Recursive Convolutional Neural Networks for Epigenomics

Aikaterini Symeonidi*, Anguelos Nicolaou†, Frank Johannes* and Vincent Christlein†
*Technical University of Munich
Liesel-Beckmann-Str. 2, 85354 Freising, Germany,
Email: ksymeonidh@gmail.com and f.johannes@tum.de
†Friedrich-Alexander University Erlangen-Nürnberg
Martensstr. 3, 91058 Erlangen, Germany
Email: anguelos.nicolaou@gmail.com and vincent.christlein@fau.de

***Abstract*—Deep learning methods have proved to be powerful classification tools in the fields of structural and functional genomics. In this paper we introduce a Recursive Convolutional Neural Networks (RCNN) for the anlaysis of epigenomic data. We focus on the task of predicting gene expression from the intensity of histone modifications. The proposed RCNN architecture can be applied to data of an arbitrary size, and has a single meta-parameter that quantifies the models capacity, thus making it flexible for experimenting. The proposed architecture outperforms state-of-the-art systems, while having several orders of magnitude fewer parameters.**

I. Introduction

All somatic cells of eukaryotic organisms contain the same DNA sequence, yet there are vast functional differences between cells across tissue types. It has become increasingly clear that these functional differences are partly determined by epigenetic factors that alter gene expression patterns without changing the underlying DNA sequence [1], [2]. Two key epigenetic mechanisms are DNA methylation [3] and modifications of histone proteins [4], [5], [6]. Histone modifications, are chemical modifications of the histone proteins around which the DNA is wrapped and have an indirect influence on gene expression by changing the accessibility of the DNA [7], [5]. The distribution of these epigenetic marks along the genome acts like a regulating mechanism that defines the functional state of a cell at a given time. As an example, it has been shown that circadian programming of cells, is realised through such epigenomic mechanisms, more specifically H3K27me3 and H3K4me3 regulate core circadian genes[8]. Numerous plant and animal studies have collected epigenomic data from different tissue types, developmental stages, disease states and environmental treatments [9], [10], [1], and have shown that this information is instrumental in elucidating key functional changes during cellular differentiation, disease pathology and for annotating causal mutations from genome-wide association (GWAS) mapping results [9], [11], [12]. To date, more than 100 histone modifications have been described [1], [4], [7], [13]. The large number of marks has led to the notion of a "histone code", a layer of epigenetic information that is encoded by the combinatorial presence/absence patterns of specific histone modifications along the genome [6], [14]. Indeed, molecular studies have shown that histone modifications often act cooperatively in recruiting various chromatin readers and writers as well as transcriptional activators to particular genomic regions [15], [16], indicating that their combinatorial presence is required for proper gene regulation. Recent advances in next-generation sequencing (NGS) technologies now allow scientists to profile gene expression data along with histone modifications and other genomic and epigenomic signals at an unprecedented resolution and scale. The signal obtained from NGS technologies is a discrete signal over each position of the genome. The accumulation of such big datasets in the public domain provides novel opportunities for data mining.

Several computational models have been proposed, using regression models [17], [18], support vector machines [19], random forests [20], or rule-based models [21], for the task of predicting gene expression from histone modifications. However, the above approaches do not capture the combinatorial effect of histone modifications and require a great deal of effort in order to optimize the various model parameters. In recent developments it has been shown that deep learning trumps in performance aforementioned approaches[22], [23].

The principal contributions of this paper are: The introduction of recursive networks for epigenomic data analysis. The restriction of free parameters bellow the cardinality of of the train-set making over-fitting practically impossible, and large-scale cross-domain experimentation demonstrating the proposed models tend to model generic biological phenomena rather than dataset specific correlations.

II. Related Work

A. Deep Epigenomic Analysis

Neural networks have already been used for the task of gene expression prediction from histone modification marks. Singh et. al [22] proposed DeepChrome, a classical Convolutional Neural Network (CNN), with one convolutional layer and two fully connected layers. Dropout was employed to reduce overfitting to the training data. The network consists in total of four layers and approximately 650 thousand free parameters and outperformed other machine learning approaches. Follow on work by the same group, proposed a second architecture, AttentiveChrome[23], a Recurrent Neural Network (RNN)

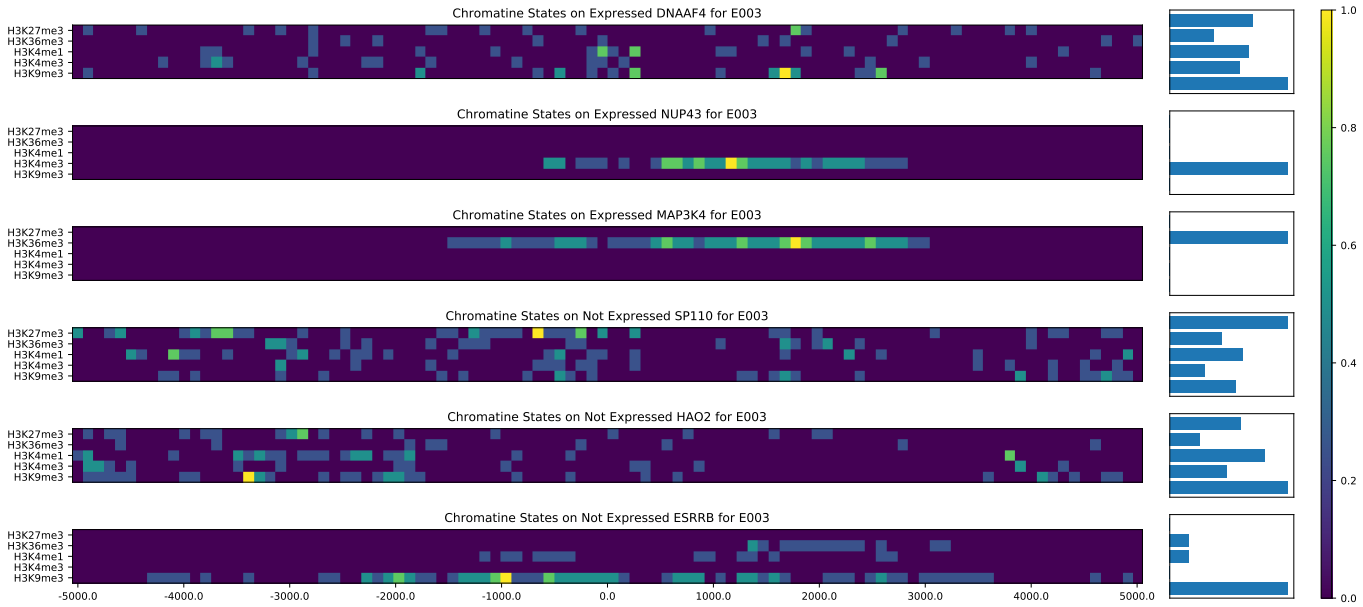

Fig. 1. Chromatine states around the Transcription Start Site (TSS) of expressed and non-expressed genes. Barplots contain the sum per channel.

with two levels of soft attention. This network also consists of four layers and reduces the free parameters to approximately 55 thousands. These models were applied to a collection of 56 human datasets each one containg the histone marks and the gene-expression for a specific cell-line. DeepChrome achieved an average AUC score 80.1%[23] and AttentiveChrome an average AUC score of 81.3% for the same datasets. We consider the experimental process a de-facto standard for comparing models that predict gene expression from histone marks. More details about the employed data and their analysis can be seen in IV-A. In Fig.1 examples of both expressed and not expressed samples as employed in DeepChrome and AttentiveChrome can be seen.

B. Recursive Networks

Recursive Networks and Back-Propagation Through Structure (BPTS) were proposed in 1996 by Goller et al. [24] as a generalization of Back-Propagation Through Time (BPTT). They were introduced as a means of modeling information structured as directed acyclic graphs. The similarities between recursive and recurrent neural networks are very large and have occasionally been confused in older literature, since both have the acronym RNN. Most importantly, they both suffer from vanishing and exploding gradients [25]. In 2011, recursive networks were used for scene and language parsing [26] and achieved state-of-the art performance for those tasks. The model was not directly producing a global tree-structure; the optimal structure was estimated at inference-time with a greedy algorithm. Convolutional variants of recursive networks were first introduced by Socher et.al [27] for 3D object detection. Recursive convolutional networks allow for a fixed tree-structure and thus remove the need for a greedy algorithm at inference-time. Kim et. al [28]. employ a recursive CNN to the problem of image super-resolution. They propose to apply intermediate losses to each recursion layer, similar to the inception network [29] in order to alleviate vanishing/exploding gradients. Additionally, they employ skip connections from the input to the output which is beneficial for super-resolution.

III. Methodology

We model the problem as one of sequence classification similar to other approaches[22], [23]. Specifically, we model the input as a five channel continuous 1D signal, and gene expression as a two-class problem: *"is expressed"* or *"is not expressed"*. The input signal is a quantification of histone modification marks on the genome, which is obtained by mapping snippets of these marks on the genome and counting their amount. Similarly, the expression signal is a quantification of RNA and is obtained by counting snippets on the genes of the genome and normalizing for factors like gene length and depth of sequencing. The genome samples consist of a couple of thousands of genes; depending on the data partitioning, this number can reach ten thousand examples while the input features are orders of magnitude larger. Therefore, one of the main problems is overfitting, which commonly occurs for models using genomic data or other biological data of equivalent cardinalities such as epigenomics. Romero et. al [30] proposed that the ratio of number of inputs to the number of parameters is alarming and must be countered by reducing the number of free model parameters as established regularization techniques such as dropout are probably inefficient. Therefore, they proposed architectures that were primarily designed to drastically reduce the free model parameters. On the other hand approaches such as DeepChrome [22] and AttentiveChrome [23] reduce the inputs by binning them at 100 bases by bin.

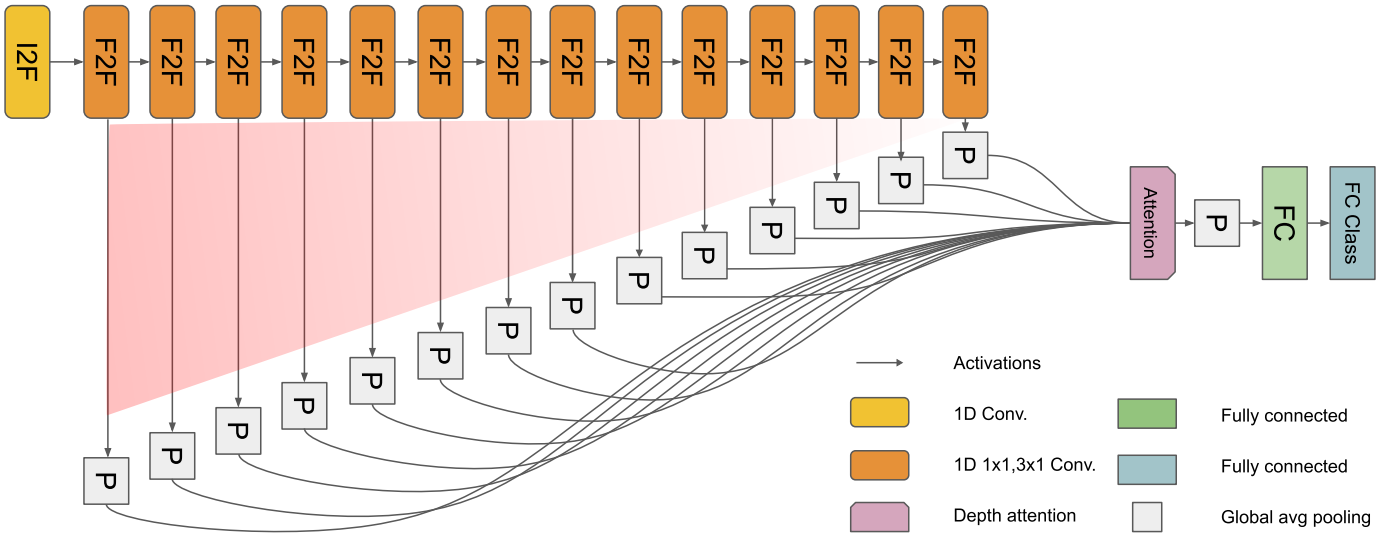

Fig. 2. Model architecture consisting of one convolutional layer followed by a recursive block and a reception-like layer, "feeding" an attention-like mechanism, followed by two fully connected layers . Layers with shared weights have the same color. The activation's spatial resolution is halved after each F2F layer.

Our proposed architecture was designed with the following needs in mind: first, to drastically reduce the parameter of the model; second, to provide a large receptive field; third, to minimize the meta-parameter search; and finally to minimize the introduction of implicit assumptions about biological phenomena.

The proposed architecture consists of three stages similar to the architecture described by Kim et. al [28], see Fig. 2. The first stage consists of a 1x1 convolution, which maps the histone channels to the internal channel representation. In the second stage, a block of layers with N input and N output channels is applied recursively as often as the (theoretical) receptive field we want to achieve. The third stage collates the outputs from each application of the second stage to a multiscale representation, pools across it, and feeds it into a fully connected layer followed by the output layer. N is the only metaparameter of the proposed architecture and can be perceived as the network's capacity. The second stage recursive-block consists of a cascade of an inception-like layer ([29]), a ReLU non-linearity, and a batch normalization ([31]) layer; thus the output-features have the same size as the input-features and follow a normal distribution. The inception-like layer consists of N/2 1x1 convolutions and N/2 3x1 convolutions, padded as needed to ensure the output having the same size as the input. Each application of the recursive-block is followed by a striding of 2 allowing for an exponentially growing receptive field (RF) given by the formula:

$$\begin{aligned} RF(1) &= 1 \\ RF(n) &= RF(n-1) * 2 + 2 \end{aligned} \tag{1}$$

In the proposed architecture, the depth was fixed at a maximum of n=14 so that the receptive field of the ultimate convolutional layer is 24,574. The design that allows an exponentially growing receptive field, is inspired by Wavenet [32]. In Fig. 2 orange boxes represent the successive applications of the recursive block. The output of each application of the recursive-block are pooled into a single feature vector representing a particular scale. In the third stage these vectors are weighted by an attention-like mechanism and pooled into a single vector representing the full signal which is finally classified by the fully connected layers.

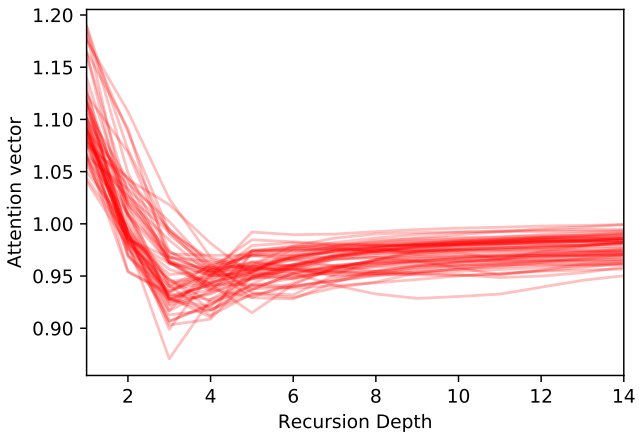

Fig. 3. Attention-like vectors as a function of recursion depth based on the proposed architecture, trained for all 56 human cell lines (each line corresponds to the attention of a given cell line). First application of recursive layer is highly important; all applications converge to 1.

Fig. 3 illustrates that the first application of the recursive layer is emphasized, the following early applications contribute less and then they converge to one. As can be seen in Fig. 3 inference is realised through many alternative paths and the recursive-block can be perceived as the router.

IV. Experiments

We performed experiments in order to compare the proposed networks to the state-of-the-art and also in order to make

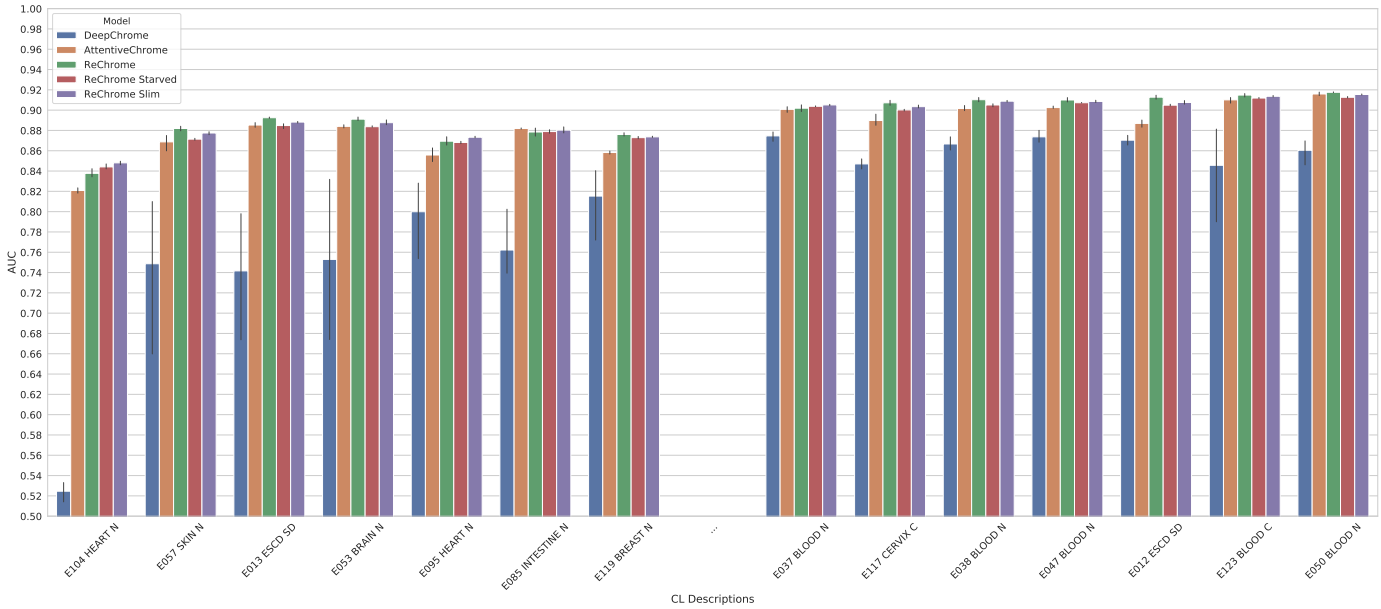

Fig. 4. Performance of the different variants of ReChrome compared with the state-of-the-art, on the 7 cell lines with the lowest performance and the 7 cell lines with the highest. The thin variant is performing better than the slim and starved.

TABLE I
HISTONE MODIFICATION MARKS AND THEIR FUNCTIONALITY.

Histone Modification	Association	Function
H3K27me3	Polycomp repressor	Repressor mark
H3K36me3	Transcribed region	Structural mark
H3K4me1	Enhancer region	Distal mark
H3K4me3	Promoter region	Promoter mark
H3K9me3	Heterochromatin region	Repressor mark

informed decisions during the design of the proposed method. We limited the use of the test-sets to the proposed models and the state-of-the-art models. The validation set was used for the determination of the network architecture and its parameters. The test-set was only used after we had committed to the proposed networks in order to evaluate them. For comparing our proposed architecture with the state-of-the-art, we implemented the network described in DeepChrome ([22]) and trained, validated and tested it with our dataset. In addition, we used the publicly available model and train of AttentiveChrome ([23]) and tested it on our dataset, as described in IV-B.

A. Dataset

In order to obtain uniform data of a large collection of chromatin marks, as well as quantification data of RNA expression in a variety of different samples, we used the REMC database ([9]). From the REMC database we obtained data for five core histone modification marks (H3K4me1/3, H3K36me3, H3K27me3 and H3K9me3) in 56 cell lines and the corresponding normalized RPKM ([33]) expression data for these cell lines. We selected these histone marks not only because of their availability, but also because they are implicated in the gene expression regulation, acting either as promoters, suppressors, or mediators ([4], [34], [35], [7]), see Table I. To place the genes on the hg19 genome we used an updated annotation of the genes, the Ensembl GRCh37 version 87 ([36]), in order to have the data reflect a current state of gene annotation, removing the expression data for the genes that were not included in the version 87 annotation. During this process approximately 100 genes were removed from the initial dataset. For each cell line, we partitioned 19796 genes as listed by [9] of the human genome into three equal sets for training, validation, and testing process, making sure that the three sets were completely isolated from each other. Genomic sequences that overlapped, either because they are located on opposite strands or overlap within the same genomic context, i.e. area around Transcription Start Site (TSS), were placed in the same set, in order to avoid samples in different partitions being near. We specifically guarantied a distance of 100,000 bases between any pair of samples coming from different partitions ([37]). The test-set was only used for comparing our model to the state of the art, while we used the validation set for any other experiment, tuning etc. For the labeling of the genes in the two categories of expressed or not-expressed, in each cell line, the gene expression values were discretized using the median normalized expression value of all genes in the cell line, in accordance with ([22], [23]). Genomic elements, such as genes and histone modifications, have a set of coordinates containing the chromosome on which the element is located and the starting and ending coordinates of the element on the given chromosome. During pre-processing, genomic coordinates are transformed into a uniform continuous space, transferring the information of the location into a two-element structure.

This strategy can also be applied to different organisms with varying number of chromosomes and varying chromosome lengths. For each gene in the dataset, we focused on a genome context of maximum 15,000 bases upstream and downstream of the Transcription Start Site (TSS), thus remaining within the range of reported short range interactions between histone modifications and gene expression ([38], [39]). Within our implementation the histone signal is calculated in bins of size 1 base or higher, without restrictions on the size of the bins. In accordance to previous studies ([22], [23]), we made extensive experiments using a genome context of 5 000 bases upstream and downstream of the TSS and a bin size of 100 bases, although we also made experiments with higher genome context and variable bin sizes.

B. Baselines

a) DeepChrome: This classical CNN architecture is designed specifically for the task of chromatine sequence classification ([22]). For the experiments, we used our implementation of the exact model described in DeepChrome ([22]) and trained it with the given training parameters, ie a learning rate of 0.001, for 100 epochs, dropout of 0.5 was applied on the convolutional layer. DeepChrome receives inputs of 100 values for 5 channels. Each input value has the average histone signal for 100 bases so it covers 5 000 bases upstream and 5 000 bases downstream from the TSS. As the model has no global-pooling it can only operate on inputs of a fixed length of 100 bins.

b) AttentiveChrome: A Recurent Neural Network with two attention mechanisms that was proposed as a means of improving interpretabillity through the attention mechanism. Performance-wise it is inconclusive whether AttentiveChrome or DeepChrome performs better. We incorporated the publicly available implementation into our experimental pipeline; specifically we used the model definition and the training source code as they were and we simply adapted our data-loading and evaluation code to use the model. We used the exact training protocol of the public code. For both DeepChrome and AttentiveChrome, the validation and test set AUC reported is the best achieved during training. Although this is favorable to larger models that suffer from over-fitting, it can be argued that it is equivalent to early stopping. It can be inferred from the public source code that this is the employed evaluation protocol. The inputs to the network are the same as in DeepChrome.

C. Thin, Slim, and Starved networks

The proposed architecture is designed with the goal of limiting the meta-parameter search. We did not observe any effect on performance when the recursion depth increased. We used a depth of 14 as we wanted to achieve a receptive field greater than 10 000 base-pairs. The only meta-parameter the proposed model has is the channel-count N which functions as a quantification of the model's capacity. The parameter count of the proposed model is $iN + 3N^2 + 5N + 2$ where i is the number of input channels, which is equivalent to $O(N^2)$ as i can be considered a constant. In our data it was 5. Three variants of the network are proposed: a thin variant that has 100 channels resulting in 31 016 parameters, a slim variant having 30 channels resulting in 3 076 parameters, and a starved variant that has only 10 channels resulting in 416 parameters. The fact that our networks only vary in capacity, allows for experiments on the capacity-performance trade-off which is a way to quantify the hardness of the problem.

D. Comparison with the state-of-the-art

The proposed ReChrome network performs better than both DeepChrome and AttentiveChrome over the test-set as can be seen in table II. For every cell line of the 56 we used in experiments, networks were trained four times on the train genes, used the validation genes for early stopping and evaluated the ROC-AUC. This allows to have an estimate on the significance of performance measurements for each cell line independently. In Fig. 4 each model's performance in each cell line can be seen; the standard deviation for the four replicates is marked as a black error range in the bars. ReChrome variants demonstrate smaller deviation than AttentiveChrome, while DeepChrome demonstrates to be more volatile. As there are 56 cell lines, each one being a dataset for gene expression in a different biological context, statistical testing can allow for safe estimates on the performance of the proposed ReChrome and the state-of-the-art (DeepChrome and AttentiveChrome). All ReChrome variants, ReChrome, ReChrome-Slim, and ReChrome-Starved outperform DeepChrome with p-values 1.63E-45, 1.43E-46 and 3.12E-42 respectively. ReChrome and ReChrome-Slim outperform AttentiveChrome as well, with p-values 2.67E-05 and 8.83E-06 respectively. AttentiveChrome outperforms ReChrome-Starved variant, although not significantly (p-value 0.007). While DeepChrome and AttentiveChrome were reported to have similar performance by their authors, we find that AttentiveChrome is more consistent, robust, and less susceptible to over-fitting. The early-stopping algorithm we used, was to train each model for 20 epochs and keep the model at the point were it reached it's top performance on the validation set. While this algorithm is good for well regularised models such as ReChrome and AttentiveChrome, if a model is not regularised such as DeepChrome, this algorithm might lead to overfitting the validation-set.

E. Cross Dataset Generalization

In order to assess how well the networks generalize, we conducted experiments where each network is trained in one cell line and then tested in all remaining 55 cell-lines. This experiment provides intuitions on the importance of biological context for determining the relationship between chromatin marks and gene-expression. As can be seen in the last column of Table II, on average performance of Rechrome and ReChrome-Slim trained in one cell-line and tested on an other only dropped by 1.6% and 1.2% respectively while ReChrome-Starved had no observable performance drop at all. These findings validate our hypothesis that a drastic reduction in the neural networks capacity, will force the models to learn more

Fig. 5. Gradient ascent visualization of the proposed ReChrome and DeepChrome models on both expressed and non-expressed samples of E003 (Embryonic Stem Cells).

TABLE II
AVERAGE AUC OF THE DIFFERENT MODELS OVER VALIDATION, TEST AND CROSS-TEST SET.

Model	Parameters	Val. set	Test set	Cross test set
DeepChrome	644,177	87.36%	81.43%	79.78
AttentiveChrome	55,681	86.36%	86.8%	NA
ReChrome	31,016	**87.54**%	87.73%	86.13%
ReChrome Slim	3,076	87.06%	**87.75**%	**86.56**%
ReChrome Starved	416	86.55%	86.45%	86.45%

TABLE III
AVERAGE AUC OF THE DIFFERENT BIN-SIZES AND BIN-COUNT ON VALIDATION.

Model	bin size	bin count	TSS context (bases)	AUC %
ReChrome	1	30,000	$\pm 15,000$	85.35
ReChrome	100	100	$\pm 5,000$	87.54
ReChrome	150	66	$\pm 4,950$	87.64
ReChrome	150	200	$\pm 15,000$	87.63
ReChrome	300	34	$\pm 10,200$	88.05
ReChrome	300	100	$\pm 15,000$	88.07
ReChrome	400	26	$\pm 5,200$	**88.14**
ReChrome	400	76	$\pm 15,200$	88.09

generic associations. ReChrome-Starved, has 416 parameters and was trained on 6598 samples of size 5×100, and manages to generalise perfectly between different biological contexts.

F. Versatility

A benefit of the recursive architecture of ReChrome, is that it allows the exact same model to be employed on data of arbitrary size. We performed several experiments by altering the sampling parameters of ReChrome's input. In Table II all networks had 100 inputs per channel each input bin had the average of 100 bases proving a context of $\pm$ 5,000 bases around the TSS of each gene. In Table III the average performance of ReChrome on the validation-set of each cell-line with alternative sampling parameters can be seen. We tried bin sizes of 1, 100, 150, 300, and 400 bases for contexts of approximately 10,000 and 30,000 bases. It can be seen that on several occasions ReChrome benefits and even improves it's performance on validation set compared to the standard sampling parameters.

G. Visualization

In order to gain meaningful visualizations of what ReChrome actually learns, we employed gradient ascent to construct optimal synthetic inputs that maximize the network's confidence for both expressed and non-expressed classes. In Fig. 5 we can see optimal inputs for both classes for the DeepChrome baseline and ReChrome. The gradient ascent was obtained with a learning rate of 0.01, weight decay coefficient λ 0.001 and 600 iterations. It is worth remarking that the proposed ReChrome has observed very clearly the relationship of the existence and absence of chromatin marks and gene expression. Presence of structural, distal and promoter marks, and absence of repressor marks is found to be the optimal input for expressed genes, while the exact opposite situation is found to be the optimal input for non-expressed genes. ReChrome is in agreement with known functional relationships between gene expression and specific histone modifications Table I, while it seems DeepChrome focuses on complicated interactions at the expense of the dominant linear ones. This observation demonstrates that ReChrome, while sensitive to non-linear variable interactions that allow it to outperform the state-of-the-art, tends to produce classifications that are also consistent with the fact that H3K4me1, H3K4me3 and H3K36me3 are known activators and H3K27me3 and H3K9me3 known supressors of gene expression. [40], [4], [7], [38], [14], [6], [41].

V. CONCLUSIONS

A. Findings

In this work, we proposed a recursive convolutional neural network that drastically reduces the number of parameters and meta-parameters and at the same time, manages a small but significant improvement over the state-of-the-art. The proposed architecture is well regularised because of its constrained capacity and proved to be impervious to over-fitting

of any kind. We consider that the value of predicting gene expression from histone marks does not lie in using machine learning to reduce the cost of analysis but rather, on providing a valid quantification and a lower bound of the interaction between these two phenomena.

The proposed architecture dealt very effectively with the presented data and proved to be highly stable, one might argue that this is because the proposed model can specifically discover the suppressor or promoter roles each histone mark has as seen in Fig. 5. We should point out that established linear and non-linear methods can not cope as well as deep learning [22] with this kind of data. Yet the scope of domains to which ReChrome can successfully be employed remains a question.

B. Discussion

While the proposed model outperformed significantly the state-of-the-art; the error rate of all alternative architectures, is very large for a binary classifier and there might be a limit to the performance machine learning techniques can achieve with the specific data used in these experiments. Apart from the histone modifications that we have explored, several others exist that may also have an impact on gene expression. The addition of more modifications could also improve the prediction of gene expression. DNA methylation ([42]) is another layer of epigenetic regulation, which also impacts gene expression. It gives an additional layer of information and another channel that could be used as input to our model. We used a genomic context of 10,000 bases around each gene's TSS; but there are histone marks that are modulating gene enhancers, some of which might be even millions of bases further away from the genes that they regulate [43], [40], [44]. In fact, a gene can have an accumulation of active marks and still be non-active, because the long range enhancer that participates in the regulation of the gene is not in an active state [45].

While these enhancers are far away from the genes if we regard the DNA as a 1D structure, this does not reflect the reality, because the DNA has a 3D structure [46], specifically it is a thread folded into knots several times over; even-more the knots fold and unfold dynamically and present day acquisition protocols average over different such states. Nonetheless, the addition of data regarding the structure of the DNA, such as Hi-C data [47], would be valuable.

C. Future Work

The work presented in this paper, was principally aimed at establishing that ReChrome is a reliable tool for analysis of epigenomic data across different configurations. Our experiments were principally aimed at comparing fairly ReChrome to the state-of-the-art instead of aiming at biologically meaningful insights. The binary expressed, not expressed classes obtained by thresholding expression values to above and bellow median, leaves much to be desired. Three classes might make more sense as under-expressed, normal, over-expressed genes. These classes should not be defined by ranked order statistics as this forces assumptions on the population of gene expression. Although experiments in this paper demonstrated that ReChrome works and generalises across several different cell-types of humans, experiments for other species including plants and other mammals are planned.

Acknowledgment

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan V GPU used for this research.

References

[1] B. E. Bernstein, A. Meissner, and E. S. Lander, "The mammalian epigenome," *Cell*, vol. 128, no. 4, pp. 669–681, 2007.

[2] W. Reik, "Stability and flexibility of epigenetic gene regulation in mammalian development," *Nature*, vol. 447, no. 7143, pp. 425–432, 2007.

[3] J. A. Law and S. E. Jacobsen, "Establishing, maintaining and modifying dna methylation patterns in plants and animals," *Nature Reviews Genetics*, vol. 11, no. 3, pp. 204–220, 2010.

[4] S. L. Berger, "The complex language of chromatin regulation during transcription," *Nature*, vol. 447, no. 7143, p. 407, 2007.

[5] A. J. Bannister and T. Kouzarides, "Regulation of chromatin by histone modifications," *Cell research*, vol. 21, no. 3, p. 381, 2011.

[6] T. Jenuwein and C. D. Allis, "Translating the histone code," *Science*, vol. 293, no. 5532, pp. 1074–1080, 2001.

[7] T. Kouzarides, "Chromatin modifications and their function," *Cell*, vol. 128, no. 4, pp. 693–705, 2007.

[8] J. A. Ripperger and M. Merrow, "Perfect timing: epigenetic regulation of the circadian clock," *FEBS letters*, vol. 585, no. 10, pp. 1406–1411, 2011.

[9] A. Kundaje, W. Meuleman, J. Ernst, M. Bilenky, A. Yen, A. Heravi-Moussavi, P. Kheradpour, Z. Zhang, J. Wang, M. J. Ziller *et al.*, "Integrative analysis of 111 reference human epigenomes," *Nature*, vol. 518, no. 7539, p. 317, 2015.

[10] D. Adams, L. Altucci, S. E. Antonarakis, J. Ballesteros, S. Beck, A. Bird, C. Bock, B. Boehm, E. Campo, A. Caricasole *et al.*, "Blueprint to decode the epigenetic signature written in blood," *Nature biotechnology*, vol. 30, no. 3, pp. 224–226, 2012.

[11] B. E. Bernstein, J. A. Stamatoyannopoulos, J. F. Costello, B. Ren, A. Milosavljevic, A. Meissner, M. Kellis, M. A. Marra, A. L. Beaudet, J. R. Ecker *et al.*, "The nih roadmap epigenomics mapping consortium," *Nature biotechnology*, vol. 28, no. 10, pp. 1045–1048, 2010.

[12] M. Kasowski, S. Kyriazopoulou-Panagiotopoulou, F. Grubert, J. B. Zaugg, A. Kundaje, Y. Liu, A. P. Boyle, Q. C. Zhang, F. Zakharia, D. V. Spacek *et al.*, "Extensive variation in chromatin states across humans," *Science*, vol. 342, no. 6159, pp. 750–752, 2013.

[13] V. W. Zhou, A. Goren, and B. E. Bernstein, "Charting histone modifications and the functional organization of mammalian genomes," *Nature Reviews Genetics*, vol. 12, no. 1, pp. 7–18, 2011.

[14] B. D. Strahl and C. D. Allis, "The language of covalent histone modifications," *Nature*, vol. 403, no. 6765, p. 41, 2000.

[15] I. O. Torres and D. G. Fujimori, "Functional coupling between writers, erasers and readers of histone and dna methylation," *Current opinion in structural biology*, vol. 35, pp. 68–75, 2015.

[16] E. T. Wiles and E. U. Selker, "H3k27 methylation: a promiscuous repressive chromatin mark," *Current opinion in genetics & development*, vol. 43, pp. 31–37, 2017.

[17] R. Karlić, H.-R. Chung, J. Lasserre, K. Vlahoviček, and M. Vingron, "Histone modification levels are predictive for gene expression," *Proceedings of the National Academy of Sciences*, vol. 107, no. 7, pp. 2926–2931, 2010.

[18] I. G. Costa, H. G. Roider, T. G. do Rego, and F. d. A. de Carvalho, "Predicting gene expression in t cell differentiation from histone modifications and transcription factor binding affinities by linear mixture models," *BMC bioinformatics*, vol. 12, no. 1, p. S29, 2011.

[19] C. Cheng, K.-K. Yan, K. Y. Yip, J. Rozowsky, R. Alexander, C. Shou, and M. Gerstein, "A statistical framework for modeling gene expression using chromatin features and application to modencode datasets," *Genome biology*, vol. 12, no. 2, p. R15, 2011.

[20] X. Dong, M. C. Greven, A. Kundaje, S. Djebali, J. B. Brown, C. Cheng, T. R. Gingeras, M. Gerstein, R. Guigó, E. Birney *et al.*, "Modeling gene expression using chromatin features in various cellular contexts," *Genome biology*, vol. 13, no. 9, p. R53, 2012.

[21] B. H. Ho, R. M. K. Hassen, and N. T. Le, "Combinatorial roles of dna methylation and histone modifications on gene expression," in *The National Foundation for Science and Technology Development (NAFOSTED) Conference on Information and Computer Science*. Springer, 2014, pp. 123–135.

[22] R. Singh, J. Lanchantin, G. Robins, and Y. Qi, "Deepchrome: deep-learning for predicting gene expression from histone modifications," *Bioinformatics*, vol. 32, no. 17, pp. i639–i648, 2016.

[23] R. Singh, J. Lanchantin, A. Sekhon, and Y. Qi, "Attend and predict: Understanding gene regulation by selective attention on chromatin," in *Advances in neural information processing systems*, 2017, pp. 6785–6795.

[24] C. Goller and A. Kuchler, "Learning task-dependent distributed representations by backpropagation through structure," in *Proceedings of International Conference on Neural Networks (ICNN'96)*, vol. 1. IEEE, 1996, pp. 347–352.

[25] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *International conference on machine learning*, 2013, pp. 1310–1318.

[26] R. Socher, C. C. Lin, C. Manning, and A. Y. Ng, "Parsing natural scenes and natural language with recursive neural networks," in *Proceedings of the 28th international conference on machine learning (ICML-11)*, 2011, pp. 129–136.

[27] R. Socher, B. Huval, B. Bath, C. D. Manning, and A. Y. Ng, "Convolutional-recursive deep learning for 3d object classification," in *Advances in neural information processing systems*, 2012, pp. 656–664.

[28] J. Kim, J. Kwon Lee, and K. Mu Lee, "Deeply-recursive convolutional network for image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1637–1645.

[29] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[30] A. Romero, P. L. Carrier, A. Erraqabi, T. Sylvain, A. Auvolat, E. Dejoie, M.-A. Legault, M.-P. Dubé, J. G. Hussin, and Y. Bengio, "Diet networks: thin parameters for fat genomics," *arXiv preprint arXiv:1611.09340*, 2016.

[31] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, F. Bach and D. Blei, Eds., vol. 37. Lille, France: PMLR, 07–09 Jul 2015, pp. 448–456.

[32] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[33] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, "Mapping and quantifying mammalian transcriptomes by rna-seq," *Nature methods*, vol. 5, no. 7, p. 621, 2008.

[34] A. Eberharter and P. B. Becker, "Histone acetylation: a switch between repressive and permissive chromatin: Second in review series on chromatin dynamics," *EMBO reports*, vol. 3, no. 3, pp. 224–229, 2002.

[35] G. J. Filion, J. G. van Bemmel, U. Braunschweig, W. Talhout, J. Kind, L. D. Ward, W. Brugman, I. J. de Castro, R. M. Kerkhoven, H. J. Bussemaker *et al.*, "Systematic protein location mapping reveals five principal chromatin types in drosophila cells," *Cell*, vol. 143, no. 2, pp. 212–224, 2010.

[36] B. L. Aken, C. García Girón, D. N. Murphy, F. J. Martin, K. Billis, P. Flicek, R. Nag, T. Hourlier, D. Barrell, M. Schuster, F. Kokocinski, S. M. J. Searle, V. Curwen, S. White, S. Ayling, A. Zadissa, K. Howe, L. Clarke, M. Ruffier, S. Fairley, Y. A. Tang, J. Fernandez Banet, A. Kähäri, and J.-H. Vogel, "The Ensembl gene annotation system," *Database*, vol. 2016, 06 2016.

[37] F. Cao and M. J. Fullwood, "Inflated performance measures in enhancer–promoter interaction-prediction methods," vol. 51, no. 8, pp. 1196–1198. [Online]. Available: https://www.nature.com/articles/s41588-019-0434-7

[38] C. M. Koch, R. M. Andrews, P. Flicek, S. C. Dillon, U. Karaöz, G. K. Clelland, S. Wilcox, D. M. Beare, J. C. Fowler, P. Couttet *et al.*, "The landscape of histone modifications across 1% of the human genome in five human cell lines," *Genome research*, vol. 17, no. 6, pp. 691–707, 2007.

[39] A. Sekhon, R. Singh, and Y. Qi, "Deepdiff: Deep-learning for predicting differential gene expression from histone modifications," *Bioinformatics*, vol. 34, no. 17, pp. i891–i900, 2018.

[40] J. Yan, S.-A. A. Chen, A. Local, T. Liu, Y. Qiu, K. M. Dorighi, S. Preissl, C. M. Rivera, C. Wang, Z. Ye *et al.*, "Histone h3 lysine 4 monomethylation modulates long-range chromatin interactions at enhancers," *Cell research*, vol. 28, no. 2, p. 204, 2018.

[41] P. Chi, C. D. Allis, and G. G. Wang, "Covalent histone modifications—miswritten, misinterpreted and mis-erased in human cancers," *Nature reviews cancer*, vol. 10, no. 7, p. 457, 2010.

[42] P. A. Jones, "Functions of dna methylation: islands, start sites, gene bodies and beyond," *Nature Reviews Genetics*, vol. 13, no. 7, p. 484, 2012.

[43] I. Chepelev, G. Wei, D. Wangsa, Q. Tang, and K. Zhao, "Characterization of genome-wide enhancer-promoter interactions reveals co-expression of interacting genes and modes of higher order chromatin organization," *Cell research*, vol. 22, no. 3, p. 490, 2012.

[44] J. Dekker and T. Misteli, "Long-range chromatin interactions," *Cold Spring Harbor perspectives in biology*, vol. 7, no. 10, p. a019356, 2015.

[45] M. P. Creyghton, A. W. Cheng, G. G. Welstead, T. Kooistra, B. W. Carey, E. J. Steine, J. Hanna, M. A. Lodato, G. M. Frampton, P. A. Sharp *et al.*, "Histone h3k27ac separates active from poised enhancers and predicts developmental state," *Proceedings of the National Academy of Sciences*, vol. 107, no. 50, pp. 21 931–21 936, 2010.

[46] L. J. Alberts B, Johnson A, *Molecular Biology of the Cell. 4th edition*. Garland Science, 2002.

[47] M. Forcato, C. Nicoletti, K. Pal, C. M. Livi, F. Ferrari, and S. Bicciato, "Comparison of computational methods for hi-c data analysis," *Nature methods*, vol. 14, no. 7, p. 679, 2017.