

Interpretable Deep Neural Networks for Enhancer Prediction

Seong Gon Kim, Nawanol Theera-Ampornpunt, Ananth Grama, and Somali Chaterji

Department of Computer Science, Purdue University

West Lafayette, IN USA

{kim1871, ntheeraa, ayg, schaterj}@purdue.edu

Abstract—Enhancers are short DNA sequences that modulate gene expression patterns. Recent studies have shown that enhancer elements could be enriched for certain histone modification combinatorial codes, leading to interest in developing computational models to predict enhancer locations. Here we present *EP-DNN*, a protocol for predicting enhancers based on chromatin features, in two different cell types, a human embryonic (H1) and a human lung fibroblast (IMR90) cell line. Specifically, we use a deep neural network (DNN)-based architecture to extract enhancer signatures. We train *EP-DNN* using distal p300 binding sites, as enhancers, and TSS and random non-DNase-I hypersensitivity sites, as non-enhancers. We find that *EP-DNN* has superior accuracy relative to other state-of-the-art algorithms, such as DEEP-EN and RFECS, and also scales well to large number of predictions. Then, we surmount the problem that DNN results are not interpretable and develop a method to interpret which histone modifications are important, and within that, which spatial features proximal or distal to the enhancer site, are important. We uncover that the important histone modifications vary between cell types. Further, whether the important features are clustered around the enhancer peak or more spread out also differs among the different histone modifications. Thus, we bring forth a new paradigm for automatically determining the important features and the important histone modifications, rather than the current computational standard of using the *same* fixed number of features from *all* the histone modifications for *all* cell types. Our results have implications for computational scientists who can now do feature selection for their classification task and for biologists who can now experimentally collect data only for the relevant histone modifications.

Keywords—Enhancer predictions, deep neural networks, histone modification signatures, DNN interpretability.

I. INTRODUCTION

Distinct cell phenotypes are largely modulated by unique gene expression patterns, stemming from the interaction of the genome with its environment. Such crosstalk is mediated by specialized cis-regulatory modules (CRMs), including enhancers (1), silencers, promoters, and insulators (2-4). Among these, enhancers constitute the most prominent class of gene expression regulators. Early experiments indicated that sequences located far from the gene promoters are often responsible for mediating gene transcription (5). Such genetic elements are called *enhancers* and they are defined as short DNA sequences regulating temporal and cell-type specific basal gene-transcription levels at distances ranging

from hundreds of bases to, in rare cases, even megabases (6-8). Knowing their properties, regulatory activity, and genomic targets is crucial to the functional understanding of cellular events, ranging from cellular homeostasis to differentiation. Recent genome-wide investigation of epigenomic marks has indicated that enhancer elements could be enriched for certain epigenomic marks, such as complex, albeit predictive, histone combinatorial grammar codes. Our efforts in this paper are motivated by these recent advances in epigenomic profiling methods, which have uncovered enhancer-associated chromatin features in different cell types and organisms (9-12). Through analysis of the prior approaches, including experimental runs, we conclude that the difficulty of computationally predicting enhancer sites is because of two primary factors. First, they did not use the full spectrum of available features, i.e., all the histone modifications and their enrichment values in a wide region around the hypothesized enhancer site, denoted by the enhancer peak. Second, they did not use a highly expressive classifier, one that can extract the distinguishing features from a complex landscape.

We address both of these problems, the first by starting with an (almost) exhaustive set of features and then doing feature selection through an innovative mechanism, to identify a top- k most relevant features¹. For the second shortcoming, we use recent state-of-the-art Deep Learning methods and develop a deep neural network (DNN)-based architecture (13-15) to predict the presence and types of enhancers in the human genome, “learning” from the combinatorial histone modification codes. We call our system “EP-DNN”, an acronym for “*Enhancer Prediction through a Deep Neural Network*”. Our sanity check result shows the improvement in validation rate of EP-DNN compared to the two state-of-the-art computational approaches—RFECS (16) and DEEP-EN (17) (Figure 1)². We use the metric “validation rate” instead

¹ Empirically, the full set has 480 features and the value of k is 150 for two distinct human cell lines—H1 and IMR90.

² For the dataset, we used the H1 and IMR90 datasets that were generated as part of the NIH Roadmap Epigenome Project and that have been released to the public prior to publication (<http://www.genboree.org/epigenomeatlas/multiGridViewerPublic.rhtml>). Briefly, 24 chromatin modifications in human embryonic stem cell (H1) and primary lung fibroblast cells (IMR90) were generated by the Ren lab and deposited under the NCBI GEO accession number GSE16256. Additionally, two replicates of H3K9me3 datasets deposited under GEO accession numbers GSM18057 and GSM42829 were used. Genome-wide

of the more common “recall” and “precision” metrics because the ground truth is not known for some data samples. Hence, based on prior work (16), we use validation rate, which is defined as the percentage of predicted enhancers overlapping with DNase-I hypersensitivity sites (DHS) and binding sites of p300 and a few sequence-specific transcription factors (TFs) known to function in each cell type (true positive markers, TPMs).

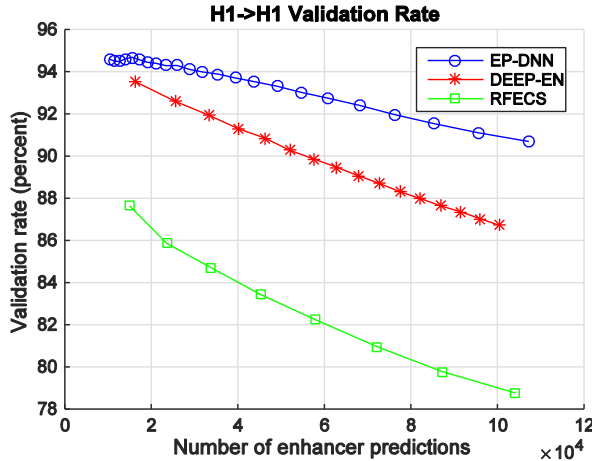


Figure 1: Comparison of EP-DNN against two state-of-the-art computational approaches for prediction of enhancer locations with respect to validation rate.

Looking further into the challenges of computational identification of enhancers, we identify several reasons (18). First, the search space for enhancers is large—billions of DNA base pairs—scattered across 98% of the non-coding genome. Second, while enhancers regulate genes in *cis*, they do not display distinct locational or orientation-centric signals relative to the genes that they regulate—potentially located upstream, downstream, or even in introns of the genes that they modulate, often regulating multiple genes (19). Enhancers function at a distance from their target genes via chromatin loops that bring the enhancers and target genes into proximity (20, 21), or via direct eRNA transcription from the enhancer DNA sequences (22).

Biological experiments for enhancer identification

Several high-throughput experimental approaches exist to identify enhancers (23, 24). The first is mapping specific transcription factor binding sites (TFBS) through ChIP-seq. This stems from the fact that short enhancer DNA sequences serve as binding sites for TFs, and the combined regulatory cues of all bound TFs determine ultimate enhancer activity (25, 26). However, this approach requires the knowledge of the TF subset that is not only expressed but also occupies all active enhancer regions in the spatio-temporal setting, such as in a specific cell type at a point of time (27). Therefore,

binding data for p300 in H1 and IMR90, and transcription factors NANOG, SOX2, and OCT4 in H1 were generated in the Ren lab using ChIP-seq and deposited under accession numbers GSE37858, GSE18292, and GSE17917 respectively.

predicting enhancer activity from sequence-based information, such as from the TF motif content, remains challenging (26, 28). In addition, this approach is limited by the lack of available ChIP-grade antibodies that specifically recognize these subsets. The second approach is based on mapping transcriptional co-activator binding sites (e.g., histone acetyltransferase HAT, p300) (29, 30). However, not all enhancers are marked by a set of co-activators and also often lack available ChIP-grade antibodies. The third approach relies on identifying open chromatin regions by DNase-I hypersensitivity (DHS) mapping (31), which lacks specificity due to the fact that the identified regions can correspond to other CRMs. Finally, the fourth approach involves histone modification signatures produced by ChIP-seq that consistently mark enhancer regions (16, 32-35). Due to their consistency in marking enhancers, we use histone modifications to make *in-silico* predictions of enhancer signatures.

Related Work: Previous Methods based on Histone Modifications

Several computational methods that use histone modification signatures to identify enhancer regions have been developed. Won *et al.* proposed the usage of Hidden Markov Models (HMMs) to predict enhancers using three primary histone modifications (33). Firpi *et al.* focused on the importance of recognizing the histone modification signals through data transformation and employed Time-Delayed Neural Networks (TDNNs) using a set of histone marks selected through simulated annealing (34). Fernández *et al.* used Support Vector Machines (SVMs) on an optimized set of histone modifications found through Genetic Algorithms (35). RFECS (Random Forest based Enhancer identification from Chromatin States) improved upon the limited number of training samples in previous approaches using Random Forests (RFs), in order to determine the optimal set of histone modifications to predict enhancers (16).

EP-DNN’s Contributions

In this paper, we solve the classification problem of whether a histone combinatorial code represents an enhancer element, or not, using our deep learning-based classifier, EP-DNN. Our main contributions in this paper are as follows:

- We have developed an efficient DNN-based classifier to identify enhancers in two distinct cell types, namely the human embryonic stem cell type (H1) and a primary lung fibroblast cell line (IMR90). We demonstrate that DNNs work well in extracting features automatically from a set of 480 features in 24 histone modifications, and then, in predicting enhancers. We take care to optimize the DNN while avoiding overfitting on the training data. We also show that DNN-based classification is much faster than prior computational methods, which often use an ensemble of models.
- We have extended and applied a previous analysis method for Neural Networks to work for DNNs to identify which features are important, thus solving the problem that DNNs are typically not interpretable. Then,

using a carefully designed approach that avoids the combinatorial explosion, we have determined which features need to be included in the classifier to obtain validation comparable to using the full feature set. In doing so, we have pruned the feature space to top-150 features from the original 480 features.

- We identify that not all histone modifications are equally important, and that the important histone modifications are different for the two different cell types, elucidating cell-type specific differences in the histone combinatorial code. Further, for an important histone modification, which subset of its features is important, varies from one histone modification to another. Thus, in this paper, we introduce a new paradigm where we do not have to use a fixed number of features, that is, fixed window sizes, for all histone modifications. Rather, we zoom into the important histone modifications and the important features within these modifications. Thanks to our solution, the experimentalists need only collect the data for the important histone modifications for any given cell type.

II. METHOD

A. Overview

We use histone modification signatures that are known to mark enhancer regions, as input features to our classifier in order to determine whether a particular region is an enhancer or not. There are many possible choices for a classifier for predicting which will be the enhancers. In this specific case, we could have chosen from many possible options for taking the histone modification features and predicting which indicate, with high likelihood, the presence of an enhancer. “Simple” classifiers include naïve Bayes and Support Vector Machine (SVM), and Decision Tree, while relatively more complex classifiers include Bayesian Networks, Hidden Markov Models, and Neural Networks. Ensemble methods that combine multiple base models have also been used in prior work—Random Forest, which uses multiple Decision Trees, as employed in RFECS (16), and ensemble of SVMs as used in DEEP-ENCODE (DEEP-EN (17)). In our work, we chose Deep Neural Networks (DNNs), arguably one of the more complex classifiers. Our design choice was based on the well-known advantages of DNNs—they provide feature extraction abilities without requiring manual feature engineering or transformation of the data. Further, they are generally less computationally expensive than larger ensemble methods that combine multiple algorithms or multiple models of the same kind. This has to be weighed against the disadvantages of DNNs, none of which are insurmountable. The first is that there is a need for suitable training of the DNN and the parameter space to explore can be quite large. The second is that DNN results are not as easily interpretable as say, that of a Decision Tree. We mitigate the first problem by performing an extensive training with a large amount of data and sweeping through the parameter space (the number of layers in the DNN, the number of neurons in each layer, the transfer function used) to decide on the optimal DNN architecture. We find empirically that the optimal DNN architecture does not

change significantly when we use the subsets of features. For the second problem, we employ a method to calculate the importance of each input feature based on the learned weights.

Figure 1 shows the comparison against previous methods RFECS (16) and DEEP-EN (17) in predicting enhancers in H1 cell type and demonstrates the superior prediction performance of DNN.

We give a high level overview of our solution approach in **Figure 2**, in which we show separately the initial training phase we use to create an optimal DNN with all features, the feature selection phase where useful features are separated from the noises, and the prediction phase where we use a subset of histone modifications and associated features to predict whether a regulatory motif is an enhancer. In the training phase, we start by finding the optimal DNN for all features. Next, we compute the features’ importance scores based on the learned weights in the optimal DNN. The features are then sorted by their importance scores and multiple DNNs are trained using only the top- k features, for varying values of k . We increase k in increments of 10, thereby avoiding a combinatorial explosion in the number of combinations of features that we need to consider. A subsequent evaluation of the DNNs using validation rate identifies the optimal DNN to use for the prediction phase.

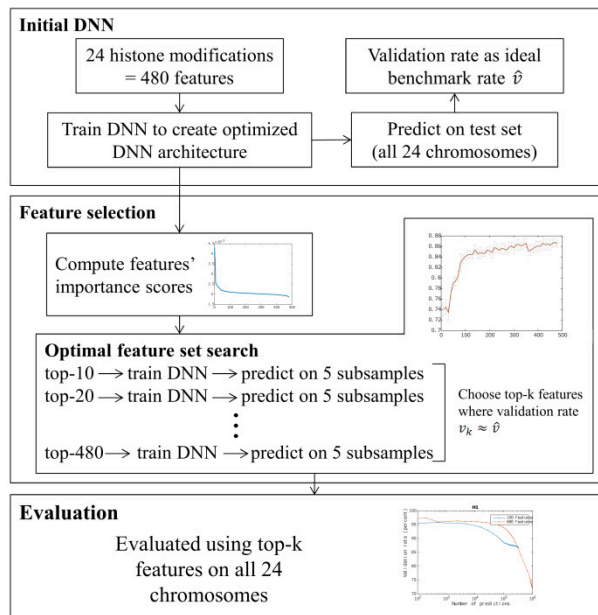


Figure 2: High level overview of our solution approach—EP-DNN. It shows the training phase, the feature sub-selection, and the final evaluation phase with the top- k features.

B. Optimization of our DNN’s feature space

The availability and difficulty of obtaining data as well as the loading and pre-processing time to read this data becomes a significant factor as the input feature space gets larger, especially when dealing with big data such as epigenomics data. Within the context of enhancer prediction, there are generally two approaches to alleviate this problem. The first is reducing the number of histone modifications

used as input to include the minimal amount necessary to make accurate predictions. However, the optimal histone modification set for each cell type is not well known and omitting certain histone modifications can potentially lead to a loss of important data. Previous work has pointed out some important histone modifications, though it has not rigorously justified which modification is important in which cell type. Another approach is to reduce the input window size. That is, for each histone modification, the range of base pairs used for prediction that are adjacent to each particular site being predicted can be reduced. However, a window size that is too large can include unnecessary noise, while a window size that is too small can lead to information loss. This problem is compounded by the fact that all prior works have used a fixed window size for *all* histone modifications. In reality, as we find through this current study, each modification signature contains different areas of significance around a particular site (**Figure 5**).

We develop the solution of reducing the number of features that need to be input to the DNN. In the full generality, there are 24 histone modifications and with the largest window size from prior works, 20 windows around the target site and correspondingly 20 features for each histone modification, thus leading to a total of 480 features. We posit that only a small fraction of this total number of features is important to the classification task, distributed in varying proportions, across the combinatorial histone modification signatures. To come up with the subset of features, we follow a two-step process.

Step 1: Analyze the weights of the edges in the DNN and come up with an approximate ranking of the importance of these features (see below for more details). Let us call this rank ordered list OL (Ordered List).

Step 2: Use the validation performance with all 480 features as the ideal goal (V_{opt}) and the benchmark. Now, perform classification by taking a cumulative group of 10 features from the rank-ordered list OL. Thus, first start with the top-10 features and by running the classifier on 5 different subsamples of the test set, see what the validation rate is. The subsample sets each contain 480K data points, and each was drawn randomly from the entire dataset of 21,247,949 data points without any overlaps. Then, repeat the process with the top-20 features, top-30 features, and so on. Observe when the validation rate plateaus and check that it reaches close to the ideal goal V_{opt} . Let us say that is achieved with the top- k features. That implies that we can perform further classification tasks with these k features, rather than all 480 features. Empirically, we find that for both cell lines—H1 and IMR90—the value of k is 150. Thus, this leads to a significant reduction in the feature size that our DNN needs to consider.

C. Interpretation of DNN results

We present a way to use the learned DNN connection weights to calculate the importance of each input feature and determine not only the histone modifications, but also the range of signal features within these histone modifications that influence enhancers. These contributions taken together make DNNs more interpretable than other methods, and

solve the aforementioned problems with DNNs. The trained weights contain information regarding the histone modification feature inputs and enhancers. In order to extract this information, we took a previous feature selection method (36) that determines feature importance from shallow Neural Network architecture connection weights, then expanded it for deep architectures and applied it to our initial-480 feature DNN. The equation used is given below, where i is a neuron whose importance score we are calculating, and N_i is the set of neurons in the next layer (closer to output) that i feeds into. The importance score of neuron i , denoted S_i , is computed as

$$S_i = \sum_{j \in N_i} |w_{ij}| S_j$$

Beginning with the output neuron’s importance scores set to 1, the importance scores are distributed to each neuron in the previous layers to find its importance score, in a backpropagation-like manner, down the DNN till the input features are reached. With only 1 output in our DNN, we further normalize each neuron’s weights to sum to 1, such that the final resulting feature importance scores sum to 1.

III. EXPERIMENTAL APPROACH

A. Datasets

We use 24 histone modifications for our enhancer prediction task. This choice is dictated by the fact that a genomics laboratory has previously generated this data and deposited it in the NCBI database under NCBI GEO accession number GSE16256. The 24 histone modifications are H2AK5ac, H2BK120ac, H2BK12ac, H2BK15ac, H2BK20ac, H2BK5ac, H3K14ac, H3K18ac, H3K23ac, H3K27ac, H3K27me3, H3K36me3, H3K4ac, H3K4me1, H3K4me2, H3K4me3, H3K56ac, H3K79me1, H3K79me2, H3K9ac, H3K9me3, H4K20me1, H4K5ac, and H4K91ac, in two distinct human cell types, embryonic stem cells (H1) and primary lung fibroblasts (IMR90), which were generated as a part of the NIH Epigenome Roadmap Project (10, 37). These histone modifications comprise a superset of all that are hypothesized to be possibly relevant biologically to the presence (or absence) of enhancers (16).

The ChIP-seq reads of these histone modifications give us the enhancement level of the modification. These were binned into 100 base pair (bp) intervals and normalized against their corresponding inputs by using an RPKM (reads per kilobase per million) measure (38). Multiple replicates of histone modifications were used to minimize batch-related differences, and the RPKM-level of the replicates was averaged to produce a single RPKM measurement per histone. The RPKM-levels were further zero-meaned and normalized by the standard deviation (Z-score) of the training set. The same mean and standard deviation from the training set were used to normalize the test set before prediction as well. We will refer to this normalized enrichment level of a histone modification as its signature. The histone modification signatures of each bin location are then used as input to the DNN.

To train our DNN, we first select distal p300 co-activator binding sites through ChIP-seq, then further select though overlapping DNase-I hypersensitive sites (DHSs) that are distal to TSS, as regions representing enhancers. Of these, 5,899 p300 peak calls were selected for H1 and 6,000 peaks for the IMR90 cell line to represent enhancers for the training set. However, p300 co-activators also bind to distal Transcription Start Sites (TSS), which are not enhancers. Therefore, we also select TSS that overlap with DHS, as well as random 100 bp bins that are distal to known DHS or TSS to represent non-enhancers. We include 9,299 TSS peaks from H1 and 8,000 peaks from IMR90 in our training set to distinguish between p300 binding sites that are enhancers and TSS that are not, and 31,994 random distal background sites were selected for H1 and 34,000 for IMR90 to represent non-enhancers for training.

For testing the DNN, we used all known distal p300 and CBP co-activator and Transcription Factor (TF) binding sites that overlap with DHS as positive enhancer sites, and TSS as non-enhancer sites.

B. Initial 480-feature DNN

We trained a fully-connected DNN with 480 inputs, 1 output, and softplus activation functions for each neuron to make enhancer predictions. Each input sample consists of twenty-four 20-dimensional vectors of 100 bp bin RPKM-levels, windowed from -1 to +1 kb at each bin location. Training was done in mini-batches of 100 samples through stochastic gradient descent backpropagation. To prevent overfitting, “dropout” training (39) was applied with a dropout rate of 0.5, along with a weight decay of 0.9. An optimal architecture of three hidden layers, comprising of 600 neurons in the first layer, 500 in the second, and 400 in the third, was found through cross-validation on half the training data selected randomly, and the full training set was used to train the model before prediction. A convergence on the mean squared error could be achieved with only 5 epochs of training. This extensive training mechanism was found to be suitable to optimize the DNN, even though a DNN has a fairly large parameter space.

C. Feature Importance

To show that our previously described method produces importance scores that are indicative of each feature’s *actual importance* in predicting enhancers, the input features were sorted by their scores. Then we chose of a set of the top 10 features and trained a DNN. This DNN was tested against 5 subsample sets each consisting of 480K random data points sampled without replacement across all 24 chromosomes. We continuously repeated this procedure by incrementing the size of the set in batches of 10 features, then training and testing a new DNN for the entire set of 480 features.

The results show that a plateau is reached at a certain input feature set size. This feature set was used to train a DNN, identical to the initial DNN but with a reduced number of features with the highest importance scores, then the performance of the initial DNN using all 480 features as input and the reduced DNN using selected features was compared for evaluation.

D. Validation

The standard precision and recall metrics misrepresent actual prediction performance on real data, since there are many more unknown functional sites than just the p300, CBP, NANOG, SOX2, OCT4 binding enhancers or TSS. Ideally, we would have to evaluate performance on all these sites that are unaccounted for. However, most are not experimentally verified and are unknown. Thus, there is not enough data to make an accurate evaluation of the precision and recall of any computational model. This observation has been made by prior computational approaches for enhancer prediction, such as RFECS. Consequently, they have also not used the standard precision and recall metrics in their evaluation.

Furthermore, functional enhancers are experimentally verified by single peak locations. However, in reality, enhancers exist in various levels (heights) and sizes (widths) that more or less gradually decrease around the peaks. These peaks are not available during prediction on real data because we are trying to predict for locations that have not yet been experimentally verified. Therefore, any computational model must be able to predict for the peak as well as the surrounding non-peak regions. Further, the evaluation method must synthesize some criterion to determine what is the ground truth (is it an enhancer or not) for any genic region away from the peak location. Therefore, the traditional evaluation using precision and recall metrics cannot be used in this case.

Consequently, RFECS introduced the notion of validation to solve this problem by checking and validating a prediction against criteria mentioned at the end of this section. If a location has histone modification enrichment signatures similar to that of an enhancer and a prediction is made on that location, we can say the prediction is validated, given that the location is sufficiently close to either a known enhancer peak marker or an open chromatin site. However, the RFECS method of evaluation has one drawback. This method singles out TSSs as misclassifications, while omitting known insulators, promoters, and other functional non-enhancer sites, and lumping them together as ‘Unknown’. TSSs alone only make up a tiny portion of non-enhancers, and are not truly representative of the real overall misclassifications that a prediction algorithm makes. Furthermore, if enhancers are a subset of DHS, it is safe to assume that the unknown sites are, at the very least, not enhancers of any kind and should be considered invalid as well. They should not be “unknown” from an enhancer prediction point of view because we “know” they are not enhancers. Based on these observations, the RFECS validation method was refined to classify predicted enhancers as either “validated” or “invalid”, based on the following criteria. True Positive markers (TPM) refer to distal DHS sites, p300 or CBP coactivator binding sites, and TFBS that are greater than 1kb away from TSS.

- If a predicted enhancer lies within 2.5kb of a TPM, then the prediction is “validated”. In this case, we know that this site is either a known enhancer, or an unknown

enhancer that we can safely assume to be an enhancer since it overlaps with a DHS site.

- Otherwise, it is “invalid”. This means that it is either a TSS or an Unknown, but we know for a fact it is not an enhancer and hence, the prediction is incorrect.

IV. RESULTS

We trained an initial DNN using all 480 features from the 24 histone modifications, then calculated the importance score for each feature from the learned weights of the DNN according to Equation 1. The features were then sorted by their respective importance scores and different subsets of features were used to train DNNs, starting with the top 10 features, top 20 features, and so on, ending at top 480. The architecture is fixed, with 400 neurons in the first layer, 300 in the second, and 200 in the third. Each DNN was tested against 5 random subsamples of 480K data points each, chosen from among *all* the chromosomes. **Figure 3** shows the features sorted by their importance score in descending order, plotted along with the subsample validation rate of of

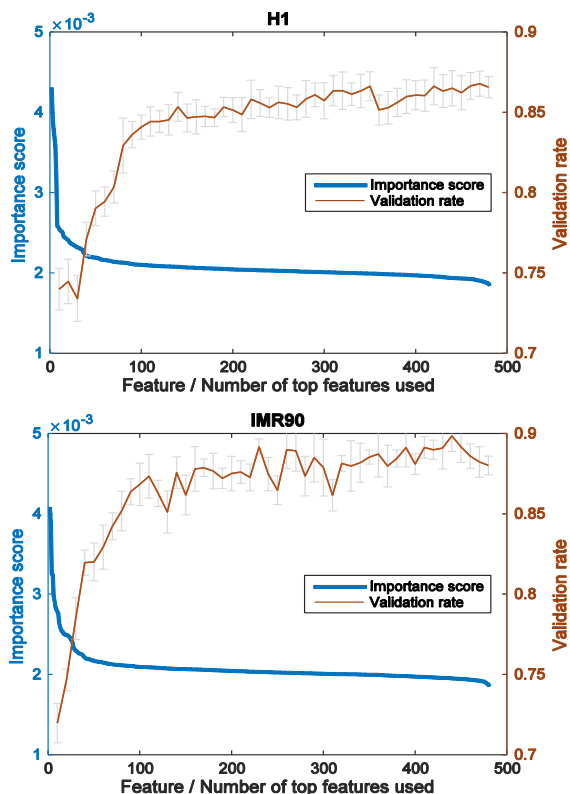


Figure 3: Feature importance score and validation rate when only a subset of features is used, for cell types H1 and IMR90.

each DNN, first for the H1, and next, for the IMR90 cell line.

We can see that validation rate increases sharply as important features are added. However, the rate increases more gradually after the top-150 features are added, for both cell types. This confirms that our weight analysis method

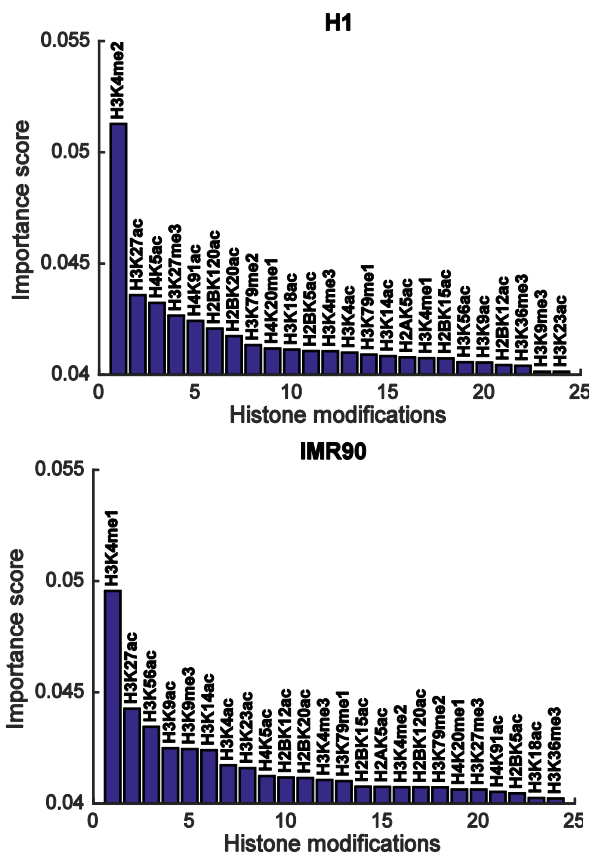


Figure 4: Importance scores of each histone modification, calculated as the sum of its 20 features. The important histone modifications are different for the two cell types H1 and IMR90, albeit with some overlap.

does indeed find the most important features for DNN and allows us to use the reduced subset of 150 features for the final system.

Next, we find the importance of each histone modification by summing up the importance scores of its 20 features. The results are shown in **Figure 4**.

The most important histone modifications we found to confirm previous reports of H3K4me1 (32, 40), H3K27ac (41, 42), and H3K4me2 (16), being the most important ones, in various combinations, overall in global enhancer prediction. However, comparing the histone modification importance of H1 and IMR90 also reveals cell-type specific differences. While H3K4me2 and H3K27ac are the most important histone modifications for H1, for IMR90, H3K4me1 and H3K27ac are the most important. We can see that the important histone modifications are different for the two cell types (with some overlap). This information can help computational scientists when building models to make predictions on specific cell types. Further, it can also help life-science researchers optimize their experiments and collect the features for the most important histone modifications, for the cell type that they are focusing on.

Figure 5 shows importance scores of features within each histone modification. We selected four histone modifications to show the four distinct feature-importance patterns that we observe in the data. We omit the IMR90 results since they were similar to the H1 results presented here. This reveals that the most important features within a histone modification are not always centered at the enhancer site location, and consequently, it is detrimental to use fixed window sizes around the enhancer location, as all prior computational approaches have done. Window sizes that are too small can lead to important features being excluded, while large window sizes will include noise in the data that can be detrimental to prediction accuracy. Furthermore, certain “unimportant” histone modifications do contain relatively important features. This is why omitting histone modifications, altogether, even though they were reported to be unimportant can hurt the classifier’s performance. Thus, analysis at this finer granularity of features, rather than the coarser granularity of histone modifications used in prior approaches, is needed.

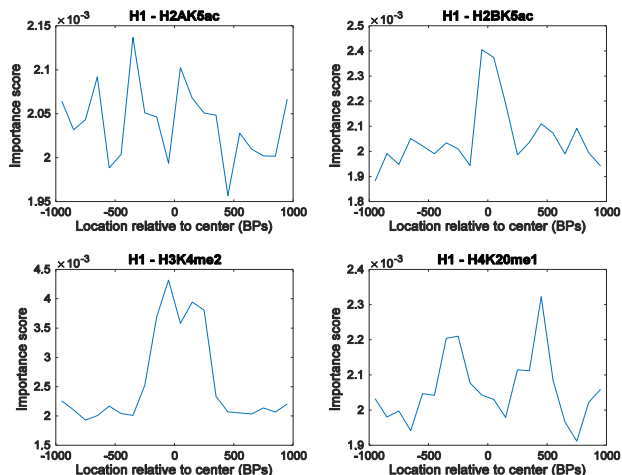


Figure 5: Importance scores for each feature of different histone modifications. The four specific histone modifications shown here represent the different patterns found through our analysis.

Sorting by the feature importance allows us to select only the most important and necessary features for prediction, instead of a fixed window size that has been used with previous methods. This allows us to reduce the number of input features necessary without a significant loss in its actual performance. **Figure 6** shows the comparison between the full 480-feature DNN and the DNN with the selected top-150 features. For generating this figure, the threshold that is used as a comparison point for the output of the DNN is varied. Thus, as the threshold is raised, a fewer number of enhancer predictions are done. First, note that a realistic range for predicting prominent enhancer activity is approximately until 10^5 predictions; beyond that the predictions are too uncertain due to marginal enhancers or sites exhibiting weak enhancer signatures being predicted. Within this operational range, the validation performance with the reduced 150 features is no more than 5% worse than with the full feature set. For much of the operational region, the difference is approximately 2% or less. Thus, we see that

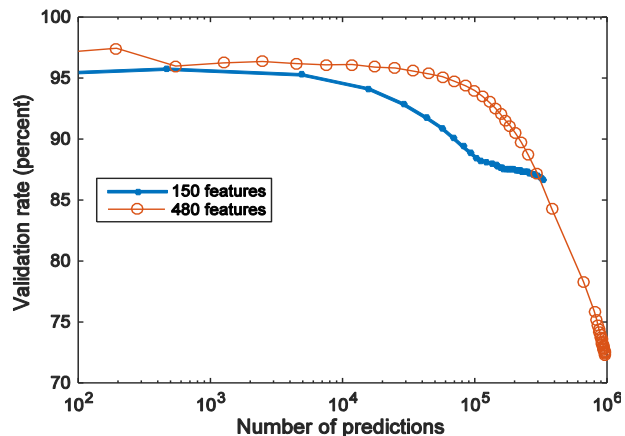


Figure 6: Validation rate comparison between using only the top 150 features and using all 480 features for H1 cell type. We find that for the operational region (less than 10^5 enhancer predictions), the system with the reduced feature set performs comparably to that with the full feature set.

the reduction in the feature space, which reduces the cost of biological experiments to collect the data and the size of input data that a DNN has to be trained and tested with, does not hurt the enhancer prediction performance significantly. The interpretability of the DNN comes as another benefit of our process of reducing the feature set based on the importance scores of the features as calculated by our method.

We compare the runtime of our approach with DEEP and RFECS. The results are shown in **Table 1**. For EP-DNN, the top 150 features were used. We see that for prediction, EP-DNN performs 7.7 times faster than DEEP and 14.8 times faster than RFECS. For training, which is typically less time critical, EP-DNN falls in between RFECS (best) and DEEP (worst).

Table 1: Comparison of training and prediction runtime for EP-DNN, DEEP, and RFECS. Number of predictions is fixed at 40,000.

Method	Training Time (seconds)	Prediction Time (seconds)
EP-DNN	57.90	1.88
DEEP	1,473.88	14.52
RFECS	5.12	27.80

V. CONCLUSION

Enhancers are short DNA sequences that modulate gene expression patterns. Recent studies have shown that enhancer elements could be enriched for certain histone modification combinatorial codes, leading to interest in developing computational models to predict enhancer locations. However, prior attempts had suffered from either low accuracy of prediction or lack of interpretability of the results about which histone modifications are biologically significant. In this paper, we developed a Deep Neural Network based method, called EP-DNN, which addressed

both of these issues. We find validation rates of above 90% for the operational region of enhancer prediction. We developed a method to interpret which histone modifications are important, and within that, which features proximal or distal to the enhancer site, are important. We uncovered that the important histone modifications vary from between cell types. Our results have implications for computational scientists who can now do feature selection for their classification task and for biologists who can now experimentally collect data only for the relevant histone modifications. In ongoing work, we are experimenting with simpler classification schemes, parallelizing our computational approach, and investigating further cell types to uncover possible groupings among cell types.

REFERENCES

- Bulger M, Groudine M. Enhancers: the abundance and function of regulatory sequences beyond promoters. *Developmental biology*. 2010;339(2):250-7.
- Bushey AM, Dorman ER, Corces VG. Chromatin insulators: regulatory mechanisms and epigenetic inheritance. *Molecular cell*. 2008;32(1):1-9.
- Riethoven J-JM. Regulatory regions in DNA: promoters, enhancers, silencers, and insulators. *Computational Biology of Transcription Factor Binding*; Springer; 2010. p. 33-42.
- Noonan JP, McCallion AS. Genomics of long-range regulatory elements. *Annual review of genomics and human genetics*. 2010;11:1-23.
- Banerji J, Rusconi S, Schaffner W. Expression of a β -globin gene is enhanced by remote SV40 DNA sequences. *Cell*. 1981;27(2):299-308. doi: 10.1016/0092-8674(81)90413-x.
- Lettice LA, Heaney SJ, Purdie LA, Li L, de Beer P, Oostra BA, et al. A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Human molecular genetics*. 2003;12(14):1725-35.
- Bulger M, Groudine M. Functional and mechanistic diversity of distal transcription enhancers. *Cell*. 2011;144(3):327-39.
- Ong C-T, Corces VG. Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nature reviews Genetics*. 2011;12(4):283-93. doi: 10.1038/nrg2957. PubMed PMID: PMC3175006.
- Consortium TEP. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012;489(7414):57-74. \
- Romanoski CE, Glass CK, Stunnenberg HG, Wilson L, Almouzni G. Epigenomics: Roadmap for regulation. *Nature*. 2015;518(7539):314-6. doi: 10.1038/518314a.
- Barrett T, Troup DB, Wilhite SE, Ledoux P, Evangelista C, Kim IF, et al. NCBI GEO: archive for functional genomics data sets—10 years on. *Nucleic acids research*. 2011;39(suppl 1):D1005-D10. doi: 10.1093/nar/gkq1184.
- Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, et al. The NIH Roadmap Epigenomics Mapping Consortium. *Nature biotechnology*. 2010;28(10):1045-8. doi: 10.1038/nbt1010-1045. PubMed PMID: PMC3607281.
- Deng L, Hinton G, Kingsbury B, editors. New types of deep neural network learning for speech recognition and related applications: An overview. *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*; 2013: IEEE.
- Hinton G, Li D, Dong Y, Dahl GE, Mohamed A, Jaitly N, et al. Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. *Signal Processing Magazine, IEEE*. 2012;29(6):82-97. doi: 10.1109/MSP.2012.2205597.
- Ngiam J, Coates A, Lahiri A, Prochnow B, Le QV, Ng AY, editors. On optimization methods for deep learning. *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*; 2011.
- Rajagopal N, Xie W, Li Y, Wagner U, Wang W, Stamatoyannopoulos J, et al. RFECS: a random-forest based algorithm for enhancer identification from chromatin state. *PLoS computational biology*. 2013;9(3):e1002968.
- Kleftogiannis D, Kalnis P, Bajic VB. DEEP: a general computational framework for predicting enhancers. *Nucleic acids research*. 2014. doi: 10.1093/nar/gku1058.
- Pennacchio LA, Bickmore W, Dean A, Nobrega MA, Bejerano G. Enhancers: five essential questions. *Nat Rev Genet*. 2013;14(4):288-95.
- Mohrs M, Blankespoor CM, Wang Z-E, Loots GG, Afzal V, Hadeiba H, et al. Deletion of a coordinate regulator of type 2 cytokine expression in mice. *Nature immunology*. 2001;2(9):842-7.
- Levine M, Cattoglio C, Tjian R. Looping back to leap forward: transcription enters a new era. *Cell*. 2014;157(1):13-25.
- Krivega I, Dean A. Enhancer and promoter interactions—long distance calls. *Current Opinion in Genetics & Development*. 2012;22(2):79-85.
- Wang D, Garcia-Bassets I, Benner C, Li W, Su X, Zhou Y, et al. Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. *Nature*. 2011;474(7351):390-4.
- Shlyueva D, Stampfel G, Stark A. Transcriptional enhancers: from properties to genome-wide predictions. *Nat Rev Genet*. 2014;15(4):272-86. doi: 10.1038/nrg3682.
- Visel A, Blow MJ, Li Z, Zhang T, Akiyama JA, Holt A, et al. ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature*. 2009;457(7231):854-8.
- Spitz F, Furlong EE. Transcription factors: from enhancer binding to developmental control. *Nature Reviews Genetics*. 2012;13(9):613-26.
- Yáñez-Cuna JO, Dinh HQ, Kvon EZ, Shlyueva D, Stark A. Uncovering cis-regulatory sequence requirements for context-specific transcription factor binding. *Genome research*. 2012;22(10):2018-30.
- Heintzman ND, Ren B. Finding distal regulatory elements in the human genome. *Current Opinion in Genetics & Development*. 2009;19(6):541-9.
- Berman BP, Pfeiffer BD, Laverty TR, Salzberg SL, Rubin GM, Eisen MB, et al. Computational identification of developmental enhancers: conservation and function of transcription factor binding-site clusters in *Drosophila melanogaster* and *Drosophila pseudoobscura*. *Genome biology*. 2004;5(9):R61.
- Chan HM, La Thangue NB. p300/CBP proteins: HATs for transcriptional bridges and scaffolds. *Journal of cell science*. 2001;114(13):2363-73.
- Korzus E, Torchia J, Rose DW, Xu L, Kurokawa R, McInerney EM, et al. Transcription factor-specific requirements for coactivators and their acetyltransferase functions. *Science*. 1998;279(5351):703-7.
- Boyle AP, Song L, Lee B-K, London D, Keefe D, Birney E, et al. High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. *Genome research*. 2011;21(3):456-64.
- Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet*. 2007;39(3):311-8.
- Won K-J, Chepelev I, Ren B, Wang W. Prediction of regulatory elements in mammalian genomes using chromatin signatures. *BMC bioinformatics*. 2008;9(1):547.
- Firpi HA, Ucar D, Tan K. Discover regulatory DNA elements using chromatin signatures and artificial neural network. *Bioinformatics*. 2010;26(13):1579-86.
- Fernández M, Miranda-Saavedra D. Genome-wide enhancer prediction from epigenetic signatures using genetic algorithm-optimized support vector machines. *Nucleic acids research*. 2012;40(10):e77-e.
- Leray P, Gallinari P. Feature selection with neural networks. *Behaviormetrika*. 1999;26(1):145-66.
- Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, et al. The NIH roadmap epigenomics mapping consortium. *Nature biotechnology*. 2010;28(10):1045-8.
- Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods*. 2008;5(7):621-8.
- Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*. 2014;15(1):1929-58.
- Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, Harp LF, et al. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*. 2009;459(7243):108-12.
- Cotney J, Leng J, Oh S, DeMare LE, Reilly SK, Gerstein MB, et al. Chromatin state signatures associated with tissue-specific gene expression and enhancer activity in the embryonic limb. *Genome research*. 2012;22(6):1069-80.
- Teng L, Firpi HA, Tan K. Enhancers in embryonic stem cells are enriched for transposable elements and genetic variations associated with cancers. *Nucleic acids research*. 2011;39(17):7371-9.