Deciphering the Genetic Code behind Single-Cell Chromatin Accessibility:

Interpreting scBasset: a Sequence-based Convolutional Neural Network

Tianhao Luo[1], Peter Koo[2]

[1]University of Pennsylvania & [2]Cold Spring Harbor Laboratory

INTRODUCTION

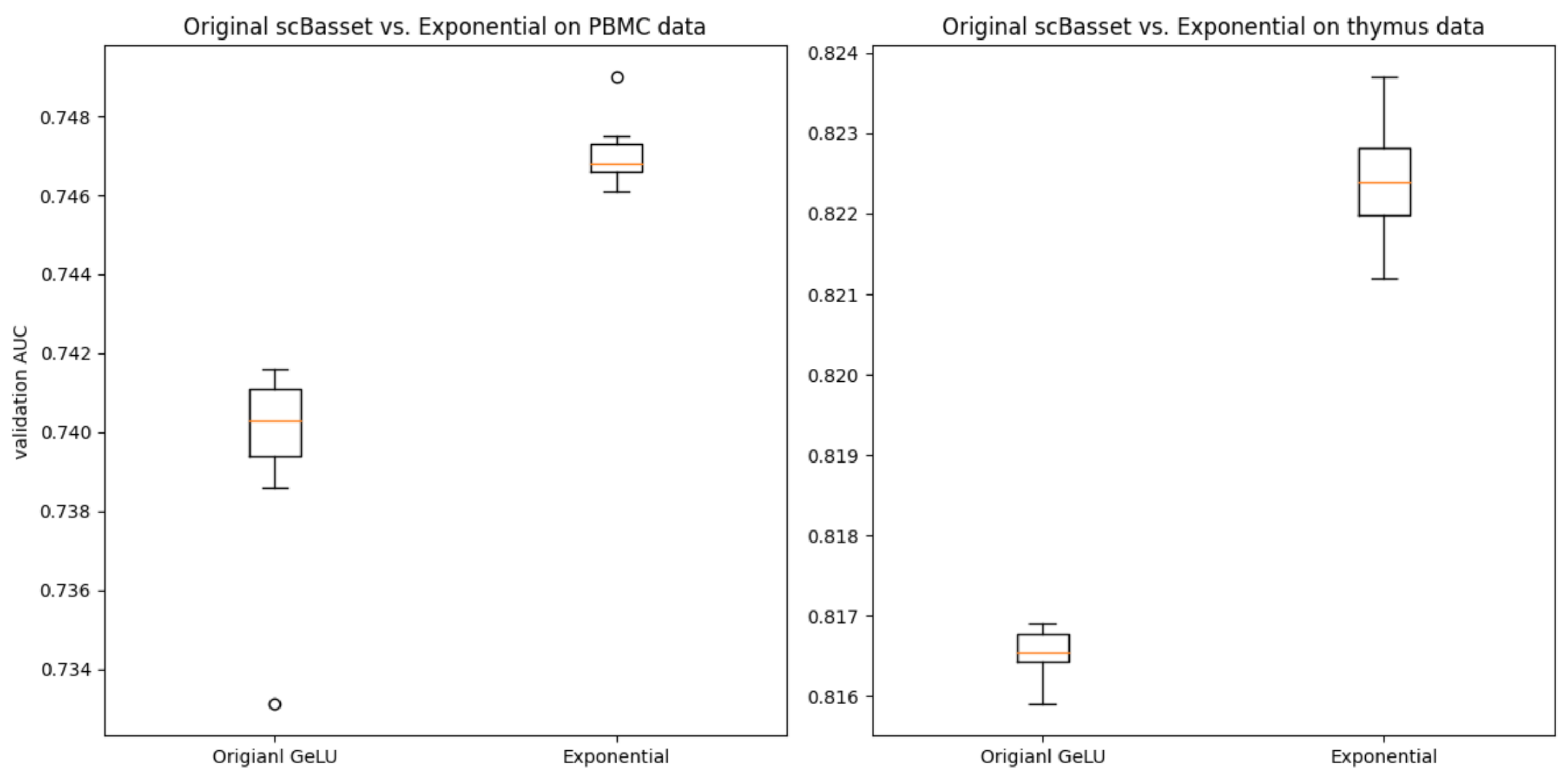
Chromatin accessibility refers to the degree to which the DNA in a cell's nucleus is accessible and open for various cellular processes, such as transcription and DNA replication.

scBasset, a **sequence-based convolutional neural network (CNN) model**, aims to predict chromatin accessibility (a binary classification task) based on single-cell ATAC-seq data.

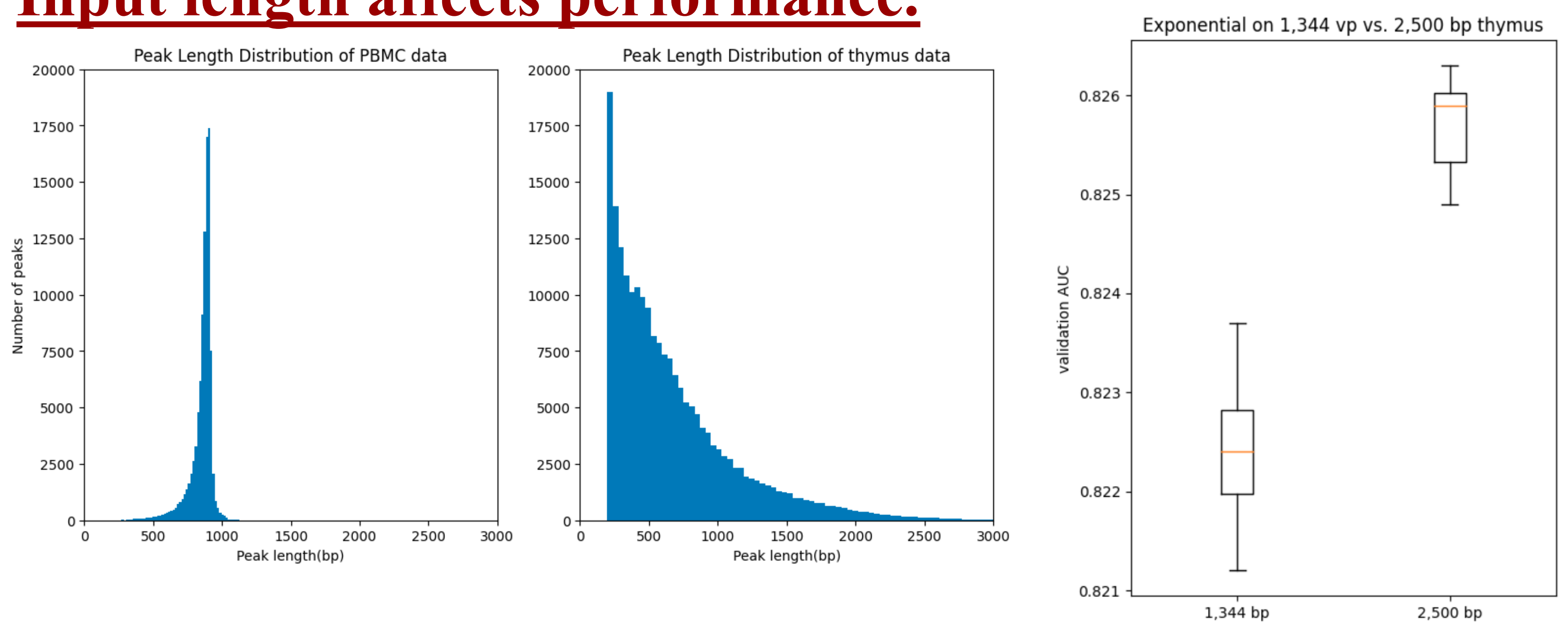
In this study, we present a comprehensive evaluation and improvement of scBasset's performance and interpretability.

RESULTS

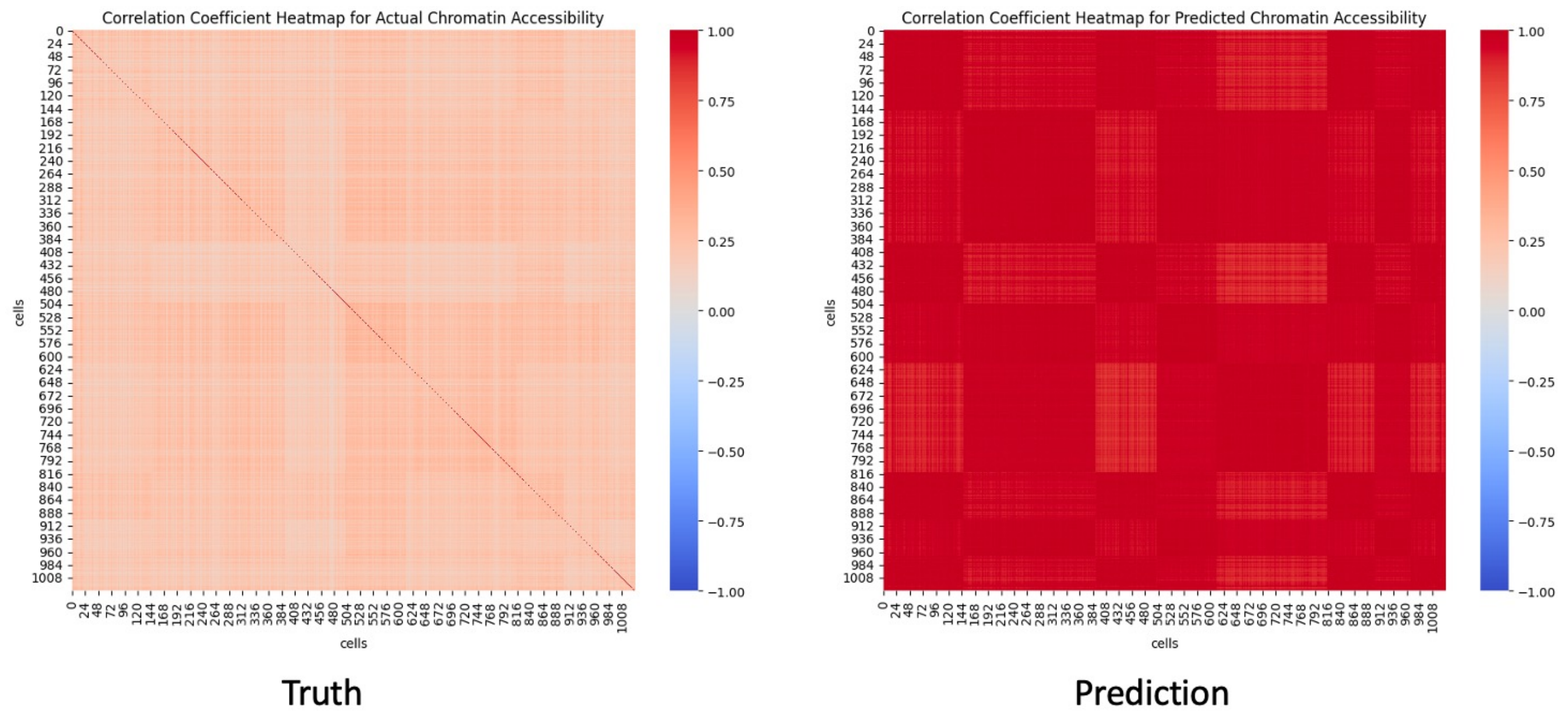
Exponential activation improves performance.



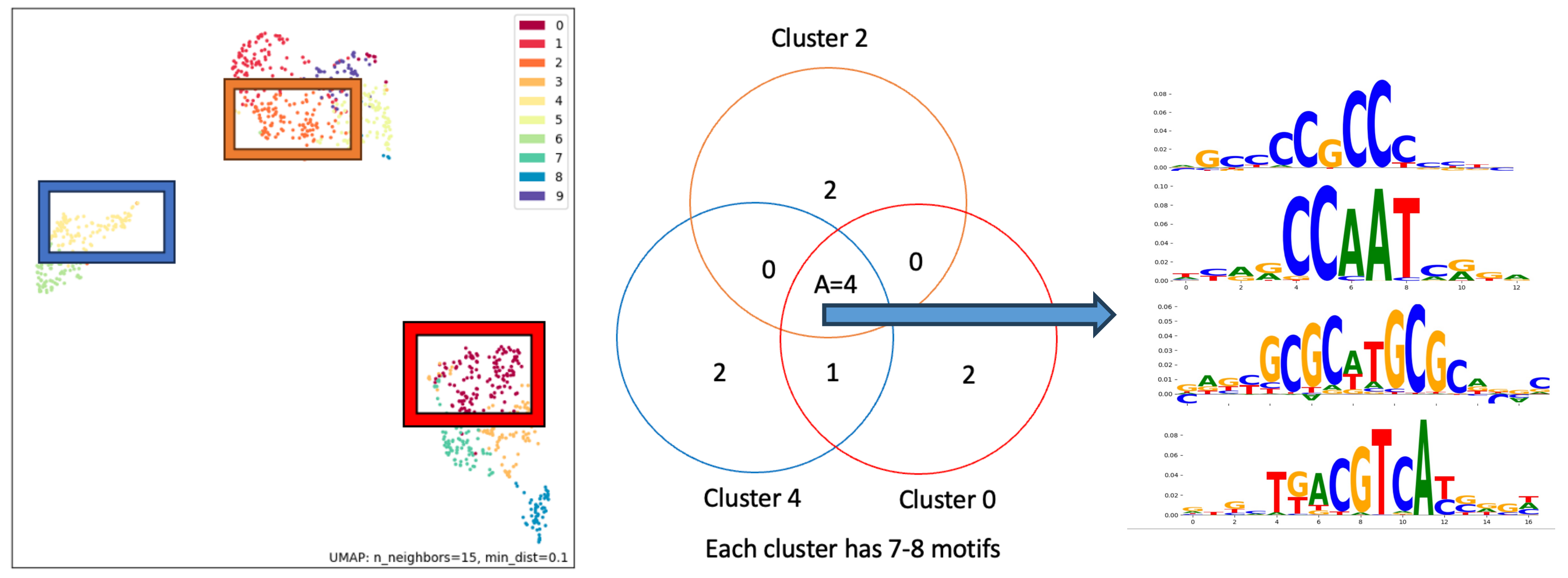
Input length affects performance.



Predictions have a much higher correlation than actual data,



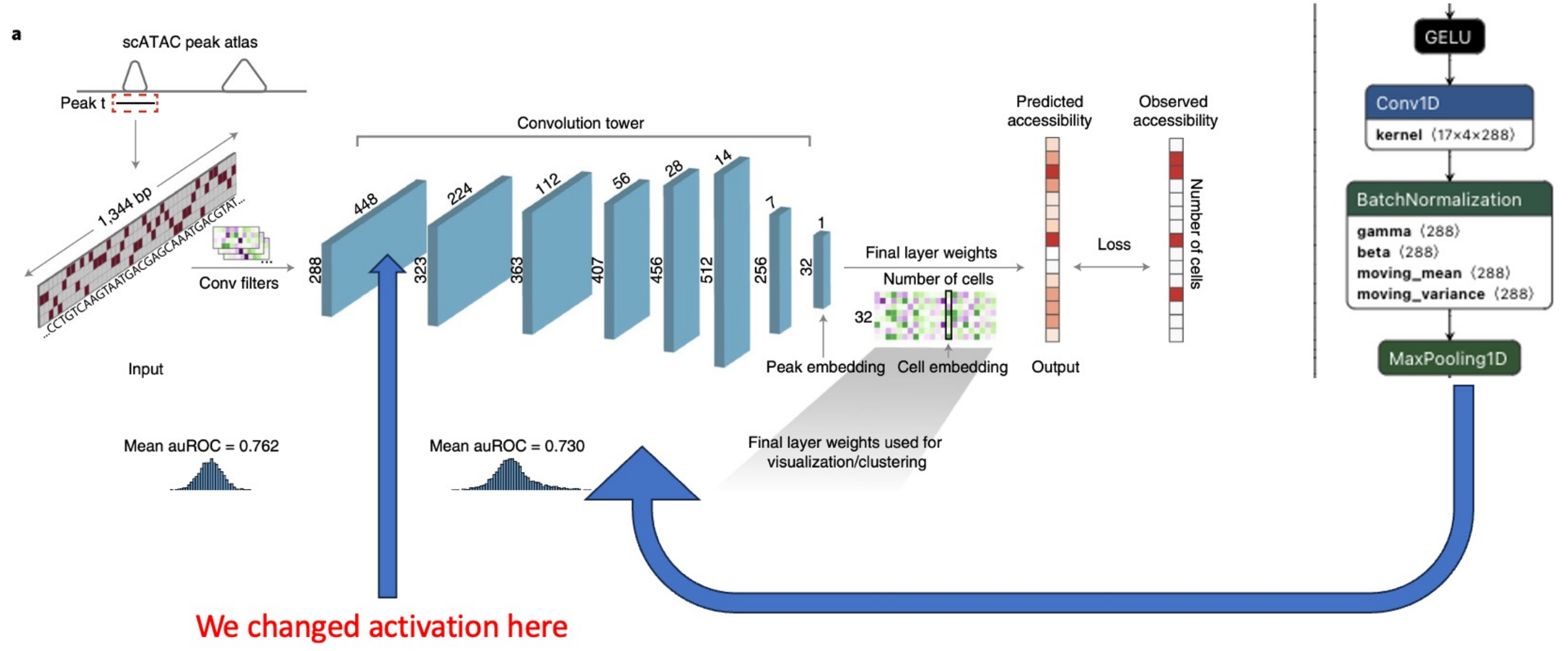
TF-MoDISco finds same motifs in different clusters.



Each cluster has 7-8 motifs

METHODS

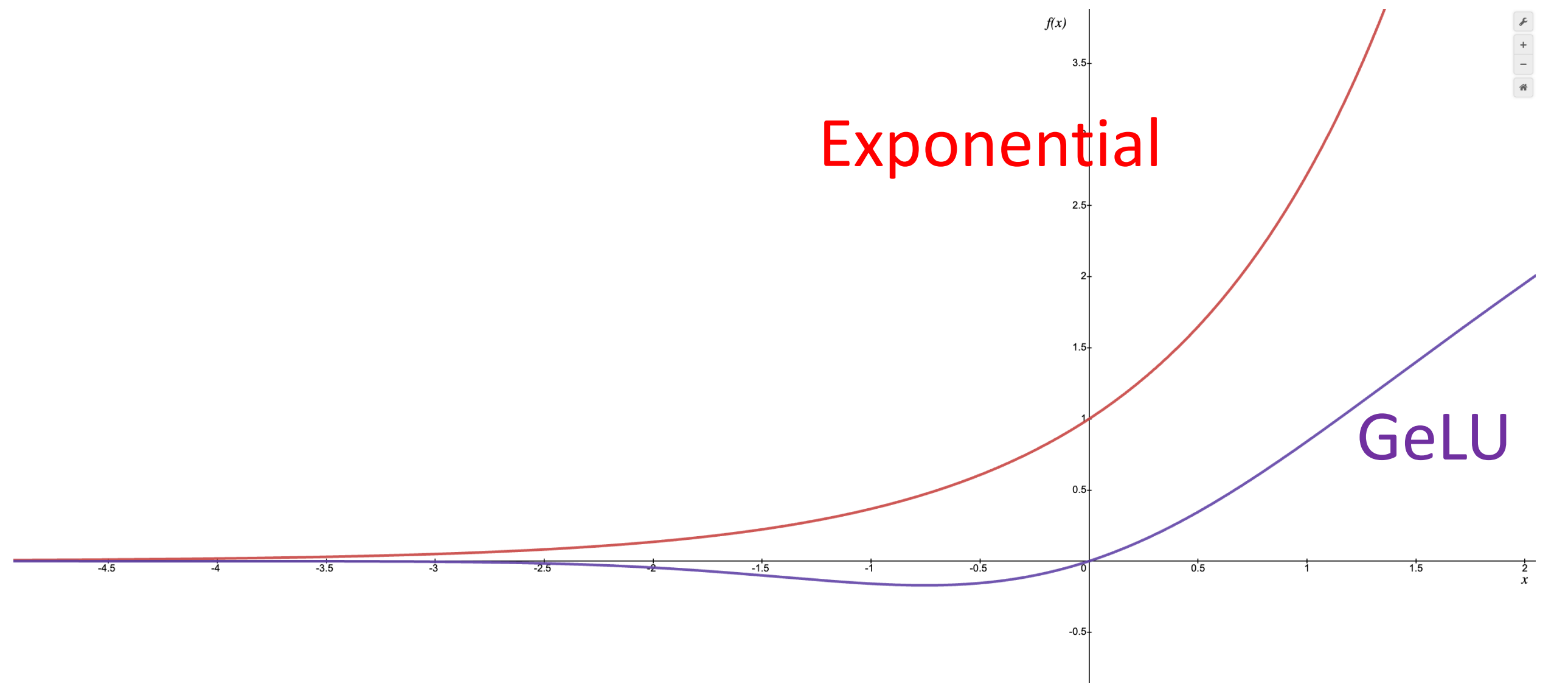
scBasset Architecture.



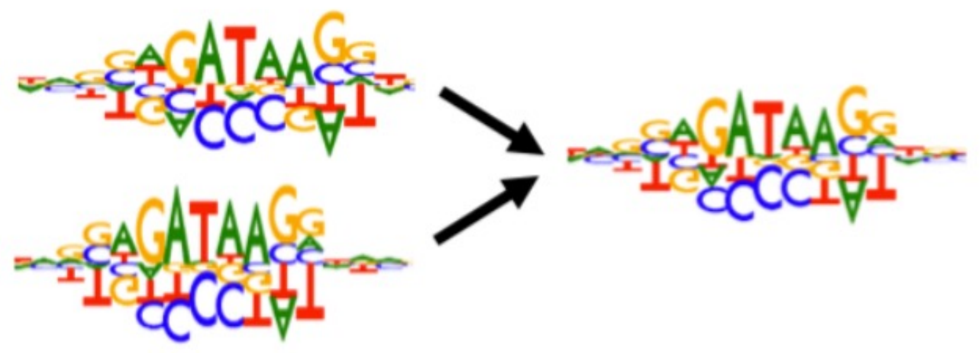
We changed activation here

Task: Predict a binary chromatin accessibility (binary classification)

Exponential activation.



TF-MoDISco.



Integrate saliency maps into motifs

CONCLUSIONS & FUTURE

- scBasset makes predictions by aggregating information across clusters; i.e., it does take the sample heterogeneity & advantages of single-cell data into account;
- The binary classification does not fully reflect the nature of chromatin accessibility
- To fully utilize the advantages of single-cell data, more consideration of
- Future single-cell models should improve clustering capability

REFERENCES

[Buenrostro] Buenrostro, J. D., Wu, B., Litzenburger, U. M., Ruff, D., Gonzales, M. L., Snyder, M. P., ... & Greenleaf, W. J. (2015). Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, *523*(7561), 486-490.

[Chen] Chen, H., Lareau, C., Andreani, T., Vinyard, M. E., Garcia, S. P., Clement, K., ... & Pinello, L. (2019). Assessment of computational methods for the analysis of single-cell ATAC-seq data. *Genome biology*, *20*(1), 1-25.

[Koo1] Koo, P. K., & Eddy, S. R. (2019). Representation learning of genomic sequence motifs with convolutional neural networks. *PLoS computational biology*, *15*(12), e1007560.

[Koo2] Koo, P. K., Majdandzic, A., Ploenzke, M., Anand, P., & Paul, S. B. (2021). Global importance analysis: An interpretability method to quantify importance of genomic features in deep neural networks. *PLoS computational biology*, *17*(5), e1008925.

[Koo3] Koo, P. K., & Ploenzke, M. (2021). Improving representations of genomic sequence motifs in convolutional networks with exponential activations. *Nature machine intelligence*, *3*(3), 258-266.

[Lee] Lee, N. K., Tang, Z., Toneyan, S., & Koo, P. K. (2023). EvoAug: improving generalization and interpretability of genomic deep neural networks with evolution-inspired data augmentations. *Genome Biology*, 24(1), 105.

[Maj1] Majdandzic, A., Rajesh, C., & Koo, P. K. (2023). Correcting gradient-based interpretations of deep neural networks for genomics. *Genome Biology*, *24*(1), 1-13.

[Maj2] Majdandzic, A., Rajesh, C., Tang, Z., Toneyan, S., Labelson, E. L., Tripathy, R. K., & Koo, P. K. (2022, December). Selecting deep neural networks that yield consistent attribution-based interpretations for genomics. In *Machine Learning in Computational Biology* (pp. 131-149). PMLR.

[Nov] Novakovsky, G., Dexter, N., Libbrecht, M. W., Wasserman, W. W., & Mostafavi, S. (2023). Obtaining genetics insights from deep learning via explainable artificial intelligence. *Nature Reviews Genetics*, *24*(2), 125-137.

[Parks] Park, S., Koh, Y., Jeon, H., Kim, H., Yeo, Y., & Kang, J. (2020). Enhancing the interpretability of transcription factor binding site prediction using attention mechanism. *Scientific reports*, *10*(1), 13413.

[Quang] Quang, D., & Xie, X. (2016). DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic acids research*, *44*(11), e107-e107.

[Shrik] Shrikumar, A., Tian, K., Avsec, Ž., Shcherbina, A., Banerjee, A., Sharmin, M., ... & Kundaje, A. (2018). Technical note on transcription factor motif discovery from importance scores (TF-MoDISco) version 0.5. 6.5. *arXiv preprint arXiv:1811.00416*.

[Sri] Srivastava, D., Aydin, B., Mazzoni, E. O., & Mahony, S. (2021). An interpretable bimodal neural network characterizes the sequence and preexisting chromatin predictors of induced transcription factor binding. *Genome biology*, 22(1), 1-25.

[Tal] Talukder, A., Barham, C., Li, X., & Hu, H. (2021). Interpretation of deep learning in genomics and epigenomics. *Briefings in Bioinformatics*, *22*(3), bbaa177.

[Tone] Toneyan, S., Tang, Z., & Koo, P. K. (2022). Evaluating deep learning for predicting epigenomic profiles. *Nature machine intelligence*, *4*(12), 1088-1100.

[Yan] Yan, F., Powell, D. R., Curtis, D. J., & Wong, N. C. (2020). From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. *Genome biology*, 21, 1-16.

[Yuan] Yuan, H., & Kelley, D. R. (2022). scBasset: sequence-based modeling of single-cell ATAC-seq using convolutional neural networks. *Nature Methods*, *19*(9), 1088-1096.

[Zhou] Zhou, Y., Booth, S., Ribeiro, M. T., & Shah, J. (2022, June). Do feature attribution methods correctly attribute features?. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 36, No. 9, pp. 9623-9633).

ACKNOWLEDGEMENT & CONTACT

I am deeply grateful to Dr. Peter Koo, my faculty advisor, for his invaluable help and step-by-step guidance throughout my project, both in terms of the research and my personal growth. I also thank everyone in the Koo Lab for fostering a collaborative and friendly environment. Special thanks to URP program coordinators for organizing this wonderful research opportunity at the Cold Spring Harbor Laboratory. I am honored to be supported by the Dorcas Cummings Scholar and 30th Anniversary URP Scholar, which made this research possible.

For contact, please email me at: **lth888@wharton.upenn.edu.**