Multi-modal Self-supervised Pre-training for Large-scale Genome Data

Anonymous Author(s)
Affiliation
Address
email

Abstract

Open genomic regions, being accessible to regulatory proteins, could act as the on/off switch or amplifier/attenuator of gene expression, and thus reflect the defining characteristics of cell types. Many previous models make predictions from the sequence to the regulatory region, but the interaction between regulatory regions and genes could be complex and differ between cell types. Moreover, current models usually only perform well on the cell types in the training set, which are not generalizable to data-scarce scenarios. In this work, we propose a simple yet effective approach for pre-training genome data in a multi-modal and self-supervised manner, which we call **GeneBERT**. Specifically, we simultaneously take the 1d sequence of genome data and a 2d matrix of (transcription factors × regions) as the input, where three pre-training tasks are proposed to improve the robustness and generalizability of our model. We pre-train our model on the ATAC-seq dataset with 17 million gene sequences. We evaluate our GeneBERT on various downstream tasks, including promoter prediction, transaction factor binding sites prediction, disease risks estimation, and RNA-Splicing. Extensive experiments demonstrate the effectiveness of multi-modal and self-supervised pre-training for large-scale genome data.

1 Introduction

In recent years, some works [1, 2] have been proposed to explore the genome data, which only perform well on the cell types in the training set. Typically, Enformer [1] combines dilated CNN and transformer architecture as well as multi-head output for gene-related tasks, such as expression, epigenomic marks, etc. However, there is no objective term for unsupervised pre-training and thus is less transferable to data-scarce scenarios. More recently, DNABERT [2] is introduced to formulate the whole DNA sequence as a sentence of nucleotide k-mers and utilize BERT to model the sequence generatively. However, DNABERT is only applied to downstream tasks such as core promoter prediction or TFBS-prediction in a single cell type, where no cell-type specificity was considered. Furthermore, no pre-trained models have been developed to model the regulation mechanism across various cell types in the human body. Interactions between regulatory regions and genes are not well captured, and thus cannot generalize well to different cell types.

Integration of genome data modalities across different cell types could help to build a more holistic model of gene expression regulation and benefit downstream applications such as mutation impact evaluation and disease risk prediction, as well as promoting our understanding of cell-type-specific regulatory programs and various development processes and disease etiology. Inspired by this fact, in this work, we present a simple yet effective method called GeneBERT, for pre-training large-scale genome data in a multi-modal and self-supervised manner. Specifically, we simultaneously take the 1d modality (*i.e.* sequence) and a 2d modality (*i.e.* regulatory region) of genome data as the input, where three pre-training tasks are proposed to improve the robustness and generalizability of our model. 1)

Submitted to NeurIPS 2021 AI for Science Workshop.

masked sequence modeling: we randomly mask some parts of the input k-mers with a special token (i.e., [MASK]), and the model is trained to predict the masked k-mer. 2) next sequence prediction: we train the model using the embedding [CLS] to classify whether a pair of given sequences are two consecutive sequences in a cell. 3) sequence-region matching: a sequence-region matching mechanism is proposed to capture the multi-modal alignment between sequence and regulatory region of genome data.

We pre-train our GeneBERT on the ATAC-seq dataset with 17 million gene sequences. Furthermore, we conduct extensive experiments to evaluate our GeneBERT on four downstream tasks, including promoter prediction, transaction factor binding sites prediction, disease risks estimation, and RNA-Splicing. Comprehensive ablation studies demonstrate the effectiveness of multi-modal and self-supervised pre-training for large-scale genome data.

The main contributions of this work are summarized as follows:

- We propose a simple yet effective method named GeneBERT, for large-scale genome data pre-training in a multi-modal and self-supervised manner.
- We are the first to incorporate different genome data modalities across various cell types into the pre-training for large-scale genome data.
- Extensive experiments demonstrate the effectiveness of our model on four downstream tasks.

2 Related Work

Language/Vision Pre-training. Self-supervised pre-training models such as GPT [3], BERT [4], RoBERTa [5], and ERNIE [6] have led to dramatic improvement on a variety of natural language processing tasks in the past few years, significantly surpassing the traditional context-independent language model such as Word2Vec. RoBERTa [5] uses dynamic MLM and discards NSP, spends a long time to train the model. ERNIE [6] masks entities and phrases, this method expects to learn more context relations. Multi-modal pre-training has recently addressed researchers' attention to learning meaningful representations. Typically, Previous methods [7, 8] learn visual representations from text paired with images in unsupervised, self-supervised, weakly supervised, and supervised ways. Since language and vision can share a similar semantic meaning, CLIP [7] is a commonly-used neural network trained on a variety of (image, text) pairs for learning transferable visual representations from natural language supervision. Huo *et al.* [8] apply a cross-modal contrastive learning framework called BriVL for image-text pre-training. However, in this work, we leverage the multi-modal self-supervised pre-training on the genome data to improve the robustness and generalizability of pre-trained models used for data-scarce scenarios.

Genome data pre-training. Transformer models have been recently established to better understand the genotype-phenotype relationships [1, 2]. DNABERT uses the human genome to pre-train a BERT-based model, trying to decipher the regulatory code related to gene expression [2]. In order to adapt the DNA scenario, sequences are split into 5 to 510 base-pair long and tokenized to 3- to 6-mers representations. After the pre-training, the model was fine-tuned on three downstream tasks related to gene regulation: prediction of promoters, transcription factor binding sites (TFBSs), and splice sites. Furthermore, by analyzing the attention maps, DNABERT could visualize the important regions contributing to the model decision, which improved the interpretability of the model. Different from DNABERT, we incorporate different genome data modalities across various cell types into the pre-training for large-scale genome data.

3 Method

3.1 Preliminary: BERT

Masked language model (MLM) and next sentence prediction (NSP) are two core self-supervised tasks of BERT, and BERT relies on them for pre-training. MLM is called a cloze task in the literature, where we select some percentage of random tokens in the sequence and replace them with masked tokens to predict the masked tokens. BERT randomly selects 15% of the input tokens as possible objects. Among the selected tokens, 80% are replaced by mask, 10% with randomly selected tokens, and 10% left unchanged. NSP is used for binary classification of context relationship between sequences, which predicts whether two fragments in the original sequence are related to each other.

3.2 GeneBERT

In this section, we propose a simple yet effective approach for pre-training genome data in a multi-modal and self-supervised manner, as shown in Figure 1.

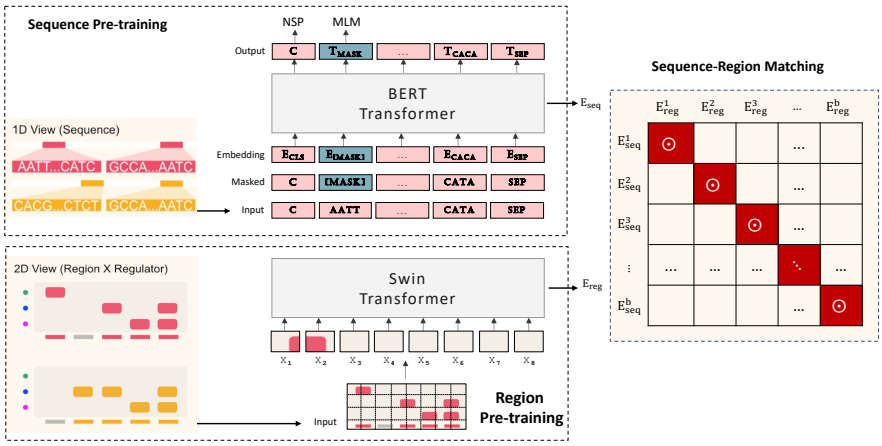

Figure 1: The overall framework of our proposed GeneBERT model.

Sequence Pre-training. For sequence embeddings in the pre-training, we input three types of embeddings: 1) a k-mer embedding $\mathbf{e}^t$ for each k-mer in a sequence; 2) a segment embedding $\mathbf{e}^s$ indicating which part of the sequence the k-mer is from; 3) a position embedding $\mathbf{e}^p$ for the position of the k-mer in the sequence. The k-mer refers to a sequence with length k, *i.e.*, for a sequence AGTCAG, the 3-mers are {AGT, GTC, TCA, CAG}, and the 4-mers are {AGTC, GTCA, TCAG}. Then we sum up all three embeddings in a contextual representation $\mathbf{e}^n, n \in \{1, 2, ..., N\}$, where N denotes the number of k-mers in the sequence. After being fed into a BERT-based transformer, those contextual embeddings become $\mathbf{E}_{\text{seq}}$. We adopt two similar objectives as BERT, including masked language modeling (MLM) $\mathcal{L}_{mlm}$ and next sentence prediction (NSP) $\mathcal{L}_{nsp}$. For the former objective, we randomly mask some parts of the input k-mers with a special token (i.e., [MASK]), and the model is trained to predict the masked kmer. As for NSP, we train the model using the embedding [CLS] to classify whether a pair of given sequences are consecutive in a cell.

Region Pre-training. For the region features in the pre-training, we consider a strong backbone (*i.e.* Swin [9]) transformer as the encoder to extract representations $\mathbf{E}_{\text{reg}}$. Specifically, we apply the Swin transformer pre-trained on ImageNet to the region input directly to generate $\mathbf{E}_{\text{reg}}$. During the pre-training, we do not fix the parameters of Swin transformer and update them for learning better regional representations. In the pre-training setting, each region input corresponds to each sequence such that we can capture the multi-modal alignment between sequence and region of genome data.

Sequence-Region Matching. In order to learn the alignments between sequence and region of genome data, we propose a sequence-region matching mechanism to sequence embeddings $\mathbf{E}_{\text{seq}}$ and region embeddings $\mathbf{E}_{\text{reg}}$. Specifically, we calculate the cosine similarity between each pair of linguistic embeddings $\mathbf{E}^i_{\text{seq}}$ and visual embeddings $\mathbf{E}^i_{\text{reg}}$ in a batch of size b, where $i \in 1, 2, ..., b$. Then, those similarities are jointly learned for alignments between the whole sequence and each region in the same batch, where we maximize the cosine similarity of the sequential and regional embeddings of the b correct pairs in the batch while minimizing the cosine similarity of the embeddings of the b^2b false pairings. We apply a sequence-region matching loss over these similarities scores for optimization, and the loss is defined as:

$$\mathcal{L}_{srm} = -\log \frac{\sum_{i=1}^{b} \mathbf{E}^i_{\text{seq}} \cdot \mathbf{E}^i_{\text{reg}}}{\sum_{i=1}^{b} \sum_{j=1}^{b} \mathbf{E}^i_{\text{seq}} \cdot \mathbf{E}^j_{\text{reg}}} \tag{1}$$

where b is the batch size. In this way, we maximize the cosine similarity of sequential and regional embeddings from correct pairs while minimizing the cosine similarity of embeddings of false pairs. Intuitively, alignments between the whole sequence and each region are learned via our Gene-BERT in the pre-training process. Thus, the overall objective is defined as $\mathcal{L} = \mathcal{L}_{mlm} + \mathcal{L}_{nsp} + \lambda \cdot \mathcal{L}_{srm}$. We set λ = [0.01, 1] to perform the parameter study for λ, and observe that the performance of our model is stable when λ = [0.5, 1]. In our experiments, we set λ = 0.5.

4 Experiments

4.1 Pre-training Data & Settings

For pre-training data, we process public human fetal cerebrum single-cell chromatin accessibility data in the Descartes database [10] to generate pseudo-bulk accessibility tracks for each cell type (Seurat cell clustering provided by the original paper). Specifically, we take the provided 'Peak Count Sparse Matrices' and summed up columns (cells) according to cell type definition, producing a regions $\times$ cell-types matrix. Then we binarize the matrix and use only non-zero entries (accessible regions) for each cell type. The corresponding sequence for each region is then retrieved from hg19 human reference genome. While the motif scanning for each region is either retrieved from the Descartes database or scanned following the same approach using JASPAR 2018 [11] vertebrate transcription factor binding site motifs. In total, we use 17 cell types and the union of all accessibility track includes 1,000,029 accessible regions across the genome, covering 504,657,456 base pairs. For the 1D modality, we group 10 consecutive accessible regions into one sample, which corresponds to a (10 $\times$ number of TFs) matrix for the 2D modality. Following previous works [2], we pre-train the model for 120k steps with a warm-up learning rate of 4e-4 and batch size of 2000. 15% of k-mers in each sequence are masked in the first 100k steps, and 20% for the last 20k steps.

4.2 Downstream Tasks

We evaluate our GeneBERT on four downstream tasks: promoter classification, Transcription Factor Binding Sites (TFBS) classification, splicing, and disease-related regions identification. See more experimental results in the Appendix.

Promoter Classification. Promoters are the elements responsible for regulating the initial transcription of the gene, which is located near the transcription start site (TSS). As the promoters play an important role in gene regulation, using machine learning methods to predict promoter sites accurately is one of the most popular problems in bioinformatics. Here we first used the promoter core dataset from [2], which are the 70bp sequences centered around TSS. Promoter core is the key part of the promoter flanking region which is sufficient to direct accurate initiation of transcription [12]. Here we fine-tune our GeneBERT model to predict the promoter core sequences. We report the experimental results in Table 1. From the results, we can see that our model can predict promoter core accurately.

Table 1: Comparison results on promoter and TFBS classification.

Task	Method	Precision	Recall	AUC
Promoter	DNABERT	0.675	0.637	0.693
	GeneBERT (ours)	**0.805**	**0.803**	**0.894**
CTCF_A549_CTCF_UW	DNABERT	0.250	0.500	0.542
	GeneBERT (ours)	**0.925**	**0.921**	**0.983**
CTCF_AG04450_CTCF_UW	DNABERT	0.250	0.500	0.501
	GeneBERT (ours)	**0.929**	**0.925**	**0.987**

TFBS Classification. Predicting TFBS is an important step in studying gene regulation. Sequencing technologies like ChIP-seq can provide information on the in vivo binding sequences, which improve the identification of gene regulatory regions. There are several previous studies that tried to predict TFBSs using traditional machine learning [13] and deep learning methods [14]. By incorporating the multi-modal pre-training, the prediction of TFBSs can be further improved. Although we utilize the motif information during the region pre-training, we do not provide any matching information to the model, which avoids leaking information about the actual motif of a specific TF. Here we fine-tune our model for predicting TFBSs from the ChIP-seq data, using 497 TF ChIP-seq uniform peak profiles from ENCODE Consortium [15]. We take the peak sequences of each TF as the positive set and generated a corresponding negative set by randomly shuffling the nucleotides in each positive sequence while preserving dinucleotide frequencies. Table 1 reports the comparison results and those results demonstrate the advantage of our GeneBERT over DNABERT.

5 Conclusion

In this work, we present the GeneBERT, a multi-modal self-supervised framework for large-scale genome data pre-training. Specifically, we leverage sequence pre-training, region pre-training and sequence-region matching together to improve the robustness and generalizability of our model. Extensive experiments on four main downstream tasks demonstrate the effectiveness of our GeneBERT via multi-modal and self-supervised pre-training for large-scale genome data.

References

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

[2] Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V Davuluri. Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. *Bioinformatics*, 37(15):2112–2120, 2021.

[3] Tim Salimans Alec Radford, Karthik Narasimhan and Ilya Sutskever. Improving language under-standing by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/research- covers/languageunsupervised/language understanding paper.pdf*, 2018.

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

[5] Ott M. Goyal N. Du J. Joshi M. Chen D. Levy O. Lewis M. Zettlemoyer L. Liu, Y. and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[6] Yukun Li Shikun Feng Xuyi Chen Han Zhang Xin Tian Danxiang Zhu Hao Tian Yu Sun, Shuohuan Wang and Hua Wu. Ernie: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223*, 2019.

[7] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

[8] Yuqi Huo, Manli Zhang, Guangzhen Liu, Haoyu Lu, Yizhao Gao, Guoxing Yang, Jingyuan Wen, Heng Zhang, Baogui Xu, Weihao Zheng, Zongzheng Xi, Yueqian Yang, Anwen Hu, Jinming Zhao, Ruichen Li, Yida Zhao, Liang Zhang, Yuqing Song, Xin Hong, Wanqing Cui, Danyang Hou, Yingyan Li, Junyi Li, Peiyu Liu, Zheng Gong, Chuhao Jin, Yuchong Sun, Shizhe Chen, Zhiwu Lu, Zhicheng Dou, Qin Jin, Yanyan Lan, Wayne Xin Zhao, Ruihua Song, and Ji-Rong Wen. WenLan: Bridging vision and language by large-scale multi-modal pre-training. *arXiv preprint arXiv:2103.06561*, 2021.

[9] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021.

[10] 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, Michael A. Zager, Ian A. Glass, Frank J. Steemers, Dan Doherty, Cole Trapnell, Darren A. Cusanovich, and Jay Shendure. A human cell atlas of fetal chromatin accessibility. *Science*, 370(6518):eaba7612, November 2020. Publisher: American Association for the Advancement of Science.

[11] Aziz Khan, Oriol Fornes, Arnaud Stigliani, Marius Gheorghe, Jaime A Castro-Mondragon, Robin van der Lee, Adrien Bessy, Jeanne Chèneby, Shubhada R Kulkarni, Ge Tan, Damir Baranasic, David J Arenillas, Albin Sandelin, Klaas Vandepoele, Boris Lenhard, Benoît Ballester, Wyeth W Wasserman, François Parcy, and Anthony Mathelier. JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Research*, 46(D1):D1284, January 2018.

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

[13] Chenyang Hong and Kevin Y Yip. Flexible k-mers with variable-length indels for identifying binding sequences of protein dimers. *Briefings in Bioinformatics*, 21(5):1787–1797, 2020.

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

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

[16] Kishore Jaganathan, Sofia Kyriazopoulou Panagiotopoulou, Jeremy F. McRae, Siavash Fazel Darbandi, David Knowles, Yang I. Li, Jack A. Kosmicki, Juan Arbelaez, Wenwu Cui, Grace B. Schwartz, Eric D. Chow, Efstathios Kanterakis, Hong Gao, Amirali Kia, Serafim Batzoglou, Stephan J. Sanders, and Kyle Kai-How Farh. Predicting Splicing from Primary Sequence with Deep Learning. *Cell*, 176(3):535–548.e24, January 2019.

A Appendix

In this appendix, we provide more experimental results on TFBS Classification, disease risks estimation, and RNA-Splicing.

TFBS Classification. We report the comparison results in Table 2. We can observe that our GeneBERT outperforms DNABERT by a large margin in terms of all protein types TFBSs classification. This further demonstrates the effectiveness of our GeneBERT in boosting the performance via incorporating the multi-modal pre-training.

Table 2: Comparison results on Transcription Factor Binding Sites classification.

Protein	Method	Precision	Recall	AUC
CTCF_A549_CTCF_UT-A	DNABERT	0.250	0.500	0.501
	GeneBERT (ours)	**0.908**	**0.899**	**0.983**
CTCF_A549_CTCF_UW	DNABERT	0.250	0.500	0.542
	GeneBERT (ours)	**0.925**	**0.921**	**0.983**
CTCF_AG04449_CTCF_UW	DNABERT	0.250	0.500	0.523
	GeneBERT (ours)	**0.907**	**0.894**	**0.983**
CTCF_AG04450_CTCF_UW	DNABERT	0.250	0.500	0.501
	GeneBERT (ours)	**0.929**	**0.925**	**0.987**
CTCF_AG09309_CTCF_UW	DNABERT	0.250	0.500	0.545
	GeneBERT (ours)	**0.931**	**0.927**	**0.987**
CTCF_AG09319_CTCF_UW	DNABERT	0.250	0.500	0.529
	GeneBERT (ours)	**0.924**	**0.919**	**0.983**
CTCF_AG10803_CTCF_UW	DNABERT	0.250	0.500	0.535
	GeneBERT (ours)	0.944	0.942	0.991
CTCF_AoAF_CTCF_UW	DNABERT	0.250	0.500	0.531
	GeneBERT (ours)	**0.917**	**0.913**	**0.982**
CTCF_BE(2)-C_CTCF_UW	DNABERT	0.250	0.500	0.540
	GeneBERT (ours)	**0.937**	**0.935**	**0.989**

Disease Risks Estimation. GeneBERT could provide more interpretations of complex genetic diseases. On the one hand, while the disease status and genomic mutations were available, by integrating the 2D-data, the relationships among regulatory regions of genes could be captured, which allowed us to estimate the disease risk more accurately. As shown in Table 3, GeneBERT can precisely predict Hirschsprung Disease (HSCR), which is known as a genetic disorder with complex patterns of inheritance. On the other hand, similar to DNABERT, by comparing the attention maps of mutant and wild-type, disease-related regions could be identified and ranked based on the attention scores, which could be seen as the candidates of treatment target sites and proceeded to the medical experimental validation.

Table 3: Comparison results on disease risks estimation.

Data	Method	Precision	Recall	AUC
HSCR-RET	DNABERT	0.265	0.500	0.500
	GeneBERT (ours)	**0.770**	**0.519**	**0.562**
HSCR-RET-Long	DNABERT	0.252	0.500	0.462
	GeneBERT (ours)	**0.768**	**0.513**	**0.541**

Table 4: Comparison results on Splicing datasets.

Data	Method	Top-k Accuracy	PR-AUC
SpliceAI-80nt	dilated CNN	0.57	0.60
	GeneBERT (ours)	**0.83**	**0.89**
SpliceAI-256nt	dilated CNN	-	-
	GeneBERT (ours)	**0.93**	**0.95**
SpliceAI-400nt	dilated CNN	0.90	0.95
	GeneBERT (ours)	**0.95**	**0.98**
SpliceAI-2k	dilated CNN	0.93	0.97
	GeneBERT (ours)	**0.97**	**0.99**

RNA-Splicing Sites Prediction. RNA Splicing is an important post-transcription processing to remove introns from pre-mRNA sequences and generate mature mRNA for protein translation. Previously, dilated CNN models have been used to predict splice junction across the genome and evaluate the impact of genomics variants on splicing sites [16]. In particular, for each nucleotide in a given sequence for splicing site prediction, we follow the previous approach and include a context sequence around the nucleotide, which could potentially capture the sequence specificity features

of RNA-binding proteins and splicing machinery. Since open chromatin regions and splicing sites does not always overlap with each other, among all 548,000 splicing sites in the GTEx pre-mRNA transcripts data, our pre-training sequence only fully covers the entire (in the 256nt context setting) sequence of 72,500 sites. In total, 26.7% of nucleotides in context and splicing site sequence where included in the open chromatin region we used for pre-training. Following the same training/testing split scheme and classification metric as in the SpliceAI study [16], we are able to achieve similar or better results in different context settings without including a extremely long context sequence. This task clearly demonstrated the capacity and generalizablilty of our pre-training model. By integrating sequence binding features of RNA binding proteins, we might be able to further extended our model to enable cell-type specific splicing junction prediction in the future.