Convolutional Additive Models: a fully interpretable approach to Deep Learning in Genomics

Manu Saraswat[1,†], Gherman Novakovsky[1,†], Etienne Meunier[2], Oriol Fornes[1], Sara Mostafavi[3✉], Wyeth W. Wasserman[1,✉]

[1]Centre for Molecular Medicine and Therapeutics, Department of Medical Genetics, BC Children's Hospital Research Institute, University of British Columbia, Vancouver, BC V5Z 4H4, Canada

[2]INRIA, Centre Rennes - Bretagne Atlantique, France

[3]Paul G. Allen School of Computer Science and Engineering, University of Washington (UW), Seattle, USA

[†]MS and GN contributed equally to this work

[✉]Correspondence to SM (saramos@cs.wshington.edu) and WWW (wyeth@cmmt.ubc.ca)

Introduction:

Convolutional neural networks (CNNs) and hybrid architectures with recurrent neural networks (RNNs) are highly effective deep learning tools in genomics. They have successfully been applied to predict transcription factor binding sites (TFBSs) [1–4], chromatin accessibility [5,6], and the functional impact of non-coding variants [3,7].

Deep learning methods have increased predictive accuracy however they lack interpretability. There are two main approaches currently used to tackle the problem of poor interpretability for deep neural networks: attribution based and visualisation of filters. Attribution based approaches quantify the importance of individual base pairs for each input sequence [8]. The importance scores are clustered into motifs which can be compared with binding motifs of known transcription factors (TFs) to gain insights into the biological interplay behind the task at hand [9,10]. Unfortunately, these approaches do not quantify importance on a global level [11,12] and can sometimes fail to provide reliable importance scores [13,14]. The second approach to interpretability involves visualising filters/kernels in the first layer of the convolutional network as position weight matrices(PWMs). This is achieved by aligning the sequences activated by those filters [5,6]. However, the efficacy of these approaches is dependent on architecture choices [15]. In order to gain interpretability at a global level, filters must be nullified sequentially, which is both computationally intensive and dependent on arbitrary thresholds. Furthermore, these approaches may miss out on the importance of features assembled in the later convolutional layers.

Recently, Agarwal et al. proposed Neural Additive models (NAMs)[16] as an interpretable alternative to feedforward neural networks. NAMs are a form of generalised linear model, where the prediction is computed as a linear combination of multiple feedforward neural network outputs, each of which is attended by a particular input (Figure 1A).

Mathematically:

$$Y = \sum(\omega_i \times f_i(x_i)) + \beta$$

Where f_i is the neural network associated with input x_i .

NAMs have been shown to have similar performance to feedforward neural networks and gradient boosted trees for tabular input data.

In this paper, we present Convolutional Additive Models (CAMs). CAMs are a modification to NAMs optimised for genomics tasks trained on one-hot encoded sequences. The prediction is computed as a linear combination of outputs from multiple independent CNNs, each of which consists of a single kernel and two fully connected layers (Figure 1C). We argue that if we are able to assign a biological interpretation to each of these convolutional units (for example, by comparing the filter activations with known TF motifs), we will obtain a global view of how predictions are being made for the task at hand by visualizing the weights of the output layer; thereby overcoming the limitations of the existing interpretation methods. We demonstrate that CAMs achieve a performance comparable to that of Multi-layer CNNs in predicting TF binding from DNA sequence data. We use Genetic Algorithm (GA) [17] to generate artificial sequences that maximise each independent CNN unit to overcome the limitations associated with filter to PWM conversion. Further, by combining the output layer weights with the motifs learned by each independent CNN unit, we recovered patterns of cooperativity between TFs. Lastly, we trained CAMs to predict chromatin accessibility in 81 immune cell types in mouse using DNA sequence data and recovered the role of known TFs and their cooperative action in driving lineage specification. We show that CAMs provide an easy to use, fully interpretable approach to deep learning in genomics without sacrificing accuracy.

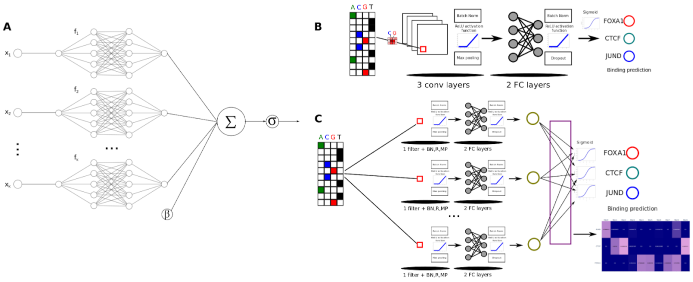

Figure 1 Model architectures used in this study

A. Neural Additive Model. Each input x_i is fed into its corresponding fully-connected neural network f_i. The final output is computed as a sum of outputs from each fully-connected network.

B. The CNN architecture used in this work as a baseline is inspired by Basset and AI-TAC. Specifically, three convolutional layers, each followed by batch normalization, ReLU activation function, and max pooling, followed by two fully-connected layers and one output layer.

C. The architecture of the proposed Convolutional Additive Model (CAM). Input sequences are operated on by independent

CNNs, each of which consists of one convolutional filter of length 20, followed by batch normalization (BN), ReLU activation (R), max pooling (MP), and two fully-connected layers of 100 and 1 neuron(s) respectively. The final output is computed as the weighted sum of outputs from each independent CNN unit. These weights (purple rectangle) can be visualized to provide global interpretability of the model.

Results

CAMs accurately predict TF binding

As proof of concept, we implemented a CAM with 10 independent CNN units to predict the binding to DNA of CTCF, JUND, and FOXA1. To facilitate model interpretation, we applied the ReLU activation function after each independent CNN unit and restricted the weights from the final layer to be non-negative. As a baseline comparison, we trained a multilayer CNN-based model following specifications from [5] and [6], as shown in Figure 1B. Both models were trained on a custom dataset combining TF binding data from ReMap 2018[18] and UniBind[19] with ENCODE DNase I hypersensitive sites [20,21] in a cell and tissue type agnostic manner, as described here[31]**.** Overall, the CAM and baseline models achieved comparable performance levels (Figure 2A). We observed that the CAM performance improved with the number of CNN units used (Figure 2B). However, the training time also increased linearly with the number of CNN units (Figure 2C).

To interpret each independent CNN unit, we converted the convolutional filters of the first layer to PWMs and compared them to TF DNA-binding profiles from the JASPAR database [22] using Tomtom [23], as described in [5]. As expected, most filter PWMs had significant similarities to the profiles of CTCF, JUND or FOXA1 (Figure 2D; left). Next, we visualised the weights from the final layer of the CAM to highlight the contribution of each independent CNN unit (and thereby its corresponding TF motif) to the prediction of CTCF, JUND and FOXA1 (Figure 2D; right). This provides a global picture of how predictions are made for each output without performing the computationally intensive exercise of iteratively nullifying each filter. To achieve interpretability at the level of individual sequences (hereafter referred to as local interpretability), we visualised the weighted outputs from each independent CNN unit (Figure 2E). This provides an understanding of how predictions are made for each sequence while highlighting which features are important.

A potential solution to overcome the increasing training time is the use of larger batch sizes. This however could lead to a decrease in accuracy and is a subject of further study.

Overall, we demonstrate that CAM approach achieves global and local interpretability without compromising predictive accuracy.

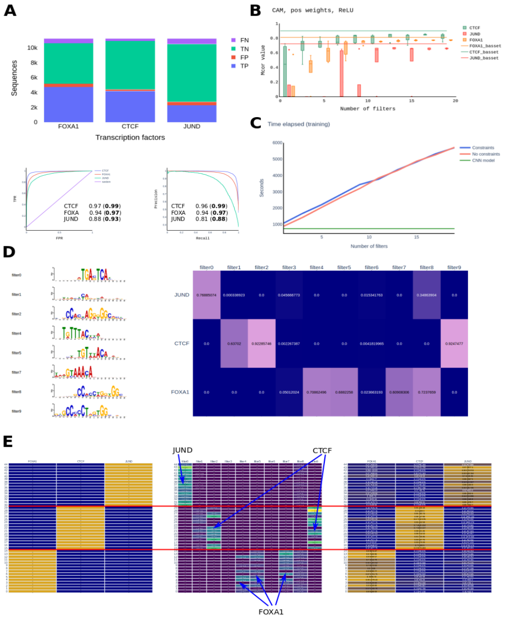

Figure 2. CAMs achieve global and local interpretability without comprising on accuracy

A. Performance of the CAM model on TF binding prediction task. Confusion matrix and Precision Recall/Receiver Operating Characteristics curve for CTCF, JUND and FOXA1. AUROC/AUPRC for the baseline CNN model are shown in bold for comparison.

B. CAM achieves similar performance (measured by Matthews Correlation Coefficient) to the baseline model as the number of CNNs increases.

C. Time required for training a CAM scales linearly with the number of independent CNN units. Constraints denote non-negative weight restrictions in the final layer.

D. Global interpretability with CAM. On the left: filters from independent CNN units correspond to binding motifs of CTCF, JUND and FOXA1, except filter one.

On the right: visualisation of final layer weights of CAM highlights the importance of different CNN units for prediction of each TF.

E. Local Interpretability with CAM.

On the left: True labels for a selected set of sequences.

In the middle: Weighted output from each CNN unit for the selected set of sequences. On the right: CAM prediction computed by sigmoid operation on sum of weighted outputs from each CNN unit.

Artificial sequences generated using Genetic Algorithms provide an alternative to filter-PWM conversion

As shown in the section above, not all filter PWMs corresponded to a known TF motif. For instance, filter one in Figure 2A did not match any JASPAR profile, but it was important for the prediction of CTCF binding. On the other hand, filter eight in Figure 2A resembled the motif of CTCF, yet it was not necessary for its prediction. We hypothesized that since each CAM filter was followed by two fully-connected layers, focusing only on the motifs activated by the filter might not be enough to visualize the important features learnt by the CNN unit.

To overcome these limitations, we implemented a genetic algorithm (GA) to generate artificial sequences that maximise the output of each CNN unit (Figure 3A). Briefly, each CNN unit was treated as an oracle to evaluate the "fitness" of a population of sequences. Next, sequences with the highest fitness underwent crossover and mutations, resulting in the next population of sequences. This was repeated until the average fitness of the population of sequences converged. Then, we compared the ratio of known TF motif occurrences between this set of sequences and random sequences. Using this approach, we generated artificial sequences for units one and eight. Sequences optimized for unit one were enriched for C2H2 zinc finger TFs such as PLAG1, ZNF263, and RREB1 (Figure 3C), explaining the importance of this unit for the prediction of CTCF binding, which is also a C2H2 zinc finger.

However, the sequences optimized for unit eight did not show significant enrichment for any TF motif. We hypothesised that this unit might detect the absence of CTCF by yielding a low output when a sequence contained a CTCF motif. In turn, this would help predict the binding of FOXA1 and JUND, as suggested by the importance of unit eight to the prediction of both TFs (Figure 3B). To test this hypothesis, we used the GA to generate artificial sequences that minimize the output from unit eight. As expected, the resulting sequences were enriched for the CTCF motif (Figure 3D). Moreover, these sequences yielded a high output from units two and nine, both of which are important for the prediction of CTCF.

Overall, we demonstrate that using GA with CAMs overcomes the limitations associated with motif generation from convolutional filters and provides a deeper understanding of what each independent CNN unit is learning.

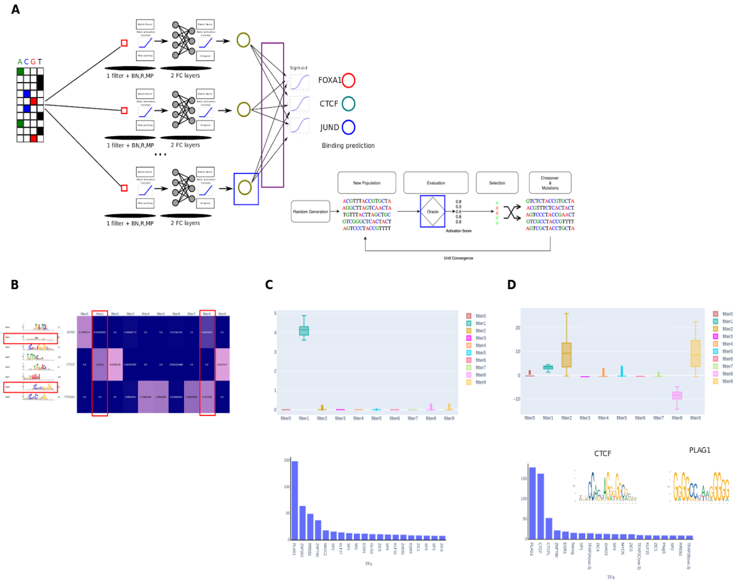

Figure 3. Artificial Sequences from Genetic Algorithm as an alternative to filter-to-PWM conversion

A . Each independent CNN unit is treated as an oracle to evaluate the fitness of a sequence. High fitness sequences are selected for crossover and mutation to obtain the next generation. This process is repeated until the average fitness of the population converges.

B. Filter one from CAM for TF binding does not correspond to motifs of any TF. Filter eight corresponds to CTCF motif but is not important for predicting CTCF.

C. Top: Distribution of outputs for artificial sequences optimised for CNN unit one.

Bottom: Artificial sequences optimised for CNN unit one are enriched for zinc finger factors such as PLAG1, ZNF263, and RREB1.

D. Top: Distribution of outputs for artificial sequences optimised to produce low output from CNN unit eight.

Bottom: Enrichment of PLAG1 and CTCF motifs in these artificial sequences, indicating the usage of this unit for the detection of an absence of CTCF motif.

CAMs identify TF cooperativity

TFs can bind to adjacent TFBSs to co-regulate gene expression in a phenomenon known as cooperativity [24]. We hypothesized that cooperative TFs would be predicted by the same set of CNN units. To test this hypothesis, we trained a CAM with 20 CNN units to predict the binding of six TFs: the chromatin remodelers CTCF, REST, and ZNF143 [25,26] and the erythropoietic factors TAL1 and GATA1, and their co-factor TEAD4 [27,28]. The CAM achieved adequate performance levels for all TFs by means of AUCROC and AUCPR (Figure 4A). Filter visualization revealed the individual motifs of CTCF and GATA1, and a composite TAL1-GATA1 motif. Furthermore, visualization of the weights from the final layer on a heatmap separated the chromatin remodelers from the erythropoietic factors into two distinct clusters (Figure 4B). These clusters were replicated with sequence level predictions (Figure 4C).

Thus clustering based on CAM final layer weights can be used to identify potential cooperativity between TFs

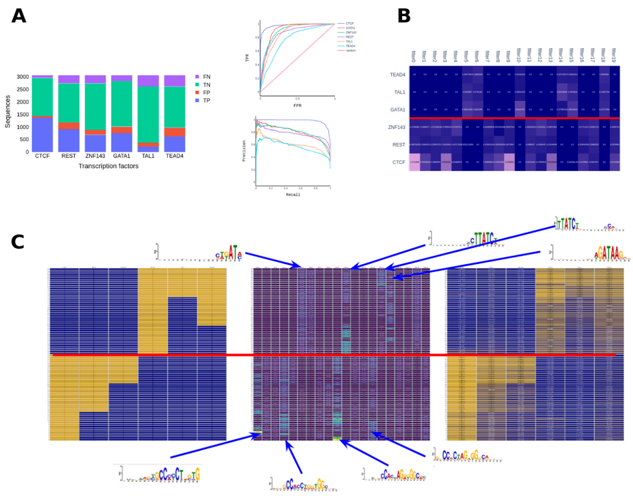

Figure 4. CAMs identify potential cooperative TFs

A. Performance of CAM on the TF binding prediction task of the selected cofactors. Confusion matrix and Precision Recall/Receiver Operating Characteristics curve for CTCF, GATA1, ZNF143, REST, TAL1 and TEAD4.

B. Visualisation of the output layer weights from CAM. Weights are clustered into two groups, corroborating with known cooperativity.

C. Sequence level predictions with CAM (as in Figure 2E). Motif representations are highlighted for the most important units.

CAMs recover the regulatory landscape of immune cell differentiation

In a recent work, Maslova *et al.* used an extensively optimized deep learning model, namely AI-TAC, to study immune cell differentiation in mouse [6]. The authors recovered known regulators using model interpretation techniques based on filter-to-PWM conversion. Specifically, they established a hierarchy of TFs and their interactions responsible for lineage specification by iteratively nullifying each filter.

To evaluate the interpretation capabilities of our approach, we trained a CAM with 300 CNN units on the same dataset (Figure 5A). We trained the CAM on ~300K ATAC-seq open chromatin regions (OCRs) to predict their accessibility in 81 stem and immune cell types. We observed that the performance of our approach (denoted by correlation between predicted and actual accessibility values) was similar to that of AI-TAC (Figure 5B) without any hyperparameter tuning. Moreover, filters from different CNN units corresponded to the motifs of known regulators of immune cell differentiation such as PAX5, EBF1, CEBP, SFPI1 and NF-kB (Figure 5C). These motifs were recovered consistently across multiple training iterations of the model.

Visualisation of the weights from the final layer revealed clusters of TFs responsible for lineage specification (Figure 5D). Specifically, PAX5 and EBF1 were identified as important for B cells, SPI1 and CEBP for myeloid cell lineages, and TBX20 for NK cells. The T cell lineage was not characterised by any strong cluster of TFs, which is in agreement with the findings from AI-TAC. More exciting was the identification of a group of TFs as important for stem cell lineage, NFIX, HOX and Ascl2 (teal coloured box in Figure 5D) [29,30], which was not detected by AI-TAC. Interestingly, this cluster is also important for early stages of B and abT cell lineages. Using the approach described in the previous section, we looked at TF clusters to identify potential cooperative factors. The analysis revealed known cooperative partners such as PAX5 and EBF1 (red coloured box in Figure 5D), which further demonstrates the utility of CAMs in identifying TF cooperativity.

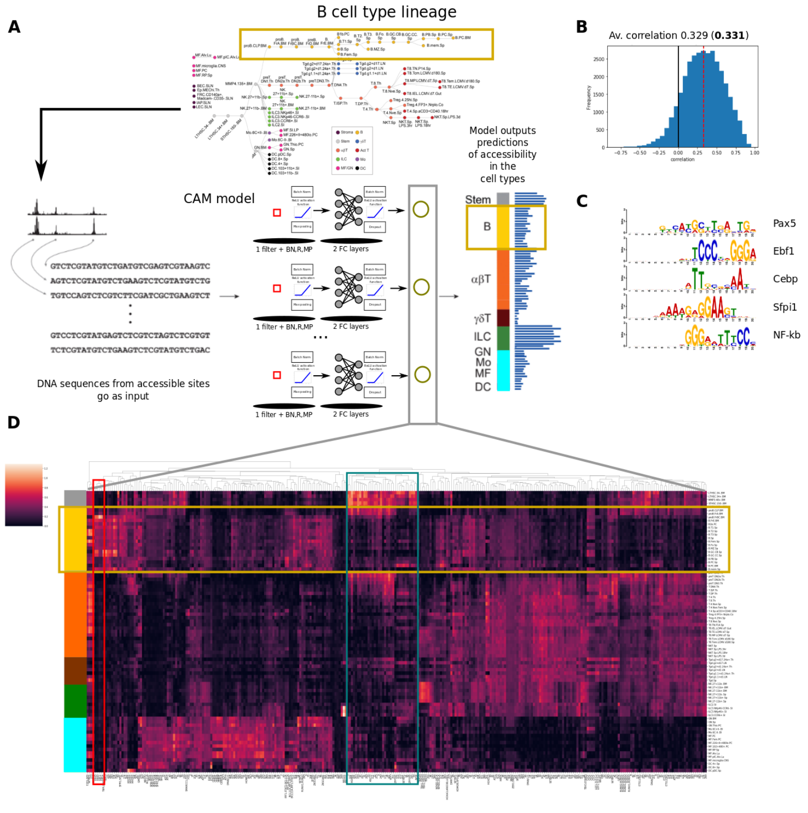
A
B cell type lineage
Model outputs predictions of accessibility in the cell types
CAM model
1 filter + BN,R,MP
2 FC layers
GTCTCGTATGTCTGATGTCGAGTCGTAAGTC
AGTCTCGTATGTCTGAAGTCTCGTATGTCTG
TGTCCAGTCTCGTCTTCGATCGCTGAAGTCT
GTCCTCGTATGAGTCTCGTCTAGTCTCGTGT
TCTCGTATGTCTGAAGTCTCGTATGTCTGAC
DNA sequences from accessible sites go as input
Stem
B
αβT
γδT
ILC
GN
Mo
MF
DC
B
Av. correlation 0.329 (0.331)
Frequency
correlation
C
Pax5
Ebf1
Cebp
Sfpi1
NF-kb
D

Figure 5. CAMs recover regulatory landscape of immune cell differentiation
A. Schematic diagram of CAM to predict chromatin accessibility across 81 stem and immune cell types in mouse. Highlighted in yellow is B cell lineage. The CAM is trained with 300 CNN units and takes 251 bp long DNA sequences as input.
B. The performance is evaluated by calculating the correlation between predicted and actual ATAC-seq activity for each input sequence. The distribution of this correlation is shown for the test set sequences. For comparison, the number in bold denotes the average correlation of the multilayer AI-TAC model on test sequences.
C. Conversion of filters from independent CNN units of CAM discovers motifs of known regulators of immune lineage specificity including PAX5, EBF1, Cebp, Sfpi1 and NK-kb. These motifs were consistently obtained in multiple training iterations.
D. Visualisation of final layer weights reveals clusters of TFs responsible for lineage specification. Each column in the heatmap represents a CNN unit of CAM, some of which are annotated to TFs based on motif similarity. Highlighted in the teal box is a cluster of TFs active in stem cell lineage - NFIX, HOX, Ascl2. This cluster was not obtained from filter nullification experiments in the AI-TAC model. PAX5 and EBF1 are clustered together (highlighted in red box) and are known cooperative partners specifying B cell lineage.

Conclusions

CAMs achieve global and local interpretability through visualisation of final layer weights and outputs without compromising on accuracy. By generating synthetic sequences optimised for each unit of CAM, we overcome the limitations associated with filter-to-PWM conversion. Clustering of final layer weights of CAM units can identify potential cooperative TFs, without iteratively nullifying combinations of filters. Finally, CAMs successfully recover the regulatory landscape of mouse immune cell differentiation without applying any hyperparameter optimisation or computationally intensive interpretation techniques. Overall, CAMs offer an easy to use, fully interpretable approach to deep learning in genomics.

References

1. Alipanahi, B., Delong, A., Weirauch, M. T. & Frey, B. J. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.* **33**, 831–838 (2015).
2. Quang, D. & Xie, X. FactorNet: A deep learning framework for predicting cell type specific transcription factor binding from nucleotide-resolution sequential data. *Methods* **166**, 40–47 (2019).
3. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods* **12**, 931–934 (2015).
4. Quang, D. & Xie, X. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Res.* **44**, e107 (2016).
5. Kelley, D. R., Snoek, J. & Rinn, J. L. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res.* **26**, 990–999 (2016).
6. Maslova, A. *et al.* Learning immune cell differentiation. *bioRxiv* 2019.12.21.885814 (2019) doi:10.1101/2019.12.21.885814.
7. Kelley, D. R. *et al.* Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Res.* **28**, 739–750 (2018).
8. Shrikumar, A., Greenside, P. & Kundaje, A. Learning Important Features Through Propagating Activation Differences. *arXiv [cs.CV]* (2017).
9. Shrikumar, A. *et al.* Technical Note on Transcription Factor Motif Discovery from Importance Scores (TF-MoDISco) version 0.5.6.5. *arXiv [cs.LG]* (2018).
10. Avsec, Ž. *et al.* Deep learning at base-resolution reveals motif syntax of the cis-regulatory code. 737981 (2019) doi:10.1101/737981.
11. Koo, P. K. & Ploenzke, M. Deep learning for inferring transcription factor binding sites. *Curr Opin Syst Biol* **19**, 16–23 (2020).
12. Koo, P. K., Ploenzke, M., Anand, P., Paul, S. B. & Majdandzic, A. Global Importance Analysis: A Method to Quantify Importance of Genomic Features in Deep Neural Networks. 2020.09.08.288068 (2020) doi:10.1101/2020.09.08.288068.
13. Sixt, L., Granz, M. & Landgraf, T. When Explanations Lie: Why Many Modified BP Attributions Fail. *arXiv [cs.LG]* (2019).
14. Koo, P. K. & Ploenzke, M. Improving representations of genomic sequence motifs in convolutional networks with exponential activations. *bioRxiv* (2020).
15. Koo, P. K. & Eddy, S. R. Representation learning of genomic sequence motifs with convolutional neural networks. *PLoS Comput. Biol.* **15**, e1007560 (2019).
16. Agarwal, R., Frosst, N., Zhang, X., Caruana, R. & Hinton, G. E. Neural Additive Models: Interpretable Machine Learning with Neural Nets. *arXiv [cs.LG]* (2020).
17. Mitchell, M. *An Introduction to Genetic Algorithms*. (MIT Press, 1998).

18. Chèneby, J., Gheorghe, M., Artufel, M., Mathelier, A. & Ballester, B. ReMap 2018: an updated atlas of regulatory regions from an integrative analysis of DNA-binding ChIP-seq experiments. *Nucleic Acids Res.* **46**, D267–D275 (2018).

19. Gheorghe, M. *et al.* A map of direct TF-DNA interactions in the human genome. *Nucleic Acids Res.* **47**, e21 (2019).

20. Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**, 75–82 (2012).

21. Lee, C. M. *et al.* UCSC Genome Browser enters 20th year. *Nucleic Acids Res.* **48**, D756–D761 (2020).

22. Fornes, O. *et al.* JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **48**, D87–D92 (2020).

23. Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L. & Noble, W. S. Quantifying similarity between motifs. *Genome Biol.* **8**, R24 (2007).

24. Panne, D., Maniatis, T. & Harrison, S. C. An atomic model of the interferon-beta enhanceosome. *Cell* **129**, 1111–1123 (2007).

25. Barisic, D., Stadler, M. B., Iurlaro, M. & Schübeler, D. Mammalian ISWI and SWI/SNF selectively mediate binding of distinct transcription factors. *Nature* **569**, 136–140 (2019).

26. Ye, B. *et al.* ZNF143 in Chromatin Looping and Gene Regulation. *Front. Genet.* **11**, 338 (2020).

27. Capellera-Garcia, S. *et al.* Defining the Minimal Factors Required for Erythropoiesis through Direct Lineage Conversion. *Cell Rep.* **15**, 2550–2562 (2016).

28. Lu, R., Mucaki, E. J. & Rogan, P. K. Discovery and validation of information theory-based transcription factor and cofactor binding site motifs. *Nucleic Acids Res.* **45**, e27 (2017).

29. Holmfeldt, P. *et al.* Nfix is a novel regulator of murine hematopoietic stem and progenitor cell survival. *Blood* **122**, 2987–2996 (2013).

30. Bhatlekar, S., Fields, J. Z. & Boman, B. M. Role of HOX Genes in Stem Cell Differentiation and Cancer. *Stem Cells Int.* **2018**, 3569493 (2018).

31. Novakovsky, G., Saraswat, M., Fornes, O., Mostafavi, S., Wasserman, W.Biologically relevant transfer learning improves transcription factor binding prediction. *Workshop on Computational Biology at International Conference on Machine Learning* **2020**,https://icml-compbio.github.io/2020/papers/WCBICML2020_paper_66.pdf (2020).