Characterization of Histone Modification Using Machine Learning

by
Dylan Setiawan

Mentor Names:
Dr. Xuehua Zhong, PhD
Carl H. Simmons

Department of Biology, Washington University in St. Louis, St. Louis, MO, 63130, USA

Spring 2025

Abstract

Understanding the functional impact of histone modifications on gene expression remains a major challenge in epigenetics due to the vast number of possible modification combinations and the confounding effects of proteins that regulate their deposition or removal. Traditional knockout experiments can struggle to isolate the specific effects of individual modifications in addition to its costly and tedious nature, making quantification difficult. To address this, we developed a pipeline that characterizes histone modification effects in *Arabidopsis* using explainable machine learning. By leveraging feature importance and effect analysis, the pipeline requires only ChIP-seq data to quantify the influence of a given histone modification on gene expression. This new pipeline was then tested on a previously uncharacterized monomethylation of lysine 23 on histone 3 (H3K23me1) to quantify its effects on gene expression and compare its impact to other histone modifications.

1. Introduction

Epigenetics is the study of heritable changes in gene expression that occur without alterations to the underlying DNA sequence (1). These changes are mediated by chemical modifications to DNA and histones which regulate chromatin accessibility and transcription (2). Histone modifications such as methylation, acetylation, phosphorylation, and ubiquitination, serve as key regulators of gene activity by altering chromatin structure and recruiting specific protein complexes (3).

In plants, epigenetic regulation is essential for controlling gene expression during development and in response to environmental cues. Histone modifications, in particular, guide cell differentiation (4), control flowering time (5), and regulate responses to stresses such as drought (6) or soil salinity (7). Unlike genetic mutations, epigenetic modifications are dynamic and reversible, allowing adaptations to changing conditions.

Despite the advances in epigenetics, there are still major challenges in understanding the functional impact of histone modifications. 1) There are a large number of histone modifications and most of their functions remain unknown. Because of the diversity of these modifications, along with their potential interactions and placements, it makes assigning a clear functional role to each modification difficult (3, 8). Many histone marks still remain uncharacterized and their biological significance unknown. 2) Characterizing histone modifications can be expensive and time consuming. Traditionally histone modification functions are studied using a combination of ChIP sequencing and knockout or overexpression experiments to determine its effects on overall gene expression (8). However, these approaches are time-consuming and expensive. Generating mutants for each modification or associated protein is impractical. 3) There are confounding effects that limit traditional knockout experiments. Many proteins (writers and erasers) may be responsible for depositing or removing multiple modifications meaning that removing a single enzyme can affect multiple histone marks

(9). This confounding issue makes it difficult to attribute the observed effects to a specific modification often leading to ambiguous results.

To overcome these challenges, this study proposes a computational approach that uses sequencing data and machine learning to characterize histone modifications. Because of recent advancements in high-throughput sequencing, vast amounts of epigenomic data are published online. This offers an opportunity to analyze histone modifications at a genome-wide scale using already published ChIP-seq data. Although extracting meaningful biological insights from these datasets alone remains a challenge, multiple datasets can be combined (RNA-seq and Bisulfite-seq) to understand the ChIP-seq data better and obtain the insights that we seek. By applying explainable machine learning techniques, we can quantify the impact of specific histone modifications and characterize their effects on gene expression without relying on knockout models. This approach enables the functional characterization of both well-studied and previously uncharacterized histone marks.

2. Methods

2.1 Data Collection

Various sequencing data was collected from previously published papers and re-analyzed using custom bioinformatic pipelines. Sequencing data was matched as closely as possible for consistency with ChIP-seq data from Arabidopsis seedlings used for the data analysis. 10-day Arabidopsis seedlings were used whenever possible with 2 datasets used with 12-day seedlings and 14-day seedlings respectively. Bisulfite-seq data was collected to provide the model with a more robust dataset to learn about chromatin states as much as possible. Meanwhile, RNA-seq data was collected to be used as the target variable for our model. A total of twenty-two features were collected including various histone modifications using ChIP-seq, DNA methylation data from bisulfite sequencing, and one target variable: FPKM (*Fragments Per*

Kilobase per Million) values from RNA-seq data (10). The data was formatted as a gene by modification matrix where all different Arabidopsis genes aligned to the TAIR10 Col-0 genome (11) were put into the rows of the data matrix and the features were the different histone and DNA modifications.

2.1.1 ChIP-seq Pipeline

The ChIP-seq sequencing data was analyzed based on methods of Nakato et al. (12). All raw FASTQ files underwent quality control and adapter trimming using fastp (13). Trimmed reads were then aligned to the reference genome using Bowtie2 in local alignment mode (19). Paired end reads were aligned as read pairs, and single-end reads were aligned independently. Aligned SAM files were then converted to BAM format using Samtools (14), followed by sorting to arrange alignments based on genomic coordinates. To eliminate PCR duplicates, Picard MarkDuplicates was used (15). This was done to reduce the artificial signal inflation in ChIP-seq data. Deduplicated files were then indexed using Samtools. It was then converted to BigWig and a matrix of signals were produced using deepTools (16). Greenscreen was used when converting files to BigWig to remove false positive signals (17).

2.1.2 Bisulfite-seq Pipeline

The bisulfite sequencing pipeline utilized Bismark for most of the analysis (18). Raw bisulfite sequencing reads were first processed using fastp for adapter removal and quality trimming (13). Trimmed reads were then aligned to a bisulfite-converted TAIR10 reference genome using Bismark which internally uses Bowtie2. Samples were then deduplicated to remove PCR duplicates to reduce amplification bias (15). Bismark then calculates the methylation values for all cytosine contexts in the genome (CpG, CHG, CHH), which can then

be used by deepTools to calculate average methylation values for each gene in the final matrix for downstream analysis.

2.1.3 RNA-seq Pipeline

Raw RNA-seq reads were also first subjected to quality control and adapter trimming using fastp (13). A reference index was then created using RSEM (RNA-Seq by Expectation Maximization) (20). Trimmed reads were then quantified using RSEM. Expression results were the output as a gene-level expression matrix with values such as expected counts, TPM (*Transcripts Per Million*), and FPKM for the sample.

2.2 Data Pre-processing

After the initial bioinformatics analyses, matrices from each of the sequencing pipelines were merged together to produce the dataset. To ensure robust model performance, the skewness in the data was addressed. The histone modification data exhibited left-skewed distributions, where there were a small number of genes with extremely high modification and expression levels. Because left-skewed data can often introduce bias and deteriorate model performance, a log transformation was applied compressing larger values whilst expanding smaller ones making the distribution of the data more symmetrical (Fig. 1). This transformation will enhance the interpretability of the histone modification effects and lower the variance of the model. Feature engineering was also implemented as many genes had zero gene expression. This was addressed by creating a new feature named expression category that specified whether that gene expressed anything at all further helping the model performance and decorrelating some of the model data.

In addition to the transformation, Isolation Forest, an unsupervised outlier detection learning algorithm was implemented. It detects ‘anomalies’ by ‘isolating’ observations by randomly selecting features and split values. This random partitioning will produce a forest of random trees with the idea being that outliers are more easily split by requiring less partitions compared to an observation that is not an outlier. This unsupervised method was chosen because of the non-linear behavior of sequencing data and the computational efficiency of producing random trees compared to estimating distributions of a high dimensional dataset. (Fig. 2) The final dataset had a total of 25045 genes and a total of 23 features.

2.3 Chosen Model: XGBoost

The model used to predict gene expression was XGBoost, an efficient and scalable algorithm from the gradient-boosted decision tree family (21). Unlike traditional decision trees or random forests, XGBoost builds trees sequentially, with each new tree learning to correct the errors of the previous ones. This approach enables it to model complex relationships with high accuracy. Given the high dimensionality and non-linear feature interactions present in epigenomic data, XGBoost was particularly well suited for this task. Its ability to handle heterogeneous feature scales and capture subtle, non-linear patterns made it an ideal choice for modeling gene regulation from sequencing data.

XGBoost was also chosen for its low computational costs and extreme efficiency due to its compatibility with GPUs. This makes training models significantly faster than other tree based models especially when processing large datasets due to XGBoost using parallelized tree construction unlike other boosting algorithms that rely on single-threaded execution. Additionally, XGBoost has built in regularization techniques and high hyperparameter customizability. This combination of efficiency, scalability, and predictive power makes XGBoost an ideal choice for this non-linear dataset.

2.3.1 Considerations of Other Models

Other models were also considered in the selection process. A combination of R-squared score, RMSE, SHAPley value accuracy (compared to known research data) (22), and computational efficiency were used as important metrics for model selection. Models range from simpler ones such as K-Neighbors Regression (KNNR) and Support Vector Regression (SVR), to more complicated ones such as LightGBM (23), custom Multi Layered Perceptrons, and XGBoost. The most important aspect of model selection, however, was computational efficiency.

The focus on computational efficiency was due to the cost of calculating SHAPley values (discussed below) and the number of iterations required to optimize the model hyperparameters. All models listed above, with the exception of LightGBM, are GPU compatible to speed up the process of training and testing.

Once models were initialized, they were trained on the dataset using an 80-20 split. Root mean squared error and r-squared values were compared to choose the model with the best performance. R-squared was chosen to represent a simple way to see how much variation base models are able to capture from the dataset while root mean squared error was used to see how accurate model predictions can be. A combination of the two would be ideal as the model will need to capture as much of the underlying heterogeneous pattern of the dataset and although feature effects are prioritized more in this study, a good prediction score is also needed to ensure that the model does not get entirely incorrect conclusions.

Boosted models performed the best compared to other models. KNNR and SVR did not perform as well, likely due to the high dimensionality of the dataset paired with the highly non-parametric nature of the quantification of sequencing data. The custom multi-layered perceptron was underperforming likely due to the lack of data. The dataset only contains 27000

genes. Moreover, the MLP was also computationally expensive despite leveraging the GPU. In the end, both XGBoost and LightGBM performed the best out of all the models above with the highest r-squared value and the lowest RMSE on their base models. However, XGBoost was chosen over LightGBM due to its slightly faster computation time and better metrics.

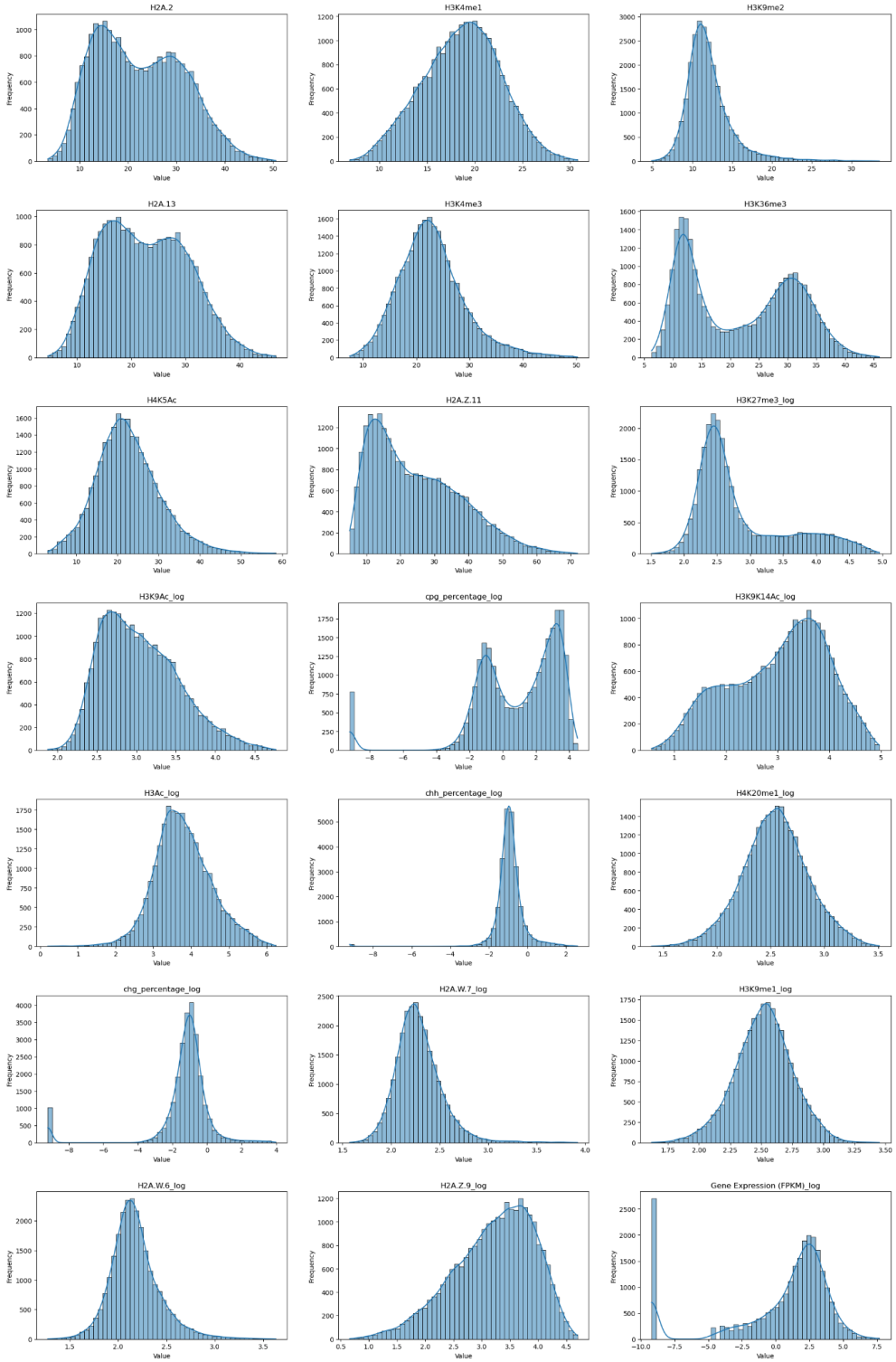

Fig 1 : Distributions of data after the log transformations were done. Some were left untransformed because their distributions were already normal. Note that not all features were log transformed and for our analysis, because SHAP values measure the average contribution towards the prediction, transformations will not affect the final results.

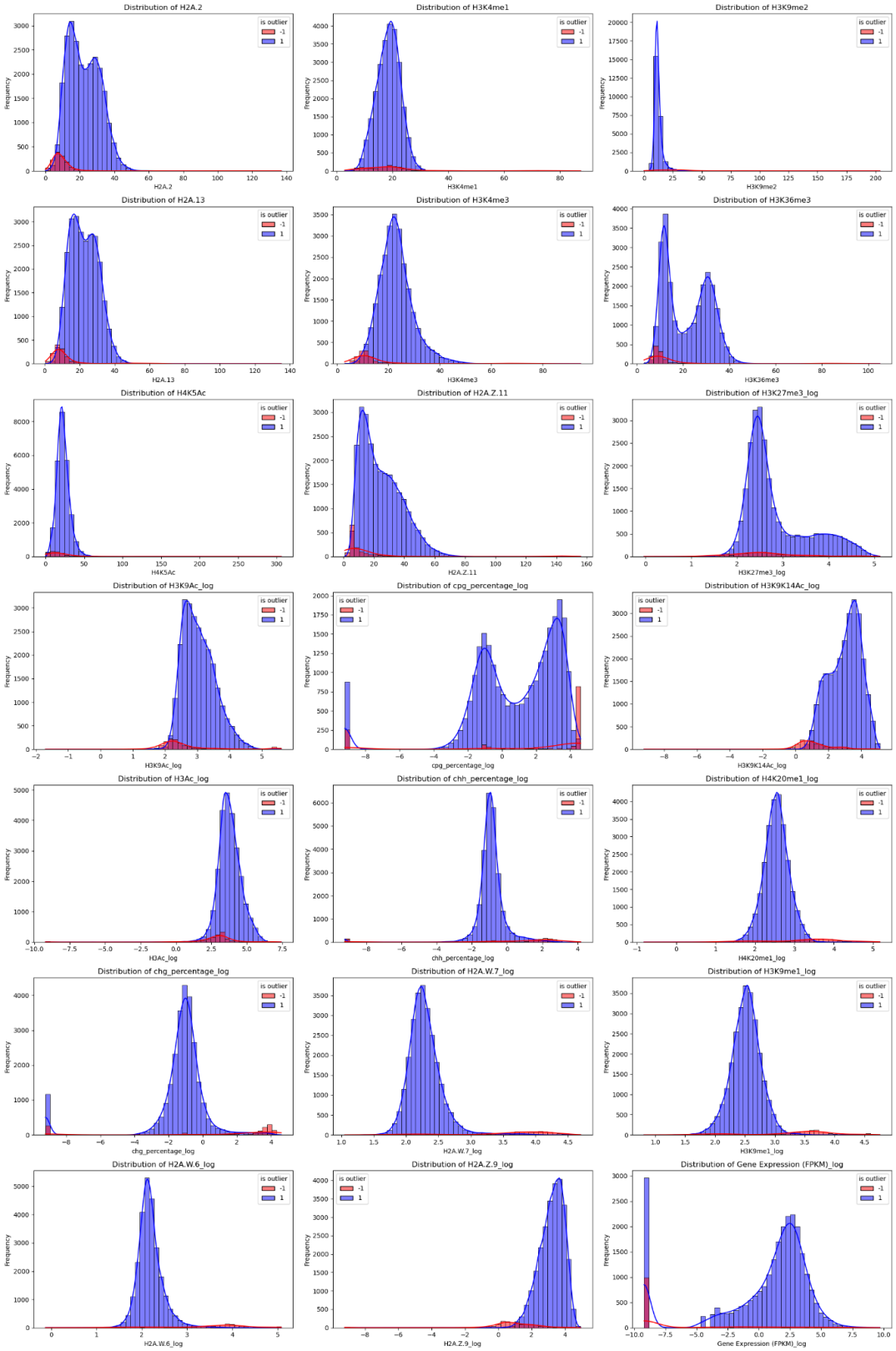

Fig 2 : Outlier removed according to Isolation Forest method. (Red = Outliers, Blue = Inliers). Outliers were determined according to a non-parametric partitioning approach, taking into account all the dimensions of the dataset.

2.4 Explainable AI (XAI) and SHAPley Values

The fundamental objective of this study is to determine a method to characterize different unknown histone modifications on how they influence gene expression. While building a predictive model is essential, an accurate prediction of gene expression does not provide meaningful insights into the effects of different histone modifications and their underlying effects on gene regulation. Instead, a method to quantify the contribution of each histone modification to gene expression is needed, allowing the assessment of their relative impact. This requires more than just model accuracy.

To address this, we leverage explainable artificial intelligence (XAI), a field dedicated to making machine learning models more interpretable (24, 32). Traditional machine learning models, especially complex ones such as XGBoost, operate as “black boxes,” where the relationships they learn remain opaque. Explainability methods help bridge this gap by providing insights into which features drive predictions and to what extent. In this study, SHAP (SHapley Additive exPlanations) values were used, a mathematically grounded approach for quantifying feature importance and effect in predictive models (25).

SHAP values, derived from cooperative game theory, assign a contribution score to each feature by considering all possible combinations of features in the model. Unlike traditional feature importance metrics, which only measure correlation, SHAP values assess the marginal impact of each histone modification on gene expression by evaluating how the model’s prediction changes when the feature is included or excluded. This ensures a fair and consistent attribution of importance, allowing the deconstruction of a model’s decision-making process at both the global (dataset-wide) and local (individual gene) levels. However, calculating all possible combinations of feature values can be extremely computationally expensive, therefore an efficient model was chosen to compensate.

While a well-performing model is necessary to ensure reliable predictions, SHAP values serve as the primary tool for quantifying the impact of histone modifications. This approach allows the ranking of different histone modifications and their relative magnitude of effect on gene expression. More importantly, this also allows the characterization of unknown histone modifications. Ultimately, SHAP values bridge the gap and shine a light on model decisions that allow the characterization of the feature impacts on the target variable. In this case, it allows the characterization of histone modifications on gene expression.

2.5 Leave One out Feature Cross Validation (LOOFCV)

To ensure that the pipeline can correctly characterize histone modifications, a leave one out feature cross validation method was employed. There were two methods involved in making the pipeline.

2.5.1 Method 1: LOOFCV Without Feature Selection

The first was to perform a simple leave one out cross validation but using the features instead of individual observations by measuring the average of the SHAP values for each feature (cross validation using the columns instead of the rows). One feature would be randomly taken out of the dataset and the model would be trained on all other features. Once the optimal model is found, the feature would then be added back in and the model would be tasked to quantify the effect of the feature that was added back in. Feature impacts of each previously well researched histone modification effects were then compared to the concluding information online, whether they would be activating or repressive modifications. In this case, if the average large observation SHAP values calculated for the histone modifications matched previously known functions of those histone modifications, the model would mark those histone modifications as correctly classified while mismatches between the average SHAP values and

known functions of those histone modifications would be marked as incorrectly classified. This will then provide us with an accuracy score that we can use to test whether the method can correctly classify the histone modifications into the proper category.

2.5.2 Method 2: LOOFCV With Feature Selection

The second method is the same as above, but with the added technique of feature selection. It would be similar to LOOFCV but at each iteration of cross validation for each feature, it would drop the most mismatched feature one by one based on the highest SHAP values of the incorrectly classified histone modification. It would then repeat this process until all features were correctly classified, slowly paring down the list of misclassified features, before adding in the feature the model is currently tested to quantify (Fig 3). Although this would require a lot more computation time, this method should further decrease the variance of the model from feature to feature.

Fig 3 : A visual representation of leave one out feature cross validation (LOOFCV). The feature of interest is first isolated from the data set. The chosen model is then trained on the dataset without the feature of interest and optimized. After every step of optimization, incorrectly matched features were discarded and the optimization step is repeated again. The discarding of features is repeated until the model correctly classifies all features. Once all features are correctly classified, the unknown feature is then added back into the model to be characterized using the parameters trained on the model.

2.6 Model Optimization

If left alone, XGBoost and SHAPley would perform at an acceptable rate. However, to further optimize the performance of the model, Optuna was employed to iteratively find the best hyperparameters possible for each model and each iteration of the cross validation (26). It works by using intelligent samplers where a distribution of the hyperparameter space is estimated for each model. It uses non-parametric density estimators to model that distribution space and iteratively searches for the best hyperparameters by maximizing the ratio of good

hyperparameters compared to bad parameters. For Method 1 where feature selection was not included, Optuna was used at every iteration of cross validation for each feature of interest and repeated for each round of cross validation. In method two, Optuna was used at each round of cross validation including the rounds where mismatched features were eliminated one by one. Each model was given forty iterations of Optuna to determine the best hyperparameter configuration possible. Although more iterations would be best to search the hyperparameter space more thoroughly, each Optuna iteration requires the calculation of SHAP values which are computationally very expensive. This further emphasizes the computational efficiency needed in our model even further.

The metric that Optuna used is the accuracy: $\frac{No.\ Correct\ Previously\ Known\ Modifications}{All\ Previously\ Known\ Modifications}$. Specifically, accuracy was used as the metric that Optuna is tasked to maximize. By doing so, each iteration of cross validation is optimized and because the dataset for the feature of interest is withheld from the model when it is trained, data snooping was avoided.

3. Results

The two methods were tested to gauge the accuracy of characterizing histone modifications this way. The first was the characterization of histone modifications using LOOFCV alone. This proved to be a promising method boasting an accuracy rate of 82%. However, when LOOFCV was employed with further feature selection, the accuracy rate of the second methodology increased to 88%. This makes LOOFCV paired with feature selection a promising method to characterize histone modifications and their effects on gene expression. SHAPley values were able to match previous research and the magnitudes of the modifications measured by the absolute magnitude of SHAPley values were also accurate with the combination of tri-methylation on lysine 36 of histone 3 (H3K36me3) to be an important activating mark involved in transcriptional elongation (27) and the combination of multiple acetylating marks on lysine 9 and 14 on histone 3 (H3K9K14Ac) to also be one of the more

impactful activating marks (28, 29). These all point towards an accurate model to characterize important histone modifications. Moreover, by using SHAP values and comparing the histone modification effects alongside each other, we are able to determine the relative magnitudes of each histone modification when compared to each other in terms of gene expression.

3.1 Histone Modification Magnitude Results

When comparing the results of the magnitude of histone modifications, it looks like H3K36me3 and H3K9K14Ac have the biggest activating roles in terms of increasing gene expression while the isoform of H2A.Z.11 has the biggest impact on decreasing gene expression. Through measuring each SHAPley value and plotting its distributions of contributions to gene expression, we were able to show the relative ranking of importance of histone modifications. This method will prove to be useful in quantitatively validating known histone modifications but also in uncovering potential novel regulatory interactions between modifications that were previously not known.

In the dataset that was used, H2A.Z.11 seems to be a promising histone isoform to explore. Although some of H2A.Z's isoforms have been studied before (30, 31), H2A.Z.11 has not been extensively studied. Given that H2A.Z variants are known to influence chromatin accessibility in both transcription and repression and that the model has ranked it high in terms of functional impact of gene expression, it seems likely that H2A.Z.11 can play an important role in the regulatory processes of transcriptional repression.

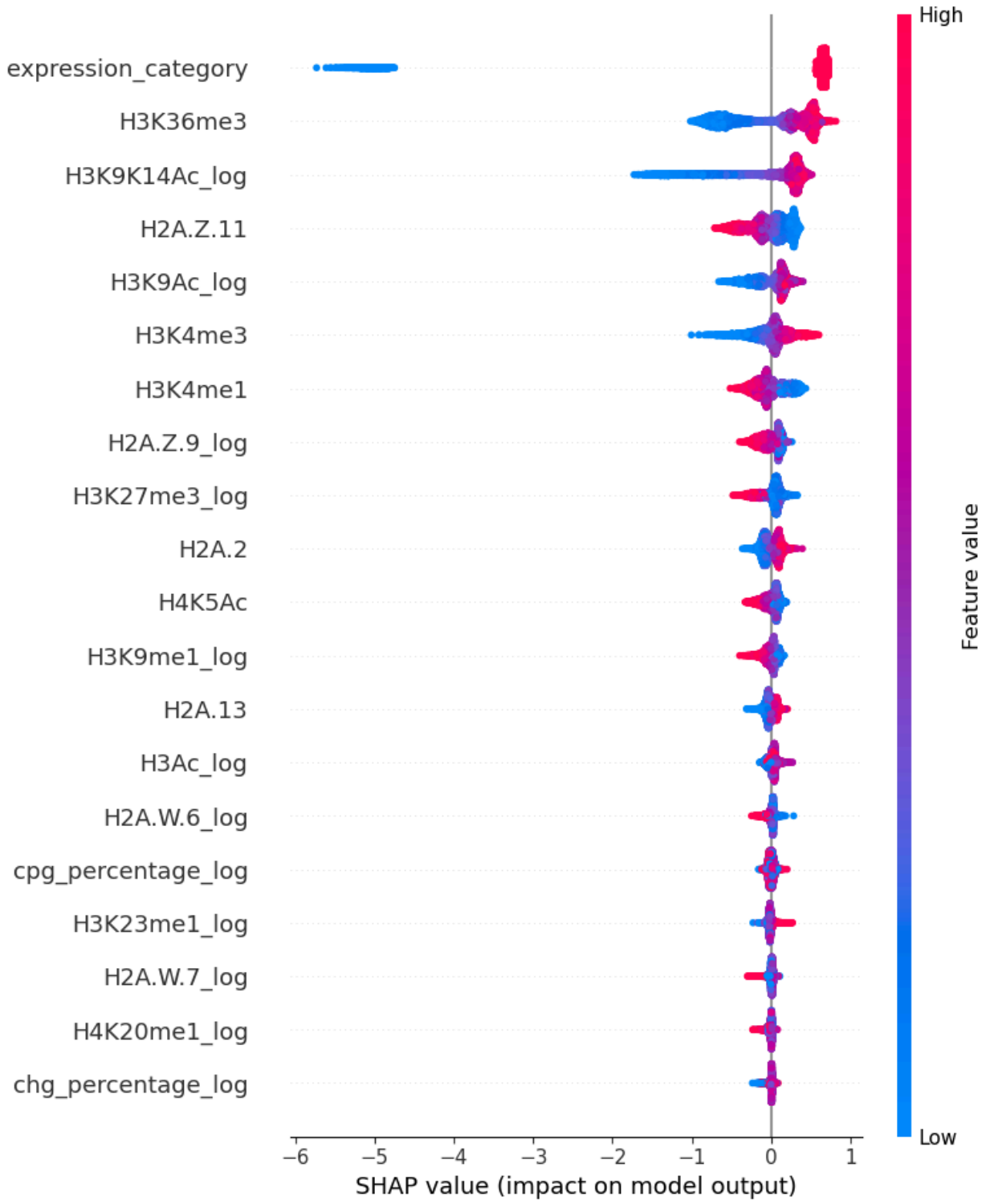

Fig 4 : SHAP beeswarm plot to quantify feature effect magnitude of Method 1. Red points mean high feature values and blue points mean low feature values. Their position on the X axis determines their positive or negative effects on the prediction. *Method 2 beeswarm was not shown because each feature it trained on had values it feature selected out.

3.2 Unknown, Confounded Histone Modification

Because the method has achieved an 88% accuracy rate, a previously unknown histone modification was added into the LOOFCV to be quantified. Histone H3 lysine 23 monomethylation (H3K23me1) is a post-translational modification chosen for quantification using this model. It was selected because of its dependence on KRYPTONITE, a histone methyltransferase that also controls H3K9 methylation (33). This dual role presents a significant confounding issue in traditional genetic studies, as knocking out KRYPTONITE not only disrupts H3K23me1 levels, but also H3K9 methylation levels making it impossible to isolate the specific effects of H3K23me1 on gene expression. Because of this limitation, an alternative approach is needed to quantify the functional impact of H3K23me1 without genetic perturbation.

By using explainable AI, we avoided the confounding effects associated with traditional knockout models and instead used SHAP values to quantify the independent contribution of H3K23me1 to gene expression. According to the model (Fig. 4), H3K23me1 has a relatively low overall importance compared to other histone modifications. However, on average, the SHAP values for H3K23me1 are slightly positive, indicating that when this mark is present, the model tends to predict higher gene expression. Furthermore, H3K23me1 is known to be rarely deposited in the genome, which likely explains its low relative importance in the model's predictions (33). Overall, these results suggest that while H3K23me1 presence may be weakly associated with increased expression prediction in this context, its biological impact is likely limited due to its low abundance. Moreover, H3K23me1 also seems to primarily mark transposable elements and not as many genes which might also be another reason that the model seems to show a weak association with gene expression as the dataset the model was trained on does not contain transposable element data.

4. Discussion

Histone modifications are essential regulators of gene expression, chromatin structure, and overall genome stability. These modifications heavily influence how DNA is packaged and accessed by transcriptional machinery. These activating or repressive marks play a pivotal role in developmental transitions, differentiation, stress responses, and transposon silencing by ensuring precise and context-dependent gene regulation. Specifically in plants, they control key processes like flowering time, stem cell maintenance, and drought tolerance. Despite their importance however, many histone modifications still remain poorly characterized due to their complex interactions and confounding regulatory networks. Paving the way to understand the functions of unknown histone modifications is crucial for further uncovering the regulatory mechanisms controlling gene expression and, ultimately, the adaptability and survival of plants.

The model above offers a solution to help identifying important histone modifications and characterizing previously unknown modifications. The ability to characterize previously unknown histone modifications and rank their relative importance solely through sequencing data solves the problem above. Traditional approaches can struggle with confounding effects making it difficult to isolate the function of individual histone marks when they are regulated by the same enzymatic machinery. By using a machine learning-driven framework, the model can quantify the impact of histone modifications without the need for genetic perturbations. This not only accelerates the discovery of novel regulatory roles but can also guide researchers to prioritize modifications based on their functional significance as provided by the magnitude of the SHAP values. The ability to rank histone modifications enables a deeper understanding of how different marks contribute to gene expression, guiding future experimental research toward the most biologically relevant modifications.

With the model performing with an accuracy of 88% on previously known modifications and achieving a low RMSE value, the model proves to be a valuable tool in helping inform future

researchers on the characterization of histone modifications. More specifically, unknown histone modifications can be characterized faster and their impacts measured before deciding whether to pursue the exploration of such unknown modifications further. Because the model was also able to isolate the function of individual histone marks controlled by a shared enzyme, H3K23me1 was characterized to be a minor but positive influence on gene expression. This highlights the strengths of such a model as no knockout experiments were needed, and only one form of sequencing data was used.

5. Future Work

While this study successfully demonstrates the ability of explainable machine learning to characterize histone modifications and quantify their effects on gene expression, several key improvements could further enhance the model's performance and reliability. One primary avenue for future work involves expanding the dataset to allow the model to learn from a broader and more diverse set of histone modifications. Increasing the number of samples involved in the dataset such as including transposable elements or even including mutant data could help improve model performance. More specifically, mutant datasets where specific histone modifiers are knocked out or overexpressed could provide the model causal insights into how the modifications affect the gene expression helping the model learn from a larger breadth of data. This could lead to more robust and generalizable predictions. The challenge with this would be to ensure that the RNA-seq data is consistent as RNA sequencing is highly variable from experiment to experiment.

Another challenge that remains is the correlation between histone modifications and in some cases, genes closer to each other or even overlapping with each other. Because many marks can interact with each other due to histone crosstalk, the interdependence can make it difficult to disentangle the individual contributions of each modification. More work could be

done to decorrelate the features through techniques like residualization or smart feature engineering. Similarly, because genes located close in proximity can share regulatory elements and chromatin environments, incorporating spatial decorrelation methods could help reduce biases introduced from local chromatin effects. This would ensure that observations are more independent, leading to a clearer understanding of each histone modification's true effect.

Beyond improving the dataset and feature independence, model evaluation and optimization could also be refined by developing a custom performance metric that Optuna could maximize (or minimize). While the accuracy metric provides a measure of success, metrics combining total average correct SHAP values or a penalization of RMSE could help improve the model performance even more.

By addressing these improvements: expanding the dataset, reducing feature correlations, and refining the evaluation metrics, this can further strengthen the ability of the model to better uncover the functional roles of histone modifications, providing a more powerful framework for epigenetic analysis.

Data Availability

The following previously published datasets were used. Sequencing data from the studies below were downloaded and analyzed independently. Raw data can be found online on NCBI's GEO database with accession numbers: GSE116068 (1), PRJNA552176 (2), GSE167288 (3), GSE226469 (4), GSE231408 (4). Unpublished H3K23me1 ChIP-seq data from 10-day Arabidopsis seedlings was generated and provided by the Zhong Lab.

1. Tan, L.M., Zhang, C.J., Hou, X.M., Shao, C.R., Lu, Y.J., Zhou, J.X., Li, Y.Q., Li, L., Chen, S. and He, X.J., 2018. The PEAT protein complexes are required for histone deacetylation and heterochromatin silencing. *The EMBO journal*, 37(19), p.e98770.
2. Lin, J., Hung, F.Y., Ye, C., Hong, L., Shih, Y.H., Wu, K. and Li, Q.Q., 2020. HDA6-dependent histone deacetylation regulates mRNA polyadenylation in Arabidopsis. *Genome research*, 30(10), pp.1407-1417.
3. Zhou, X., He, J., Velanis, C.N., Zhu, Y., He, Y., Tang, K., Zhu, M., Graser, L., de Leau, E., Wang, X. and Zhang, L., 2021. A domesticated Harbinger transposase forms a complex with HDA6 and promotes histone H3 deacetylation at genes but not TEs in Arabidopsis. *Journal of integrative plant biology*, 63(8), pp.1462-1474.
4. Jamge, B., Lorković, Z.J., Axelsson, E., Osakabe, A., Shukla, V., Yelagandula, R., Akimcheva, S., Kuehn, A.L. and Berger, F., 2023. Histone variants shape chromatin states in Arabidopsis. *Elife*, 12, p.RP87714.

All scripts and code used for data processing, model training, and analysis are available on GitHub at: https://github.com/d-setiawan/characterization-of-histone-modification-using-ML

References

1. Holliday, R., 2006. Epigenetics: a historical overview. *Epigenetics*, 1(2), pp.76-80.
2. Klemm, S.L., Shipony, Z. and Greenleaf, W.J., 2019. Chromatin accessibility and the regulatory epigenome. *Nature Reviews Genetics*, 20(4), pp.207-220.
3. Peterson, C.L. and Laniel, M.A., 2004. Histones and histone modifications. *Current Biology*, 14(14), pp.R546-R551.
4. Pi, L., Aichinger, E., van der Graaff, E., Llavata-Peris, C.I., Weijers, D., Hennig, L., Groot, E. and Laux, T., 2015. Organizer-derived WOX5 signal maintains root columella stem cells through chromatin-mediated repression of CDF4 expression. *Developmental cell*, 33(5), pp.576-588.
5. Whittaker, C. and Dean, C., 2017. The FLC locus: a platform for discoveries in epigenetics and adaptation. *Annual review of cell and developmental biology*, 33(1), pp.555-575.
6. Van Dijk, K., Ding, Y., Malkaram, S., Riethoven, J.J.M., Liu, R., Yang, J., Laczko, P., Chen, H., Xia, Y., Ladunga, I. and Avramova, Z., 2010. Dynamic changes in genome-wide histone H3 lysine 4 methylation patterns in response to dehydration stress in Arabidopsis thaliana. BMC plant biology, 10, pp.1-12.
7. Sani, E., Herzyk, P., Perrella, G., Colot, V. and Amtmann, A., 2013. Hyperosmotic priming of Arabidopsis seedlings establishes a long-term somatic memory accompanied by specific changes of the epigenome. *Genome biology*, 14, pp.1-24.
8. Jamge, B., Lorković, Z.J., Axelsson, E., Osakabe, A., Shukla, V., Yelagandula, R., Akimcheva, S., Kuehn, A.L. and Berger, F., 2023. Histone variants shape chromatin states in Arabidopsis. *Elife*, 12, p.RP87714.
9. Trejo-Arellano, M.S., Mahrez, W., Nakamura, M., Moreno-Romero, J., Nanni, P., Köhler, C. and Hennig, L., 2017. H3K23me1 is an evolutionarily conserved histone modification

associated with CG DNA methylation in Arabidopsis. *The Plant Journal*, 90(2), pp.293-303.

10. Chatterjee, A., Ahn, A., Rodger, E.J., Stockwell, P.A. and Eccles, M.R., 2018. A guide for designing and analyzing RNA-Seq data. *Gene expression analysis: methods and protocols*, pp.35-80.
11. Berardini, T.Z., Reiser, L., Li, D., Mezheritsky, Y., Muller, R., Strait, E. and Huala, E., 2015. The Arabidopsis information resource: making and mining the “gold standard” annotated reference plant genome. *genesis*, 53(8), pp.474-485.
12. Nakato, R. and Shirahige, K., 2017. Recent advances in ChIP-seq analysis: from quality management to whole-genome annotation. *Briefings in bioinformatics*, 18(2), pp.279-290.
13. Chen, S., Zhou, Y., Chen, Y. and Gu, J., 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, 34(17), pp.i884-i890.
14. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. and 1000 Genome Project Data Processing Subgroup, 2009. The sequence alignment/map format and SAMtools. *bioinformatics*, 25(16), pp.2078-2079.
15. Picard (2018) Picard Tools - *By Broad Institute*. Available at: https://broadinstitute.github.io/picard/ (Accessed: 24 March 2025).
16. Ramírez, F., Dündar, F., Diehl, S., Grüning, B.A. and Manke, T., 2014. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic acids research*, 42(W1), pp.W187-W191.
17. Klasfeld, S., Roulé, T. and Wagner, D., 2022. Greenscreen: A simple method to remove artifactual signals and enrich for true peaks in genomic datasets including ChIP-seq data. *The Plant Cell*, 34(12), pp.4795-4815.
18. Krueger, F. and Andrews, S.R., 2011. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *bioinformatics*, 27(11), pp.1571-1572.

19. Langmead, B. and Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. *Nature methods*, 9(4), pp.357-359.
20. Li, B. and Dewey, C.N., 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics*, 12, pp.1-16.
21. Chen, T. and Guestrin, C., 2016, August. Xgboost: A scalable tree boosting system. *In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).
22. Winter, E., 2002. The shapley value. *Handbook of game theory with economic applications*, 3, pp.2025-2054.
23. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. and Liu, T.Y., 2017. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30.
24. Minh, D., Wang, H.X., Li, Y.F. and Nguyen, T.N., 2022. Explainable artificial intelligence: a comprehensive review. *Artificial Intelligence Review*, pp.1-66.
25. Lundberg, S.M. and Lee, S.I., 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.
26. Akiba, T., Sano, S., Yanase, T., Ohta, T. and Koyama, M., 2019, July. Optuna: A next-generation hyperparameter optimization framework. *In Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 2623-2631).
27. Kouzarides, T., 2007. Chromatin modifications and their function. *Cell*, 128(4), pp.693-705.
28. Igolkina, A.A., Zinkevich, A., Karandasheva, K.O., Popov, A.A., Selifanova, M.V., Nikolaeva, D., Tkachev, V., Penzar, D., Nikitin, D.M. and Buzdin, A., 2019. H3K4me3, H3K9ac, H3K27ac, H3K27me3 and H3K9me3 histone tags suggest distinct regulatory evolution of open and condensed chromatin landmarks. *Cells*, 8(9), p.1034.

29. Roth, S.Y., Denu, J.M. and Allis, C.D., 2001. Histone acetyltransferases. *Annual review of biochemistry*, 70(1), pp.81-120.

30. Lei, B. and Berger, F., 2020. H2A variants in Arabidopsis: versatile regulators of genome activity. *Plant Communications*, 1(1).

31. Gómez-Zambrano, Á., Merini, W. and Calonje, M., 2019. The repressive role of Arabidopsis H2A. Z in transcriptional regulation depends on AtBMI1 activity. *Nature Communications*, 10(1), p.2828.

32. Novakovsky, G., Dexter, N., Libbrecht, M.W., Wasserman, W.W. and Mostafavi, S., 2023. Obtaining genetics insights from deep learning via explainable artificial intelligence. *Nature Reviews Genetics*, 24(2), pp.125-137.

33. Trejo-Arellano, M.S., Mahrez, W., Nakamura, M., Moreno-Romero, J., Nanni, P., Köhler, C. and Hennig, L., 2017. H3K23me1 is an evolutionarily conserved histone modification associated with CG DNA methylation in Arabidopsis. *The Plant Journal*, 90(2), pp.293-303.