Unravelling epigenetic regulation of gene expression with explainable AI - a case study leveraging degron data

Kashyap Chhatbar[1,2], Adrian Bird[2], Guido Sanguinetti[3]

Abstract

Transcriptional regulation involves complex interactions involving chromatin–associated proteins, but disentangling these mechanistically remains challenging. Here, we generate deep learning models to predict RNA Pol–II occupancy from chromatin–associated protein profiles in unperturbed conditions. We evaluate the suitability of Shapley Additive Explanations (SHAP), a widely used explainable AI (XAI) approach, to infer causality and analyse regulatory mechanisms across diverse datasets and to validate these insights using data from degron–based perturbation experiments. Remarkably, genes ranked by SHAP importance predict direct targets of perturbation even from unperturbed data, enabling inference without costly experimental interventions. Our analysis reveals that SHAP not only predicts differential gene expression but also captures the magnitude of transcriptional changes. We validate the cooperative roles of SET1A and ZC3H4 at promoters and distinct regulatory contributions of ZC3H4 at gene bodies in influencing transcription. Cross-dataset validation uncovers unexpected connections between ZC3H4, a component of the Restrictor complex, and INTS11, part of the Integrator complex, suggesting crosstalk mediated by H3K4me3 and the SET1/COMPASS complex in transcriptional regulation. These findings highlight the power of integrating predictive modelling and experimental validation to unravel complex context-dependent regulatory networks and generate novel biological hypotheses.

1 School of Informatics, University of Edinburgh, Edinburgh, United Kingdom. kashyap.c@ed.ac.uk
2 Institute of Cell Biology, University of Edinburgh, Edinburgh, United Kingdom.
3 International School for Advanced Studies (SISSA), Trieste, Italy. gsanguin@sissa.it

Kashyap Chhatbar was supported by a scholarship from College of Science and Engineering, University of Edinburgh.

Introduction

Understanding how cells regulate gene expression during physiological or disease processes remains one of the most important open problems in biology, with potentially immense implications for biomedicine and biotechnology. While the fundamental process of regulation by transcription factor binding has been extensively studied[1], the complexity of sequence–dependent signal integration in a crowded chromatin environment remains largely unknown. Through efforts to tackle this outstanding problem, the last decade has witnessed the development of several large–scale international efforts to profile multiple epigenomic marks across a variety of cellular contexts[2–4]. The large amounts of data made publicly available by these projects raised the prospect of a complementary, data–driven approach to understanding general mechanisms by which gene expression is regulated. Numerous studies have used statistical models to predict gene expression or other molecular features from genomic sequence and epigenomic marks[5–11]. More recently, advances in artificial intelligence (AI) using a variety of neural architectures have been brought to bear, allowing increasingly accurate prediction and imputation of gene expression as well as other molecular features[12–15]. A critical limitation of this data–driven approach is its entirely correlative nature. Because the vast majority of data sets are derived from cells at steady-state, every prediction will demand independent validation via explicit perturbation experiments, a requirement that has rarely been met.

Since complex models generally trade off biological interpretability against statistical accuracy, so-called explainable AI (XAI) techniques are adopted to provide human interpretable explanations for specific model predictions and to hypothesize potential underpinning biological mechanisms[16–18]. In this paper, we set out to benchmark the efficacy of XAI approaches in identifying potentially causal regulatory processes by taking advantage of the recent availability of dynamical measurements provided after rapid perturbations via auxin-inducible degron (AID) and degradation tag (dTAG) systems[19]. We set out to predict RNA Polymerase II occupancy from several chromatin associated proteins and epigenomic marks, quantifying the influence of each protein or mark on each gene using SHAP (Shapley Additive Explanations)[17], one of the most widely used XAI approaches. We then validate the efficacy of the learnt explanations by systematic comparisons of direct target and non-target gene sets following degradation of one or more factors. Our findings demonstrate the utility of combining predictive modeling with SHAP analysis to uncover complex biological interactions and infer causal mechanisms in transcriptional regulation. By leveraging this integrative approach, we dissected the roles of chromatin-associated proteins, uncovered unexpected functional relationships, and validated the robustness of SHAP across datasets.

Results

Data sets used for modeling

To ensure a robust and comprehensive analysis of transcriptional regulation, we utilized datasets from three independent studies spanning different experimental designs and chromatin–associated

protein targetes: Dobrinic et al. (2021)[20], Hughes et al. (2023)[21], and Wang et al. (2023)[22] (Table 1). Dobrinic and colleagues[20] investigated the roles of PRC1[23, 24] and PRC2[25, 26] complexes in Polycomb-mediated transcriptional repression, using ChIP-seq data for RING1B, SUZ12, and associated histone modifications such as H3K27me3 and H2AK119ub1 to evaluate their contributions. Additionally, this study included degron-based perturbation experiments for RING1B and SUZ12, enabling a direct assessment of acute transcriptional changes using nuclear RNA-seq assays. Hughes and colleagues[21] focused on transcriptional regulation mediated by SET1A[27] and ZC3H4[28], two chromatin-associated proteins with well-characterized roles in activating and restricting transcription, respectively. This study leveraged ChIP-seq data for SET1A, ZC3H4, and H3K4me3, alongside degron-based perturbation experiments, to uncover the interplay between these factors to regulate transcription. Wang and colleagues[22] explored the role of H3K4me3 in transcription initiation and RNA Pol–II pause—release, utilizing ChIP-seq data for the SET1 complex components DPY30 and RBBP5[27], Integrator complex member INTS11[29, 30], and factors involved in transcription elongation and Pol II pause—release such as CDK9[31], BRD4[32], HEXIM1, NELFA[33], and AFF4[34]. Transcriptional dynamics after acute perturbations were assayed using TT-chem-seq[21, 22]. Together, these datasets provided a rich resource for examining the relationships between chromatin modifications and transcriptional regulation.

Modelling of protein occupancy predicts transcription output

In order to understand the relationships between protein occupancy profiles and transcriptional regulation, we built deep neural network (DNN) models, specifically multilayer perceptrons (MLPs), to predict RNA Pol–II occupancy and steady-state mRNA levels from ChIP-seq data (Figure 1A). MLPs are prioritized in this study for their competitive performance and efficiency in regression tasks[35, 36], along with their scalability and potential for improved interpretability using XAI approaches. To enhance the resolution of our predictions, normalized ChIP signals were quantified separately from promoter regions and gene body regions, enabling us to investigate the distinct contributions of chromatin proteins and marks to transcription regulation across these regions. By integrating these separate measurements, we aimed to identify specific factors or chromatin features that predominantly influence transcription at promoters or gene bodies. For all data sets, the occupancy of RNA Pol–II was predicted with significantly higher precision (R^2 ranging between 0.9–0.95) (Figure 1B) compared to steady state mRNA levels (R^2 ranging between 0.7–0.8) (Figure 1C). This disparity likely reflects the nature of the data: RNA Pol–II occupancy is assayed by ChIP-seq, which also profiles the occupancy of other transcription-associated proteins and histone modifications. In contrast, steady-state mRNA levels are influenced by additional post-transcriptional processes, such as RNA processing and degradation, making mRNA level predictions inherently more complex. We chose to concentrate on models which predict RNA Pol–II occupancy for further analysis.

We hypothesize that the model, trained on data assayed from unperturbed conditions, captures biologically meaningful relationships between protein occupancy and transcriptional output. To interpret these relationships, we use SHAP (SHapley Additive exPlanations)[17], an XAI approach that quantifies the contribution of each input feature to the model's predictions while potentially

accounting for complex, non-linear interactions (Figure 1A). Through comparing SHAP values between direct transcriptional targets (genes that exhibit significant transcriptional changes following rapid protein degradation) and non-target genes, we aim to assess the relative importance of specific protein occupancy features in driving transcriptional outcomes. For robust comparison, we calculate the median absolute SHAP values for each group, estimating error through random subsampling. Since the model is trained on data assayed before any perturbations, differences in SHAP values between direct targets and non-target genes may indicate SHAP's potential to infer causality.

To benchmark SHAP's effectiveness against traditional approaches, we also conducted separate linear regression analyses for target and non-target genes. Notice that linear regression is given access to some of the information from the experiment whose outcome it is trying to predict (i.e., which genes will be differentially expressed upon degradation). This unfair advantage makes linear regression a stronger competitor in our comparison. Linear regression coefficients can provide only global insights, revealing general trends across the entire set of target or non-target genes. In contrast, SHAP analysis calculates feature importance scores for each gene individually, offering unprecedented resolution into how different regulatory features influence transcription in a gene-specific manner. We anticipated that this gene-by-gene analysis would reveal complex, context-dependent interactions between features that would be masked by global linear trends. This comparative approach was designed to test whether SHAP's ability to capture gene-specific regulatory patterns could provide deeper mechanistic insights than traditional analyses.

Strengths and limitations of SHAP analysis in resolving co-occupying transcriptional regulators

To evaluate SHAP's ability to accurately attribute importance to individual proteins in complex regulatory systems, we generated a model to predict RNA Pol–II occupancy using PRC1 and PRC2 ChIP-seq data. PRC1, through its core subunit RING1B[23, 24], functions as the primary transcriptional repressor[37], while PRC2 reinforces repression by depositing the H3K27me3 histone mark[38]. The extensive co-occupancy of PRC1 and PRC2 at target loci[39] provided an ideal system to test SHAP's performance in disentangling overlapping regulatory inputs. As expected, RING1B promoter occupancy emerged as significantly more important at direct target gene loci across all time points (2h, 4h, 8h, and 24h; Figure S1A, B, C, and D), following RING1B depletion, aligning with extensive research that has established PRC1's dominant role in early transcriptional repression. This consistency represents a significant achievement, as it demonstrates SHAP's ability to capture biologically meaningful features, even in a highly co-linear regulatory environment. However, SHAP analysis at the 4h and 8h time points assigned higher importance to SUZ12, a PRC2 subunit, over RING1B (Figure S1B, C). It is important to note that the model was trained on unperturbed data, and thus it cannot reasonably infer or capture the temporal ordering of regulatory effects. The dynamics of feedback mechanism between PRC1 and PRC2[25, 40–42] makes it plausible for SUZ12, a PRC2 subunit, to show higher importance at these time points.

Interestingly, SHAP also highlighted the importance of H3K4me3 at target gene loci (Figure S1A, B, C, D), raising the possibility that the SET1 complex, which deposits this histone mark[27] is

recruited to polycomb-repressed genes. This observation aligns with the established roles of CFP1, a SET1 complex component that binds CpG islands (CGIs)[43], and KDM2B, a PRC1 complex recruiter which also binds CGIs[44]. These findings provide further evidence of a dynamic interplay between the SET1 and PRC complexes, wherein CFP1 and KDM2B continuously sample CGIs but ultimately yield to the reinforcing feedback mechanisms of PRC1 and PRC2[45], maintaining transcriptionally repressed loci. These findings also raise the possibility that these loci represent "bivalent" promoters, where H3K4me3 and H3K27me3 coexist[46]. This dual marking is thought to safeguard proper and robust differentiation by maintaining genes in a poised state, balancing activation and repression until lineage-specific cues drive transcriptional commitment[46, 47], further emphasizing the dynamic regulatory interplay revealed by our analysis.

Overall, these results demonstrate the utility of SHAP in consistently identifying biologically validated regulators like RING1B as key players, even in complex and co-linear regulatory systems. Surprisingly, SUZ12 importance is highlighted earlier at 4h and 8h, well ahead of its observed deposition of the H3K27me3 mark[48].

Interpretable modelling analysis aids in inferring the combined role of SET1A and ZC3H4 in transcriptional regulation

A recent study[21] revealed the combined role of SET1A and ZC3H4 in transcriptional regulation, highlighting their interplay. Building on these datasets, we aimed to determine whether XAI approaches could infer causal relationships between these proteins directly from predictive models trained on unperturbed data. Understanding such cooperative roles demands methodologies that go beyond traditional bioinformatic techniques, which often rely on summary statistics and correlation analyses to infer biological mechanisms. For instance, SET1A, a histone methyltransferase responsible for H3K4 trimethylation[27], co-occupies promoters with H3K4me3, as demonstrated by ChIP-seq analysis (Figure S2A). SET1A, in complex with WDR82, functions as a transcriptional activator, while ZC3H4, also interacting with WDR82, is part of the restrictor complex that regulates non-coding transcription[28, 49]. ChIP-seq data further revealed co-occupancy of SET1A and ZC3H4 at promoters (Figure S2A), suggesting potential interplay between these two proteins.

To shed light on their roles in transcriptional regulation, we utilised our highly accurate model (Figure 1B) which predicts RNA Pol–II occupancy using the occupancy profiles of SET1A, ZC3H4, and H3K4me3. SHAP (SHapley Additive exPlanations) analysis of the model revealed that ZC3H4 exhibited higher importance than H3K4me3 at promoters where SET1A was most significant (Figure S2B). This finding provides support for a cooperative relationship between SET1A and ZC3H4 in regulating transcription, aligning well with experimental evidence[21]. To evaluate potential causality, we performed SHAP analysis on gene sets following the acute degradation of SET1A, ZC3H4, and both proteins simultaneously. SHAP values were consistently higher for SET1A and ZC3H4 promoter occupancy in target genes compared to non-target genes (Figure 2A, B, C), indicating that the model captures biologically relevant effects of these proteins on transcriptional regulation. Additionally, SHAP analysis highlighted a higher importance of ZC3H4 gene body (gb) occupancy in target genes, whereas SET1A showed no significant increase in gene body importance. As a control comparison,

we trained separate linear regression models on target and non-target genes to evaluate whether the relationships between input features and transcriptional regulation differ across conditions. The coefficients derived from these models showed no significant differences (Figure S3A), indicating that the predictive relationships remain consistent between target and non-target gene sets. These results demonstrate that XAI approaches, such as SHAP, provide valuable insights into the cooperative roles of SET1A and ZC3H4 at promoters. This finding aligns with established experimental evidence[21], validating the robustness of our approach. Notably, the distinct regulatory role of ZC3H4 at gene body represents a potentially novel phenomenon, extending beyond published conclusions and uncovering novel aspects of ZC3H4's contribution to transcriptional regulation.

SHAP analysis validates the importance of INTS11 at target genes following RBBP5 and DPY30 degradation

DPY30 and RBBP5 are critical components of the SET1 complex[22, 27], which facilitates the deposition of the H3K4me3 histone mark, a key epigenetic modification associated with active transcription. INTS11, a core component of the Integrator complex[29], is recruited to active promoters through H3K4me3[22]. This interaction highlights the crosstalk between the SET1 complex and Integrator complex in regulating transcriptional activity. To further investigate these relationships, we utilized our model, employing SHAP to quantify the contributions of DPY30, RBBP5 and INTS11 along with CDK9, HEXIM1, BRD4, NELFA, and AFF4, each playing crucial roles in transcription elongation and Pol II pause—release[31, 32, 34, 50].

Our analysis revealed that INTS11 promoter occupancy consistently exhibited high SHAP values at target gene loci, regardless of whether INTS11, DPY30, or RBBP5 was degraded (Figure 2D, E, F). This finding underscores INTS11's central role, likely facilitated through its interaction with H3K4me3 at active promoters. Notably, H3K4me3, DPY30, and NELFA promoter occupancies were also identified as significant contributors to RNA Pol–II occupancy, particularly at INTS11 target gene loci (Figure 2A). Despite DPY30 and RBBP5 being part of the same SET1 complex depositing H3K4me3, SHAP analysis revealed a lack of significant importance for RBBP5 (Figure 2D, E, F), an observation that stands in contrast to expectations. Our analysis highlights the importance of DPY30, NELFA and H3K4me3 promoter occupancy in the regulation of target genes, supporting the broader role of H3K4me3 in regulating RNA Pol–II pausing[22]. Notably, our results independently validate the central role of INTS11 in regulating transcription at target loci, emphasizing its critical function in transcriptional regulation through H3K4me3 recruitment.

Cross-dataset validation of SHAP analysis reveals unexpected insights into ZC3H4 and INTS11 in transcription termination

To investigate the generalizability of our findings across datasets, we performed a cross-dataset comparison by analyzing gene sets derived from the degradation of DPY30, RBBP5, and INTS11 alongside predictive models trained on independent ChIP-seq data[21]. We also conducted the reverse analysis, using gene sets identified as likely direct targets of SET1A, ZC3H4, and the simultaneous degradation of SET1A and ZC3H4 to validate predictive models trained on ChIP-seq data[22]. This

bipartite approach enabled us to evaluate whether SHAP values derived from one dataset could reliably capture the transcriptional regulatory roles in an independent dataset, providing a stringent test of both the predictive model and the SHAP framework.

As predicted, SET1A and ZC3H4 emerged as having higher SHAP importance at DPY30 and RBBP5 target gene loci (Figure 3A) and DPY30 emerged as having higher SHAP importance at SET1A and ZC3H4 target gene loci (Figure 3B). This finding aligns with the expectation for DPY30 and RBBP5 target genes, given their well-characterized roles as core components of the SET1 complex[27] and their contributions to SET1A-mediated transcriptional regulation. However, the intriguing observation of increased SHAP importance for SET1A and ZC3H4 in the INTS11 degron gene sets (Figure 3A) and increased SHAP importance of INTS11 in SET1A and ZC3H4 gene sets (Figure 3B) was unexpected.

To probe this potential connection further, we analyzed the overlap of target genes between INTS11 and ZC3H4. This analysis revealed that INTS11 and ZC3H4 share a significant number of target loci (Figure 3C), with highly correlated transcriptional regulatory patterns (Figure 3D) suggesting a functional relationship between these two proteins. Existing evidence supports the notion that transcriptional machineries often co-occupy the same loci[49], influencing one another's activity. INTS11 is reported to be recruited by H3K4me3[22] which could explain this unexpected overlap. This reciprocal analysis further reinforced the robustness of SHAP-based interpretations, demonstrating their ability to identify transcriptional regulatory mechanisms consistently across diverse experimental conditions.

Integrative modelling retains SHAP evaluations

To further evaluate the robustness of SHAP evaluations, we combined ChIP-seq datasets from Hughes and Wang studies[21, 22] to create a unified data set that includes profiles of key regulatory chromatin proteins and modifications, including SET1A, ZC3H4, INTS11, H3K4me3, DPY30 and RBBP5. This integrative approach allowed us to assess whether SHAP-derived feature importance remains consistent across multiple studies and to probe for additional combinatorial effects in modelling transcriptional regulation.

SHAP analysis consistently identified SET1A promoter occupancy as the most important feature for target genes across all degradation experiments (Figure S4A, B, C, D, E, F), reinforcing its critical role in transcriptional regulation. This trend was observed regardless of the specific perturbation, suggesting SET1A's crucial importance in maintaining RNA Pol–II occupancy. Following SET1A, ZC3H4 promoter occupancy emerged as the next key contributor to transcriptional regulation at target loci. This result further supports the cooperative role of SET1A and ZC3H4 at promoters[21]. While INTS11 promoter occupancy remained prominent for target genes, its median SHAP importance value was consistently lower than that of SET1A, ZC3H4, and DPY30. This observation suggests that although INTS11 plays a significant role, its importance is secondary to SET1A and ZC3H4 in embryonic stem cells under these conditions. ZC3H4 gene body occupancy continued to show notable SHAP importance for target genes. This finding highlights ZC3H4's dual role in transcriptional

regulation, functioning both at promoters and along gene bodies, where it may influence transcription elongation.

SHAP importance predicts direct targets and correlates with magnitude of differential gene expression

We next sought to determine whether SHAP importance values could better predict transcriptional changes compared to normalized ChIP signals. To evaluate this, genes were ranked based on their summed SHAP importance values for highly relevant marks (SET1A, INTS11, ZC3H4, DPY30 and H3K4me3) and compared to rankings based on normalized ChIP signal of corresponding proteins or marks. Our analysis revealed that genes ranked by SHAP importance consistently exhibited a higher fraction of differentially expressed genes than those ranked by normalized ChIP signals (Figure 4A). Additionally, SHAP importance values showed stronger correlations with differential gene expression compared to correlations with normalized ChIP signal values (Figure 4B). Importantly, SHAP values were consistently higher for target genes compared to non-target genes, as demonstrated in Figure S4. We argue that the analysis of SHAP values from relevant marks, where SHAP values are significantly higher for target genes compared to non-target genes, is analogous to assessing the functional impact from context-dependent influences.

Discussion

A key strength of this study lies in the ability of SHAP analysis to identify biologically meaningful relationships directly from data derived from unperturbed cell systems. The model, trained solely on unperturbed conditions, captures the relationships between protein occupancy and transcriptional output without prior knowledge of perturbations. Remarkably, by ranking genes based on SHAP importance values, direct targets of perturbation can be accurately predicted. This represents a potentially significant achievement, enabling context-dependent inference and complex interplay of regulatory targets without the need for time–consuming and expensive perturbation experiments. We also evaluated differences in SHAP–derived importance for target gene sets derived from perturbation experiments to infer regulatory roles of individual proteins. For instance, the higher SHAP importance of ZC3H4 at promoters and gene bodies of target genes suggests distinct regulatory roles for ZC3H4 and SET1A, highlighting how the model trained on unperturbed systems can reveal insights under perturbed conditions. This observation aligns with recent experimental evidence that ZC3H4 recruits the PNUTS phosphatase[51], and directly controls the phosphorylation of proteins involved in regulating pause release and elongation[52, 53], providing further insight into ZC3H4-mediated transcriptional regulation. In parallel, the consistent importance of INTS11 promoter occupancy, even in target genes following DPY30 and RBBP5 depletion, highlights its critical role in transcriptional regulation via H3K4me3 recruitment[22]. Cross-dataset validation and integrative modelling reinforced the utility of SHAP-based interpretations, revealing unexpected overlaps between ZC3H4 and INTS11 target loci. This functional connection, supported by shared transcriptional regulatory patterns, raises the intriguing possibility of cooperation between the

Restrictor and Integrator complexes (Figure 5), potentially converging on a common transcriptional termination pathway[49]. Whether the recruitment of INTS11 via H3K4me3[22] or the antagonistic interplay between ZC3H4 and SET1A[21], which deposits H3K4me3, stimulates a cascade of transcriptional regulation at the same genes remains an open question. Future experiments will be essential to unravel the precise mechanisms underlying these interactions and to determine whether they act sequentially, interdependently, or as part of a larger coordinated regulatory pathway.

The analysis of PRC1 and PRC2 regulatory systems highlights both the strengths and challenges of SHAP in disentangling contributions of co–occupying proteins. SHAP consistently identified RING1B, the core PRC1 subunit, as a key contributor to transcriptional repression, aligning with research that established PRC1's dominant role at early time points[20, 37, 38]. This result demonstrates SHAP's robustness in capturing biologically meaningful relationships, even in highly co-linear systems. The predominant contribution of EZH2 at some time points signifies the complex interplay between PRC1 and PRC2[25, 40–42]. Notably, SHAP's ability to identify H3K4me3 importance at polycomb-repressed loci suggests potential recruitment of the SET1 complex, uncovering new avenues for understanding the interplay between activation and repression machinery[45]. It could also indicate identification of "bivalent" promoters[46, 47], which are a peculiar feature of embryonic stem cells. Overall, these results reaffirm SHAP's utility as a powerful tool for identifying key regulators and generating novel biological hypotheses, while acknowledging the challenges posed by systems with overlapping regulatory inputs.

Our case study highlights the significant advantages of SHAP in providing gene–by–gene explanations for individual predictions, enabling insights into transcriptional regulation that would be difficult to achieve using traditional methods. By integrating predictive modelling with explainable AI, we offer a scalable framework capable of exploring both well-characterized and poorly understood transcriptional processes. When paired with complementary experimental approaches, such as time-resolved transcriptional measurements or orthogonal validation assays, SHAP can refine its insights further, providing a powerful tool for decoding complex regulatory networks. Overall, this study underscores the transformative potential of combining computational modelling with explainable AI to advance our understanding of transcriptional regulation and generate novel biological hypotheses.

Materials and methods

Preprocessing of sequencing data

All bioinformatic analyses were performed using Snakemake workflows[54], ensuring reproducibility and scalability of the analysis pipeline. These workflows include detailed specifications of software versions, command-line parameters, and dependencies for each analysis step. The complete workflows are publicly available on GitHub at `https://github.com/kashyapchhatbar/SHAP-analysis`.

ChIP-seq data analysis

Raw sequencing data for ChIP-seq experiments were downloaded from the Gene Expression Omnibus (GEO). Reads were aligned to the reference genome using Bowtie2[55], ensuring high alignment accuracy. To improve data quality, reads mapping to ENCODE blacklisted regions, which are repetitive and highly mappable, were removed. Peak calling and signal track generation were performed with MACS2[56]. Treatment and control pileup tracks were generated, and the `bdgcmp` command was used to calculate fold enrichment tracks. To quantify the signal, bigWigAverageOverBed[57] was used to calculate the mean signal across defined genomic regions, such as promoters and gene bodies.

RNA-seq and TT-seq data analysis

Raw sequencing data for RNA-seq and TT-seq experiments were also downloaded from GEO. Reads were first aligned to the transcriptome using STAR[58] with a two-pass mapping approach to ensure accurate alignment. Reads that failed to map to the transcriptome were subsequently aligned to the reference genome using Bowtie2[55]. PCR duplicates were identified and removed using SAMtools markup[59] to eliminate artifacts due to library preparation. Gene-level read counts were quantified using featureCounts[60] from the Subread package. Differential gene expression analysis was performed using DESeq2[61] in R, applying default parameters to identify significantly differentially expressed genes.

ENCODE RNA-seq

Transcripts per million (TPM) values from polyA RNA-seq experiments in mouse embryonic stem cells were obtained from the ENCODE database[2].

Sequencing data sets

Table 1. Sequencing data sets considered for interpretable modelling

Accession	Source	Assays	ChIP antibodies
GSE199805[21]	mESC	ChIP-seq, TT-seq	SET1A, ZC3H4, H3K4me3, Pol–II
GSE181714[22]	mESC	ChIP-seq, TT-seq	INTS11, DPY30, NELFA, H3K4me3, CDK9, AFF4, HEXIM1, BRD4, RBBP5, Pol–II
GSE159400[20]	mESC	ChIP-seq, RNA-seq	RING1B, SUZ12, H2AK119ub1, H3K27me3, H3K4me3, Pol–II

Data preparation and modeling

All modeling analyses, including data preparation, MLP regression, SHAP analysis, and linear regression, were performed using Python[62]. The source code is publicly available on the GitHub repository `https://github.com/kashyapchhatbar/SHAP-analysis`.

Data preparation

Fold enrichment values derived from ChIP-seq data were summarized for promoters and gene bodies using bigWigAverageOverBed[57]. These values were log2-transformed and standardized to have a mean of 0 and a standard deviation of 1. Similarly, transcripts per million (TPM) values from polyA RNA-seq experiments in mouse embryonic stem cells, obtained from the ENCODE database, were log10-transformed and standardized with a mean of 0 and a standard deviation of 1. Depending on the prediction variable (e.g., Pol-II occupancy or mRNA levels), the input and output data were split into training, testing, and validation datasets using Scikit-learn's data-splitting functionality.

Linear models

Linear regression models were fitted separately for perturbed and unperturbed genes to analyze the relationships between input features and the prediction variables. Regression coefficients for all features were compared between the two groups.

Deep learning models

A multilayer perceptron (MLP) regressor was employed to model the data. Hyperparameter tuning, including optimization of the number of hidden layers and selection of the best-performing activation function, was conducted using Optuna. Model performance was assessed using the R^2 score from Scikit-learn[62]. All trained model parameters were stored, and the best-performing model was selected for SHAP analysis to interpret feature importance and contributions.

SHAP analysis

SHAP values were calculated using the KernelExplainer[17] method to interpret the contributions of individual features to model predictions. A comparison between SHAP values of features for perturbed and unperturbed genes was performed to test whether the model accurately captures biologically meaningful patterns and responses to perturbations.

Figures

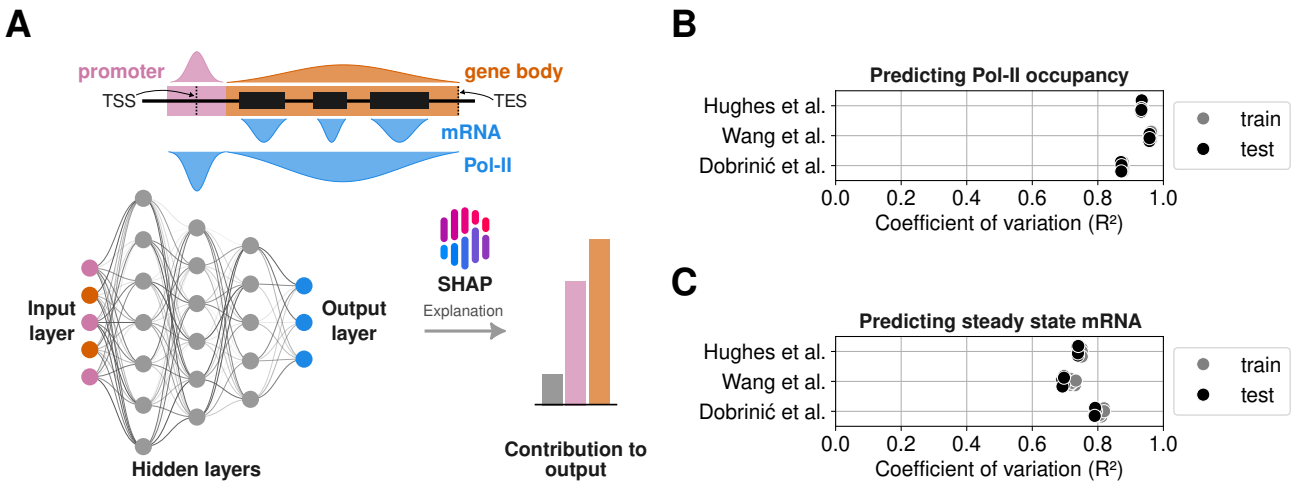

Figure 1. Overview of predictive modeling for RNA Pol-II occupancy and mRNA levels. (A) Schematic of the multilayer perceptron (MLP) model architecture used to predict RNA Pol-II occupancy and steady-state mRNA levels based on chromatin-associated protein profiles. Input features include protein occupancy data from ChIP-seq experiments, and outputs represent predicted transcriptional outcomes. (B) Model performance for predicting RNA Pol-II occupancy based on chromatin-associated protein profiles across datasets. Coefficients of determination (R^2) are shown for both training and test datasets. (C) Model performance for predicting steady-state mRNA levels based on chromatin-associated protein profiles across datasets. Coefficients of determination (R^2) are shown for both training and test datasets.

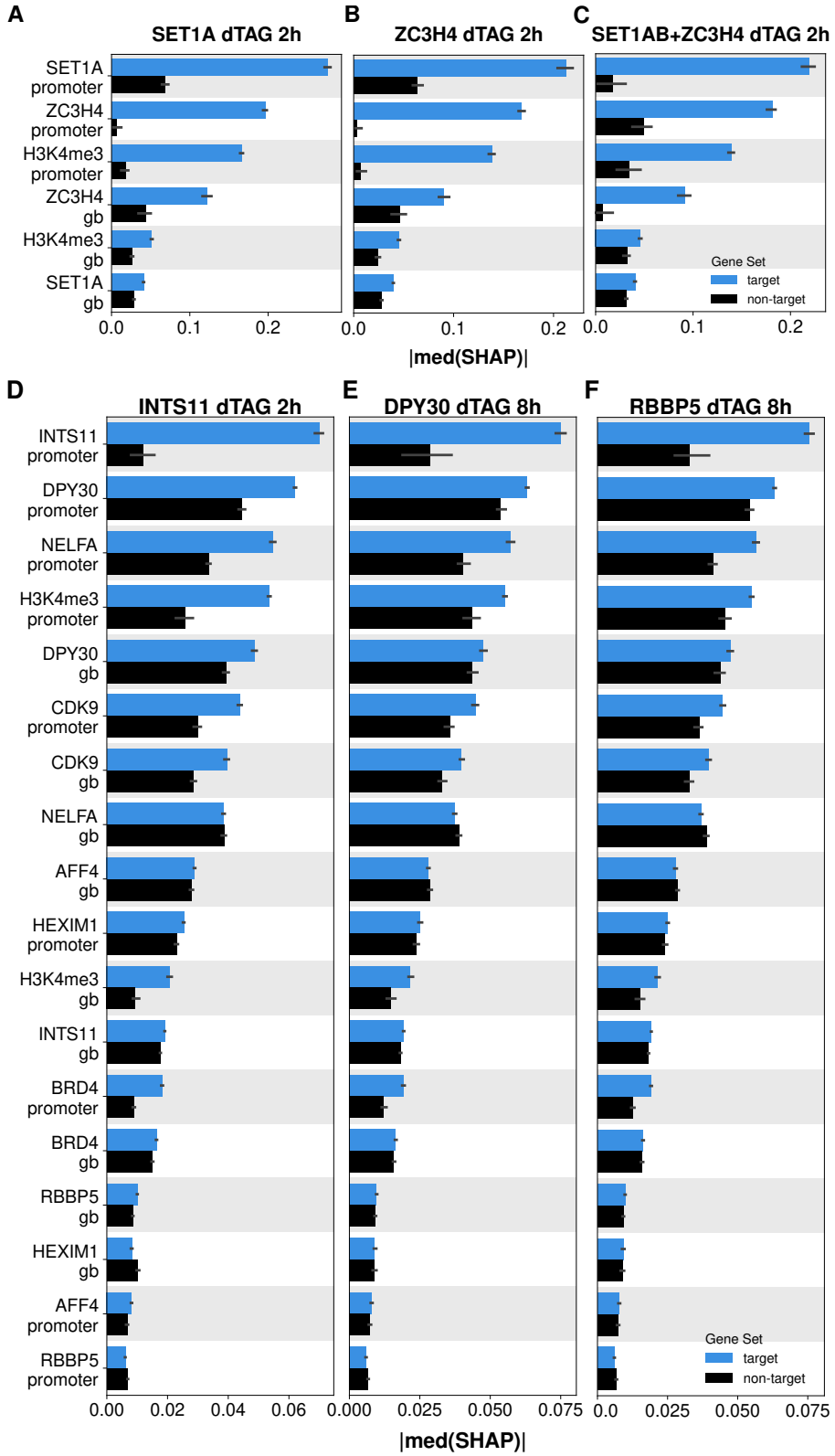

Figure 2. SHAP analysis reveals cooperative roles of SET1A and ZC3H4 in transcriptional regulation. INTS11 is a key regulator for target gene loci following DPY30 and RBBP5 degradation. Median SHAP values for promoter and gene body (gb) occupancy calculated for target and non-target genes. (A) SHAP analysis following acute degradation of SET1A. (B) SHAP analysis following acute degradation of ZC3H4. (C) SHAP analysis following combined degradation of SET1A and ZC3H4. (D) SHAP analysis for genes following INTS11 degradation. (E) SHAP analysis for genes following DPY30 degradation. (F) SHAP analysis for genes following RBBP5 degradation.

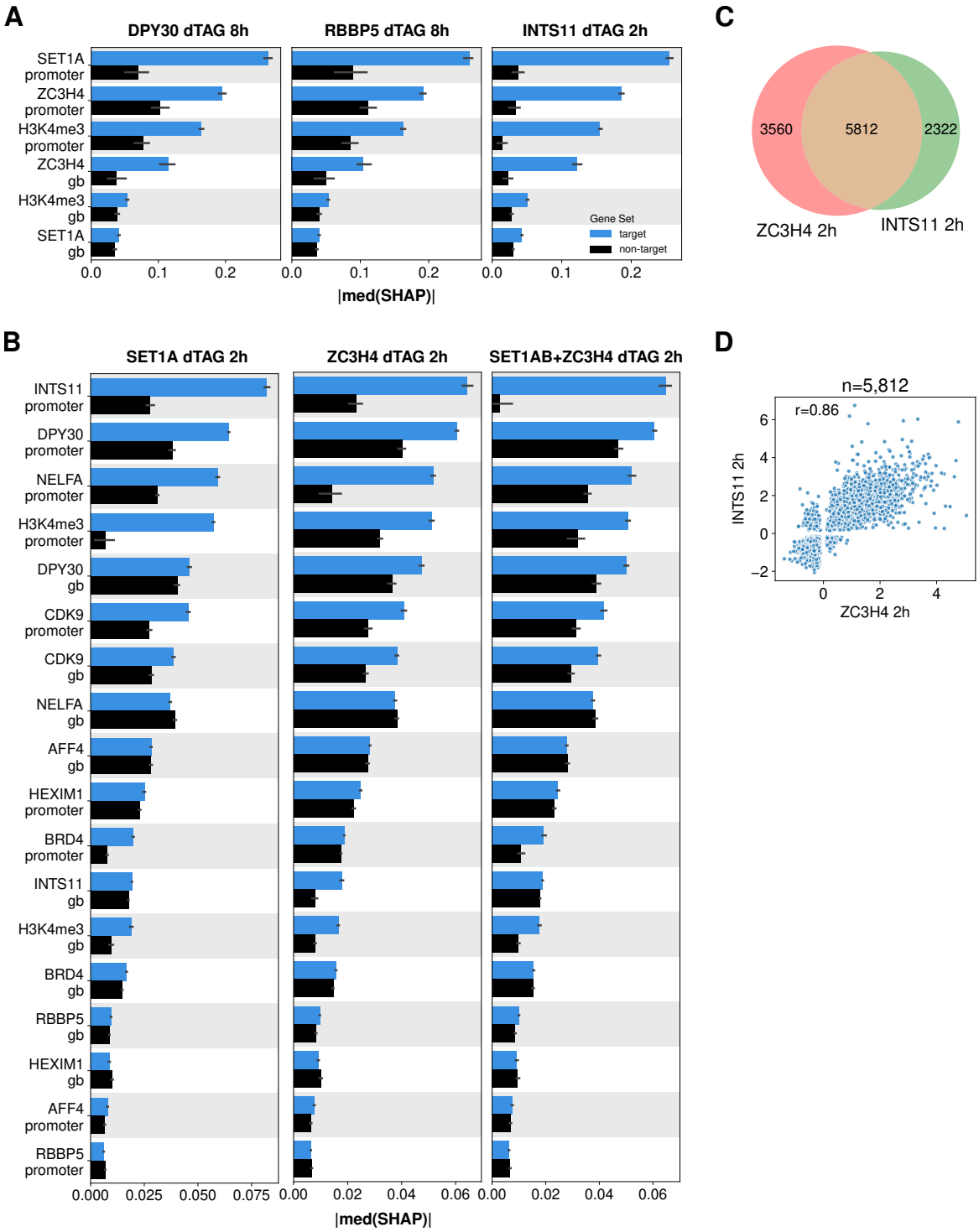

Figure 3. Cross-dataset SHAP analysis reveals unexpected overlaps between ZC3H4 and INTS11 target genes. (A) SHAP importance values for SET1A, ZC3H4, and H3K4me3 at target gene loci following DPY30, RBBP5, and INTS11 degradation. (B) SHAP importance values for INTS11 and related factors at SET1A- and ZC3H4-target gene loci. (C) Overlap of target genes after degradation of ZC3H4 and INTS11. (D) Correlation between ZC3H4 and INTS11 transcriptional regulatory patterns.

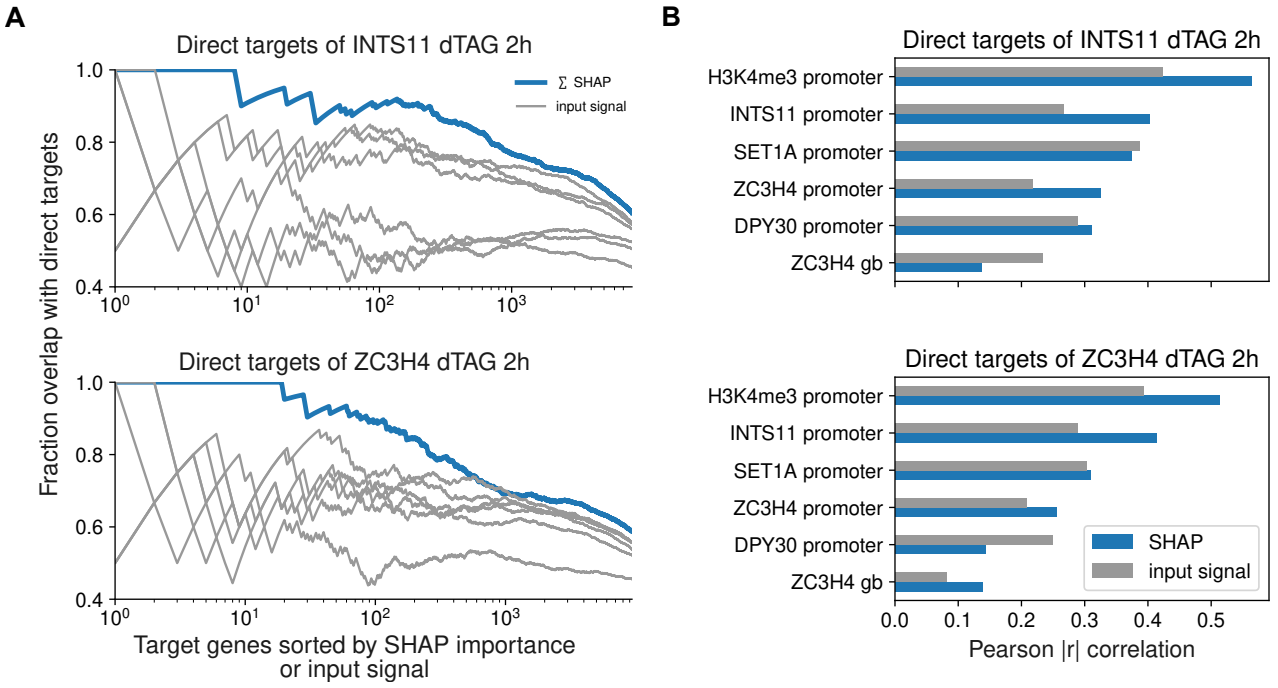

Figure 4. SHAP-based analysis identifies direct targets and predicts transcriptional changes with higher accuracy. (A) Overlap of genes ranked by summed SHAP importance values for highly relevant marks (SET1A, INTS11, ZC3H4, DPY30, H3K4me3) with target genes following INTS11 and ZC3H4 depletion. (B) Pearson correlations with the magnitude of differential gene expression of SHAP values and raw input signals.

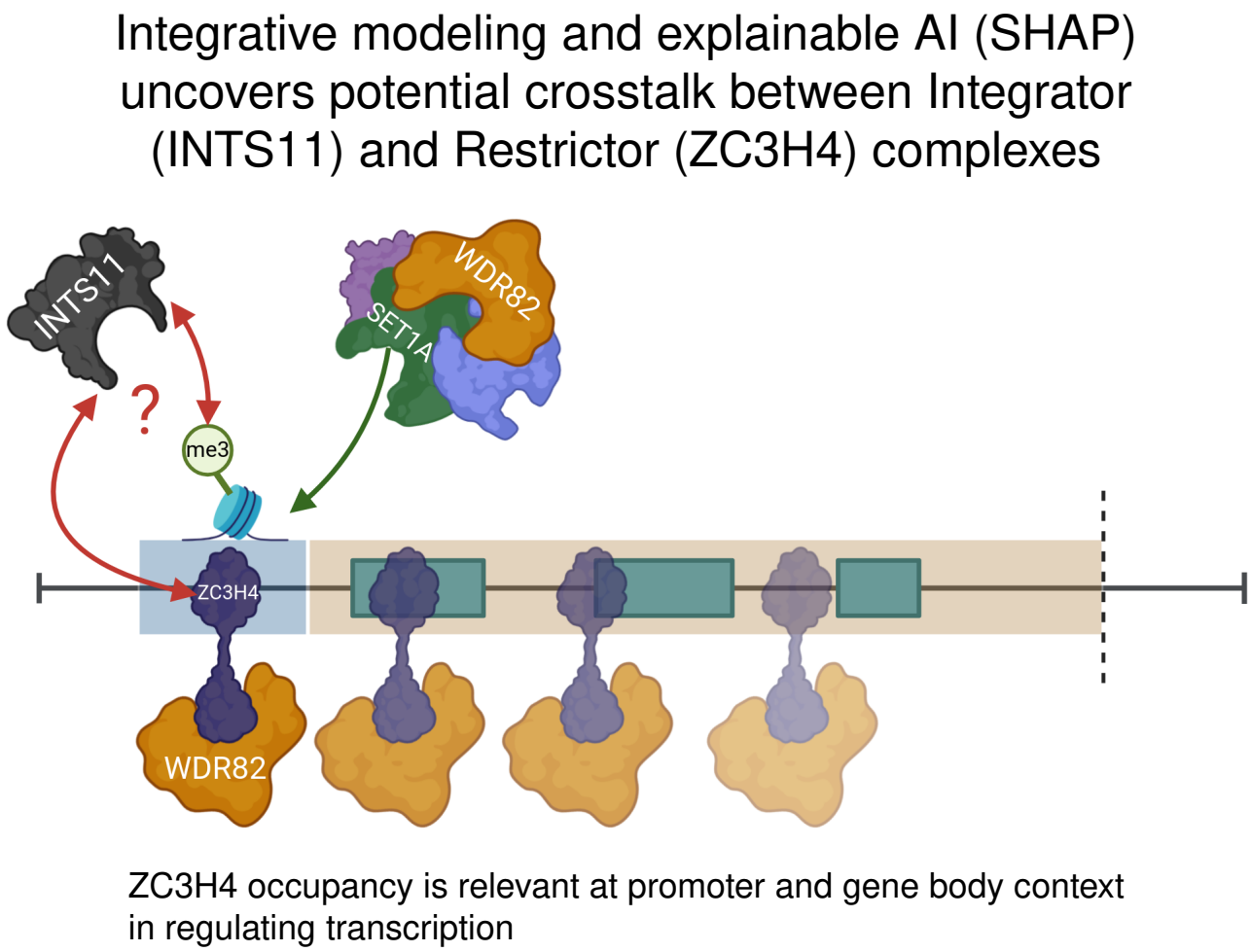

Figure 5. Potential crosstalk between Integrator and Restrictor complexes in transcriptional regulation

Supplemental Figures

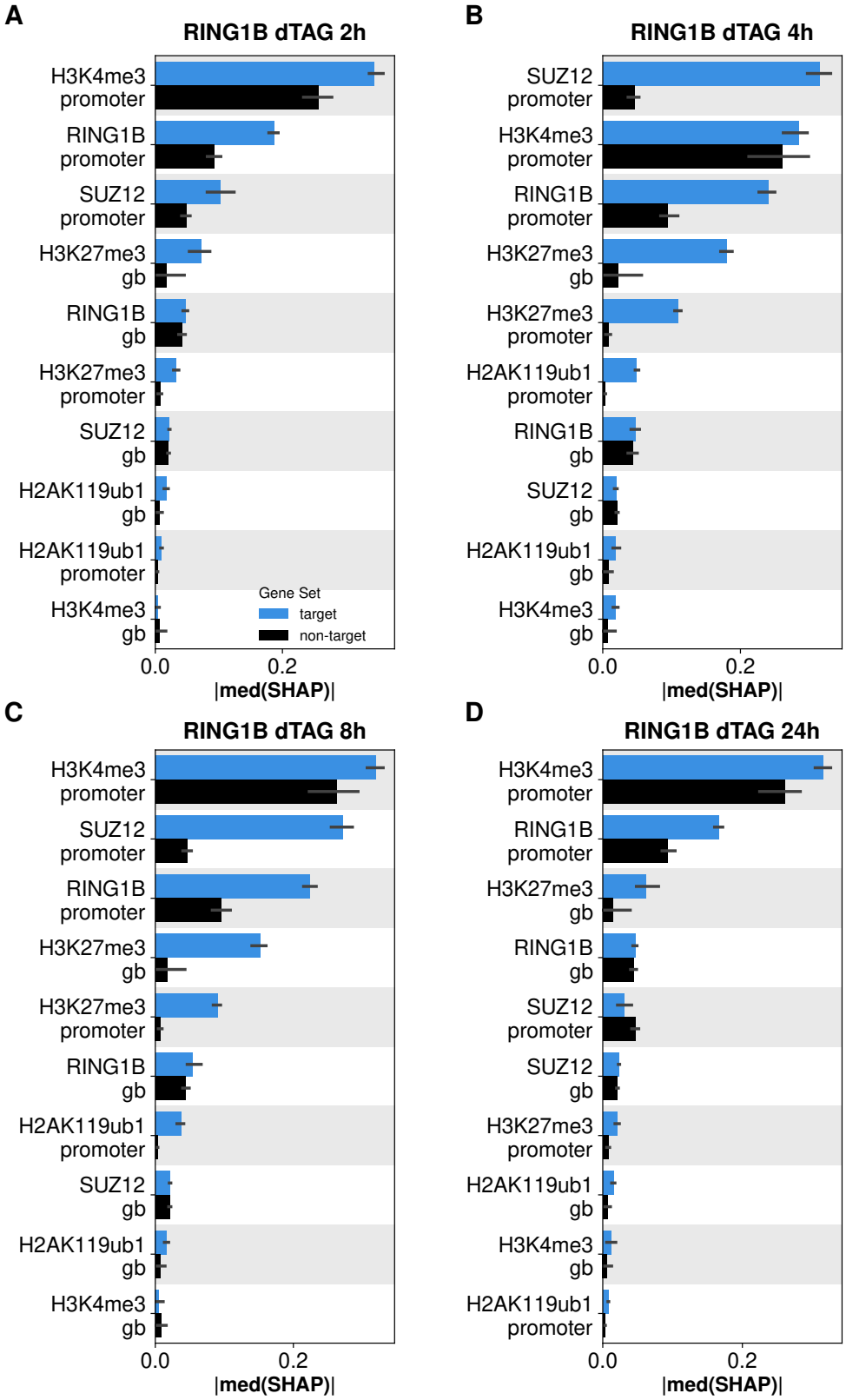

Figure S1. SHAP values for co-occupying PRC1 (RING1B) and PRC2 (SUZ12) components across multiple time points following RING1B degradation. (A-D) Median SHAP importance for transcriptional regulators at 2h, 4h, 8h, and 24h post-RING1B degradation.

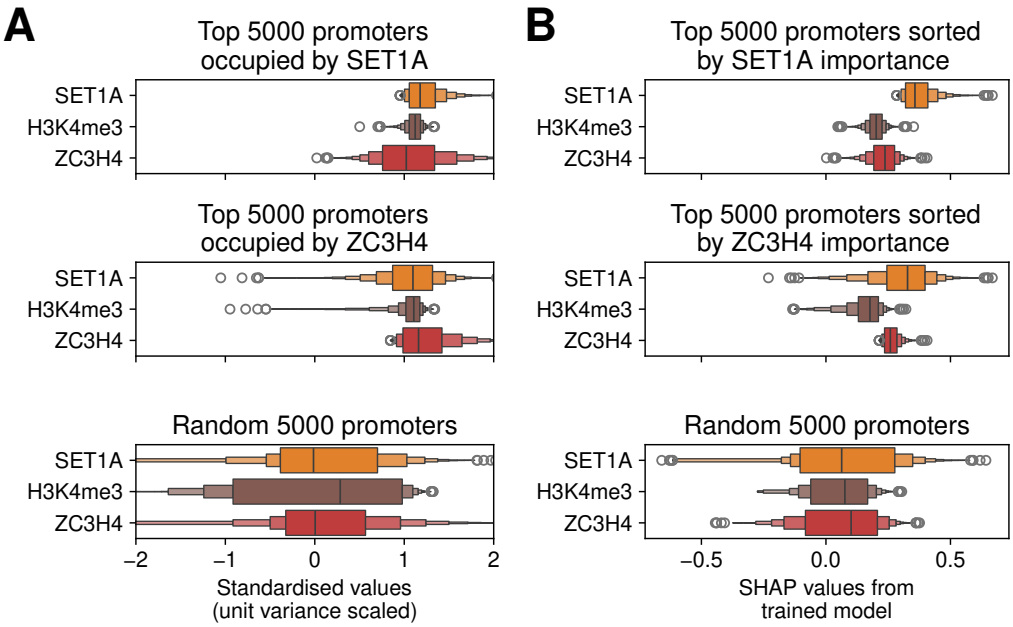

Figure S2. Promoter Occupancy and SHAP values of SET1A, H3K4me3 and ZC3H4. (A) Summary statistics of ChIP-seq occupancy profiles for SET1A, H3K4me3, and ZC3H4 at the top 5000 promoters ranked by SET1A and ZC3H4 occupancy, as well as at 5000 randomly selected promoters. (B) SHAP importance values for SET1A, H3K4me3, and ZC3H4 at the top 5000 promoters ranked by SET1A importance, ZC3H4 importance, and at 5000 randomly selected promoters.

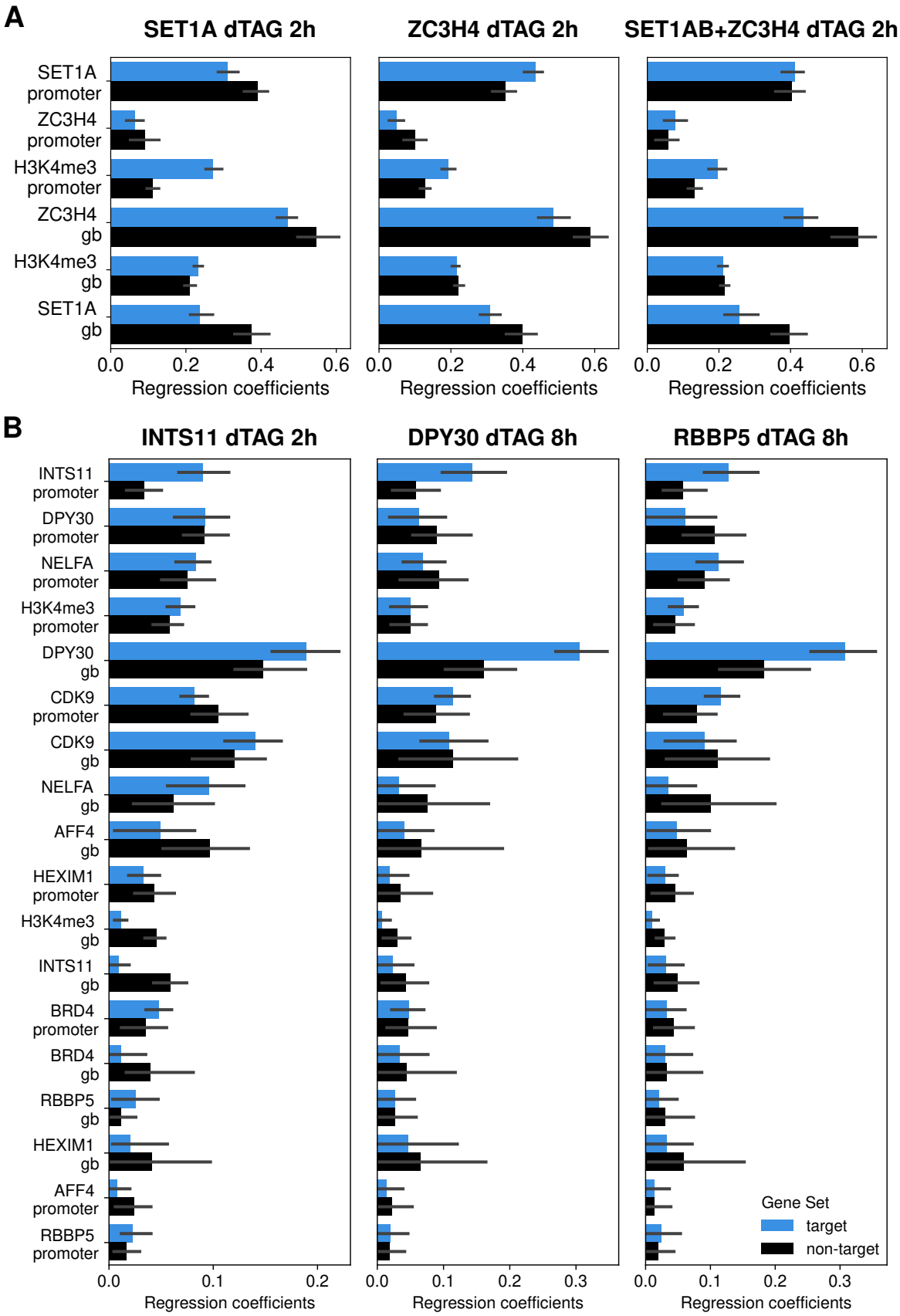

Figure S3. Linear Regression coefficients for chromatin-associated proteins in predicting RNA Pol-II occupancy at perturbed and unperturbed gene loci. (A) Coefficients for genes affected by SET1A degradation, ZC3H4 degradation and combined degradation of SET1A and ZC3H4. (B) Coefficients for genes affected by INTS11 degradataion, DPY30 degradatiaon and RBBP5 degradation.

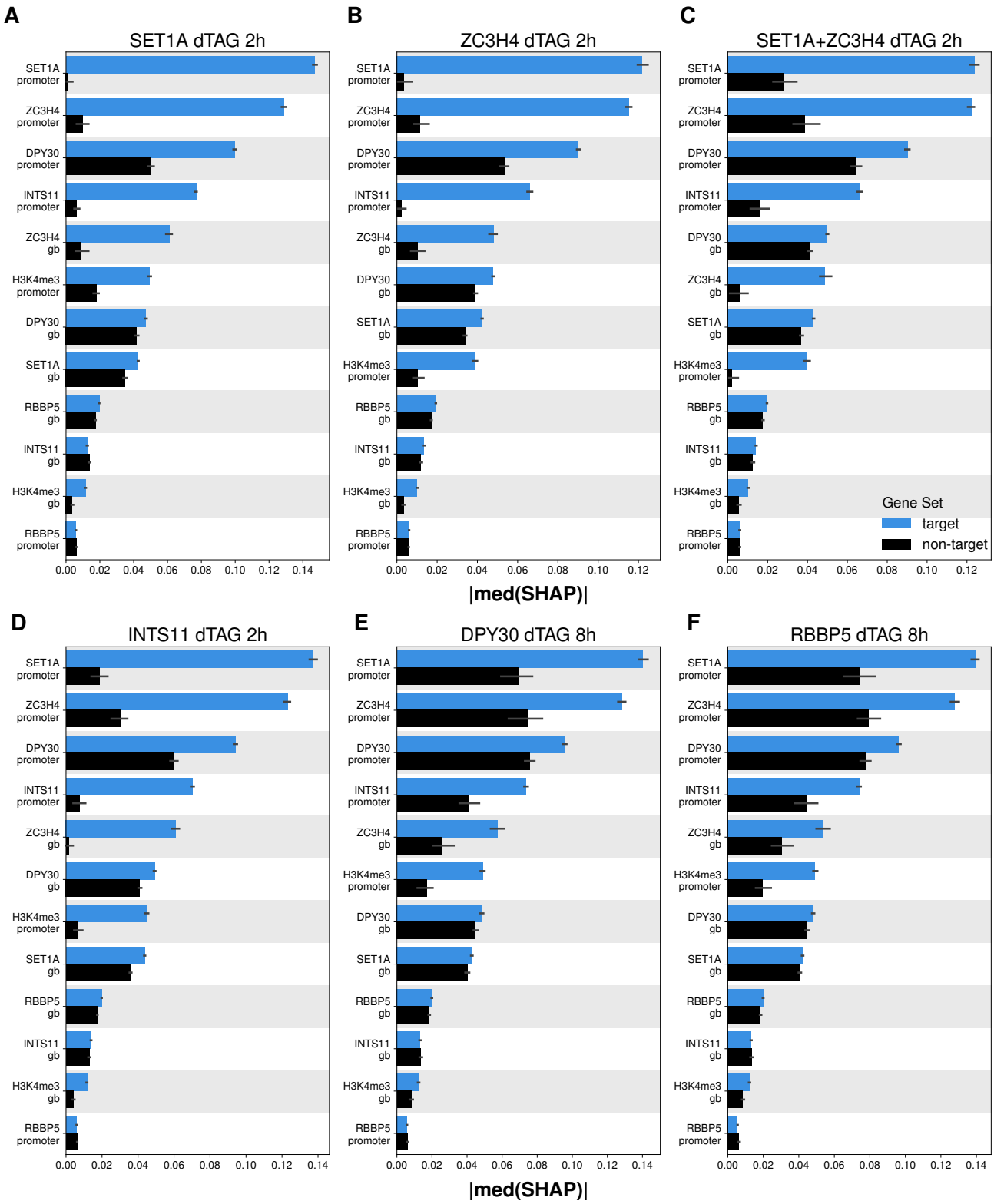

Figure S4. Integrated SHAP analysis highlights key contributors across multiple perturbation experiments. Median SHAP values for promoter and gene body (gb) occupancy of SET1A, ZC3H4, INTS11, DPY30, RBBP5, and H3K4me3, evaluated across target and non-target gene sets following degron-based experiments. (A) SHAP analysis after SET1A degradation at 2 hours (dTAG 2h). (B) SHAP analysis following ZC3H4 degradation at 2 hours (dTAG 2h). (C) SHAP analysis after combined degradation of SET1A and ZC3H4 (dTAG 2h) (D) SHAP analysis after INTS11 degradation (dTAG 2h). (E) SHAP analysis after DPY30 degradation at 8 hours (dTAG 8h). (F) SHAP analysis following RBBP5 degradation at 8 hours (dTAG 8h).

References

[1] Mark Ptashne and Alexander Gann. "Imposing specificity by localization: mechanism and evolvability". In: *Current Biology* 8.22 (Nov. 5, 1998). Publisher: Elsevier, R812–R822. ISSN: 0960-9822. DOI: 10.1016/S0960-9822(07)00508-8.

[2] Ian Dunham et al. "An integrated encyclopedia of DNA elements in the human genome". In: *Nature* 489.7414 (Sept. 2012). Publisher: Nature Publishing Group, pp. 57–74. ISSN: 1476-4687. DOI: 10.1038/nature11247.

[3] Joost H. A. Martens and Hendrik G. Stunnenberg. "BLUEPRINT: mapping human blood cell epigenomes". In: *Haematologica* 98.10 (Oct. 1, 2013). Number: 10, pp. 1487–1489. ISSN: 1592-8721. DOI: 10.3324/haematol.2013.094243.

[4] Hendrik G. Stunnenberg et al. "The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery". In: *Cell* 167.5 (Nov. 17, 2016). Publisher: Elsevier, pp. 1145–1149. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2016.11.007.

[5] Dan Benveniste et al. "Transcription factor binding predicts histone modifications in human cell lines". In: *Proceedings of the National Academy of Sciences* 111.37 (Sept. 16, 2014). Publisher: Proceedings of the National Academy of Sciences, pp. 13367–13372. DOI: 10.1073/pnas.1412081111.

[6] Michael A. Beer and Saeed Tavazoie. "Predicting Gene Expression from Sequence". In: *Cell* 117.2 (Apr. 16, 2004), pp. 185–198. ISSN: 0092-8674. DOI: 10.1016/S0092-8674(04)00304-6.

[7] David M. Budden, Daniel G. Hurley, and Edmund J. Crampin. "Predictive modelling of gene expression from transcriptional regulatory elements". In: *Briefings in Bioinformatics* 16.4 (July 1, 2015), pp. 616–628. ISSN: 1467-5463. DOI: 10.1093/bib/bbu034.

[8] Bartek Wilczynski et al. "Predicting Spatial and Temporal Gene Expression Using an Integrative Model of Transcription Factor Occupancy and Chromatin State". In: *PLOS Computational Biology* 8.12 (Dec. 6, 2012). Publisher: Public Library of Science, e1002798. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1002798.

[9] Zhengqing Ouyang, Qing Zhou, and Wing Hung Wong. "ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells". In: *Proceedings of the National Academy of Sciences* 106.51 (Dec. 22, 2009). Publisher: Proceedings of the National Academy of Sciences, pp. 21521–21526. DOI: 10.1073/pnas.0904863106.

[10] Chao Cheng and Mark Gerstein. "Modeling the relative relationship of transcription factor binding and histone modifications to gene expression levels in mouse embryonic stem cells". In: *Nucleic Acids Research* 40.2 (Jan. 1, 2012), pp. 553–568. ISSN: 0305-1048. DOI: 10.1093/nar/gkr752.

[11] Chao Cheng et al. "A statistical framework for modeling gene expression using chromatin features and application to modENCODE datasets". In: *Genome Biology* 12.2 (Feb. 16, 2011), R15. ISSN: 1474-760X. DOI: 10.1186/gb-2011-12-2-r15.

[12] Christof Angermueller et al. "DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning". In: *Genome Biology* 18.1 (Apr. 11, 2017), p. 67. ISSN: 1474-760X. DOI: 10.1186/s13059-017-1189-z.

[13] Jacob Schreiber et al. "Avocado: a multi-scale deep tensor factorization method learns a latent representation of the human epigenome". In: *Genome Biology* 21.1 (Mar. 30, 2020), p. 81. ISSN: 1474-760X. DOI: 10.1186/s13059-020-01977-6.

[14] Timothy J. Durham et al. "PREDICTD PaRallel Epigenomics Data Imputation with Cloud-based Tensor Decomposition". In: *Nat Commun* 9.1 (Apr. 11, 2018). Publisher: Nature Publishing Group, p. 1402. ISSN: 2041-1723. DOI: 10.1038/s41467-018-03635-9.

[15] Jason Ernst and Manolis Kellis. "Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues". In: *Nat Biotechnol* 33.4 (Apr. 2015). Publisher: Nature Publishing Group, pp. 364–376. ISSN: 1546-1696. DOI: 10.1038/nbt.3157.

[16] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ""Why Should I Trust You?": Explaining the Predictions of Any Classifier". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and*

Data Mining. KDD '16. New York, NY, USA: Association for Computing Machinery, Aug. 13, 2016, pp. 1135–1144. ISBN: 978-1-4503-4232-2. DOI: 10.1145/2939672.2939778.

[17] Scott M Lundberg and Su-In Lee. "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc., 2017.

[18] Grégoire Montavon et al. "Layer-Wise Relevance Propagation: An Overview". In: *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Ed. by Wojciech Samek et al. Cham: Springer International Publishing, 2019, pp. 193–209. ISBN: 978-3-030-28954-6. DOI: 10.1007/978-3-030-28954-6_10.

[19] Martin G. Jaeger and Georg E. Winter. "Fast-acting chemical tools to delineate causality in transcriptional control". In: *Molecular Cell* 81.8 (Apr. 15, 2021). Publisher: Elsevier, pp. 1617–1630. ISSN: 1097-2765. DOI: 10.1016/j.molcel.2021.02.015.

[20] Paula Dobrinić, Aleksander T. Szczurek, and Robert J. Klose. "PRC1 drives Polycomb-mediated gene repression by controlling transcription initiation and burst frequency". In: *Nat Struct Mol Biol* 28.10 (Oct. 2021). Publisher: Nature Publishing Group, pp. 811–824. ISSN: 1545-9985. DOI: 10.1038/s41594-021-00661-y.

[21] Amy L. Hughes et al. "A CpG island-encoded mechanism protects genes from premature transcription termination". In: *Nat Commun* 14.1 (Feb. 9, 2023). Publisher: Nature Publishing Group, p. 726. ISSN: 2041-1723. DOI: 10.1038/s41467-023-36236-2.

[22] Hua Wang et al. "H3K4me3 regulates RNA polymerase II promoter-proximal pause-release". In: *Nature* 615.7951 (Mar. 2023). Publisher: Nature Publishing Group, pp. 339–348. ISSN: 1476-4687. DOI: 10.1038/s41586-023-05780-8.

[23] Mariana de Napoles et al. "Polycomb Group Proteins Ring1A/B Link Ubiquitylation of Histone H2A to Heritable Gene Silencing and X Inactivation". In: *Developmental Cell* 7.5 (Nov. 1, 2004), pp. 663–676. ISSN: 1534-5807. DOI: 10.1016/j.devcel.2004.10.005.

[24] Hengbin Wang et al. "Role of histone H2A ubiquitination in Polycomb silencing". In: *Nature* 431.7010 (Oct. 2004). Publisher: Nature Publishing Group, pp. 873–878. ISSN: 1476-4687. DOI: 10.1038/nature02985.

[25] Ru Cao et al. "Role of Histone H3 Lysine 27 Methylation in Polycomb-Group Silencing". In: *Science* 298.5595 (Nov. 2002). Publisher: American Association for the Advancement of Science, pp. 1039–1043. DOI: 10.1126/science.1076997.

[26] Jürg Müller et al. "Histone Methyltransferase Activity of a *Drosophila* Polycomb Group Repressor Complex". In: *Cell* 111.2 (Oct. 18, 2002), pp. 197–208. ISSN: 0092-8674. DOI: 10.1016/S0092-8674(02)00976-5.

[27] Trissa Miller et al. "COMPASS: A complex of proteins associated with a trithorax-related SET domain protein". In: *Proceedings of the National Academy of Sciences* 98.23 (Nov. 6, 2001). Publisher: Proceedings of the National Academy of Sciences, pp. 12902–12907. DOI: 10.1073/pnas.231473398.

[28] Chris Estell et al. "ZC3H4 restricts non-coding transcription in human cells". In: *eLife* 10 (Apr. 29, 2021). Ed. by Torben Heick Jensen, James L Manley, and Joan Steitz. Publisher: eLife Sciences Publications, Ltd, e67305. ISSN: 2050-084X. DOI: 10.7554/eLife.67305.

[29] David Baillat et al. "Integrator, a Multiprotein Mediator of Small Nuclear RNA Processing, Associates with the C-Terminal Repeat of RNA Polymerase II". In: *Cell* 123.2 (Oct. 21, 2005), pp. 265–276. ISSN: 0092-8674. DOI: 10.1016/j.cell.2005.08.019.

[30] Chad B. Stein et al. "Integrator endonuclease drives promoter-proximal termination at all RNA polymerase II-transcribed loci". In: *Molecular Cell* 82.22 (Nov. 17, 2022). Publisher: Elsevier, 4232–4245.e11. ISSN: 1097-2765. DOI: 10.1016/j.molcel.2022.10.004.

[31] Saskia Gressel et al. "CDK9-dependent RNA polymerase II pausing controls transcription initiation". In: *eLife* 6 (Oct. 10, 2017). Ed. by Katherine A Jones. Publisher: eLife Sciences Publications, Ltd, e29736. ISSN: 2050-084X. DOI: 10.7554/eLife.29736.

[32] Shwu-Yuan Wu and Cheng-Ming Chiang. "The Double Bromodomain-containing Chromatin Adaptor Brd4 and Transcriptional Regulation*". In: *Journal of Biological Chemistry* 282.18 (May 4, 2007), pp. 13141–13145. ISSN: 0021-9258. DOI: 10.1074/jbc.R700001200.

[33] Jian Li et al. "Kinetic Competition between Elongation Rate and Binding of NELF Controls Promoter-Proximal Pausing". In: *Molecular Cell* 50.5 (June 6, 2013), pp. 711–722. ISSN: 1097-2765. DOI: 10.1016/j.molcel.2013.05.016.

[34] Hojoong Kwak and John T. Lis. "Control of Transcriptional Elongation". In: *Annual Review of Genetics* 47 (Volume 47, 2013 Nov. 23, 2013). Publisher: Annual Reviews, pp. 483–508. ISSN: 0066-4197, 1545-2948. DOI: 10.1146/annurev-genet-110711-155440.

[35] William L. Tong and Cengiz Pehlevan. *MLPs Learn In-Context on Regression and Classification Tasks.* Sept. 26, 2024. DOI: 10.48550/arXiv.2405.15618. arXiv: 2405.15618[cs].

[36] Hanxiao Liu et al. *Pay Attention to MLPs.* June 1, 2021. DOI: 10.48550/arXiv.2105.08050. arXiv: 2105.08050[cs].

[37] Neil P. Blackledge et al. "PRC1 Catalytic Activity Is Central to Polycomb System Function". In: *Molecular Cell* 77.4 (Feb. 20, 2020), 857–874.e9. ISSN: 1097-2765. DOI: 10.1016/j.molcel.2019.12.001.

[38] Neil P. Blackledge and Robert J. Klose. "The molecular principles of gene regulation by Polycomb repressive complexes". In: *Nat Rev Mol Cell Biol* 22.12 (Dec. 2021). Publisher: Nature Publishing Group, pp. 815–833. ISSN: 1471-0080. DOI: 10.1038/s41580-021-00398-y.

[39] Adrian P. Bracken et al. "Genome-wide mapping of Polycomb target genes unravels their roles in cell fate transitions". In: *Genes Dev.* 20.9 (May 1, 2006). Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab, pp. 1123–1136. ISSN: 0890-9369, 1549-5477. DOI: 10.1101/gad.381706.

[40] Neil P. Blackledge et al. "Variant PRC1 Complex-Dependent H2A Ubiquitylation Drives PRC2 Recruitment and Polycomb Domain Formation". In: *Cell* 157.6 (June 5, 2014), pp. 1445–1459. ISSN: 0092-8674. DOI: 10.1016/j.cell.2014.05.004.

[41] Jinrong Min, Yi Zhang, and Rui-Ming Xu. "Structural basis for specific binding of Polycomb chromodomain to histone H3 methylated at Lys 27". In: *Genes Dev.* 17.15 (Jan. 8, 2003). Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab, pp. 1823–1828. ISSN: 0890-9369, 1549-5477. DOI: 10.1101/gad.269603.

[42] Liangjun Wang et al. "Hierarchical Recruitment of Polycomb Group Silencing Complexes". In: *Molecular Cell* 14.5 (June 4, 2004), pp. 637–646. ISSN: 1097-2765. DOI: 10.1016/j.molcel.2004.05.009.

[43] Thomas Clouaire et al. "Cfp1 integrates both CpG content and gene activity for accurate H3K4me3 deposition in embryonic stem cells". In: *Genes Dev.* 26.15 (Aug. 1, 2012). Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab, pp. 1714–1728. ISSN: 0890-9369, 1549-5477. DOI: 10.1101/gad.194209.112.

[44] Anca M Farcas et al. "KDM2B links the Polycomb Repressive Complex 1 (PRC1) to recognition of CpG islands". In: *eLife* 1 (Dec. 18, 2012). Ed. by Kevin Struhl. Publisher: eLife Sciences Publications, Ltd, e00205. ISSN: 2050-084X. DOI: 10.7554/eLife.00205.

[45] Robert J. Klose et al. "Chromatin Sampling—An Emerging Perspective on Targeting Polycomb Repressor Proteins". In: *PLOS Genetics* 9.8 (Aug. 22, 2013). Publisher: Public Library of Science, e1003717. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1003717.

[46] Bradley E. Bernstein et al. "A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells". In: *Cell* 125.2 (Apr. 2006), pp. 315–326. ISSN: 00928674. DOI: 10.1016/j.cell.2006.02.041.

[47] Philipp Voigt, Wee-Wei Tee, and Danny Reinberg. "A double take on bivalent promoters". In: *Genes Dev.* 27.12 (June 15, 2013), pp. 1318–1338. ISSN: 0890-9369, 1549-5477. DOI: 10.1101/gad.219626.113.

[48] Ozgur Oksuz et al. "Capturing the Onset of PRC2-Mediated Repressive Domain Formation". In: *Molecular Cell* 70.6 (June 2018), 1149–1162.e5. ISSN: 10972765. DOI: 10.1016/j.molcel.2018.05.023.

[49] Chris Estell et al. "A restrictor complex of ZC3H4, WDR82, and ARS2 integrates with PNUTS to control unproductive transcription". In: *Molecular Cell* 83.13 (July 6, 2023). Publisher: Elsevier, 2222–2239.e5. ISSN: 1097-2765. DOI: 10.1016/j.molcel.2023.05.029.

[50] Ruichuan Chen et al. "Brd4 and HEXIM1: Multiple Roles in P-TEFb Regulation and Cancer". In: *BioMed Research International* 2014.1 (2014). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1155/2014/232870, p. 232870. ISSN: 2314-6141. DOI: 10.1155/2014/232870.

[51] Benjamin Erickson et al. *PP1 PNUTS binds the "restrictor" and dephosphorylates RNA pol II CTD Ser5 to stimulate transcription termination.* Pages: 2024.07.12.603302 Section: New Results. July 13, 2024. DOI: 10.1101/2024.07.12.603302.

[52] Jessica R. Kelley et al. "The PNUTS phosphatase complex controls transcription pause release". In: *Molecular Cell* 84.24 (Dec. 19, 2024). Publisher: Elsevier, 4843–4861.e8. ISSN: 1097-2765. DOI: 10.1016/j.molcel.2024.10.045.

[53] Zhenning Wang et al. "The phosphatase PP1 sustains global transcription by promoting RNA polymerase II pause release". In: *Molecular Cell* 84.24 (Dec. 19, 2024). Publisher: Elsevier, 4824–4842.e7. ISSN: 1097-2765. DOI: 10.1016/j.molcel.2024.10.046.

[54] Johannes Köster and Sven Rahmann. "Snakemake—a scalable bioinformatics workflow engine". In: *Bioinformatics* 28.19 (Oct. 1, 2012), pp. 2520–2522. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bts480.

[55] Ben Langmead and Steven L. Salzberg. "Fast gapped-read alignment with Bowtie 2". In: *Nat Methods* 9.4 (Apr. 2012). Publisher: Nature Publishing Group, pp. 357–359. ISSN: 1548-7105. DOI: 10.1038/nmeth.1923.

[56] Yong Zhang et al. "Model-based Analysis of ChIP-Seq (MACS)". In: *Genome Biology* 9.9 (Sept. 17, 2008), R137. ISSN: 1474-760X. DOI: 10.1186/gb-2008-9-9-r137.

[57] W. J. Kent et al. "BigWig and BigBed: enabling browsing of large distributed datasets". In: *Bioinformatics* 26.17 (Sept. 1, 2010), pp. 2204–2207. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btq351.

[58] Alexander Dobin et al. "STAR: ultrafast universal RNA-seq aligner". In: *Bioinformatics* 29.1 (Jan. 1, 2013), pp. 15–21. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bts635.

[59] Heng Li et al. "The Sequence Alignment/Map format and SAMtools". In: *Bioinformatics* 25.16 (Aug. 15, 2009), pp. 2078–2079. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btp352.

[60] Yang Liao, Gordon K. Smyth, and Wei Shi. "featureCounts: an efficient general purpose program for assigning sequence reads to genomic features". In: *Bioinformatics* 30.7 (Apr. 1, 2014), pp. 923–930. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btt656.

[61] Michael I. Love, Wolfgang Huber, and Simon Anders. "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". In: *Genome Biology* 15.12 (Dec. 5, 2014), p. 550. ISSN: 1474-760X. DOI: 10.1186/s13059-014-0550-8.

[62] Fabian Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12.85 (2011), pp. 2825–2830. ISSN: 1533-7928.