Prediction of TF Binding Based on DNA Physical Properties

Hugo Paulat

hugo.paulat@mail.mcgill.ca

CS Department, McGill University, Canada

Introduction

Transcriptional regulation acts as the operating system of the cell, governing essential processes such as development and homeostasis. This regulation is driven by Transcription Factors (TFs), proteins that bind to specific genomic coordinates to modulate gene expression (He et al. 2023). While classical approaches attempt to predict these binding sites using sequence-based Position-Weight Matrices (PWMs), sequence matching alone is insufficient (Lavezzo et al. 2024). The recognition between a TF and its cognate site is a complex biophysical event; TFs often recognize specific DNA 3D shapes that are complementary to their binding domains, even in the absence of a perfect sequence match (Laughton and Luisi 1999).

Consequently, a major challenge in genomics is distinguishing functional binding sites from ”decoy” sequences that appear viable biologically but remain unoccupied in vivo. Currently, the ”gold standard” for mapping these sites, ChIP-seq, offers high accuracy but is resource-intensive, requiring high-quality antibodies and significant biological input (Park 2009). This makes scaling to new cell types or species difficult. Conversely, computational alternatives are often dominated by complex Deep Learning models (e.g., MLSNet) (Zhang et al. 2024). While powerful, these models act as ”black boxes,” requiring massive datasets and computational power while offering little biological interpretability regarding why a specific site is bound.

To bridge this gap, this study proposes a streamlined, interpretable alternative: predicting TFBS occupancy by modeling the local structure of DNA rather than relying solely on sequence or heavy neural networks. We focus on identifying bound sites on human Chromosome 1 using structural features extracted via the DNAShape package, specifically *Minor Groove Width (MGW)*, *Propeller Twist (ProT)*, *Roll*, and *Helix Twist (HelT)*, alongside *GC content*. We hypothesize that these structural signals contain sufficient information to distinguish true binding sites from non-functional candidates. By employing interpretable Logistic Regression and lightweight Multilayer Perceptron (MLP) models, we aim to demonstrate that accessible, efficient screening tools can achieve robust performance, democratizing TFBS prediction for resource-constrained research.

Methodology

Data Processing

Regulatory region annotations and transcription factor (TF) binding sites were first processed as `BED files`. Using the `intersect` functionality in the `bedtools` suite, both datasets were restricted to chromosome 1 and then intersected with each other to generate all overlapping regulatory regions and TF binding events. Training examples were generated using a sliding-window procedure. For each TF binding site, a 101 bp window centered on the binding site midpoint was extracted. This choice was based on other studies in this field (Zhou et al. 2019). This choice standardizes the input length despite variable TFBS sizes and captures the immediate flanking sequence known to influence binding affinity. All remaining regulatory regions that did not overlap any positive window were designated as negative examples. In total, this yielded 93,598 positive windows and 16,144 negative windows.

Feature extraction was then performed for each DNA shape modality. Directly computing the mean feature value within each window would be computationally expensive, so cumulative-sum arrays were constructed for each feature track. After converting these tracks to `NumPy` arrays, window means could be obtained by simple indexed lookups on the prefix-sum arrays, making the computation tractable at scale. To ensure consistency across features, the genomic interval of valid coverage shared by *MGW, ProT, Roll, Opening*, and *Buckle* was identified, and any window outside this common interval was discarded. This filtering step reduced the dataset to 49,870 windows. All features were then aggregated into a single dataframe, and GC content was computed for each window. A final quality-control check was performed to identify and remove any windows containing NaN values, resulting in 42,887 positive class samples and 6,982 negative class samples.

Feature Analysis

Pairwise Scatterplots These visualizations make it possible to detect structure that simple correlation coefficients may obscure. For several feature pairs, such as *MGW/Roll, ProT/Opening, GC Content/ProT, and Opening/GC Content* the points formed narrow, elongated bands, indicating strong linear relationships. In other cases, such as *Opening/Buckle*, the plots showed a bit more diffused clouds, with some possible visible boundaries, suggesting weaker associations (Figure 1).

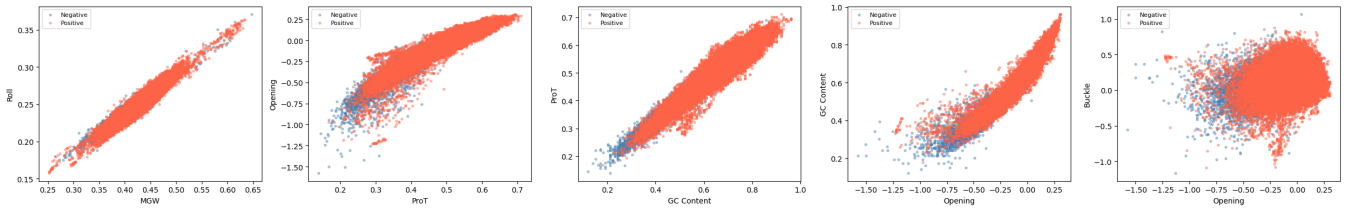


Figure 1: Pairwise scatterplots illustrating the relationships between DNA shape features

Correlation Analysis Graph observations were confirmed (Table 1). No meaningful nonlinear patterns or multimodal clusters appeared, implying that the relationships among the DNA shape features are largely monotonic and that the observed correlations are not driven by localized outliers or highly complex interactions. This is expected: physicochemical properties of DNA are intrinsically couples to its nucleotide

Feature 1	Feature 2	Correlation
MGW	Roll	0.948319
ProT	Opening	0.940132
ProT	GC Content	0.958051
Opening	GC Content	0.946098

Table 1: Correlation values for selected pairs of DNA shape and sequence features.

composition. For example, an increase in GC content directly dictates an increase in melting temperature and alters local DNA stiffness (Liu et al. 2007).

The distribution of values for each DNA shape feature were also examined. These plots allowed me to identify whether the values were tightly concentrated or broadly dispersed, and whether any long tails or irregular shapes suggested measurement artifacts or localized genomic effects. The distributions were generally smooth and unimodal, indicating that each feature spans a consistent range across the genome and does not contain extreme outliers that could disproportionately influence downstream models.

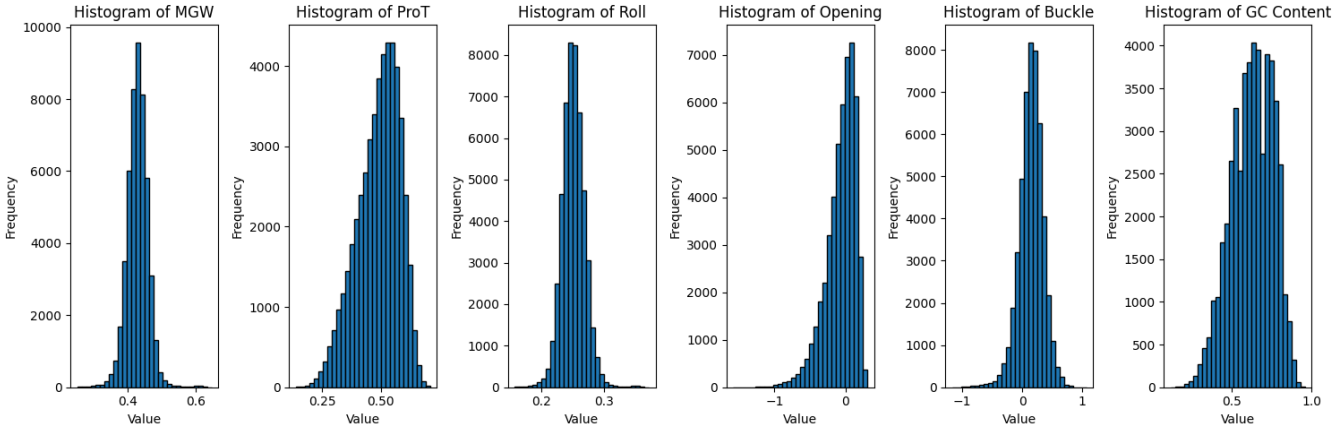


Figure 2: Distributions of the extracted DNA shape features and GC content across all windows.

Model Pipeline and Preprocessing

To ensure a rigorous comparison, both the Logistic Regression and Multi-Layer Perceptron (MLP) classifiers were developed using a unified pipeline structure designed to stabilize training and mitigate class imbalance. The feature matrix and target labels were first partitioned into training and test sets using stratified sampling with `test_size = 0.20`. This stratification preserved the original class distribution, ensuring that both partitions provided representative coverage of minority and majority windows.

Both classifiers were embedded in a `Pipeline` with a `StandardScaler`. Standardization is critical for these algorithms, albeit for distinct mathematical reasons: for logistic regression, it prevents features with large numerical ranges from disproportionately influencing the L2 penalty term; for the MLP, it prevents poorly conditioned loss surfaces that can destabilize gradient-based optimization.

Given the substantial class imbalance, both models incorporated inverse-frequency weighting to force

the optimization algorithms to penalize misclassification of the minority class. This was achieved using the `class_weight=balanced` parameter for logistic regression and `compute_sample_weight (balanced)` for the MLP.

Classifier Configurations

Logistic Regression: This linear model was selected to serve as a baseline for computational efficiency and interpretability. The classifier was configured with an L2 penalty—well-suited for handling the expected collinearity in the feature space—and optimized using the LBFGS solver.

Multi-Layer Perceptron (MLP): To capture non-linear feature interactions, an MLP was trained using ReLU activations and the Adam optimizer. The architecture used hidden layers of width 64, selected to balance model capacity against overfitting risk while remaining computationally tractable. To further prevent overfitting to the training partition, early stopping was enabled (`early_stopping = True`) with a patience of ten epochs (`n_iter_no_change = 10`), halting training when validation loss failed to improve.

Hyperparameter Optimization For both models this was performed using an exhaustive grid search within a 5-fold stratified cross-validation framework. This procedure ensured that class ratios were preserved across all validation folds, providing stable and representative performance estimates.

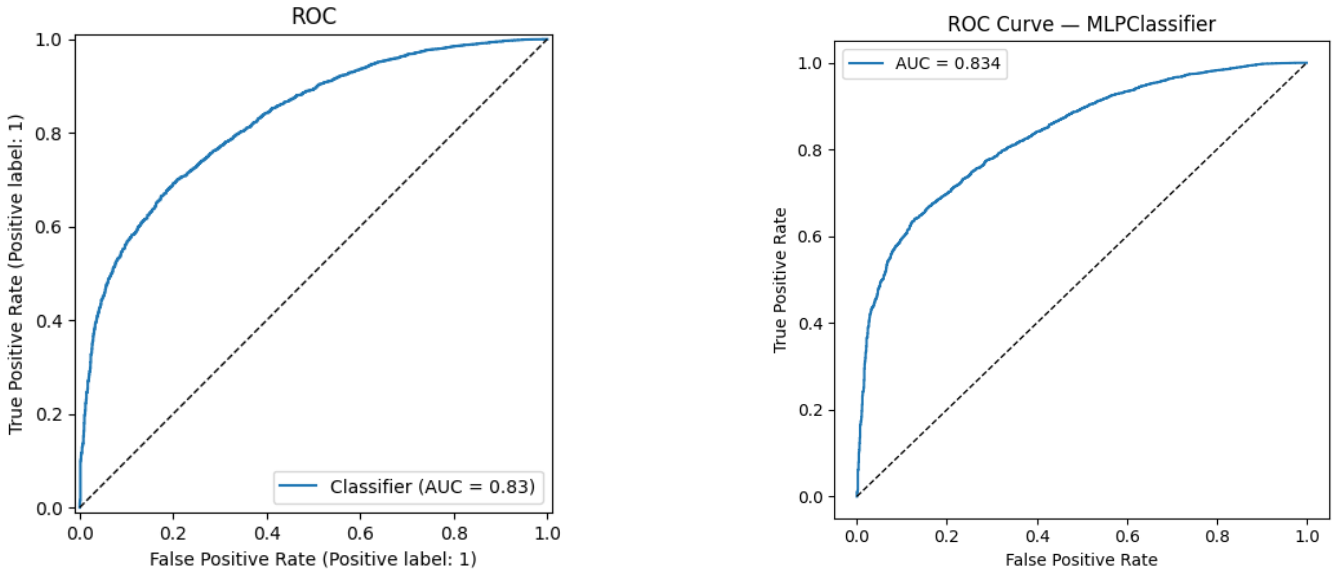
For Logistic Regression, the search space focused on regularization strength, evaluating 30 logarithmically spaced values of the inverse penalty parameter $C \in [10^{-4}, 10^{2}]$. For the MLP, the grid explored both architectural complexity and regularization, testing single and dual hidden-layer configurations (`(64,)` and `(64, 64)`) alongside L2 penalty parameters $\alpha \in [10^{-4}, 10^{1}]$.

Class Imbalance Model selection for both classifiers relied on maximizing balanced accuracy. While auxiliary metrics such as ROC AUC and average precision were tracked, maximizing balanced accuracy ensured that the final models equally prioritized sensitivity and specificity, preventing the selection of trivial classifiers that merely predict the majority class.

Results and Robustness Analysis

Hyperparameters $C = 0.5298$ for the Logistic Regression model and an optimal architecture of two hidden layers (64, 64) with L2 regularization strength $\alpha = 0.1$ for the MLP.

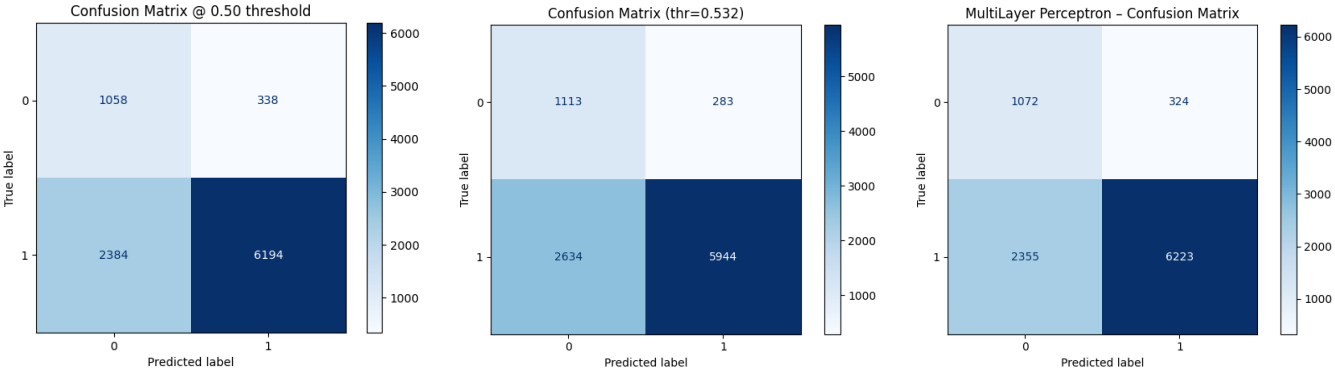
Generalization Performance Logistic Regression achieved a test balanced accuracy of 0.74, while the MLP showed a marginal improvement with a balanced accuracy of 0.75. Probabilistic predictions yielded nearly identical discriminative power: the Logistic Regression obtained an ROC AUC of 0.83, comparable to the MLP's ROC AUC of 0.834 (Figure 3). Given the significant class imbalance, Balanced Accuracy and ROC AUC were prioritized as the primary performance metrics, as they are independent of class prevalence and provide a robust assessment of discriminative power across both minority and majority classes.



(a) Log. Reg. ROC **(b)** MLP ROC

Figure 3: Comparative analysis of performance curves. (a) and (b) show the sensitivity analysis for the baseline Logistic Regression model, alongside (c) the ROC curve for the MLP classifier.

To further refine the Logistic Regression baseline, the decision threshold was optimized using `TunedThresholdClassifierCV`. This identified a probability cutoff of 0.532 that maximized balanced accuracy on validation folds. This slight adjustment from the default 0.50 decision rule yields a more equitable trade-off between false positives and false negatives, as illustrated by the shift in error distribution shown in the confusion matrices (Figure 4).



(a) Log. Reg. (Default) **(b)** Log. Reg. (Tuned) **(c)** MLP Classifier

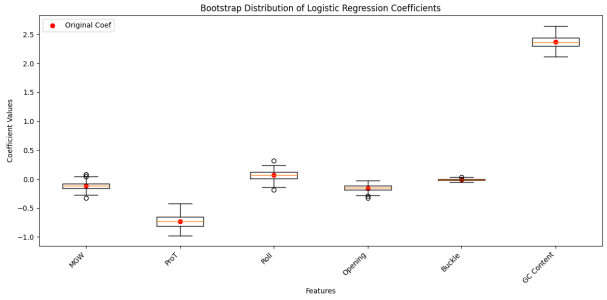
Figure 4: Confusion Matrix comparison. (a) Logistic Regression at default threshold (0.50), (b) Logistic Regression at optimized threshold (0.532), and (c) the MLP classifier. The tuned threshold in (b) shifts the error distribution to prioritize balanced accuracy.

Bootstrap To assess the reliability of feature contributions in the Logistic Regression model, we analyzed parameter uncertainty using bootstrapping ($n = 200$), as shown in Figure 5. The low variance

in coefficient estimates across bootstrap iterations confirms the stability of the model's feature rankings, reinforcing confidence in the biological interpretation of the weights.

Feature	Orig	Mean	Std	95% CI
MGW	-0.118	-0.113	0.070	$[-0.240, 0.028]$
ProT	-0.728	-0.730	0.110	$[-0.918, -0.515]$
Roll	0.070	0.065	0.083	$[-0.100, 0.208]$
Opening	-0.155	-0.153	0.053	$[-0.261, -0.050]$
Buckle	-0.010	-0.009	0.016	$[-0.040, 0.021]$
GC Content	2.367	2.366	0.103	$[2.160, 2.555]$

(a) Feature Statistics



(b) Logistic Regression Uncertainty

Figure 5: Combined analysis of feature statistics and model stability. (a) A summary of the input features, showing the original value, bootstrap mean, standard deviation, and 95% confidence interval. (b) The distribution of prediction uncertainty (standard deviation of predicted probabilities) for the Logistic Regression model across 100 bootstrap iterations, indicating highly stable predictions.

In the case of the MLP, where individual weights lack direct interpretability, we assessed model reliability by analyzing prediction uncertainty via bootstrapping ($n = 100$). The resulting distribution of predicted probabilities exhibited a low mean standard deviation ($\sigma \approx 0.014$). This stability indicates that the MLP has learned robust decision boundaries and produces consistent predictions impervious to minor fluctuations in the training data, arguing against overfitting.

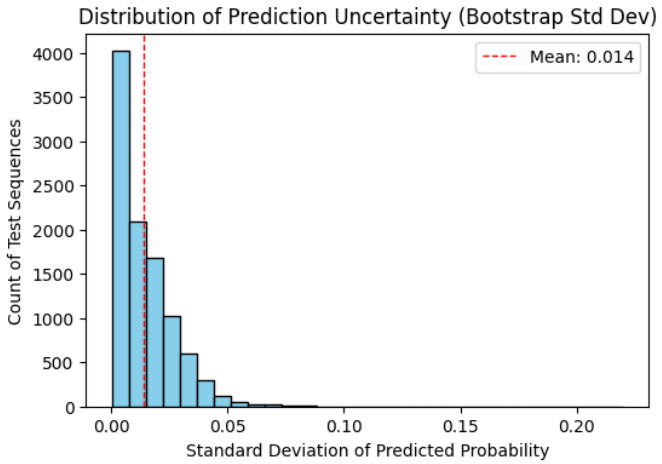


Figure 6: MLP Prediction Uncertainty. The histogram displays the standard deviation of predicted probabilities across 100 bootstrap iterations. The low mean standard deviation ($\sigma \approx 0.014$) indicates that the MLP produces highly stable predictions and is robust to variations in the training data.

Discussion and Future Work

Analysis of the Logistic Model's weights revealed GC Content as the dominant positive predictor (mean coefficient ≈ 2.366), suggesting a strong global preference for GC-rich motifs among the TFs on Chromosome 1. While this aligns with the known stability of G-C hydrogen bonding, it is important to con-

textualize this finding; as noted by Dror et al. 2015, TF binding preferences are heterogeneous, with distinct families favoring AT-rich environments. The model's heavy reliance on GC content likely reflects an aggregate average across the diverse TFs present in our dataset, potentially obscuring family-specific preferences.

Regarding structural features, Buckle and Roll exhibited coefficients near zero. However, this negligible influence should be interpreted with caution. Since different transcription factors require distinct—and often opposing—structural deformations (e.g., some induce bending while others stabilize rigid DNA), aggregating all TFs into a single class likely causes these structural signals to cancel out. Interestingly, ProT, Opening, and MGW displayed consistent negative weights. This inverse relationship corroborates the model's preference for GC content: AT-rich tracts are biophysically associated with narrow minor grooves (low MGW) and high propeller twist. By penalizing these features, the model effectively reinforces its selection against AT-rich regions in this specific dataset.

The performance convergence between the linear Logistic Regression and the non-linear Multilayer Perceptron (MLP) suggests that the current feature set does not harbor complex, non-linear dependencies that the linear model failed to capture. Instead, the limitation likely lies in the feature engineering itself. We utilized mean shape values across the specific sites; however, DNA shape readout is often driven by the flanking regions surrounding the core motif.

To address these limitations, future studies should move beyond aggregate modeling. We propose clustering TFs into structural families (e.g., Helix-Turn-Helix vs. Zinc Fingers) to determine if shape features become significant predictors within specific groups. Additionally, expanding the feature vector to include shape data from the flanking nucleotide positions—rather than the core site alone—may capture the local structural context necessary for high-specificity binding prediction.

Data and Code Availability

Most of the data used in this study has been previously published and is publicly available. The sequence of the human genome (assembly hg19) is found on UCSC's genome assembly browser (https://hgdownload.soe.ucsc.edu/downloads.html). The set of genomic regions that are active regulatory regions (GM12878) was provided by Professor Blanchette. The set of genomic coordinates of actual transcription factor binding sites for several transcription factors was also found on UCSC's website, in the Golden Path database (https://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/). Finally, the predicted structural properties for every human genome position used in this study was found using the DNAShape R package (https://www.bioconductor.org/packages/devel/bioc/html/DNAshapeR.html).

The code can be found on github: (https://github.com/hpaulat/Prediction-TF-Binding-Sites-Based-on-DNA-Physical-Properties).

References

Dror, Iris, Tamar Golan, Carmit Levy, Remo Rohs, and Yael Mandel-Gutfreund. 2015. "A widespread role of the motif environment in transcription factor binding across diverse protein families." *Genome Research* 25 (9): 1268–1280. https://doi.org/10.1101/gr.184671.114. https://doi.org/10.1101/gr.184671.114.

He, Hehe, Mingfei Yang, Siyu Li, Gaoyang Zhang, Zhongyang Ding, Liang Zhang, Guiyang Shi, and Youran Li. 2023. "Mechanisms and biotechnological applications of transcription factors." *Synthetic and Systems Biotechnology* 8 (4): 565–577. https://doi.org/10.1016/j.synbio.2023.08.006. https://doi.org/10.1016/j.synbio.2023.08.006.

Laughton, Charles A, and Benisi Luisi. 1999. "The mechanics of minor groove width variation in DNA, and its implications for the accommodation of ligands." *Journal of Molecular Biology* 288 (5): 953–963. https://doi.org/10.1006/jmbi.1999.2733. https://doi.org/10.1006/jmbi.1999.2733.

Lavezzo, Guilherme Miura, Marcelo de Souza Lauretto, Luiz Paulo Moura Andrioli, and Ariane Machado-Lima. 2024. "Position Weight Matrix or Acyclic Probabilistic Finite Automaton: Which model to use? A decision rule inferred for the prediction of transcription factor binding sites." *Genetics and Molecular Biology* 46 (4): e20230048. https://doi.org/10.1590/1678-4685-GMB-2023-0048. https://doi.org/10.1590/1678-4685-GMB-2023-0048.

Liu, Fang, Eivind Tøstesen, Jostein K. Sundet, Tor-Kristian Jenssen, Christoph Bock, Geir Ivar Jerstad, William G. Thilly, and Eivind Hovig. 2007. "The Human Genomic Melting Map." *PLOS Computational Biology* 3, no. 5 (May): e93. https://doi.org/10.1371/journal.pcbi.0030093. https://doi.org/10.1371/journal.pcbi.0030093.

Park, Peter J. 2009. "ChIP-seq: advantages and challenges of a maturing technology." *Nature Reviews Genetics* 10 (10): 669–680. https://doi.org/10.1038/nrg2641. https://www.nature.com/articles/nrg2641.

Zhang, Yuchuan, Zhikang Wang, Fang Ge, Xiaoyu Wang, Yiwen Zhang, Shanshan Li, Yuming Guo, Jiangning Song, and Dong-Jun Yu. 2024. "MLSNet: a deep learning model for predicting transcription factor binding sites." *Briefings in Bioinformatics* 25, no. 6 (September): bbae489. https://doi.org/10.1093/bib/bbae489. https://doi.org/10.1093/bib/bbae489.

Zhou, Jiyun, Qin Lu, Lin Gui, Ruifeng Xu, Yunfei Long, and Hongpeng Wang. 2019. "MTTFsite: cross-cell type TF binding site prediction by using multi-task learning." *Bioinformatics* 35, no. 24 (June): 5067–5077. https://doi.org/10.1093/bioinformatics/btz451. https://doi.org/10.1093/bioinformatics/btz451.