EpiBinder: a multimodal deep learning model at base-resolution to analyze in vivo Transcription Factor-DNA Binding

Ruben Solozabal [1] **Albert Baichorov** [1] **Tamir Avioz** [2] **Le Song** [1] **Martin Takáč** [1] **Ariel Afek** [2]

Abstract

Predicting in vivo transcription factor (TF) binding remains challenging due to interactions with cofactors and dynamic chromatin remodeling that modulate site accessibility. In this regard, accurate characterization of TF binding loci is essential for predictive performance. We introduce EpiBinder, a multimodal deep neural network that augments base-resolution DNA sequence models with single-nucleotide epigenetic information—cytosine methylation levels from whole-genome bisulfite sequencing and chromatin accessibility from DNase I hypersensitivity—to improve in vivo TF binding predictions. Trained on human cell-lines, EpiBinder demonstrates that integrating epigenetic information in a cell-type–specific manner reduces the epistemic uncertainty that current sequence-only DNA models suffer. Our model achieves up to a 14-point gain in area under the precision–recall curve compared the state-of-the-art. Our code is publicly available at GitHub.

1. Introduction

While cytosine methylation was previously thought to universally disrupt transcription factors (TFs), recent experiments *in vitro* reveal a wider map of methylation effects on TFs binding sites (Hernandez-Corchado & Najafabadi, 2022). In this work, we learn the impact of cytosine methylation from in-vivo data and use it to analyze its effect on TF binding. We extract binding occupancy de-novo from TF chromatin immunoprecipitation experiments (ChIP-seq) and connect it to key factors affecting in vivo binding, such as chromatin accessibility and base-resolution cytosine methylation levels. The extensive experimental data available in the ENCODE (de Souza, 2012) and ROADMAP (Kundaje et al., 2015) projects is used to build a cell-line specific deep-learning model that enables this association. Our work offers a biophysical interpretation of *in vivo* binding and suggests that current deep learning models can benefit from epigenetic markers currently available.

In order to learn the complex rules governing in vivo TF interactions with DNA in various cell conditions, we develop a deep learning model that integrates epigenetics markers at single base resolution. This model aims to improve current DNA model that only rely on the DNA sequence by incorporating cell-specific epigenetic information such as cytosine methylation and chromatin accessibility. We particularize our analysis on human cell-lines GM12878, K562, and HepG2, for which extensive ChIP-seq data for numerous TFs is available in the ENCODE project. Specifically, methylation levels are obtained from Whole-Genome Bisulfite Sequencing (WGBS) experiments for CpG, CHG and CHH sites; and the accessibility data from DNase I-hypersensitive (see Appendix A).

Our approach significantly surpasses current deep learning models in predicting in vivo TF binding accuracy that rely only on the DNA sequence (Ji et al., 2021; Zaheer et al., 2020; Zhou & Troyanskaya, 2015), revealing that epistemic uncertainty in current models can be reduced by incorporating epigenetic data. Furthermore, in this work, we use this model to investigate the effects of DNA methylation on transcription factor binding sites (TFBS) and analyze its impact on downstream gene expression regulation, causal variant interpretation, and epigenome-wide association studies (EWAS). A comprehensive overview of the experimental setup is provided in Figure 1.

2. Related work

Several works in the literature have approached the problem of predicting in vivo TF binding; however, the complexity of this problem resides in the interactions with other proteins, cofactors and the cell-type specific chromatin state that defines the physical accessibility of the binding site. DeepBind (Alipanahi et al., 2015) set a precedent by demonstrating that a deep learning approach

[1]MBZUAI - Mohamed bin Zayed University of Artificial Intelligence, UAE [2]Weizmann Institute of Science, Israel. Correspondence to: Ruben Solozabal <ruben.solozabal@mbzuai.ac.ae>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

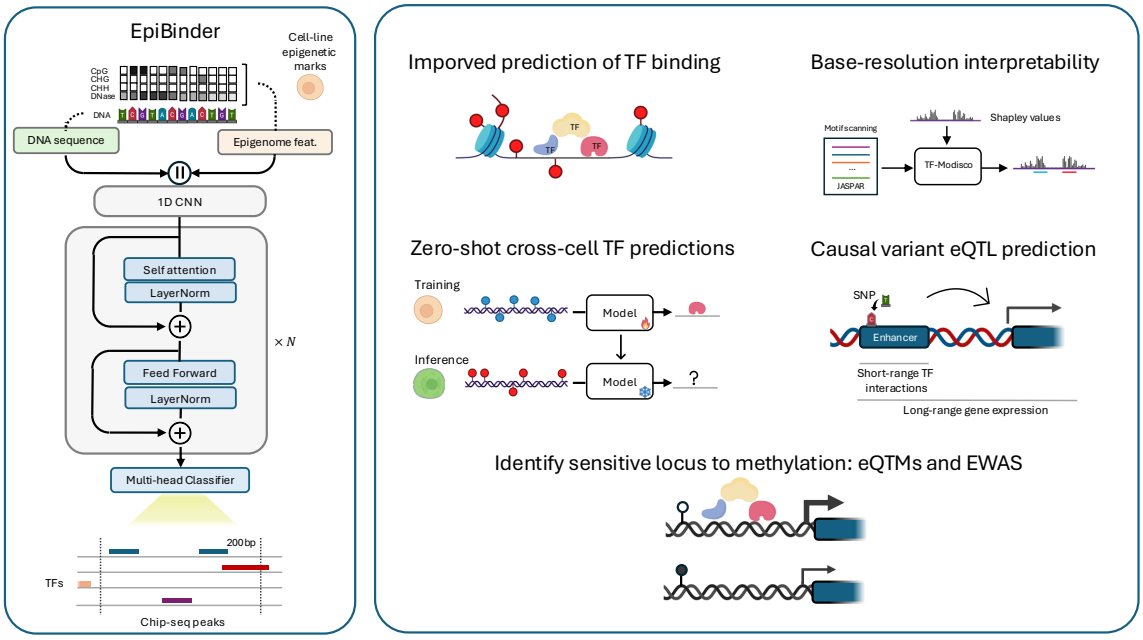


Figure 1. Overview of the EpiBinder framework and downstream tasks. **Left:** Multimodal architecture that encodes base-resolution DNA sequence alongside single-nucleotide epigenetic tracks (cytosine methylation and DNase I hypersensitivity) via parallel encoder streams, which are then fused to predict transcription factor binding. **Right:** Overview of the downstream tasks evaluated in this work, including zero-shot cross-cell TF prediction, causal variant eQTL prediction, and identification of methylation-sensitive loci via eQTMs and EWAS.

outperformed previous machine-learning methods in the DREAM5 competition (Marbach et al., 2012). Along these lines, DeepSEA (Zhou & Troyanskaya, 2015) is regarded as the first foundational model for transcriptional prediction, framing the binding prediction as a multi-task binary classification problem by training a single network on an extensive set of transcription factors. Novel approaches utilizing Masked Language Models (MLM) have been introduced into the problem as DnaBERT (Ji et al., 2021; Zhou et al., 2023) or the Nucleotide Transformer (Dalla-Torre et al., 2023). These approaches use a genome-wide pre-trained model, which is then fine-tuned for each of the downstream genomic tasks. Long context-aware models have also been utilized as Enformer (Avsec et al., 2021a) or HyenaDNA (Nguyen et al., 2024) with 200k bp sequence lengths and 1 million tokens, respectively. Although increasing the context length—and consequently the number of parameters—of these models substantially benefits genomics tasks that rely on long-range interactions (e.g., gene expression prediction), it offers no advantage for chromatin-state prediction, which depends on localized sequence context (Nguyen et al., 2024). In these scenarios, models with a localized context excel, e.g., HyenaDNA shows the best performance in the chromatin prediction task is achieved using a reduced 1kbp context.

Recent efforts to integrate epigenetic information into TF-binding prediction include analytical tools such as SEMplME (Nishizaki & Boyle, 2022), which correlate ChIP-seq peaks with local methylation levels to generate "methylation effect" logos. More closely related to our approach, (Hernandez-Corchado & Najafabadi, 2022) proposed a linear model that incorporates CpG methylation to predict TF ChIP-seq peaks. Other studies focus on directly predicting DNA methylation itself (e.g., DeepCpG (Angermueller et al., 2017)) or employ methylation-aware masked language models such as CpGPT (de Lima Camillo et al., 2024) and MethylGPT (Ying et al., 2024). In this work, we harness multiple epigenetic markers to uncover the mechanistic principles governing regulatory regions at single base resolution.

3. Main Results

3.1. Epigenetic markers enhance TF binding prediction

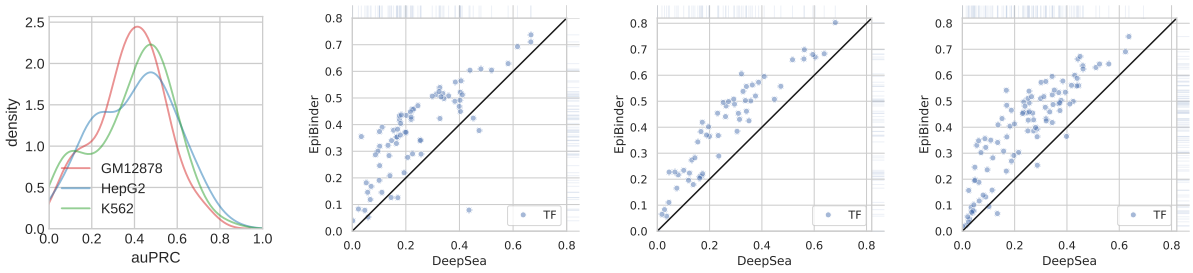
DNA methylation is a fundamental epigenetic mark that governs gene expression and chromatin organization. In this paper, we enrich the features used to train DNA models with cell-specific epigenetic information. In that sense, we train a deep learning model to learn the cell-specific dynamics. In order to evaluate this approach, we use a subset of the dataset introduced in (Zhou & Troyanskaya, 2015). This dataset is compiled from 919 chromatin features collected from

Table 1. Mean auPRC per cell-line grouped for Transcription factor binding (TF) and Polymerase (Pol). The same testing conditions on chromosomes 8 and 9 are preserved from (Zhou & Troyanskaya, 2015).

MODEL	LEN	GM12878		HepG2		K562	
		TF	Pol	TF	Pol	TF	Pol
DNA sequence							
DeepSea	1kbp	26.1	35.7	29.3	36.1	25.8	31.8
HyenaDNA	1kbp	26.9	37.6	30.5	33.3	26.5	32.3
DNABERT	512bp	26.7	35.3	29.1	35.7	26.0	31.3
BigBird	8kbp	27.4	34.8	31.7	34.4	26.7	30.9
DNA sequence + epigenetics							
EpiBinder	1kbp	**39.1**	**48.4**	**43.1**	**45.4**	**41.1**	**44.7**

Table 2. Histone mark predictions on K562 cell-line reported as auPRC. Testing conditions are preserved from (Zhou & Troyanskaya, 2015).

	DeepSea —40M—	HyenaDNA —7M—	DNABERT —86M—	BigBird —110M—	EpiBinder —46M—
H2AZ	0.45	0.43	0.42	0.44	**0.65**
H3K27ac	0.46	0.44	0.42	0.46	**0.70**
H3K27me3	0.09	0.08	0.08	0.09	**0.25**
H3K36me3	0.14	0.13	0.15	0.15	**0.42**
H3K4me1	0.33	0.34	0.31	0.35	**0.67**
H3K4me2	0.56	0.54	0.52	0.58	**0.78**
H3K4me3	0.58	0.56	0.54	0.60	**0.78**
H3K79me2	0.33	0.32	0.29	0.35	**0.47**
H3K9ac	0.52	0.50	0.49	0.53	**0.75**
H3K9me1	0.04	0.04	0.04	0.04	**0.08**
H3K9me3	0.06	0.06	0.10	0.07	**0.12**
H4K20me1	0.13	0.13	0.12	0.13	**0.21**



(a) auPRC per cell-line (b) GM12878 (c) HepG2 (d) K562

Figure 2. Results on TF-binding prediction. **(a)** Density distributions of per-cell auPRC for EpiBinder. **(b–d)** Scatterplots of per-TF auPRC comparing DeepSEA (x-axis) vs EpiBinder (y-axis) perdictions on (b) GM12878, (c) HepG2, and (d) K562 cell-lines.

Encode (de Souza, 2012) and Roadmap (Bernstein et al., 2010) projects. This collection includes 690 TF binding profiles for 160 different TFs collected from 148 different human cell lines. In this work, we particularized our study on the 3 most overrepresented cells in the dataset, GM12878, K562, and HepG2; which combined represent 47.4% of the dataset.

Despite the fact that many works target multiple cells in a multi-task learning approach (Zhou & Troyanskaya, 2015; Avsec et al., 2021a;b). We argue that learning a common representation from various cells is not always advantageous. For instance, while promoter regions tend to be more consistent across different cells, enhancers are more dependent on cell-specific conditions (Shigaki et al., 2019). Therefore, models that account for the unique chromatin conditions within each cell can significantly enhance performance.

Our results (see Tables 1 and Fig. 2), show that our multimodal approach improves the area under the precision-recall curve (auPRC) metric when compared to current state-of-the-art models on TF binding prediction as HyenaDNA (Nguyen et al., 2024), DNABert (Ji et al., 2021), BigBird (Zaheer et al., 2020) or DeepSEA (Zhou & Troyanskaya, 2015). We select auPRC metric as it is more appropriate for imbalanced datasets, rather than auROC reported in these works (see Appendix C), as the Chip-seq peaks are sparse, on average presenting ∼17k peaks genome-wide. As observed, despite the fact that each of our single cell-lines models is trained on significantly fewer data (i.e. chip-seq of TF belonging to the same cell), the auPRC on binding prediction increases significantly in almost every TF. Furthermore, Table 2 demonstrates that EpiBinder's multimodal integration also boosts the prediction of histone-mark occupancy. These results confirm that incorporating cell-specific epigenetic information substantially reduces the epistemic uncertainty inherent to sequence-only models, and this translates into improved accuracy not only for TF-binding but also for histone-modification prediction.

3.2. Zero-shot across cell-line predictions for TF binding

Currently, most of the TFs ChIP-seq data available belong to a limited number of cell-lines that have been well characterized in the literature, while for most of the cell-lines usually only a few key TFs are profiled. Therefore, transferring the knowledge from well-characterized to uncharacterized cell lines would a valuable asset. In that regard, we evaluate the zero-shot TF binding prediction capabilities of the model on

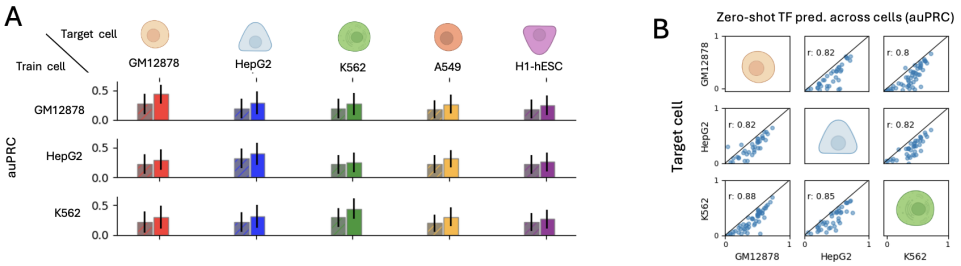


Figure 3. A. Zero-shot TF binding prediction across different cell-lines for the model including the epigenetic features (in color) or shadowed for the model relying only the nucleotide sequence. B. Correlation on the zero-shot predictions for a subset of TF concurrently available in the studied cell-lines.

unseen cell lines. We test a model trained on one reference cell in predicting TFs for a different cell line under different epigenetic conditions from the training ones. For each pair of cell lines, for all TFs whose data are shared between all of them, we evaluate auPRC obtained in zero-shot predictions. As summarized in Fig. 3, our model shows the ability to predict well, showing that epigenome markers account for a significant proportion of the TF prediction. Furthermore, as reflected in Table S6 for TFs with a small number of binding instances, zero-shot from a different cell-line could lead to better predictions when compared to a model directly trained on the target cell.

3.3. Identify CpGm loci associated with regulation

In this section, we apply EpiBinder to identify CpGm loci where methylation plays a critical role in transcription. For each CpG in the reference genome, we compare the predicted TF-binding score before and after setting its methylation level to zero, and flag as "high-sensitivity" those sites with Δ(binding score) exceeding a predefined threshold. Across GM12878, K562, and HepG2, we recover two classes of loci: (i) demethylation-enhanced sites, where loss of methylation substantially increases binding affinity, and (ii) demethylation repressing sites, which are rare and predominantly observed in K562 and HepG2.

To identify these loci, we integrated our model's predictions with data from the Illumina EPIC BeadChip—a high-throughput platform that profiles DNA methylation at over 850,000 CpG sites across the genome. By cross-referencing the EPIC BeadChip data with our model predictions, we pinpoint regions of potential regulatory significance.

In this experiment we evaluate whether methylation-driven changes in TF binding predict corresponding shifts in gene expression. We retrieve CpG–expression associations from the EWAS Atlas (Li et al., 2019), and—for each locus—compare the sign of the modeled Δ(binding score) to the sign of the observed methylation–expression correlation provided on that study. We particularize the study on CpGm

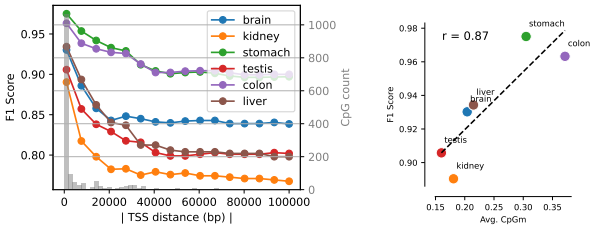


Figure 4. **Left:** F1 score for gene expression predictions based on distance to the nearest transcription start site (TSS). **Right:** F1 score as a function of average CpG methylation level in the tissue.

loci detected for HepG2. We summarize this agreement via a confusion matrix between predicted TF-binding effects and EWAS-reported expression changes.

Results are summarized in Figure 4(left), our *in silico* demethylation screen achieves F1-scores above 0.9 for close-by genes in all tissues evaluated. Predictive accuracy gradually declines with increasing distance from the transcription start site. Moreover, as shown in Figure 4(right), the magnitude of the predicted methylation effect correlates strongly with experimentally measured CpG methylation levels ($r = 0.87$) on the studied CpGs.

4. Conclusions

In this work, we demonstrated that augmenting base-resolution DNA sequence models with cell-line specific epigenetic information substantially enhances the predictive performance in transcription factor binding. Our multimodal framework effectively reduces epistemic uncertainty and yields more precise maps of regulatory interactions. We further showed its robust transferability to unseen cell types and its capacity to pinpoint functionally relevant methylation-sensitive loci. These results underscore the importance of embedding rich cellular context into current models and pave the way for future extensions to additional epigenetic marks and deeper effect prediction tasks, thereby deepening our mechanistic understanding of genome regulation.

References

Alipanahi, B., Delong, A., Weirauch, M. T., and Frey, B. J. Predicting the sequence specificities of dna-and rna-binding proteins by deep learning. *Nature biotechnology*, 33(8):831–838, 2015.

Angermueller, C., Lee, H. J., Reik, W., and Stegle, O. Deepcpg: accurate prediction of single-cell dna methylation states using deep learning. *Genome biology*, 18:1–13, 2017.

Avsec, Ž., Agarwal, V., Visentin, D., Ledsam, J. R., Grabska-Barwinska, A., Taylor, K. R., Assael, Y., Jumper, J., Kohli, P., and Kelley, D. R. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature methods*, 18(10):1196–1203, 2021a.

Avsec, Ž., Weilert, M., Shrikumar, A., Krueger, S., Alexandari, A., Dalal, K., Fropf, R., McAnany, C., Gagneur, J., Kundaje, A., et al. Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nature genetics*, 53(3):354–366, 2021b.

Bernstein, B. E., Stamatoyannopoulos, J. A., Costello, J. F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M. A., Beaudet, A. L., Ecker, J. R., et al. The nih roadmap epigenomics mapping consortium. *Nature biotechnology*, 28(10):1045–1048, 2010.

Dalla-Torre, H., Gonzalez, L., Mendoza-Revilla, J., Carranza, N. L., Grzywaczewski, A. H., Oteri, F., Dallago, C., Trop, E., de Almeida, B. P., Sirelkhatim, H., et al. The nucleotide transformer: Building and evaluating robust foundation models for human genomics. *BioRxiv*, pp. 2023–01, 2023.

de Lima Camillo, L. P., Sehgal, R., Armstrong, J., Higgins-Chen, A. T., Horvath, S., and Wang, B. Cpgpt: a foundation model for dna methylation. *bioRxiv*, pp. 2024–10, 2024.

de Souza, N. The encode project. *Nature methods*, 9(11): 1046–1046, 2012.

GENA_LM. GENA_LM. https://github.com/AIRI-Institute/GENA_LM/tree/main/downstream_tasks/DeepSea, 2024.

Hernandez-Corchado, A. and Najafabadi, H. S. Toward a base-resolution panorama of the in vivo impact of cytosine methylation on transcription factor binding. *Genome Biology*, 23(1):151, 2022.

Ji, Y., Zhou, Z., Liu, H., and Davuluri, R. V. Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. *Bioinformatics*, 37(15):2112–2120, 2021.

Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Moussavi, A. H., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., Amin, V., Whitaker, J. W., Schultz, M. D., Ward, L. D., Sarkar, A., Quon, G. T., Sandstrom, R. S., Eaton, M. L., Wu, Y., Pfenning, A. R., Wang, X., Claussnitzer, M., Liu, Y., Coarfa, C., Harris, R. A., Shoresh, N., Epstein, C. B., Gjoneska, E., Leung, D., Xie, W., Hawkins, R. D., Lister, R., Hong, C., Gascard, P., Mungall, A. J., Moore, R. A., Chuah, E., Tam, A., Canfield, T. K., Hansen, R. S., Kaul, R., Sabo, P. J., Bansal, M. S., Carles, A., Dixon, J. R., Farh, K. K., Feizi, S., Karlic, R., Kim, A., Kulkarni, A., Li, D., Lowdon, R. F., Elliott, G., Mercer, T. R., Neph, S. J., Onuchic, V., Polak, P., Rajagopal, N., Ray, P., Sallari, R. C., Siebenthall, K. T., Sinnott-Armstrong, N. A., Stevens, M., Thurman, R. E., Wu, J., Zhang, B., Zhou, X., Beaudet, A. E., Boyer, L. A., Jager, P. L. D., Farnham, P. J., Fisher, S. J., Haussler, D., Jones, S. J. M., Li, W., Marra, M. A., McManus, M. T., Sunyaev, S. R., Thomson, J. A., Tlsty, T. D., Tsai, L., Wang, W., Waterland, R. A., Zhang, M. Q., Chadwick, L. H., Bernstein, B. E., Costello, J. F., Ecker, J. R., Hirst, M., Meissner, A., Milosavljevic, A., Ren, B., Stamatoyannopoulos, J. A., Wang, T., and Kellis, M. Integrative analysis of 111 reference human epigenomes open. *Nat.*, 518(7539):317–330, 2015.

Li, M., Zou, D., Li, Z., Gao, R., Sang, J., Zhang, Y., Li, R., Xia, L., Zhang, T., Niu, G., Bao, Y., and Zhang, Z. EWAS atlas: a curated knowledgebase of epigenome-wide association studies. *Nucleic Acids Res.*, 47(Database-Issue): D983–D988, 2019.

Lipinski, J. Build DeepSEA training dataset. https://github.com/jakublipinski/build-deepsea-training-dataset, 2024.

Marbach, D., Costello, J. C., Küffner, R., Vega, N. M., Prill, R. J., Camacho, D. M., Allison, K. R., Kellis, M., Collins, J. J., et al. Wisdom of crowds for robust gene network inference. *Nature methods*, 9(8):796–804, 2012.

Nguyen, E., Poli, M., Faizi, M., Thomas, A., Wornow, M., Birch-Sykes, C., Massaroli, S., Patel, A., Rabideau, C., Bengio, Y., et al. Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. *Advances in neural information processing systems*, 36, 2024.

Nishizaki, S. S. and Boyle, A. P. Semplme: a tool for integrating dna methylation effects in transcription factor binding affinity predictions. *BMC bioinformatics*, 23(1): 317, 2022.

Shigaki, D., Adato, O., Adhikari, A. N., Dong, S., Hawkins-Hooker, A., Inoue, F., Juven-Gershon, T., Kenlay, H., Martin, B., Patra, A., et al. Integration of multiple epigenomic marks improves prediction of variant impact in

saturation mutagenesis reporter assay. *Human mutation*, 40(9):1280–1291, 2019.

Ying, K., Song, J., Cui, H., Zhang, Y., Li, S., Chen, X., Liu, H., Eames, A., McCartney, D. L., Marioni, R. E., et al. Methylgpt: a foundation model for the dna methylome. *bioRxiv*, pp. 2024–10, 2024.

Zaheer, M., Guruganesh, G., Dubey, K. A., Ainslie, J., Alberti, C., Ontanon, S., Pham, P., Ravula, A., Wang, Q., Yang, L., et al. Big bird: Transformers for longer sequences. *Advances in neural information processing systems*, 33:17283–17297, 2020.

Zhou, J. and Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning–based sequence model. *Nature methods*, 12(10):931–934, 2015.

Zhou, Z., Ji, Y., Li, W., Dutta, P., Davuluri, R., and Liu, H. Dnabert-2: Efficient foundation model and benchmark for multi-species genome. *arXiv preprint arXiv:2306.15006*, 2023.

A. Model design and training data

The dataset compiled in (Zhou & Troyanskaya, 2015) comprises 919 chromatin features across 148 human cell types, collected from ENCODE (de Souza, 2012) and ROADMAP (Bernstein et al., 2010). These include 690 transcription factor (TF) binding profiles covering 160 TFs, 125 DNase I hypersensitivity site (DHS) profiles, and 104 histone modification (HM) profiles. Each sample consists of a 1000-base-pair (bp) sequence from the hg19 human reference genome. Labels indicate the presence or absence of a peak for a given chromatin feature within the central 200 bp of the sequence, while the 400 bp flanking regions provide broader contextual information. The training and testing sets are split by chromosome to ensure strict non-overlap. In total, the dataset contains 2.2 million training samples, with 227,512 held-out samples from chromosomes 8 and 9 used for testing.

We augmented the dataset with epigenetic features corresponding to base-resolution methylation levels at CpG, CHH, and CHG derived from ENCODE Whole-Genome Bisulfite Sequencing, alongside DNase I hypersensitivity profiles (DNase-seq) to quantify chromatin accessibility. Details of these datasets, including accession identifiers are summarized in the following Table 3.

Table 3. References for epigenetic data on ENCODE project.

CELL		WGBS	DNASE-SEQ
GM12878	CpG	ENCFF570TIL	ENCFF264NMW
	CHH	ENCFF187KAK	
	CHG	ENCFF910HOG	
HepG2	CpG	ENCFF817LMT	ENCFF867UYB
	CHH	ENCFF158FZM	
	CHG	ENCFF101UQI	
K562	CpG	ENCFF660IHA	ENCFF352SET
	CHH	ENCFF294NMQ	
	CHG	ENCFF571AGF	
A549	CpG	ENCFF948WVD	ENCFF674WCO
	CHH	ENCFF589SGV	
	CHG	ENCFF461YYK	
H1-hESC	CpG	ENCFF434CNG	ENCFF233CHA
	CHH	ENCFF036NWK	
	CHG	ENCFF780ECA	

Model description

EpiBinder processes input tensors of shape $(B, L, 8)$ through a three-stage 1D convolutional backbone, where each stage comprises a convolutional layer with ReLU activation, followed by pooling and dropout. This design gradually reduces the sequence length while projecting features into a shared embedding space. Linear positional embeddings are then added to the resulting feature sequence, which is passed through a single Transformer encoder layer with 8 attention heads. Finally, the encoded representation is flattened and fed into a multilayer perceptron (MLP) to produce the class logits.

Table 4. Model config.

Stage	**Layer(s) & Configuration**	**Output Shape**
Input	`One-hot nucleotide + 4 normalized epigenetic`	$(B, L, 8)$
Conv Stage 1	`Conv1d(8→320, k=8, padding=4)`	$(B, \lfloor L/4 \rfloor, 320)$
Conv Stage 2	`Conv1d(320→480, k=8, padding=4)`	$(B, \lfloor L/16 \rfloor, 480)$
Conv Stage 3	`Conv1d(480→d_model, k=8, padding=4)`	$(B, L', d_{\text{model}})$
Positional Embedding	`Embedding(num_embeddings=`L'`, embedding_dim=`d_{model}`)`	$(B, L', d_{\text{model}})$
Transformer Encoder	$N\times$ `EncoderLayer(`d_{model}`,` n_{heads}`)`	$(B, L', d_{\text{model}})$
Flatten	—	$(B, L' \times d_{\text{model}})$
Classifier MLP	`MLP(`$L' \times d_{\text{model}} \to d_{\text{clf}} \to d_{\text{clf}} \to n_{\text{labels}}$`)`	(B, n_{labels})

Ablation experiments on performance

Ablation analysis of TF-binding prediction performance. Incorporating cytosine methylation or DNase I hypersensitivity individually into the sequence-only model each individually increases the auPRC. The combined "sequence + methylation

+ DNase" configuration yields the highest auPRC, demonstrating that methylation and chromatin-accessibility features contribute complementary improvements.

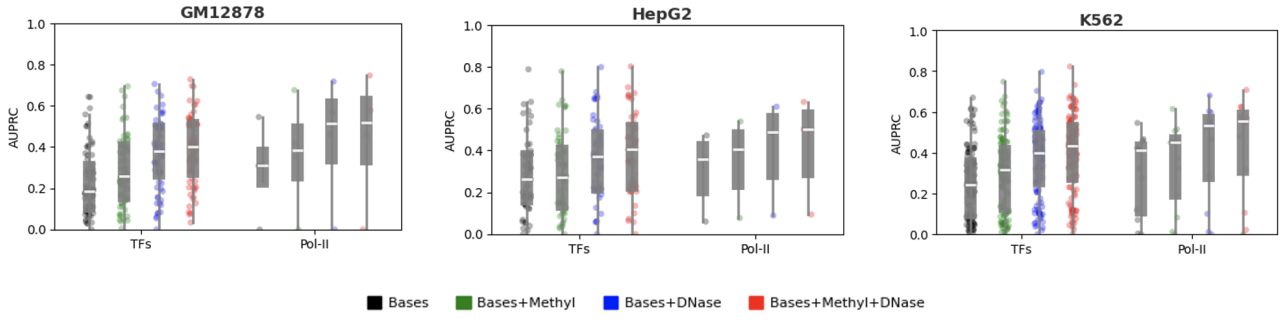


Figure 5. Performance of the single cell model in ablation studies: AUPRC obtained adding methylation and DNase features to base sequence training.

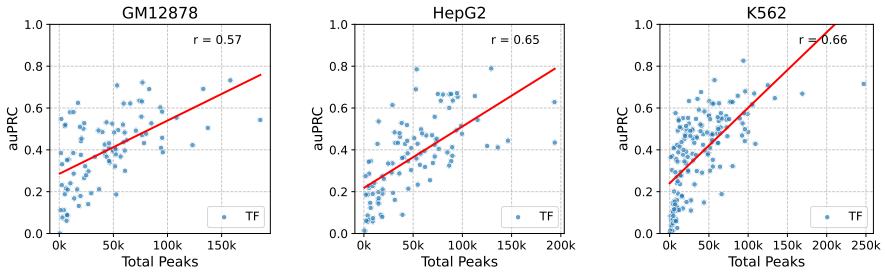


Figure 6. Scatter plot of transcription factor prediction performance versus training data availability. The x-axis shows the number of ChIP-seq peaks available for each TF, and the y-axis shows the corresponding in-domain auPRC. In red, the best-fit regression, highlighting the positive correlation between peak count and predictive accuracy.

B. Tools

To ensure all data reside on a common reference, we first converted each raw BED file from hg38 to hg19 coordinates using the UCSC liftOver utility with the hg38ToHg19 chain file. For our signal and annotation tracks in BigWig format, we applied CrossMap to perform the equivalent coordinate transformation directly on those indexed files. Finally, the merged, hg19-lifted BedGraph outputs were converted into indexed, binary BigWig coverage tracks using the UCSC bedGraphToBigWig tool.

C. Reproducing efforts on the Chromatin profile prediction task.

DeepSEA We reproduce the dataset construction and retrain the model following (Lipinski, 2024).

HyenaDNA. We fine-tune a HyenaDNA-7M model to reproduce the results obtained in the original work (Nguyen et al., 2024). For the downstream task of chromatin profile prediction, we utilize a pre-trained Hyena encoder in combination with sequence-level pooling and a fully connected decoder to perform multilabel sequence classification. In the original work, two variants with sequence lengths of 1k and 8k were fine-tuned. As reported in the paper, the 1k model outperforms the 8k model in predicting short-range tasks such as transcription factor (TF) binding. Therefore, we reproduce this setting for performance comparison.

GENA-LM We finetune a Gena-LMs Bert and BigBrid model using the downstream code provided in (GENA_LM, 2024).

Table 5. Performance comparison between reported in the original work and reproduced efforts on the Chromatin profile prediction task. Mean area under the ROC curve (auROC), averaged across three assay types: transcription factor binding profiles (TF), DNase I–hypersensitive sites (DHS), and histone modifications (HM).

Model	Params	Len	O/R	AUROC		
				TF	DHS	HM
Deepsea	40M	1kbp	orignal	94.7	91.5	85.2
			reprod.	93.4	90.7	84.2
HyenaDNA	7M	1kbp	orignal	96.4	93.0	86.3
			reprod.	95.0	91.6	84.9
DNABERT	86M	512bp	original	96.3	92.7	86.1
			reprod. (GENA-LM)	96.5	92.8	86.6
BigBird	110M	8kbp	original	96.1	92.1	88.7
			reprod. (GENA-LM)	96.8	92.0	85.3

D. Additional results.

Table 6. **Comparison of zero-shot performance across different cell-lines.** The top label indicates the cell-line the model was trained on, whereas the bottom label indicates the cell-line being tested. *InCell* refers to the scenario in which the model is both trained and tested on the same cell-line, included here for comparison. As observed, some cases show an improvement when transferring prediction from one cell-line to another.

Source cell	GM12878			HepG2			K562		
Target cell	*in-domain*	HepG2	K562	*in-domain*	GM12878	K562	*in-domain*	HepG2	GM12878
SMC3	**0.73**	0.60	0.69	**0.67**	0.65	0.60	**0.73**	0.59	0.70
CTCF	**0.65**	0.64	0.55	**0.72**	0.54	0.53	**0.65**	0.63	0.55
Rad21	**0.63**	0.61	0.55	**0.65**	0.57	0.56	**0.64**	0.59	0.54
GABP	0.61	**0.64**	0.47	**0.65**	0.57	0.46	0.58	**0.59**	0.50
TAF1	0.51	**0.56**	0.49	**0.70**	0.39	0.47	0.58	**0.62**	0.45
MAZ	**0.57**	0.44	0.52	**0.52**	0.50	0.49	**0.67**	0.26	0.39
ELF1	**0.62**	0.36	0.41	**0.50**	0.42	0.41	**0.53**	0.44	0.49
NRSF	**0.25**	**0.25**	0.15	**0.48**	0.34	0.25	**0.23**	0.09	0.03
Nrf1	**0.51**	0.32	0.37	**0.42**	0.28	0.36	**0.41**	0.40	0.35
USF1	**0.50**	0.36	0.35	**0.56**	0.37	0.38	**0.48**	0.41	0.35
Max	**0.50**	0.37	0.39	**0.50**	0.45	0.41	**0.61**	0.42	0.38
YY1	**0.34**	0.20	0.17	**0.52**	0.17	0.31	0.46	**0.50**	0.19
USF2	**0.53**	0.39	**0.53**	**0.57**	0.40	0.49	**0.57**	0.42	0.41
Mxi1	**0.51**	0.48	0.26	**0.58**	0.45	0.26	0.38	**0.52**	0.47
CEBPB	**0.30**	0.03	0.04	**0.53**	0.04	0.22	**0.48**	0.31	0.07
TBP	**0.42**	0.37	0.27	**0.44**	0.35	0.32	**0.48**	0.41	0.31
JunD	**0.34**	0.05	0.08	**0.54**	0.04	0.23	**0.57**	0.23	0.05
SP1	**0.54**	0.18	0.34	**0.45**	0.28	0.15	**0.45**	0.23	0.43
c-Myc	**0.37**	0.10	0.21	0.26	0.20	**0.31**	**0.50**	0.25	0.15
ATF3	**0.30**	0.28	0.19	**0.35**	0.23	0.28	**0.39**	0.24	0.12
CHD2	**0.51**	0.15	0.26	0.23	**0.40**	0.27	0.34	0.18	**0.38**
BHLHE40	**0.43**	0.21	0.25	**0.31**	0.29	0.24	**0.42**	0.24	0.30
p300	**0.45**	0.09	0.08	**0.38**	0.21	0.10	**0.36**	0.13	0.19
RFX5	**0.28**	0.20	0.05	**0.29**	0.19	0.07	0.10	**0.21**	0.13
ZNF274	0.03	0.00	**0.10**	0.00	**0.30**	0.00	**0.39**	0.00	0.00
COREST	**0.20**	0.13	0.10	**0.22**	0.11	0.13	**0.39**	0.09	0.06
TR4	**0.11**	0.06	0.01	**0.20**	0.02	0.04	0.04	**0.15**	0.02
SRF	**0.20**	0.04	0.09	0.06	**0.10**	0.08	**0.14**	0.05	**0.14**
ZBTB33	**0.13**	0.08	0.05	**0.20**	0.10	0.05	**0.13**	0.06	0.07