DCU
Ollscoil Chathair
Bhaile Átha Cliath
Dublin City University

CONSILIO 1784 MANUQUE
RCSI UNIVERSITY
OF MEDICINE
AND HEALTH
SCIENCES

Using Neural Networks To Understand Gene Regulation In the MCF-7 Breast Cancer Cell Line

Student Name: Faith Ogundimu

Student Number: 21715715

Course: Genetics and Cell Biology

Module: BIO1018 Research Project

Supervisor: Senior Lecturer Simon Furney

Co-Supervisor: Assistant Professor Linda Holland

Student Declaration of Academic Integrity

I understand that the University regards breaches of academic integrity and plagiarism as grave and serious.

I have read and understood the DCU Academic Integrity and Plagiarism Policy. I accept the penalties that may be imposed should I engage in practice or practices that breach this policy.

I have identified and included the source of all facts, ideas, opinions and viewpoints of others in the assignment references. Direct quotations, paraphrasing, discussion of ideas from books, journal articles, internet sources, module text, or any other source whatsoever are acknowledged and the sources cited are identified in the assignment references.

I declare that this material, which I now submit for assessment, is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

I have used the DCU library referencing guidelines (available at https://www101.dcu.ie/library/Citing&ReferencingGuide/player.html) and/or the appropriate referencing system recommended in the assignment guidelines and/or programme documentation.

By signing this form or by submitting material for assessment online I confirm that I have read and understood the DCU Academic Integrity and Plagiarism Policy (available at http://www.dcu.ie/registry/examinations/index.shtml).

Signature: Faith Ogundimu

Table of Contents

List of Abbreviations

1. **AI** – Artificial Intelligence
2. **ATAC-seq** – Assay for Transposase-Accessible Chromatin using sequencing
3. **AUC** – Area Under the Curve
4. **AUC-ROC** – Area Under the Receiver Operating Characteristic Curve
5. **AUPRC** – Area Under the Precision-Recall Curve
6. **BED** – Browser Extensible Data
7. **ChIP-seq** – Chromatin Immunoprecipitation sequencing
8. **chr** – chromosome
9. **COMPASS** – Complex of Proteins Associated with Set1
10. **DNase-seq** – DNase I hypersensitive site sequencing
11. **ER+** – Oestrogen Receptor Positive
12. **ESR1** – Oestrogen Receptor 1
13. **FOXA1** – Forkhead Box A1
14. **GATA3** – GATA Binding Protein 3
15. **GRCh38** – Genome Reference Consortium Human Build 38
16. **Hi-C** – High-throughput chromosome conformation capture
17. **HOMER** – Hypergeometric Optimization of Motif EnRichment
18. **HMMRATAC** – Hidden Markov ModeleR for ATAC-seq
19. **H3K4me1** – Histone 3 Lysine 4 Monomethylation
20. **H3K4me2** – Histone 3 Lysine 4 Dimethylation
21. **H3K4me3** – Histone 3 Lysine 4 Trimethylation
22. **H3K9ac** – Histone 3 Lysine 9 Acetylation
23. **H3K9me3** – Histone 3 Lysine 9 Trimethylation
24. **H3K27ac** – Histone 3 Lysine 27 Acetylation
25. **H3K27me3** – Histone 3 Lysine 27 Trimethylation
26. **H3K36me3** – Histone 3 Lysine 36 Trimethylation
27. **KDM1A** – Lysine Demethylase 1A
28. **KDM1B** – Lysine Demethylase 1B

29. **KMT2C** – Lysine Methyltransferase 2C
30. **KMT2D** – Lysine Methyltransferase 2D
31. **LOCO** – Leave-One-Chromosome-Out
32. **LSD1** – Lysine-Specific Demethylase 1
33. **LSD2** – Lysine-Specific Demethylase 2
34. **MCC** – Matthews Correlation Coefficient
35. **MCF-7** – Michigan Cancer Foundation 7
36. **OLS** – Ordinary Least Squares
37. **PARP** – Poly (ADP-ribose) Polymerase
38. **RNA-seq** – RNA sequencing
39. **SCLC** – Small-Cell Lung Cancer
40. **SHAP** – SHapley Additive exPlanations
41. **TNBC** – Triple-Negative Breast Cancer
42. **TFs** - transcription factors
43. **PR+** - progesterone receptor-positive
44. **CNAs** - copy number alterations
45. **TP** - True Positives
46. **TN** - True Negatives
47. **FP** - False Positives
48. **FN** - False Negatives
49. **AML** - acute myeloid leukemia
50. **BRCA1** - Breast Cancer gene 1
51. **MYC** - MYC proto-oncogene, bHLH transcription factor
52. **YBX1** - Y-box binding protein 1
53. **PI3K** - Phosphatidylinositol-3-kinase
54. **AKT** - Protein Kinase B
55. **WDR5** - WD repeat domain 5

Lay Abstract

Genes are like instruction manuals for cells. However, these instructions are not always accessible, some are "open" and easy to read, while others are "closed" and harder to access. This “accessibility” plays a crucial role in determining which genes are turned on or off, affecting how cells work. Changes in gene accessibility are especially important in diseases like cancer, where cells behave abnormally.

In this study, artificial intelligence (AI) models were developed on a breast cancer cell line called MCF-7 to predict whether a gene’s instructions are open or closed based on chemical signals in the cell called histone modifications, and protein signals called transcription factor signals. These signals help control gene activity. Machine learning and deep learning techniques were used to train these models using publicly available biological data. The deep learning model was compared against an existing tool that predicts chromatin accessibility using similar information (Zhao et al., 2022). Even though it used fewer features, the model in this study performed better, showing that it can find open and closed chromatin regions without needing DNA sequence data. Among the different chemical signals tested, H3K4me1 was found to be the most important for making correct predictions.

Future work will look at more types of chromatin states, not just open or closed, and will use RNA data to understand which genes are actually turned on. It will also test other breast cancer cell lines to see if the patterns hold true. By combining different types of data, this approach could help find new drug targets by showing which proteins are controlling gene activity in tumours. A new idea is to predict how chromatin might change over time, which could help spot early signs of cancer progression before they happen.

Scientific Abstract

Chromatin accessibility is a key determinant of gene regulation, influencing transcription factor binding and transcriptional activation. Predicting accessible chromatin regions from histone modifications and transcription factor signals has major implications for understanding epigenetic mechanisms and cancer-specific regulation. However, studies often rely solely on AUC-ROC for evaluation, overlooking metrics like MCC and F1 Score, which are critical for imbalanced cancer datasets.

Chromatin accessibility is highly cell type-specific, with the most predictive histone marks and transcription factors varying by context. Capturing these cancer-specific dynamics is essential, as regulatory mechanisms differ between normal and malignant cells. This study develops and evaluates machine learning and deep learning models using histone modification, transcription factor binding, and ATAC-seq data from the MCF-7 breast cancer cell line (ENCODE). A deep learning model was built using shared feedforward layers to process histone and transcription factor inputs.

The model was benchmarked against an existing predictor that includes histone marks, TF motifs and DNA sequence (Zhao *et al.*, 2022). Despite using fewer features, the model in this study outperformed it, showing strong predictive power for chromatin accessibility. This neural network appears to be the first to focus solely on histone modifications and transcription factor binding signals to study mechanistic drivers of cancer gene regulation. Feature importance analysis identified H3K4me1, the enhancer priming mark, as most predictive, consistent with known chromatin biology.

Future work will explore chromatin state classification, integrate RNA-seq to link accessibility with gene expression and apply the approach to additional breast cancer cell lines. An exploratory aim is to model chromatin velocity to predict future chromatin states.

Introduction

Cancer’s Unique Chromatin Landscape

Cancer has continued to be one of the leading causes of death worldwide (Ritchie, Spooner and Roser, 2018). Artificial intelligence is an upcoming approach that has looked to tackle the complexity of cancer by prioritising the understanding of its underlying mechanisms.

Gene regulation is controlled by complex and dynamic interactions between chromatin structure, transcription factors (TFs) and histone modifications. Chromatin accessibility, in particular, serves as a critical determinant of whether genomic regions are permissive to transcriptional activation or remain silenced (Mansisidor and and Risca, 2022). Abnormal chromatin states are a hallmark of cancer, where disruptions in regulatory networks contribute to uncontrolled cell proliferation and disease progression (Hanahan and Weinberg, 2011; Locke *et al.*, 2015; Hanahan, 2022).

Characteristics of the MCF-7 Breast Cancer Cell Line

The Michigan Cancer Foundation 7 (MCF-7) cell line is a well-characterised model of luminal A breast cancer, which is oestrogen receptor-positive (ER+), progesterone receptor-positive (PR+) and HER2-negative. It was derived from a human breast donor via pleural effusion in 1973 and has since been widely used to investigate hormone-responsive breast cancer due to its dependency on oestrogen for proliferation and tumour formation in vivo (Welsh, 2013). It remains a gold-standard model for evaluating endocrine therapies such as tamoxifen and exploring ER-mediated signalling pathways (Holliday and Speirs, 2011; Beaver *et al.*, 2013).

Genomically, MCF-7 harbours a number of hallmark mutations and copy number alterations (CNAs) representative of luminal breast cancers. These include a hotspot *PIK3CA* mutation (E545K) that activates the PI3K/AKT pathway, a frameshift mutation in *GATA3* impacting transcriptional regulation, and a homozygous deletion of *CDKN2A* (p16INK4a) which contributes to cell cycle deregulation (Beaver *et al.*, 2013; Liang *et al.*, 2018). Uniquely, MCF-7 retains wild-type *TP53*, aligning with many primary luminal A tumours. It also exhibits high-level amplifications at loci including 1p13.1-p21.1, 17q22-q24.3, and 20q13.33 which are regions frequently amplified in ER+ cancers (Hampton *et al.*, 2009). These features make

MCF-7 a robust model for studying epigenetic regulation, chromatin accessibility and therapeutic vulnerabilities in ER-positive breast cancer.

Histone Modifications and their Influence on the Regulatory Epigenome

Among the histone modifications studied in this project, H3K4me1 is typically found at enhancer elements and is particularly enriched at regions that are poised for activation. It serves as a marker of potential regulatory activity and often works in tandem with other activating marks, such as H3K27ac (Creyghton *et al.*, 2010). H3K4me3 is a hallmark of active promoters and is found near transcription start sites, reflecting ongoing or recent gene transcription (Vakoc *et al.*, 2006). H3K27ac, an acetylation mark, is also found in enhancers and promoters, but in contrast to H3K4me1, it marks enhancers that are actively engaged in gene activation. The combination of H3K4me1 and H3K27ac is widely used to distinguish between poised and active enhancers (Creyghton *et al.*, 2010).

In contrast, H3K27me3 is a repressive mark deposited by Polycomb group proteins. It is commonly found in regions of facultative heterochromatin and is involved in long-term gene silencing during development and differentiation (Young *et al.*, 2011). H3K9me3 is associated with constitutive heterochromatin and marks regions of the genome that remain stably repressed, such as pericentromeric domains (Padeken, Methot and Gasser, 2022). Finally, H3K36me3 is found within gene bodies of actively transcribed genes and is thought to play a role in transcription elongation and co-transcriptional RNA processing (Vakoc *et al.*, 2006).

The balance and spatial arrangement of these histone marks contribute to the overall chromatin state, influencing whether a genomic region is accessible or closed. Their combinatorial patterns define epigenomic landscapes that are dynamic, cell-type specific, and tightly linked to gene regulatory networks. Understanding these marks provides essential context for modelling chromatin accessibility and identifying the regulatory mechanisms behind transcriptional control in cancer.

Oestrogen Receptor Positive (ER$^+$)-Associated Breast Cancer Transcription Factors and their Role in Defining the Transcriptional Landscape

Transcription factors are sequence-specific DNA-binding proteins that control the transcription of genetic information from DNA to messenger RNA. They act as central regulators of gene expression by interacting with promoter and enhancer elements, often in coordination with chromatin-modifying complexes. Beyond simply recognising DNA motifs, transcription factors can influence chromatin architecture by recruiting co-activators, co-repressors and chromatin remodelling enzymes (Weidemüller *et al.*, 2021).

In the context of breast cancer, several transcription factors play pivotal roles in defining the transcriptional landscape of hormone receptor-positive tumours. Oestrogen Receptor 1 (ESR1) is one of the most studied transcription factors in breast cancer and functions as a ligand-activated nuclear receptor. Upon binding oestrogen, ESR1 translocates to the nucleus and binds to oestrogen response elements in the genome, where it recruits co-regulators and chromatin remodelers that facilitate gene activation. It governs a wide range of cellular processes including proliferation, differentiation and survival, and is a key driver in the luminal subtype of breast cancer (Hua *et al.*, 2018).

FOXA1 is a pioneer transcription factor that can bind to condensed chromatin and facilitate the recruitment of other transcription factors such as ESR1. It plays a crucial role in remodelling the chromatin landscape and enabling hormone-dependent transcriptional activity. FOXA1 is essential for luminal lineage specification and its expression correlates with better prognosis in ER-positive breast cancers (Augello, Hickey and Knudsen, 2011).

GATA3 is another critical luminal-specific transcription factor that acts both independently and in tandem with ESR1 and FOXA1. It helps maintain epithelial cell identity and regulates genes involved in differentiation and proliferation. GATA3 mutations are common in breast cancer and often affect its DNA-binding domain, altering its regulatory functions (Adomas *et al.*, 2014).

Together, these transcription factors orchestrate complex regulatory networks that are tightly linked to chromatin accessibility. Their binding sites are enriched in accessible regions of the genome and their activity is often reflected in changes to the surrounding histone modification landscape. By including their binding signals in predictive models, one can capture a mechanistically informative snapshot of the regulatory environment governing gene expression in breast cancer cells.

Current Research on Chromatin Accessibility

High-throughput techniques such as ATAC-seq and ChIP-seq have revolutionised the ability to interrogate chromatin accessibility and the epigenomic signatures underlying gene regulation (Park, 2009; Buenrostro *et al.*, 2015). However, traditional methods of analysing such data often lack the scalability and predictive capacity required to infer regulatory patterns across the genome and in unseen biological contexts (Yan *et al.*, 2020).

The mapping of accessible regions across the genome using assays like ATAC-seq has significantly advanced our ability to identify functionally relevant genomic elements. However, while ATAC-seq and similar techniques provide a high-resolution readout of chromatin openness, they do not offer mechanistic insight into why certain regions are accessible (Zhang *et al.*, 2008; McCarthy and O'Callaghan, 2014; Tarbell and Liu, 2019). To address this gap, researchers are increasingly turning to computational models to infer chromatin accessibility from underlying molecular features such as histone marks and transcription factor binding signals (Zhao *et al.*, 2022).

Mechanistic Modelling of Chromatin Accessibility in Breast Cancer - Objectives, State of the Art and Implications for Cancer Epigenomics

Few models exist that rely solely on mechanistic epigenetic features to predict accessibility and to my knowledge, none have done so specifically in the context of breast cancer. This study addresses that gap by developing a neural network model to predict chromatin accessibility using only histone modification and transcription factor ChIP-seq signal data, without incorporating DNA sequence information. To ensure robustness and assess generalisability across the genome, a leave-one-chromosome-out (LOCO) validation strategy was employed. This approach held out chromosome 1 for testing while training on all others, thereby reducing the risk of data leakage and overfitting to local sequence contexts. Compared to traditional random-split methods, LOCO validation better simulates how models would perform on entirely unseen genomic regions, making it a more stringent and biologically relevant validation technique (Mbatchou *et al.*, 2021).

In addition to the neural network, baseline models including Extra Trees Classifier and Histogram Gradient Boosting Classifier from the *sklearn.ensemble* package (Pedregosa *et al.*, 2011) were implemented to benchmark performance. A linear regression model was also included to confirm the significance of selected features.

This work has several implications for cancer research. Firstly, it assesses whether chromatin accessibility can be effectively modelled from mechanistic data alone. Secondly, by identifying key histone marks that regulate accessibility, this approach contributes to our understanding of epigenetic dysregulation in breast cancer and highlights candidate features for further investigation. This research will also pave the way for future applications in other cancer types and cell lines, enabling comparative analyses of epigenomic landscapes across disease states.

Ultimately, this study builds a foundation for mechanistic, non-sequence-based modelling of chromatin accessibility in cancer. It underscores the value of integrating epigenetic signals to decipher regulatory dynamics and provides a practical framework for future work aimed at linking accessibility to gene expression, targeted drug development and clinical phenotypes in tumour evolution.

The Importance of Choosing the Correct ML Evaluation Metrics when Answering Biological Questions

Class imbalance occurs when one class is represented far less than the other, in binary classification models. In genomic and epigenomic studies, class imbalance is common, for instance in this study, when open chromatin represents only a small fraction (~2%) of the entire genome. Under these conditions, traditional accuracy becomes misleading, as high scores can be achieved simply by predicting the dominant class (i.e. closed chromatin). Metrics such as precision, recall, F1-score, MCC and AUC-ROC are more informative. Precision assesses how many predicted positives are correct, while recall (or sensitivity) measures how many actual positives have been identified. The F1-score balances precision and recall and is commonly used in imbalanced binary classification tasks; however, it ignores true negatives and can overestimate performance in cases where false positives also carry weight, particularly when addressing biological questions with potential clinical relevance (Chicco and Jurman, 2020; Rauschert *et al.*, 2020).

The Matthews Correlation Coefficient (MCC) addresses these limitations by incorporating all four values of the confusion matrix (TP, TN, FP, FN), offering a balanced, single-number summary. Unlike F1 or AUC-ROC, MCC penalises disproportionate errors on either class, making it especially well-suited for cancer epigenetics, where both sensitivity and specificity are critical. A model predicting only the dominant class in a highly imbalanced dataset may show high accuracy but yield an MCC of zero, accurately reflecting its lack of predictive power (Chicco and Jurman, 2020; Newsham *et al.*, 2024). The use of all metrics outlined ensures the neural network performs robustly across both classes, providing a stringent and interpretable assessment of model quality under biologically realistic imbalance.

Materials and Methods

Data Acquisition

To construct the deep learning models to predict chromatin accessibility, the well-characterised breast cancer cell line, MCF-7 was used. To ensure high-quality data on chromatin accessibility, transcription factor binding and histone modifications, datasets from ENCODE were selected (*Applications of ENCODE data to systematic analyses via data integration - ScienceDirect*, 2018). The chromatin accessibility data was obtained by using bulk ATAC-seq and downloaded from the ENCODE database as a BED file. The accession ID is ENCFF821OEF. The ChIP-seq profiles of histone modifications and transcription factors were also downloaded from ENCODE as bigWig files. A limitation of this study is that only a single ATAC-Seq peak file (ENCFF821OEF) was used, as no additional independent replicates were available on ENCODE. However, given the high-quality standards of ENCODE data processing, including rigorous peak calling and reproducibility checks, this is unlikely to significantly impact the reliability of the findings. All files were mapped to the GRCh38 genome. The annotation of all files is summarised in Supp. Table 1. All ChIP-Seq signal values were reported as p-values, representing the statistical significance of enrichment at each genomic position. ATAC-Seq peaks were defined using a pseudoreplicated peak-calling approach, ensuring reproducibility across isogenic replicates.

ATAC-Seq Data Processing and Visualisation

The distribution of chromatin accessibility was visualised across the genome, ATAC-Seq peak density was plotted for each chromosome (chr) (chr1–chr22 and chrX). Peaks from the ATAC-Seq dataset were binned into genomic intervals, and density histograms were generated to assess the frequency and distribution of accessibility sites along the genome, Supp. Figure 1. This allowed for the identification of chromatin accessibility patterns and potential sequencing biases or coverage errors in peak distribution before model training.

Assignment of Chromatin Accessible and Non-Accessible Regions

To define chromatin accessibility regions, ATAC-Seq peaks were used to label genomic bins as open (1) or closed (0) chromatin. The human genome (GRCh38) was segmented into non-overlapping 1000 base pairs bins, covering all autosomes and chrX (excluding chrY due to the MCF-7 cell line's female origin). Each bin was initially assigned a default closed chromatin (0) state, and bins overlapping ATAC-Seq peaks were labeled as open chromatin (1). This approach allowed for structured representation of chromatin accessibility across the genome while reducing the sparsity of peak-based methods.

A bin size of 1000 base pairs was chosen as a balance between resolution and computational efficiency. This binning strategy ensures that chromatin accessibility is quantified at a biologically relevant scale, approximately corresponding to the size of regulatory elements such as enhancers and promoter regions. Following preprocessing, chromatin accessibility was quantified, yielding 143,817 open chromatin regions (1) and 2,661,272 closed chromatin regions (0), corresponding to ~5.13% overall chromatin accessibility. Previous studies have reported genome-wide accessibility estimates of ~2–3%, primarily in non-cancerous cell types (Klemm, Shipony and Greenleaf, 2019). However, the higher accessibility observed in this study is likely due to the cancerous nature of MCF-7 cells and binning at 1000 base pairs resolution, which may capture broader regulatory activity compared to base-pair-level analyses.

To ensure the biological accuracy of chromatin accessibility peak assignment, peaks were annotated using Hypergeometric Optimization of Motif EnRichment's (HOMER) *annotatePeaks.pl* command, Supp. Figure 2.

Assignment and Normalisation of Histone and Transcription Factor Signals to 1 Kilobase Bins

To systematically assign both histone modification and transcription factor (transcription factor) binding signals to 1 kb genomic bins, ChIP-Seq, bigWig signal data were preprocessed and mapped using a standardised workflow. Signal data from six histone marks (H3K4me1, H3K4me3, H3K27ac, H3K27me3, H3K9me3, H3K36me3) and three transcription factors (ESR1, FOXA1, GATA3) were first converted from bigWig to bedGraph format using the *bigWigToBedGraph* command. These features were selected based on their established roles in chromatin regulation and transcriptional activity in breast cancer (Jin *et al.*, 2020)

The genomic bins and signal files were sorted by chromosome and genomic position to ensure accurate alignment. Using *bedtools map*, mean signal intensity for each feature was computed within each 1 kb bin by mapping the bedGraph files to the genomic bins. This process ensured that both histone modification and transcription factor signals were assigned to the correct genomic regions.

Missing values, typically arising from regions with undetectable signals, were replaced with zero to maintain data consistency. Additionally, rows containing duplicate signals across multiple bins were removed to prevent data leakage during model training.

Baseline Models for Benchmarking Chromatin Accessibility Prediction

To establish benchmark performance for chromatin accessibility prediction, traditional ensemble-based, machine learning models, Extra Trees Classifier and Histogram Gradient Boosting were implemented from the *sklearn.ensemble* package (Pedregosa *et al.*, 2011). Logistic Regression, specifically Ordinary Least Squares (OLS) regression, was used as a statistical framework for evaluating feature importance from p-values and regression coefficients.

Mechanistic Neural Network for Predicting Chromatin Accessibility

Model Architecture

A feedforward neural network was implemented to predict chromatin accessibility using histone modification and transcription factor binding signal data. The input layer received a feature vector containing nine signal intensities (six histone marks and three transcription factor binding signals) per genomic bin. The first hidden layer comprised 64 neurons activated by the ReLU function, followed by a dropout layer (0.3 probability) to prevent overfitting. A second hidden layer with 32 ReLU-activated neurons was introduced, followed by another dropout layer to further regularise the model. The final output layer contained a single neuron with a sigmoid activation function, producing a probability score for chromatin accessibility, where a probability threshold of >0.7 denoted open chromatin (1). All hyperparameters were chosen following grid search hyperparameter tuning, Supp. Table 2. The model was compiled using binary cross-entropy loss and optimised with the Adaptive Moment Estimation (Adam) optimiser (Kingma and Ba, 2017), at a learning rate of 0.001.

Training Strategy, Handling Class Imbalance and Leave-One-Chromosome-Out (LOCO) Validation

To address class imbalance, the majority class (closed chromatin) was downsampled to five times the number of open chromatin regions before training. Features were normalised using Min-Max Scaling (0–1 range), with scaling parameters computed solely on the training set to prevent data leakage. The final model evaluation was conducted using a Leave-One-Chromosome-Out (LOCO) Validation framework. In this setup, chromosome 1 was excluded from training and used exclusively as the test set, ensuring that model predictions were evaluated on completely unseen genomic regions. This approach prevents the model from overfitting to chromosome-specific patterns and better reflects the real-world scenario in which regulatory features must be predicted in new, unobserved genomic contexts.

During training, Stratified K-Fold Cross-Validation (K=5) was employed within the training set to ensure robust performance estimation. Early stopping was implemented, monitoring validation loss with a patience of five epochs, ensuring that training halted before overfitting occurred.

Class weights were adjusted to compensate for remaining class imbalance, assigning a 2x higher weight to open chromatin regions to improve sensitivity for the minority class.

Results

Exploratory Data Analysis to Assess Feature Relevance in Open and Closed Chromatin States

The hypothesis for this study was that the features selected (six histone marks and three transcription factors) are associated with chromatin accessibility and would have adequate predictive capability to enable chromatin accessibility prediction. The hypothesis was validated by assessing the feature signals between open and closed chromatin states using a Mann Whitney U Test. To ensure unbiased evaluation, chromosome 1 was excluded from the analysis, as it was reserved for final model validation, preventing data leakage and ensuring that statistical comparisons were not influenced by the test set, Table 1.

Table 1: Mann-Whitney U Test Statistics for Histone Modifications and Transcription Factor Signals in Open and Closed Chromatin

Feature	U-Statistic	p-value	Interpretation
H3K4me1	2.89×10^{11}	0	Strong association with open chromatin, corresponding with its role as an enhancer mark.
H3K4me3	2.69×10^{11}	0	Strong association with open chromatin, corresponding with its role as a promoter mark.
H3K27ac	2.74×10^{11}	0	Strong association with open chromatin, corresponding with its role in active enhancers and promoters.
H3K27me3	1.07×10^{11}	0	Strongly enriched in closed chromatin, confirming its role as a repressive histone mark.
H3K9me3	1.22×10^{11}	0	Consistently enriched in heterochromatin and repressed genomic regions.

H3K36me3	1.65×10^{11}	2.75 x 10^{-90}	Associated with transcriptional elongation regions, where the lower p-value highlights its mixed enrichment based on chromatin context.
ESR1	2.40×10^{11}	0	Strong enrichment in open chromatin, indicating active regulatory roles.
FOXA1	2.35×10^{11}	0	Enriched in open chromatin, supporting its function in enhancer accessibility.
GATA3	2.31×10^{11}	0	Strong enrichment in open chromatin, consistent with its role in gene regulation.

To visualise the genome-wide distribution of histone modifications in open and closed chromatin regions, histone signals were normalised by chromosome length and densities were plotted for each chromatin state, Supp. Figure 3. Each histone mark showed distinct patterns, providing insight into histone enrichment patterns and their relevance for chromatin state classification. Additionally, to investigate the distribution of histone modification signals across different genomic annotations, histone signals were normalised by chromosome length and analysed within promoter, intergenic, exon, and intron regions, highlighting localisation patterns, Fig. 2.

Evaluation Metrics

All model performances were assessed using multiple evaluation metrics. The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) measured discriminatory power, while the Matthews Correlation Coefficient (MCC) provided a robust assessment of classification quality in imbalanced datasets. Precision, recall, and F1-score were calculated to evaluate predictive performance.

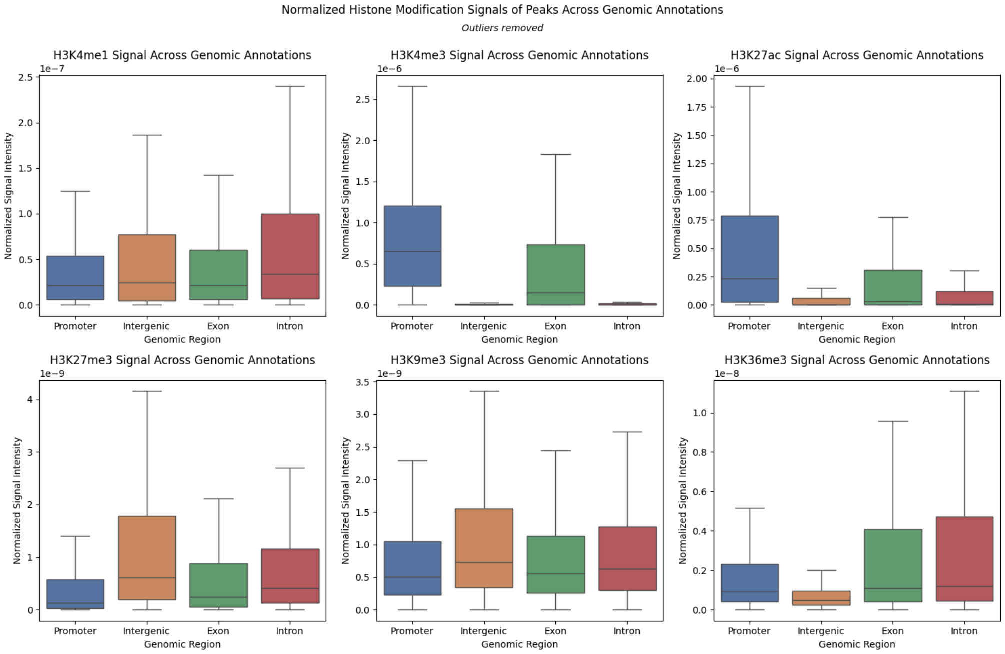


Figure 2: Boxplots depicting the distribution of normalised histone modification signals across genomic annotations. Each panel represents a distinct histone mark, with signals normalised by chromosome length. Promoters exhibit the highest enrichment of H3K4me3 and H3K27ac, while repressive marks H3K27me3 and H3K9me3 are more abundant in intergenic and intronic regions. Outliers have been removed for clarity.

Baseline Models for Benchmarking Predictive Performance

To assess the performance of the neural network, two ensemble-based baseline models, Extra Trees and Histogram Gradient Boosting from the *sklearn.ensemble* package were employed (Pedregosa *et al.*, 2011). Performance was assessed using a leave-one-chromosome-out (LOCO) validation framework, with chromosome 1 held out as an unseen test set. This rigorous approach ensured that predictions were not biased by local genomic context and better simulated real-world application.

Both models demonstrated strong predictive power, achieving AUC scores of 0.934 and 0.938 on the held-out chromosome, respectively. Both models also received scores of over 0.52 for their F1 and a Matthews Correlation Coefficient (MCC) scores, with Histogram Gradient Boosting slightly outperforming. Recall values were high for both classifiers (>0.78), while precision scores remained lower (<0.42), indicating a higher rate of false positives. All performance metrics can be viewed in Fig. 3.

Mechanistic Neural Network Performance on Predicting Chromatin Accessibility

The neural network trained on histone modification and transcription factor binding signal data successfully predicted chromatin accessibility in the MCF-7 breast cancer cell line. This model was trained and evaluated using the same input features and LOCO validation framework as the baseline models. The model architecture consisted of a fully connected neural network, illustrated in Fig. 4.

The final model achieved high performance across all evaluation metrics, an AUC-ROC of 0.927, an AUPRC of 0.657, an MCC score of 0.580 and Recall and Precision scores of 0.6077 and 0.6048, respectively. These results demonstrate the model's ability to sensitively and specifically identify open chromatin regions using only mechanistic epigenomic inputs.

Histone Marks Are Sufficient Predictors of Accessibility

To further identify the relative contribution of different feature types, two additional models were trained using only transcription factor signals or only histone modification signals. While both transcription factor-only and histone-only models used the same architecture and training procedure, their performance differed significantly. The transcription factor-only model showed reduced accuracy, with an MCC of 0.327 and an F1 score of just 0.319. In contrast, the histone-only model closely matched the full model in performance, achieving an MCC of 0.576 and an F1 score of 0.601, Fig. 5.

H3K4me1 Emerges as the Dominant Regulatory Feature

Multiple interpretability approaches were applied to identify which features the model relied on most for decision-making. Permutation importance revealed that H3K4me1, a histone mark associated with enhancers, consistently contributed the most to model performance, highlighted by the largest drop in MCC Score following its removal, Fig. 6. These findings support the role of H3K4me1 in marking regions of accessible chromatin, particularly enhancers, which are crucial regulators of gene expression in hormone-responsive cancers like breast cancer.

SHAP Analysis for False Positive Predictions

To assess feature contributions in incorrect predictions, SHAP summary analysis was performed on false positive cases for each input feature from the held-out chromosome 1 test set. Each dot in the plot represents a genomic bin, with SHAP values on the x-axis indicating the magnitude and direction of each feature's influence on the model's false positive predictions. The colour gradient (low (blue) to high (red)) reflects the original feature signal value. Features were ranked by their overall importance, as measured by the average absolute SHAP value, Fig. 7.

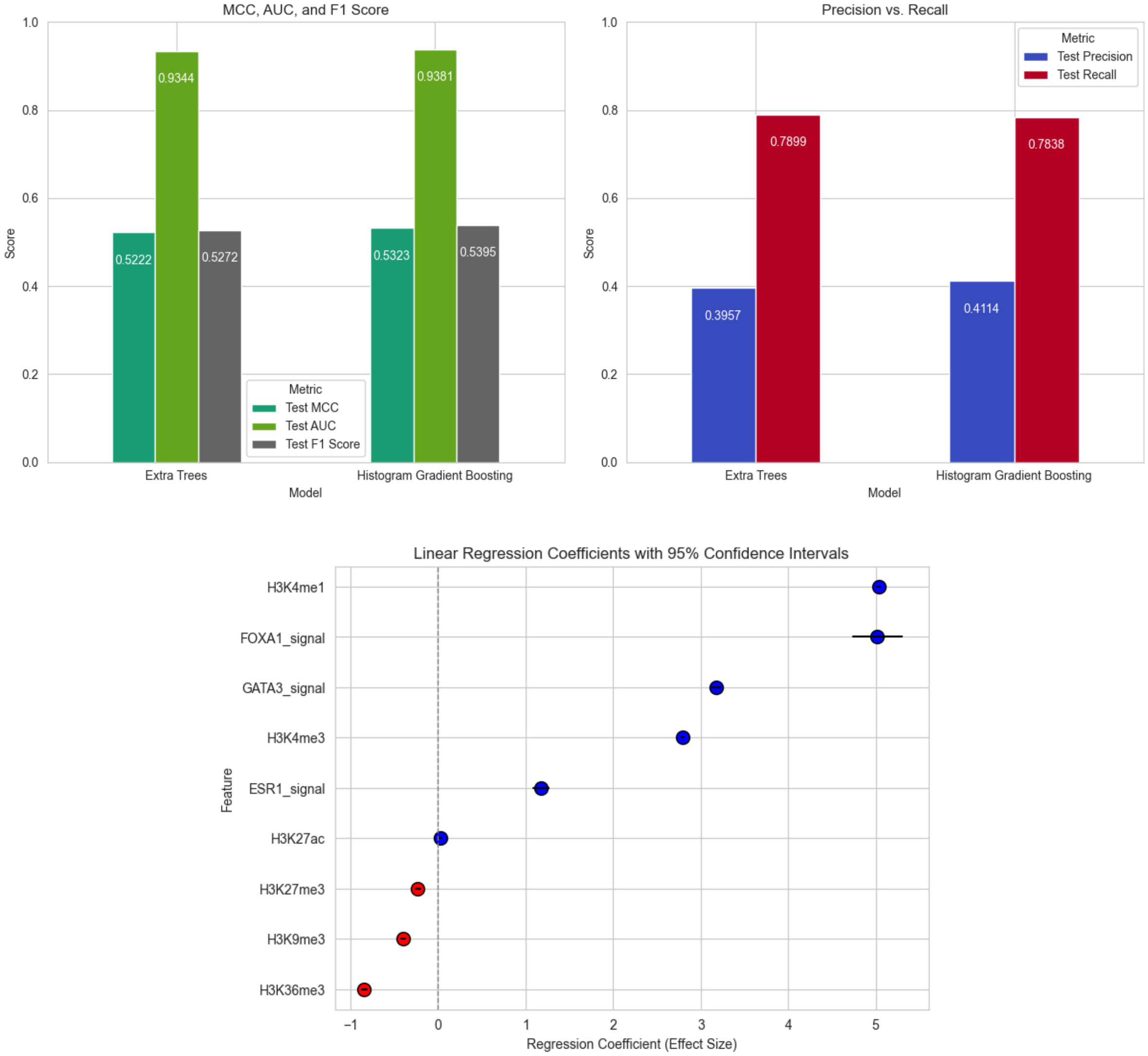


Figure 3: Performance Metrics and Feature Importance from Machine Learning and Linear Models on Held-out Chromosome 1. This figure summarises the performance of the Extra Trees and Histogram Gradient Boosting classifiers evaluated on chromosome 1 using a leave-one-chromosome-out (LOCO) validation strategy. Stratified K-Fold Cross-Validation (K=5) was applied during training to address class imbalance and ensure robust model evaluation. Both ensemble models achieved high AUC scores (>0.93) and strong recall (>0.78), with moderate MCC and F1 scores (>0.52), despite lower precision (<0.42). Additionally, an Ordinary Least Squares (OLS) linear regression model was used to estimate feature importance via regression coefficients. Feature contributions are visualised as effect sizes, with blue indicating positive association with open chromatin and red indicating association with closed chromatin. This analysis highlights the relative influence of each histone modification and transcription factor in predicting chromatin accessibility.

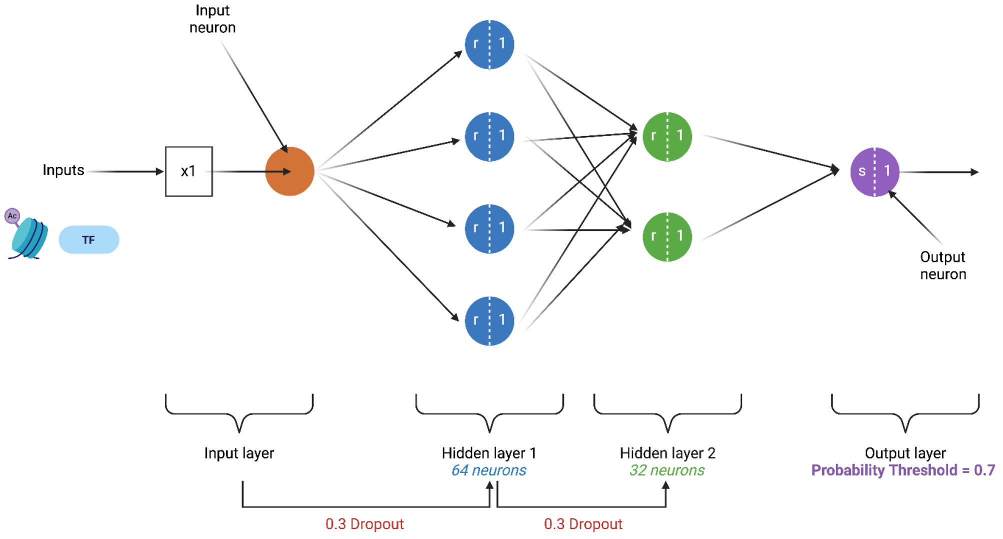


Figure 4: Neural network architecture for predicting chromatin accessibility. The model is a fully connected feedforward neural network. The input layer receives nine features per genomic bin (six histone marks and three transcription factor signals). This is followed by two hidden layers with 64 and 32 neurons, respectively, each using ReLU activation functions and dropout layers (dropout rate = 0.3) to reduce overfitting. The output layer consists of a single neuron with a sigmoid activation function, generating a probability score for chromatin accessibility. Hyperparamaters were determined by grid search hyperparameter tuning (see Supp. Table 2). The model was trained with binary cross-entropy loss and optimised using the Adam optimiser at a learning rate of 0.001.

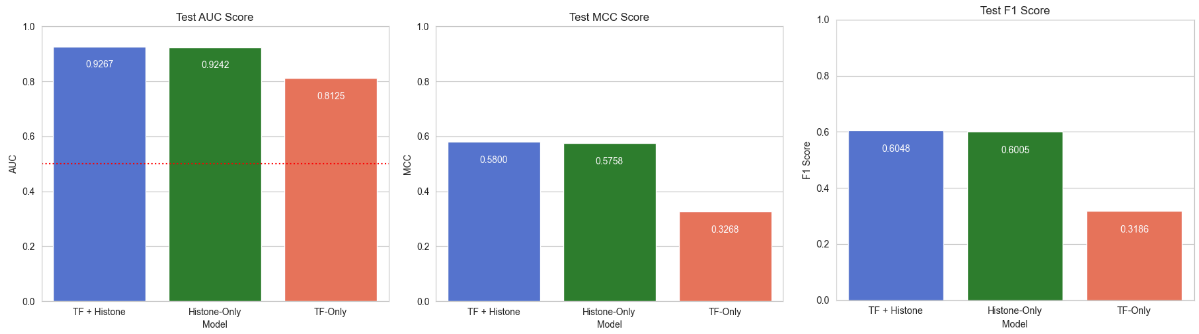


Figure 5: Comparative Performance of Histone-only and Transcription Factor-only Models in Predicting Chromatin Accessibility. This figure compares the performance of two models trained separately using either histone modification signals or transcription factor binding signals to predict chromatin accessibility. Both models used the same neural network architecture and training strategy as the full model. The histone-only model demonstrated strong predictive performance, closely matching the full feature set, with a Matthews Correlation Coefficient (MCC) of 0.576 and an F1 score of 0.601. In contrast, the transcription factor-only model showed reduced predictive power, with an MCC score of 0.327 and an F1 score of 0.319. These results highlight the greater standalone predictive value of histone modification signals in determining chromatin state.

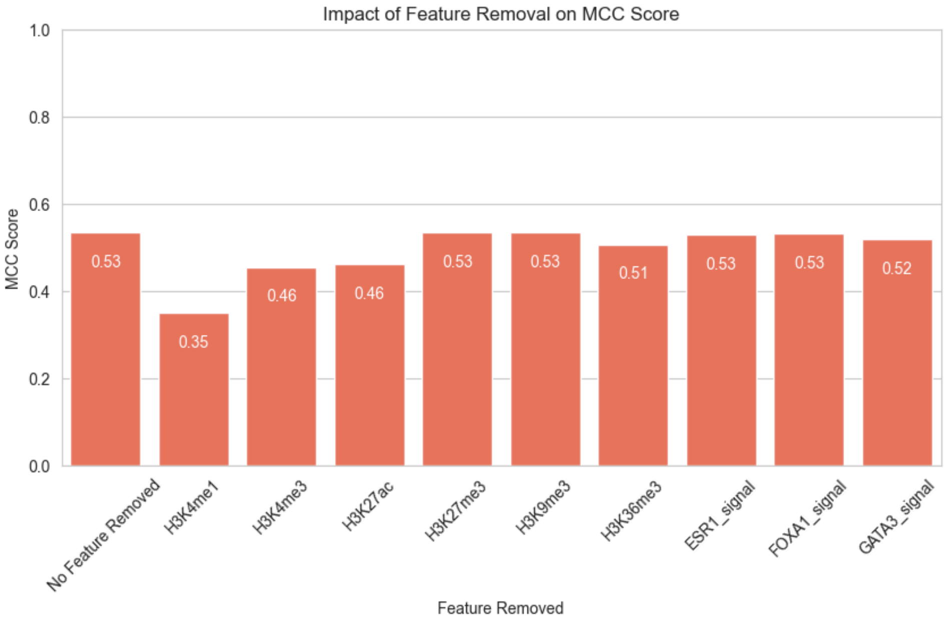


Figure 6: Permutation Importance of Epigenetic Features Based on MCC Score. This figure shows the permutation importance of each input feature, measured by the change in Matthews Correlation Coefficient (MCC) when the feature values are randomly shuffled. A greater drop in MCC indicates higher importance. H3K4me1 exhibited the largest decrease in MCC score, suggesting it played the most critical role in model predictions. Other features, including H3K27ac and H3K4me3, also showed moderate contributions. This analysis highlights the relative impact of individual histone modifications and transcription factor signals in predicting chromatin accessibility.

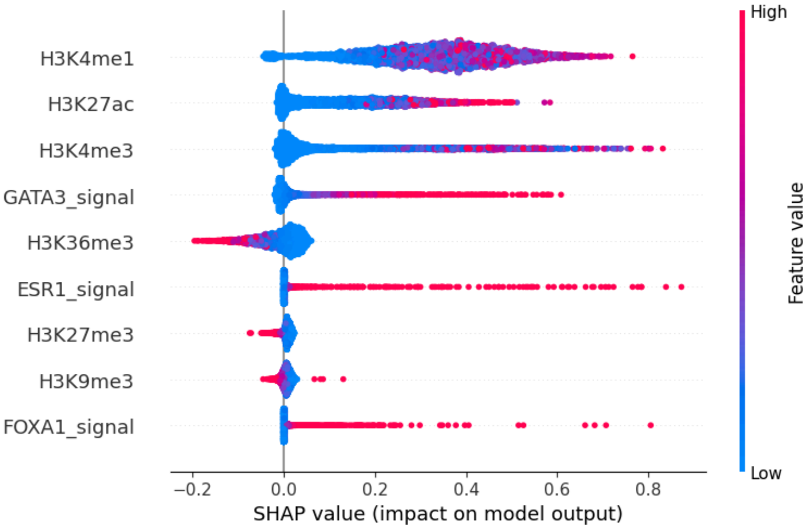


Figure 7: SHAP Summary Plot of Feature Contributions for False Positive Predictions. This plot displays SHAP values for all input features derived from false positive predictions on the held-out chromosome 1 test set. Each point represents a single 1 kb genomic bin where the model predicted open chromatin incorrectly. The x-axis indicates the SHAP value, which quantifies the impact of each feature on the model's output for that instance. The y-axis lists the features, including histone modification signals and transcription factor binding signals. Points are coloured by the original feature value, with blue representing low values and red representing high values. Features are ordered by their mean absolute SHAP value, indicating their relative importance in driving misclassification.

Discussion

H3K4me1 as a Distinctive Chromatin Accessibility Predictor in Breast Cancer

Early and recent studies have shown that chromatin accessibility can be accurately predicted using epigenetic features like histone modifications and transcription factor binding, often without relying on DNA sequence. A pioneering analysis by (Cui *et al.*, 2013) used support vector regression models on ENCODE data to quantify how histone marks and transcription factor binding correlate with chromatin “openness” (DNase hypersensitivity). They found that these features are highly predictive of accessibility and largely redundant, in fact, a small subset of histone marks and transcription factors could achieve very high predictive power. This foreshadowed later machine learning approaches indicating that a core group of histone modifications largely determine whether chromatin is accessible in a given cell context.

More recent work has leveraged deep learning for this task. (Zhao *et al.*, 2022) built a two-layer model integrating DNA sequence, transcription factor ChIP-seq binding, transcription factor motifs and histone modification ChIP-seq signals to predict ATAC-seq accessibility in the HepG2 and GM12878 human cell lines. Their results showed that DNA sequence alone has limited predictive power (AUC ≈0.6), whereas models using histone marks or transcription factor binding data each achieved high accuracy (AUC ≈0.8–0.84) in classifying open vs closed chromatin.

Notably, combining histone modifications and transcription factor features did not greatly improve accuracy over using either alone, indicating these features carry overlapping information. This was also seen in this study where the OLS regression analysis had a condition number of 12500, which might indicate that there is strong multicollinearity between features. This was further validated with a Spearman correlation matrix of all features, Supp. Figure 4. (Zhao *et al.*, 2022) identified five core histone modifications (H2AFZ, H3K4me2, H3K27ac, H3K9ac and H3K4me3) that explain most of the accessibility signals across both cell types. This aligns with the earlier finding by (Cui et al., 2013), that only a small number of chromatin features are needed for robust predictions. In other words, active histone marks (like H3K4me3 or H3K27ac at promoters) and the binding of key transcription factors tend to co-occur at open chromatin, making either data type a sufficient proxy for predicting accessibility.

However, unlike in this study, (Zhao *et al.*, 2022)'s results show H3K4me1 as a poor predictor of chromatin accessibility, achieving a relative importance score of less than ten, which is five times smaller than the strongest predictive features in HepG2 and GM12878 (H2A.Z1 and H3K4me2, respectively). In (Cui et al., 2013)'s study H3K4me1 received a prediction power score of ≈0.4 compared to the strongest feature, H3K4me2, with a score of ≈0.7. In this study, however, H3K4me1 was the most predictive feature with its removal causing the largest drop in MCC score (see Fig. 6). While these findings suggest that H3K4me1 may serve as a distinctive predictor of chromatin accessibility in breast cancer, definitive conclusions cannot be drawn without evaluating additional features such as H3K4me2 and H2A.Z1.

These studies illustrate a trend: by feeding epigenetic input features into machine learning models, one can achieve highly accurate predictions of ATAC-seq and DNase-seq peaks, often exceeding the accuracy of DNA-sequence-based models. Importantly, such models also provide biological interpretability, highlighting which histone marks are most influential in opening chromatin, thus, bridging predictive performance with mechanistic insight.

Therapeutic Targeting of H3K4me1 and Its Writers in Cancer

KMT2C and KMT2D

Histone H3 lysine 4 monomethylation (H3K4me1) is a chromatin mark typically enriched at gene enhancers and "poised" regulatory regions. It is deposited primarily by the methyltransferases MLL3 and MLL4 (also known as KMT2C and KMT2D) as part of the COMPASS family complexes. Dysregulation of these "writers" of H3K4me1 has been implicated in cancer, making them attractive targets for precision therapy. However, direct inhibitors of MLL3/4 are not yet available clinically (Yao *et al.*, 2024). Designing small-molecule inhibitors for the SET domain of KMT2D is an active area of research. One study reported virtual screening hits that bind the KMT2D catalytic domain, but with only micromolar affinity (Yu *et al.*, 2020).

To date, no specific MLL3/4 inhibitor has reached the market, reflecting both the complexity of these large enzymes and the fact that in many cancers they function as tumor suppressors rather than oncogenic drivers (Yao et al., 2024). For example, KMT2C (MLL3) is frequently mutated or lost in breast cancers and other tumors, and its loss is associated with poor prognosis and therapy resistance (Liu *et al.*, 2021; Batalini *et al.*, 2023). In such cases it would be ill-advised to inhibit MLL3 further as the loss of H3K4me1 may contribute to increased tumour progression. By contrast, KMT2D (MLL4) has recently been shown to act as a context-dependent oncogenic co-factor in certain settings. (Yao et al., 2024) also identified that in triple-negative breast cancer (TNBC), KMT2D is often overexpressed and drives enhancer activation. It was found to promote H3K4me1 deposition at enhancers of oncogenes like *MYC*, thereby facilitating tumour growth and metastasis. Additionally, the study also identified YBX1 as a novel "reader" protein that recognizes H3K4me1 marks deposited by KMT2D, hence, the KMT2D–H3K4me1–YBX1 axis was shown to epigenetically activate *MYC* and other pro-tumour genes in TNBC. Notably, high KMT2D and YBX1 levels correlated with poorer survival in breast cancer patients and disrupting this axis significantly impeded TNBC cell growth and metastasis in preclinical models. These findings suggest that inhibiting the H3K4me1 writer (KMT2D) or its reader (YBX1) could be a viable therapeutic strategy in aggressive, enhancer-driven breast cancers.

LSD1/KDM1A

Another way to target H3K4me1 levels is to inhibit the enzymes that remove this mark (i.e. histone demethylases). The LSD1/KDM1A enzyme specifically demethylates H3K4me1 and H3K4me2 (it can convert H3K4me1 to unmethylated lysine) (Fang, Liao and Yu, 2019). Elevated LSD1 in basal tumors correlates with downregulation of BRCA1 and was associated with increased sensitivity to PARP inhibitors (Nagasawa et al., 2015). This suggests LSD1-overexpressing cancers have a "BRCA-reducing" phenotype that could be therapeutically exploited. Several LSD1 inhibitors are now in clinical trials or advanced development, including ORY-1001 (iadademstat), GSK-2879552, IMG-7289, INCB059872, and CC-90011 (Fang, Liao and Yu, 2019). These compounds were initially tested in AML (acute myeloid leukemia) and SCLC (small-cell lung cancer), where LSD1 is critical for maintaining the undifferentiated, stem-like state of the cancer cells. By inhibiting LSD1's demethylase activity, such drugs increase H3K4me1/2 levels at differentiation genes, thereby reactivating suppressed gene programs and inhibiting tumor growth. Given LSD1's role in breast cancer progression and therapy resistance, there is interest in evaluating these inhibitors in breast cancer as well (Verigos *et al.*, 2019; Yang *et al.*, 2022).

LSD2/KDM1B

Another H3K4me1/2 demethylase, LSD2/KDM1B, is less studied but has been reported to demethylate enhancer marks and thereby silence genes like TP53; overactive LSD2 can promote cancer cell proliferation by epigenetically repressing p53 and other targets (Wang, Ma and Yu, 2023). While no LSD2-specific inhibitors are in clinics, the success of LSD1 programs suggests that targeting H3K4 methylation dynamics, either by inhibiting writers or erasers, is a promising approach in oncology.

Implications for Breast Cancer Prognosis and Model Interpretability

Epigenetic Breast Cancer Stratification for Precision Oncology

These advances carry important implications for breast cancer prognosis and for the interpretability of chromatin-based machine learning models. First, the status of H3K4me1-associated regulators is emerging as a biomarker of prognosis in breast cancer. For instance, high LSD1 expression (indicative of aggressive biology with low H3K4me1 on certain gene enhancers/promoters) is associated with significantly worse survival in basal-like breast cancer (Nagasawa et al., 2015). Conversely, loss-of-function mutations in MLL3 (which decrease H3K4me1) are linked to endocrine therapy resistance and poor outcome in ER-positive breast cancers (Liu et al., 2021; Batalini et al., 2023). As noted previously, overexpression of MLL4 and its H3K4me1 mark correlates with poor prognosis in TNBC (Yao et al., 2024). This opens discussion to investigate such relationships in oestrogen-amplified breast cancers.

These correlations suggest that measuring chromatin marks or the expression of their writers/erasers could refine risk stratification. For example, a high H3K4me1 enhancer signature might identify tumors reliant on active enhancers (prone to metastasis), whereas low H3K4me1 in a normally MLL3-dependent context might flag a more therapy-resistant tumor. Such knowledge can inform treatment decisions, for example, considering LSD1 inhibitor trials for patients with LSD1-overexpressing tumors, or PARP inhibitors for those with the LSD1–low BRCA1 axis.

Mechanistic Evaluation of Breast Cancer Opens Doors to New Research Avenues

Secondly, using epigenetic features in predictive models enhances model interpretability, which can yield biological insights. This neural network in this study and in (Zhao et al., 2022)'s not only predicts accessibility but also highlights which features are driving the prediction. If a model learns that H3K4me1 and H3K27ac are the top predictors of open chromatin in a breast cancer cell line, this reinforces the concept that active enhancer marks underlie the accessible chromatin landscape of that tumour. In the future this could lead to the discovery of a shared, fundamental epigenetic code for open chromatin. Identifying these key marks can direct researchers to the master regulators of the cell's epigenome, which could further influence targeted drug development, as discussed above.

For example, if H3K4me1 is consistently important in a model, one might investigate the upstream MLL3/4 complexes or associated co-factors (like menin or WDR5) in that context. The redundancy observed between transcription factor bindings and histone marks in the model is also informative; it suggests that open chromatin is a concerted state maintained by both transcription factors and histone modifications together. For therapy, this means clinicians could either target the transcription factor (perhaps with a small-molecule inhibitor or degrader) or the chromatin modifier (an epigenetic drug) to disrupt a given accessible region. In the TNBC example, one could aim at YBX1 (the transcription factor reader) or KMT2D (the writer) to collapse an oncogenic enhancer.

From a systems biology view, chromatin-feature-based models act as feature selectors, pointing to which epigenetic signals are most critical. This aids interpretability and cross validation with experiments. A model's top features can be validated in the lab, such as by CRISPR-editing a histone modifier or treating cells with an epigenetic inhibitor to see if chromatin accessibility and gene expression change as predicted.

In summary, a growing body of work demonstrates that chromatin accessibility can be accurately predicted from epigenetic profiles. These models have identified key histone marks (like

H3K4me1) as fundamental determinants of open chromatin, providing mechanistic insights that complement sequence-based predictions.

In parallel, pharmacological targeting of the H3K4me1 pathway is being pursued in oncology. While direct MLL3/4 inhibitors remain under development, inhibitors of associated factors (menin, WDR5) and demethylases (LSD1) are showing promise in clinical trials. The convergence of these research avenues suggests an exciting precision oncology paradigm. In the future, clinicians could use chromatin-based models to identify tumor-specific epigenetic vulnerabilities, and then apply epigenetic drugs to selectively target the aberrant chromatin states that drive a given patient's cancer. In breast cancer, this means the prospect of tailoring treatments that intervene in enhancer activation programs or chromatin modifications (such as aberrant H3K4me1 patterns), potentially improving outcomes for subtypes with poor prognosis and offering new strategies to combat therapy resistance.

The Implementation of RNA-Seq to Enable Multi-omic Breast Cancer Analyses

Incorporating RNA-seq data into chromatin accessibility prediction frameworks offers a critical next step in linking regulatory potential with transcriptional output. While histone modifications and transcription factor binding profiles provide a mechanistic basis for predicting open chromatin, RNA-seq captures the functional consequence of these regulatory events. By integrating transcriptomic data, future models could go beyond binary chromatin states to predict which accessible regions are actively contributing to transcription. For example, enhancers marked by H3K4me1 that also correlate with upregulated nearby genes would offer strong evidence of functional activation, allowing models to better prioritise biologically meaningful regulatory elements. This could also assist in distinguishing poised enhancers from active ones in breast cancer subtypes, refining the interpretability of models and improving precision in identifying therapeutic targets.

Additionally, RNA-seq integration would enable the exploration of enhancer–promoter interactions and regulatory network dynamics specific to cancer phenotypes. This study highlighted H3K4me1 as a key predictive marker, but does not capture downstream expression changes that drive tumour behaviour. By connecting accessible chromatin regions to their gene targets via co-expression or enhancer–promoter proximity frameworks by using Hi-C-informed assignments, one could construct interpretable gene regulatory networks in breast cancer. This could also help identify transcriptional programmes driven by specific histone-modifying enzymes (such as KMT2D or LSD1), opening avenues for stratifying patients based on expression signatures linked to epigenetic vulnerabilities. In doing so, RNA-seq–guided models could support the identification of synthetic lethal interactions or epigenetic–transcriptional dependencies that are targetable in precision oncology.

Predicting Future States of Chromatin Accessibility - Chromatin Velocity

While this model's prediction of chromatin accessibility from histone marks and transcription factor binding is significantly biologically relevant, it largely relies on static snapshots of the epigenome. However, tumour evolution is inherently dynamic, and recent innovations in RNA velocity, which estimates the future transcriptional state of individual cells by analysing the ratios of spliced and unspliced RNA, offers an untapped opportunity to integrate temporal dynamics into chromatin modelling (Tang *et al.*, 2023). The study leveraged RNA velocity within their comboSC pipeline to inform drug prioritisation by inferring the likely trajectories of immune and tumour cells, enabling personalised therapeutic optimisation at single-cell resolution.

Extending this framework, an exploratory direction would involve adapting RNA velocity to develop a concept of "chromatin velocity", an extended computational model that not only infers current transcriptomic dynamics but also anticipates future transcriptional states based on chromatin accessibility or histone modification patterns. In breast cancer, this could identify enhancer reprogramming events before they manifest, potentially forecasting epigenomic vulnerabilities in aggressive or therapy-resistant subpopulations. Such predictions could be integrated with histone modification-based models to dynamically prioritise epigenetic targets like H3K4me1-modifying enzymes, thereby enabling pre-emptive therapeutic interventions before phenotypic transitions occur, advancing the goals of precision oncology by targeting chromatin state transitions unique to malignant cells.

Conclusion

This study demonstrates the feasibility and biological relevance of predicting chromatin accessibility using only histone modification and transcription factor binding signals in the MCF-7 breast cancer cell line. By developing and benchmarking a mechanistic neural network against baseline models, it was shown that histone features, particularly H3K4me1, are sufficient for accurate chromatin state prediction, with performance comparable or superior to models that incorporate more expansive inputs. Importantly, the neural network model outperformed existing frameworks despite using fewer features and no DNA sequence data, thereby offering a scalable, interpretable alternative for inferring regulatory landscapes from ChIP-seq profiles alone.

These findings support a precision oncology paradigm in which chromatin-based models are used not only for prediction, but for identifying epigenetic dependencies specific to malignant states. The prominence of H3K4me1 as a predictive feature, contrasted with its lower importance in other cell types, suggests that enhancer priming mechanisms may play a unique role in oestrogen receptor-positive breast cancer. Furthermore, mechanistic insights derived from this modelling approach provide a foundation for therapeutic intervention.

Future directions include expanding classification to additional chromatin states (e.g. poised or permissive regions), integrating transcriptomic data from RNA-seq to link accessibility with gene expression, and validating findings in other breast cancer cell lines. An exploratory avenue also lies in adapting RNA velocity frameworks to derive “chromatin velocity” measures, predicting future accessibility dynamics in response to tumour evolution. By moving towards temporally-aware, multi-omic modelling, there is substantial potential to pre-empt regulatory shifts that underlie resistance and metastasis. Ultimately, this work highlights how AI models trained on interpretable epigenetic features can advance our understanding of chromatin regulation in cancer, refine prognostic tools, and lay the groundwork for selectively targeting epigenomic vulnerabilities in breast cancer.

Bibliography

Adomas, A.B. *et al.* (2014) 'Breast tumor specific mutation in GATA3 affects physiological mechanisms regulating transcription factor turnover', *BMC Cancer*, 14(1), p. 278. Available at: https://doi.org/10.1186/1471-2407-14-278.

Applications of ENCODE data to systematic analyses via data integration - ScienceDirect (2018). Available at: https://www.sciencedirect.com/science/article/abs/pii/S2452310018300593?via%3Dihub (Accessed: 9 March 2025).

Augello, M.A., Hickey, T.E. and Knudsen, K.E. (2011) 'FOXA1: master of steroid receptor function in cancer', *The EMBO Journal*, 30(19), pp. 3885–3894. Available at: https://doi.org/10.1038/emboj.2011.340.

Batalini, F. *et al.* (2023) 'Association of KMT2C loss-of-function mutations in circulating tumor DNA and prolonged response to the combination of PARPi and PI3Ki.', *Journal of Clinical Oncology*, 41(16_suppl), pp. e13003–e13003. Available at: https://doi.org/10.1200/JCO.2023.41.16_suppl.e13003.

Beaver, J.A. *et al.* (2013) 'PIK3CA and AKT1 mutations have distinct effects on sensitivity to targeted pathway inhibitors in an isogenic luminal breast cancer model system', *Clinical cancer research : an official journal of the American Association for Cancer Research*, 19(19), pp. 5413–5422. Available at: https://doi.org/10.1158/1078-0432.CCR-13-0884.

Buenrostro, J.D. *et al.* (2015) 'ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide', *Current Protocols in Molecular Biology*, 109(1), p. 21.29.1-21.29.9. Available at: https://doi.org/10.1002/0471142727.mb2129s109.

Chicco, D. and Jurman, G. (2020) 'The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation', *BMC Genomics*, 21(1), p. 6. Available at: https://doi.org/10.1186/s12864-019-6413-7.

Creyghton, M.P. *et al.* (2010) 'Histone H3K27ac separates active from poised enhancers and predicts developmental state', *Proceedings of the National Academy of Sciences*, 107(50), pp. 21931–21936. Available at: https://doi.org/10.1073/pnas.1016071107.

Cui, P. *et al.* (2013) 'A Quantitative Analysis of the Impact on Chromatin Accessibility by Histone Modifications and Binding of Transcription Factors in DNase I Hypersensitive Sites', *BioMed Research International*, 2013, p. 914971. Available at: https://doi.org/10.1155/2013/914971.

Fang, Y., Liao, G. and Yu, B. (2019) 'LSD1/KDM1A inhibitors in clinical trials: advances and prospects', *Journal of Hematology & Oncology*, 12(1), p. 129. Available at: https://doi.org/10.1186/s13045-019-0811-9.

Hampton, O.A. *et al.* (2009) 'A sequence-level map of chromosomal breakpoints in the MCF-7 breast cancer cell line yields insights into the evolution of a cancer genome', *Genome Research*, 19(2), pp. 167–177. Available at: https://doi.org/10.1101/gr.080259.108.

Hanahan, D. (2022) *Hallmarks of Cancer: New Dimensions | Cancer Discovery | American Association for Cancer Research*. Available at: https://aacrjournals.org/cancerdiscovery/article/12/1/31/675608/Hallmarks-of-Cancer-New-DimensionsHallmarks-of (Accessed: 26 March 2025).

Hanahan, D. and Weinberg, R.A. (2011) 'Hallmarks of Cancer: The Next Generation', *Cell*, 144(5), pp. 646–674. Available at: https://doi.org/10.1016/j.cell.2011.02.013.

Heinz, S. *et al.* (2010) 'Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities', *Molecular Cell*, 38(4), pp. 576–589. Available at: https://doi.org/10.1016/j.molcel.2010.05.004.

Holliday, D.L. and Speirs, V. (2011) 'Choosing the right cell line for breast cancer research', *Breast Cancer Research*, 13(4), p. 215. Available at: https://doi.org/10.1186/bcr2889.

Hua, H. *et al.* (2018) 'Mechanisms for estrogen receptor expression in human cancer', *Experimental Hematology & Oncology*, 7(1), p. 24. Available at: https://doi.org/10.1186/s40164-018-0116-7.

Jin, W. *et al.* (2020) 'Effect of the key histone modifications on the expression of genes related to breast cancer', *Genomics*, 112(1), pp. 853–858. Available at: https://doi.org/10.1016/j.ygeno.2019.05.026.

Kingma, D.P. and Ba, J. (2017) 'Adam: A Method for Stochastic Optimization'. arXiv. Available at: https://doi.org/10.48550/arXiv.1412.6980.

Klemm, S.L., Shipony, Z. and Greenleaf, W.J. (2019) 'Chromatin accessibility and the regulatory epigenome', *Nature Reviews Genetics*, 20(4), pp. 207–220. Available at: https://doi.org/10.1038/s41576-018-0089-8.

Liang, J. *et al.* (2018) 'CDKN2A inhibits formation of homotypic cell-in-cell structures', *Oncogenesis*, 7(6), pp. 1–8. Available at: https://doi.org/10.1038/s41389-018-0056-4.

Liu, X. *et al.* (2021) 'KMT2C is a potential biomarker of prognosis and chemotherapy sensitivity in breast cancer', *Breast Cancer Research and Treatment*, 189(2), pp. 347–361. Available at: https://doi.org/10.1007/s10549-021-06325-1.

Locke, W.J. *et al.* (2015) 'Coordinated epigenetic remodelling of transcriptional networks occurs during early breast carcinogenesis', *Clinical Epigenetics*, 7(1), p. 52. Available at: https://doi.org/10.1186/s13148-015-0086-0.

Mansisidor, A.R. and and Risca, V.I. (2022) 'Chromatin accessibility: methods, mechanisms, and biological insights', *Nucleus*, 13(1), pp. 238–278. Available at:

https://doi.org/10.1080/19491034.2022.2143106.

Mbatchou, J. *et al.* (2021) ‘Computationally efficient whole-genome regression for quantitative and binary traits’, *Nature Genetics*, 53(7), pp. 1097–1103. Available at: https://doi.org/10.1038/s41588-021-00870-7.

McCarthy, M.T. and O’Callaghan, C.A. (2014) ‘PeaKDEck: a kernel density estimator-based peak calling program for DNaseI-seq data’, *Bioinformatics*, 30(9), pp. 1302–1304. Available at: https://doi.org/10.1093/bioinformatics/btt774.

Nagasawa, S. *et al.* (2015) ‘LSD1 Overexpression Is Associated with Poor Prognosis in Basal-Like Breast Cancer, and Sensitivity to PARP Inhibition’, *PLOS ONE*, 10(2), p. e0118002. Available at: https://doi.org/10.1371/journal.pone.0118002.

Newsham, I. *et al.* (2024) ‘Early detection and diagnosis of cancer with interpretable machine learning to uncover cancer-specific DNA methylation patterns’, *Biology Methods & Protocols*, 9(1), p. bpae028. Available at: https://doi.org/10.1093/biomethods/bpae028.

Padeken, J., Methot, S.P. and Gasser, S.M. (2022) ‘Establishment of H3K9-methylated heterochromatin and its functions in tissue differentiation and maintenance’, *Nature Reviews Molecular Cell Biology*, 23(9), pp. 623–640. Available at: https://doi.org/10.1038/s41580-022-00483-w.

Park, P.J. (2009) ‘ChIP–seq: advantages and challenges of a maturing technology’, *Nature Reviews Genetics*, 10(10), pp. 669–680. Available at: https://doi.org/10.1038/nrg2641.

Pedregosa, F. *et al.* (2011) ‘Scikit-learn: Machine Learning in Python’, *J. Mach. Learn. Res.*, 12(null), pp. 2825–2830.

Rauschert, S. *et al.* (2020) ‘Machine learning and clinical epigenetics: a review of challenges for diagnosis and classification’, *Clinical Epigenetics*, 12(1), p. 51. Available at: https://doi.org/10.1186/s13148-020-00842-4.

Ritchie, H., Spooner, F. and Roser, M. (2018) ‘Causes of death’, *Our World in Data* [Preprint]. Available at: https://ourworldindata.org/causes-of-death (Accessed: 9 October 2022).

Tang, C. *et al.* (2023) ‘Personalized tumor combination therapy optimization using the single-cell transcriptome’, *Genome Medicine*, 15(1), p. 105. Available at: https://doi.org/10.1186/s13073-023-01256-6.

Tarbell, E.D. and Liu, T. (2019) ‘HMMRATAC: a Hidden Markov ModeleR for ATAC-seq’, *Nucleic Acids Research*, 47(16), p. e91. Available at: https://doi.org/10.1093/nar/gkz533.

Vakoc, C.R. *et al.* (2006) ‘Profile of Histone Lysine Methylation across Transcribed Mammalian Chromatin’, *Molecular and Cellular Biology* [Preprint]. Available at: https://doi.org/10.1128/MCB.01529-06.

Verigos, J. *et al.* (2019) ‘The Histone Demethylase LSD1/KDM1A Mediates Chemoresistance in Breast Cancer via Regulation of a Stem Cell Program’, *Cancers*, 11(10), p. 1585. Available at: https://doi.org/10.3390/cancers11101585.

Wang, N., Ma, T. and Yu, B. (2023) ‘Targeting epigenetic regulators to overcome drug resistance in cancers’, *Signal Transduction and Targeted Therapy*, 8(1), pp. 1–24. Available at: https://doi.org/10.1038/s41392-023-01341-7.

Weidemüller, P. *et al.* (2021) ‘Transcription factors: Bridge between cell signaling and gene regulation’, *PROTEOMICS*, 21(23–24), p. 2000034. Available at: https://doi.org/10.1002/pmic.202000034.

Welsh, J. (2013) ‘Chapter 40 - Animal Models for Studying Prevention and Treatment of Breast Cancer’, in P.M. Conn (ed.) *Animal Models for the Study of Human Disease*. Boston: Academic Press, pp. 997–1018. Available at: https://doi.org/10.1016/B978-0-12-415894-8.00040-3.

Yan, F. *et al.* (2020) ‘From reads to insight: a hitchhiker’s guide to ATAC-seq data analysis’, *Genome Biology*, 21(1), p. 22. Available at: https://doi.org/10.1186/s13059-020-1929-3.

Yang, G.-J. *et al.* (2022) ‘A state-of-the-art review on LSD1 and its inhibitors in breast cancer: Molecular mechanisms and therapeutic significance’, *Frontiers in Pharmacology*, 13. Available at: https://doi.org/10.3389/fphar.2022.989575.

Yao, B. *et al.* (2024) ‘KMT2D‐mediated H3K4me1 recruits YBX1 to facilitate triple‐negative breast cancer progression through epigenetic activation of c‐Myc’, *Clinical and Translational Medicine*, 14(7), p. e1753. Available at: https://doi.org/10.1002/ctm2.1753.

Young, M.D. *et al.* (2011) ‘ChIP-seq analysis reveals distinct H3K27me3 profiles that correlate with transcriptional activity’, *Nucleic Acids Research*, 39(17), pp. 7415–7427. Available at: https://doi.org/10.1093/nar/gkr416.

Yu, Q. *et al.* (2020) ‘Small molecule inhibitors of the prostate cancer target KMT2D’, *Biochemical and Biophysical Research Communications*, 533(3), pp. 540–547. Available at: https://doi.org/10.1016/j.bbrc.2020.09.004.

Zhang, Y. *et al.* (2008) ‘Model-based Analysis of ChIP-Seq (MACS)’, *Genome Biology*, 9(9), p. R137. Available at: https://doi.org/10.1186/gb-2008-9-9-r137.

Zhao, Y. *et al.* (2022) ‘Computational modeling of chromatin accessibility identified important epigenomic regulators’, *BMC Genomics*, 23(1), p. 19. Available at: https://doi.org/10.1186/s12864-021-08234-5.

Appendices

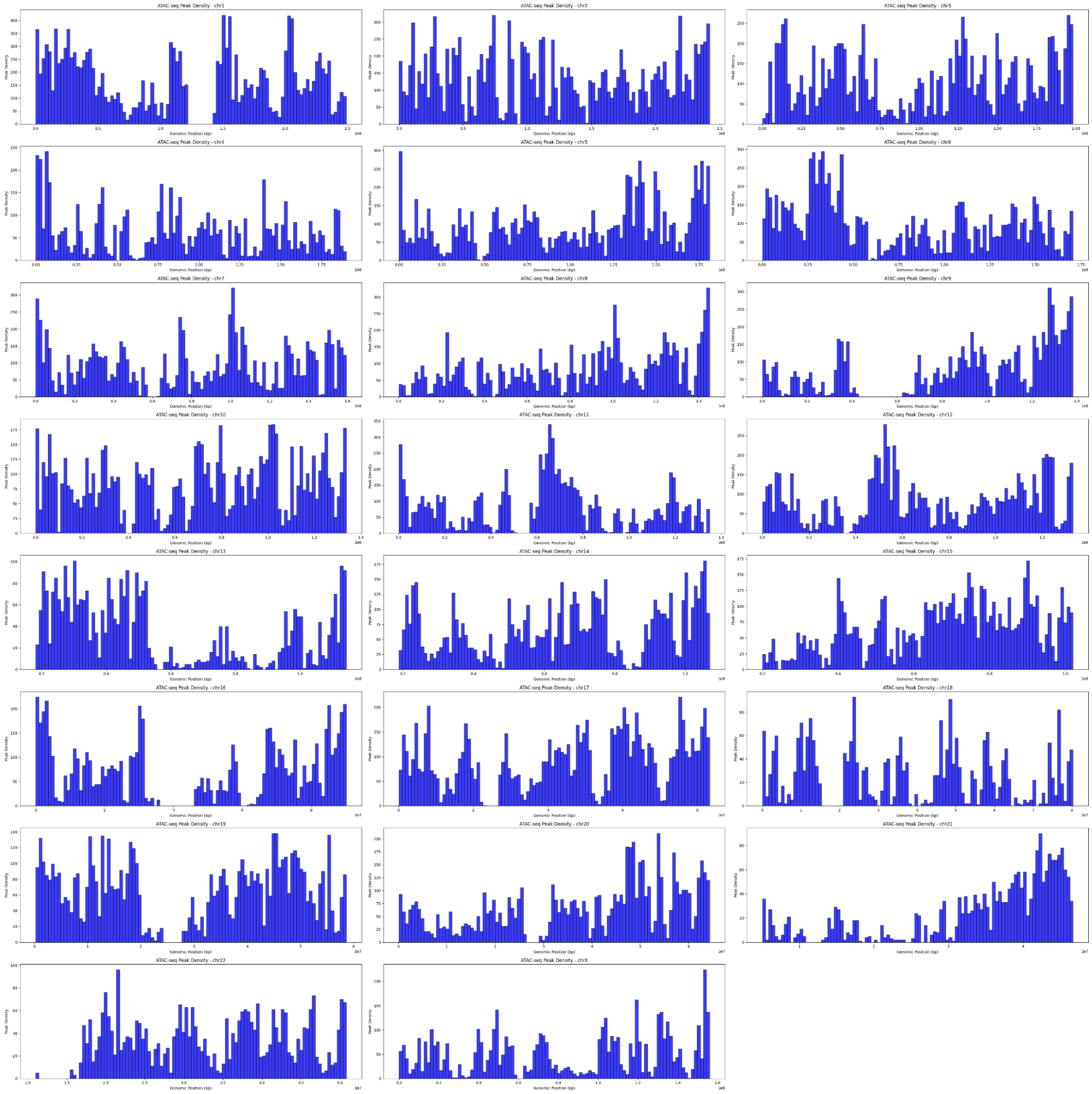
Supplementary Table 1

Dataset summary. The ENCODE accession number, experimental target, file format, replicates, genome assembly and data processing type are provided in the table below.

Supplementary Table 1: Summary of ATAC-Seq and Histone ChIP-Seq Datasets Used in This Study

Experimental Target	File Accession ID	Histone Mark / Assay Type	File Format	Replicates	Genome Assembly	Data Processing
ATAC-Seq	ENCFF821OEF	Chromatin Accessibility	BED (gzipped)	Isogenic Rep 1, 2	GRCh38	Pseudoreplicated Peaks
ChIP-Seq	ENCFF025QZH	H3K27me3	bigWig	Isogenic Rep 1, 2	GRCh38	Signal p-value output
ChIP-Seq	ENCFF372GMC	H3K4me1	bigWig	Isogenic Rep 1, 2	GRCh38	Signal p-value output
ChIP-Seq	ENCFF138YNG	H3K27ac	bigWig	Isogenic Rep 1, 2	GRCh38	Signal p-value output
ChIP-Seq	ENCFF163MXP	H3K4me3	bigWig	Isogenic Rep 1, 2	GRCh38	Signal p-value output
ChIP-Seq	ENCFF910BRP	H3K36me3	bigWig	Isogenic Rep 1, 2	GRCh38	Signal p-value output
ChIP-Seq	ENCFF481DZL	H3K9me3	bigWig	Isogenic Rep 1, 2	GRCh38	Signal p-value output
Chromatin State	ENCFF506GEX	ChromHMM Annotations	BED (gzipped)	Isogenic Rep 1, 2	GRCh38	Semi-automated genome annotation

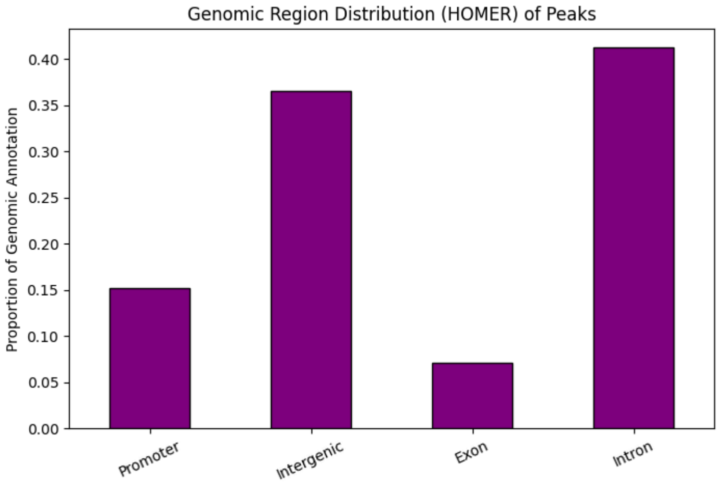
Supplementary Figure 1



Supplementary Figure 1: ATAC-Seq Peak Density Across All Chromosomes. Histograms displaying the distribution of ATAC-Seq peak density across all chromosomes (chr1–chr22 and chrX) in the MCF-7 cell line. Peaks were identified using a pseudoreplicated peak-calling approach, ensuring reproducibility across isogenic replicates. The x-axis represents the genomic position in base pairs, while the y-axis represents the frequency of peaks within 100 evenly spaced genomic bins. Variability in peak density across chromosomes reflects differences in chromatin accessibility, with certain regions exhibiting higher regulatory activity. Peaks on chrX provide insights into the regulation of sex chromosome-associated genes in this breast cancer-derived cell line.

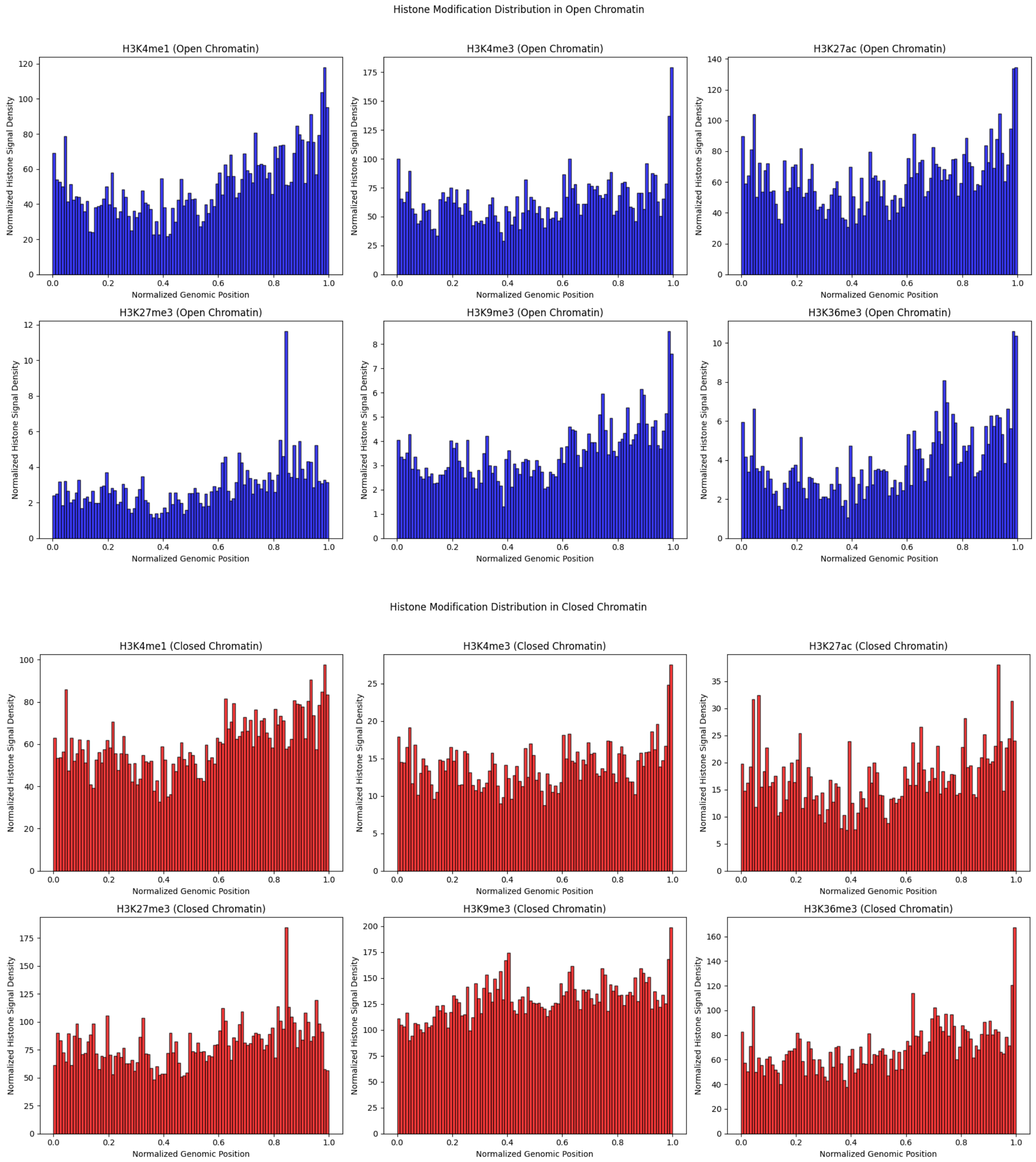
Supplementary Figure 2

HOMER is a widely used software suite designed for the identification and annotation of regulatory elements in genomic data, including promoters, enhancers, and other functional regions (Heinz *et al.*, 2010). Genomic bins from the chromatin accessibility dataset were mapped to functional genomic annotations based on HOMER's precomputed reference databases. Each bin was classified into one of four genomic categories: Promoter, Intergenic, Exon, or Intron, based on overlap with annotated peaks. A bin was assigned to a category if its genomic coordinates overlapped with a corresponding HOMER annotation.



Supplementary Figure 2: Genomic Region Distribution (HOMER) of Peaks. Bar plot displaying the proportion of ATAC-seq peaks annotated in different genomic regions using HOMER. The majority of peaks are located in introns and intergenic regions, with a smaller proportion in promoters and exons, highlighting the widespread nature of chromatin accessibility beyond promoter regions.

Supplementary Figure 3



Supplementary Figure 3: Genome-wide distribution of histone modifications in open and closed chromatin. Histograms display the normalised signal intensity of six histone modifications across the genome, separated by chromatin accessibility state. (Top) Histone signal densities in open chromatin (blue), showing enrichment of active marks such as H3K4me3 and H3K27ac. (Bottom) Histone signal densities in closed chromatin (red), where repressive marks such as H3K9me3 are more prominent. The x-axis represents the normalised genomic position, adjusting for chromosome length, while the y-axis denotes the normalised histone signal density. These plots highlight distinct histone modification patterns associated with chromatin accessibility states.

Supplementary Table 2

Grid Search Results Across Class Weights, Thresholds and Downsampling Ratios. The findings support the utility of histone modifications and transcription factor binding signals as predictive features for chromatin accessibility modeling. To ensure unbiased evaluation, chromosome 1 was excluded from the analysis, as it was reserved for final model validation, preventing data leakage and ensuring that statistical comparisons were not influenced by the test set.

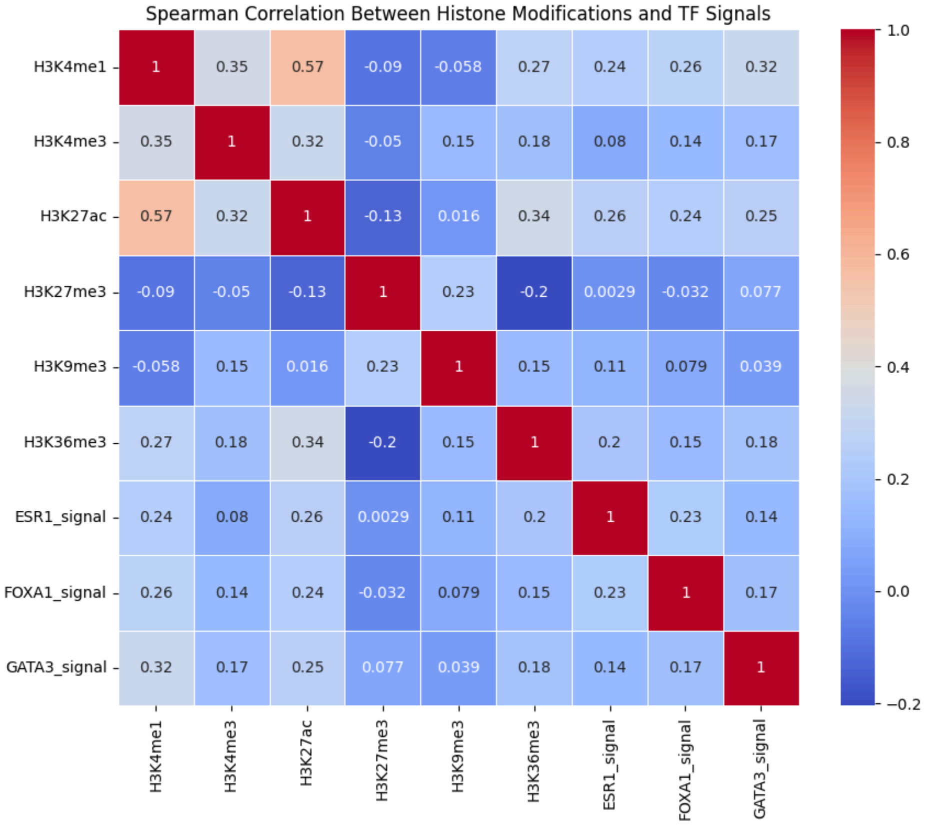
Supplementary Table 2: Grid Search Hyperparameter Tuning of LOCO Validation Model

Class Weights	Threshold	Downsampling of Majority Class (0)	AUC	AUPRC	Precision	Recall	MCC
{0: 1, 1: 2}	0.5	15	0.927101	0.663065	0.651528	0.571450	0.587534
{0: 1, 1: 2}	0.7	10	0.926923	0.659017	0.739667	0.490572	0.583029
{0: 1, 1: 2}	0.7	5	0.926472	0.654996	0.604997	0.603256	0.579429
{0: 1, 1: 16.00}	0.9	10	0.925037	0.652168	0.624370	0.582052	0.578955
{0: 1, 1: 16.00}	0.9	15	0.923326	0.646632	0.639239	0.564786	0.577575
{0: 1, 1: 2}	0.5	10	0.926072	0.656677	0.570452	0.631882	0.574069
{0: 1, 1: 2}	0.7	15	0.927039	0.663149	0.810857	0.427565	0.571843
{0: 1, 1: 2}	0.9	5	0.926639	0.656444	0.808316	0.415146	0.562221
{0: 1, 1: 2}	0.3	15	0.927548	0.660471	0.508380	0.682242	0.559354
{0: 1, 1: 16.00}	0.9	5	0.924519	0.648529	0.448371	0.716850	0.533427
{0: 1, 1: 2}	0.3	10	0.926711	0.661218	0.436271	0.737448	0.533039
{0: 1, 1: 2}	0.5	5	0.925032	0.651616	0.426799	0.737372	0.526013

{0: 1, 1: 16.00}	0.7	15	0.922552	0.642624	0.401362	0.749867	0.511550
{0: 1, 1: 2}	0.9	10	0.929306	0.667085	0.924780	0.278379	0.494132
{0: 1, 1: 2}	0.9	15	0.922231	0.644067	0.898820	0.282544	0.490011
{0: 1, 1: 16.00}	0.7	10	0.923884	0.650270	0.336945	0.803029	0.477306
{0: 1, 1: 2}	0.3	5	0.926406	0.654101	0.310036	0.825369	0.460066
{0: 1, 1: 16.00}	0.5	15	0.926075	0.653978	0.274070	0.848921	0.431814
{0: 1, 1: 16.00}	0.7	5	0.925609	0.651542	0.261812	0.856872	0.421261
{0: 1, 1: 16.00}	0.5	10	0.926664	0.656907	0.234348	0.882393	0.397914
{0: 1, 1: 16.00}	0.3	15	0.925068	0.652029	0.217604	0.889284	0.379644
{0: 1, 1: 16.00}	0.5	5	0.923603	0.644035	0.215452	0.888451	0.376728
{0: 1, 1: 16.00}	0.3	10	0.925818	0.650988	0.191236	0.906172	0.349859
{0: 1, 1: 16.00}	0.3	5	0.922798	0.641753	0.173266	0.914881	0.326493

Supplementary Figure 4

Spearman Correlation Matrix of Histone Modifications and Transcription Factor Binding Signals.



Supplementary Figure 4: Spearman Correlation Matrix of Histone Modifications and Transcription Factor Binding Signals. This heatmap shows the pairwise Spearman correlation coefficients between six histone modification signals and three transcription factor binding signals used as input features for chromatin accessibility prediction. Moderate positive correlations were observed between H3K4me1, H3K4me3, and H3K27ac, three marks associated with active regulatory elements. Transcription factors ESR1, FOXA1, and GATA3 also showed modest correlation with these active histone marks, particularly H3K4me1 and H3K27ac. These patterns indicate potential redundancy in predictive features, aligning with prior observations that combining histone marks and transcription factor signals yields minimal performance gain over using histone modifications alone (see Fig. 5). This redundancy was further evidenced by a high condition number (12500) in the Ordinary Least Squares (OLS) regression analysis, further compounding multicollinearity among features. Such overlap supports the notion that a core subset of chromatin features is sufficient to capture accessibility patterns, consistent with findings from (Cui et al., 2013) and (Zhao et al., 2022).