Learning putatively causal gene regulatory programs using permutation-equivariant neural networks

Sanjit Singh Batra[1]
Computer Science Division
University of California, Berkeley
sanjitsbatra@berkeley.edu

Jeffrey Spence[1]
Department of Genetics
Stanford University School of Medicine
jspence@stanford.edu

Yun S. Song[2]
Computer Science Division and Department of Statistics
University of California, Berkeley
Chan Zuckerberg Biohub
yss@berkeley.edu

[1]**Equal contribution** [2]**Corresponding author**

Abstract

Epigenetic modifications regulate gene expression across different cell types and understanding the mechanisms by which such gene regulation happens is an important question in genomics. Epigenetic data across different cell types can shed light on some of these mechanisms. Further, by looking at mechanisms that are common across multiple cell types, it is possible to extract putatively causal mechanisms via which epigenetic features regulate gene expression. In this work, we design a permutation-equivariant convolutional neural network to learn a cell-type agnostic mapping between epigenetic markers and gene expression. We then perform *in-silico epigenesis* to interpret the regulatory mechanisms learnt by the neural network. Finally, we explore how our network can be used to optimize the epigenetic features to achieve a desired level of gene expression.

1 Introduction

All cells within a multicellular organism have the same genetic sequence up to a miniscule number of somatic mutations. Yet, a menagerie of cell types exist with diverse morphologies and functions. In order to understand how such differences arise and are maintained, a considerable number of experiments to assay aspects of the epigenome (such as transcription factor binding, histone modifications, DNA methylation, and chromosome conformation) have been developed. Two large consortia, ENCODE [1] and the NIH Roadmap Epigenomics Project [2], have either performed an extensive number of assays in a small number of cell types (ENCODE) or a small number of assays across many cell types (Roadmap). A key challenge is to understand how these epigenetic markers modulate gene expression using these datasets.

Recent studies have tried to predict gene expression using combinations of such epigenetic markers and DNA sequence [3, 4, 5, 6, 7]. DeepChrome [8] attempts to classify gene expression into two classes (corresponding to *high* or *low* gene expression) using a convolutional neural network with histone modification ChIP-seq data as the input. Since the epigenetic markers can be different across cell types [9], DeepChrome learns an independent model for each cell type. It has recently been shown that a relationship exists between invariant representations and causality [10]. We hypothesize

Preprint. Under review.

that the mechanisms via which gene expression is regulated by epigenetic markers that are common across different cell types, could be putatively causal. A naive way to find them would be to treat each gene in each cell type as an independent training data point while predicting gene expression from epigenetic markers. A neural network consisting of only permutation-equivariant layers (as described in the lower part of Figure 1) would be equivalent to this formulation. Equipping the neural network with additional permutation-invariant layers would permit the exploration of a more general class of functions [11]. Since, the resulting neural network, which consists of permutation-equivariant layers as well as permutation-invariant layers, is itself permutation-equivariant, we refer to it as a permutation-equivariant convolutional neural network and the resulting architecture is shown in Figure 1.

In this study, we leverage epigenetic data across different cell types generated by the ENCODE project, to unravel putative causal regulatory programs [12]. In order to do so, we predict gene expression values (as opposed to binary classification into *high* or *low* gene expression [8]) using six histone modification markers from twelve different cell types (Table 1).

2 Methods

2.1 Data preparation

We obtained $-\log_{10}(\text{p-value})$ ChIP-seq tracks created by running the MACS2 peak-caller [13] on read count data, from the ENCODE Imputation Challenge [1, 14, 15]. The histone modifications and cell types used in this study are outlined in Table 1. For three tracks where data were not available, we downloaded Avocado [16] imputations from the ENCODE data portal [1, 15]. We binned each epigenetic track at 100bp resolution and pre-processed them with an additional $\log$ operation (using the numpy `log1p` function [17]) before inputting them into the network.

We downloaded polyA-plus RNA-seq gene expression TPM values for each of the 12 cell types in Table 1, from the ENCODE data portal [1, 15] and preprocessed them with a $\log$ operation (using the numpy `log1p` function [17]).

Cell Type	H3K36me3	H3K27me3	H3K27ac	H3K4me1	H3K4me3	H3K9me3
IMR-90	T	T	T	T	T	T
H1-hESC	T	T	T	T	T	T
trophoblast cell	T	T	T	T	T	T
neural stem progenitor cell	T	T	T	T	T	T
K562	T	T	T	T	T	T
heart left ventricle	T	T	T	T	T	T
adrenal gland	T	A	T	T	T	T
endocrine pancreas	T	T	T	T	T	T
peripheral blood mononuclear cell	T	T	T	T	T	T
amnion	T	T	T	T	T	T
myoepithelial cell of mammary gland	T	T	A	T	T	T
chorion	T	T	T	T	T	A

Table 1: ChIP-seq $-\log_{10}(\text{p-values})$ were obtained from the ENCODE Imputation Challenge [1, 14, 15] where the ground truth data were available (corresponding to entries labeled **T** in the table). Avocado [16] imputations were downloaded from the ENCODE data portal [1, 15], where ground truth data were not available (corresponding to entries labeled **A** in the table).

2.2 Permutation-equivariant convolutional neural networks

The network architecture we used in this study is summarized in Figure 1. Since the order of the cell types in the input tensor is arbitrary, permuting the cell types should lead to the output gene expression being permuted in the same manner. Such permutation-equivariance can be encoded into the architecture via a permutation-equivariant layer (bottom of Figure 1, which performs 2D convolutions with a kernel of height = 1 (along the cell type dimension) and width = 11 (along the position dimension), with the different epigenetic markers being considered as input channels. The permutation-invariant layer (top of Figure 1) first performs a similar 2D convolution, followed by taking the mean across the cell type dimension to enforce invariance along this dimension. We then tile the resulting tensor along the cell type dimension. The final permutation-equivariant layer's

convolution operation is performed on the concatenation (along the epigenetic markers' dimension) of feature maps obtained from both the permutation-invariant and permutation-equivariant layers.

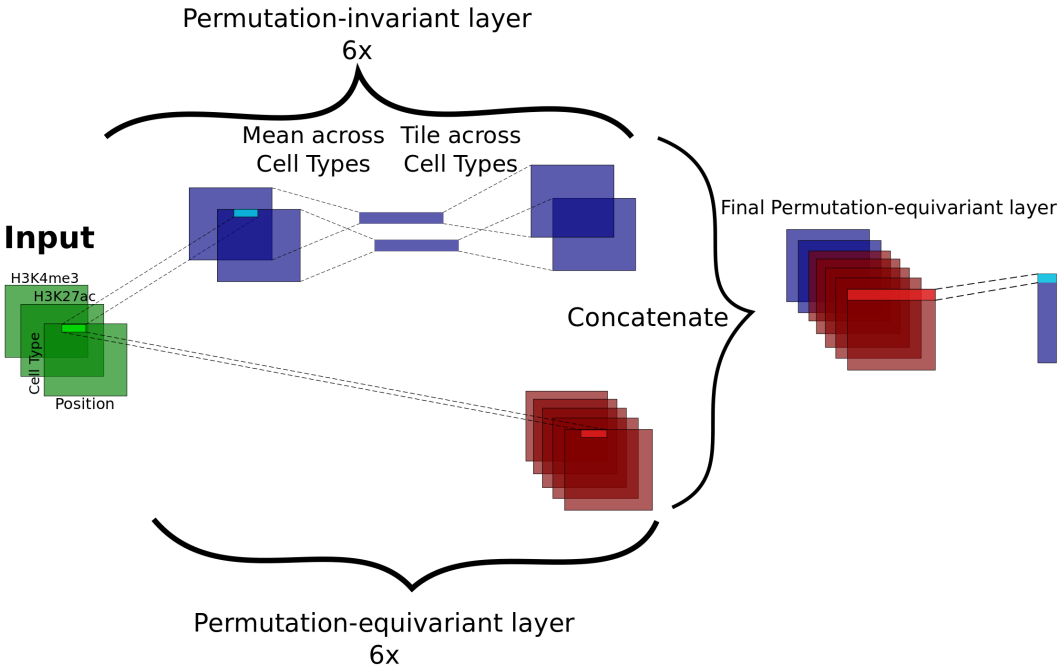

Figure 1: **Architecture of the permutation-equivariant convolutional neural network**. The input for each gene is a tensor in $\mathbb{R}^{12\times 6\times 200}$ corresponding to the 12 cell types, 6 histone modifications and 200 bins at 100bp resolution, corresponding to an input context size of $+/-$ 10Kbp, centered at the TSS of each gene. Each bin contains the average $\log(1-\log_{10}(\text{p-value}))$ of the histone modification's ChIP-seq peak calls in that 100bp bin. The output of the network is a vector in $\mathbb{R}^{12}$ corresponding to the $\log(1+\text{TPM})$ gene expression values for the 12 cell types.

2.3 Optimizing gene expression

Let f denote our trained neural network that predicts gene expression from epigenetic features. Given a gene with epigenetic features W_0 and a desired expression level vector y_{desired}, we find epigenetic features W such that $f(W) \approx y_{\text{desired}}$ by optimizing the following objective function via gradient descent, with W initialized to W_0:

$$\min_{W} \|f(W) - y_{\text{desired}}\|_2^2 \tag{1}$$

3 Results

We trained a permutation-equivariant convolutional neural network, as described in Figure 1, on the genes in the first 12 chromosomes, to predict gene expression using histone modification data described in Table 1. Figure 2A shows that the model achieves Spearman's rank correlation ~ 0.49 and Pearson's correlation ~ 0.65, on average, across the twelve different cell types.

Using the trained model as an oracle, we investigated the rules that the network had learnt in order to predict gene expression, using occlusion experiments similar to those in computer vision [18]. In order to do so, we artificially ablate all *peaks* from a particular histone modification's track, marginally, for a given gene, and then slide a *peak* of length 1Kb (corresponding to a p-value of 10^{-20}) along the 20Kbp input context, with stride 1Kb. Figure 2B summarizes the output of such a process for a representative gene using the amnion cell type's gene expression.

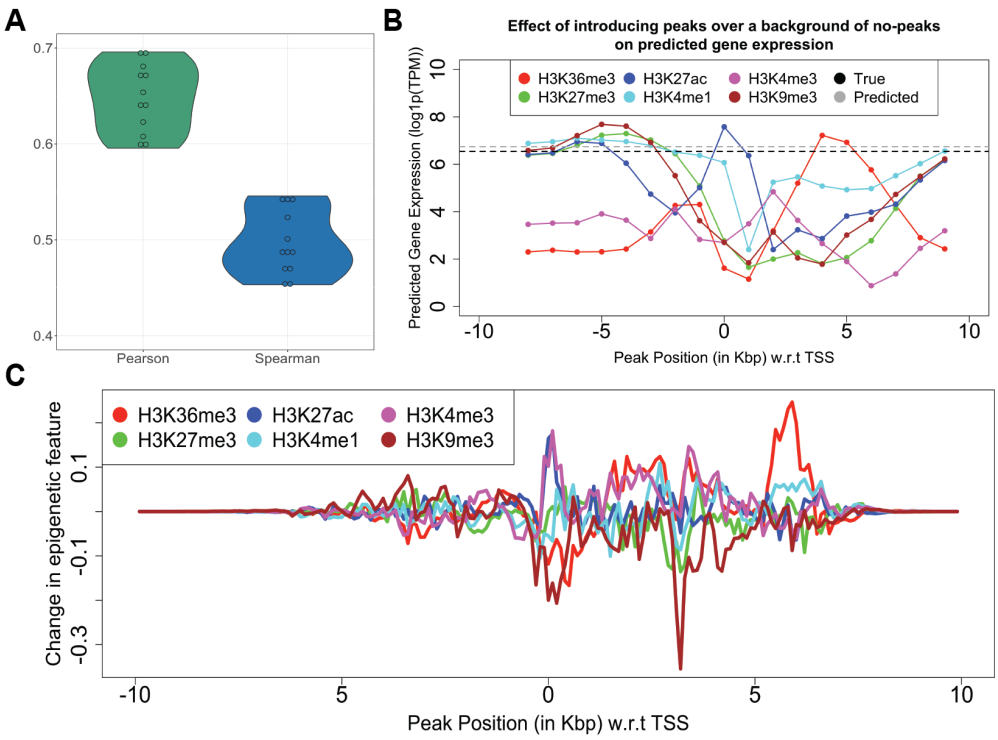

Figure 2: **Results of the permutation-equivariant neural network.** (A) The correlation between the true and predicted gene expression values for held-out chromosomes across different cell types is shown here. The average pearson correlation across the different cell types is ~ 0.65 and the average spearman correlation is ~ 0.49. (B) *in-silico epigenesis*: sliding *peaks* along a background of no peaks, for each epigenetic feature marginally, leads to different values of predicted gene expression. (C) $W - W_0$ after optimizing the gene expression of all cell types to increase by 100-fold, for a randomly chosen gene.

We observe that the H3K36me3 mark, which is known to be present inside gene bodies of actively expressed genes [19], leads to increase in gene expression when its peak is located within the gene body. Similarly, H3K27ac, which is also known to be present at the TSS of highly transcribed genes, increases gene expression when its peak is placed at the TSS. In contrast, we observe that for H3K27me3 and H3K4me1, a peak near the TSS leads to significant reduction in gene expression, in concordance with the literature [19].

Finally, using a strategy similar to activation-maximization [20], we use the trained model to optimize over the inputs for obtaining a desired predicted expression level, as described in Section 2.3. Figure 2C shows how the network is able to achieve the desired level of gene expression by optimizing over the input W_0 to obtain a new set of epigenetic features W, whose predicted expression matches the desired gene expression level. In order to increase the gene expression by 100-fold, the network created peaks at the TSS for H3K27ac and H3K4me3, which are known to promote transcription [19]. The network has also decreased the presence of H3K9me3 downstream of the TSS. H3K9me3 is known to be present in the gene bodies of low expression genes. Similarly, H3K36me3, which is present along the coding regions of highly expressed genes, has been added in the region downstream of the TSS.

4 Discussion

We have proposed a method to learn putatively causal mechanisms by which epigenetic features regulate gene expression. Consequently, in order to obtain a desired gene expression level, recently published tools such as Ledidi [21] could be used in conjunction with our proposed optimization approach to achieve a desired gene expression level by optimizing over sequence inputs.

References

[1] ENCODE Project Consortium et al. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57–74, 2012.

[2] Bradley E Bernstein, John A Stamatoyannopoulos, Joseph F Costello, Bing Ren, Aleksandar Milosavljevic, Alexander Meissner, Manolis Kellis, Marco A Marra, Arthur L Beaudet, Joseph R Ecker, et al. The nih roadmap epigenomics mapping consortium. *Nature biotechnology*, 28(10):1045–1048, 2010.

[3] David R Kelley, Yakir A Reshef, Maxwell Bileschi, David Belanger, Cory Y McLean, and Jasper Snoek. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome research*, 28(5):739–750, 2018.

[4] Jian Zhou, Chandra L Theesfeld, Kevin Yao, Kathleen M Chen, Aaron K Wong, and Olga G Troyanskaya. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nature genetics*, 50(8):1171–1179, 2018.

[5] Vikram Agarwal and Jay Shendure. Predicting mrna abundance directly from genomic sequence using deep convolutional neural networks. *Cell Reports*, 31(7):107663, 2020.

[6] Florian Schmidt, Fabian Kern, and Marcel H Schulz. Integrative prediction of gene expression with chromatin accessibility and conformation data. *Epigenetics & chromatin*, 13(1):4, 2020.

[7] Wanwen Zeng, Yong Wang, and Rui Jiang. Integrating distal and proximal information to predict gene expression via a densely connected convolutional neural network. *Bioinformatics*, 36(2):496–503, 2020.

[8] Ritambhara Singh, Jack Lanchantin, Gabriel Robins, and Yanjun Qi. Deepchrome: deep-learning for predicting gene expression from histone modifications. *Bioinformatics*, 32(17):i639–i648, 2016.

[9] Aaron D Goldberg, C David Allis, and Emily Bernstein. Epigenetics: a landscape takes shape. *Cell*, 128(4):635–638, 2007.

[10] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[11] Jeffrey Chan, Valerio Perrone, Jeffrey Spence, Paul Jenkins, Sara Mathieson, and Yun Song. A likelihood-free inference framework for population genetic data using exchangeable neural networks. In *Advances in neural information processing systems*, pages 8594–8605, 2018.

[12] Hong Yu, Shanshan Zhu, Bing Zhou, Huiling Xue, and Jing-Dong J Han. Inferring causal relationships among different histone modifications and gene expression. *Genome research*, 18(8):1314–1324, 2008.

[13] Jianxing Feng, Tao Liu, Bo Qin, Yong Zhang, and Xiaole Shirley Liu. Identifying chip-seq enrichment using macs. *Nature protocols*, 7(9):1728–1740, 2012.

[14] `https://github.com/ENCODE-DCC/imputation_challenge`, 2018.

[15] Ian Dunham, Ewan Birney, Bryan R Lajoie, Amartya Sanyal, Xianjun Dong, Melissa Greven, Xinying Lin, Jie Wang, Troy W Whitfield, Jiali Zhuang, et al. An integrated encyclopedia of dna elements in the human genome. *Nature*, 2012.

[16] Jacob Schreiber, Timothy Durham, Jeffrey Bilmes, and William Stafford Noble. Avocado: a multi-scale deep tensor factorization method learns a latent representation of the human epigenome. *Genome biology*, 21(1):1–18, 2020.

[17] Charles R Harris, K Jarrod Millman, Stéfan J van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J Smith, et al. Array programming with numpy. *Nature*, 585(7825):357–362, 2020.

[18] Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579*, 2015.

[19] Hiroshi Kimura. Histone modifications for human epigenome analysis. *Journal of human genetics*, 58(7):439–445, 2013.

[20] Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. Visualizing higher-layer features of a deep network. *University of Montreal*, 1341(3):1, 2009.

[21] Jacob Schreiber, Yang Young Lu, and William Stafford Noble. Ledidi: Designing genome edits that induce functional activity. *bioRxiv*, 2020.