Investigating the genome-wide binding mechanisms of human CTCF protein using machine learning

Brendan Wang

(in collaboration with Gabriel Dolsten and Yuri Pritykin)

Summer 2023

Abstract

The human CTCF protein is a well characterized protein that plays a key role in establishing chromatin architecture and gene regulation in human cells. While the sequence motif occurrences of CTCF are identical across all cell types, previous work has suggested that CTCF binds to different regions within and across cell types. The mechanisms of the differential binding of CTCF to identical motif sequences within and across cell types remains to be further explored. This project investigates the impact various genomic factors have on CTCF binding, including DNA sequence, methylation, and other genomic features. We design a family of machine learning models that accurately model CTCF binding using sequence, methylation, and/or accessibility data as input. Through a series of ablation experiments, we demonstrate that adding immediate DNA sequence context flanking the CTCF motifs strongly increases model performance. Additionally, we find that adding methylation increases model performance although to a lesser extent. Collectively, these results suggest that region immediately surrounding the CTCF motifs contain essential sequence information that strongly captures the features conducive to CTCF binding that can be attributed to sequence and methylation.

Phase 1: Data Acquisition

We began by gathering the data we will use as input for our data analysis and experiments:

- **CTCF CHIP-seq:** To measure levels of CTCF binding across many cell types, we acquired many CTCF CHIP-seq bigwig files from the ENCODE database across 20 cell types. We explore these datasets together in Phase 2. For our machine learning experiments and analysis, we use 3 biological replicate experiments in tranverse colon cells. The output type of the three bigwigs were fold-change over control. These three datasets were generated by Bradley Bernstein form the Broad Institute.
- **Methylation CHIP-seq:** To measure levels of methylation, we use the whole-genome shotgun bisulfite sequencing (WGBS) CHIP-seq datasets. For our analysis, we used 2 WGBS datasets for tranverse colon cells from the ENCODE database. One dataset was for the plus strand; the other for the minus strand.
- **CTCF Motifs:** CTCF motifs were called using MEME and FIMO tools with the CTCF PWM from Jaspar and the human reference genome. In total, there were 25849 motif occurrences identified. Each CTCF motif occurrence was 19 base pairs long and had an associated p-value.

Phase 2: Exploratory Data Analysis

In this phase, we explore the data. The question we wanted to interested in answering was how CTCF "knows" to bind to different places in the genome. From this EDA step, we observed three key findings: 1) CTCF expresses differential binding in a group of identical motif occurrences or what we denote as a "motif clone" 2) at CTCF motif occurrences, there seems to be a negative correlation between

fraction of CpG sites and average methylation and 3) CTCF CHIP signal and motif methylation are negative correlated. We highlight some of the results below.

CTCF Expresses Differential Binding in Motif Clones

After gathering our data, we created a starting hypothesis:

- Null: if a CTCF motif occurrence is significant (i.e., maximizes the PWM function), it should always be bound (have a high CHIP signal) where the motif occurrence shows up within and between cell types. This implies CTCF binding is a function only of the sequence itself.
- Alternative: There exists certain significant CTCF motifs at which low levels of CTCF are bound. Explanations for this include transposable element activity, methylation, post-transcriptional modifications, etc.

To investigate this, we look into regions that only contained one CTCF motif occurrence. Then, for a motif clone, we plot the CTCF CHIP signal averaged over a window centered around the CTCF motif (across all occurrences) in the largest motif clone with 26 identical occurrences. We make this plot across all 20 cell types, shown in Figure 1.

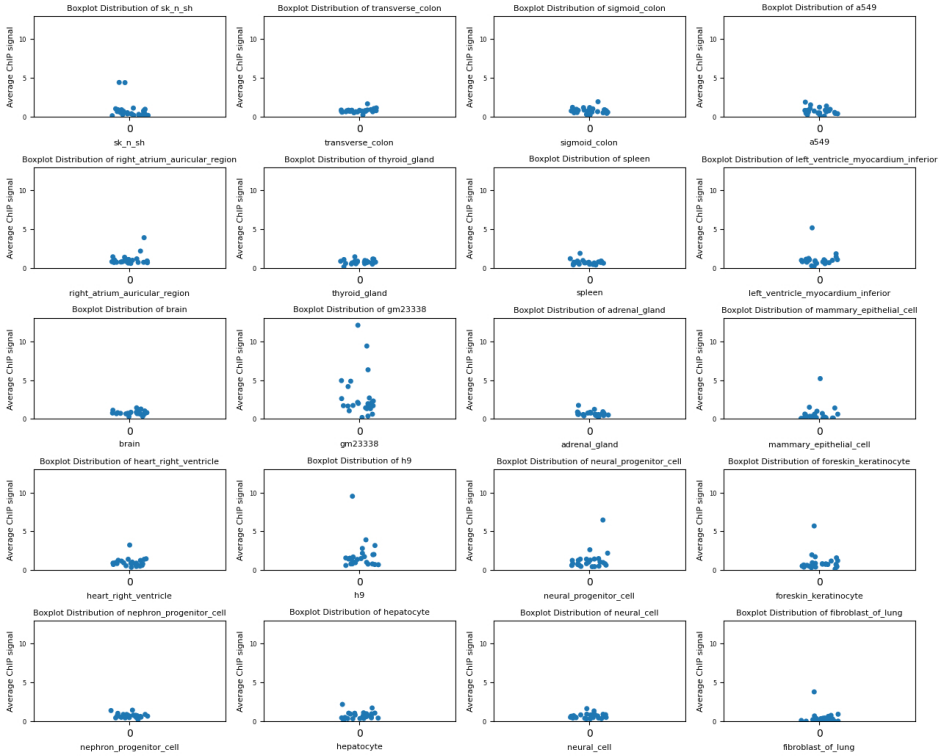


Figure 1: Average CTCF CHIP signal across 26 occurrences in largest CTCF motif clone across 20 cell types.

As Figure 1 illustrates, there are many outliers for many cell types despite the fact that the sequences across these motif occurrences are identical. This suggested that there might be other factors beyond the sequence that contribute to differences in the CHIP signals across these occurrences. However, are these differences simply due to noise? Are the differences in CHIP signal across occurrences of the same motif clone due to random chance?

To more rigorously quantify this, we performed some statistical tests using 3 replicate datasets in transverse colon cells. In particular, we wondered: for a given motif, is the difference in mean CHIP signal across occurrences statistically significant? After all, the variation in the CHIP signal could just be attributed to variation across replicates (error). Our null hypothesis is for a given motif clone, the

mean of the log-normalized CHIP-signals for each occurrences across replicates are all equal to each other. In the alternative hypothesis, at least one of the means is different across motif instances.

After performing an ANOVA test for each motif clone, we found a small subset of them (17 motif clones) where the BH-corrected p-value was significant. For these motif clones, at least one of occurrences had a mean CHIP-signal value that was different than the other means (a difference that was statistically significant). As a way of validating these results, we also performed Kruskal-Wallis test, a non-parametric analog of ANOVA which tests medians instead of means, using the same null and alternative. We find that the p-values of ANOVA and Kruskal-Wallis to quite consistent (see Figure), suggesting that the assumptions of ANOVA are met (e.g., normality and equal variances).

Overall, these results suggested that there are factors other than the sequence itself that are driving the site-specific binding patterns of CTCF. We hypothesized that one such factor was the differential methylation at the motif sites genome-wide.

Average Methylation and Fraction of CpG Sites Are Positively Correlated

Using the methylation data, we thus wondered: if two CHIP motif occurrences have the same sequence but different CHIP signal, is it because they are methylated differently? Knowing this will provide insight into why identical CTCF sequences are differentially bound by CTCF across different parts of the genome. Our hypothesis is that differentially activated CTCF sites are due to differential methylation profiles.

To investigate this, we looped over all the motif instances. For each of the 19 base pair positions of the CTCF motif, we computed the fraction of all motif occurrences that had cytosines followed by a guanine (CpG) at that position. Then we plotted the fraction of CpGs against the methylation signal averaged across the whole CTCF motif. Results are presented in Figure 2.

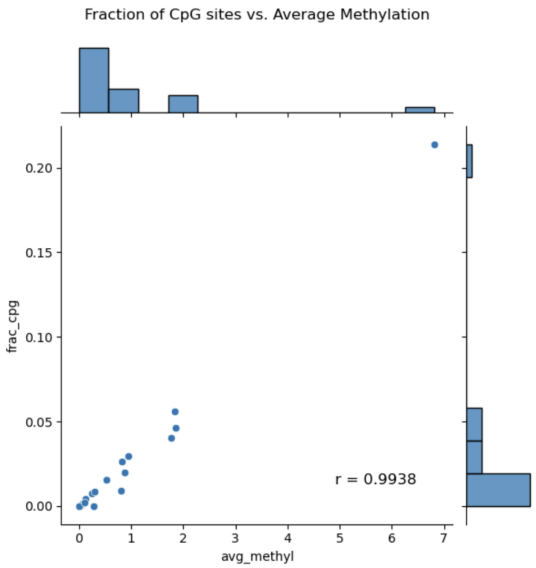


Figure 2: Scatterplot of the fraction of CpG sites and the average methylation signal across all motifs. Each point represents one of 19 positions in the CTCF motif.

As Figure 2 depicts, there is a very strong positive correlation between average methylation and fraction of CpGs in the CTCF motif. This is expected since we know that methylation happens frequently at CpG sites. In addition, we know that methylation at CpG islands contributes to gene silencing by making chromatin less accessible near the promoter regions of genes. Does a similar phenomenon

happen with CTCF motifs?

Average CHIP and Methylation at CTCF Sites Are Negatively Correlated

To explore the relationship between methylation at CpG sites in the CTCF motif and CTCF binding, we did the following procedure. We loop over each motif occurrence. Then, for each given motif occurrence x, we choose a given position y among the 19 positions; if there is a CpG site for x at position y, then we compute the average log CHIP over the entire 19 bp region as well as the methylation signal at position y. After doing this for all motif occurrences and all positions, we plot methylation at a given position with CpG sites against the log(CHIP). The correlation plot for position 14 (the position with the highest average methylation signal) is shown in Figure 3.

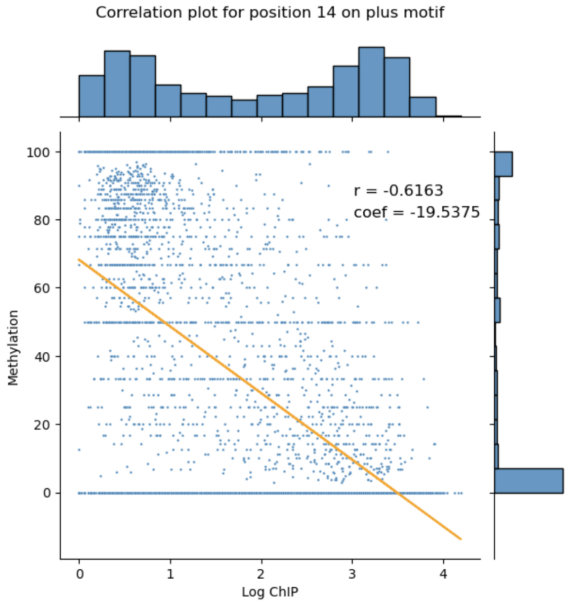


Figure 3: Scatterplot of the log(CHIP) and the average methylation signal across all motifs with a CpG at position 14. Each point represents one a motif occurrence. Also, the methylation data used corresponds to the plus strand. Line of best fit yields a negative coefficient and correlation coefficient, as is displayed on the graph.

As Figure 3 illustrates, we observe a moderate negative correlation between methylation and CTCF-CHIP at motif occurrences where position 14 has a CpG. Finally, to get a more comprehensive view of the distribution of coefficients and number of CpG sites at each position, we plotted a heatmap shown in Figure 4.

Figure 4 illustrates that methylation and CTCF CHIP have negative correlations at multiple positions in the CTCF motif where a CpG exists (e.g., at position 14 shown in Figure 3). This implies that as methylation increases, the CHIP signal tends to decrease, suggesting that methylation might inhibit the levels of CTCF binding.

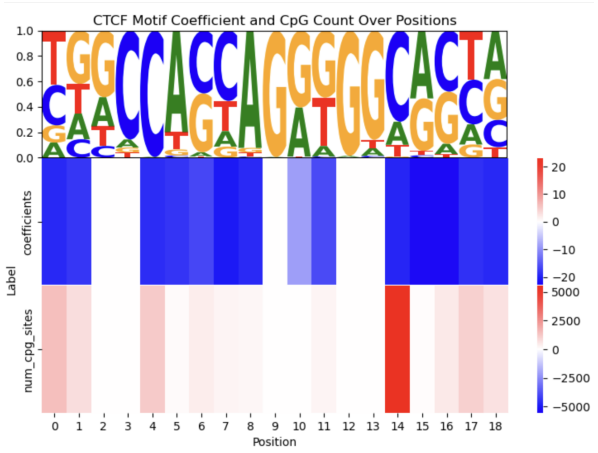


Figure 4: The first row shows the sequence logo of the CTCF motif. The second row show the coefficients of the scatterplot betwen methylation and log(CHIP) across all positions. The final row shows the distribution of number of CpG sites.

Phase 3: Modeling CTCF Binding Using Logistic Regression

Phase 2 allowed us to get a better sense of how methylation might play a role in CTCF binding. Several questions remained at this point. First, does the sequence alone already provide enough information for CTCF to recognize where to bind? (Our previous analysis on motif clones would suggest no). Second, how much of CTCF binding can be explained by sequence alone? How much by sequence and methylation patterns? To answer these questions, we started off by building a logistic regression model that classifies whether a given motif has strong or weak CTCF binding.

For our approach, we took all the CTCF motif occurrences (around 25K in total) and binarized the motif signal as "weak" (0) and "strong" (1) if the log(CHIP) was lower and higher than a threshold cutoff (determined by examining the average log(CHIP) distribution), respectively. Then, we featurized the motif examples using a one-hot encoding. For our final input dataset, we had a training, validation, and test dataset of one-hot encoded motif representations and the label (0/1). Using this data, we trained a logistic regression model with LASSO regularization, 3-fold cross validation, and GridSearch for 5 values of λ (based on AUC). After training, we obtained the best model and used it to predict the label on the held-out test sets, using a threshold of 0.5. This entire process is summarized in Figure 5 below.

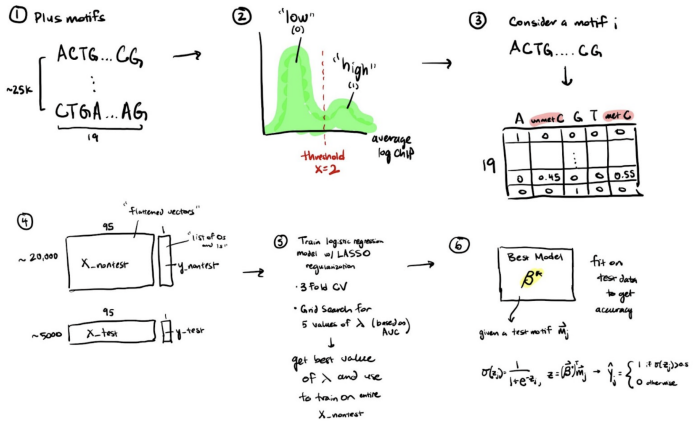


Figure 5: Flowchart schematic of training pipeline of Logistic Regression model.

Afterwards, we repeat this for more models we designed using combinations of larger context window

sizes (10bp, 20bp, and 40bp of context on either end) and methylation data (added as a fifth channel as input). When we add methylation data as an additional feature, we first check if the base at a given position is a CpG; if so, we split the C channel at that position into metC and unmetC, which represent the level of methylated and unmethylated Cs. More concretely, metC $\in [0, 1]$ and unmetC $=$ 1$-$ metC. The results of this ablation study is shown in Table 1. In addition, we plot the coefficients of the best performing Logistic Regression (LR) model with 40bp "window arm" in a heatmap shown in Figure 6.

Table 1: Accuracy of various trained LR models at classification of strong and weak CTCF CHIP motifs.

Context Arm Length (bp)	No Methylation	Methylation
0	0.7691	0.7891
10	0.8215	0.8381
20	**0.8327**	0.8456
30	0.8309	**0.8474**

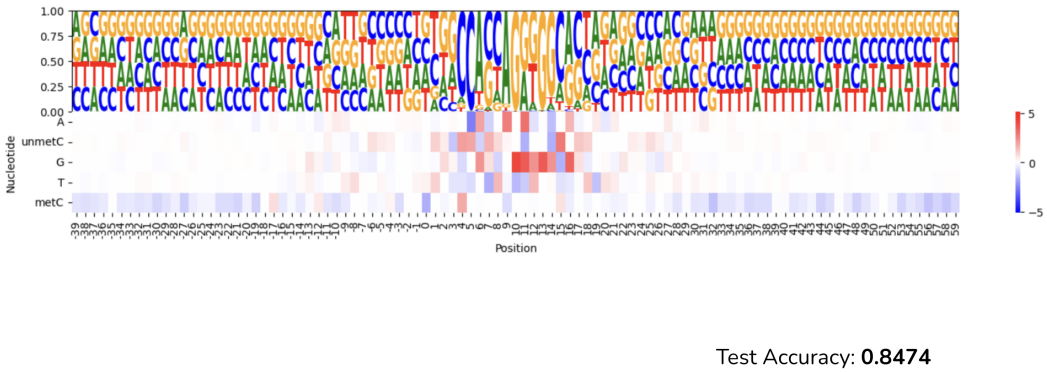


Figure 6: Sequence logo of CTCF motif surrounded by 40bp of context flanking on both ends. The heatmap displays the coefficients of the LR model trained using methylation data and 40bp of DNA context on one side.

Several interesting observations. First, the performance of the LR model increases significantly when we add more DNA context flanking the CTCF motif. Indeed, we observe a 5 to 6 % increase in accuracy just by adding 40 bp of context on both sides. This indicates that there are important features that can be learned just in the immediate genomic context of the CTCF motif.

Second, looking at Figure 6, we see that the coefficients have high magnitudes near the center positions where the CTCF motif is. This is expected since changes in the motif itself where the CTCF protein binds is likely to have a more dramatic impact on CTCF binding compared to changes in the context nearby. Additionally, we see that in the context, most of the strong coefficients are in the unmetC channel even at places far from the center. Perhaps, this provides some insight as to why context matters, since the presence of CpG sites near the CTCF motif can impact the levels of methylation in this region, which could in turn impact the level of accessibility in this region and ultimately, the level to which CTCF can bind.

Finally, it seems that although performance somewhat differs between LR models with and without methylation (assuming context is fixed), this difference is not as drastic as the performance changes attributed to context (all else equal). We conjecture that the smaller relative changes in performance attributed to methylation may be due to the fact that the sequence information already carries the bulk of "information" about the abundance and location of CpG sites in the flanking regions of the CTCF motif; hence, methylation data might not add too much additional information.

Phase 4: Modeling CTCF Binding Using Convolutional Neural Networks

Classification of Weak and Strong CTCF Binding

The Logistic Regression models we developed in Phase 3 were already performing quite well. However, the decision boundaries for predicting CTCF binding (classification task) may not be linear. As a more complex model such as a neural network or deep learning model may be able to mdoel these non-linear boundaries, we developed two classification convolutional neural network (CNN) models (denoted by Classification Net or **C**Net), which we describe below:

- **CNet1**: baseline convolutional neural network model consisting of 2 1D-convolution layers, each followed by pooling layers, followed by dense layers and softmax to output probability of 0 or 1. The first convolutional layer had 10 filters of size 20 and the second 10 filters of size 5. The input to the model is 4 by 1219 (600bp of context on each end of the CTCF motif).
- **CNet2**: similar to CNet1 except we have 3 1D-convolution layers each with 128 filters. There were 5 dense layers.

Table 2: Accuracy of two CNN-based models for classification of strong and weak CTCF CHIP motifs.

Model	**Accuracy**
CNet1	0.8449
CNet2	**0.8695**

We then train and evaluate these models on held-out test examples. Results are presented in Table 2. As the table illustrates, CNet2 has stronger performance than CNet1. Also comparing Table 1 and 2, CNet1 has comparable performance to the best LR model.

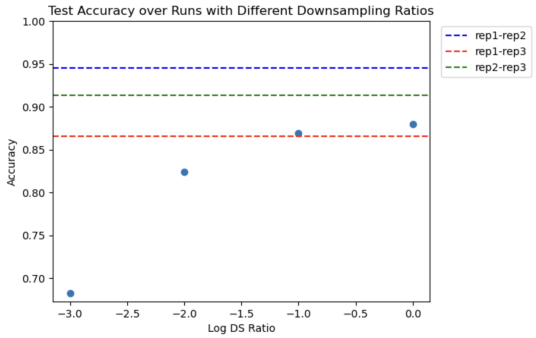
So what are these CNet models doing under the hood? The input to these models is the one-hot encoded DNA representation of the DNA sequence of size 1219. The first convolutional layer is applying a set of filters, each of which is trying to learn from the motifs themselves. Each of these filters bears some semblance of a position-weight matrix since it is trying to identify motifs in the original input. The following convolutional layers could be loosely interpreted as trying to capture patterns between groups or collections of motifs. The max pooling layers were used to reduce the size of the input while simultaneously attempting to capture the most salient information.

At this point, we hypothesized that CNet2 seemed to be the limits of what a CNN-based model could reasonably perform. To test this, we used accuracy performance between biological replicates as a reference. In particular, our model was trained on Replicate 1 on the CHIP data. We could use Replicate 2 and 3 as predictions to replicate-replicate performances; theoretically, this represents an upper bound performance of *any* model since any differences between the replicates are due to noise. In addition to computing replicate-replicate performances, we were interested in investigating whether changing the size of the input, the number of filters, the context window size, and changes to the bases in the core motif would affect model performance. We summarize our ablation procedures below:

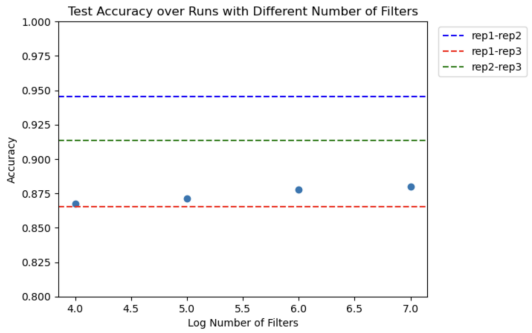
- **Size of Input**: trained and validated models trained on data that was $\frac{1}{8}, \frac{1}{4}, \frac{1}{2}, 1$ (i.e., downsampling ratios) the size of the original nontest dataset
- **Number of Filters**: trained and validated models that contained 16, 32, 64, and 128 convolutional filters.
- **Context Window Size**: trained and validated models trained on data that only contained nucleotide information based on 19, 49, 169, 319, 619, and 1219 bp. Incidentally, the window arm size is defined as (number of bp - 19)/2. For window arm sizes less than 600 bp, the bases that are not in the desired context are simply zeroed out.

- **Knockout of Bases in Core Motif**: trained and validated a model trained on context window of size 19 bp, where for all motifs, base i was replaced with the N nucleotide (one hot value of $[0, 0, 0, 0]^T$). WT means no KO was performed.

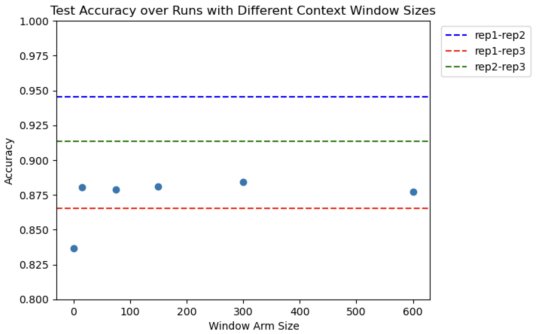
Results on the held-out test sets are summarized in Figure 7.



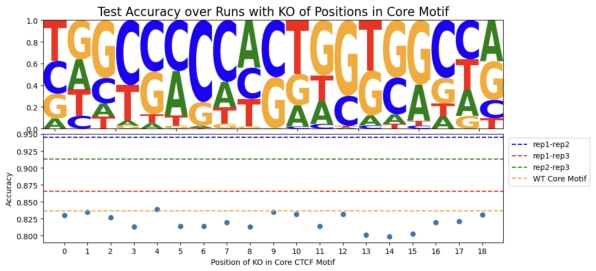
(a) Held-out performance using various DS ratios. The x-axis is log (base 2) of number of DS Ratios.



(b) Held-out performance using various number of filters. The x-axis is log (base 2) of number of filters.



(c) Held-out performance using various context window sizes.



(d) Held-out performance for KO of each position in CTCF motif.

Figure 7: Results of the four ablations. In each plot, dots represent model performance and dashed lines represent baseline performances.

When we compare Figures 7a, 7b, and 7c, the results suggest that context window size and input size of the data plays an essential role in classification of strong versus weak CTCF binding but the number of convolutional filters seems to be less significant (the model can sufficiently get enough information with just 16 filters). In addition, our models seem to be performing at their maximum capacity since many of the models are already performing within the range of the baseline performances. Finally, we observe from Figure 7d that knocking out certain positions could more significantly affect the performance. Indeed, we see that KO of position 14 which had the highest fraction of CpG sites disrupts performance the most.

Regression of Levels of CTCF Binding

From here, as we were now in the range of biological benchmarks for the classification task, we transitioned to designing CNN models for regression of the CTCF CHIP signal itself.

The data preprocessing pipeline for this step was nearly identical to that shown in Figure 5, except instead of binary labels, we used the log normalized CHIP signal at each site averaged over the CTCF motif. For the models, we developed various models to investigate how various factors impact model performance at regression of CTCF binding levels. We named the family of models Regression Net or simply "RNet," which are designed using the following specifications (unless otherwise specified, the dropout is 0.5, and the context range spanned 1219bp):

- **RNet-Base**: baseline convolutional neural network model with identical architecture to CNet2 but the output is one scalar. Uses DNA sequence only as input.
- **RNet-Met**: Identical to RNet-Base but with methylation data.
- **RNet-Met-ATAC**: Identical to RNet-Base but trained using input information from sequences, methylation, and ATAC-seq data. ATAC-seq provides a measure of accessibility at the input regions.
- **RNet-Met-ATAC-Mod**: Identical to RNet-Met-ATAC but with double number of filters, window size of 2419 bp, and dropout of 0.4. Different values of dropout via hyperparameter searching were tried to arrive at 0.4 (the optimal amount).

We concisely summarize the results of these models in Table 3. Here, we make several observations. The

Table 3: Average MSE performance and Pearson correlation performance across test examples for all the developed models. Along with model performances, replicate-replicate performances are also included.

Model	Average MSE	Pearson
RNet-Base	0.3158	0.8321
RNet-Met	0.2719	0.8593
RNet-Met-ATAC	0.1977	0.8988
RNet-Met-ATAC-Mod	0.1774	0.9093
Rep-1:Rep-2	0.0884	0.9561
Rep-1:Rep-3	0.1291	0.9357

first is that adding methylation data yields some improvements to model performance, a finding that is consistent with the classification models. Additionally, when adding information about the accessibility of the region (RNet-Met-ATAC), the model achieves must stronger performance (in terms of both MSE and Pearson correlation) relative to RNet-Met, the same model but without using ATAC-seq data. Finally, we see the best performing model RNet-Met-ATAC-Mod achieves comparable performance to replicate-replicate performances.

Discussion

In this project, we investigated the genome-wide binding mechanisms of the CTCF protein. In the exploratory data analysis phase, we found that CTCF expresses differential binding in motif clones and average CHIP and methylation within the CTCF motif are negatively correlated. To investigate how much of CTCF binding can be explained by the sequence, methylation, and accessibility patterns underlying the motifs, we developed a family of machine learning models, starting from logistic regression for classification of strong and weak CTCF binding to CNN models for regressing levels of CTCF binding. Thereafter, we ran many ablation experiments probing the effects of various input and model designs on model performance.

In our analysis, we found that context length and accessibility data most strongly improves model performance. This increase in performance attributed to both the context length and accessibility data is not too surprising to us, given that the accessibility of a genomic region is largely indicative and strong influences the binding of proteins to that region — regions with less accessibility tend to have lower levels of protein binding (including CTCF).

Incorporating methylation data into the input also improves model performance, although to a lesser extent. Perhaps, this is because the sequence information largely captures the features conducive to CTCF binding that can be attributed to methylation, and thus there is some redudancy in adding

methylation data.

Overall, while our results provide some preliminary insights as to what role the DNA sequence and methylation play in the binding mechanisms of CTCF, we are aware that there are many potential other factors that contribute to CTCF binding, such as biophysical properties of the chromosome, the TF occupancy and epigenetic landscape of the region, and more. We also recognize that the experiments that we did using machine learning models are also all done in-silico and thus, may not be able to give us the fine-grained details of binding mechanisms of CTCF at a molecular level. Areas of future work include examining at which CTCF motif occurrences does the model fall short as well as investigating if there are patterns that explain why CTCF binding is harder to predict at some regions over others (e.g., whether these regions are enriched for in repetitive regions in the genome). In addition, incorporating and ablating additional input features will be critical to advance our understanding of the genome-wide binding mechanisms of CTCF, both within and across cell types.