An Interventional Framework of Multimodal Epigenomic Regulation for Gene Expression Prediction

Zhao Yang [*1] **Yi Duan** [*1] **Jiwei Zhu** [1] **Chuan Cao** [2] **Bing Su** [1]

Abstract

Gene expression prediction through DNA sequences and multimodal epigenomic signals integration presents significant challenges. Previous methods often focus on using epigenomic signals to locate distal enhancers and incorporate these enhancers into model development through long sequence modeling. Our experiments reveal that current long sequence modeling actually decreases performance, while proximal signals near target genes prove more essential. Furthermore, we find that different signals contribute varying degrees of performance gain. Simple use of all epigenomic signals may lead models to depend excessively on widespread background signals. These background signals act as confounders, causing the model to develop spurious dependencies. To overcome these issues, we propose InFER, which employs causal intervention through backdoor adjustment to eliminate model dependencies on potential confounding background epigenomic regulation. Our experimental results show that proper modeling of epigenomic regulation with short sequences alone can achieve state-of-the-art performance in gene expression prediction.

1. Introduction

Understanding and predicting gene expression is fundamental to deciphering the complex regulatory mechanisms governing cellular functions (Pratapa et al., 2020). Accurate gene expression prediction enables breakthroughs across biomedicine (Mamoshina et al., 2016), from unraveling disease pathogenesis (Cookson et al., 2009; Emilsson et al., 2008), to enabling personalized therapeutic strategies (Blass & Ott, 2021). However, accurately predicting gene expression presents significant challenges. Previous methods primarily focused on modeling long sequences, given that regulatory elements can act over considerable distances (Avsec et al., 2021; Su et al., 2025).

Deep learning methods for gene expression prediction have rapidly advanced. DNA language models such as (Avsec et al., 2021; Dalla-Torre et al., 2024; Linder et al., 2025; Nguyen et al., 2024b; Schiff et al., 2024) predict directly from DNA sequences, but are limited by the absence of epigenomic signals. Conversely, GraphReg (Karbalayghareh et al., 2022) uses only epigenomic signals, ignoring DNA information. More recent approaches like (Lin et al., 2024; Li et al., 2023) integrate both, first identifying potential enhancers via epigenomic signals and then modeling these regions with target genes due to distal enhancer constraints. Seq2Exp (Su et al., 2025) now learns these enhancers data-drivenly, masking irrelevant regions for full-sequence input to achieve state-of-the-art (SOTA) results. Detailed related works are in Appendix A.

In this work, we first challenge the long sequence modeling paradigm (Su et al., 2025; Schiff et al., 2024). While we acknowledge that modeling long sequences is biologically necessary (Schoenfelder & Fraser, 2019), current approaches to long sequence modeling exhibit notable limitations: (1) (Lin et al., 2024; Li et al., 2023) utilize only statistically identified potential regulatory elements (Fulco et al., 2019) while ignoring other regions that may also have functional effects. (2) Seq2Exp (Su et al., 2025) and a series of DNA language models (Nguyen et al., 2024b; Schiff et al., 2024) employ Mamba (Gu & Dao, 2023) or other State Space Models (SSMs). Although SSMs can efficiently process extra-long sequences like DNA with linear complexity, their modeling capacity is questionable (Wang et al., 2025).

We conducted preliminary experiments to validate our challenge. Specifically, we trained Caduceus (Schiff et al., 2024) and Seq2Exp (Su et al., 2025) with varying input lengths centered at the transcription start site (TSS) for gene expression prediction. Caduceus uses complete sequences as input, while Seq2Exp learns masks to filter out potentially non-functional regions. According to Figure 1 (a), we observe that Caduceus's performance consistently declines after input lengths exceed 2k. Seq2Exp doesn't show a

[*]Equal contribution [1]Renmin University of China, Beijing, China [2]Zhongguancun Academy, Beijing, China. Correspondence to: Bing Su <bingsu@ruc.edu.cn>.

Proceedings of the ICML 2025 Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences, Vancouver, Canada. 2025.

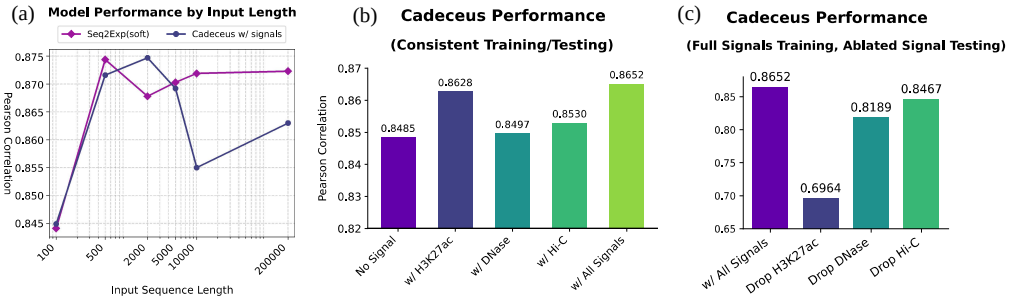

Figure 1. (a) Performance of Seq2Exp (Su et al., 2025) and Caduceus (Schiff et al., 2024) with varying input sequence lengths. (b) Caduceus performance with different epigenomic signals. (c) Performance degradation when specific signals are removed during testing from a model trained with all signals.

clear downward trend, but its performance with 200k input length remains essentially equivalent to using just 500 bases. Therefore, we can confidently conclude that using SSMs to model long sequences currently leads to performance degradation. Even Seq2Exp's masking approach only filters some irrelevant regions, limiting performance decline, but fails to leverage long sequence modeling to enhance performance. Table 2 in Appendix B demonstrates that the Seq2Exp model pre-trained on 200k sequences maintains nearly identical performance even when input sequences are shortened to 2k during the testing phase. We attribute the effectiveness of short sequences to that proximal epigenomic signals reflect the activity of distal regulatory elements through chromatin looping and spatial interactions (Plank & Dean, 2014). Therefore, rather than extending sequence length, we focus on more effectively leveraging these informative epigenomic signals to enhance prediction performance.

Previous approaches typically utilize epigenomic signals through simple feature concatenation, without considering their distinct biological roles and potential interdependencies (Lin et al., 2024; Su et al., 2025). In this work, we conducted an extensive comparative study to characterize the differential contributions of various epigenomic signals. We trained Caduceus using DNA sequence alone and with the addition of either H3K27ac, DNase, Hi-C, or all signals combined. Figure 1 (b) reveals that each signal improves performance, with H3K27ac showing the most substantial enhancement. This aligns with biological understanding, as H3K27ac directly marks active regulatory elements like promoters and enhancers (Creyghton et al., 2010), functioning as a specific "foreground" signal with direct causal influence on gene expression. In contrast, DNase and Hi-C provided smaller improvements as they predominantly serve as "background" signals, indicating chromatin accessibility (Thurman et al., 2012) and three-dimensional organization (Rao et al., 2014), respectively. Interestingly, models trained on all signals combined performed comparably to those trained solely on H3K27ac, indicating that these background signals provide almost no additional improvement when H3K27ac is already known.

More revealing is our study in Figure 1 (c), where removing background signals during testing from a model trained on all signals causes significant performance degradation. This creates a paradox with our observations in Figure 1 (b) - while background signals themselves provide minimal performance improvement, models that use these signals during training develop an over-dependence on them, resulting in substantial performance drops when these signals are removed. This asymmetric impact pattern strongly suggests these signals play a confounding role rather than a causal one. We attribute this phenomenon to background epigenomic signals acting as confounders, which are broadly present across many genes but do not causally regulate gene expression directly (Schreiber et al., 2020). The original data for Figure 1 is in Appendix C. Evidence from our case study supporting the treatment of background signals as confounders is presented in Appendix D.

These background epigenomic signals create spurious correlations by influencing both the observed epigenomic landscape and gene expression levels, while not directly participating in the causal gene regulation mechanism. To address this issue, we employ a Structural Causal Model (SCM) to characterize background epigenomic signals as confounders in gene expression prediction, and further apply backdoor adjustment (Pearl et al., 2016) to eliminate their confounding effects, better utilizing epigenomic regulation for prediction. We name our method InFER (an **In**terventional **F**ramework of **E**pigenomic **R**egulation).

2. Related Works

Sequence-to-function models are designed to predict functional genomic signals directly from DNA sequences. DeepSEA (Zhou & Troyanskaya, 2015) established this approach by utilizing convolutional neural networks (CNNs) to extract sequence features for multi-task prediction. The field has evolved through architectural innovations and expanded training datasets (Kelley et al., 2018; Zhou et al., 2018; Chen et al., 2022). Currently, Enformer (Avsec et al., 2021) represents the leading methodology, achieving exceptional performance through its hybrid Transformer-CNN

architecture. While these models simultaneously predict various outputs including epigenomic signals and gene expression levels, they typically lack specialized mechanisms for leveraging epigenomic data to enhance expression prediction specifically, treating all prediction targets as parallel outputs rather than considering their biological interdependencies.

Unsupervised DNA foundation models leverage the successful paradigm of unsupervised pre-training established in natural language processing. DNABERT (Ji et al., 2021) was the first to adapt this approach to genomics, applying BERT-like (Devlin et al., 2019) techniques to learn transferable DNA representations. Subsequent models have expanded upon this foundation (Zhou et al., 2024; Dalla-Torre et al., 2024; Li et al., 2024; Sanabria et al., 2024). In parallel, generative frameworks like Evo (Nguyen et al., 2024a) have emerged (Nguyen et al., 2024b; Brixi et al., 2025), enabling functional element design applications (Linder et al., 2025; Yang et al., 2025). Despite these advances, such models' effectiveness for gene expression prediction remains limited due to their exclusive reliance on DNA sequence information, without incorporating the critical epigenomic context that modulates gene activity.

Gene expression prediction represents a fundamental challenge in bioinformatics (Segal et al., 2002). Early approaches like Enformer (Avsec et al., 2021) attempted to predict gene expression directly from DNA sequences, facing inherent limitations, while GraphReg (Karbalayghareh et al., 2022) enhanced performance by incorporating epigenomic information through graph attention networks to model physical interactions between genomic regions. More recent methods have progressed toward integrating both sequence and epigenomic information, with Creator (Li et al., 2023) and EPInformer (Lin et al., 2024) demonstrating improved performance through this combined approach. However, these models typically rely on pre-identified regulatory elements, overlooking potential contributions from unannotated regions. Seq2Exp (Su et al., 2025) addressed this limitation through an end-to-end, data-driven methodology that simultaneously learns to identify relevant regulatory elements and predict expression with epigenomic guidance. Despite these advances, current research tends to focus predominantly on modeling distal regulatory elements through long sequence architectures, rather than optimizing the utilization of biologically interrelated epigenomic signals that directly influence gene regulation.

3. Method

3.1. Problem Formulation

Given a gene sequence $X = [x_1, x_2, \ldots, x_L]$, where for each $i \in \{1, 2, \ldots, L\}$, $x_i \in \mathbb{R}^4$ represents the one-hot encoding of a nucleotide base from the set $V = \{A, T, C, G\}$, and L denotes the sequence length surrounding the gene's TSS (Lin et al., 2024; Su et al., 2025). For each X, there are associated epigenomic signals $S = [s_1, s_2, \ldots, s_L]$, where $s_i \in \mathbb{R}^d$ with d representing the number of epigenomic signals. Our approach first employs a signal encoder $g_\theta : \mathbb{R}^{L\times d} \to \mathbb{R}^{L\times d'}$ with parameters θ to map the raw epigenomic signals S into a higher-dimensional feature space $H = g_\theta(S)$, where d' represents the dimensionality of this enriched representation following (Su et al., 2025). We then use a predictor network $h_\phi : (\mathbb{R}^{L\times 4}, \mathbb{R}^{L\times d'}) \to \mathbb{R}$ with parameters ϕ that integrates both sequence information X and encoded epigenomic features H to predict gene expression levels $Y \in \mathbb{R}$. To optimize our model parameters $\{\theta, \phi\}$, we define the following objective function:

$$\mathcal{L}_1 = \text{dist}(h_\phi(X, g_\theta(S)), Y), \tag{1}$$

where $\text{dist}(\cdot, \cdot)$ is a distance metric measuring the discrepancy between predicted and true gene expression values.

3.2. Structural Causal Model

In Section 1, we categorize H3K27ac as foreground signal and DNase/Hi-C as background signals. However, instead of relying on this biology-informed categorization, we introduce a more general confounder C representing background epigenomic regulation as a combined effect of different signal types. Drawing inspiration from computer vision, where CNNs transform RGB images into feature maps with channel combinations representing various background information (Qiang et al., 2022; Zhou et al., 2016), we model C in the high-dimensional latent space of encoded epigenomic signals H. The implementation of C is detailed in Section 3.4.

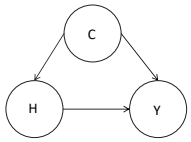

Figure 2. The proposed SCM.

Biologically, C represents background regulatory states with genome-wide effects, such as global chromatin accessibility. These background factors simultaneously influence both the epigenomic landscape and gene expression without participating in gene-specific regulation. For instance, broadly accessible chromatin regions correlate with higher expression across many genes, confounding the true causal relationship between specific regulatory elements and their targets (Schreiber et al., 2020).

Figure 2 depicts our SCM showing relationships between H, C, and Y. For clarity, we omit X from the graph, though our model ultimately uses both X and H for prediction. $H \to Y$: High-dimensional epigenomic signals H are used to predict gene expression Y, utilizing both foreground and background information. $H \leftarrow C \to Y$: Background information C directly contributes to observed signals H

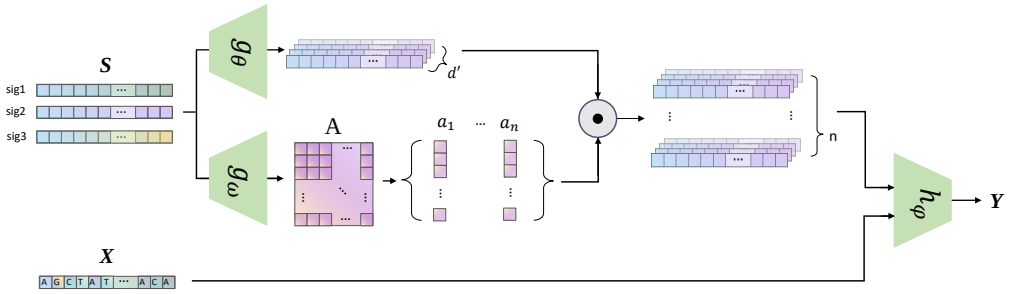

Figure 3. Architecture of InFER

.

and reflects the co-occurrence of open chromatin and highly expressed genes, creating a confounding effect.

An ideal prediction model should capture the true causal relationship between H and Y. We expect Y to be primarily influenced by foreground signals directly indicating regulatory activity, not merely by broadly accessible chromatin. However, our SCM reveals that $P(Y|H)$ includes both the direct path $H \rightarrow Y$ and spurious correlations via $H \leftarrow C \rightarrow Y$. This explains our observation in Figure 1 (c), where removing background signals during testing significantly reduces performance, indicating excessive dependency on these signals.

3.3. Causal Intervention via Backdoor Adjustment

To capture true causal relationships, we must estimate the interventional distribution $P(Y|do(H))$ rather than the conditional distribution $P(Y|H)$. The do operator (Pearl et al., 2016) represents an intervention that sets the value of H while removing its dependency on confounders, enabling us to isolate the direct causal effect of epigenomic signals on gene expression. Specifically, we can stratify the background information confounder C into n different background epigenomic regulation, i.e., $C = \{C_1, C_2, ..., C_n\}$ where n is a hyperparameter. Each C_i represents a distinct background regulatory pattern. Formally, we can formulate the backdoor adjustment for our proposed SCM as:

$$P(Y|do(H)) = \sum_{i=1}^{n} P(Y|H, C = C_i)P(C = C_i). \quad (2)$$

This intervention allows us to estimate the direct causal effect of epigenomic signals on gene expression, independent of shared background regulatory patterns. To make this computation tractable, we adopt the assumption that the latent confounding variable C follows a uniform distribution (Qiang et al., 2022): $P(C = C_i) = \frac{1}{n} \quad \forall C_i \in C$, where n is the number of different background regulatory patterns we model.

3.4. Implementation of Confounder

While the concept of background epigenomic regulation represented by C_i is sound, directly identifying specific cellular states through biological priors (Section 1) would be oversimplified, as the actual background regulatory information is likely a combination of different epigenomic signals. Instead, we implement a data-driven approach inspired by computer vision techniques, where different channels can extract hierarchical background information (Qiang et al., 2022; Zhou et al., 2016).

Shown in Figure 3, We employ a confounder encoder $g_\omega : \mathbb{R}^{L\times d} \rightarrow \mathbb{R}^{n\times d'}$ with parameters ω to extract potential background patterns from the epigenomic signals S. This network produces weight vectors $A = [a_1, a_2, \ldots, a_n]$, where each $a_i \in \mathbb{R}^{d'}$ represents a different weighting scheme over the epigenomic signal dimensions. Note that these weights are gene-wise rather than position-wise, as we assume the background regulatory patterns are shared across the entire gene region. Each weight vector a_i effectively models a distinct background regulatory state C_i by capturing a specific combination of epigenomic signal contributions. For example, one weight vector might emphasize chromatin accessibility signals for genes in active cell cycle states, while another might highlight three-dimensional organizational features in quiescent cells. Through learning these diverse weighting schemes, our model can represent the complex, multifaceted nature of background epigenomic regulation that acts as confounders in gene expression prediction. Following our backdoor adjustment formula, we directly compute:

$$\begin{aligned} P(Y|do(H)) &= \sum_{i=1}^{n} P(Y|H, C = C_i)P(C = C_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} h_\phi(X, H \odot a_i). \end{aligned} \quad (3)$$

We include X in this formulation because we assume DNA sequence X and epigenomic features H are independent (Su et al., 2025), which is why X does not appear in our SCM. Each weighted version of H is then processed through our predictor to approximate the interventional distribution. We

Table 1. Performance on Gene Expression CAGE Prediction with Standard Deviation for Cell Types K562 and GM12878.

	K562			GM12878		
	MSE ↓	MAE ↓	Pearson ↑	MSE ↓	MAE ↓	Pearson ↑
Enformer	0.2920 ± 0.0050	0.4056 ± 0.0040	0.7961 ± 0.0019	0.2889 ± 0.0009	0.4185 ± 0.0013	0.8327 ± 0.0025
HyenaDNA	0.2265 ± 0.0013	0.3497 ± 0.0012	0.8425 ± 0.0008	0.2217 ± 0.0018	0.3562 ± 0.0012	0.8729 ± 0.0010
Mamba	0.2241 ± 0.0027	0.3416 ± 0.0026	0.8412 ± 0.0021	0.2145 ± 0.0021	0.3446 ± 0.0022	0.8788 ± 0.0011
Caduceus	0.2197 ± 0.0038	0.3327 ± 0.0070	0.8475 ± 0.0014	0.2124 ± 0.0037	0.3436 ± 0.0031	0.8819 ± 0.0009
Caduceus w/ signals	0.1959 ± 0.0036	0.3187 ± 0.0036	0.8630 ± 0.0008	0.1942 ± 0.0058	0.3269 ± 0.0048	0.8928 ± 0.0017
EPInformer	0.2140 ± 0.0042	0.3291 ± 0.0031	0.8473 ± 0.0017	0.1975 ± 0.0031	0.3246 ± 0.0025	0.8907 ± 0.0011
MACS3	0.2195 ± 0.0023	0.3455 ± 0.0018	0.8435 ± 0.0013	0.2340 ± 0.0028	0.3654 ± 0.0017	0.8634 ± 0.0020
Seq2Exp-hard	0.1863 ± 0.0051	0.3074 ± 0.0036	0.8682 ± 0.0045	0.1890 ± 0.0045	0.3199 ± 0.0040	0.8916 ± 0.0027
Seq2Exp-soft	0.1856 ± 0.0032	0.3054 ± 0.0024	0.8723 ± 0.0012	0.1873 ± 0.0044	0.3137 ± 0.0028	0.8951 ± 0.0038
InFER	**0.1789 ± 0.0041**	**0.3000 ± 0.0058**	**0.8751 ± 0.0036**	**0.1759 ± 0.0054**	**0.3054 ± 0.0048**	**0.9016 ± 0.0024**

incorporate this interventional prediction as a regularization term, forming our second loss component:

$$\mathcal{L}_2 = \text{dist}\left(\frac{1}{n}\sum_{i=1}^{n} h_\phi(X, H \odot a_i), Y\right), \tag{4}$$

which follows the same structure as our standard prediction loss in Equation 1 but operates on the interventional predictions after backdoor adjustment. The complete training objective is detailed in Appendix E. The algorithm workflow is in Appendix F.

4. Experiments

Detailed experimental setup, baseline descriptions, and additional ablation studies are in Appendix G. Here we present the main results. Table 1 presents performance results across all methods for the K562 and GM12878 cell types, respectively. All baseline results are directly cited from Seq2Exp (Su et al., 2025) to ensure fair comparison. Additionally, all results reported include the mean and standard deviation from five runs using different random seeds: {2, 22, 222, 2222, 22222} following (Su et al., 2025). The best-performing method for each metric is highlighted in bold, with the second-best underlined. Notably, our InFER consistently outperforms the previous SOTA Seq2Exp-soft across all datasets and metrics.

5. Conclusion

In this work, we introduce InFER. Our causal intervention approach through backdoor adjustment effectively distinguishes genuine regulatory relationships from spurious correlations, achieving SOTA performance using only short sequences.

References

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, 2021.

Blass, E. and Ott, P. A. Advances in the development of personalized neoantigen-based therapeutic cancer vaccines. *Nature reviews Clinical oncology*, 18(4):215–229, 2021.

Brixi, G., Durrant, M. G., Ku, J., Poli, M., Brockman, G., Chang, D., Gonzalez, G. A., King, S. H., Li, D. B., Merchant, A. T., et al. Genome modeling and design across all domains of life with evo 2. *BioRxiv*, pp. 2025–02, 2025.

Chen, K. M., Wong, A. K., Troyanskaya, O. G., and Zhou, J. A sequence-based global map of regulatory activity for deciphering human genetics. *Nature genetics*, 54(7): 940–949, 2022.

Consortium, E. P. et al. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57, 2012.

Cookson, W., Liang, L., Abecasis, G., Moffatt, M., and Lathrop, M. Mapping complex disease traits with global gene expression. *Nature Reviews Genetics*, 10(3):184–194, 2009.

Creyghton, M. P., Cheng, A. W., Welstead, G. G., Kooistra, T., Carey, B. W., Steine, E. J., Hanna, J., Lodato, M. A., Frampton, G. M., Sharp, P. A., et al. Histone h3k27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences*, 107(50):21931–21936, 2010.

Dalla-Torre, H., Gonzalez, L., Mendoza-Revilla, J., Lopez Carranza, N., Grzywaczewski, A. H., Oteri, F., Dallago, C., Trop, E., de Almeida, B. P., Sirelkhatim, H., et al. Nucleotide transformer: building and evaluating robust foundation models for human genomics. *Nature Methods*, pp. 1–11, 2024.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, 2019.

Emilsson, V., Thorleifsson, G., Zhang, B., Leonardson, A. S., Zink, F., Zhu, J., Carlson, S., Helgason, A., Walters, G. B., Gunnarsdottir, S., et al. Genetics of gene expression and its effect on disease. *Nature*, 452(7186): 423–428, 2008.

Fulco, C. P., Nasser, J., Jones, T. R., Munson, G., Bergman, D. T., Subramanian, V., Grossman, S. R., Anyoha, R., Doughty, B. R., Patwardhan, T. A., et al. Activity-by-contact model of enhancer–promoter regulation from thousands of crispr perturbations. *Nature genetics*, 51 (12):1664–1669, 2019.

Gu, A. and Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

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.

Karbalayghareh, A., Sahin, M., and Leslie, C. S. Chromatin interaction–aware gene regulatory modeling with graph attention networks. *Genome Research*, 32(5):930–944, 2022.

Kelley, D. R., Reshef, Y. A., Bileschi, M., Belanger, D., McLean, C. Y., and Snoek, J. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome research*, 28(5):739–750, 2018.

Li, S., Wang, Z., Liu, Z., Wu, D., Tan, C., Zheng, J., Huang, Y., and Li, S. Z. Vqdna: Unleashing the power of vector quantization for multi-species genomic sequence modeling. In *Forty-first International Conference on Machine Learning*, 2024.

Li, Y., Ju, F., Chen, Z., Qu, Y., Xia, H., He, L., Wu, L., Zhu, J., Shao, B., and Deng, P. Creator: zero-shot cis-regulatory pattern modeling with attention mechanisms. *Genome Biology*, 24(1):266, 2023.

Lin, J., Luo, R., and Pinello, L. Epinformer: a scalable deep learning framework for gene expression prediction by integrating promoter-enhancer sequences with multimodal epigenomic data. *bioRxiv*, pp. 2024–08, 2024.

Linder, J., Srivastava, D., Yuan, H., Agarwal, V., and Kelley, D. R. Predicting rna-seq coverage from dna sequence as a unifying model of gene regulation. *Nature Genetics*, pp. 1–13, 2025.

Mamoshina, P., Vieira, A., Putin, E., and Zhavoronkov, A. Applications of deep learning in biomedicine. *Molecular pharmaceutics*, 13(5):1445–1454, 2016.

Nguyen, E., Poli, M., Durrant, M. G., Kang, B., Katrekar, D., Li, D. B., Bartie, L. J., Thomas, A. W., King, S. H., Brixi, G., et al. Sequence modeling and design from molecular to genome scale with evo. *Science*, 386(6723): eado9336, 2024a.

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, 2024b.

Pearl, J., Glymour, M., and Jewell, N. P. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

Plank, J. L. and Dean, A. Enhancer function: mechanistic and genome-wide insights come together. *Molecular cell*, 55(1):5–14, 2014.

Pratapa, A., Jalihal, A. P., Law, J. N., Bharadwaj, A., and Murali, T. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nature methods*, 17(2):147–154, 2020.

Qiang, W., Li, J., Zheng, C., Su, B., and Xiong, H. Interventional contrastive learning with meta semantic regularizer. In *International conference on machine learning*, pp. 18018–18030. PMLR, 2022.

Rao, S. S., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., Sanborn, A. L., Machol, I., Omer, A. D., Lander, E. S., et al. A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7):1665–1680, 2014.

Sanabria, M., Hirsch, J., Joubert, P. M., and Poetsch, A. R. Dna language model grover learns sequence context in the human genome. *Nature Machine Intelligence*, 6(8): 911–923, 2024.

Schiff, Y., Kao, C. H., Gokaslan, A., Dao, T., Gu, A., and Kuleshov, V. Caduceus: Bi-directional equivariant long-range dna sequence modeling. In *International Conference on Machine Learning*, pp. 43632–43648. PMLR, 2024.

Schoenfelder, S. and Fraser, P. Long-range enhancer–promoter contacts in gene expression control. *Nature Reviews Genetics*, 20(8):437–455, 2019.

Schreiber, J., Durham, T., Bilmes, J., and Noble, W. S. Avocado: a multi-scale deep tensor factorization method learns a latent representation of the human epigenome. *Genome biology*, 21:1–18, 2020.

Segal, E., Barash, Y., Simon, I., Friedman, N., and Koller, D. From promoter sequence to expression: a probabilistic framework. In *Proceedings of the sixth annual international conference on Computational biology*, pp. 263–272, 2002.

Su, X., Yu, H., Zhi, D., and Ji, S. Learning to discover regulatory elements for gene expression prediction. In *The Thirteenth International Conference on Learning Representations*, 2025.

Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., Sheffield, N. C., Stergachis, A. B., Wang, H., Vernot, B., et al. The accessible chromatin landscape of the human genome. *Nature*, 489 (7414):75–82, 2012.

Wang, P., Cai, R., Wang, Y., Zhu, J., Srivastava, P., Wang, Z., and Li, P. Understanding and mitigating bottlenecks of state space models through the lens of recency and oversmoothing. In *The Thirteenth International Conference on Learning Representations*, 2025.

Wang, T. and Isola, P. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, pp. 9929–9939. PMLR, 2020.

Yang, Z., Su, B., Cao, C., and Wen, J.-R. Regulatory DNA sequence design with reinforcement learning. In *The Thirteenth International Conference on Learning Representations*, 2025.

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W., et al. Model-based analysis of chip-seq (macs). *Genome biology*, 9:1–9, 2008.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2921–2929, 2016.

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, J., Theesfeld, C. L., Yao, K., Chen, K. M., Wong, A. K., and Troyanskaya, O. G. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nature genetics*, 50(8):1171–1179, 2018.

Zhou, Z., Ji, Y., Li, W., Dutta, P., Davuluri, R. V., and Liu, H. Dnabert-2: Efficient foundation model and benchmark for multi-species genomes. In *The Twelfth International Conference on Learning Representations*, 2024.

A. Related Works

Sequence-to-function models are designed to predict functional genomic signals directly from DNA sequences. DeepSEA (Zhou & Troyanskaya, 2015) established this approach by utilizing convolutional neural networks (CNNs) to extract sequence features for multi-task prediction. The field has evolved through architectural innovations and expanded training datasets (Kelley et al., 2018; Zhou et al., 2018; Chen et al., 2022). Currently, Enformer (Avsec et al., 2021) represents the leading methodology, achieving exceptional performance through its hybrid Transformer-CNN architecture. While these models simultaneously predict various outputs including epigenomic signals and gene expression levels, they typically lack specialized mechanisms for leveraging epigenomic data to enhance expression prediction specifically, treating all prediction targets as parallel outputs rather than considering their biological interdependencies.

Unsupervised DNA foundation models leverage the successful paradigm of unsupervised pre-training established in natural language processing. DNABERT (Ji et al., 2021) was the first to adapt this approach to genomics, applying BERT-like (Devlin et al., 2019) techniques to learn transferable DNA representations. Subsequent models have expanded upon this foundation (Zhou et al., 2024; Dalla-Torre et al., 2024; Li et al., 2024; Sanabria et al., 2024). In parallel, generative frameworks like Evo (Nguyen et al., 2024a) have emerged (Nguyen et al., 2024b; Brixi et al., 2025), enabling functional element design applications (Linder et al., 2025; Yang et al., 2025). Despite these advances, such models' effectiveness for gene expression prediction remains limited due to their exclusive reliance on DNA sequence information, without incorporating the critical epigenomic context that modulates gene activity.

Gene expression prediction represents a fundamental challenge in bioinformatics (Segal et al., 2002). Early approaches like Enformer (Avsec et al., 2021) attempted to predict gene expression directly from DNA sequences, facing inherent limitations, while GraphReg (Karbalayghareh et al., 2022) enhanced performance by incorporating epigenomic information through graph attention networks to model physical interactions between genomic regions. More recent methods have progressed toward integrating both sequence and epigenomic information, with Creator (Li et al., 2023) and EPInformer (Lin et al., 2024) demonstrating improved performance through this combined approach. However, these models typically rely on pre-identified regulatory elements, overlooking potential contributions from unannotated regions. Seq2Exp (Su et al., 2025) addressed this limitation through an end-to-end, data-driven methodology that simultaneously learns to identify relevant regulatory elements and predict expression with epigenomic guidance. Despite these advances, current research tends to focus predominantly on modeling distal regulatory elements through long sequence architectures, rather than optimizing the utilization of biologically interrelated epigenomic signals that directly influence gene regulation.

B. Shortening Input Sequence Length at Test Time

In Figure 1 (a) of Section 1, we have confirmed that training with longer sequences from scratch does not provide additional benefits. Further, we aim to investigate whether shortening the input length at test time would decrease the performance of a model pre-trained on longer sequences. Specifically, we tested the Seq2Exp model (Su et al., 2025) pre-trained on 200k sequences to evaluate if reducing context during inference affects performance. As shown in Table 2, we found that Seq2Exp, despite being trained on 200k inputs, shows minimal performance degradation when extra context is removed during testing. The performance difference between using 2100 tokens and 200k tokens is negligible. Interestingly, however, there is a significant performance drop when inputs are shortened to 2000 tokens.

Table 2. Performance of Seq2Exp (Su et al., 2025) when testing with shortened input sequences on the K562 cell line.

Input Length	MAE ↓	MSE ↓	Pearson ↑
2000	0.6485	0.6183	0.8084
2100	0.3471	0.2301	0.8603
2500	0.3174	0.1996	0.8674
3000	0.3134	0.1943	0.8698
8000	0.3082	0.1864	0.8747
10000	0.3074	0.1855	0.8751
200000	0.3054	0.1856	0.8723

Based on these observations and comparing with Figure 1 (b), we can conclude that input context length has a much smaller impact on model performance than epigenomic signals. Removing epigenomic signals during testing substantially hurts performance, while shortening sequence length has minimal effect. This finding motivates our focus on modeling

epigenomic information effectively.

C. Experimental Data of Table 1

We provide comprehensive numerical results corresponding to Figure 1 in the main text, including complete performance metrics and ablation studies.

C.1. Sequence Length Sensitivity

Table 3 compares the performance stability of Seq2Exp and Caduceus across different input lengths.

Table 3. Performance comparison with varying input lengths (left: Seq2Exp (Su et al., 2025), right: Caduceus (Schiff et al., 2024))

Length	MAE	MSE	Pearson r
100	0.3394	0.2233	0.8441
500	0.3096	0.1879	0.8744
2000	0.3150	0.1971	0.8678
5000	0.3098	0.1949	0.8703
10000	0.3088	0.1897	0.8719

Length	MAE	MSE	Pearson r
100	0.3385	0.2200	0.8449
500	0.3096	0.1889	0.8716
2000	0.3036	0.1831	0.8747
5000	0.3170	0.1941	0.8692
10000	0.3235	0.2029	0.8550

C.2. Epigenomic Signal Contributions

Table 4 demonstrates that combining all epigenomic signals yields optimal performance, with H3K27ac showing the strongest individual impact.

Table 4. Caduceus performance with different epigenomic signal configurations

Configuration	MSE	MAE	Pearson r
No signals	0.2163	0.3325	0.8485
+H3K27ac	0.1873	0.3080	0.8628
+DNase	0.2089	0.3227	0.8497
+Hi-C	0.2135	0.3264	0.8530
All signals	0.1886	0.3079	0.8652

C.3. Ablation Study

Table 5 reveals critical signal dependencies. Removing H3K27ac during testing from a model trained on all signals degrades performance most severely (22.3% MAE increase), while Hi-C removal has minimal effect (4.7% MAE increase).

Table 5. Performance degradation from signal removal (trained with all signals)

Condition	MAE	MSE	Pearson r
Drop H3K27ac	0.5653	0.6115	0.6964
Drop DNase	0.3890	0.2962	0.8189
Drop Hi-C	0.3548	0.2280	0.8467
Baseline (all signals)	0.3078	0.1886	0.8652

D. Case Study Revealing the Existence of Background Confounding Factors

To further support our hypothesis that background epigenomic signals (e.g. DNase and Hi-C) act as confounders rather than direct regulators of gene expression, we present a representative case (Figure 4). In this region, both DNase and Hi-C signals exhibit broad and high activation suggests strong chromatin accessibility and extensive three dimensional genome

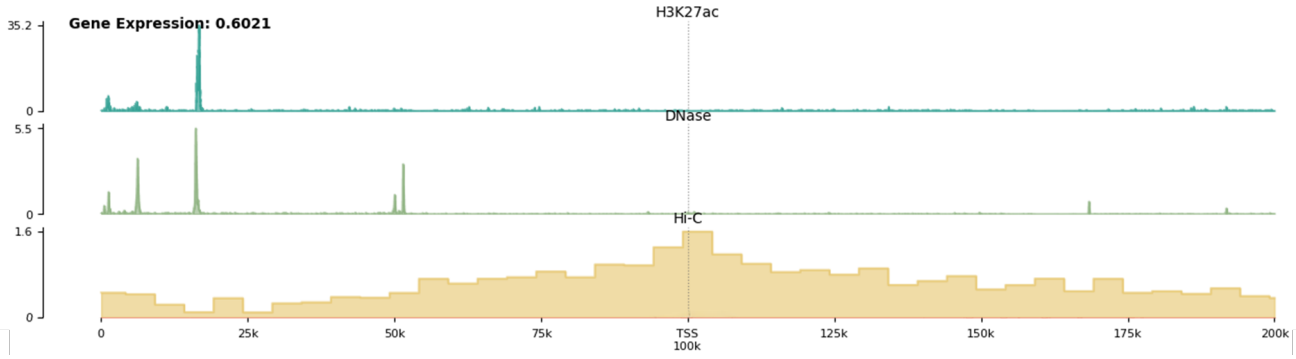

Figure 4. A representative genomic locus(Entrez ID: ENSG00000080561) where DNase and Hi-C signals are broadly active, but H3K27ac shows no enrichment. Despite strong chromatin accessibility and spatial contacts, gene expression remains low (0.6021), suggesting that background epigenomic signals alone are insufficient to drive transcription. This supports the hypothesis that such signals act as confounders rather than causal regulators.

interactions. However, H3K27ac, a known marker of active enhancers and promoters, shows little to no enrichment at the same locus. Notably, the gene expression level is very low (0.6021)

This case demonstrates that high background signal activity alone is insufficient to drive gene expression. Despite an accessible chromatin state (DNase and spatial proximity enabled by chromatin looping Hi-C), the absence of H3K27ac likely indicates that key regulatory elements are inactive, resulting in minimal transcriptional output.

Such case may reflect a permissive but not necessarily active regulatory environment. They also reinforce the necessity of discausal foreground signals like H3K27ac from background confounders when designing gene expression prediction models. This motivates our approach of modeling background signals through a Structural Causal Model and applying backdoor adjustment to correct for their confounding effects, thereby improving interpret ability and prediction accuracy.

E. Training Objective

To ensure that our model captures diverse and meaningful background patterns, we encourage the weight vectors to be distinct from each other through a uniform loss function (Wang & Isola, 2020). This loss penalizes similarity between background representations, promoting diversity in the learned weights:

$$\mathcal{L}_3 = \log \left(\sum_{i \neq j} \exp(2t \cdot a_i^T a_j - 2t) \right), \tag{5}$$

where t is a temperature parameter that controls the sharpness of the penalty.

Our final training objective combines the standard prediction loss, intervention-based regularization, and uniform diversity loss:

$$\mathcal{L} = \mathcal{L}_1 + \alpha \mathcal{L}_2 + \beta \mathcal{L}_3, \tag{6}$$

where α and β are hyperparameters controlling the relative importance of the intervention regularization and uniform diversity constraint, respectively. By jointly optimizing these three objectives, our model learns to make accurate predictions while effectively disentangling gene-specific regulatory signals from background confounding effects. The complete algorithm workflow for our InFER framework is provided in Appendix F.

F. Algorithm Workflow

Here we provide the complete algorithm workflow for our InFER framework in Algorithm 1. The algorithm initializes three neural networks: the signal encoder g_θ, the predictor network h_ψ, and the confounder encoder g_ω. During training, we

compute both standard and interventional predictions, then optimize the model using three objectives: prediction loss $\mathcal{L}_1$, intervention loss $\mathcal{L}_2$, and uniform diversity loss $\mathcal{L}_3$.

Algorithm 1 Interventional Framework for Gene Expression Prediction (InFER)

Require: Gene sequence X, epigenomic signals S, gene expression Y, hyperparameters α, β, t, n
Ensure: Trained model parameters θ, ψ, ω
  Initialize parameters θ, ψ, ω randomly
  while not converged **do**
    Forward Pass:
    $H = g_\theta(S)$ {Encode epigenomic signals}
    $\hat{Y} = h_\psi(X, H)$ {Standard prediction}
    $A = g_\omega(S)$ {Extract background patterns}
    Interventional Prediction:
    $\hat{Y}_{\text{int}} = \frac{1}{n}\sum_{i=1}^{n} h_\psi(X, H \odot a_i)$ {Apply backdoor adjustment}
    Loss Computation:
    $\mathcal{L}_1 = \text{dist}(\hat{Y}, Y)$ {Standard prediction loss}
    $\mathcal{L}_2 = \text{dist}(\hat{Y}_{\text{int}}, Y)$ {Intervention loss}
    $\mathcal{L}_3 = \log\left(\sum_{i,j} \exp(2t \cdot A^T A - 2t)\right)$ {Uniform diversity loss}
    $\mathcal{L} = \mathcal{L}_1 + \alpha\mathcal{L}_2 + \beta\mathcal{L}_3$ {Total loss}
    Backward Pass:
    Update parameters θ, ψ, ω using gradient descent on $\mathcal{L}$
  end while

G. Experiments Setup and More Results

G.1. Experimental Setup

Datasets. To evaluate gene expression prediction, we adopt Cap Analysis of Gene Expression (CAGE) values as our prediction proxy, in line with established approaches (Avsec et al., 2021; Lin et al., 2024; Su et al., 2025). Our study focuses on two well-characterized human cell lines that represent distinct cellular lineages: K562 and GM12878, both of which are extensively characterized in genomic research. We use CAGE measurements obtained from the ENCODE (Consortium et al., 2012). Following the experimental framework established in previous studies (Lin et al., 2024; Su et al., 2025), we evaluate our model across 18,377 protein-coding genes.

For input data, we utilize both DNA sequences and epigenomic signals. The DNA sequences are derived from the human genome HG38 project, while the epigenomic signals were carefully selected to capture different aspects of gene regulation: **H3K27ac** marks histone acetylation at active enhancers and promoters. Our experiments in Section 1 identify this as the most influential signal that directly pinpoints regulatory elements, thus we classify it as a foreground signal. **DNase** measures chromatin accessibility in genomic regions, often coinciding with but not causally determining regulatory elements. Our experiments confirm this, showing smaller performance improvements from DNase compared to H3K27ac. **Hi-C** quantifies contact frequencies between genomic positions and the target TSS, processed using the ABC pipeline (Fulco et al., 2019). Like DNase, we categorize Hi-C as a background signal representing the broader chromatin environment rather than specific regulatory elements.

Furthermore, we incorporate additional features such as mRNA half-life and promoter activity from previous studies (Lin et al., 2024) following (Lin et al., 2024; Su et al., 2025). These features are simply concatenated to the final linear predictor and are not part of our core modeling approach for epigenomic signals. Detailed data processing procedures can be found in (Lin et al., 2024; Su et al., 2025).

Baselines. We benchmark our InFER against the following baselines: Enformer (Avsec et al., 2021), a CNN-Transformer hybrid architecture designed to predict epigenomic signals and gene expression from sequences, here used solely for CAGE prediction; HyenaDNA (Nguyen et al., 2024b), Mamba (Gu & Dao, 2023), and Caduceus (Schiff et al., 2024), three recently developed DNA foundation models leveraging efficient long-sequence modeling capabilities through SSMs as prediction backbones; EPInformer (Lin et al., 2024), which extends the Activity-By-Contact (ABC) model (Fulco et al., 2019) by utilizing DNase-seq peaks to define potential regulatory regions and applying attention mechanisms to aggregate enhancer

signals; and Seq2Exp (Su et al., 2025), a recent SOTA method that applies information bottleneck principles to learn regulatory element masks, available in hard (binary) and soft (continuous) encoding variants.

We also include Caduceus w/signal, which incorporates epigenomic signals directly into Caduceus's encoder, and MACS3 (Zhang et al., 2008), which differs from Seq2Exp by using MACS3-identified regulatory elements instead of learned masks. Most baseline models process raw DNA sequences from the input region, while EPInformer operates on potential enhancer candidates extracted based on DNase-seq measurements following the ABC model (Fulco et al., 2019).

Evaluation Metrics. We assess model performance using three complementary metrics following (Su et al., 2025): Mean Squared Error (MSE) for measuring prediction variance with emphasis on larger errors; Mean Absolute Error (MAE) for quantifying average prediction deviation in expression units; and Pearson Correlation for evaluating how well models capture expression patterns and gene rankings regardless of absolute scale. These metrics together provide a balanced assessment of both prediction accuracy and pattern preservation capabilities.

Implementation Details. We partition datasets by chromosome for training, validation, and testing, following (Su et al., 2025). Specifically, chromosomes 3 and 21 serve as the validation set, while chromosomes 22 and X are reserved for testing. The inclusion of chromosome X provides a more stringent evaluation of model robustness due to its distinct biological characteristics compared to autosomes.

Our signal encoder g_θ is implemented as a simple linear layer, while the confounder encoder g_ω utilizes a lightweight 1D-CNN, with detailed configurations in Appendix H. For the predictor h_ϕ, we adopt Caduceus (Schiff et al., 2024) as our backbone model, following Seq2Exp (Su et al., 2025). Notably, we maintain the same training hyperparameters (learning rate, batch size, and other configurations) established in Seq2Exp (Su et al., 2025). Further performance gains could likely be achieved through hyperparameter fine-tuning specific to our approach. We use the L1 function as our prediction loss function, while the best model is selected based on the MSE metric on the validation set following (Su et al., 2025). All experiments were conducted on NVIDIA A40 and A100 GPUs. While most baseline models process inputs of length 200k, our InFER implementation operates on sequences of just 2k base pairs. We also present results for Caduceus w/signal at 2k length in Section G.2. Additional experimental details can be found in Appendix H.

G.2. Ablation Study

Our method introduces several hyperparameters: the number of confounder patterns n, and coefficients α and β in the training objective (Equation 6) that balance different loss components. We conducted ablation studies on the K562 cell type to evaluate our model's sensitivity to these hyperparameters. First, we fixed $\alpha = \beta = 1.0$ and varied the number of confounder patterns with $n \in \{1, 2, 4\}$. As shown in Table 6 (a), $n = 2$ yields the best performance. Given that computational overhead increases with larger values of n, we selected $n = 2$ for our final model. Next, we examined the impact of $\alpha \in \{0.0, 0.01, 0.1, 1.0, 10.0\}$, which controls the weight of the intervention loss. Note that $\alpha = 0.0$ disables our proposed InFER framework, and its mediocre performance demonstrates the effectiveness of our approach. Table 6 (b) shows that performance deteriorates significantly when $\alpha = 10.0$ (excessive intervention weight) and $\alpha = 0.0$ (no intervention). The model performs comparably for intermediate values, leading us to select $\alpha = 1.0$ for our final configuration. Finally, we investigated the influence of β, which encourages diversity among learned confounder weights. Table 6 (c) reveals that $\beta = 0.0$ (no diversity constraint) produces the worst results, likely due to weight collapse where all patterns converge to similar values. Once the diversity constraint is applied, the model demonstrates robust performance across different β values. We therefore selected $\beta = 1.0$ for our final implementation.

Table 6. Ablation studies on hyperparameters of InFER: (a) number of confounder patterns n, (b) intervention loss weight α, and (c) diversity constraint weight β.

n	MSE ↓	MAE ↓	Pearson ↑
1	0.1847	0.3047	0.8740
2	**0.1789**	**0.3000**	**0.8751**
4	0.1884	0.3079	0.8710

(a) Patterns n (α=β=1.0)

α	MSE ↓	MAE ↓	Pearson ↑
0.0	0.1910	0.3077	0.8691
0.01	0.1829	0.3046	0.8738
0.1	0.1843	0.3055	0.8750
1.0	**0.1789**	**0.3000**	**0.8751**
10.0	0.1974	0.3215	0.8695

(b) Weight α (n=2, β=1.0)

β	MSE ↓	MAE ↓	Pearson ↑
0.0	0.1893	0.3107	0.8714
0.01	0.1854	0.3039	0.8741
0.1	0.1807	0.3074	0.8723
1.0	**0.1789**	**0.3000**	**0.8751**
10.0	0.1860	0.3019	0.8725

(c) Weight β (n=2, α=1.0)

Table 7. Exploring the performance of Caduceus w/signal on short inputs for Cell Type K562

	MSE ↓	MAE ↓	Pearson ↑
Caduceus w/ signals (200k input)	0.1959 ± 0.0036	0.3187 ± 0.0036	0.8630 ± 0.0008
Caduceus w/ signals (2k input)	0.1863 ± 0.0035	0.3092 ± 0.0050	0.8713 ± 0.0023
Seq2Exp-soft	0.1856 ± 0.0032	0.3054 ± 0.0024	0.8723 ± 0.0012
InFER	**0.1789 ± 0.0041**	**0.3000 ± 0.0058**	**0.8751 ± 0.0036**

G.3. Do we really need more input length?

In Figure 1 (a), our preliminary experiments demonstrate that for Caduceus, performance plateaus at 2k sequence length, with further extensions leading to performance degradation. This limitation likely stems from the inherent inability of linear models like SSMs to effectively model long sequences - as input length increases, the model progressively loses focus on critical proximal regions. Even Seq2Exp (Su et al., 2025), which learns additional masks to mitigate performance degradation on long sequences, fails to derive benefits from extended inputs. To further investigate this phenomenon, we conducted a fair comparison between Caduceus w/signals (2k), Caduceus w/signals (200k), Seq2Exp-soft, and our InFER model. Shown in Table 7, the results reveal that Seq2Exp-soft shows negligible improvement over Caduceus w/signals (2k), while our InFER demonstrates significant performance gains. This finding reinforces our hypothesis that blindly extending sequence length yields diminishing returns with current SSM-based architectures. Instead, the critical factor for improved performance lies in effectively leveraging epigenomic signals rather than processing longer genomic contexts.

G.4. Parameter Overhead

Our confounder encoder is designed to be lightweight while delivering substantial performance improvements. We compare the additional parameters introduced by InFER and Seq2Exp (Su et al., 2025) relative to the base Caduceus (Schiff et al., 2024). As shown in Table 8, InFER adds only 11K trainable parameters to the base model. Our lightweight confounder encoder g_ω introduces minimal parameter overhead, whereas Seq2Exp's mask generator causes its parameter count to double compared to Caduceus. Notably, our approach outperforms Seq2Exp across all metrics while maintaining an almost unchanged parameter count compared to Caduceus.

Table 8. Parameter comparison between models.

Model	**Trainable Parameters**
Caduceus w/signals	574K
Seq2Exp-soft	1.1M
InFER	585K

H. More Implementation Details

H.1. Training Settings

Our training framework is implemented using PyTorch Lightning. All training-related hyperparameters were adopted directly from Seq2Exp (Su et al., 2025), which means we did not perform extensive parameter tuning for our specific approach. Consequently, there is potential for further performance improvements through careful hyperparameter optimization. The complete set of hyperparameters used in our experiments is presented in Table 9.

H.2. Implementation Details of Confounder Encoder

Our confounder encoder g_ω is implemented as a lightweight 1D-CNN that maps raw epigenomic signals $S \in \mathbb{R}^{L\times d}$ to weight vectors $A \in \mathbb{R}^{n\times d'}$. The architecture consists of a three-layer CNN followed by a projection layer:

- **Layer 1:** Conv1D (in_channels=d, out_channels=8, kernel_size=7) followed by BatchNorm, ReLU, and MaxPool (kernel_size=4)
- **Layer 2:** Conv1D (in_channels=8, out_channels=16, kernel_size=5) followed by BatchNorm, ReLU, and MaxPool (kernel_size=4)

Table 9. Hyperparameter values following Seq2Exp (Su et al., 2025).

Hyperparameters	Values
Layers of Generator	4
Hidden dimensions	128
Max training steps	50000
Batch size	8
Learning rate	5e-4
Scheduler strategy	CosineLR with Linear Warmup
Initial warmup learning rate	1e-5
Min learning rate	1e-4
Warmup steps	5,000
Validation model selection criterion	validation MSE

- **Layer 3:** Conv1D (in_channels=16, out_channels=32, kernel_size=3) followed by BatchNorm, ReLU, and MaxPool (kernel_size=4)
- **Global Pooling:** AdaptiveAvgPool1D(1) followed by Flatten
- **Projection:** Linear layer mapping the flattened features (32 dimensions) to $n \times d'$ dimensions

The progressive reduction in sequence length through max pooling operations (by a factor of 64 in total) efficiently captures patterns at different genomic scales while significantly reducing the computational overhead. After obtaining the raw weights, we apply a sigmoid activation function to constrain the values between 0 and 1, making them suitable for weighting the epigenomic signals via the Hadamard product operation. This lightweight design adds minimal parameters to the overall model while effectively modeling the background epigenomic regulatory patterns. The entire encoder requires only 11K parameters, which is negligible compared to the backbone model's parameter count.

I. Limitations

Our framework has two main limitations. First, we did not explore scaling up the model parameters (e.g., deeper layers or wider hidden dimensions) to assess if larger architectures could further boost performance. Second, the experiments were limited to two cell types (K562 and GM12878), and broader validation across diverse cell lines or tissues is needed to confirm generalizability. Future work should address these constraints to strengthen the method's applicability.