MoDNA: Motif-Oriented Pre-training For DNA Language Model[*]

Weizhi An
weizhi.an@mavs.uta.edu
University of Texas at Arlington
Arlington, Texas, USA

Yuzhi Guo
University of Texas at Arlington
Arlington, Texas, USA
yuzhi.guo@mavs.uta.edu

Yatao Bian
Tencent AI Lab
Shenzhen, Guangdong, China
yatao.bian@gmail.com

Hehuan Ma
University of Texas at Arlington
Arlington, Texas, USA
hehuan.ma@mavs.uta.edu

Jinyu Yang
University of Texas at Arlington
Arlington, Texas, USA
jinyu.yang@mavs.uta.edu

Chunyuan Li
University of Texas at Arlington
Arlington, Texas, USA
chunyuan.li@mavs.uta.edu

Junzhou Huang[†]
University of Texas at Arlington
Arlington, Texas, USA
jzhuang@uta.edu

ABSTRACT

Obtaining informative representations of gene expression is crucial in predicting various downstream regulatory-related tasks such as promoter prediction and transcription factor binding sites prediction. Nevertheless, current supervised learning with insufficient labeled genomes limits the generalization capability of training a robust predictive model. Recently researchers model DNA sequences by self-supervised training and transfer the pre-trained genome representations to various downstream tasks. Instead of directly shifting the mask language learning to DNA sequence learning, we incorporate prior knowledge into genome language modeling representations. We propose a novel Motif-oriented DNA (MoDNA) pre-training framework, which is designed self-supervised and can be fine-tuned for different downstream tasks MoDNA effectively learns the semantic level genome representations from enormous unlabelled genome data, and is more computationally efficient than previous methods. We pre-train MoDNA on human genome data and fine-tune it on downstream tasks. Extensive experimental results on promoter prediction and transcription factor binding sites prediction demonstrate the state-of-the-art performance of MoDNA.

CCS CONCEPTS

• **Computing methodologies** → *Neural networks*; • **Applied computing** → **Computational genomics**; *Bioinformatics.*

KEYWORDS

Transformer; Self-supervised Learning; Genome Analysis

ACM Reference Format:
Weizhi An, Yuzhi Guo, Yatao Bian, Hehuan Ma, Jinyu Yang, Chunyuan Li, and Junzhou Huang. 2022. MoDNA: Motif-Oriented Pre-training For DNA Language Model. In *13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (BCB '22), August 7–10, 2022, Northbrook, IL, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3535508.3545512

1 INTRODUCTION

Mining the non-coding regulatory genome to reveal the function of regulatory elements which coordinate with gene expression has recently attracted focus. Only a tiny fraction of DNA is utilized for coding proteins in DNA sequences, and the other 98% is non-coding DNA. The non-coding DNA contains sequences related to regulatory elements, determining when and where genes turn on and off. These elements provide sites for specialized proteins to bind, activate or repress gene expression. Recently, deep learning computational methods have been attractive for non-coding regulatory predictions. Furthermore, the researchers find that the non-coding DNA sequences have certain statistical features in common with natural languages [15]. The similarities between non-coding DNA and human language attract people to implement language models to decipher the non-coding DNA language.

Many convolutional neural network (CNN)-based methods ImaGene [19], DanQ [18], [20] have been applied to genomics sequences to study the gene regulatory components. Enformer [2] develops a new model based on Transformers [21] to make use of self-attention mechanisms that can integrate much longer DNA context, which takes advantage of the Transformers to read long-range dependencies. Despite the fruitful progress, two issues still limit the discovery of the DNA language model. (1) It is challenging for CNN to capture long-range dependencies within the genome context. (2) Supervised learning limits the generalization capability, and the models only adapt to a specific task with labeled data. Consequently, most previous sequence models are trained on limited downstream genome scenario data.

[*] This work was partially supported by the NSF CAREER grant IIS-1553687 and Cancer Prevention and Research Institute of Texas (CPRIT) award (RP190107).
[†] Corresponding author.

BCB '22, August 7–10, 2022, Northbrook, IL, USA
© 2022 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9386-7/22/08.
https://doi.org/10.1145/3535508.3545512

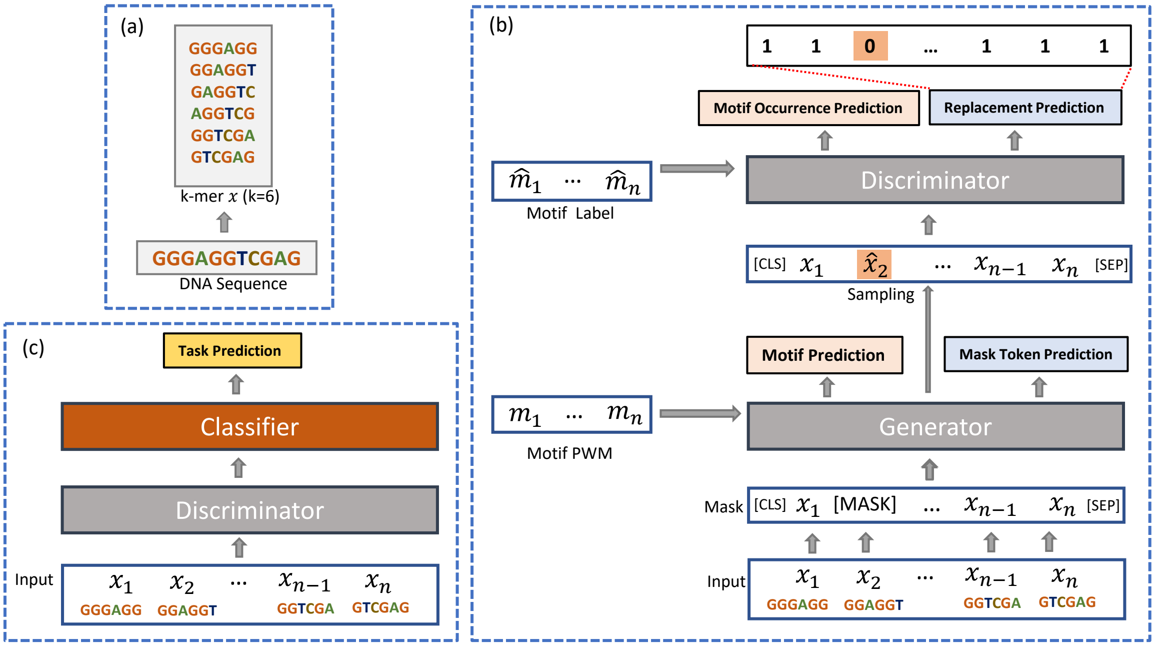
(a)
GGGAGG
GGAGGT
GAGGTC
AGGTCG
GGTCGA
GTCGAG
k-mer x (k=6)
GGGAGGTCGAG
DNA Sequence
(b)
1 1 0 … 1 1 1
Motif Occurrence Prediction
Replacement Prediction
$\hat{m}_1$ ⋯ $\hat{m}_n$
Motif Label
Discriminator
[CLS] x_1 $\hat{x}_2$ ⋯ x_{n-1} x_n [SEP]
Sampling
Motif Prediction
Mask Token Prediction
m_1 ⋯ m_n
Motif PWM
Generator
Mask [CLS] x_1 [MASK] ⋯ x_{n-1} x_n [SEP]
Input x_1 x_2 ⋯ x_{n-1} x_n
GGGAGG GGAGGT GGTCGA GTCGAG
(c)
Task Prediction
Classifier
Discriminator
Input x_1 x_2 ⋯ x_{n-1} x_n
GGGAGG GGAGGT GGTCGA GTCGAG

Figure 1: The Framework Overview of MoDNA. (a) The representation of DNA sequence k-mers x (k=6). (b) MoDNA Pre-training Pipelines. The input DNA sequence k-mers are randomly masked, and we denote x_2 as the mask token. We tokenize the input and add special tokens to construct our DNA tokens. The tokens are fed into a generator with two objectives: mask genome prediction and motif pattern prediction. Meanwhile, the generator produces sampling $\hat{x}_2$ to replace the mask-out token [MASK]. The replaced samplings with the other original tokens are fed into the discriminator. The discriminator is trained to distinguish whether the token has been replaced. By giving the motif occurrence label, the discriminator also learns to predict the motif occurrence. (c) MoDNA Fine-tuning Pipelines. We initialize the weights of the pre-trained discriminator. Then additional multiple layer perception is implemented for fine-tuning on different downstream tasks.

Recently, DNABERT [13] adapts Bidirectional Encoder Representations from Transformers (BERT) model [7] to genomics DNA settings and develops a general model which expands various sequence-related tasks. Although human language has similarities with gene language, directly considering DNA sequences as equivalent to human words is still misleading [11]. Unlike human language, which has well-defined explicit words, the DNA sequences exist in long sequences without spaces [12]. For example, if we directly apply BERT to DNA language, it is unclear whether a vocabulary has a biological function. GeneBERT [16] recently proposes an effective multi-modal pre-training model which is pre-trained from 17 million DNA sequences on the ATAC-seq dataset [9]. GeneBERT conducts sequences and regions in a multi-modal manner and further aligns them in the pre-training stage. However, sometimes the downstream tasks only contain sequence data, which limits the expressive power of GeneBERT when fine-tuning with sequence learning exclusively.

To address the above issues, we propose a novel framework called MoDNA, which introduces common DNA functional motifs as domain knowledge. Eukaryotic genomes contain a variety of structured patterns: repetitive elements, binding sites of DNA and RNA associated proteins, splice sites, and so on [4]. Such structured patterns can be formalized as motifs. In DNA, motif pattern functionally indicates sequences specific binding sites for proteins [8]. Instead of solely predicting the masks from the original k-mer DNA vocabulary, we jointly use a motif pattern task to learn the DNA representations. MoDNA constructs two types of prediction tasks in the self-supervised pre-train stage: k-mer tokens prediction and motif prediction. MoDNA can use such functional motifs to optimize the DNA representation embedding. To prove the effectiveness of our approach, we implement the MoDNA on the human genome data and conduct our MoDNA on two downstream tasks, which include promoter prediction and transcription factor binding sites prediction.

Our contributions can be summarized as follows: (1) by incorporating the domain knowledge, MoDNA extracts both semantic representations of DNA and predicts the occurrence of motifs which contributes to learning-rich semantic information from DNA sequences. (2) To the best of our knowledge, it is the first time to implement ELECTRA [5] on DNA sequences representation, which outperforms the BERT both in computing efficiency and evaluation performance. (3) We implement large-scare unlabeled genome data and leverage the pre-trained MoDNA on promoter and transcription factor binding site prediction tasks. On the downstream tasks, MoDNA achieves improvement compared with current state-of-the-art previous works.

2 METHODS

2.1 Pre-training

In the pre-training stage, MoDNA consists of two networks: a generator G and a discriminator D. First, the k-mers tokens from DNA sequences are randomly masked with [MASK] in a probability. Then, our generator learns from the masked tokens to perform the mask genome prediction. Meanwhile, G generates new samples at the position of the masked tokens, and then replaces the [MASK] tokens with the samplings. Correspondingly, by replacing the masked DNA k-mers with generated samples, the discriminator D is trained to predict whether each token is the same as the original one [22].

It is crucial to design self-supervision tasks for pre-training. In DNA, motifs are recurrent short sequences among the input genome data. We propose novel self-supervision tasks on the generator and discriminator networks, respectively: motif pattern PWM prediction and motif occurrence prediction. We consider motifs as labels and formulate the motif prediction tasks as a multi-label classification problem.

2.1.1 *Genome Token Prediction.*

In the pre-training stage, we take k-mer representation as our unlabeled pre-training data format as DNABERT. For example, six-mer of the DNA sequence *TTGGAAAT* includes: *TTGGAA*, *TGGAAA*, *GGAAAT*. The vocabulary of the k-mer representation contains all the permutations of the k-mer with additional five special tokens. We implement k-mer which is denoted as $x = [x_1, ..., x_n]$ into tokenization embedding E with correspondent $k^4 + 4$ token vocabulary.

Similar to ELECTRA, we jointly train a generator and a discriminator. Before feeding k-mer x into tokenization embedding, we randomly mask six consecutive positions of six-mer x as the generator's input. Formally, we denote the random mask positions as c where $c = [c_1, ..., c_t]$. The token x at the random mask position c is replaced by $r = replace(x, c, [MASK])$.

The generator encodes the masked input r into contextual embedding and performs mask genome modeling (MGM) to predict the original identities of the masked tokens. The MGM loss function is:

$$L_G = -\sum_{i \in c} log p_G(x_i | r) \tag{1}$$

We sample the replacement from MGM prediction output p and generate a sample distribution $\hat{x} \sim p_G(\hat{x}|r)$. Meanwhile, the inputs of the discriminator replace the masked x_t with generated samples $\hat{x}$ by $x^R = replace(x, c, \hat{x})$. The discriminator D is used to distinguish whether x_t^R comes from the original sequence or not. x_t^R is the input of the discriminator, which represents the generated samples for the masked tokens produced by the generator. The corresponding loss functions are:

$$L_D = \begin{cases} -\log D(x_t^R, x_R) & x_t^R = x^R \\ -\log(1 - D(x_t^R, x_R)) & x_t^R \neq x^R \end{cases} \tag{2}$$

2.1.2 *Motif Prediction.*

The motif in a nucleotide sequence pattern has biological significance. From the view of biological insight, associating motifs with gene regulation can interpret the biological role of the sequence's representation. We generate motifs pattern from the original DNA sequences by MEME [3] and denote such motifs PWM as $m = [m_1, ..., m_n]$ which is corresponded to the unmasked x. Each motif PWM can then be viewed as a weight vector. Since the PWM contains meaningful biological functions, we expect our generator G to learn the motif distribution. We introduce the PWMs m as our motif learning labels.

The last hidden layer representation $h_G(x)$ predicts the motif patterns by Kullback Leibler Divergence (KL) loss [14], which measures the matching of two different probability distributions. To perform motif distribution learning optimization, we train the generator using the last hidden layer $h_G(x)$ to learn the motif representation. The loss function of the motif learning is calculated by:

$$L_{G_{motif}} = \sum_{x \in c} m_i \log \frac{m_i}{W^T(h_G(x_i))} \tag{3}$$

Where c represents the masked k-mer genome tokens, and the generator learns the motif representation with the last hidden layer $h_G(x)$ to predict the motif pattern supervised by PWMs m.

Another component D_{motif} is used to predict the occurrence of the motif pattern from the x^R, which is replaced by the generated sample. We expect our discriminator to predict the locations of the motif occurrences in the given genome tokens. We make use of motif pattern occurrence labels $\hat{m} \in (0, 1)$ for the motif occurrence prediction task.

$$\begin{aligned} L_{D_{motif}} = &-\sum_{i \in n} \hat{m}_i \log sigmoid(h_D(x_i)) \\ &+ (1 - \hat{m}_i) \log(1 - sigmoid(h_D(x_i))) \end{aligned} \tag{4}$$

2.1.3 *Pre-training Objectives.*

We jointly train the generator and the discriminator by adopting the above two types of self-supervised tasks. We combine all the G and D losses and the final objective function for our MoDNA, which is formulated as:

$$L = L_G + \alpha L_D + \beta(L_{G_{motif}} + L_{D_{motif}}) \tag{5}$$

Where α and β are two hyperparameters to be adjusted.

It is worthy to note that the discriminator D is performed on all the sequences tokens, which is different from BERT. The masked token prediction in BERT only predicts the maskedx_t (15% of the tokens), while the discriminator of ELECTRA predicts all the input tokens x^R, which is partially sampled by a generator. Furthermore, our MoDNA combines the motif pattern both in generator and discriminator, which enriches the MoDNA to access semantic and functional information, as well as optimizes the sequence representation learning.

2.2 Fine-tuning

The supervised fine-tuning stage is based on our pre-trained discriminator network. We fine-tuned our pre-trained discriminator model on different downstream tasks. Specifically, the MoDNA is fine-tuned on two downstream tasks: promoter prediction and transcription factor binding site (TFBS) prediction. We add linear classifier for each downstream task on top of the pre-trained discriminator model. All the parameters from our pre-trained model are fine-tuned according to task-specific datasets and loss functions. After fine-tuning with several epochs, the downstream tasks consistently achieve promising performance on the corresponding task predictions.

Table 1: Comparison results on promoter prediction

Method	Accuracy	AUC	F1	MCC	Precision	Recall
GeneBERT [16]	-	0.894	-	-	0.805	0.803
DNABERT [13]	0.841	0.925	0.840	0.685	0.844	0.841
MoDNA (without motif prediction)	0.857	0.929	0.857	0.714	0.858	0.857
MoDNA (**ours**)	**0.862**	**0.935**	**0.862**	**0.725**	**0.863**	**0.862**

3 EXPERIMENTAL RESULTS

3.1 Pre-training and Fine-tuning Experimental Pipelines

We process human genome DNA sequences from GRCh38 and use them as our pre-training data. Also, we implement motif scanning on pre-training genome sequences to generate the corresponding motif pattern PWM matrix. We then pre-train our MoDNA based on the human genome dataset. Following the DNABERT, we pre-train our MoDNA with 15% six-mers mask-out. Specifically, we firstly process DNA sequences (maximum length is 512) into six-mers permutations. Then these k-mers tokens are randomly masked and sent to the generator. Meanwhile, the motif pattern PWM and the motif occurrence labels also join the generator and discriminator training correspondingly as mentioned before.

Table 3: Comparison between MoDNA with and without pre-training.

Method	Accuracy	AUC	F1	MCC	Precision	Recall
Pretrain	**0.862**	**0.935**	**0.862**	**0.725**	**0.863**	**0.862**
No Pretrain	0.808	0.889	0.808	0.618	0.809	0.809

3.2 Promoter Prediction

The promoter region is located near the transcription start sites (TSS) and regulates the initiation of the transcription of DNA [17]. We use the promoter core dataset, which is built from Eukaryotic Promoter Database (EPDnew) [10] and such core DNA promoter sequences are 70bp long and centered at the TSS. Each sample includes a genome sequence and a label. In our fine-tuning experiment, we select 10% for evaluation and the rest for training. For comparison, we download the official public DNABERT pre-trained model and follow the same fine-tune strategy in the paper. GeneBERT recently proposed an effective multi-modal pre-training model, which is pre-trained on 17 million genome sequences in the ATAC-seq dataset. We adopt the same experimental setting and promoter dataset, then fine-tune it with our MoDNA.

The performance of GeneBERT, DNABERT, MoDNA without motif, and MoDNA is shown in Table 1. Our MoDNA achieves the best performance in promoter prediction from the comparison results. Although GeneBERT is trained on large-scale genome data with cell-type motifs, our pre-trained MoDNA outperforms GeneBERT on all the evaluation metrics (Accuracy, AUC, F1, MCC, Precision, and Recall). Compared with DNABERT, our MoDNA is more efficient and has surpassed DNABERT with a relative 1.5% improvement on all the metrics. The remarkable improvements validate the effectiveness of our MoDNA. We can attribute this improvement to the efficient pre-training and the help of our self-defined motif prediction tasks. Compared with GeneBERT and DNABERT, the number of parameters of MoDNA is much smaller. Unlike BERT's pre-training strategy, our MoDNA makes predictions on the whole tokens, improving our training efficiency. What's more, MoDNA alleviates the data mismatch between the pre-training and the fine-tuning stages in the previous works. Although we implement mask genome prediction in the generator's pre-training, we generate the mask tokens with samplings and feed them to the discriminator. Thus, our input tokens keep consistent during both the pre-training and fine-tuning stages. Such implementation demonstrates the power of the ELECTRA. Also, it confirms our well-defined pre-training tasks can study the implicit domain knowledge and enhance the prediction performance of downstream tasks.

To investigate the contribution of self-supervised pre-training, we also compare the performances of the pre-trained MoDNA and MoDNA without pre-training on the promoter prediction task in Table 3. We keep the two models under the same experimental setting, and the performance of MoDNA without pre-training considerably drops on all the evaluation metrics. This confirms our MoDNA can learn the implicit semantic genome representation in the pre-training stage. By learning meaningful biological knowledge, our MoDNA boosts the performance on the downstream task.

3.3 Transcription Factor Binding Sites (TFBS) Prediction

Transcription is the process when a DNA segment copies into an RNA segment. Transcription factor proteins bind to specific regulatory DNA regions to regulate the expression of the gene. These regulator proteins can bind to gene promoters, enhancers, or control transcription. Therefore, predicting transcription factor binding sites is essential for evaluating the DNA sequence representation. The 690 CHIP-Seq datasets of uniform TFBS contains 161 TFs covering 91 human cell types [6].

We fine-tune MoDNA on 690 ENCODE CHIP-Seq datasets [6] and compare them with popular existing methods in this field. DeepBind [1] is a CNN-based model which automatically learns sequence motifs representation as kernels of a convolutional layer. DeepBind has been proved to be the state-of-the-art method in the TFBS prediction. Moreover, DeepSite [24] conducts the bidirectional long short-term memory and CNN to capture long-term dependencies between the sequence motifs in DNA. DESSO [23] impressively introduces motif shape into DNA sequence representation learning based on CNNs and performs comparably with DeepBind. In

Table 2: Comparison results on CHIP-Seq ENCODE 690 datasets with DeepSEA, DanQ, DeepSite, DESSO and MoDNA

Method	Accuracy	AUC	F1	MCC	Precision	Recall
DeepSEA [25]	0.853	0.919	0.836	0.717	0.840	0.858
DanQ [18]	0.840	0.910	0.823	0.694	0.848	0.823
DeepSite [24]	0.817	0.88	0.795	0.647	0.817	0.822
DESSO [23]	0.851	0.926	0.848	0.711	0.832	**0.884**
MoDNA	**0.856**	**0.935**	**0.851**	**0.727**	**0.859**	0.856

the TFBS prediction task, we implement fine-tuning on all the 690 datasets. The experimental average results are shown in Table 2. Our MoDNA also outperforms other baseline methods on the 690 datasets. Compared with DeepSEA, DanQ, DeepSite, and DESSO, the average AUC of MoDNA achieves the best performance.

GeneBERT fine-tunes their pre-trained model on 9 CHIP-Seq profile data of CTCF sites. DeepBind removes some biases of CHIP-Seq data and evaluates their performance using 506 in ENCODE ChIP-Seq datasets. We compare MoDNA with DeepBind and GeneBERT, and our pre-trained MoDNA achieves promising performance on the limited CHIP-Seq profiles. For 506 ENCODE CHIP-Seq datasets, our mean AUC achieves 0.94, significantly surpassing DeepBind's 0.914. For the CTCF sites experiment, our mean AUC is 0.996, exceeding the GeneBERT's 0.983.

4 CONCLUSION

In this paper, we present a novel self-supervised framework MoDNA. Instead of directly shifting the NLP paradigm to DNA language, we conduct motif patterns in the pre-training stage. It overcomes data mismatch limitation in BERT by learning from the replaced tokens from the generator. Meanwhile, the discriminator of MoDNA is utilized to learn and distinguish from all the input tokens, which also benefits the learning efficiency. We construct self-supervised motif prediction tasks by incorporating the domain knowledge motif pattern into the MoDNA pre-training. We conduct motif prediction in generator and motif occurrence in discriminator, respectively. With the help of prior domain knowledge, MoDNA can learn rich semantic DNA representation. By fine-tuning MoDNA, we achieve state-of-the-art performance on both promoter prediction and transcription factors binding sites prediction downstream tasks, which proves the power of MoDNA.

REFERENCES

[1] Babak Alipanahi, Andrew Delong, Matthew T Weirauch, and Brendan J Frey. 2015. Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. *Nature biotechnology* 33, 8 (2015), 831–838.

[2] Ziga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R Ledsam, Agnieszka Grabska-Barwinska, Kyle R Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R Kelley. 2021. Effective gene expression prediction from sequence by integrating long-range interactions. *bioRxiv* (2021).

[3] Timothy L Bailey, Mikael Boden, Fabian A Buske, Martin Frith, Charles E Grant, Luca Clementi, Jingyuan Ren, Wilfred W Li, and William S Noble. 2009. MEME SUITE: tools for motif discovery and searching. *Nucleic acids research* 37, suppl_2 (2009), W202–W208.

[4] Valentina Boeva. 2016. Analysis of genomic sequence motifs for deciphering transcription factor binding and transcriptional regulation in eukaryotic cells. *Frontiers in genetics* 7 (2016), 24.

[5] Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555* (2020).

[6] ENCODE Project Consortium et al. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 7414 (2012), 57.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[8] Patrik D'haeseleer. 2006. What are DNA sequence motifs? *Nature biotechnology* 24, 4 (2006), 423–425.

[9] Silvia Domcke, Andrew J Hill, Riza M Daza, Junyue Cao, Diana R O'Day, Hannah A Pliner, Kimberly A Aldinger, Dmitry Pokholok, Fan Zhang, Jennifer H Milbank, et al. 2020. A human cell atlas of fetal chromatin accessibility. *Science* 370, 6518 (2020).

[10] René Dreos, Giovanna Ambrosini, Rouayda Cavin Périer, and Philipp Bucher. 2013. EPD and EPDnew, high-quality promoter resources in the next-generation sequencing era. *Nucleic acids research* 41, D1 (2013), D157–D164.

[11] Yuzhi Guo, Jiaxiang Wu, Hehuan Ma, Sheng Wang, and Junzhou Huang. 2020. Bagging msa learning: Enhancing low-quality pssm with deep learning for accurate protein structure property prediction. In *International Conference on Research in Computational Molecular Biology*. Springer, 88–103.

[12] Yuzhi Guo, Jiaxiang Wu, Hehuan Ma, Sheng Wang, and Junzhou Huang. 2021. EPTool: a new enhancing PSSM tool for protein secondary structure prediction. *Journal of Computational Biology* 28, 4 (2021), 362–364.

[13] Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V Davuluri. 2021. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics* 37, 15 (2021), 2112–2120.

[14] Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics* 22, 1 (1951), 79–86.

[15] Rosario N Mantegna, Sergey V Buldyrev, Ary L Goldberger, Shlomo Havlin, Chung-Kang Peng, M Simons, and H Eugene Stanley. 1994. Linguistic features of noncoding DNA sequences. *Physical review letters* 73, 23 (1994), 3169.

[16] Shentong Mo, Xi Fu, Chenyang Hong, Yizhen Chen, Yuxuan Zheng, Xiangru Tang, Zhiqiang Shen, Eric P Xing, and Yanyan Lan. 2021. Multi-modal Self-supervised Pre-training for Regulatory Genome Across Cell Types. *arXiv preprint arXiv:2110.05231* (2021).

[17] Mhaned Oubounyt, Zakaria Louadi, Hilal Tayara, and Kil To Chong. 2019. DeePromoter: robust promoter predictor using deep learning. *Frontiers in genetics* 10 (2019), 286.

[18] Daniel Quang and Xiaohui Xie. 2016. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic acids research* 44, 11 (2016), e107–e107.

[19] Luis Torada, Lucrezia Lorenzon, Alice Beddis, Ulas Isildak, Linda Pattini, Sara Mathieson, and Matteo Fumagalli. 2019. ImaGene: a convolutional neural network to quantify natural selection from genomic data. *BMC bioinformatics* 20, 9 (2019), 1–12.

[20] Ramzan Umarov, Hiroyuki Kuwahara, Yu Li, Xin Gao, and Victor Solovyev. 2019. Promoter analysis and prediction in the human genome using sequence-based deep learning models. *Bioinformatics* 35, 16 (2019), 2730–2737.

[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[22] Sheng Wang, Yuzhi Guo, Yuhong Wang, Hongmao Sun, and Junzhou Huang. 2019. SMILES-BERT: large scale unsupervised pre-training for molecular property prediction. In *Proceedings of the 10th ACM international conference on bioinformatics, computational biology and health informatics*. 429–436.

[23] Jinyu Yang, Anjun Ma, Adam D Hoppe, Cankun Wang, Yang Li, Chi Zhang, Yan Wang, Bingqiang Liu, and Qin Ma. 2019. Prediction of regulatory motifs from human Chip-sequencing data using a deep learning framework. *Nucleic acids research* 47, 15 (2019), 7809–7824.

[24] Yongqing Zhang, Shaojie Qiao, Shengjie Ji, and Yizhou Li. 2020. DeepSite: bidirectional LSTM and CNN models for predicting DNA–protein binding. *International Journal of Machine Learning and Cybernetics* 11, 4 (2020), 841–851.

[25] Jian Zhou and Olga G Troyanskaya. 2015. Predicting effects of noncoding variants with deep learning–based sequence model. *Nature methods* 12, 10 (2015), 931–934.