Distribution Agreement

Kihoon Alan Kang 4/8/2025

Characterizing Chromatin Changes Upon Inhibition of Chromatin Remodeling Complexes

by

Kihoon Alan Kang

David Gorkin, Ph.D.
Adviser

Biology

David Gorkin, Ph.D.

Adviser

Yana Bromberg, Ph.D.

Committee Member

Arri Eisen, Ph.D.

Committee Member

2025

Characterizing Chromatin Changes Upon Inhibition of Chromatin Remodeling Complexes

By

Kihoon Alan Kang

David Gorkin, Ph.D.
Adviser

An abstract of
a thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Science with Honors

Biology

2025

Abstract

Characterizing Chromatin Changes Upon Inhibition of Chromatin Remodeling Complexes

By Kihoon Alan Kang

BRG1/BRM Associated Factors (BAF) complexes are ATP-dependent chromatin remodelers which control chromatin accessibility genome-wide. Mutations in BAF subunits can cause neurodevelopmental disease and cancer. While BAF complexes are known to regulate chromatin accessibility, the specific mechanisms by which they target genomic regions, and the downstream effects of their inhibition remain incompletely understood. In particular, it is unclear which transcription factors or chromatin features determine a region's sensitivity to BAF activity. Addressing this gap is critical for interpreting how mutations in BAF subunits contribute to disease. Motivated by this, my thesis research aimed to systematically characterize the chromatin-level consequences of BAF inhibition and identify the molecular features predictive of such changes. To do this, I inhibited BAF activity using a small molecule targeting the ATPase subunit of the complex. Using ATAC-seq to profile chromatin accessibility, we observed widespread loss of accessible chromatin regions upon BAF inhibition. Machine learning models, including a random forest classifier and ridge regression, were then trained to predict accessible chromatin sensitive or insensitive to BAF inhibition based on transcription factor binding and histone modification profiles. A random forest classifier achieved accuracies above 78% with high AUROC values, while feature importance analyses from linear regression models highlights distinct roles for promoter-associated factors, CTCF/cohesin subunits and lineage-specific transcription factors (e.g., RUNX3, BATF, JUNB, SPI1) in understanding chromatin response to BAF inhibition. Analysis of known protein-protein interactions in StringDB indicates that transcription factors which bind to BAF subunits are predictive of chromatin accessibility loss upon BAF inhibition, suggesting that these TFs may function to recruit BAF complexes to chromatin via protein-protein interactions. Overall, this work establishes an analytical framework for fundamentally understanding the effects of BAF activity on chromatin

Characterizing Chromatin Changes Upon Inhibition of Chromatin Remodeling Complexes

By

Kihoon Alan Kang

David Gorkin, Ph.D.
Adviser

A thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Science with Honors

Biology

2025

Table of Contents

1. Abstract

BRG1/BRM Associated Factors (BAF) complexes are ATP-dependent chromatin remodelers which control chromatin accessibility genome-wide. Mutations in BAF subunits can cause neurodevelopmental disease and cancer. While BAF complexes are known to regulate chromatin accessibility, the specific mechanisms by which they target genomic regions, and the downstream effects of their inhibition remain incompletely understood. In particular, it is unclear which transcription factors or chromatin features determine a region's sensitivity to BAF activity. Addressing this gap is critical for interpreting how mutations in BAF subunits contribute to disease. Motivated by this, my thesis research aimed to systematically characterize the chromatin-level consequences of BAF inhibition and identify the molecular features predictive of such changes. To do this, I inhibited BAF activity using a small molecule targeting the ATPase subunit of the complex. Using ATAC-seq to profile chromatin accessibility, we observed widespread loss of accessible chromatin regions upon BAF inhibition. Machine learning models, including a random forest classifier and ridge regression, were then trained to predict accessible chromatin sensitive or insensitive to BAF inhibition based on transcription factor binding and histone modification profiles. A random forest classifier achieved accuracies above 78% with high AUROC values, while feature importance analyses from linear regression models highlights distinct roles for promoter-associated factors, CTCF/cohesin subunits and lineage-specific transcription factors (e.g., RUNX3, BATF, JUNB, SPI1) in understanding chromatin response to BAF inhibition. Analysis of known protein-protein interactions in StringDB indicates that transcription factors which bind to BAF subunits are predictive of chromatin accessibility loss upon BAF inhibition, suggesting that these TFs may function to recruit BAF complexes to chromatin via protein-protein

interactions. Overall, this work establishes an analytical framework for fundamentally understanding the effects of BAF activity on chromatin.

2. Introduction

2.1. The packaging of Eukaryotic DNA into chromatin

In eukaryotic cells, DNA is tightly packaged into a dynamic and hierarchical structure known as chromatin. This packaging is essential for organizing the genome within the nucleus and plays a critical role in regulating access to genetic information. The fundamental unit of chromatin is the nucleosome, which consists of ~147 base pairs of DNA wrapped around a histone octamer. Nucleosomes are further organized into higher-order structures, creating a physical barrier to the transcriptional machinery and other DNA-binding proteins (Li et al., 2007). Chromatin structure is not static; rather, it is continuously remodeled in response to cellular signals, developmental cues, and environmental stimuli. These structural changes are mediated by chromatin remodelers, histone modifiers, and non-coding RNAs, which collectively modulate DNA accessibility and genome function (Kouzarides, 2007).

2.2 Chromatin accessibility and transcriptional regulation.

The degree to which DNA within chromatin is exposed and available for interaction with regulatory proteins is referred to as chromatin accessibility. Regions of DNA unbound by nucleosomes, known as **accessible chromatin**, are typically associated with active gene expression, as they allow transcription factors and the transcriptional machinery to bind to DNA. Conversely, regions of DNA bound by nucleosomes, or **inaccessible chromatin**, are often linked to gene silencing due to the obstruction of DNA-protein interaction (Tsompana and Buck, 2014).

Transcription factors (**TFs**) are proteins that bind to specific DNA sequences to regulate the transcription of genes. They play a crucial role in turning genes on or off by facilitating or hindering the recruitment and stabilization of RNA polymerase to gene promoters. TFs often bind to cis-regulatory elements **(cREs)**, which are regions of non-coding DNA which control the regulation of genes. These elements include promoters, enhancers, silencers, and insulators, and they function as binding sites for TFs to modulate gene expression.

The interplay between chromatin accessibility, cREs, and TFs is fundamental to transcriptional regulation. Accessible chromatin regions often correspond to active cREs where TFs can bind and initiate or enhance transcription. Conversely, in regions where chromatin is less accessible, TF binding is hindered, leading to reduced gene expression. This dynamic regulation allows cells to fine-tune gene expression programs in response to developmental cues and environmental stimuli, highlighting the importance of chromatin structure in controlling cellular function and identity.

2.3. The BRG1/BRM Associated Factors (BAF) Complexes regulate chromatin accessibility at cis-regulatory sequences.

Proper regulation of chromatin accessibility is critical for both development and disease, as the regions of the genome that remain open directly determine which genes are expressed. When genes are inappropriately activated or repressed due to misregulated chromatin accessibility, it can lead to a variety of diseases (Kouzarides, 2007). The BRG1/BRM Associated Factors (BAF) family of complexes are ATP-dependent chromatin remodelers that establish and maintain accessible chromatin. BAF can achieve this function by displacing histone octamers or by shifting nucleosome placement, which in turn influences which parts of the genome remain accessible (Alfert et al., 2019).

BAF complexes are highly modular and can consist of up to 15 different subunits. The main catalytic component of the complex is the ATPase subunit, which comes in one of two mutually exclusive forms: either BRG1 (SMARCA4) or BRM (SMARCA2). Without the activity of this ATPase, the BAF complex cannot remodel chromatin (Mashtalir et al., 2018). In addition to the ATPase, the complex also contains several other proteins that help recognize histones and specific histone modifications. For example, some subunits contain bromodomains or double PHD finger (DPF) domains which allows BAF to bind directly to modified histones. Together, these subunits work to regulate chromatin accessibility, ensuring that the correct regions of the genome are opened during development and homeostasis. Consequently, loss-of-function mutations that inactivate BAF-mediated chromatin remodeling are often associated with diseases like cancer and various neurodevelopmental disorders (Hodges et al., 2016, Mathur and Roberts, 2018).

2.4. Transcription Factors as Potential Recruiters of BAF

Although BAF contains subunits which help it bind to chromatin (such as a bromodomain), it lacks subunits for sequence specificity, meaning that it has no way of knowing from DNA-sequence which areas of the genome to bind to. This raises a central question in chromatin biology: how do BAF complexes get recruited to specific genomic loci?

A leading hypothesis is that sequence-specific transcription factors (TFs) act as recruiters for BAF complexes. Because TFs bind to defined DNA motifs, they can guide BAF complexes to discrete regulatory regions, thus resulting in specificity (Ho et al., 2019). For example, studies have demonstrated that TFs such as the AP1 TF, JUNB interact with components of the BAF complex during key developmental processes (Vierbuchen et al., 2017). Moreover, posttranslational modifications and additional cofactors are thought to further fine-tune these interactions, ensuring that chromatin remodeling is both context-dependent and precisely regulated. Despite these advances, the full spectrum of transcription factors involved in recruiting BAF complexes, remains an active area of research, and a deeper understanding of this process is crucial.

2.5. GM12878 as a system to study chromatin dynamics upon BAF inhibition

GM12878 is a lymphoblastoid cell line derived from B lymphocytes that have been transformed by the Epstein–Barr virus. It’s widely used to study gene regulation and chromatin biology. As an ENCODE Tier 1 cell line, GM12878 has been studied in depth, and there are plenty of public datasets available on chromatin accessibility, transcription factor binding, and histone modifications through the ENCODE database (ENCODE Project Consortium, 2012; Thurman et al., 2012).

Since GM12878 cells grow easily in suspension, they are a practical model for conducting techniques like Hi-C, ATAC-seq, and ChIP-seq to track changes in chromatin structure. All of this makes GM12878 a useful tool for understanding basic chromatin dynamics and the role of chromatin remodelers like the BAF in regulating chromatin.

2.6. Accessible chromatin regions display heterogeneous responses to BAF inhibition

Previously, researchers have used a small molecule inhibitor of the BAF ATPase subunit, BRM014, on different cell lines to characterize chromatin accessibility changes upon loss of BAF function. These studies concluded that upon BAF loss-of-function, there was global chromatin accessibility loss and a drastic change in transcription. However, these previous studies also found that not all accessible chromatin reacted the same to BAF loss-of-function. For example, cis-regulatory elements (cREs) like enhancers were found to be more sensitive to BAF inhibition compared to other areas of open chromatin (Schick et al., 2021, Iurlaro et al., 2021). cREs are non-coding regulatory regions that serve as binding sites for TFs, which further recruit additional transcriptional machinery to control gene expression.

Although it is known that certain broad classes of cREs, such as those bound by CTCF or found in promoter-associated regions, remain accessible when BAF is inhibited in some cell types (Bao et al., 2015), there is still a lack of detailed methods to connect the combined effects of transcription factor binding and histone modifications with changes in chromatin accessibility upon BAF inhibition. In this study, I propose using a machine-learning feature-analysis based approach to (1) identify regions of open chromatin that are either sensitive or insensitive to BAF activity and (2) determine which features - such as specific transcription factor binding profiles or histone modification patterns - explain these differences. By developing this analytical framework, I also hope to reveal general patterns of BAF activity and better understand its role in development and disease.

3. Methods

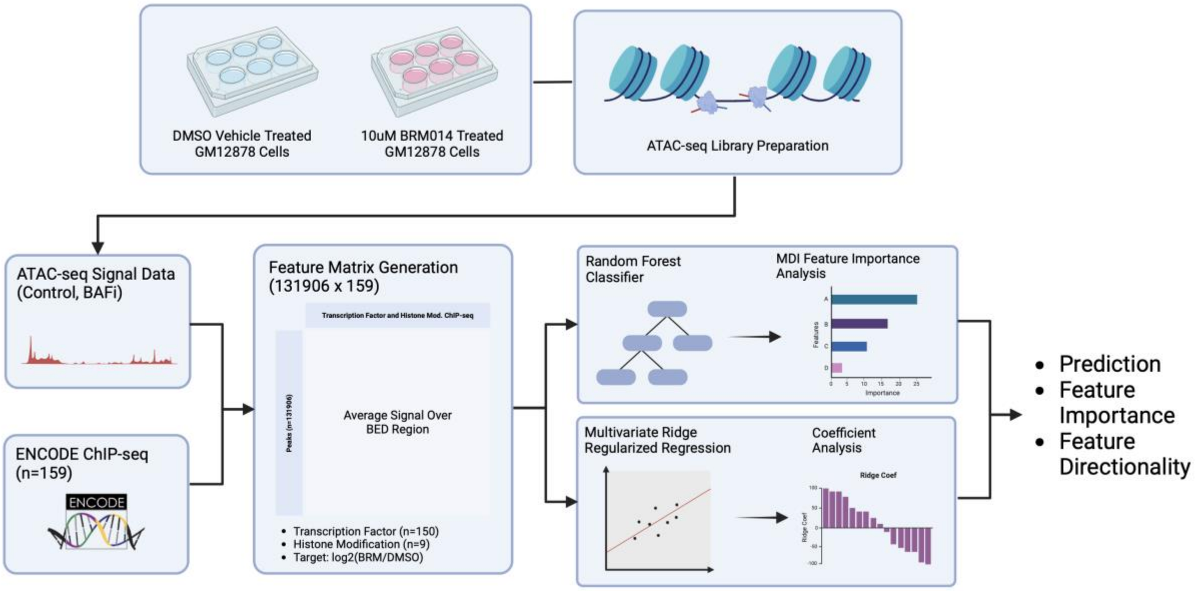
DMSO Vehicle Treated GM12878 Cells
10uM BRM014 Treated GM12878 Cells
ATAC-seq Library Preparation
ATAC-seq Signal Data (Control, BAFi)
ENCODE ChIP-seq (n=159)
ENCODE
Feature Matrix Generation (131906 x 159)
Transcription Factor and Histone Mod. ChIP-seq
Peaks (n=131906)
Average Signal Over BED Region
Transcription Factor (n=150)
Histone Modification (n=9)
Target: log2(BRM/DMSO)
Random Forest Classifier
MDI Feature Importance Analysis
A
B
C
D
Features
0
5
10
15
20
25
Importance
Multivariate Ridge Regularized Regression
Coefficient Analysis
Ridge Coef
100
50
0
-50
-100
Ridge Coef
Prediction
Feature Importance
Feature Directionality

Figure 1. Overview of the experimental and computational workflow used to assess chromatin accessibility changes upon BAF inhibition in GM12878 cells. This diagram details key steps including cell treatment with the BRG1/BRM inhibitor BRM014, ATAC-seq library preparation, construction of a binary feature matrix from transcription factor binding sites and histone modification data, and the application of machine learning models (random forest and ridge regression) to classify BAF sensitivity and conduct feature analysis.

3.1. Culturing of GM12878 cells

GM12878 is a suspension lymphoblastoid cell line which can be obtained from Coriell. After cells were brought out from long-term liquid nitrogen storage at passage #7, they were passaged 2 times in in 20mL of culture media which comprised of 85% RPMI media supplemented with 15% fetal bovine serum and 1X penicillin/streptomycin in upright T25 flasks following Coriell's culture guidelines. The cells were incubated in 37°C and at 5% carbon dioxide. All cell culture work was done in a sterile environment inside a biosafety cabinet.

3.2. BAF-ATPase inhibition of GM12878 cells

To inhibit BAF-ATPase activity, I used a small-molecule dual-inhibitor of BRG1/BRM known as BRM014 (obtained from MedChemExpress Cat. No.: HY-119374) which is dissolved and aliquoted in DMSO at a concentration of 10mM. At passage # 9, GM12878 cells were transferred into a 6-well plate (obtained from Corning Cat. No.: CLS353046) at a concentration of 1 million cells per 3mL of culture media per well. Then two wells were treated with 30uL of DMSO, and 2 wells were treated with 30uL of BRM014 diluted in DMSO to reach an effective concentration of 10nM. This gives us 2 replicates BRM014 treated GM12878 cells each paired with a vehicle treated group of cells. The cells were given its treatment group and incubated in 37°C at 5% carbon dioxide for 24 hours.

3.3. Assay for Transposase Accessible Chromatin with Sequencing (ATAC-seq)

After GM12878 cells were incubated under appropriate treatment conditions, 250K cells were collected, spun down, and flash frozen in 1mL of freezing media (90% RPMI and 10% DMSO) to be stored in –80°C conditions. The ATAC-seq protocol was adapted from Buenrostro et al., 2015.

3.4. Library preparation and sequencing

For barcoding of samples, I used primers from IDT’s Nextera Index XT Kit v2 which provides dual i5 and i7 IDs for each sample. Barcoded and amplified libraries were sent off to Novogene for sequencing on Illumina’s NovaSeq6000 machine. Information generated from sequencing was downloaded onto the Lab server.

3.5. ATAC-seq data processing

The raw ATAC-seq sequencing data obtained from Novogene was processed using the standardized ENCODE ATAC-seq pipeline. This pipeline consists of adapter trimming, alignment to the human reference genome (GRCh38), PCR duplicate removal, and peak calling. Adapter trimming was performed with trim galore to remove Illumina-specific sequencing adapters. Alignment was executed using BWA, generating aligned BAM files (Li and Durbin, 2009).

Duplicate reads, indicative of PCR amplification bias, were identified and removed with samtools (Li et al., 2009). Peak calling was then performed with MACS3 to generate narrowPeak files, identifying regions of accessible chromatin (Zhang et al., 2008). Quality control metrics, including those that result from fastQC and FRiP scores (Fraction of Reads in Peaks), were calculated to ensure high data quality and reproducibility across replicates.

3.6. ENCODE BED file parsing and download

Relevant publicly available BED files representing transcription factor binding sites and histone modification peaks for GM12878 were identified and retrieved using the ENCODE REST API. Files were filtered to select datasets with the highest FRiP values, ensuring the use of data with the highest signal-to-noise ratio. These datasets included ChIP-seq peak files for transcription factors (e.g., CTCF, JUNB, SPI1) and histone modifications (e.g., H3K27ac, H3K4me3, H3K27me3) critical for understanding chromatin accessibility and regulatory element activity. In total, there were 150 DNA-protein interaction CHiP-Seq BED files and 9 Histone Modification BED files available for us to use.

3.7. Implementation of machine learning algorithms

3.7.1. Feature Matrix Generation

For this analysis, replicated GM12878 peaks from both BRM-treated and DMSO-treated cells were obtained to ensure reproducibility in peak detection. These peaks formed the basis for integrating additional genomic features. A binary feature matrix was constructed by assessing the overlap between transcription factor (TF) and histone modification BED files and ATAC-seq peaks, allowing us to encode the presence or absence of key regulatory elements.

The target variable was defined as the $\log_2$ fold change in signal enrichment over the BED regions, calculated using BigWigAverageOverBed from the UCSC Genome Browser Utilities

(Perez et al., 2025). A two-fold decrease in signal enrichment ($\log_2$ fold change = -1) was used as the cutoff to categorize peaks as BAF-sensitive, while peaks above this threshold were considered BAF-insensitive.

To ensure robust model performance, the feature matrices were class balanced by under sampling the majority class. This balancing step mitigated bias and enhanced the overall accuracy of the predictive models.

3.7.2. Random Forest Classifier

We implemented a random forest classifier using scikit-learn (Pedregosa et al. 2012). The dataset was randomly partitioned into training, validation, and test subsets, typically at a ratio of 80% training and 20% test. A model consisting of 100 trees was trained on the training set and the final performance evaluation was conducted on the unseen test set. Feature importance was analyzed to determine the contribution of individual genomic features to the classification accuracy using Mean Decrease Impurity (MDI):

$$MDI(X_j) = \frac{1}{N_T} \sum_{T} \sum_{t \in T : v(s_t) = X_j} p(t) \Delta i(s_t, t)$$

Where N_t is the number of trees, T represents individual trees, s_t is the split at node t, $v(s_t)$ is the feature used in split s_t, p(t) is the proportion of samples reaching node t, and $\Delta i(s_t,t)$ is the impurity decrease resulting from split s_t.

3.7.3. Ridge Linear Regression

Multiple ridge linear regression models were implemented using scikit-learn's Ridge module with an alpha penalty value of 0.5 (Pedregosa et al. 2012). A sampling approach was adopted in which 70% of the dataset was randomly sampled multiple times (e.g., 100 iterations). Rather than focusing solely on predictive performance, we extracted and analyzed the regression coefficients (beta values) from each model iteration. These coefficients provided insights into feature importance, reflecting the magnitude and direction of each feature's relationship with the response variable. Ridge regression was specifically chosen because its regularization term addresses multicollinearity among predictors by shrinking coefficient estimates, thereby reducing variance and improving the model's stability and interpretability.

$y = X_i\beta + \epsilon$ where $\hat{\beta}_k = (X'X + kI)^{-1}X'y$

3.7.4. Evaluation Metrics

We evaluated model performance using standard metrics such as Accuracy, Area Under the Receiver Operating Characteristic Curve (AUROC), and Area Under the Precision Recall Curve (AUPRC). The formulas are as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$TRP = \frac{TP}{TP + FN}, \qquad FPR = \frac{FP}{FP + TN}$$

$$AUROC = \int_0^1 TPR(FPR)\, d(FPR)$$

$$\text{Precision} = \frac{TP}{TP + FP}, \qquad \text{Recall} = \frac{TP}{TP + FN}$$

$$AUPRC = \int_0^1 P(R)\, dR$$

3.8. StringDB to predict BAF-Protein Interaction

StringDB is a comprehensive database that integrates known and predicted protein-protein interactions from various sources, including experimental data, curated databases, and computational predictions (Szklarczyk et al., 2023). For this analysis, I input all relevant protein features alongside the BAF subunits into StringDB. This allowed for a broad exploration of potential interactions between the BAF complex and other proteins of interest.

To refine the predictions, I focused on selecting the highest interaction score for each BAF-related interaction. Specifically, when multiple scores were available for interactions involving a BAF subunit, the maximum value was retained. This approach ensured that the strongest and most confident predictions were considered for downstream analyses.

4. Results

4.1. Treatment of BRM014 to GM12878 cells results in genome wide loss in accessibility

Upon processing of BRM014 and DMSO treated GM12878 cell ATAC-seq libraries, DeepTools was used to plot the enrichment of sequencing signal over bed regions corresponding to accessible chromatin in both treatment groups (Figure 2a). Such plotting demonstrates that there is a noticable genome-wide loss of chromatin accessibility, and that inhibition of the BAF-ATPase subunit in GM12878 achieves similar observations to previous studies in different cell types such as HAP1 and mESCs (Sheick et al., lurlaro et al.).

Furthermore, loading the generated signal tracks onto a UCSC genome browser, certain peaks (often in close proximity to each other) were seen to be differentially affected by BAF inhibition (figure 2b). This further confirms the findings in Sheick et al. and Lurlaro et al. a demonstrates the GM12878 is an appropriate model cell line to investigate differential sensitivity of accessible chromatin to BAF inhibition.

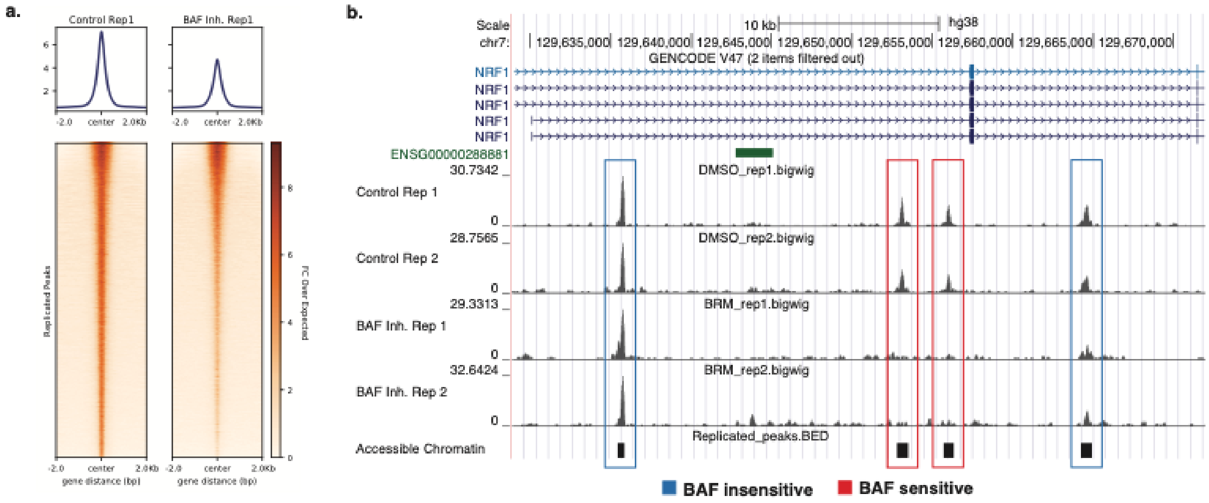

Figure 2. (a) A heatmap of ATAC-seq signals over all replicated called peaks from both BRM014 and DMSO treated GM12878 cells. A significant decrease in chromatin accessibility can be seen from the control to the BAF inhibited sample. (b) A UCSC genome browser shot demonstrating how even within the same gene, accessible chromatin reacts to BAF inhibition in different manners. The regions boxed in blue are BAF insensitive chromatin, while the regions boxed in red are BAF sensitive chromatin.

4.2. Machine learning predicts BAF sensitivity of open chromatin at high performance

Creation of the binary feature matrix yields a matrix of 107,335 samples (corresponding to a region of accessible chromatin) and 162 features comprised of 151 TFs/proteins and 11 Histone Modifications. Here we filtered out peaks that had a log2FC > 1 as these peaks represent accessibility gaining peaks which is interesting, but not part of the biological question we hope to pursue. Upon drawing the BAF-sensitivity cutoff at log2FC = -1, there were 45,517 samples in the positive case (corresponding with BAF sensitivity) and 57,274 samples in the negative case (corresponding with BAF insensitivity). After under-sampling the negative class to balance the data, a final feature matrix of 91,034 samples and 162 features were obtained.

After creating the binary feature matrix, a scikit-learn random forest classifier comprising of 100 trees was trained with the following parameters: [criterion='squared_error', max_depth=None, min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features=1.0, max_leaf_nodes=None]. On a held back testing set of 20% of the total data, the trained model achieves a high performance of 78.80%. Furthermore, the model achieved an AUROC of 0.86 and an AUPRC of 0.84. The model then attempted to predict on a second biological replicate of GM12878 cells where it achieved an even better performance at accuracy = 81.71%, AUROC = 0.88, and AUPRC = 0.86 (Figure 3a,b). This gives us confidence that our model is not overfitting to noise from batch effect.

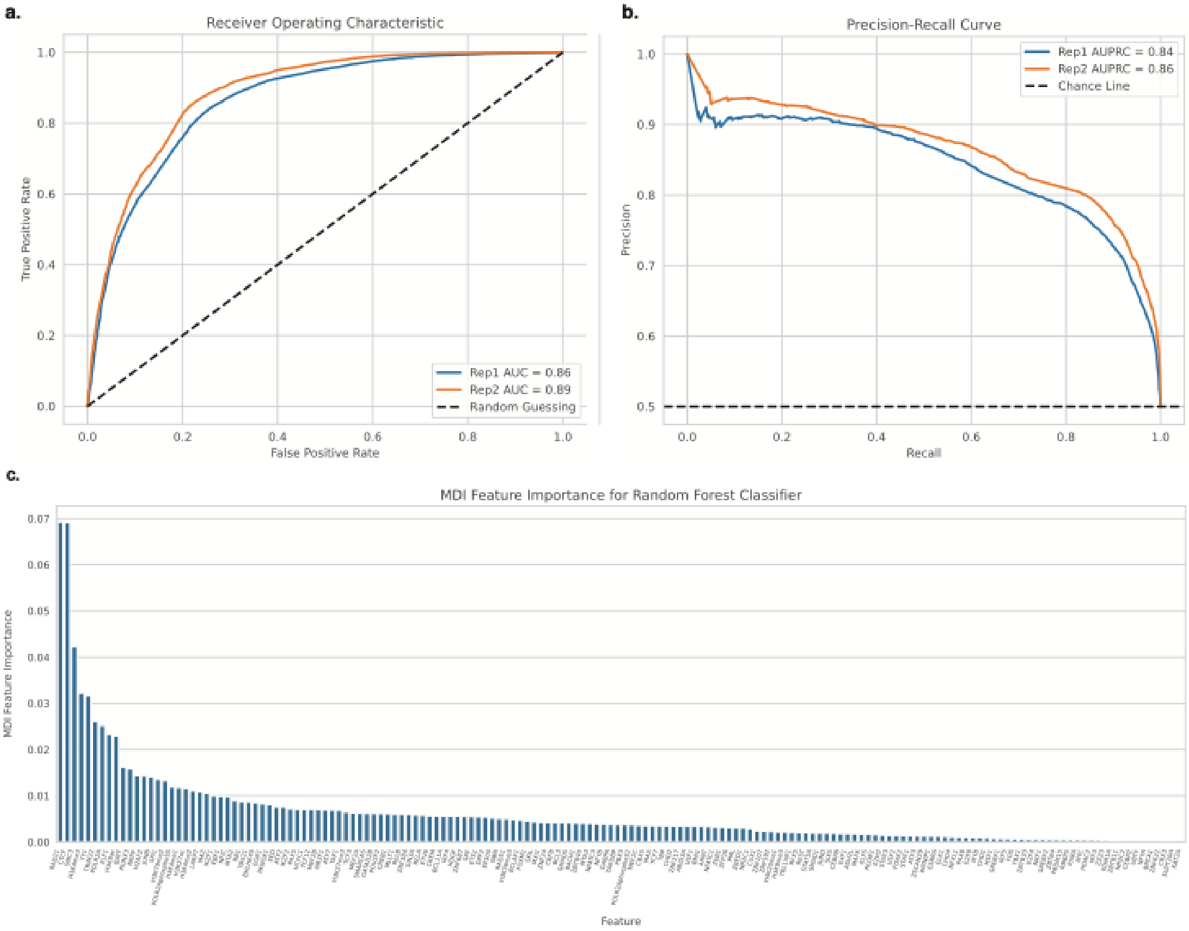

Figure 3. Performance evaluation and feature importance analysis of the random forest classifier. (a) Receiver Operating Characteristic (ROC) curve demonstrating an AUROC of 0.86, which reflects the model's ability to distinguish between BAF-sensitive and BAF-insensitive chromatin regions. (b) Precision-Recall (PR) curve illustrating the balance between precision and recall across different thresholds, with the area under the curve (AUPRC) indicating robust predictive performance. (c) Feature importance plot based on Mean Decrease Impurity (MDI), highlighting key genomic features - including transcription factor binding profiles and histone modifications - that drive the classification.

4.3. Mean Decrease in Impurity (MDI) index highlight important features in ML prediction

After training the Random Forest Classifier, an MDI feature importance analysis was carried out on the model using the formula outlined in Methods: Random Forest Classifier. Ranking the most important features to the least, features identified as important in previously publish studies were identified such as CTCF and Cohesin subunits (RAD21 & SMC3), as well as promoter associated features such as H3K4me4, H3K9ac, and RNA Polymerase II (insert citations). Interestingly however, certain key transcription factors also get parsed out of this analysis. Mainly, YY1, ELF1, RUNX3, BATF, NKRF, JUNB, and SPI1 appeared at the top of the feature importance analysis (Figure 3c). These are all transcription factors which are important for gene regulation in immune cells which is unsurprising given the fact that GM12878 is a lymphoblastoid cell line.

4.4. Linear regression helps distinguish features that predict BAF sensitivity vs insensitivity

While an exploration of RF feature importance can provide us with the importance of the feature to BAF-sensitive accessible chromatin classification, it cannot provide us the directionality of the feature. In other words, if a feature exist in a certain sample, does that mean that it is indicative of a BAF-sensitive chromatin or BAF-insensitive chromatin? To answer this question, a ridge regularized linear regression model was trained on the same feature matrix used to train the random forest classifier. This time, however, the target was not converted to binary as a regression is able to fit to a continuous target.

From here, the sign and magnitude of the β values inform us about the directionality of predictiveness for each individual features. Positive values indicate predictiveness for BAF-insensitivity (accessible chromatin that remain accessible upon BAF inhibition) while negative values predict BAF-sensitivity (accessible chromatin that lose accessibility upon BAF inhibition). After fitting the ridge regularized linear regression, plotting of the β values from most negative to most positive result in many of the features appearing on each pole of the plot. As expected, CTCF and cohesion subunits as well as promoter associated features appear to congregate to the positive Beta values. Interestingly, much of the TFs identified as important in the MDI feature importance appears to have the most negative β values (Figure 4a).

Importantly, the magnitude of the Beta values for each feature correlate strongly with the MDI feature importance upon normalization of feature values, which give us confidence that the feature importance we are calculating agrees across methods (Figure 4b).

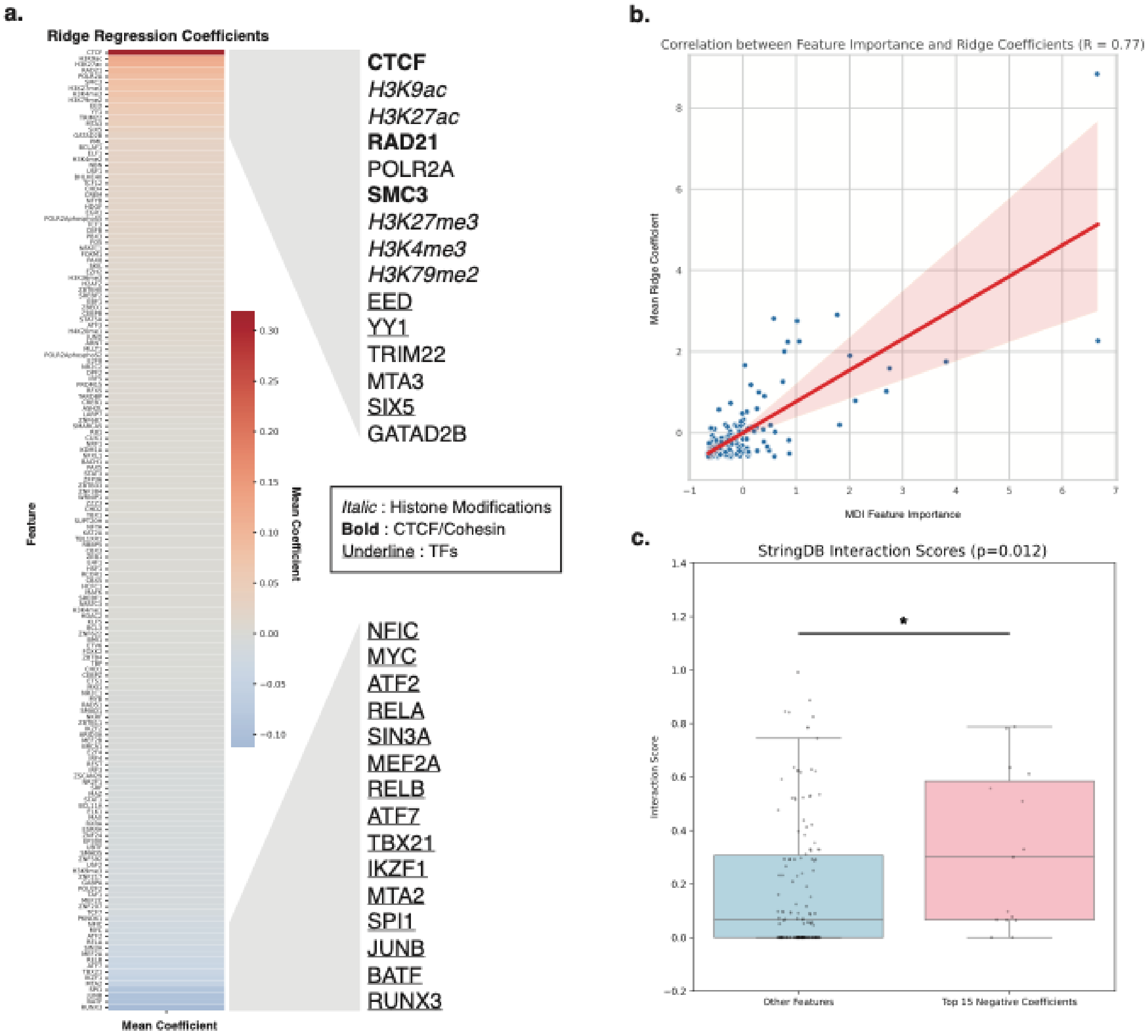

Figure 4. Feature analysis and protein interaction enrichment. (a) Ridge regression coefficients, ordered from most positive to most negative, with the 15 most positive (indicative of BAF-insensitivity) and 15 most negative (indicative of BAF-sensitivity) coefficients highlighted. (b) Scatter plot illustrating the correlation between Mean Decrease Impurity (MDI) from the random forest classifier and the absolute value of ridge regression coefficients, demonstrating strong concordance between the two importance metrics. (c) Boxplot comparing StringDB interaction scores for the 15 features with the most negative ridge coefficients against all other features, revealing a statistically significant enrichment of interactions among the top negative predictors, suggestive of their potential role as BAF recruiters.

4.5. BAF-sensitivity predicting features are enriched for BAF interaction

To assess whether features predictive of BAF-sensitive chromatin are functionally linked to BAF complexes, we used StringDB to retrieve protein-protein interaction scores between each transcription factor or cofactor feature and known BAF subunits, retaining the highest interaction score per feature. We then compared the top 15 features with the most negative Beta values (i.e., strongest predictors of chromatin accessibility loss upon BAF inhibition) against all other non-histone features. These top features showed a statistically significant enrichment for high StringDB interaction scores with BAF subunits (p = 0.021; Figure 4c).

5. Discussion

5.1. Machine learning cam predict the loss of chromatin accessibility upon BAF inhibition based on protein binding and histone modification

A random forest model trained on chromatin features such as TF binding and histone modifications can accurately predict the BAF-sensitivity of a region of accessible chromatin at a very high performance. This means that the chromatin landscape is indeed informative in distinguishing between BAF-sensitive and insensitive chromatin. This finding provides motivation to examine histone modifications and TFs in relation to BAF activity.

5.2. CTCF/cohesin and promoter-associated modifications are predictive of chromatin accessibly retention

CTCF, cohesin subunits, and promoter associated factors are implicated as highly important and highly predictive of BAF-insensitivity in both the MDI feature importance and the linear regression analysis. This observation is significant as it serves to confirm that certain trends which are common across different cell types (HAP1 and mESCs) can also be seen in GM12878. This finding also points towards the fact that loss-of-function mutations in BAF are most likely not affecting these regions of the genome.

5.3. Lineage Determining and Enhancer associated TFs are predictive of chromatin accessibility loss

On the other side of the spectrum, TFs such as RUNX3, BATF, JUNB, SPI1, and IKZF1 are implicated as important to the RF classification and highly predictive of BAF-sensitivity. Many of these transcription factors, such as RUNX3, BATF, and SPI1 are lineage determining genes and are indicative of enhancer binding.

5.4. TFs that predict BAF sensitivity point to potential mechanisms of BAF recruitment to chromatin.

There is a statistically significant enrichment of experimentally derived StringDB scores against BAF subunits in the 15 most negative features compared to all other proteins screened in this study. Among these negative features include many TFs of the AP1 family (such as JUNB, BATF, ATF2, and ATF7), of which JUNB has been experimentally shown to have BAF recruiting activity (Vierbuchen et al., 2017).

Altogether, in this paper I present a data analysis pipeline where ML is used to predict BAF-sensitivity of chromatin based on chromatin features. Upon feature analysis of ML models, certain patterns can be seen. These patterns point towards a biological significance of BAF sensitive cREs as binding sites for TFs which may function as BAF recruiters.

5.5. Limitations and Future directions

While the current study offers valuable insights into the chromatin dynamics following BAF inhibition, several limitations remain. First, my analysis is confined to the GM12878 cell line, which may not fully capture the characteristics present in other cell types. Moreover, the machine learning models rely on a predetermined set of chromatin features; additional epigenetic marks or transcription factors not included in the current feature matrix might also contribute to BAF sensitivity. Lastly, although I have proposed certain transcription factors as potential BAF recruiters, experimental validation is still required to confirm these interactions and prove their functional significance.

Future research should extend this analysis to multiple cell types, including both pluripotent and differentiated cells, to enhance the generalizability of the predictive models. Incorporating additional multi-omics data, such as RNA-seq and ChIP-seq for a broader range of histone modifications, could refine the feature matrix and improve model performance. Furthermore, systematic laboratory-based assays are needed to screen and validate the predicted BAF recruiters, thereby bridging the gap between computational predictions and biological function.

References

1. Bao, X. *et al.* A novel ATAC-seq approach reveals lineage-specific reinforcement of the open chromatin landscape via cooperation between BAF and p63. *Genome Biol* **16**, 284 (2015).

2. Schick, S. *et al.* Acute BAF perturbation causes immediate changes in chromatin accessibility. *Nat Genet* **53**, 269–278 (2021).

3. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

4. Vierbuchen, T. *et al.* AP-1 Transcription Factors and the BAF Complex Mediate Signal-Dependent Enhancer Selection. *Molecular Cell* **68**, 1067-1082.e12 (2017).

5. Buenrostro, J. D., Wu, B., Chang, H. Y. & Greenleaf, W. J. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *CP Molecular Biology* **109**, (2015).

6. Kouzarides, T. Chromatin Modifications and Their Function. *Cell* **128**, 693–705 (2007).

7. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

8. Iurlaro, M. *et al.* Mammalian SWI/SNF continuously restores local accessibility to chromatin. *Nat Genet* **53**, 279–287 (2021).

9. Zhang, Y. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137 (2008).

10. Mashtalir, N. *et al.* Modular Organization and Assembly of SWI/SNF Family Chromatin Remodeling Complexes. *Cell* **175**, 1272-1288.e20 (2018).

11. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. (2012) doi:10.48550/ARXIV.1201.0490.

12. Mathur, R. & Roberts, C. W. M. SWI/SNF (BAF) Complexes: Guardians of the Epigenome. *Annu. Rev. Cancer Biol.* **2**, 413–427 (2018).

13. Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**, 75–82 (2012).

14. Alfert, A., Moreno, N. & Kerl, K. The BAF complex in development and disease. *Epigenetics & Chromatin* **12**, 19 (2019).

15. Hodges, C., Kirkland, J. G. & Crabtree, G. R. The Many Roles of BAF (mSWI/SNF) and PBAF Complexes in Cancer. *Cold Spring Harb Perspect Med* **6**, a026930 (2016).

16. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

17. Szklarczyk, D. *et al.* The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Research* **51**, D638–D646 (2023).

18. Perez, G. *et al.* The UCSC Genome Browser database: 2025 update. *Nucleic Acids Research* **53**, D1243–D1249 (2025).

19. Ho, P. J., Lloyd, S. M. & Bao, X. Unwinding chromatin at the right places: how BAF is targeted to specific genomic locations during development. *Development* **146**, dev178780 (2019).

20. Maria & Buck, M. J. Chromatin accessibility: a window into the genome. *Epigenetics & Chromatin* **7**, 33 (2014).

21. Li, B., Carey, M. & Workman, J. L. The Role of Chromatin during Transcription. *Cell* **128**, 707–719 (2007).