Transcription Factor Binding Prediction through DNA Physical Properties

Eric Haoran Huang
eric.h.huang@mail.mcgill.ca
McGill University
261275889

Connor Clark-Baba
connor.clark-baba@mail.mcgill.ca
McGill University
261276360

Neda Esfehani
neda.esfehani@mail.mcgill.ca
McGill University
261224369

Abstract

Detailing which open chromatin regions are bound by a target transcription factor (TF) is highly important for understanding gene regulation. This remains a difficult task as TF binding sites (TFBSs) depend on the physical properties of DNA. As developments in structured data mining have allowed for large corpi of these properties to be readily available for use, a natural step is in utilizing them to predict whether an open chromatin region is bound. In this paper, we argue that for chromatin regions that cannot be predicted solely using TF motif scores require more enriched input from sequence and structural data. We compare different Machine Learning methodologies to infer the complex relationships between the enriched data and TFBSs. In addition, we provide a robust, modular, and extensible framework on our GitHub[1].

1 Introduction

Transcription factors are a class of proteins which regulate gene expression by binding specific DNA motifs known as transcription factor binding sites (TFBSs) (Lambert et al., 2018). Whether a TFBS is bound depends strongly on the cellular context. In certain cancers, altered cell states may arise through changes in DNA methylation and histone modifications, which modify chromatin accessibility. Consequently, CCCTC-binding factor (CTCF) affinity to its binding sites can be altered, despite both the underlying motif and TF sequences remaining unchanged (Gridina and Fishman, 2022).

In this manner, the same TFBS may be bound in one condition and unbound in another, driven by changes in chromatin accessibility, DNA shape, or other structural factors (Levo et al., 2015). As such, alterations in TF-DNA affinity underlie numerous diseases and distinguishing bound from unbound TFBSs is important in understanding disease regulatory mechanisms. Although experimental techniques such as chromatin immunoprecipitation coupled to sequencing (ChIP-seq) provide high-throughput maps of bound sites, they remain costly and condition-specific, motivating the need for accurate computational prediction methods (Park, 2009).

Traditional approaches rely on sequence-level motif scores derived from Position Weight Matrices (PWMs), typically inferred from enriched motifs within ChIP-seq peaks. However, PWM-based scores alone often fail to explain *in vivo* binding because they ignore chromatin context and structural properties of the underlying DNA (Maienschein-Cline et al., 2012). Computational tools such as DNAShape and GBShape predict local DNA geometry at single-nucleotide resolution, including structural features such as minor groove width (MGW), propeller twist (ProT), helix twist (HelT), hydroxyl radical cleavage intensity (OC2), and roll (Zhou et al., 2013; Chiu et al., 2015). These structural features capture subtle variations in the geometry of the DNA double helix which influence how accessible and favorable the DNA is for protein binding (Inukai et al., 2017). Because these shape variations often correlate with regions of more open or flexible chromatin, they provide complementary information to PWM scores, allowing models to better distinguish bound sites from unbound sites that share similar motif sequences.

In this paper, we explore how DNA structural features can enrich the prediction of whether a candidate site is bound by a TF. We demonstrate this on the PAX5 TF in the GM12878 lymphoblast cell line. We show that both classical Machine Learning (ML) methods and modern deep learning approaches greatly improve on simple regression by better leveraging sequence information, with modest improvement from including structural information.

[1] https://github.com/ehuan2/tf-binding

1.1 Challenges

A recurrent theme across the literature is that motif strength alone is insufficient in explaining *in vivo* TF binding. We reproduce this trend on PWM scores derived from PAX5. Comparing PWM scores on all positive samples to scores on genome-wide negatives yields clearly separable distributions which should be easily distinguishable by a linear classifier (Fig. 1). However, when negatives are restricted to overlapping regions which contain the PAX5 motif but show no ChIP-seq enrichment (i.e. sequences that are motif-positive but unbound) the score distributions become far less distinguishable (Fig. 2).

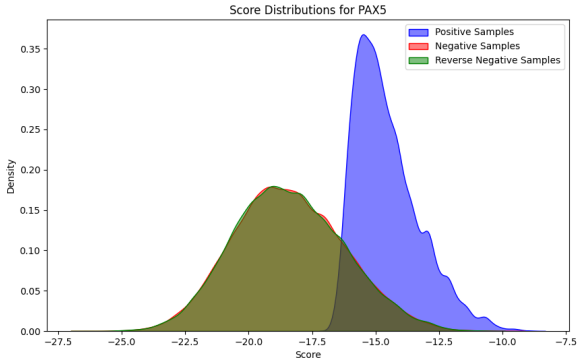

Figure 1: Distribution of top scores for genome-wide negative samples (forward and reverse) vs. the positive samples

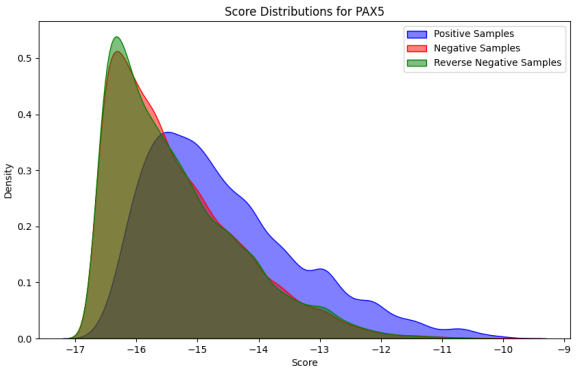

Figure 2: Distribution of top scores for restricted set of overlapping negative samples (forward and reverse) vs. the positive samples

We further demonstrate the limitations of motif strength by evaluating a simple linear classifier constructed directly from kernel density estimates of the PWM score distributions. Despite yielding a modest AUROC of 0.70 and AUPRC of 0.52, the overall performance remained limited (accuracy of 0.61, precision of 0.44, recall of 0.74, and F1 score of 0.56). These results reinforce that even though PWM scores can partially separate bound from unbound TFBSs, they fail to capture more nuanced factors that influence TF binding, motivating the inclusion of additional features such as DNA shape.

1.2 Objectives

In this report, we aim to evaluate the extent to which DNA structural features improve TFBS prediction beyond what is captured by sequence motifs alone. Building on our observation that PWM scores only partially distinguish bound from unbound sites, we use PAX5 in GM12878 cells as a case study to systematically compare several classifiers and determine whether structural features improve predictive performance.

We ultimately show that while classic ML methods do improve upon linear classifiers, when using general structural information, they remain limited in fully leveraging that data. Similarly, we show that deep learning techniques can match classic ML methods, but require further exploration.

2 Background

2.1 Case Study: PAX5

PAX5 is a paired-box TF that plays a central role in B-cell lineage commitment and maintenance. It regulates a broad network of genes acting as both an activator and repressor, and often cooperates with other B-cell-specific factors (Bullerwell et al., 2021). PAX5 contains a paired DNA-binding domain composed of two helix-turn-helix (HTH) subdomains and a linker region which bind three independent sites on a flexible 15-bp motif (Revilla-i Domingo et al., 2012). Specifically, the linker region binds the minor groove while the HTH subdomains interact with bases on the major groove. Each subdomain of PAX5 has relatively low specificity resulting in the 15-bp motif being highly degenerate (Revilla-i Domingo et al., 2012). Thus, successful binding does not depend only on the sequence but also on the relative positioning of the paired domains modules and local chromatin accessibility, proving PWMs alone to be insufficient. These properties make PAX5 particularly well-suited for evaluating whether the inclusion of DNA structural features can provide complementary information and improve predictive performance.

2.2 Related Work

ML approaches have transformed regulatory genomics by enabling models to learn complex sequence-function relationships directly from raw DNA. Early architectures such as Basset and DanQ apply convolutional neural networks (CNNs) and hybrid convolutional-recurrent neural networks (CNN-RNNs) to extract richer sequence representations and successfully predict cell-type-specific chromatin accessibility (Kelley et al., 2016; Quang and Xie, 2016). While such models do not perform TF-specific TFBS prediction, they established the effectiveness of deep architectures in capturing higher-order regulatory sequence features.

In parallel to advancements in sequence-based models, many studies have exploited DNA shape features to improve classification performance. Mathelier and colleagues, for example, used a gradient boosting classifier and reported measurable gains when DNA shape features were added alongside PWMs or raw sequence features (Mathelier et al., 2016). More recent deep learning approaches integrate shape directly into the input representation, such as CRTPS, which concatenates DNA-shape channels with one-hot sequence inputs within a CNN–RNN hybrid model (Wang et al., 2021).

3 Methodology

3.1 Data

We focus on predicting the binding of PAX5 in GM12878 lymphoblastoid cells. All input data was provided through the course repository and linked public resources. We used five main data sources; **i.** Human reference genome (hg19), **ii.** Active regulatory regions in GM12878 (ChIP-seq), **iii.** Genome-wide motif hits for multiple TFs (Factorbook), **iv.** PWM for PAX5, and **v.** Predicted DNA structural properties (MGW, HelT, ProT, OC2, and Roll from GBShape). More details on our available preprocessed data can be found on the GitHub's ReadMe[2]. While training, we ensure that 20% of the data is held-out for testing purposes only, while using the same data split (set by a random seed) across models.

3.1.1 Interval Construction and Labeling

The preprocessing code provided in the repository extracts candidate genomic sequences from open chromatin regions to label them as positive or negative intervals. We defined positive intervals as windows that overlap both a PAX5 motif hit and a corresponding PAX5 ChIP-seq peak. Negative intervals were restricted by only taking GM12878 regulatory regions that contain a PAX5 motif but show no ChIP-seq evidence of binding, done through a three-step filtering process:

1. We take all open regions that do not overlap with any positive interval.
2. We find the best TF-length interval according to its motif score on both the forward and reverse strands.
3. We filter out any remaining intervals whose motif score does not overlap with the range of possible positive interval motif scores shown in Figure 2.

This ensures that the negative samples are non-trivial, as they share the correct motif sequence but remain unbound *in vivo*, providing a more biologically realistic and challenging classification task.

3.1.2 Feature Representation

Each interval sample is represented as a fixed-length window around a candidate binding site using a combination of sequence, DNA shape, and motif features. For the sequence-based representation, we used a simple one-hot encoding of the nucleotides (A,C,G,T) across the window. For the structural representation, we used the precomputed bigWig tracks for MGW, Roll, HelT, OC2, and ProT, then extracted the corresponding values over each interval, of which we allow the extension of its context window on either side of the TF of both positive and negative samples to capture flanking shape information. This information is known to influence binding stability and accessibility. Finally, we include the PWM-based motif score for each nucleotide in the window, as well as the total motif score (single aggregate number). Then, on a per-model basis, we concatenate these features differently. For classical ML methods, we use a classifier on the single-concatenated vector (`[one-hot (sequence), DNAShape features, PWM scores]`), while concatenating differently or separating for deep learning models.

3.1.3 Preprocessing and Configuration

The GitHub repository was built to be modular, extensible, and easy to use. To this end, we use

[2] https://github.com/ehuan2/tf-binding/blob/main/ReadMe.md

MLflow (Chen et al., 2020) framework to capture multiple different runs of models, and allowing for easy model configuration through the use of YAML files. These configuration files can be used to specify the TF of interest (PAX5), the list of structural features to use (`MGW`, `Roll`, `HelT`, `ProT`, `OC2`), and flags that enable ablation studies (such as not including sequence data). Using the provided dataloader, the PAX5 dataset is split into 36,362 training intervals and 9,091 test intervals, maintaining the positive/negative class imbalance determined by the preprocessing pipeline.

3.2 Classifiers

3.2.1 Classical ML: Scikit-learn Models

We implemented traditional ML models as scikit-learn (v1.7.2) pipelines (Pedregosa et al., 2012). These include logistic regression, gradient-boosted decision trees via XGBoost (v3.1.2) (Chen and Guestrin, 2016), Random Forests (Breiman, 2001), and Support Vector Machines (SVMs). This was done for completeness to benchmark performance across simple non-linear, kernel-based, and ensemble architectures. All models were fit on the full training set and evaluated on the held-out test set. For model specific details, refer to the GitHub models' folder.

3.2.2 Deep Learning: PyTorch Models

Deep learning models were implemented in PyTorch (v2.9.1) (Paszke et al., 2019). Based off output from the traditional ML models, we found that raw one-hot encoded sequence inputs contributed minimally beyond PWM and DNA shape features. Additionally, these inputs scale poorly with deeper architectures. Therefore, PyTorch models were trained using only DNA structural features and PWM scores.

We also test simple multi-layer perceptron (MLP), 1D Convolutional Neural Network (CNN), 2D-CNN, and variational autoencoder (VAE) frameworks to assess whether increased model capacity or alternative feature transformations improve performance. We focus on MLPs and 1D-CNNs. MLPs provide the simplest architecture, while 1D-CNNs are the clearest method to best utilize localized patterns in structural features for TFBS prediction. Full implementation details for all models are provided in Appendix A.2.

3.3 Evaluation Metrics

We assessed model performance using several standard binary classification metrics computed from the predicted probabilities and thresholded class labels. These include accuracy, F1 score, and the area under both the receiver operating characteristic curve (AUROC) and the precision–recall curve (AUPRC).

Since our dataset exhibits a moderate class imbalance (15,023 positives vs. 30,430 negatives), metrics that directly account for this are particularly informative. In this context, AUPRC provides a more realistic depiction of performance than accuracy or AUROC as it emphasizes each model's ability to correctly identify true binding events despite the imbalance. Similarly, F1 score, the harmonic mean between precision and recall, is critical for our application where we place high value on maintaining a strong true positive rate while minimizing false negatives.

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} = \frac{\text{TP}}{\text{TP} + \frac{1}{2}(\text{FP} + \text{FN})}$$

From a biological perspective, correctly identifying true TF binding events is the most important task as false negatives can obscure meaningful regulatory interactions and compromise downstream conclusions. Accuracy and AUROC remain useful for overall characterization, but as they are more sensitive to imbalanced classes, they are interpreted primarily as secondary metrics. For each run, we additionally generate and save the ROC and PR curves.

4 Results

In our results, we first show that solely using motif probability information results in a sub-par classifier, and that by including both sequence information and more importantly structural information, a stronger classifier can be built. At the same time, we argue that the best traditional ML classifier, XGBoost, improves the most by including structural information as opposed to only using sequence information, as we see the best F1 score from including larger structural contexts. However, this still only results in modest improvements, motivating the use of more complex deep learning techniques that can better leverage this data structure. Overall, we show that including structural information is important in TFBS prediction, though remains difficult to fully exploit. We conclude by motivating

the development of better robust methods through deep learning architectures.

4.1 Importance of Structural Information

We ablate on including different types of input data to our various models, starting from simple motif and PWM scores, then adding either only sequence data, structural information, or both. By comparing the improvements from including each data modality, we can conclude whether that information is useful and should be used to further develop better classifiers.

In particular, Table 1 displays that on average, the included structure information helps all classical ML models across all metrics. Of note, structural data with motif probabilities is the second best modality, only behind including all data. It is ahead of sequence with motif probabilities, suggesting that sequence information itself needs to be supplemented with structural information.

Due to the complexity and the high dimensionality of these one-hot encoding sequence data, we thus decided to not include sequence data in deep learning techniques (see Section 4.3) to reduce the complexity of the input. For a more detailed breakdown of each method and their performance across modalities, see appendix A.1.

It is clear that while structural information provides the most benefit, its improvement remains modest, only increasing the F1 score by 0.02 in comparison from pure motif probability scores for the best performing method XGBoost. To further improve accuracy, we need to either enhance our data or improve upon our modelling techniques.

4.2 Increasing Context Window

Ample data is crucial in ML to ensure models have enough information to correctly predict specific cases. To enhance the data that exists, we experiment on increasing the amount of data that is passed to the model by adding structural information surrounding each candidate TFBS. Specifically we run ablations on only increasing the context for structural information and compare its effect on F1 score. We add structural information from n base pair positions before and after the TF window for XGBoost, while maintaining the same PWM and sequence context length. Notably, adding corresponding PWM scores for the extended context is not relevant as these are motif-specific scores which are nearly random outside of the core motif and could even harm performance.

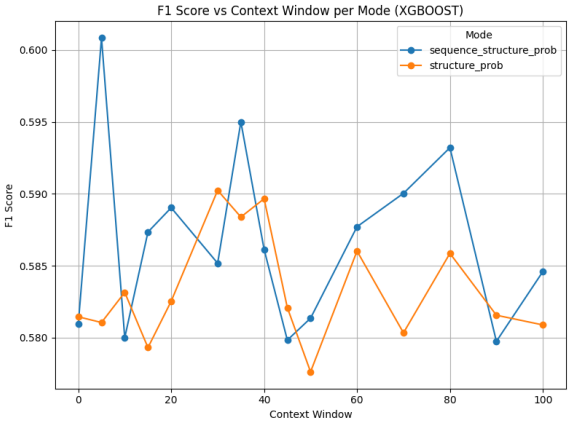

Figure 3: F1 Score for XGBoost on different context window lengths. Outlined in blue is the performance on including sequence data, in orange outlines the performance of not including it.

As shown in Figure 3, by increasing the context window, XGBoost does improve modestly. In particular, it improves the most when including sequence information as well, meaning that sequence information is still important to include when considering larger contexts. There is no clear trend however, of what context window remains the best, as there seems to be high variance in the outcome. The context window that seems to work consistently well for both methods are windows between 20 and 40 base pairs, which seems biologically plausible considering local effects of DNA shape. The overall improvement remains marginal with only 0.02 improvement for F1 score (in comparison with 0 context window) when using a context window of size 5. This motivates the exploration of more complex models that can better utilize the given input data.

4.3 Deep Learning Methods

Deep learning techniques attempt to build function approximators (such as classification labels) by updating tuneable weights. There are many architectures, and in Table 2 we compare the best runs across hyperparameters of different deep learning models, with XGBoost as a baseline. The results here suggest that deep learning architectures can be on-par with classic ML methods, however, currently do not improve nor overtake them. The architecture details are found in Appendix A.2, explaining MLPs, 1D CNNs, 2D CNNs, and VAEs.

In addition to measuring the best runs per model architecture, we ablate the context window length, and how much to penalize incorrectly classify-

Mode	Accuracy	ROC AUC	F1 Score	PR AUC
Motif Score	0.6751	0.6888	0.3516	0.5019
PWM	0.7098	0.7635	0.4573	0.5822
Structure + PWM	*0.7206*	*0.7849*	*0.4936*	*0.6117*
Sequence + PWM	0.7119	0.7764	0.4769	0.5921
Structure + Sequence + PWM	**0.7265**	**0.7917**	**0.5125**	**0.6226**
Mode (XGBoost)	**Accuracy**	**ROC AUC**	**F1 Score**	**PR AUC**
Motif Score	0.6899	0.6999	0.2755	0.5148
PWM Vector + Score	0.7458	0.8148	0.5594	0.6605
Structure + PWM	**0.7533**	**0.8279**	**0.5815**	**0.6874**
Sequence + PWM	0.7455	0.8183	0.5694	0.6666
Structure + Sequence + PWM	*0.7520*	*0.8277*	*0.5809*	**0.6874**
Linear Classifier (all data)	0.6082	0.7007	0.5557	0.5165

Table 1: Average scikit-learn model performance, and XGBoost-specific performance across different inputs modalities. Motif score indicates the scalar PWM score, PWM refers to the joint motif score and per-position PWM probability, structure refers to all five MGW, HelT, ProT, Roll, and OC2 structural data, and sequence refers to the one-hot encoding per nucleotide. **Bold** indicates the best value in each column, and *italic* indicates the second best. The final line is the linear classifier results from linearly separating kernel densities as shown in Figure 2.

Arch	Ctx	Acc	ROC-AUC	F1	PR-AUC
MLP	5	0.7037	**0.7931**	**0.6276**	**0.6445**
CNN	10	0.6788	0.7789	0.6257	0.6104
2D-CNN	0	0.6622	0.7447	0.6008	0.5624
VAE	0	0.5852	0.5719	0.4230	0.3882
XGBOOST	5	0.7746	0.8469	0.6021	0.7327

Table 2: Best model performance of deep learning architectures, in comparison with the best run of XGBoost. Note that ablations were not run on 2D-CNN nor VAE to find the best context window size.

ing positive samples during training (either no upweighting or upweighting by a factor of 2, due to a ratio of 1:2 positive to negative samples). We will focus on two methods, the MLP and 1D-CNN. By training on five epochs for both the MLP and the CNN, and varying between 0, 5, and 10 context window lengths, we get the following metrics found in Table 3.

The results suggest that all deep learning techniques should account for class imbalance. While the accuracies drop, the F1 scores (which are more important) do tend to improve significantly. Furthermore, the explored deep learning techniques either cannot utilize context windows well enough, or that we have not explored enough of the hyperparameter space. Which case follows remains an open question, and requires us to do more careful hyperparameter tuning, and dive deeper into the training dynamics of the chosen models.

Arch	Ctx	Acc	ROC-AUC	F1	PR-AUC
MLP	0	**0.7294**	0.7900	0.5572	0.6426
MLP	5	0.7197	0.7872	0.5716	0.6435
MLP	10	0.7272	0.7905	0.5565	0.6401
CNN	0	0.7138	0.7714	0.4696	0.6128
CNN	5	0.7293	0.7881	0.5027	0.6380
CNN	10	0.7208	0.7854	0.4439	0.6322
With upweighting positive classes					
MLP	0	0.5923	0.7741	0.6044	0.6134
MLP	5	0.7037	**0.7931**	**0.6276**	**0.6445**
MLP	10	0.6954	0.7827	0.6261	0.6230
CNN	0	0.7146	0.7630	0.5961	0.5238
CNN	5	0.7128	0.7779	0.5965	0.6124
CNN	10	0.6788	0.7789	0.6257	0.6104

Table 3: Model performance across architectures and context windows, and with upweighting positive intervals to fix the class imbalance.

5 Discussion

In this study, we evaluated how incorporating nucleotide-resolution DNA structural features influences TFBS prediction for PAX5 using a range of traditional ML and deep learning frameworks. Across all settings, we observed that DNA shape information provided moderate though consistent benefits, while additional sequence information (for traditional ML models only) displayed little improvement with high variance. Below, we interpret these findings in terms of underlying biology, model behaviour, and future methodological directions.

5.1 Biological Interpretation

Generally, our findings align with the known biology of PAX5. The canonical binding motif of PAX5 is highly degenerate and may exhibit considerable sequence variability. We expected this degeneracy to result in the nucleotide sequence itself providing a weak or ambiguous signal meaning many true binding sites may not closely resemble the consensus PWM score while false sites have an increased chance of randomly matching the motif (Mathelier et al., 2016). In contrast, DNA structural features offered a more stable and biologically meaningful representation of PAX5 binding preferences. PAX family TFs are known to rely on local DNA shape to establish favourable TF-DNA contacts, particularly MGW, roll, and other related helical parameters (Jolma, 2015). Thus, even when the PAX5 motif sequence varies, shape features remain predictive because similar structural conformations can arise from diverse nucleotide sequences. Additionally, increasing the context window enabled structural characterization of flanking regions which are known to influence DNA flexibility and the shape of the binding pocket (Zhou, 2015). However, the benefits of naively including more structural information through this expanded window in our analyses remain inconclusive.

5.2 Interpretation of Model Behaviours

5.2.1 Traditional ML Models

Among traditional ML methods, we focus on XGBoost as a representative model since it consistently achieved the strongest performance across all metrics. Consistent with our biological interpretation, structural features contributed more reliably to performance gains than sequence-based features when each modality was independently included alongside PWM scores. A more nuanced trend emerges when we examine how structural features interact with context window size. While the most biologically relevant DNA shape information for TFs is typically confined to around 15 bp surrounding the core motif, this still results in a relatively small feature set for training XGBoost (Mathelier et al., 2016). Extending the structural context window to 20-40 bp therefore increases the amount of correlated structural signal to aid in classification, even if not all added positions are shown to explicitly influence local DNA geometry.

Indeed, we observe a general increase in F1 score up to a 40 bp context window, followed by a sharp decrease and substantial variance as the window is further expanded to 100 bp (Figure 3). Importantly, the F1 score remains tightly constrained between approximately 0.575 and 0.590, indicating that increased structural context has only a limited effect on model performance. When sequence features at the core motif are additionally included, slightly higher peak F1 scores are observed though this is accompanied by a drastic increase in variance. This suggests that any gain is highly inconsistent and may reflect sensitivity to noise rather than robust sequence-driven signal.

Taken together, these results suggest that structural information provides modest gains while using classic ML methods. Based on existing literature, larger context windows should increase classification performance, however these methods struggle to best leverage them. From these points, we conclude that classic ML methods **cannot** fully capture the complexity of TFBSs and their structural properties.

5.2.2 Deep Learning Models

Deep learning models, including MLP and 1D-CNN, achieved comparable performance to XGBoost under the experimental setting tests. While both architectures benefited marginally from the inclusion of structural context, neither demonstrated clear advantages over traditional models, despite their increased capacities. Our results are not entirely consistent with prior successes of deep learning in regulatory genomics and highlights several important TFBS-prediction-task-specific constraints.

In particular, we have three possible explanations for the underperforming:

1. We did not perform enough of a hyperparameter search.
2. Sequence data was omitted completely.
3. PAX5 itself is a difficult TF to characterize due to its underlying biology.

In deep learning, there are many different parameters such as training length, network depth, number of parameters in a network, learning rate, and many others. All of which may significantly impact final performance metrics, leading to underperforming if not well executed.

Furthermore, all deep learning models were trained exclusively on structural features and PWM information, with sequence data being omitted

based on earlier ablation results. While this design choice reduced input dimensionality and noise brought about through sparse one-hot encoded representations, it may also have limited the advantage of deeper, more expressive architectures, which have proven to be effective at learning rich representations directly from raw sequence (Kelley et al., 2016; Quang and Xie, 2016).

Finally, the lack of substantial improvement for CNNs is unexpected given their prior success in capturing local regulatory dependencies. Previous CNN methods Basset and DanQ are trained on prediction tasks that for intervals ranging from 600-1000 base pairs (Kelley et al., 2016; Quang and Xie, 2016). In contrast, our problem of TFBS binding classification focuses on narrow genomic intervals where distinguishable features are sparse and subtle. This makes the PAX5 binding prediction task especially difficult as its motif window is short (16 base pairs), and has a weak motif signal.

5.3 Limitations

Despite these insights, several limitations remain. Generally, TFBS genomic intervals are inherently small and may not always contain sufficient contextual information for deeper architectures. For sequence-based information, this is especially true when the motif signal is highly weak or variable. Although we extend the context window to combat this, it remains limited to structural features and does not account for any distal regulatory interactions that may be present, resulting in marginal gains at best. Additionally, we rely entirely on predicted DNA shape features which may introduce noise in challenging regions. While DNAShape and GBShape are quite accurate on average, high-resolution experimental data could help validate or refine these predictions (Zhou et al., 2013; Chiu et al., 2015).

We report no significant gains from including one-hot sequence encodings in the TFBS prediction task. However, other models such as Basset or DanQ, which rely entirely on sequence information, display commendable results in cell-type-specific chromatin accessibility prediction (Kelley et al., 2016; Quang and Xie, 2016). This suggests two constraints in our current framework: (1) our models are not deep enough to capture complex sequence signals that larger architectures can learn, and (2) the inherently small sequence windows for TFBS prediction task, combined with the weak and degenerate PAX5 motif cannot provide ample sequence signal for deeper models.

Finally, our evaluation is limited to a single TF, PAX5. We were primarily interested in determining the usefulness of including DNA shape features, a setting where PAX5 is an appropriate test case due to its degenerate motif and shape-sensitive binding mechanisms. However, our results may not generalize to TFs with stronger sequence specificity or different structural binding dynamics and additional TFs should be examined.

5.4 Future Work

Future work should address the limitations described. In particular, there should be better hyperparameter searching, sequence data inclusion, and experimenting on other important TFs from the lymphoblast cell line.

Further exploration on deeper architectures would also prove beneficial, as they can better incorporate sequence information. Such models include transformers or dilated CNNs, as they provide deeper architectures that can likely uncover more meaningful sequence patterns than the shallower architectures tested here. This would enable the sequence to also be expanded across the context window and allow characterization of larger DNA regions or additional genomic signals beyond the local context of the core motif to capture more factors that influence TF binding. Regarding the structural information, leveraging or benchmarking against experimentally measured DNA shape data, when available, could help assess how much model performance depends on limitations in current structural prediction tools.

Finally, once a more reliable classifier is established, *in silico* perturbation studies can be performed on DNA physical properties to predict how certain conditions may affect TFBS binding and potentially uncover disease mechanisms or progression.

6 Conclusion

In conclusion, this work demonstrates that DNA structural properties can capture meaningful aspects of PAX5 binding that are not fully explained by sequence alone. Our results reinforce biological findings where shape contributes substantially to recognition, particularly at degenerate motifs. By systematically evaluating sequence and shape features across multiple ML frameworks, we show that structural information provides consistent pre-

dictive value, while sequence-only models yielded mixed and ultimately inconclusive results. Together, these findings highlight the importance of integrating structural signals when modeling TF-DNA interactions. While deeper models capable of leveraging sequence data, experimental shape measurements, and broader TF coverage remain promising directions, our results establish a clear baseline that DNA shape is indeed a critical component of TFBS prediction.

References

Leo Breiman. 2001. Random forests. *Mach. Learn.*, 45(1):5–32.

Charles E. Bullerwell, Philippe Pierre Robichaud, Pierre M. L. Deprez, Andrew P. Joy, Gabriel Wajnberg, Darwin D'Souza, Simi Chacko, Sébastien Fournier, Nicolas Crapoulet, David A. Barnett, Stephen M. Lewis, and Rodney J. Ouellette. 2021. Ebf1 drives hallmark b cell gene expression by enabling the interaction of pax5 with the mll h3k4 methyltransferase complex. *Scientific Reports*, 11(1):1537.

Andrew Chen, Andy Chow, Aaron Davidson, Arjun DCunha, Ali Ghodsi, Sue Ann Hong, Andy Konwinski, Clemens Mewald, Siddharth Murching, Tomas Nykodym, Paul Ogilvie, Mani Parkhe, Avesh Singh, Fen Xie, Matei Zaharia, Richard Zang, Juntai Zheng, and Corey Zumar. 2020. Developments in mlflow: A system to accelerate the machine learning lifecycle. In *Proceedings of the Fourth International Workshop on Data Management for End-to-End Machine Learning*, DEEM '20, New York, NY, USA. Association for Computing Machinery.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. *CoRR*, abs/1603.02754.

Tsu-Pei Chiu, Lin Yang, Tianyin Zhou, Bradley J Main, Stephen C J Parker, Sergey V Nuzhdin, Thomas D Tullius, and Remo Rohs. 2015. GBshape: a genome browser database for DNA shape annotations. *Nucleic Acids Res.*, 43(Database issue):D103–9.

Maria Gridina and Veniamin Fishman. 2022. Multilevel view on chromatin architecture alterations in cancer. *Frontiers in Genetics*, Volume 13 - 2022.

Sachi Inukai, Kian Hong Kock, and Martha L. Bulyk. 2017. Transcription factor-dna binding: beyond binding site motifs. *Current opinion in genetics & development*, 43:110–119.

Arttu et al. Jolma. 2015. Dna-dependent formation of transcription factor pairs alters their binding specificity. *Nature*, 527:384–388.

David R. Kelley, Jasper Snoek, and John Rinn. 2016. Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks. *bioRxiv*.

Samuel A. Lambert, Arttu Jolma, Laura F. Campitelli, Pratyush K. Das, Yimeng Yin, Mihai Albu, Xiaoting Chen, Jussi Taipale, Timothy R. Hughes, and Matthew T. Weirauch. 2018. The human transcription factors. *Cell*, 172(4):650–665.

Michal Levo, Einat Zalckvar, Eilon Sharon, Ana Carolina Dantas Machado, Yael Kalma, Maya Lotam-Pompan, Adina Weinberger, Zohar Yakhini, Remo Rohs, and Eran Segal. 2015. Unraveling determinants of transcription factor binding outside the core binding site. *Genome Research*, 25(7):1018–1029.

Mark Maienschein-Cline, Aaron R. Dinner, William S. Hlavacek, and Fangping Mu. 2012. Improved predictions of transcription factor binding sites using physicochemical features of dna. *Nucleic Acids Research*, 40(22):e175–e175.

Anthony Mathelier, Beibei Xin, Tsu-Pei Chiu, Lin Yang, Remo Rohs, and Wyeth W. Wasserman. 2016. Dna shape features improve transcription factor binding site predictions in vivo. *Cell Systems*, 3(3):278–286.e4.

Peter J. Park. 2009. Chip–seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics*, 10(10):669–680.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, and 2 others. 2019. Pytorch: An imperative style, high-performance deep learning library. *Preprint*, arXiv:1912.01703.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake VanderPlas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. 2012. Scikit-learn: Machine learning in python. *CoRR*, abs/1201.0490.

Daniel Quang and Xiaohui Xie. 2016. Danq: a hybrid convolutional and recurrent deep neural network for quantifying the function of dna sequences. *Nucleic Acids Research*, 44(11):e107–e107.

Roger Revilla-i Domingo, Ivan Bilic, Bojan Vilagos, Hiromi Tagoh, Anja Ebert, Ido M. Tamir, Leonie Smeenk, Johanna Trupke, Andreas Sommer, Markus Jaritz, and Meinrad Busslinger. 2012. The b-cell identity factor pax5 regulates distinct transcriptional programmes in early and late b lymphopoiesis. *The EMBO Journal*, 31(14):3130–3146.

F Rosenblatt. 1958. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol. Rev.*, 65(6):386–408.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. 1986. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536.

Siguo Wang, Qinhu Zhang, Zhen Shen, Ying He, Zhen-Heng Chen, Jianqiang Li, and De-Shuang Huang. 2021. Predicting transcription factor binding sites using dna shape features based on shared hybrid deep learning architecture. *Molecular Therapy - Nucleic Acids*, 24:154–163.

Tianyin Zhou, Lin Yang, Yan Lu, Iris Dror, Ana Carolina Dantas Machado, Tahereh Ghane, Rosa Di Felice, and Remo Rohs. 2013. Dnashape: a method for the high-throughput prediction of dna structural features on a genomic scale. *Nucleic Acids Research*, 41(W1):W56–W62.

Ting et al. Zhou. 2015. Quantitative modeling of transcription factor binding specificities using dna shape. *PNAS*, 112(15):4654–4659.

A Appendix

A.1 Scikit-Learn Model Performance

Table 4 outlines the different traditional ML methods and their performance on different input modalities. XGBoost performs the best, while structural, sequence, and PWM or structural and PWM modalities are the best types of input.

Architecture	Accuracy	ROC AUC	F1 Score	PR AUC
Motif Score				
LOGREG	0.6856	0.6976	0.3409	0.5123
RANDOM_FOREST	0.6391	0.6603	0.4492	0.4682
SVM	0.6856	0.6976	0.3409	0.5123
XGBOOST	0.6899	0.6999	0.2755	0.5148
PWM				
LOGREG	0.6806	0.7204	0.3643	0.5141
RANDOM_FOREST	0.7318	0.7975	0.5409	0.6388
SVM	0.6810	0.7213	0.3646	0.5155
XGBOOST	0.7458	0.8148	0.5594	0.6605
Struct + PWM				
LOGREG	0.6951	0.7493	0.4389	0.5479
RANDOM_FOREST	0.7399	0.8118	0.5210	0.6626
SVM	0.6942	0.7504	0.4331	0.5489
XGBOOST	**0.7533**	**0.8279**	**0.5815**	**0.6874**
Seq + PWM				
LOGREG	0.6849	0.7450	0.4013	0.5322
RANDOM_FOREST	0.7320	0.7957	0.5360	0.6361
SVM	0.6852	0.7466	0.4008	0.5336
XGBOOST	0.7455	0.8183	0.5694	0.6666
Struct + Seq + PWM				
LOGREG	0.6955	0.7540	0.4477	0.5559
RANDOM_FOREST	0.7405	0.8082	0.5200	0.6585
SVM	0.7180	0.7769	0.5014	0.5885
XGBOOST	*0.7520*	*0.8277*	*0.5809*	**0.6874**

Table 4: Unified comparison across all input modes and classical ML architectures. Motif score indicates the scalar PWM score, PWM refers to the joint motif score and per-position PWM probability, structure refers to all five MGW, HelT, ProT, Roll, and OC2 structural data, and sequence refers to the one-hot encoding per nucleotide. **Bold** indicates the best value in each column, and *italic* indicates the second best.

A.2 Deep Learning Architectures

A.2.1 Multi-Layer Perceptron

MLPs are universal function approximators, learning the proper function through seeing enough training data (Rumelhart et al., 1986; Rosenblatt, 1958). TFBS is one such task that takes in a high-dimensional vector and classifies to either binded or not, represented as either a 0 or 1. For the multi-layer perceptron, we create separate MLPs per structural feature and probability weight matrix scores that is activated with ReLU and is 2 layers deep, with user-specified hidden size. In the ablations, this hidden size is set to 128. Then each of these features are concatenated alongside the aggregated motif score to predict a final class using a 3-layer MLP with a final hidden size of 64. Refer to the MLP model class for more details.

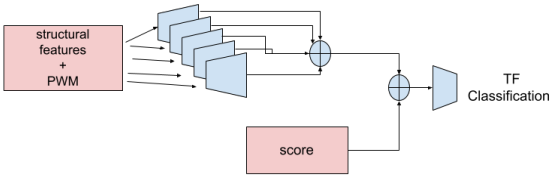

Figure 4: MLP Architecture

A.2.2 Variational Autoencoders

Shown in Figure 5, the variational autoencoder has a two-step approach, where we first pretrained an autoencoder, then used the embedding space with an MLP layer to do classification. However, this did not reveal anything interesting, most likely because although the input features were somewhat high dimensional, they were not so complicated that they needed a separate embedding space. Refer to the VAE model class for more details.

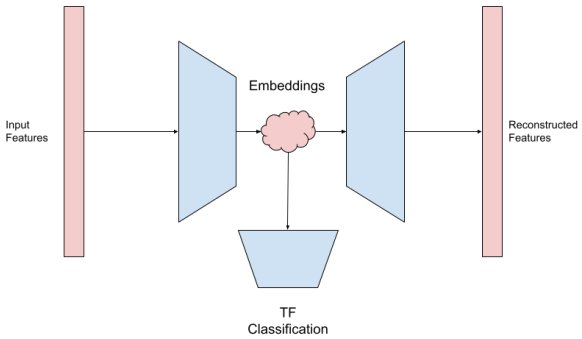

Figure 5: VAE Architecture

A.2.3 1D Convolutional Neural Network

CNNs can also be used for classification, famously applied to structured data where local context is important, such as the MNIST image classification. Due to the importance of local effects in DNA

structure, we hypothesized that using CNNs will be able to better leverage the context window lengths.

As depicted in Figure 6, the 1D convolutional neural network learns separate convolutions per feature, leading to separate feature vectors which are further convolved on. Then, by aggregating a final 1D feature vectors, we run it through an MLP for classification. Refer to the CNN model class for more details.

A.2.4 2D Convolutional Neural Network

Similar to 1D convolutional neural networks, the 2D convolutional neural networks (Figure 7) instead have a 2D convolution, creating 2D vectors at each stage. Refer to the 2D CNN model class for more details.

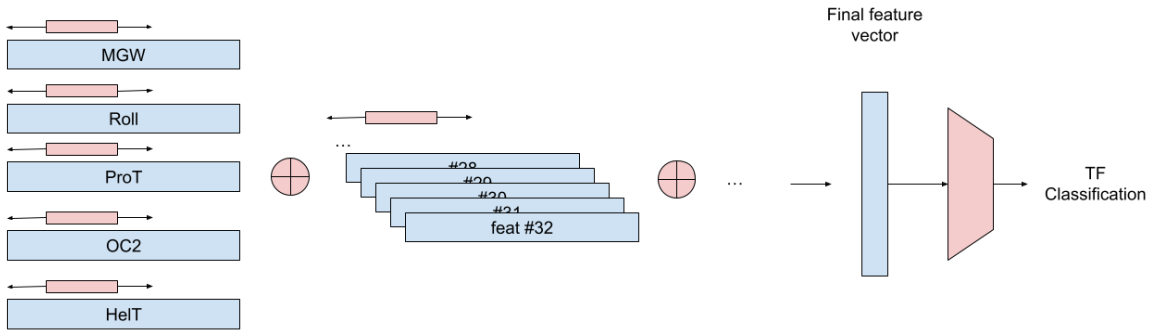

Figure 6: 1D CNN Architecture

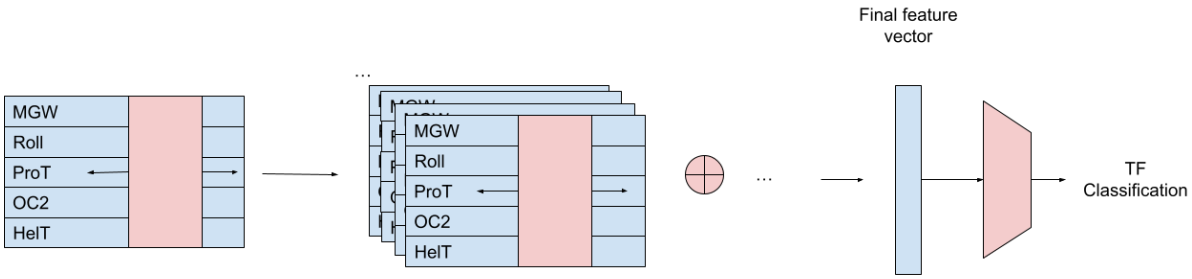

Figure 7: 2D CNN