An Attention-Based Model for Transcription Factor Binding Site Prediction

Gunjan Baid

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2018-83
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2018/EECS-2018-83.html

May 19, 2018

An Attention-Based Model for Transcription Factor Binding Site Prediction

by

Gunjan Baid

A thesis submitted in partial satisfaction of the
requirements for the degree of
Master of Science

in

Electrical Engineering and Computer Sciences

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Anthony D. Joseph, Advisor
Professor Joseph E. Gonzalez, Second Reader

Spring 2018

Abstract

An Attention-Based Model for Transcription Factor Binding Site Prediction

by

Gunjan Baid

Master of Science in Electrical Engineering and Computer Sciences

University of California, Berkeley

Professor Anthony D. Joseph, Advisor

We propose an attention-based approach for accurately predicting transcription factor binding sites. Our method combines DNA sequence with partially observed labels from epigenetic experiments to impute the values of missing labels, allowing for better predictions as more label information is known beforehand. We train and evaluate this model on cell lines from the ENCODE consortium [5, 14] and show that our model performs well on standard prediction tasks and further improves when partial data becomes available. The main contributions of our approach are generalization to unseen cell types and informed experimental design. Our model is able to reliably predict binding sites for cell types never seen during training. In addition, we use a beam search to identify the set of experimental labels that maximize prediction accuracy on missing data. The results of this beam search

can be used to inform cost-efficient experimental design under limited resources.

Contents

List of Figures

List of Tables

Acknowledgments

I want to thank my advisor, Professor Anthony Joseph, for his guidance throughout the past year. I feel fortunate to have had the opportunity to work with him and the Big Data Genomics group in the RISELab. I would also like to thank Professor Joseph Gonzalez, my second reader, for his feedback at every step of this work. Professor Nir Yosef was also heavily involved with the project and I am thankful for his thoughtful insights and intuitions. Thank you to Ashish Vaswani at Google Brain for the many Sunday night conference calls held to discuss this work.

This project was done in collaboration with Alyssa Morrow, Alexander Ku, and Weston Hughes. Thanks to Alyssa for assembling this group and being our fearless leader, to Alex for his uplifting enthusiasm, and to Weston for his wealth of ideas. I am lucky to have found a group that shares my interests and have learned so much from all of you over the last year.

I am also grateful to Nishant Desai, who has been both my closest companion and a mentor throughout my time at Berkeley. Most importantly, I want to thank my parents for their unwavering support of my education and the countless hours spent driving me to and from Berkeley over the past six years. Finally, I would like to dedicate this work to my younger sister, who I hope will someday read enough of it to find this dedication.

Chapter 1

Introduction

Transcription, the biological process through which DNA is transcribed into RNA, is heavily regulated by DNA-binding transcription factors. Transcription factors are proteins that can increase or decrease the rate of transcription, directly affecting gene expression in cells. As a result, characterizing the interplay between transcription factors and DNA is an important step in understanding gene expression. Chromatin immunoprecipitation sequencing (ChIP-seq) is a wet lab procedure that can be used to map out the global binding sites for many proteins of interest, including transcription factors. However, ChIP-seq experiments are time and resource-intensive, as a separate assay must be performed for each transcription factor and cell type combination of interest. This process can be limiting, thus motivating the need for a computational approach. Deep learning is a cost-efficient alternative that offers the scale needed to handle the massive amounts of data generated by high-throughput sequencing.

We can frame transcription factor binding site prediction as a multi-label classification task for which the input is a DNA sequence of length n, and the output is a binary vector of m predictions, with each prediction corresponding to a different epigenetic mark. The predictions are not independent, as it is well-known that DNA binding is highly correlated for transcription factors that interact through the formation of protein complexes or contain similar DNA-binding domains [8].

Previous models for transcription factor binding site prediction rely solely on sequence and do not aim to capture the dependencies among labels. While these approaches have been successful, their predictive power does not increase with the availability of more information. Specifically, when a subset of labels is known beforehand, these models are unable to incorporate them into the prediction. Previous approaches are also unable to generalize to new cell types not seen during training. DeepSEA, the current state-of-the-art approach, is limited to the fixed set of cell types used during training, as the model's label space consists of transcription factor and cell type pairs [17].

In this work, we develop a computational framework for transcription factor binding site prediction that uses not only DNA sequence, but also previously known labels to impute the missing labels. Our approach modifies the Transformer network, a sequence-to-sequence model introduced by Vaswani et al. [15], for the transcription factor binding site prediction task. We show how prediction accuracy increases with the availability of more experimental data. We also demonstrate the transfer learning capabilities of our model by training and testing on different cell types. Our predictions are not tied to a particular cell type, allowing us to generalize to arbitrary cell types never seen during training.

1.1 Outline

The remaining chapters in this report are structured as follows. Chapter 2 reviews background information on the process of transcription, biological assays, and the Transformer model. Chapter 3 covers related work in transcription factor binding site prediction. Chapter 4 goes into detail on our methods and model architecture. Results and discussion are presented in Chapter 5, and Chapter 6 concludes with potential areas of further exploration.

Chapter 2

Background

This chapter begins by outlining the process of transcription and the role played by transcription factors. The second section covers the ChIP-seq procedure, which is the source of our ground truth labels. Finally, we describe the Transformer network, a sequence-to-sequence architecture that uses self-attention.

2.1 Transcription

The central dogma of biology describes the flow of genetic information from DNA to RNA to proteins. During transcription, DNA is transcribed into RNA, and during translation, the RNA transcript is used to synthesize proteins. Though each cell in an organism contains the same genome, differential gene expression leads to the development of cell types.

Gene expression begins with translation, during which RNA polymerase is recruited to the gene promoter site. RNA polymerase is an enzyme that synthesizes the corresponding RNA from a DNA template. One or more transcription factors, proteins that directly bind to the DNA, may serve to either activate or block the recruitment of RNA polymerase, thereby directly upregulating or downregulating the given gene. Transcription factors can act by binding to the promoter site, which is proximal to the gene, or to distal enhancer sites. While enhancers are linearly distant from the promoter, they are brought into close spatial contact through folding of the DNA [6].

Transcription factor binding depends on the structure of DNA-binding domains, which are unique to each protein and recognize specific DNA sequences. Mutations in the target DNA sequence may weaken transcription factor binding affinity. Epigenetic factors such as the physical structure of chromatin also have an impact on binding [6]. Nuclear DNA is packed tightly in chromatin complexes, meaning that large stretches are inaccessible to outside proteins that regulate transcription. Such regions are termed closed chromatin, whereas open chromatin is physically accessible. Open chromatin is associated with active transcription, as the loose structure allows transcription factors, RNA polymerase, and other molecules to access sites of interest. Often, transcription factors do not bind independently but are part of larger protein complexes that consist of multiple transcription factors. When this is the case, the binding of one transcription factor may be related to, or even conditioned on, the binding of another one [8].

2.2 Data Types

ChIP-seq

Chromatin Immunoprecipitation with sequencing (ChIP-seq) is a technique used to map the genome-wide binding sites for a protein of interest. The ChIP procedure is used to isolate and amplify regions containing binding sites. This is done by crosslinking DNA with bound cellular proteins, after which the DNA is sheared into many fragments. The fragments containing the protein of interest can be isolated using a specific antibody which binds to the target protein. DNA fragments are then extracted from these protein-antibody complexes, sequenced using high-throughput sequencing, and analyzed with peak calling algorithms to produce a ranking of potential binding sites [7].

DNase-Seq and ATAC-Seq

While this work focuses primarily on transcription factor binding, our label space also contains binary labels from DNase I hypersensitivity sites sequencing (DNase-seq). DNase-seq is a technique used to characterize the accessibility of regulatory regions in the genome. The DNase I enzyme selectively digests open chromatin regions in the genome, which are easier to access than closed chromatin. The cut sites, also called DNase I hypersensitivity sites, indicate which genomic regions are active. Active fragments can be isolated and sequenced using high-throughput sequencing. Similar to ChIP-Seq, DNase-seq data can be analyzed using peak calling algorithms [4, 3].

Though DNase-seq has been widely used in the past, Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq) is now the preferred method of characterizing epigenetic state. ATAC-seq assays measure chromatin accessibility using transposase enzymes, which are used to cut and label accessible regions of chromatin. These regions can then be isolated and sequenced using high-throughput sequencing. ATAC-seq and DNase-seq have different experimental biases, so the data from both assays is not interchangeable. For example, DNase preferentially binds to T-rich segments of the genome, so higher peaks are seen in such regions. Simple methods for dataset normalization across both assays have recently been proposed, but further validation of these methods remains pending [11].

2.3 The Transformer Model

Deep neural networks have shown great success in many data-rich domains, such as natural language and image classification. These networks can be thought of as universal function approximators that learn complicated relationships between structured inputs and outputs. Recent advancements in wet-lab procedures and high-throughput sequencing have generated vast amounts of genomic data, opening up a new domain for the application of deep learning. Genomic data is particularly interesting because it bears similarity to both natural language and images. DNA consists of a structured alphabet and also bears known motifs against background noise. Previous works in genomics have used techniques from both computer vision, such as convolutional neural networks, and natural language, such as recurrent neural networks.

Our approach uses a Transformer network [15], a sequence-to-sequence model that was originally designed for machine translation and has also been applied to generative problems in computer vision [12]. The original design relies solely on attention to transform one sequence into another. The Transformer architecture consists of multiple layers of encoder and decoder components. Both encoder and decoder use multi-headed attention, which consists of several layers of scaled dot-product attention. Scaled dot-product attention is computed as follows, where K, V, Q are the keys, values, and queries and d is the dimension of all embeddings [15].

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V$$

The authors found it beneficial to compute several attention functions by linearly projecting the keys, values, and queries into multiple representation spaces.

In an encoder component, multi-headed self-attention is followed by a fully-connected layer. In a decoder component, multi-headed self-attention is followed by another multi-headed attention layer, which combines information from the encoder, and then a fully-connected layer. For self-attention layers, the keys, values, and queries all come from either the encoder or decoder. For the encoder-decoder attention, the keys and values come from the encoder while the queries come from the decoder [15].

Chapter 3

Related Works

3.1 DeepSEA

DeepSEA is framework that uses a convolutional neural network to jointly predict 919 chromatin marks solely from DNA sequence. Each of these labels, which include transcription factors, histone modifications, and DNase I hypersensitivity sites, corresponds to a cell type-specific prediction. The DeepSEA architecture consists of three convolutional layers of sizes 320, 480, and 690, followed by a fully connected layer. The model is hosted online and can be used to characterize the effects of single nucleotide polymorphisms (SNPs), which are single base changes in the DNA.

The model is trained on all 200bp regions of the genome, flanked by 400bp on each side. Each value in the label applies only the middle 200bp of the DNA sequence, and the 400bp flange on each side provides additional context. The DNA sequences come from the hg37 reference genome and only those sequences with at least one transcription factor binding event are included in the dataset. The dataset suffers from class imbalance as most of the labels are negative.

The authors found that increasing the flange length significantly improved model performance, suggesting that contextual information is crucial in understanding the properties of sequences. DeepSEA achieves average area under Receiver Operating Characteristic curve (AUC) values of 0.915, 0.948, and 0.852 for DNase I hypersensitivity sites, transcription factors, and histone modifications, respectively [17].

Chapter 4

Approach

4.1 Dataset

We use the dataset published by the authors of DeepSEA, which contains 4.2 million training examples from the ENCODE consortium [5, 14]. We consider only a subset of the labels present in the DeepSEA dataset. Specifically, we look at the GM12878, H1-hESC, HeLa-S3, HepG2, and K562 cell types. We train on data from the GM12878, HeLa-S3, HepG2, and K562 cell types and evaluate on held out chromosomes from these cell types. We also test the model on data from the H1-hESC cell type, which is never seen during training, to evaluate the transfer learning capabilities of our model. This setup is similar to a zero-shot learning problem, as the model is trained without any data from the test cell type. There is biological motivation for such a setup, as we seek to apply our model to newly sequenced cell populations.

4.2 Model Architecture

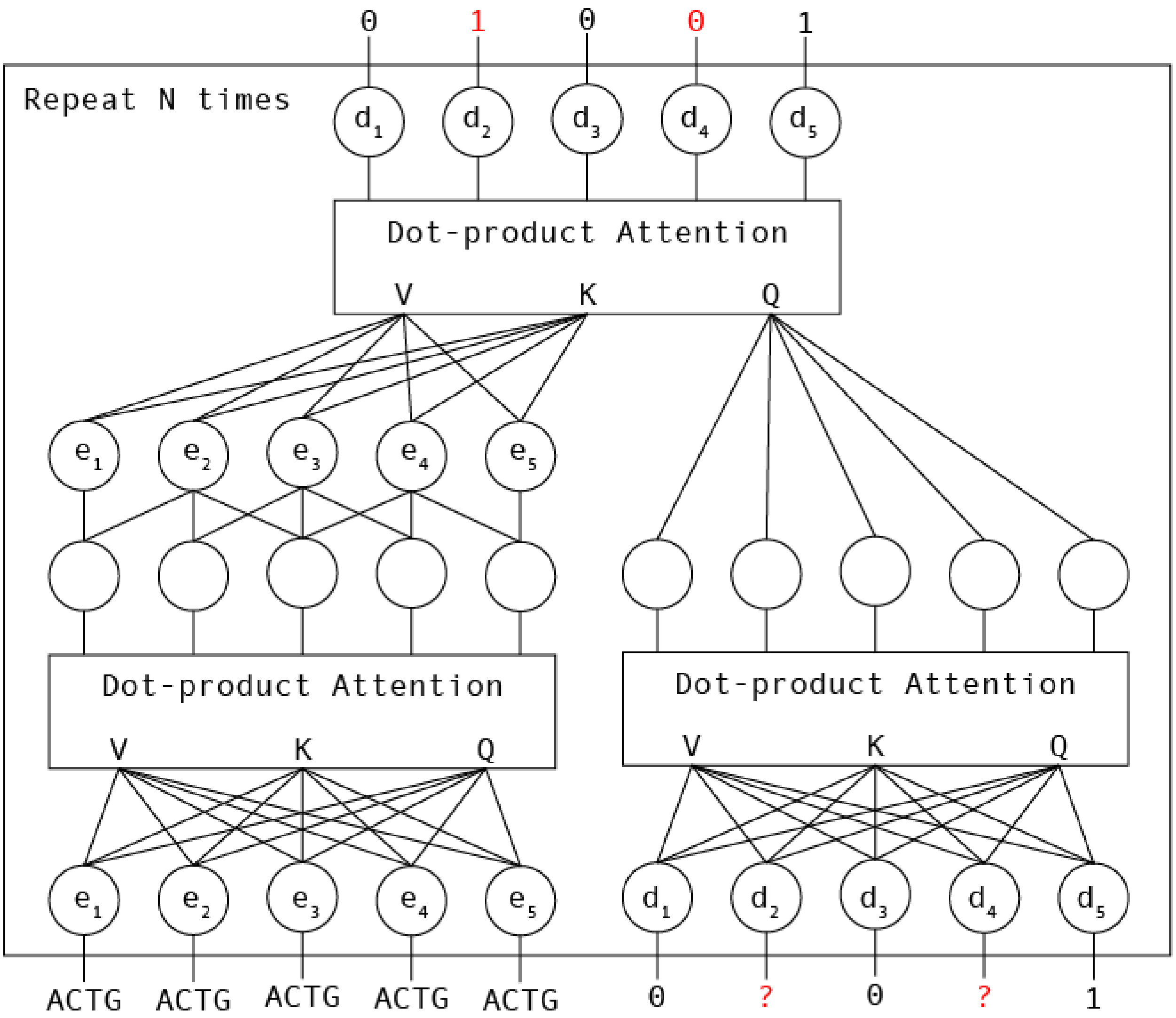

Figure 4.1: Our modified Transformer network architecture.

We modify several components of the original Transformer architecture. On the encoder side, our model takes in as input an n-gram embedded DNA sequence. We choose to work specifically with 4-grams to ensure a balance between expressive power and model complexity.

Following the multi-headed attention, the encoder uses convolutional layers, instead of fully-connected layers. We found that switching from fully-connected to convolutional layers significantly improved model performance.

On the decoder side, the network takes in as input a sequence of partially observed labels, where missing labels are denoted with a unique value, and outputs a sequence of binary labels corresponding to positive and negative predictions. All missing labels are imputed in the final output. We do not modify the fully-connected layers on the decoder side. Since there is no underlying order in the label space, we remove the positional embedding for the labels. Unlike the standard Transformer network, this imputation model does not mask decoder self-attention. The decoder generates outputs using the latent input, so we want to allow each position to attend to all positions, including the subsequent ones.

4.3 Input and Label Embeddings

For the DNA sequence input, we use the DNA encoder from the *tensor2tensor* framework [16] to represent each 4-gram with a unique token. We embed each of the 4-grams using a learned embedding matrix and use sinusoidal positional encoding, similar to what was used by Vaswani et al [15]. For the labels, we use position-wise learned embeddings, as a particular label for one binding event should be interpreted differently from the same label for another binding event. For each label, we learn an embedding for the positive, negative, and unknown values, and all learned embeddings are of dimension 512.

4.4 Training Procedure

To train a model that can perform imputation of missing data at test time, we introduce stochasticity into the training procedure by randomly discarding a subset of the labels. The proportion of labels fed to the model is a hyperparameter, p, where any particular label is kept with probability p, and discarded labels are replaced with an unknown token with probability $1 - p$. We implement this procedure by generating a random boolean mask for each batch of training examples. When evaluating the model, we modify this procedure to mask out the same labels across all examples in a batch in order to simulate the setup of missing experimental data. We compute cross entropy loss with respect to all labels, including the ones fed into the model, and we evaluate AUC only with respect to the imputed labels.

Chapter 5

Results and Discussion

We evaluate the performance of single-cell models trained on data from only GM12878 along with multi-cell models trained on data from GM12878, HeLa-S3, HepG2, and K562. Single-cell models are evaluated using only data from held out chromosomes of the same cell type, whereas multi-cell models are also applied to a new cell type. We evaluate the performance of each model using average area under Receiver Operating Characteristic curve (AUC).

5.1 Comparison to DeepSEA

We compare our model to DeepSEA, a state-of-the-art convolutional model for predicting transcription factor binding sites from DNA sequence [17]. We trained our model using the same training and validation sets as DeepSEA, but considered only a subset of their labels corresponding to the GM12878 cell type. Our label space includes 35 epigenetic marks,

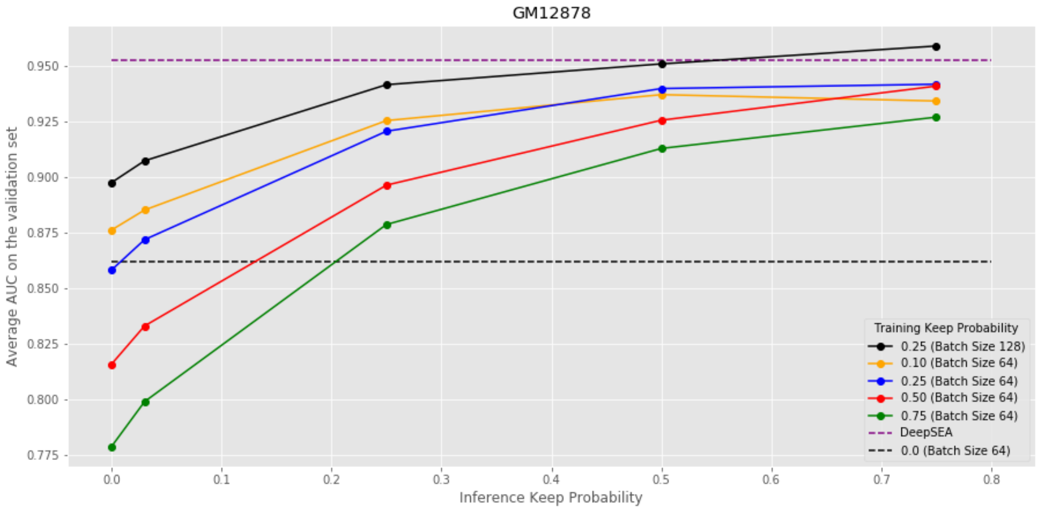

Figure 5.1: Average AUC vs. inference keep probabilities for GM12878 single-cell models.

including 34 transcription factor binding events and a binary DNase label. Each of these marks represents a separate but correlated binary prediction task. Each model is evaluated using average AUC across these 35 marks on held out chromosomes of the same cell type.

We compare both our base model with a keep probability of 0 and other models with varying keep probabilities at training and test time to DeepSEA. The keep probability is analogous to the proportion of labels used during training. Surprisingly, we see that models trained with lower keep probabilities generally perform better. For each model, the inclusion of additional data during inference leads to better performance. We also see that larger batch sizes lead to gains in performance. Though we do not currently outperform DeepSEA, we hope that further tuning of our models will show improved results.

5.2 Transfer Learning to the H1-hESC Cell Type

While restricting our training to one cell type makes comparing to existing models easier, the more interesting use case of our model involves application to a new cell type not in the training set. To this end, we train a model on 19 epigenetic marks, 18 transcription factors and DNase, across GM12878, HeLa-S3, HepG2, and K562 cell types. We test this model on the H1-hESC cell type. Cell types and transcription factors were chosen to have a large intersection of marks over a sufficient number of cell types in DeepSEA's dataset. In this context, we expect imputation to be valuable, as though all cell types have the same genetic sequence, behavior of transcription factors differs and offers a view of epigenetic behavior. We see that our model is able to generalize well to the H1-hESC data, improving as more data is supplied during inference time.

Probability of keeping a mark	Average AUC on held out chromosome in H1-hESC	Average AUC on held out chromosome in training cells
0	0.82112056	0.87444746
0.03	0.8250943	0.8833169
0.25	0.8473549	0.9187583
0.5	0.8641222	0.9331556
0.75	0.8725221	0.937384

Table 5.1: Multi-cell model trained with keep probability of 0.25 and evaluated using different keep probabilities on held out chromosomes.

5.3 Greedy Forward Subset Selection

We perform greedy forward subset selection based on beam search over our model to determine the set of experiments which maximizes prediction accuracy on imputed experiments. Beam search provides a tractable approach to exploring the space of all experimental subsets.

We define E as the set of all experiments and $S \subseteq E$ as a subset of those experiments. Starting from the root of the search tree $\mathcal{B}_0 = \{\emptyset\}$, beam search computes candidate beams $\mathcal{B}'_i = \{S \cup \{e\} : S \in \mathcal{B}_{i-1}, e \in E, e \notin S\}$. Candidates are ordered according to a heuristic value function and only the top β experiment sets are kept, thus $\mathcal{B}_i \subseteq \mathcal{B}'_i$ and $|\mathcal{B}_i| = \beta$. For our heuristic, we use the average AUC computed over all epigenetic marks when partial information from experiments in S is provided to the model during inference.

Beam search is not guaranteed to give the optimal subset of experiments, but we can always trade computational efficiency for a better solution by increasing β. When β is infinitely large, beam search is equivalent to an exhaustive breadth-first search. Exhaustively searching the space of experiment subsets of size k is $\mathcal{O}(2^k)$ in time complexity, while beam search runs in $\mathcal{O}(\beta k)$.

Subset size	Experiments	Average AUC of imputed marks on held out chromosome
0	{}	0.869723
1	{CHD2}	0.905142
2	{DNase, CHD2}	0.924376
3	{RFX5, DNase, CHD2}	0.934909
4	{RFX5, EZH2, DNase, CHD2}	0.944609

Table 5.2: Average AUC of subsets found through beam search.

5.4 Multi-perturbation Shapley Value Analysis

Although beam search provides us with an optimal set of experiments, it does not provide us with insight into how specific assays affect the efficacy of our model. We perform Multi-perturbation Shapley Value Analysis (MSA) to learn marginal and cooperative contributions of epigenetic marks learned from our imputation model to determine the effects of different assay pairs on the accuracy of our model [10]. MSA has been previously used to identify the contributions of multiple genes to the success of specific biological pathways [9]. Here, we use these marginal contributions from MSA to identify any outliers that have synergistic or adversarial affects on prediction accuracy. We then analyze two-dimensional MSA to determine cooperative contribution among marks.

The Shapley value computes the overall gain from a subset of a coalition of players, and can determine the most important players in the outcome of a game [13]. To determine marginal and cooperative contributions of epigenetic marks in the prediction accuracy of our model, we define a coalitional game as a set of experiments N of size n and a value function v that measures the contribution of a subset of experiments in the coalition. We define the value function $v(S)$ as follows, where S is a subset of epigenetic experiments, and $\texttt{avgAUC}(S)$ is the average AUC computed over all epigenetic marks when partial information from experiments in S is provided to the model during inference.

$$v(S) = \texttt{avgAUC}(S) - \texttt{avgAUC}(\emptyset) \tag{5.1}$$

We use this value function to compute the generalized Shapley value ϕ for a given subset of epigenetic marks C.

$$\phi_C(v) = \sum_{T \subseteq N \setminus C} \frac{(n - |T| - |C|)!|T|!}{(n - |C| + 1)!} \sum_{S \subseteq C} (-1)^{|C| - |S|} v(S \cup T) \tag{5.2}$$

Computing the Shapley value for each experiment requires inference to be run on all possible combinations of partial experiments from the 19 original experiments, resulting in 524,287 combinations. To decrease the amount of computation during inference, we utilize Multi-perturbation Shapley Value Analysis (MSA) to predict the value function for combinations containing more than five experiments using projection pursuit regression [10].

We perform one and two-dimensional MSA to compute the marginal contributions of 19 epigenetic marks, including 18 transcription factors and one DNase chromatin accessibility mark. Figure 5.2 shows the marginal and shared contribution of all 19 marks. One-dimensional Shapley Analysis reveals CHD2, DNase, and p300 as highest marginal contributors in the system. These marginal contribution scores closely align with the average AUC calculated from each of the 19 marks.

We use two-dimensional MSA to determine the 2-D interactions between all pairs of marks. We define 2-D information between mark i and j $I_{i,j}$ as specified in [10], where $\phi(i, j)$ is the Shapley Value of i and j, $\phi_C(i, \bar{j})$ is the Shapley value in the game containing i but not j, and $\phi_C(j, \bar{i})$ is the Shapley value in the game containing j but not i.

$$I_{i,j} = \phi(i,j) - \phi_C(i,\bar{j}) - \phi_C(\bar{i},j) \tag{5.3}$$

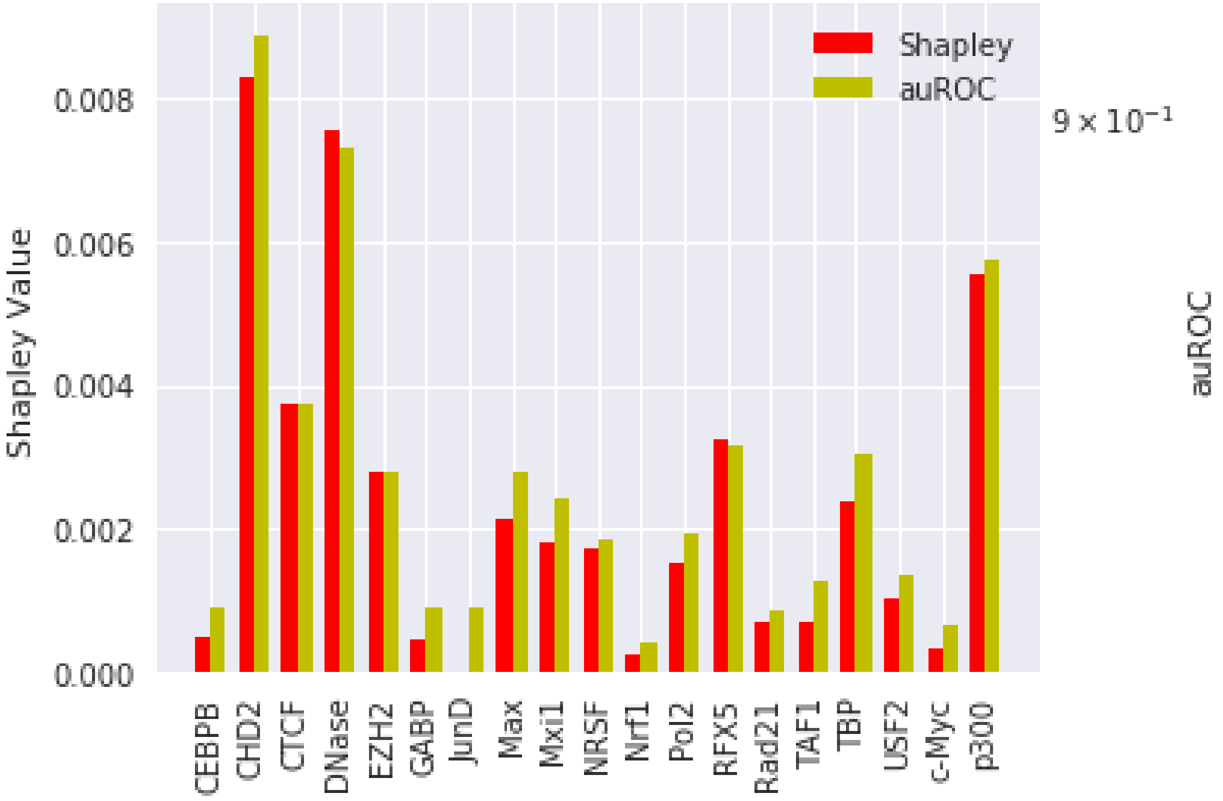

Figure 5.2: 1-D MSA.

We use the two-dimensional information to determine whether pairwise interactions are neutral, synergistic, or antagonistic. If $I_{i,j} = 0$, marks i and j are additive. If $I_{i,j} > 0$, the pairwise interactions are synergistic. If $I_{i,j} < 0$, the pairwise interactions are antagonistic.

Figure 5.3 demonstrates information computed from all pairwise combinations of the 19 marks. From this figure, we infer that JunD acts antagonistically with multiple other factors. This hypothesis is further supported in the one-dimensional MSA analysis, where

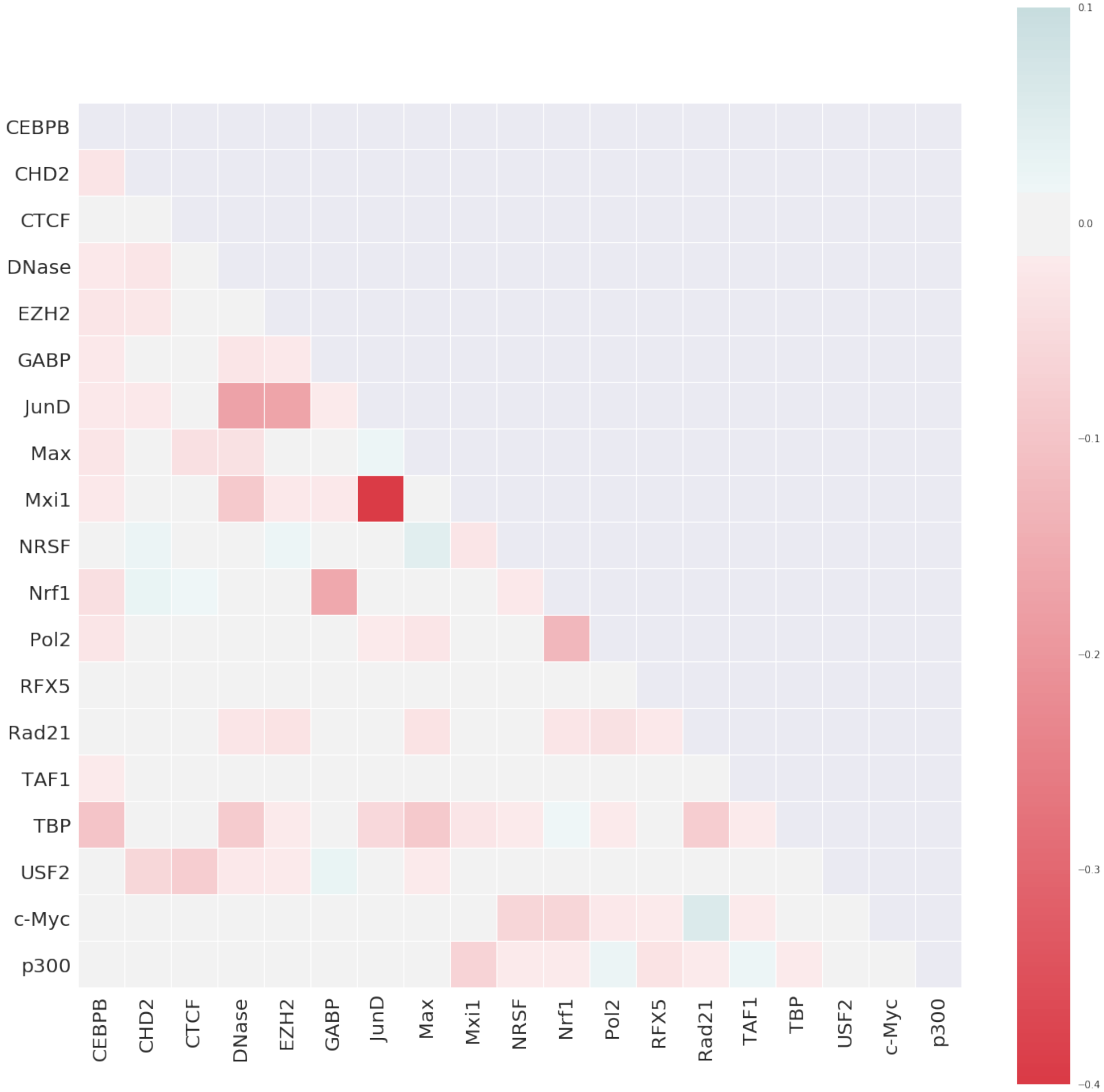

Figure 5.3: 2-D MSA.

the marginal contribution of JunD is zero.

5.5 Effects of a Weakened Decoder

Our experiments show that models trained with a lower keep probability generally perform better during evaluation. These results seem counterintuitive, as we would expect model performance to increase with the proportion of labels seen during training. One hypothesis is that models trained with higher keep probabilities overfit to the label information, thereby missing important information contained in the sequence. We investigate the effect of a weakened decoder on models trained with a keep probability of 0.50. In comparing the performance of models trained with two, four, and six decoder layers, we do not find any significant differences, so it is unclear why models trained using lower keep probabilities perform better. This experiment also serve to demonstrate the trade off between complexity and model performance. If computational power is limited, fewer decoder layers can be used at the cost of slightly worse performance, especially when the amount of available data is limited.

5.6 Network Interpretabilty

Interpretability is an important consideration for neural network models, which can often seem like a black box. Previous models such as DeepBind and Orbweaver [1, 2] have compared learned convolutional filters to position weight matrices, which are n x 4 matrices that define the likelihood of bases at each of the n positions. Learned convolutional filters are often very similar to position weight matrices for well-known motifs, thus acting as motif detectors in neural networks.

Due to the 4-gram embedding and self-attention layer that precede the convolutional layer of the encoder, the convolutional filters of the Transformer are not as easily interpretable. However, we can examine the attention maps to better understand the model. Figure 5.4 shows the encoder-decoder attention map from the final layer of the model. It can be seen that the model learns to place greater overall emphasis on the middle 200bp of the DNA sequence, which we know to be more important than the contextual flange.

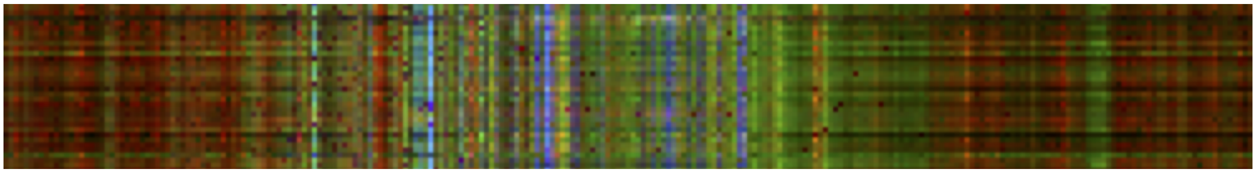

Figure 5.4: Encoder-Decoder attention map from the final layer of our model.

Chapter 6

Future Work and Conclusion

We have shown how the Transformer network can be modified to impute missing labels for the transcription factor binding site prediction task. While we do not yet achieve state-of-the-art results as measured by AUC, we hope to improve this work through further tuning. Nonetheless, our approach overcomes two limitations of previous works: the inability to leverage known experimental information and the lack of generalization to new cell types.

To address the first concern, our model is able to use any amount of known information for prediction. If no labels are known beforehand, our model uses just DNA sequence, and if some labels are known, these can be fed into the network to improve performance. Surprisingly, our experiments showed improved performance when the network was trained using a smaller proportion of the labels, and further work is needed to better understand these results.

Regarding improved generalization, we have shown preliminary results demonstrating the transfer learning capabilities of our model through testing on the H1-hESC cell type, which was never seen during training. A natural next step is application of the model to newly sequenced cell populations followed by experimental validation.

Bibliography

[1] Babak Alipanahi et al. "Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning". In: *Nature Biotechnology* 33 (July 2015), 831 EP -. URL: http://dx.doi.org/10.1038/nbt.3300.

[2] Nicholas E Banovich et al. "Impact of regulatory variation across human iPSCs and differentiated cells". In: *Genome Research* 28.1 (Jan. 2018), pp. 122–131. DOI: 10.1101/gr.224436.117. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5749177/.

[3] Alan P Boyle et al. "High-Resolution Mapping and Characterization of Open Chromatin across the Genome". In: *Cell* 132.2 (Jan. 2008), pp. 311–322. DOI: 10.1016/j.cell.2007.12.014. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2669738/.

[4] Jason D Buenrostro et al. "Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position". In: *Nature Methods* 10 (Oct. 2013), 1213 EP -. URL: http://dx.doi.org/10.1038/nmeth.2688.

[5] The ENCODE Project Consortium. “An Integrated Encyclopedia of DNA Elements in the Human Genome”. In: *Nature* 489.7414 (Sept. 2012), pp. 57–74. DOI: 10.1038/nature11247. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3439153/.

[6] N. Gheldof et al. “Cell-type-specific long-range looping interactions identify distant regulatory elements of the CFTR gene”. In: (2010). URL: https://academic.oup.com/nar/article-abstract/38/13/4325/2409392.

[7] David S. Johnson et al. “Genome-Wide Mapping of in Vivo Protein-DNA Interactions”. In: *Science* 316.5830 (2007), pp. 1497–1502. ISSN: 0036-8075. DOI: 10.1126/science.1141319. eprint: http://science.sciencemag.org/content/316/5830/1497.full.pdf. URL: http://science.sciencemag.org/content/316/5830/1497.

[8] A. Jolma et al. “DNA-dependent formation of transcription factor pairs alters their binding specificity”. In: (2015). URL: https://www.nature.com/articles/nature15518.

[9] Alon Kaufman, Martin Kupiec, and Eytan Ruppin. “Multi-knockout genetic network analysis: the Rad6 example.” In: *Proceedings. IEEE Computational Systems Bioinformatics Conference* (2001), pp. 332–340. ISSN: 1551-7497. DOI: 10.1109/CSB.2004.1332446. URL: http://www.ncbi.nlm.nih.gov/pubmed/16448026.

[10] Alon Keinan et al. “Fair attribution of functional contribution in artificial and biological networks.” In: *Neural computation* 16.9 (Sept. 2004), pp. 1887–1915. ISSN: 0899-7667. DOI: 10.1162/0899766041336387. URL: http://dx.doi.org/10.1162/0899766041336387.

[11] AndréL Martins et al. "Universal correction of enzymatic sequence bias reveals molecular signatures of protein/DNA interactions". In: *Nucleic Acids Research* 46.2 (Jan. 2018), e9–e9. DOI: 10.1093/nar/gkx1053. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5778497/.

[12] Niki Parmar et al. "Image Transformer". In: *CoRR* abs/1802.05751 (2018). arXiv: 1802.05751. URL: http://arxiv.org/abs/1802.05751.

[13] Management G. Owen?- Science and 1972. "Multilinear extensions of games". In: (1972). URL: https://pubsonline.informs.org/doi/abs/10.1287/mnsc.18.5.64.

[14] Cricket A Sloan et al. "ENCODE data at the ENCODE portal". In: *Nucleic Acids Research* 44.Database issue (Jan. 2016), pp. D726–D732. DOI: 10.1093/nar/gkv1160. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4702836/.

[15] Ashish Vaswani et al. "Attention is All you Need". In: *Advances in Neural Information Processing Systems 30* (2017).

[16] Ashish Vaswani et al. "Tensor2Tensor for Neural Machine Translation". In: *CoRR* abs/1803.07416 (2018). URL: http://arxiv.org/abs/1803.07416.

[17] Jian Zhou and Olga G Troyanskaya. "Predicting effects of noncoding variants with deep learning–based sequence model". In: *Nature Methods* 12 (Aug. 2015), 931 EP -. URL: http://dx.doi.org/10.1038/nmeth.3547.