Learning Cell-Type-Specific Gene Regulation Mechanisms by Multi-Attention Based Deep Learning with Regulatory Latent Space

Minji Kang *[1], Sangseon Lee *[2], Dohoon Lee[3], and Sun Kim †[1,2,3]

[1] *Bioinformatics Institute, Seoul National University*
[2] *Department of Computer Science and Engineering, Seoul National University*
[3] *Interdisciplinary Program in Bioinformatics, Seoul National University*

November 30, 2019

Introduction

Epigenetic gene regulation is a major control mechanism of gene expression. Most of the existing methods for modeling the control mechanism of gene expression used only a single epigenetic marker and few methods were successful in modeling the complex mechanisms of gene regulations using epigenetic markers on transcriptional regulation. Modeling gene regulations needs complex interactions among multiple epigenetic and transcriptional markers such as histone marks, DNA methylation, and transcription factors, thus single machine learning methods can hardly model such complex transcriptional control mechanisms. To address these challenges, we introduce a multi-attention based deep learning model for gene regulation mechanisms.

Materials and Methods

We propose a two-step ensemble machine learning model and the architecture is illustrated in Figure 1. At the first layer of the model, separate models vectorize each epigenetic and transcriptional marker with a different deep learning strategy, and then, at the second layer, output vectors from the models at the first layers are integrated by a multi-attention network. The epigenetic and transcriptional markers near the transcription start site (TSS) mainly involve in the gene expression, hence we focus on histone marks and DNA methylation in the gene region of 4000 base-pair (bp) around the TSS. Similarly, we utilize transcription factors in the region of 200 bp around the TSS. To begin with, separate models embed histone marks, DNA methylation, and transcription factors into a regulatory latent space. First, histone marks are embedded into the latent space by a Convolutional Neural Network (CNN) followed by Bi-directional Long Short-Term Memory (LSTM) network with attention. The CNN and LSTM capture local and sequential patterns of seven histone marks, while the attention mechanism reveals important gene loci. Second, DNA methylation is vectorized by

*co-first author

†sunkim.bioinfo@snu.ac.kr

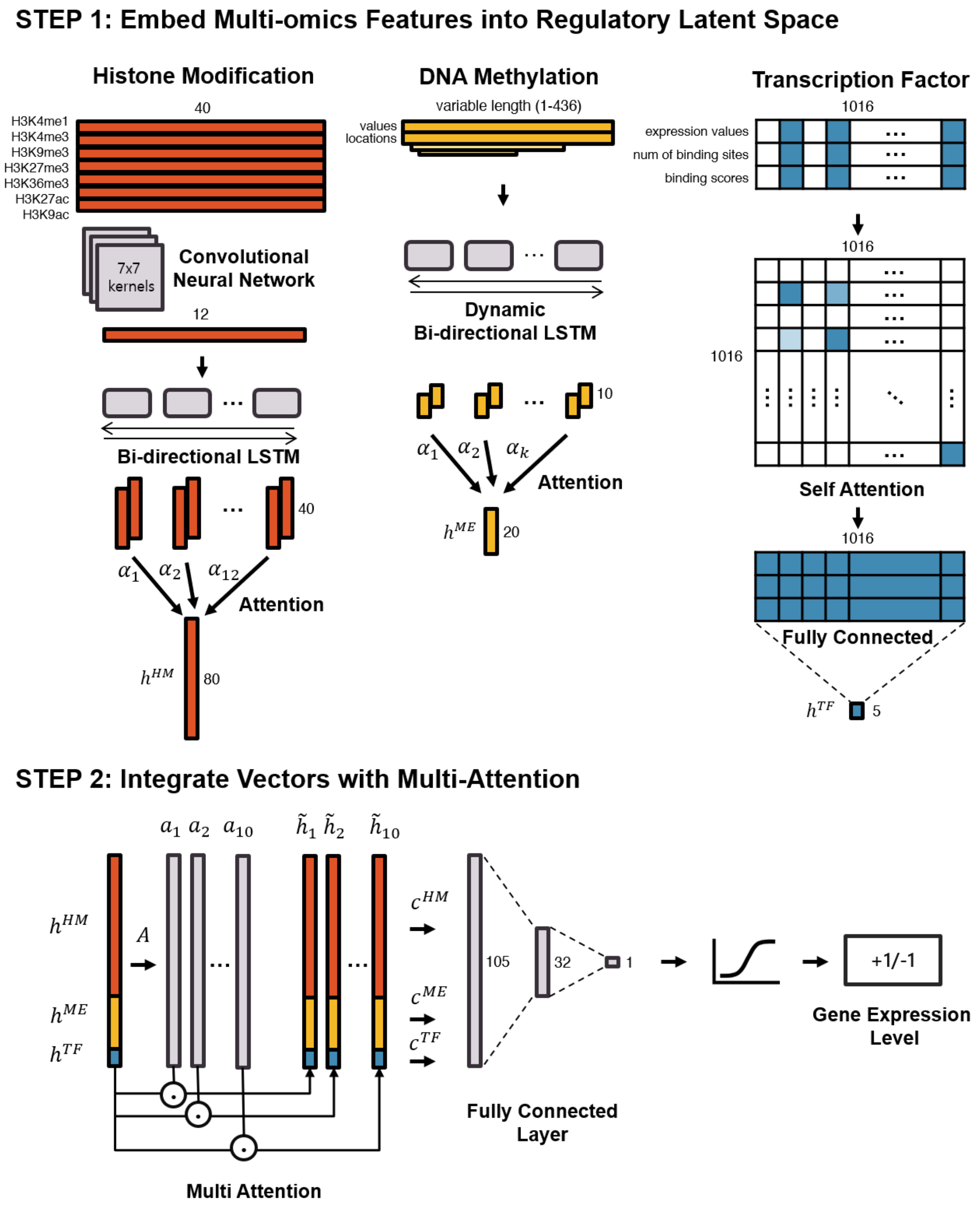


Figure 1: A multi-attention based deep learning model with regulatory latent space.

a Dynamic Bi-directional LSTM with attention. Because we use the methylation values and locations of all CpG sites near TSS as input, the input size varies for different genes and cell-types. Therefore, the ‘Dynamic’ Bi-directional LSTM is employed to capture the pattern of DNA methylation. Last, to embed transcription factors, we first select candidate binding transcription factors for each gene, based on the prior knowledge, HumanTFDB, and motif detection tool, HOMER. Then, Self-Attention Network (SAN) vectorizes the expression value of candidate transcription factors. As a result of SAN, the attention weight matrix provides vital information about relationships and interactions between transcription factors.

After embedding features in three vectors, the model combines these vectors by a multi-attention network to predict whether the gene would be expressed. While the end-to-end model as a whole is to predict the gene expression level, the multi-attention network determines which types of epigenetic markers are most influential for controlling gene expression and how epigenetic features interact with each other in each cell type.

Results

To evaluate our proposed model, we utilized three different multi-omics markers from the ENCODE project to predict the gene expression level of 18 cell lines, for which histone marks, DNA methylation, and transcription factors are available (Table 1). We split 18,070 genes into four-folds. The first and second folds were used as a test and validation set, and the rest two folds were used as a training set. Every result is averaged from 4-fold cross-validation.

Table 1: 18 cell lines, for which histone marks, DNA methylation, and transcription factors are available.

Cell Line	Group	Standardized Epigenome Name
E003	ESC	H1 Cells
E004	ES-deriv	H1 BMP4 Derived Mesendoderm Cultured Cells
E005	ES-deriv	H1 BMP4 Derived Trophoblast Cultured Cells
E006	ES-deriv	H1 Derived Mesenchymal Stem Cells
E007	ES-deriv	H1 Derived Neuronal Progenitor Cultured Cells
E011	ES-deriv	hESC Derived CD184+ Endoderm Cultured Cells
E016	ESC	HUES64 Cells
E038	Blood & T-cell	Primary T helper naive cells from peripheral blood
E047	Blood & T-cell	Primary T CD8+ naive cells from peripheral blood
E066	Other	Liver
E087	Other	Pancreatic Islets
E114	ENCODE2012	A549 EtOH 0.02pct Lung Carcinoma Cell Line
E116	ENCODE2012	GM12878 Lymphoblastoid Cells
E117	ENCODE2012	HeLa-S3 Cervical Carcinoma Cell Line
E118	ENCODE2012	HepG2 Hepatocellular Carcinoma Cell Line
E119	ENCODE2012	HMEC Mammary Epithelial Primary Cells
E120	ENCODE2012	HSMM Skeletal Muscle Myoblasts Cells
E123	ENCODE2012	K562 Leukemia Cells

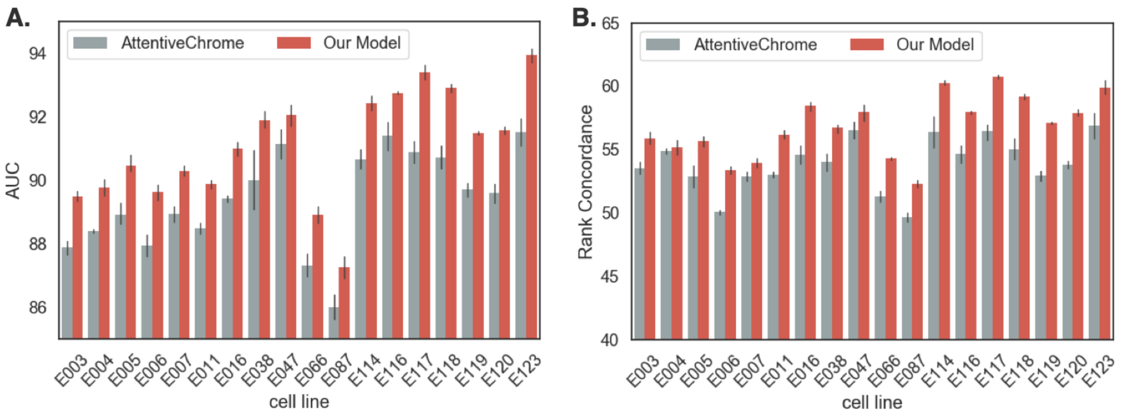


Figure 2: (A) AUC with the baseline for all cell lines. (B) Rank Concordance with the baseline for all cell lines.

Performance evaluation using histone signals only

We set a baseline with the state-of-the-art method for gene expression level prediction, AttentiveChrome [1]. As AttentiveChrome was designed for histone marks instead of multiple epigenetic features, we trained both our model and AttentiveChrome using only histone marks for a fair comparison. We evaluated them with two metrics. (1) First, we used classification for a binary value prediction of active or not. (2) Second, gene expression value prediction in terms of rank concordance between the gene expression values and the final output values of the model.

The AUC and rank concordance of our model exceeded that of AttentiveChrome for every cell line (Figure 2). On average, the proposed model achieved 91.05% of AUC, while AttentiveChrome got 89.72%. Moreover, the model demonstrated its robustness by showing higher rank concordances between the gene expression value and the final output of the model. Our model showed 56.83% of rank concordance on average, while AttentiveChrome showed only 53.85%. We guess that the performance difference is due to the difference in the model architectures. AttentiveChrome first uses an individual RNN structure for each histone mark, and then integrate histone marks with an additional RNN. On account of the individual RNN, the interactions of numerous kinds of histone marks are likely to be neglected. Consequently, AttentiveChrome was not successful to grasp the local characteristics of seven histone marks. On the other hand, our model employs both CNN and RNN, allowing to capture local and sequential features of every histone mark in a single model. Therefore, our model is suitable for understanding not only the roles of histone marks but also interactions among them.

Performance evaluation using multi-omics markers

Since our model is designed to utilize multi-omics biomarkers, we measured performance in terms of the average AUC of our models that were trained on all possible combinations of multi-omics features (Figure 3). The average AUC of the model improved with adding and integrating multi-omics features. Especially, the model with histone marks (HM, TF+HM, ME+HM, TF+ME+HM) showed remarkable levels of AUC, exceeding the AUC of AttentiveChrome. This result attributes to the fact that genes can be expressed only if histones are opened, and histone marks play a crucial

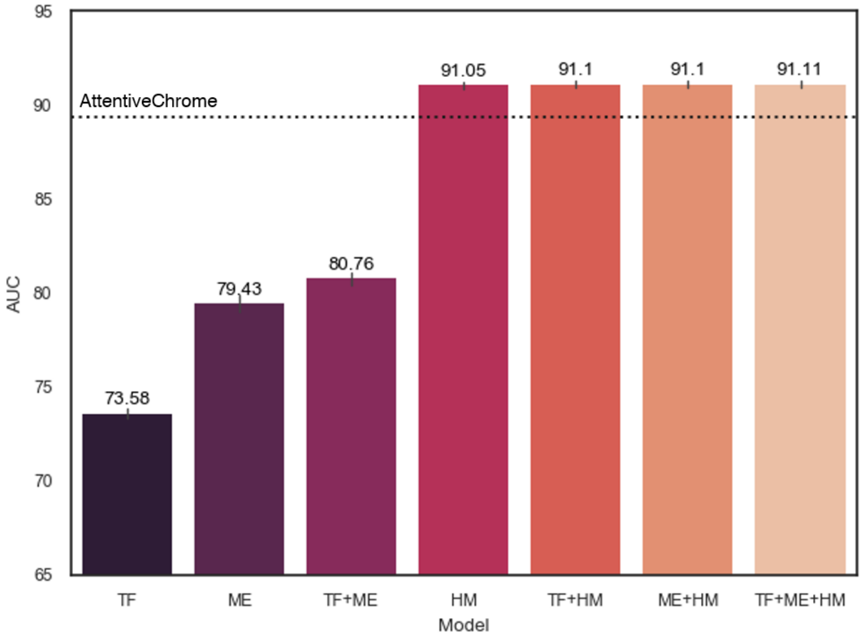


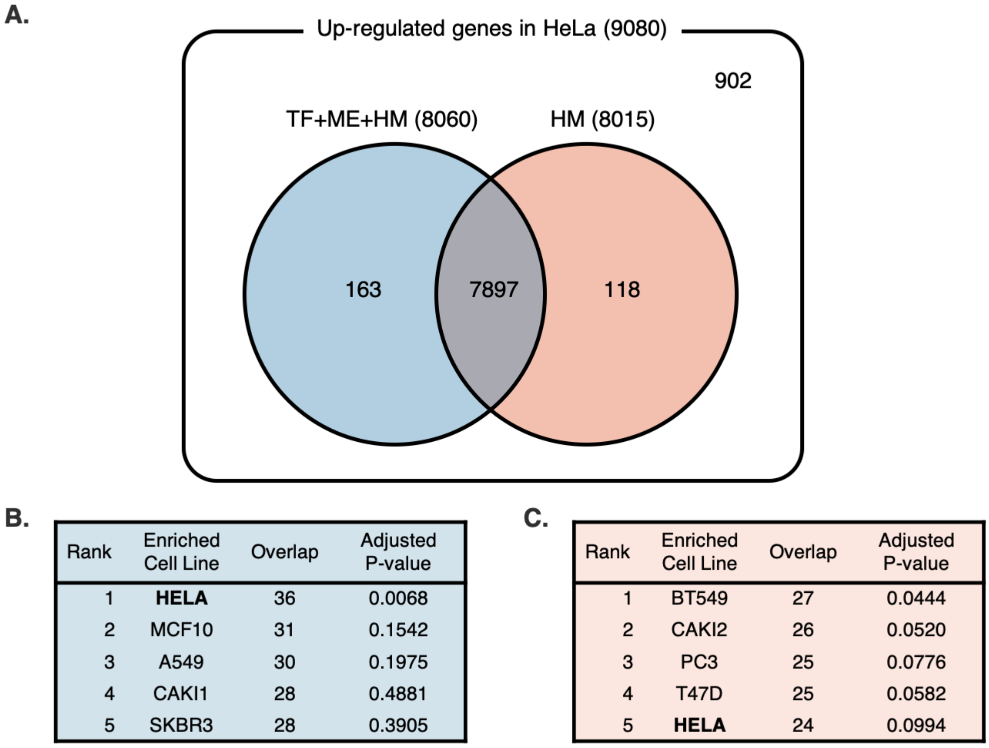
Figure 3: Average AUC of all cell lines for different combinations of multi-omics features. TF, ME, and HM stand for transcription factors, DNA methylation, and histone marks, respectively.

role in the gene regulations. This result is intuitive.

Finding cell-type-specific gene regulation mechanisms

Finally, the model was successful in discovering some cell-type-specific gene regulation mechanisms using multi-omics markers. To explore the cell-type-specificity of our model, we conducted a case study on E117, HeLa-S3 cervical carcinoma cell line.

We compared the cell-type-specificity between the model with transcription factors, DNA methylation, and histone marks (TF+ME+HM) and the model with histone marks only (HM). As up-regulated genes give each cell type its functionality, predicting them correctly is important in grasping cell-type-specific gene regulations. Thus, it is important to reduce a false negative rate. In this regard, the TF+ME+HM model had a lower false negative rate (0.1123) than that of HM model (0.1209) even though they had similar AUC (93.45 for TF+ME+HM model, and 93.39 for HM model). Therefore, the TF+ME+HM model is likely to be more meaningful in cell-type-specificity than HM model. To substantiate the cell-type-specificity of TF+ME+HM model, we investigated the set of up-regulated genes, which were predicted correctly by the model. Figure 4(A) shows a set diagram of up-regulated genes in the HeLa cell. One set contained the genes predicted correctly by TF+ME+HM model, and another set contained the genes predicted correctly by HM model. Among 9080 up-regulated genes in the HeLa cell line, 87% of genes (7,897) were predicted correctly by both the TF+ME+HM model and the HM model. However, **the focus of our analysis was on genes that two models predicted differently**, to show the roles of DNA methylation and transcription factors.



Rank	Enriched Cell Line	Overlap	Adjusted P-value
1	**HELA**	36	0.0068
2	MCF10	31	0.1542
3	A549	30	0.1975
4	CAKI1	28	0.4881
5	SKBR3	28	0.3905

Rank	Enriched Cell Line	Overlap	Adjusted P-value
1	BT549	27	0.0444
2	CAKI2	26	0.0520
3	PC3	25	0.0776
4	T47D	25	0.0582
5	**HELA**	24	0.0994

Figure 4: (A) Up-regulated genes in the HeLa-S3 cervical carcinoma cell line. Among 9,080 up-regulated genes, 163 genes were predicted as up-regulated by the TF+ME+HM model and they were predicted as down-regulated by the HM model. Meanwhile, 118 genes were predicted as up-regulated by the HM model and they were predicted as down-regulated by the TF+ME+HM model. (B) The cell-line enrichment results of 163 genes, which were predicted as up-regulated genes by the TF+ME+HM model and not by HM model. (C) The cell-line enrichment results on 118 genes, which were predicted as up-regulated genes by the model trained on HM.

We used Enrichr [2] to perform cell-line enrichment of the genes that were predicted differently by the two models. Figure 4(B) shows the top 5 enriched cell lines with 163 genes that were predicted correctly by the TF+ME+HM model and not by the HM model. Among the 163 genes, 36 genes were known highly expressed genes in the HeLa cell from the ARCHS4 Database [3]. The HeLa cell ranked first in enriched cell lines, with statistical significance (Adjusted p-value: 0.0068). On the other hand, with 118 genes, predicted correctly by the HM model, only 24 genes were overlapped with the known high expressed genes in the HeLa cell (Figure 4(C)). The HeLa cell ranked fifth, and the enrichment is not statistically significant (Adjusted p-value: 0.0994).

To sum up, the TF+ME+HM model predicted cell-type-specific genes more accurately than the HM model. This was, of course, because our model used transcription factors and DNA methylation that play important roles in the cell-type-specific gene regulations.

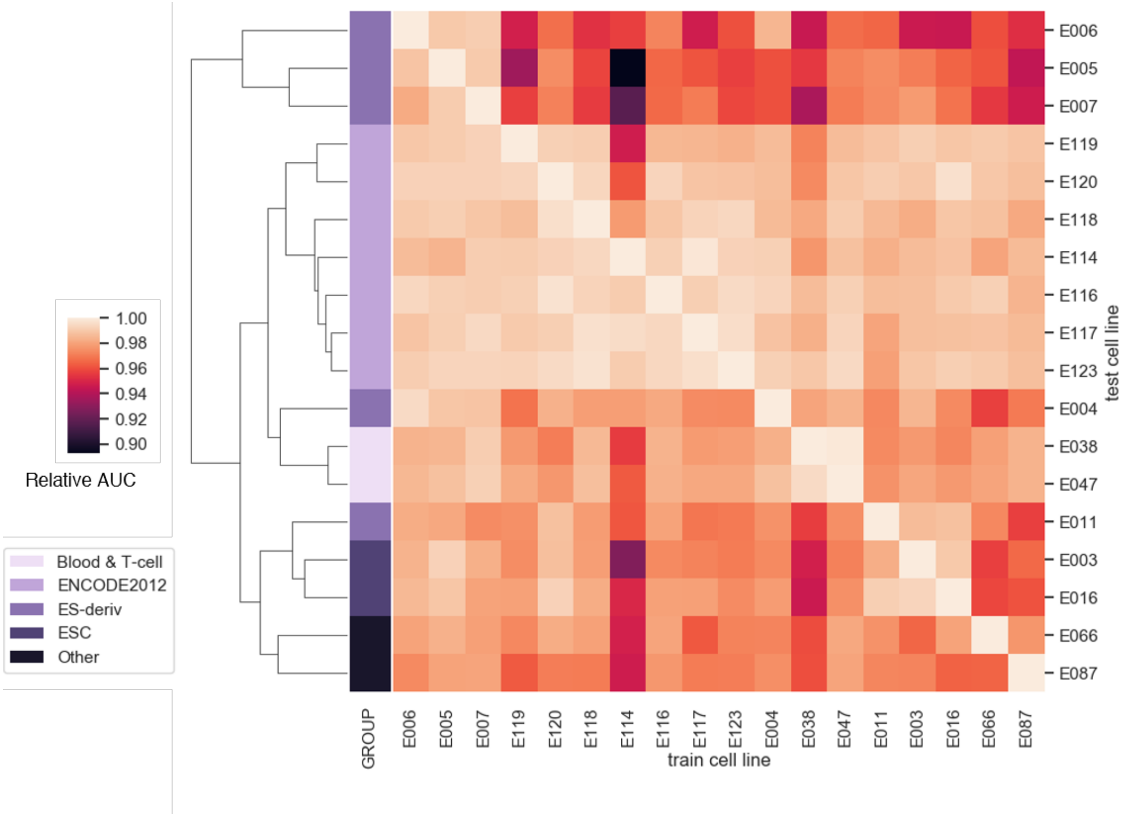


Figure 5: Compatibility test result between cell lines.

Compatibility test between 18 cell lines

Furthermore, we evaluated the cell-type-specificity and compatibility of our model, by training on one cell line and testing on other cell lines. Each cell line got the greatest AUC with the model trained on the cell line, demonstrating the cell-type-specificity. Besides, it is notable that the cell lines in the same group showed similar AUC patterns (Figure 5). By performing hierarchical clustering with a Euclidean distance, cell lines in the same group were clustered together. The result highlights the transferability between the models in the same group. In other words, each cell line can be explained well by the model of the other cell lines if they are in the same group. For instance, the blood and T-cell group, E038 and E047, showed the best AUC for each other's model. This is probably because the cell lines in the same group tend to have similar gene regulation mechanisms.

Conclusion

In summary, the proposed model is the first of its kind to use multiple epigenetic and transcriptional markers for predicting gene expressions. Constructing cell-type-specific models using the proposed methods will be very insightful in unveiling cell-type-specific gene regulation mechanisms using epigenetic and transcriptional markers.

References

[1] Singh, Ritambhara and Lanchantin, Jack and Sekhon, Arshdeep and Qi, Yanjun. Attend and predict: Understanding gene regulation by selective attention on chromatin. *Advances in neural information processing systems*, 6785–6795, 2017.

[2] Chen, Edward Y and Tan, Christopher M and Kou, Yan and Duan, Qiaonan and Wang, Zichen and Meirelles, Gabriela Vaz and Clark, Neil R and Ma'ayan, Avi. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC bioinformatics*, 14(1):128, 2013.

[3] Lachmann, Alexander and Torre, Denis and Keenan, Alexandra B and Jagodnik, Kathleen M and Lee, Hoyjin J and Wang, Lily and Silverstein, Moshe C and Ma'ayan, Avi. Massive mining of publicly available RNA-seq data from human and mouse. *Nature communications*, 9(1):1366, 2018.