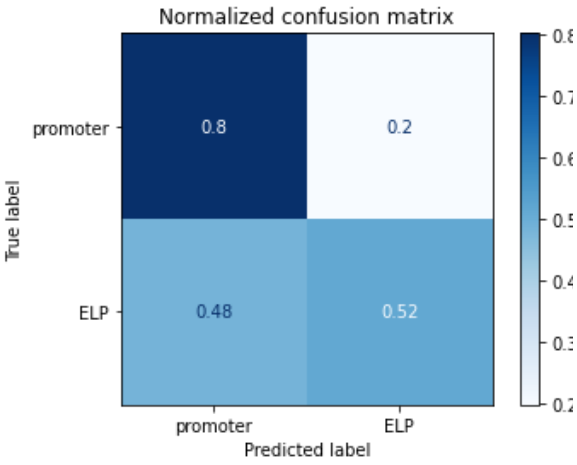




(a) Logistic Regression



(b) Decision Tree Classifier



(c) MLP classifier

(d) SVM classifier

Figure 14: Confusion matrices for HeLa S3 TF dataset

The confusion matrices indicate that the MLP classifier obtained the highest predictive accuracy for the epromoter class in both the histone modification and the transcription factor datasets. The logistic regression and decision tree classifier models performed similarly (~ 60 - 70% class prediction), whereas the SVM classifier failed to classify epromoter samples correctly. The best model trained was the MLP classifier for the TF dataset.

3.2.1.2 Precision-Recall values

Precision values represent the proportion of positive identifications of a positive class. Recall is a measure of the number of actual positives that were identified correctly (Google Developers, 2020).

Precision and recall values for a given class are calculated as follows:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

A precision value of 0.5 indicates that if a model labels a specific datapoint as positive, there is a 50% chance that it is positive. A precision value of 1.0 indicates that any point classified by the model as positive is a positive data point. Conversely, a precision value

of 0.0 indicates that any point classified by the model as positive is not positive. A higher precision value for a class indicates that the model does not mislabel other non-class datapoints as datapoints of the given class.

Recall = no. of correctly labelled class items/ (no. of correctly labelled class items + no. of incorrectly labelled class items) = TP / (TP + FN)

A recall value of 0.5 indicates that the model correctly classifies 50% of all class data points. A recall value of 1.0 indicates that the model has identified all the points belonging to the given class. Conversely, a recall value of 0.0 indicates that the model did not identify even a single data point belonging to the class. A higher recall value of a class in a model indicates that a model can find most of the data points belonging to the class.

Table 4: Precision-recall values for the classification models and HeLa datasets

Dataset	**Model**	**Precision**		**Recall**	
		Promoter class	**Epromoter class**	**Promoter class**	**Epromoter class**
HeLa Histone	Logistic Regression	0.99	0.06	0.62	0.80
	Decision Tree Classifier	0.99	0.06	0.6	0.78
	MLP	0.99	0.06	0.61	0.85
	SVM	0.98	0.06	0.74	0.57
HeLa TF	Logistic Regression	0.99	0.07	0.68	0.73
	Decision Tree Classifier	0.99	0.06	0.65	0.73
	MLP	0.99	0.07	0.65	0.80
	SVM	0.98	0.08	0.80	0.52

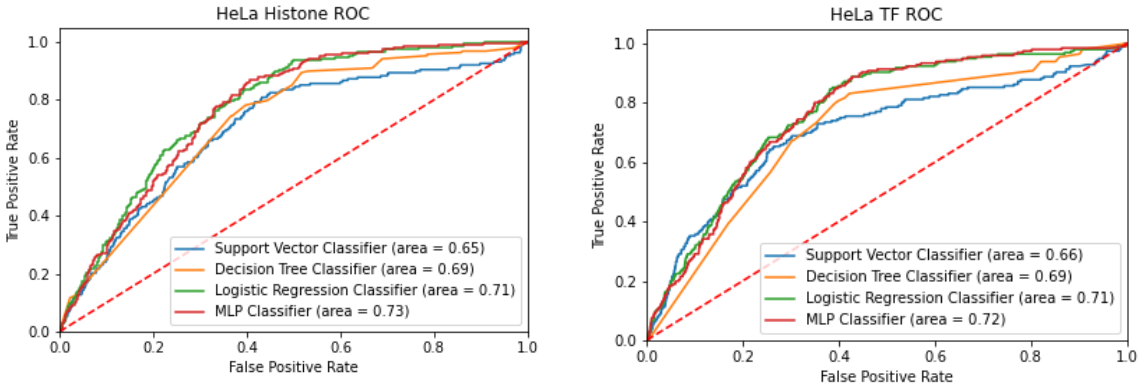
The highly skewed values for precision for the promoter and epromoter classes are expected as the testing dataset is highly imbalanced. The class imbalance means that the number of epromoters is too less compared to the number of promoters; thus, there are not many epromoters that can be mislabelled as promoters (skewing the promoter precision value towards 1.0), and there is an overwhelming number of promoters that can be misclassified as epromoters (skewing the epromoter precision value towards 0.0).

The promoter class's recall values are almost similar and constant for the MLP, Decision Tree, and Logistic Regression models. In contrast, the recall values in the SVM classifier case are higher compared to the other models. The recall values for the epromoter class are lower in SVM compared to the other models. This is indicative of the poor performance of the SVM classifier. The SVM classifier tends to classify most points as promoters, and due to the class imbalance in the test dataset, it gives rise to higher recall values for the promoter class and lower recall values for the epromoter class. The MLP model has the highest recall values for the epromoter class, proving that the MLP model is the best classifier to identify epromoter data points in a given dataset of promoters and epromoters.

3.2.1.3 ROC-AUC curves

A ROC curve is a graphical plot used to show the predictive ability of binary classifiers. It is constructed by plotting the true positive rate (TPR) against the given models' false positive rate (FPR). The true positive rate is calculated as the proportion of correctly predicted observations to be positive (number of correctly predicted epromoters) out of all the positive observations (total number correct predictions). Similarly, the false positive rate is the proportion of incorrectly predicted observations to be positive (number of promoters wrongly predicted to be epromoters) out of all negative observations (total number of incorrect predictions). The ROC curve can then be reduced to a single metric called the area under the curve (AUC). The greater the area under the curve of a given model, the better is the model. A perfect predictive model should have a curve closer to the upper left corner of the plot. In contrast, a random model with no capacity to distinguish between positive and negative classes has an AUC value of 0.5 and a curve represented by the diagonal dotted line(Sarang, 2018).

ROC plots for all the models and a given dataset were plotted (the bracket values indicate the AUC values). The plots for the different datasets are given below.



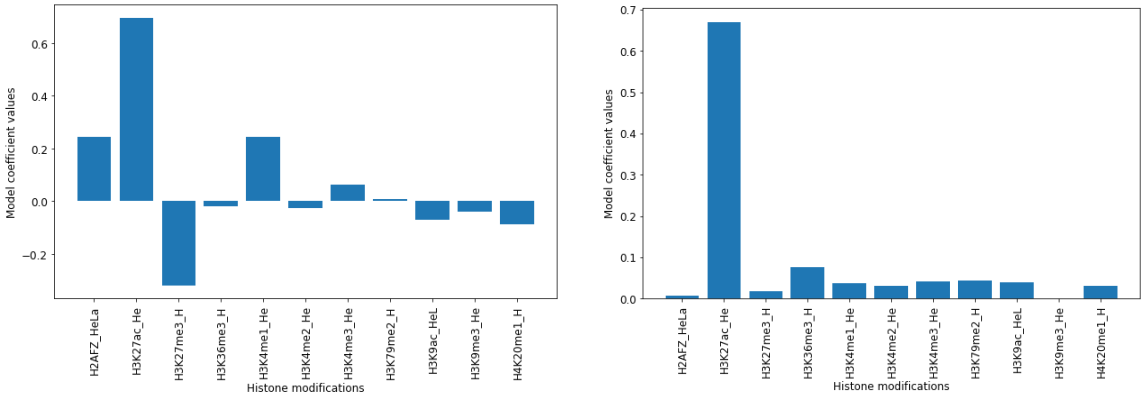
(a) HeLa Histone dataset (b) Decision Tree regressor

Figure 15: ROC-AUC plots for comparison of classification models

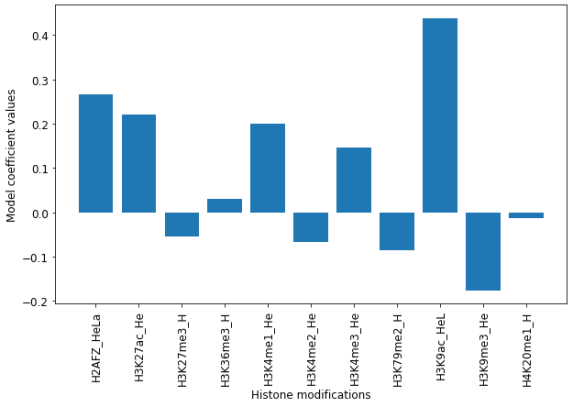
The plots for HeLa S3 (both histone and TF datasets) support the results from the confusion matrices and the precision-recall values. The plots indicate that the MLP is the best classifier, followed by the Logistic Regression and Decision Tree classifiers. The SVM is the least preferred classifier in both datasets due to its low AUC value.

3.2.2 Extracting Feature Importance from the Classification models

The aim of using classification models was to improve the regression models by extracting essential features. Using methods provided in the sci-kit learn library, the coefficients of different features were extracted from the Decision Tree, MLP, and Logistic Regression models. Features from the HeLa Histone ratio (Appendix C) dataset were not extracted due to the model's lower accuracy.



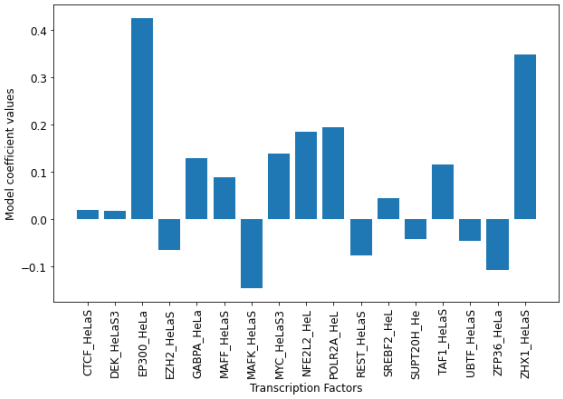
(a) Logistic Regression coefficients (b) Decision Tree Classifier coefficients



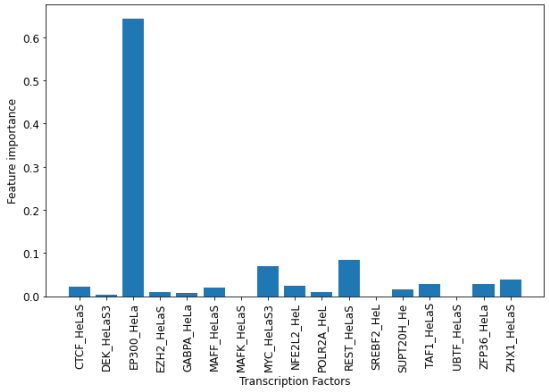
(c) MLP classifier coefficients

Figure 16: Feature importance values for the HeLa S3 histone dataset

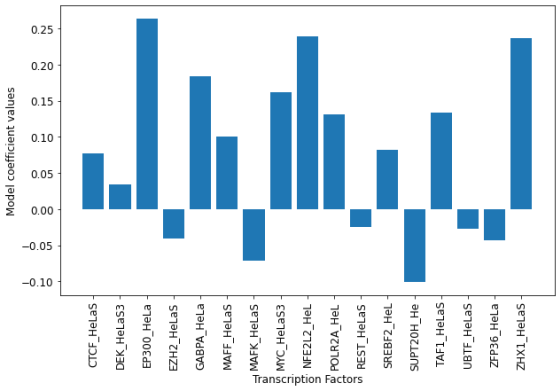
Essential features in the Logistic Regression classifier were H3K27ac and H3K27me3. In the decision tree classifier, the only most important feature is H3K27ac which is in accordance with the features from the Logistic Regression. The MLP classifier, however, has H3K9ac, H2AFZ, and H3K27ac as essential features.



(a) Logistic Regression coefficients



(b) Decision Tree Classifier coefficients



(c) MLP classifier coefficients

Figure 17: Feature importance values for the HeLa S3 TF dataset

The Logistic Regression model indicates that EP300 and ZHX1 are essential transcription factors needed to differentiate epromoters from promoters in the TF dataset. The Decision Tree classifier has one crucial feature (EP300) in accordance with the Logistic Regression model. The MLP classifier shows that EP300, NFE2L2, and ZHX1 are essential features for the differentiation of epromoters from promoters.

3.3 Conclusion

Of the four classification methods utilized, the MLP classifier and the Logistic Regression classifier performed the best in almost all the datasets. The SVM classifier was unable to classify epromoters in most of the datasets. Thus, features were only extracted from the MLP, Logistic Regression, and Decision Tree models. A literature survey revealed that the histone markers deemed essential from the models (H3K27ac, H3K27me3, and H3K9ac) served as important biological markers for distinguishing active enhancers from poised enhancers and typical promoters (Cai et al., 2021; Creyghton et al., 2010). Similarly, for the TF dataset, the most critical features (EP300, NFE2L2, and ZHX1) play an active part in gene regulatory mechanisms (Eckner et al., 1994a; Pajares et al., 2016; Yamada, Printz, Osawa, & Granner, 1999).
The extracted features have some biological relevance in distinguishing promoters and epromoters. Thus, the features extracted from the classification models can help in building more accurate regression models.

Chapter 4

Extending Analysis to K562 cell line

Building a generalized framework that can apply to different cell lines, including non-cancer mammalian cells, has been one of the thesis's main aims. Extending the analysis to a second cancer cell line is one step towards the generalized model. The methodology discussed in the previous sections has been applied to create models for enhancer potential data of the K562 cell line.

4.1 Materials and methods

4.1.1 Data Collection

Enhancer potential data for K562 cells were collected from CapStarr-seq experiments (Dao et al., 2017). Data for histone modifications and transcription factor bindings were downloaded from ENCODE. The data collected was organized as matrices as mentioned in section 2.1.1

4.1.2 Dichotomizing enhancer potential matrix

A threshold was determined for the K562 enhancer potential dataset using the algorithm given in 3.1.1. The threshold was determined to be 1.764. Any promoter having an enhancer potential greater than 1.764 was classified as an enhancer-like promoter (epromoter), and promoters with a lower value were classified as a promoter-like promoter (promoter).

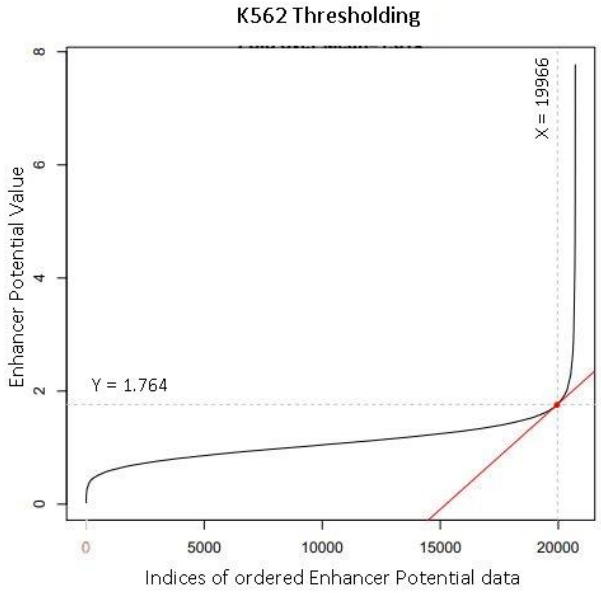


Figure 18: Enhancer potential threshold for K562 cells

4.1.3 Classification Models

The data (histone and TF matrices) was then scaled using the standard scaler (sci-kit learn) and split into training and validation datasets in the ratio 1:2. The training data was oversampled using SMOTE algorithm to balance the number of promoters and epromoters, and the same four classification models (detailed in 3.1.3) were trained on the data. The final number of training samples in the cleaned and balanced dataset are outlined in Table 5.

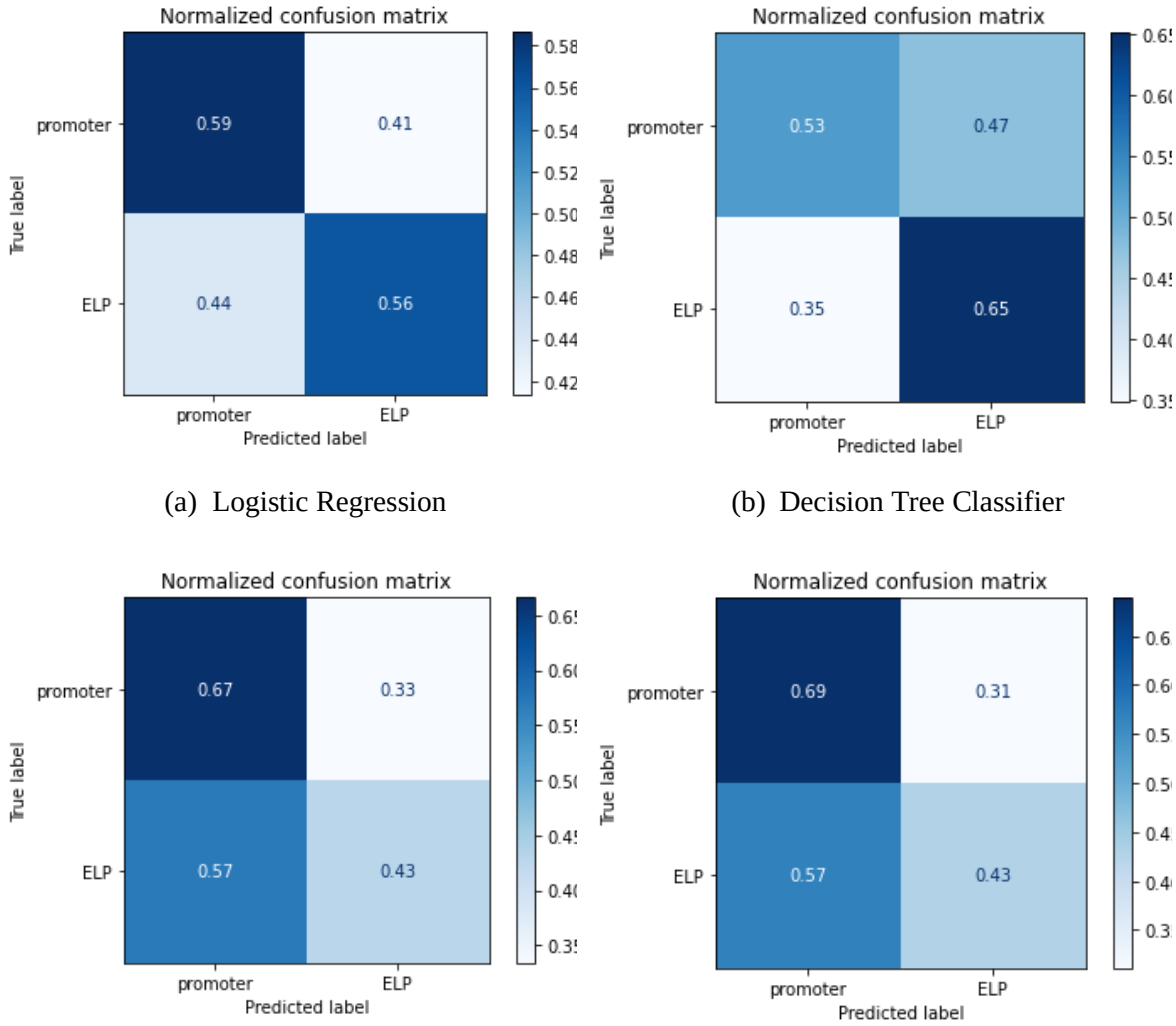
Table 5: Number of samples in cleaned K562 datasets

Cell Line	**Dataset**	**No. of Training Samples**		**No. of Validation Samples**		**No. of Features**
		No. of promoters	**No. of epromoters**	**No. of promoters**	**No. of epromoters**	
K562	Histone	13969	13969	5995	221	12
	TF	13969	13969	5995	221	14

4.2 Results

4.2.1 Model Evaluation

4.2.1.1 Confusion Matrices



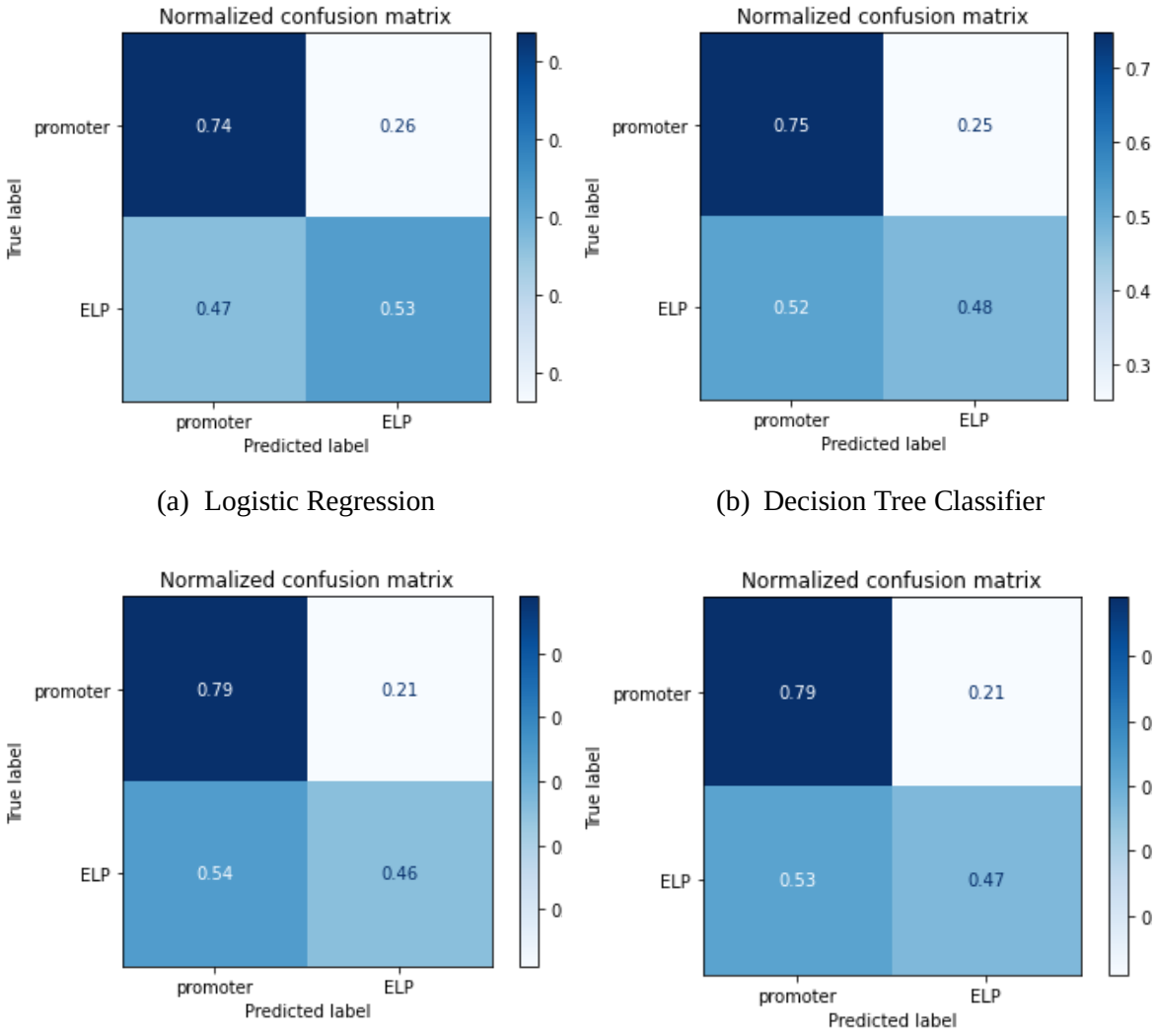
(a) Logistic Regression

(b) Decision Tree Classifier

(c) MLP classifier

(d) SVM classifier

Figure 19: Confusion matrices for K562 Histone dataset



(a) Logistic Regression

(b) Decision Tree Classifier

(c) MLP classifier

(d) SVM classifier

Figure 20: Confusion matrices for K562 TF dataset

The confusion matrices show that the classifiers for the K562 dataset are not high-performing. The best models are marginally better than a random classifier (a model that randomly spits out class labels given the datapoint). The SVM and MLP models for the histone dataset predict most of the values as promoters and thus, end up predicting epromoters as promoters. In the TF dataset, all the models perform slightly better than the K562 histone dataset models.

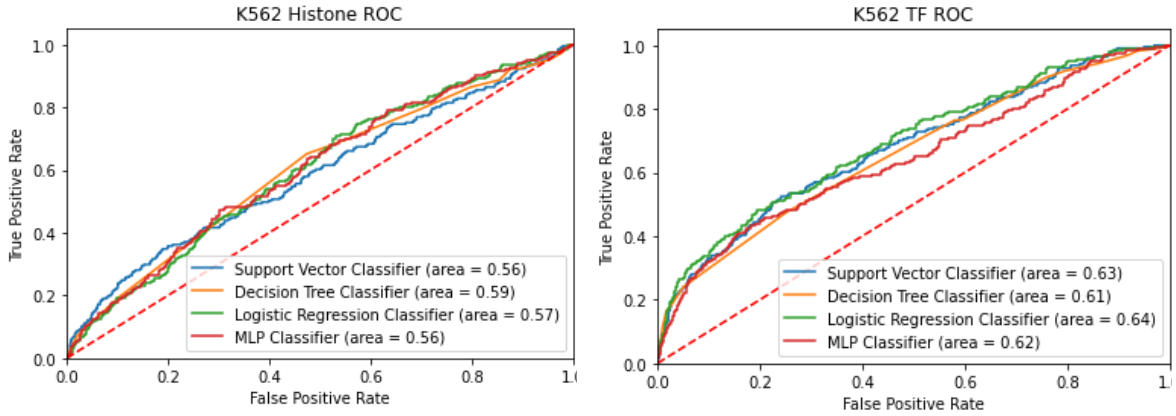
4.2.1.2 *Precision-Recall values*

Table 6: Precision-recall values for the classification models and K562 datasets

Dataset	Model	Precision		Recall	
		Promoter class	**Epromoter class**	**Promoter class**	**Epromoter class**
K562 Histone	Logistic Regression	0.97	0.05	0.59	0.56
	Decision Tree Classifier	0.98	0.05	0.53	0.65
	MLP	0.97	0.05	0.67	0.43
	SVM	0.97	0.05	0.69	0.43
K562 TF	Logistic Regression	0.98	0.07	0.74	0.53
	Decision Tree Classifier	0.97	0.06	0.75	0.48
	MLP	0.98	0.07	0.79	0.46
	SVM	0.98	0.08	0.79	0.47

In the K562 datasets, the recall values indicate that the best performing model is the Decision Tree model for the histone dataset. Although the model has a higher recall value than the other K562 models, the value is not high enough to accurately predict all epromoters in a given dataset. The rest of the models perform poorly compared to the Decision Tree model and the HeLa dataset models.

4.2.1.3 *ROC-AUC curves*



(a) Histone dataset (b) TF dataset

Figure 21: ROC-AUC curves for the K562 classification models

The plots for the K562 dataset support the notion that the models are not entirely accurate. The models (particularly the histone dataset models) are pretty close to the dotted line. The dotted line represents a model that has no discriminatory power between both classes. Thus, the ROC curve and the AUC values are indicative of the poor performance of the models.

4.3 Conclusion

The classification models performed poorly on the K562 dataset. The highest accuracy obtained was around 64% (Logistic Regression model on TF data), but this model also had a recall value of around 0.5 for the epromoter class. Indicating that it missed almost half of the epromoters from the dataset. As the different models and training methods were tested for the HeLa cell data and performed well, the low model accuracy in K562 cells can be attributed to either the dataset or the cell line. Further analysis (feature extraction) was not carried out using the models.

Chapter 5

Mathematical Modelling of promoter-promoter interactions

A mathematical framework for promoter-promoter interactions has been laid out in this chapter. The framework models interactions between spatially interacting promoters and uses the enhancer potential value for different promoters to quantify changes in different interacting promoters' activation states.

5.1 Materials and Methods

5.1.1 Data Collection

Raw fastq files containing Hi-C sequence reads of HeLa cells were downloaded from ENA (European Nucleotide Archive). The entire reference genome for hg19 was downloaded from the Genome Research Consortium using the UCSC Genome Browser.

5.1.2 Hi-C Pre-processing

Hind III was identified as the restriction enzyme used for digestion from the Hi-C data experimental paper (Lieberman-Aiden et al., 2009; Naumova et al., 2013). HindIII cuts a given DNA sequence at A^AGCTT palindromic sites producing sticky ends. A restriction digestion map of the whole hg19 genome (Chr 1-22, Chr X, Chr Y, and Chr M) was prepared using the hicup_digester script. The digested genome file contains all fragments of the genome's total digestion by Hind III restriction digestion. Bowtie2 was used to create index files for the hg19 genome. A bowtie index file is an aligned representation of the raw genome sequences. It helps reduce the whole genome's memory footprint and provides a compatible format that various tools can use for alignment and processing. SolexaQA determined the format of the fastq sequence files to be "Sanger" format.

The index files, digested genome segments, and Sanger fastq files were processed using the HiCUP (Hi-C User Pipeline) pipeline. HiCUP pipeline consists of six Perl scripts for processing Hi-C sequence data. The **hicup_truncater** cut the fastq reads at putative Hi-C ligation junctions. The **hicup_mapper** takes in output from the hicup_truncater and bowtie index files to align the sequences to the reference genome. The reference genome is digested using the **hicup_digester**. The **hicup_filter** script takes the digested genome (hicup_digester) and mapped sequences (hicup_mapper) and removes commonly encountered Hi-C artefacts. Finally, the **hicup_deduplicator** removes (retaining one copy) putative PCR duplicates. The sixth Perl script, titled "**hicup,**" executes all the other scripts sequentially to automatically produce BAM files from raw fastq reads (S. Wingett et al., 2015).

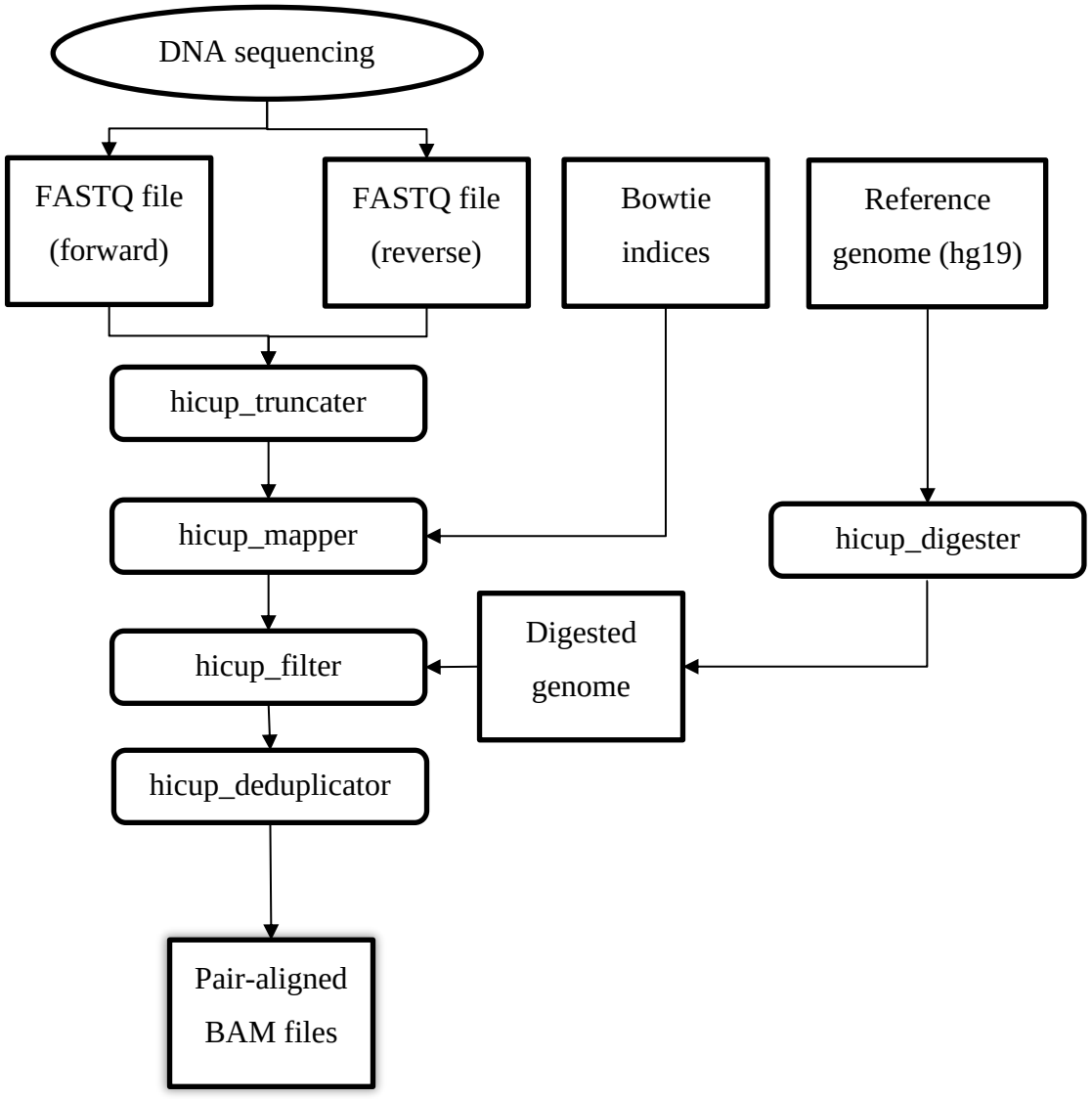


Figure 22: Pre-processing of Hi-C data using the HiCUP pipeline. Adapted from (S. W. Wingett et al., 2015)

The pipeline produces paired-read BAM files representing the filtered di-tags ("HiCUP Overview — HiCUP 0.8 documentation," n.d.). Each read pair in the BAM file corresponds to a putative Hi-C di-tag.

5.1.3 Building Interaction Matrix

HiCUP provides additional tools to convert the final processed Hi-C BAM file to different formats compatible with a wide range of post-processing tools. The hicup2gothic script (from HiCUP tools) was used to convert the BAM file into a format compatible with the GOTHiC tool. GOTHiC is a tool to extract significant interactions from the aligned pair read files using a simple binomial probabilistic model that resolves complex biases and distinguishes between true and false interactions. The tool returns a lognormal probability of the observed number of interactions vs. the expected number of interactions between two given regions at a given resolution (Mifsud et al., 2017). The final converted BAM output and the restriction digestion file were then put through the gothic pipeline. All cis-trans interactions at a resolution of 10kb were identified. GOTHiC uses a probabilistic model to identify significant interactions in the processed Hi-C data. A custom python script was written to filter all the promoters from the above interaction matrix and prepare a final interaction matrix containing only those genomic locations as rows and columns, which appeared in the range of promoter locations for the respective epigenomic dataset.

5.1.4 The Mathematical Model

The following equation was used to model the interactions of the promoters:

$$\frac{dx_i}{dt} = f(x_i) + \sum_i W_{ij} g(x_j)$$

$$g(x_j) = \frac{x_j e_j}{1 + x_j . e_j}$$

Where,

x_i = activation state of promoter i

x_j = activation state of promoter j

W_{ij} = spatial interaction between promoter i and j

$f(x_i)$ = function marking the internal dynamics of promoter i

e_j = enhancer potential of promoter j

The function g(x) that marks the external interactions between different promoters is modelled as a hill function equation (Santillán, 2008). Thus, the expected curve when using only the external interactions is a sigmoid.

5.1.5 Simulating Promoter-Promoter interactions

The equation was implemented in Python 3.8. The enhancer potential values for HeLa cells were plugged in from the CapStarr-seq data. The data from the prepared interaction matrix was used to calculate the effect of promoters on other given promoters. The function representing the internal dynamics of the promoters was ignored. The initial activation states were set to 0.1 (to replicate baseline activity). Finally, time course plots of the change in activation states were plotted for the interactions for ten iterations (timesteps).

5.2 Results

5.2.1 Promoter Interaction Map

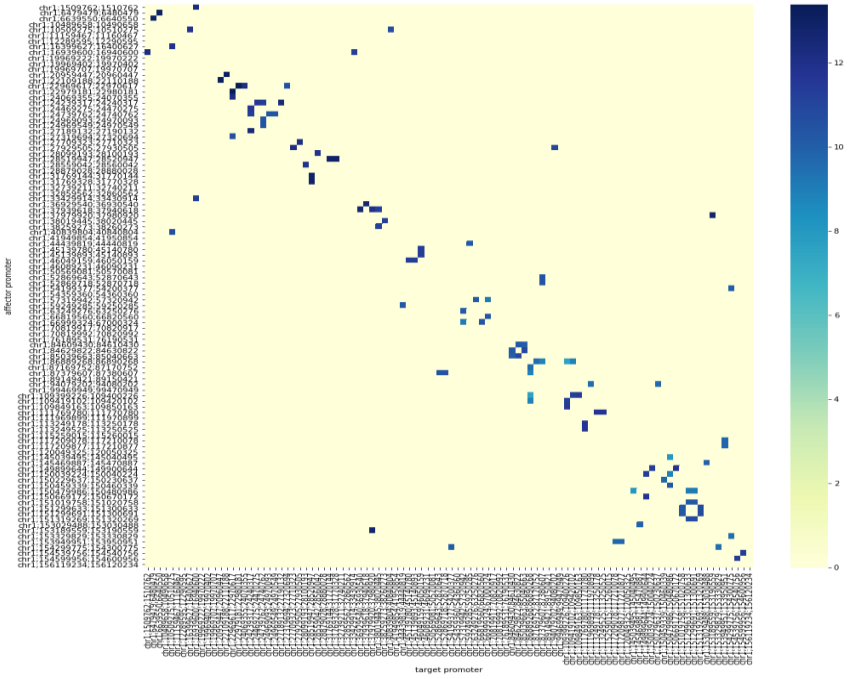


Figure 23: Promoter interaction map for the first 100 interacting promoters

The promoter interaction map is sparse and resembles a Hi-C contact matrix indicating that a given promoter only interacts with a few more promoters. There may not be very dense connections in the promoter-promoter network.

5.2.2 Characteristics of the promoter-promoter interaction graph

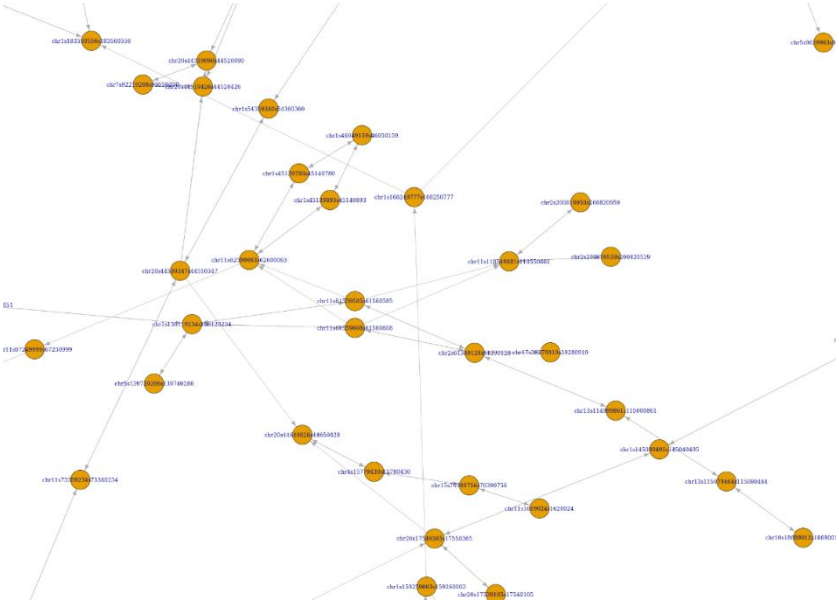


Figure 24: Enlarged view of the promoter-promoter interaction graph

The various characteristics of the promoter-promoter network are detailed in Table 7.

Table 7: Characteristics of the promoter-promoter network

Characteristic	Value
Nodes	970
Edges	1668
Graph Density	0.0017
Global clustering coefficient	0.018
Assortativity Coefficient	-0.073

The graph density is high if there are a more significant number of edges between the different nodes. A value of 0.0017 is relatively low, indicating the sparseness of the network. The assortativity coefficient represents the kind of interactions that the nodes undertake. A higher positive value means similar nodes (nodes with comparable degrees) tend to be more connected than dissimilar nodes, while a negative value indicates dissimilar nodes form connections. The assortativity coefficient's slight negative value indicates dissimilar nodes (nodes with a high difference in their degrees) tend to form more contact. However, the extent of this interaction is weak.

5.2.3 Modelling of promoter-promoter interactions

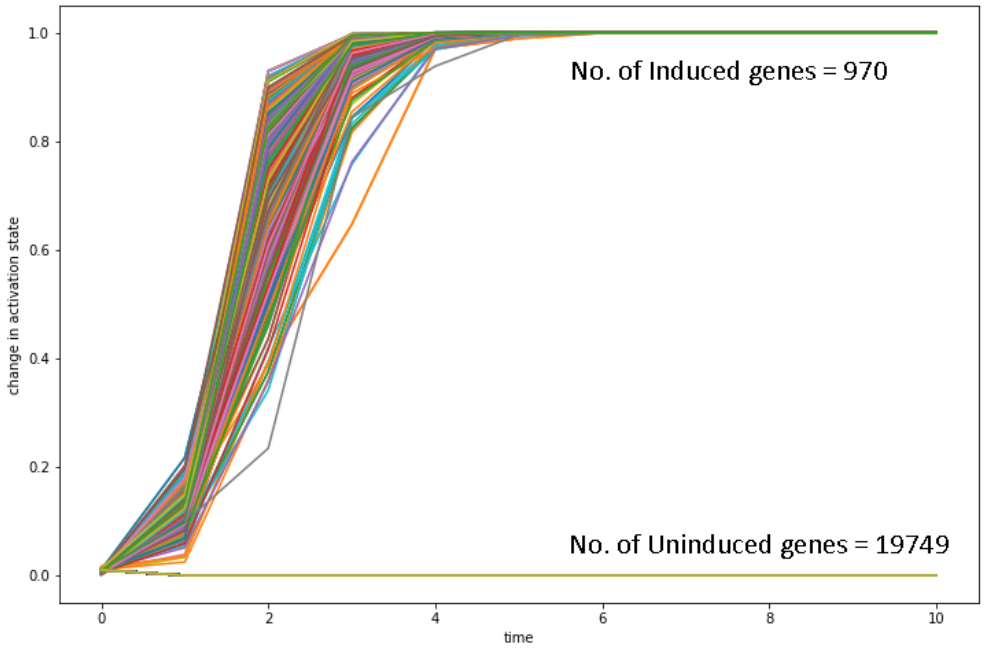


Figure 25: Change in the activation state of promoters (basal induction = 0.1)

Promoters with at least one interacting partner (a part of the promoter-promoter network) were induced to some higher value than the basal induction. As the internal dynamics term was ignored, the basal induction rate was not sustained in promoters that were not induced by other promoters. This gives us a clue that the internal dynamics equation should sustain a given basal rate of induction. The induced genes quickly rose to an activation state of 1.0 and were constant at that value. This can be attributed to the lack of deregulation (negative interactions) in the external interaction equation. As expected, the interactions produce a sigmoid curve in the induced promoters.

Discussion

Some attempts at characterising the promoters were fruitful while some were not. The error rates of the initial regression models were high, and the R^2 coefficient was relatively low. This indicates that the models did not satisfactorily predict the enhancer potential values given the data. Various treatments were used to improve the performance of the regression models (Appendix A) and it was noted that removing outlier enhancer potentials and min-max scaling of the histone matrix led to a slight improvement in the predictions. However, the improvements were not significant enough to build an accurate model. Scaling of the enhancer potential values resulted in deteriorated models indicating that it is not an appropriate approach to increasing the models' efficiency. The steep decline in the accuracy after scaling the enhancer potential values may be due to the diminishing difference between the data points resulting from the scaling.

To build better regression models, classification models were used as pipelines to extract relevant features from the datasets. The classification models determined that essential features needed to differentiate epromoter from promoters for the histone dataset were H3K27ac, H3K27me3, H3K9ac, and H2AFZ. It has been shown in the literature that the H3K27ac in a combination of H3K4me1 are histone markers used to differentiate active from inactive enhancers (Creyghton et al., 2010; Dao & Spicuglia, 2018). H3K27me3 is known to function as a silencer to repress gene expression (Cai et al., 2021) and has been known to associate with super-enhancers (Hnisz et al., 2013). It has also been acknowledged that such gene silencing regions may transform into active enhancers depending on cell line and cell developmental stage (Ngan et al., 2020). The H3K9ac histone marker has been associated with putative enhancer regions (Ernst et al., 2011). No significant difference between H2AFZ is promoters and enhancers has been described.

In the TF dataset, the most important features were determined to be EP300, NFE2L2, and ZHX1. The EP300 gene encodes for a histone acetyltransferase protein (p300) which plays an active hand in chromatin remodelling (Eckner et al., 1994b). Transcription factor NFE2L2 acts as a regulator of macroautophagy genes in mouse cells (Pajares et al., 2016). The ZHX1 transcription factor interacts with the Nf-Y complex and acts as a TSS selection mechanism in animal cells, and regulates the transcription initiation (Oldfield et al., 2019). Thus, all the transcription factors identified as essential features serve

regulatory functions and are more likely to be biologically associated with enhancer-like promoters regulating other promoters' expression.

It is evident from the literature that the features deemed necessary in the classification models have some biological relevance in distinguishing promoters and epromoters. Thus, the classifiers built can effectively extract relevant features that can then help in predicting the enhancer potential of promoters given their histone modifications and transcription factor bindings. One of the future directions of the project would be to build regression models using these extracted features and check whether the model can make accurate predictions.

The classification models performed poorly for the K562 datasets. As the models performed relatively well for similar data on HeLa cells, it is assumed that the issue lies with the K562 dataset or the cell line. Underlying causes such as poor experimental data may lead to the model's poor performance. It may also be the case that the nature of K562 cells caused the poor performance of classification models. This can be biologically rationalized if there is are pathways other than histone modifications and transcription factor bindings that can lead to enhancer-like activity in promoters.

Finally, the framework developed to study promoter-promoter interactions was successfully used to generate interaction and activation plots in a given cell type. Various improvements can be made to the model to add features such as negative regulation and sustenance of basal induction rate in non-interacting promoters. The mathematical framework allows us to infer what promoters will activate other promoters. This inference is significant in interpreting the long-range indirect genotype-phenotype association in the genome often seen in GWAS SNPs (Schierding et al., 2014). The time-course gene expression data generated from the model can be tested using observed time-course gene activation data to infer novel routes through which regulatory circuits may function.

References

Altmüller, J., Palmer, L. J., Fischer, G., Scherb, H., & Wjst, M. (2001). Genomewide scans of complex human diseases: True linkage is hard to find. *American Journal of Human Genetics*, Vol. 69, pp. 936–950. https://doi.org/10.1086/324069

Amari, S. ichi. (1993). Backpropagation and stochastic gradient descent method. *Neurocomputing*, *5*(4–5), 185–196. https://doi.org/10.1016/0925-2312(93)90006-O

Andersson, R. (2015). Promoter or enhancer, what's the difference? Deconstruction of established distinctions and presentation of a unifying model. *BioEssays*, *37*(3), 314–323. https://doi.org/10.1002/bies.201400162

Awad, M., Khanna, R., Awad, M., & Khanna, R. (2015). Support Vector Regression. In *Efficient Learning Machines* (pp. 67–80). https://doi.org/10.1007/978-1-4302-5990-9_4

Belmont, J. W., Boudreau, A., Leal, S. M., Hardenbol, P., Pasternak, S., Wheeler, D. A., … Stewart, J. (2005). A haplotype map of the human genome. *Nature*, *437*(7063), 1299–1320. https://doi.org/10.1038/nature04226

Belmont, J. W., Hardenbol, P., Willis, T. D., Yu, F., Yang, H., Ch'Ang, L. Y., … Tanaka, T. (2003). The international HapMap project. *Nature*, *426*(6968), 789–796. https://doi.org/10.1038/nature02168

Berry, M. W., Browne, M., & Omitaomu, O. A. (2006). DECISION TREES. In *Lecture Notes in Data Mining* (pp. 39–51). https://doi.org/10.1142/9789812773630_0004

Brandt, J., & Lanzén, E. (2021). *A Comparative Review of SMOTE and ADASYN in Imbalanced Data Classification*.

Bush, W. S., & Moore, J. H. (2012). Chapter 11: Genome-Wide Association Studies. *PLoS Computational Biology*, *8*(12). https://doi.org/10.1371/journal.pcbi.1002822

Cai, Y., Zhang, Y., Loh, Y. P., Tng, J. Q., Lim, M. C., Cao, Z., … Fullwood, M. J. (2021). H3K27me3-rich genomic regions can function as silencers to repress gene expression via chromatin interactions. *Nature Communications*, *12*(1), 1–22. https://doi.org/10.1038/s41467-021-20940-y

Calo, E., & Wysocka, J. (2013, March 7). Modification of Enhancer Chromatin: What, How, and Why? *Molecular Cell*, Vol. 49, pp. 825–837. https://doi.org/10.1016/j.molcel.2013.01.038

Chanklan, R., Kaoungku, N., Suksut, K., Kerdprasop, K., & Kerdprasop, N. (2018). Runoff

prediction with a combined artificial neural network and support vector regression. *International Journal of Machine Learning and Computing*, *8*(1), 39–43. https://doi.org/10.18178/ijmlc.2018.8.1.660

Chawla, N. V, Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. In *Journal of Artificial Intelligence Research* (Vol. 16).

Craig Venter, J., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., … Zhu, X. (2001). The sequence of the human genome. *Science*, *291*(5507), 1304–1351. https://doi.org/10.1126/science.1058040

Creyghton, M. P., Cheng, A. W., Welstead, G. G., Kooistra, T., Carey, B. W., Steine, E. J., … Jaenisch, R. (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(50), 21931–21936. https://doi.org/10.1073/pnas.1016071107

Dao, L. T. M., Galindo-Albarrán, A. O., Castro-Mondragon, J. A., Andrieu-Soler, C., Medina-Rivera, A., Souaid, C., … Spicuglia, S. (2017). Genome-wide characterization of mammalian promoters with distal enhancer functions. *Nature Genetics*, *49*(7), 1073–1081. https://doi.org/10.1038/ng.3884

Dao, L. T. M., & Spicuglia, S. (2018). Transcriptional regulation by promoters with enhancer function. *Transcription*, *9*(5), 307–314. https://doi.org/10.1080/21541264.2018.1486150

Drakos, G. (2019). Decision Tree Regressor explained in depth. Retrieved April 8, 2021, from https://gdcoder.com/decision-tree-regressor-explained-in-depth/

DTN. (2019). DataTechNotes: Regression Model Accuracy (MAE, MSE, RMSE, R-squared) Check in R. Retrieved April 6, 2021, from https://www.datatechnotes.com/2019/02/regression-model-accuracy-mae-mse-rmse.html

Eckner, R., Ewen, M. E., Newsome, D., Gerdes, M., DeCaprio, J. A., Lawrence, J. B., & Livingston, D. M. (1994a). Molecular cloning and functional analysis of the adenovirus E1A- associated 300-kD protein (p300) reveals a protein with properties of a transcriptional adaptor. *Genes and Development*, *8*(8), 869–884. https://doi.org/10.1101/gad.8.8.869

Eckner, R., Ewen, M. E., Newsome, D., Gerdes, M., DeCaprio, J. A., Lawrence, J. B., & Livingston, D. M. (1994b). Molecular cloning and functional analysis of the adenovirus E1A- associated 300-kD protein (p300) reveals a protein with properties of a transcriptional adaptor. *Genes and Development*, *8*(8), 869–884. https://doi.org/10.1101/gad.8.8.869

Ernst, J., Kheradpour, P., Mikkelsen, T. S., Shoresh, N., Ward, L. D., Epstein, C. B., … Bernstein, B. E. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, *473*(7345), 43–49. https://doi.org/10.1038/nature09906

Genome-wide association study - Wikipedia. (n.d.). Retrieved November 19, 2020, from https://en.wikipedia.org/wiki/Genome-wide_association_study

Google Developers. (2020). Classification: Precision and Recall | Machine Learning Crash

Course. Retrieved April 6, 2021, from Google website: https://developers.google.com/machine-learning/crash-course/classification/precision-and-recall

Haines, J. L., Hauser, M. A., Schmidt, S., Scott, W. K., Olson, L. M., Gallins, P., … Pericak-Vance, M. A. (2005). Complement factor H variant increases the risk of age-related macular degeneration. *Science*, *308*(5720), 419–421. https://doi.org/10.1126/science.1110359

Heintzman, N. D., & Ren, B. (2009, December 1). Finding distal regulatory elements in the human genome. *Current Opinion in Genetics and Development*, Vol. 19, pp. 541–549. https://doi.org/10.1016/j.gde.2009.09.006

HiCUP Overview — HiCUP 0.8 documentation. (n.d.). Retrieved April 6, 2021, from https://www.bioinformatics.babraham.ac.uk/projects/hicup/read_the_docs/html/index.html

Hill, W. G., & Robertson, A. (1968). The effects of inbreeding at loci with heterozygote advantage. *Genetics*, *60*(3), 615–628. https://doi.org/10.1093/genetics/60.3.615

Hnisz, D., Abraham, B. J., Lee, T. I., Lau, A., Saint-André, V., Sigova, A. A., … Young, R. A. (2013). XSuper-enhancers in the control of cell identity and disease. *Cell*, *155*(4), 934. https://doi.org/10.1016/j.cell.2013.09.053

Javatpoint. (2018). Support Vector Machine (SVM) Algorithm - Javatpoint. Retrieved April 8, 2021, from JavaTpoint website: https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm

Jostins, L., & Barrett, J. C. (2011). Genetic risk prediction in complex disease. *Human Molecular Genetics*, *20*(R2), 182–188. https://doi.org/10.1093/hmg/ddr378

Klein, R. J., Xu, X., Mukherjee, S., Willis, J., & Hayes, J. (2010, August 6). Successes of Genome-wide association studies. *Cell*, Vol. 142, pp. 350–351. https://doi.org/10.1016/j.cell.2010.07.026

Li, G., Ruan, X., Auerbach, R. K., Sandhu, K. S., Zheng, M., Wang, P., … Ruan, Y. (2012). Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell*, *148*(1–2), 84–98. https://doi.org/10.1016/j.cell.2011.12.014

Libioulle, C., Louis, E., Hansoul, S., Sandor, C., Farnir, F., Franchimont, D., … Georges, M. (2007). Novel Crohn disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of PTGER4. *PLoS Genetics*, *3*(4), 0538–0543. https://doi.org/10.1371/journal.pgen.0030058

Lieberman-Aiden, E., Van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., … Dekker, J. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, *326*(5950), 289–293. https://doi.org/10.1126/science.1181369

Liu, X., Xu, W., Leng, F., Hao, C., Kolora, S. R. R., & Li, W. (2020). Prioritizing long range interactions in noncoding regions using GWAS and deletions perturbed TADs.

Computational and Structural Biotechnology Journal, *18*, 2945–2952. https://doi.org/10.1016/j.csbj.2020.10.014

Lopez, F. (2021). SMOTE: Synthetic Data Augmentation for Tabular Data | by Fernando López | Mar, 2021 | Towards Data Science. Retrieved April 7, 2021, from https://towardsdatascience.com/smote-synthetic-data-augmentation-for-tabular-data-1ce28090debc

McClellan, J., & King, M. C. (2010, April 16). Genetic heterogeneity in human disease. *Cell*, Vol. 141, pp. 210–217. https://doi.org/10.1016/j.cell.2010.03.032

Mifsud, B., Martincorena, I., Darbo, E., Sugar, R., Schoenfelder, S., Fraser, P., & Luscombe, N. M. (2017). GOTHiC, a probabilistic model to resolve complex biases and to identify real interactions in Hi-C data. *PLOS ONE*, *12*(4), e0174744. https://doi.org/10.1371/journal.pone.0174744

Minsky, M., & Papert, S. A. (2019). Perceptrons. In *Perceptrons*. https://doi.org/10.7551/mitpress/11301.001.0001

Naumova, N., Imakaev, M., Fudenberg, G., Zhan, Y., Lajoie, B. R., Mirny, L. A., & Dekker, J. (2013). Organization of the mitotic chromosome. *Science*, *342*(6161), 948–953. https://doi.org/10.1126/science.1236083

Ngan, C. Y., Wong, C. H., Tjong, H., Wang, W., Goldfeder, R. L., Choi, C., … Wei, C. L. (2020). Chromatin interaction analyses elucidate the roles of PRC2-bound silencers in mouse development. *Nature Genetics*, *52*(3), 264–272. https://doi.org/10.1038/s41588-020-0581-x

Nguyen, T. A., Jones, R. D., Snavely, A. R., Pfenning, A. R., Kirchner, R., Hemberg, M., & Gray, J. M. (2016). High-throughput functional comparison of promoter and enhancer activities. *Genome Research*, *26*(8), 1023–1033. https://doi.org/10.1101/gr.204834.116

Oldfield, A. J., Henriques, T., Kumar, D., Burkholder, A. B., Cinghu, S., Paulet, D., … Jothi, R. (2019). NF-Y controls fidelity of transcription initiation at gene promoters through maintenance of the nucleosome-depleted region. *Nature Communications*, *10*(1). https://doi.org/10.1038/s41467-019-10905-7

Pajares, M., Jiménez-Moreno, N., García-Yagüe, Á. J., Escoll, M., de Ceballos, M. L., Van Leuven, F., … Cuadrado, A. (2016). Transcription factor NFE2L2/NRF2 is a regulator of macroautophagy genes. *Autophagy*, *12*(10), 1902–1916. https://doi.org/10.1080/15548627.2016.1208889

Pulst, S. M. (1999). Genetic linkage analysis. *Archives of Neurology*, *56*(6), 667–672. https://doi.org/10.1001/archneur.56.6.667

Reuter, J. A., Spacek, D. V., & Snyder, M. P. (2015, May 21). High-Throughput Sequencing Technologies. *Molecular Cell*, Vol. 58, pp. 586–597. https://doi.org/10.1016/j.molcel.2015.05.004

Risch, N., & Merikangas, K. (1996, September 13). The future of genetic studies of complex

human diseases. *Science*, Vol. 273, pp. 1516–1517. https://doi.org/10.1126/science.273.5281.1516

Ritchie, M. D., & Van Steen, K. (2018). The search for gene-gene interactions in genome-wide association studies: challenges in abundance of methods, practical considerations, and biological interpretation. *Annals of Translational Medicine*, *6*(8), 157–157. https://doi.org/10.21037/atm.2018.04.05

Santillán, M. (2008). On the Use of the Hill Functions in Mathematical Models of Gene Regulatory Networks. *Math. Model. Nat. Phenom*, *3*(2). https://doi.org/10.1051/mmnp:2008056

Sarang, N. (2018). Understanding AUC - ROC Curve | by Sarang Narkhede | Towards Data Science. Retrieved April 6, 2021, from https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

Schaffner, W. (2015, April 1). Enhancers, enhancers - From their discovery to today's universe of transcription enhancers. *Biological Chemistry*, Vol. 396, pp. 311–327. https://doi.org/10.1515/hsz-2014-0303

Schierding, W., Cutfield, W. S., & O'Sullivan, J. M. (2014). The missing story behind Genome Wide Association Studies: Single nucleotide polymorphisms in gene deserts have a story to tell. *Frontiers in Genetics*, Vol. 5. https://doi.org/10.3389/fgene.2014.00039

Simske, S. (2019). Meta-analytic design patterns. In *Meta-Analytics* (pp. 147–185). https://doi.org/10.1016/b978-0-12-814623-1.00004-6

Singh Sandhu, K., Li, G., Mei Poh, H., Ling Kelly Quek, Y., Yen Sia, Y., Qin Peh, S., … Ruan, Y. (2012). Large-Scale Functional Organization of Long-Range Chromatin Interaction Networks. *CellReports*, *2*, 1207–1219. https://doi.org/10.1016/j.celrep.2012.09.022

sklearn.neural_network.MLPRegressor — scikit-learn 0.24.1 documentation. (n.d.). Retrieved April 6, 2021, from https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPRegressor.html#sklearn.neural_network.MLPRegressor.score

Vanhille, L., Griffon, A., Maqbool, M. A., Zacarias-Cabeza, J., Dao, L. T. M., Fernandez, N., … Spicuglia, S. (2015). High-throughput and quantitative assessment of enhancer activity in mammals by CapStarr-seq. *Nature Communications*, *6*(1), 1–10. https://doi.org/10.1038/ncomms7905

Visel, A., Rubin, E. M., & Pennacchio, L. A. (2009, September 10). Genomic views of distant-acting enhancers. *Nature*, Vol. 461, pp. 199–205. https://doi.org/10.1038/nature08451

Visscher, P. M., Brown, M. A., McCarthy, M. I., & Yang, J. (2012, January 13). Five years of GWAS discovery. *American Journal of Human Genetics*, Vol. 90, pp. 7–24. https://doi.org/10.1016/j.ajhg.2011.11.029

Wei, C. L., Nicolis, S. K., Zhu, Y., & Pagin, M. (2019, August 1). Sox2-Dependent 3D

Chromatin Interactomes in Transcription, Neural Stem Cell Proliferation and Neurodevelopmental Diseases. *Journal of Experimental Neuroscience*, Vol. 13. https://doi.org/10.1177/1179069519868224

Wingett, S., Ewels, P., Furlan-Magaril, M., Nagano, T., Schoenfelder, S., Fraser, P., & Andrews, S. (2015). HiCUP: Pipeline for mapping and processing Hi-C data. *F1000Research*, *4*, 1310. https://doi.org/10.12688/f1000research.7334.1

Wingett, S. W., Ewels, P., Furlan-Magaril, M., Nagano, T., Schoenfelder, S., Fraser, P., & Andrews, S. (2015). HiCUP: pipeline for mapping and processing Hi-C data. *F1000Research*, *4*, 1310. https://doi.org/10.12688/f1000research.7334.1

Yamada, K., Printz, R. L., Osawa, H., & Granner, D. K. (1999). Human ZHX1: Cloning, chromosomal location, and interaction with transcription factor NF-Y. *Biochemical and Biophysical Research Communications*, *261*(3), 614–621. https://doi.org/10.1006/bbrc.1999.1087

Yang, C.-H., Lin, Y.-D., Wu, S.-J., Chuang, L.-Y., & Chang, H.-W. (2015). High Order Gene-Gene Interactions in Eight Single Nucleotide Polymorphisms of Renin-Angiotensin System Genes for Hypertension Association Study. *BioMed Research International*, *2015*, 1–11. https://doi.org/10.1155/2015/454091

Appendix A

Attempts at improving the performance of regression

Apart from feature extraction using classification, various other methods to improve the regression model's performance were tested. Some of the approaches implemented to improve the regression models' performance included removing outliers from the Enhancer Potential data, utilizing different scaling algorithms, and scaling the Enhancer Potential data. The results from all such approaches are outlined in the following table.

Table 8: Comparison of different methods to improve the performance of the Regression models

Model	**Treatment**	**Overall Score**	**Training statistics**			**Testing statistics**		
			RMSE	**MSE**	**MAE**	**RMSE**	**MSE**	**MAE**
MLP Regressor	Removing outliers	0.13	0.38	0.14	0.27	0.39	0.15	0.28
	Using Min-Max scaling	0.15	0.39	0.15	0.28	0.39	0.15	0.28
	Scaling the Y matrix	-4.77	0.08	0.006	0.06	0.4	0.16	0.29
SVR	Removing outliers	0.15	0.37	0.14	0.26	0.39	0.15	0.27
	Using Min-Max scaling	0.15	0.38	0.14	0.27	0.39	0.15	0.27
	Scaling the Y matrix	-4.19	0.1	0.01	0.09	0.96	0.91	0.86

Decision Tree Regressor	Removing outliers	0.14	0.39	0.15	0.28	0.39	0.15	0.28
	Using Min-Max scaling	0.14	0.39	0.15	0.28	0.39	0.15	0.28
	Scaling the Y matrix	-4.80	0.08	0.007	0.06	1.0	1.01	0.92

In Table 8, removing outliers refers to removing enhancer potential values significantly larger (>5). The min-max scaling was done on the X matrix (histone modifications), while in the final treatment (Scaling the Y matrix), the enhancer potential values were scaled using a min-max scaler.

A slight increase in performance is seen for the first two treatments (removing outliers and scaling with a min-max scaler); however, the performance severely deteriorates when the enhancer potential values are scaled using a min-max scaler.

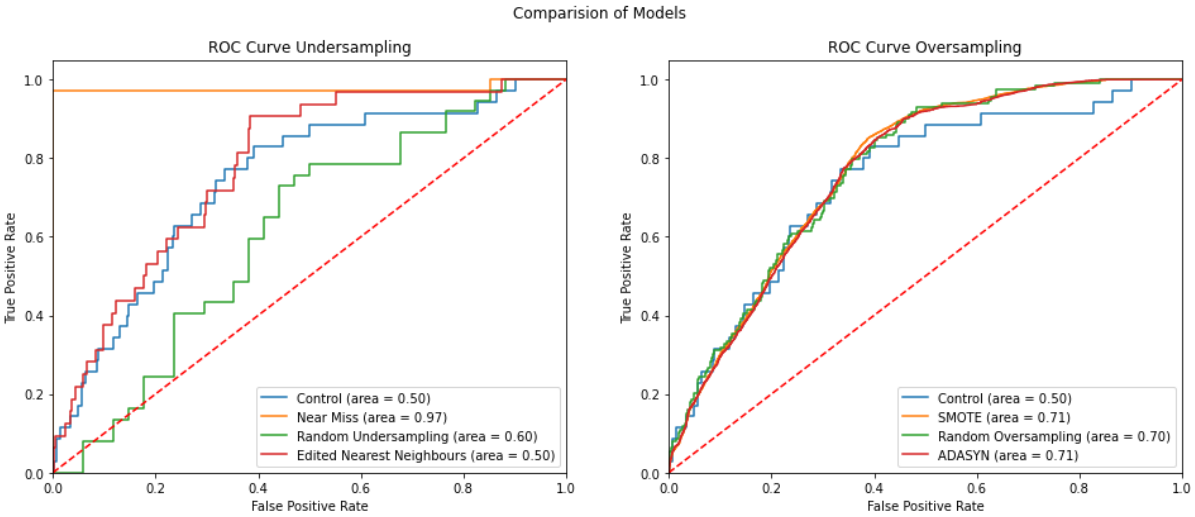
Appendix B

Choosing a sampling method to balance the dataset

The method of dichotomizing the enhancer potential values created a high imbalance in the dataset (promoters ~ 96% & epromoters ~ 4%). Many sampling methods were considered to balance the dataset:

- Random undersampling
- Near miss undersampling
- Edited nearest neighbours undersampling
- Random oversampling
- SMOTE
- ADASYN

Classification analysis was carried out using a logistic regression model on the HeLa histone data in conjunction with all the mentioned sampling methods to test out which method is the best. A control case was also maintained without any over/under-sampling. The ROC (Receiver Operator Characteristic) curves for the different models are presented below.



(a) Undersampling methods (b) Oversampling methods

Figure 26: ROC-AUC Plots of the different sampling methods

The ROC-AUC values tell us that all the methods except the Edited Nearest Neighbour sampling perform better than the control (no over/under sampling). The best performing model followed the Near Miss undersampling method followed by the SMOTE and ADASYN over samplers. Although the Near Miss method produced a significantly higher prediction rate, it does so by dropping almost 90% of the data points. As the logistic regression model is one of the simpler models (it has a smaller number of adjustable parameters), the scarcity of data points does not affect the results significantly. However, while training more complex models, the lesser number of data points starts affecting the performance. Thus, I decided not to use such an extreme undersampling method.

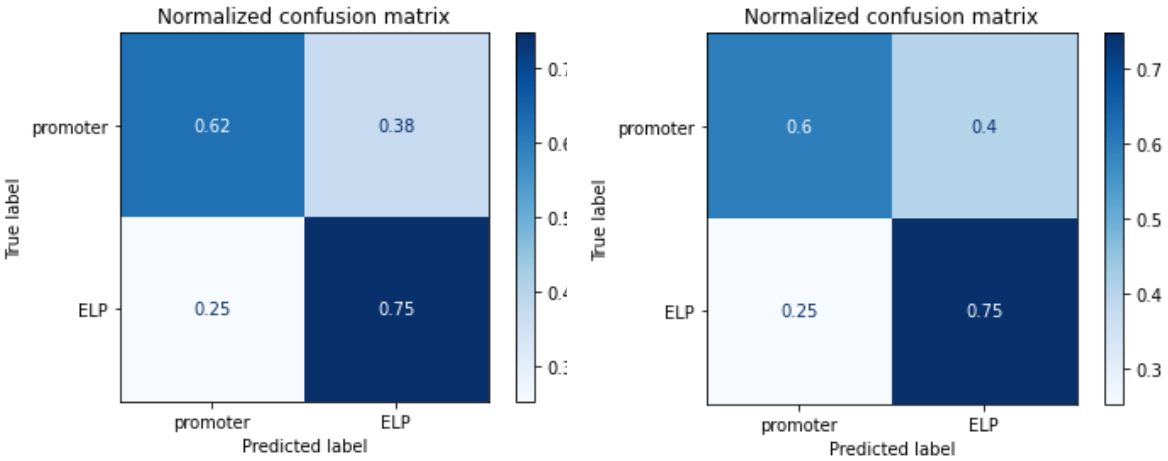
Of the above oversampling methods, SMOTE and ADASYN algorithms performed relatively better than the random oversampler and the control. Recent literature has shown that in the case of highly imbalanced datasets, SMOTE performs slightly better than ADASYN (Brandt & Lanzén, 2021). Thus, SMOTE was used as an oversampling method to balance the datasets.

Appendix C

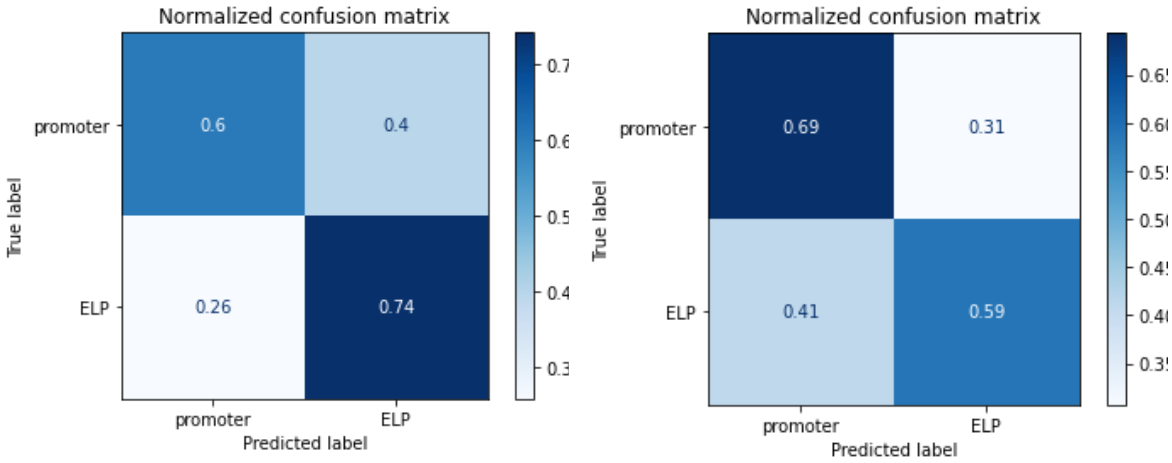
Using H3K4me1/H3K4me3 ratio to improve classification

Broadly, a high ratio of H3K4me1 to H3K4me3 is used to distinguish enhancers from promoters(Calo & Wysocka, 2013). Thus, it was suspected that the ratio of H3K4me1 to H3K4me3 might also be an essential feature to help determine whether a promoter acts as an epromoter. The HeLa histone dataset columns were modified to create a column containing the ratio of H3K4me1 to H3K4me3, and the individual columns for H3K4me1 and H3K4me3 were removed from the dataset. Taking ratio produced some NaN values in the dataset due to division by 0. Removing the NaN values, the dataset contained 20336 valid samples (promoters and epromoters).

The dataset was again processed by scaling the data, splitting it into training and validation datasets, and oversampling using the SMOTE algorithm. The following results were obtained after training the models on the different classification models.



(a) Logistic Regression (b) Decision Tree Classifier



(c) MLP classifier (d) SVM classifier

Figure 27: Confusion matrix for the H3K4me1/H3K4me3 ratio dataset

When compared to the original HeLa histone dataset confusion matrices, the confusion matrices indicate a general deterioration in the performance in three out of four models. The SVM classifier is the only model where the detection values of epromoters slightly increase.

Table 9: Precision and Recall values for the Histone ratio dataset

Dataset	**Model**	**Precision**		**Recall**	
		Promoter class	**Epromoter class**	**Promoter class**	**Epromoter class**
HeLa Histone Ratio	Logistic Regression	0.99	0.06	0.62	0.75
	Decision Tree Classifier	0.99	0.06	0.6	0.75
	MLP	0.99	0.06	0.6	0.74
	SVM	0.98	0.06	0.69	0.59

The precision-recall values also show a general deterioration of model performance. The precision values are almost similar to the original unmodified dataset, while changes are only prominent in the recall values.

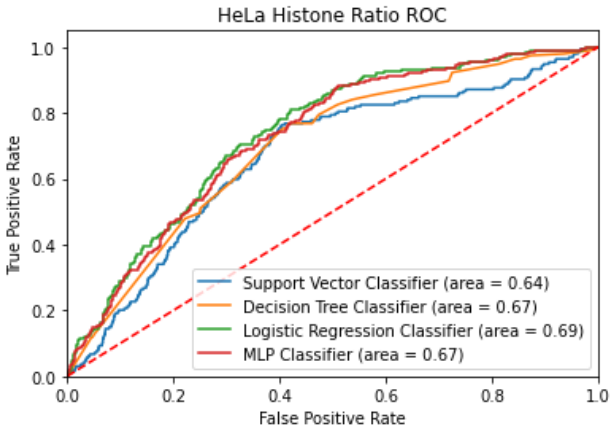


Figure 28: ROC-AUC plot for the HeLa Histone ratio dataset

The ROC plot indicates that the models perform similarly to each other, with AUC values ranging within a range of 0.05. Comparing to the ROC plot in Figure 9, the AUC values and ROC curves indicate a slight deterioration in model performance.

The deteriorating model performance is contrary to the initial idea that the ratio of H3K4me1 to H3K4me3 should increase the models' prediction accuracy. This may be attributed to differences between the supposed characteristics of enhancers and enhancer-like promoters (epromoters). It is speculated that the high ratio of H3K4me1:me3 is the result of a more significant amount of H3K4me3 present in the promoter regions. H3K4me3 is associated with the presence of initiating form of Pol II. As promoters are transcribed while enhancers are not, enhancer-like promoters may have slightly higher H3K4me3 values, which lower the H3K4me1:me3 ratio and differentiate them from enhancers (Calo & Wysocka, 2013).