Enhanced Metrics for Assessing Reliability of Deep Neural Networks in Regulatory Genomics

Sophia Chen, Farragut High School

The author thanks Dr. Peter Koo of Cold Spring Harbor Laboratory for his mentorship and guidance throughout the course of this research.

Abstract

The field of regulatory genomics aims to understand the mechanisms by which the non-coding genome controls gene expression, and deep learning has proven to be a powerful tool for such tasks. Despite their strong predictive capabilities, the black-box nature of deep neural networks presents the critical challenge of interpretability, and existing tools such as attribution analysis often result in inconsistent maps. In genomics, there exists only a single method to quantitatively assess pattern consistency across a set of attribution maps, but it is reliant on arbitrary choices. Here, this method is extended to address the limitations introduced by these choices. Attribution scores and length scaling are incorporated to ensure a multi-scale characterization of pattern consistency. Both simulated and real-world experimental data are used to demonstrate the success of this approach in identifying models that locate consistent patterns and thus are more human interpretable. This work provides a robust metric for selecting models that balance generalization performance with reliability.

Introduction

Gene regulation allows cells to control expression in response to various signals, allowing them to maintain function and adequately adapt. Dysregulation of gene expression is the underlying cause of conditions such as cancer and autoimmune disorders. The processes involved in gene regulation are mediated by transcription factors (TFs), proteins that bind specific sequences to coordinate interactions between distal regulatory elements (e.g., enhancers and silencers), which in turn regulate transcriptional output. Interactions are governed by the cis-regulatory code, which describes how TFs recognize and bind sites [1]. However, TF binding is complex due to the vast number of potential binding sites, which are represented by short sequence motifs, and understanding it remains a major goal within regulatory genomics.

Mapping TF binding sites has long relied on experimental approaches, such as chromatin immunoprecipitation followed by sequencing (ChIP-seq) [2] or transposable-accessible chromatin using sequencing (ATAC-seq) [3]. These methods, however, are limited in resolution and often identify regions of hundreds to thousands of nucleotides (nts), whereas motifs are usually between 5 and 20 nts in length [4]. High-throughput techniques such as CRISPR-based screening [5] and saturation mutagenesis [6], can improve resolution, but the outputs are still are often noisy. As a result, computational techniques have increased in usage.

Traditional computational approaches for motif discovery include tools such as MEME (Multiple Expectation maximizations for Motif Elicitation), which identifies enriched sequence patterns likely to be associated with TF activity [7]. K-mer based approaches, such as gapped k-mer support vector machines [8], analyze sequences of length k and offer more granularity. However, deep learning (Fig. 1) has emerged as the preferred approach over traditional methods due to its ability to handle larger volumes of data with greater accuracy and flexibility [9].

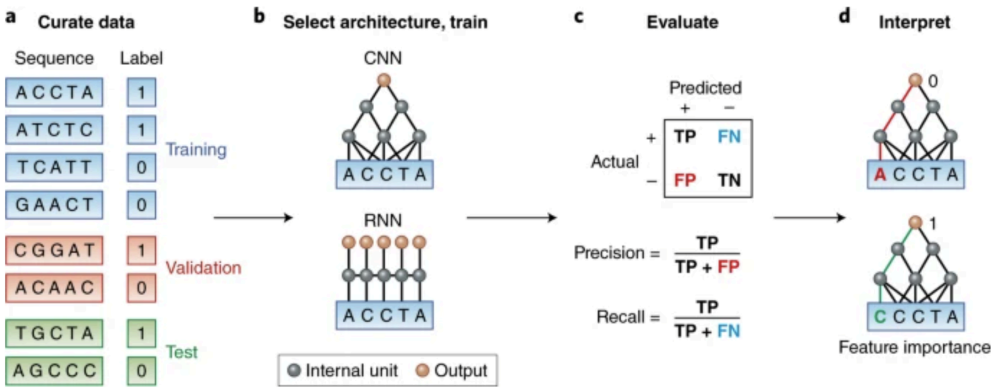

Figure 1. Workflow of deep learning in genomics. (a) Sequence data preparation involves labeling and splitting and is followed by (b) architecture selection, with popular choices including convolutional and recurrent neural networks. (c) Model evaluation and (d) feature interpretation are key steps in model selection, with a goal of balancing performance and interpretability [10].

Deep learning models such as neural networks have demonstrated improved capabilities in genomic tasks such as chromatin accessibility or regulatory motif prediction [11, 12]. These models are typically trained on data from the aforementioned experimental techniques. However, these models are highly complex, often containing layers of thousands of interconnected neurons in order to learn intricate patterns. While this gives them high predictive power, it also limits their inherent interpretability. The lack of transparency in the decision-making of these black-box models represents a barrier to practical application [13, 14].

The main approach to interpretation is attribution analysis, which highlights the parts of an input that most influence the output. In genomics, locating regions with high attribution scores can provide insight into cis-regulatory elements. However, the outputs of such analyses can vary significantly across methods or models [15]. Additionally, a common standard in model selection is to focus on generalization performance. However, it has been shown that predictive performance is not indicative of interpretability, as there is not necessarily a “trade-off” between

the two [16]. As a result, identifying models that are both high-performing and biologically insightful has become a major challenge. Saliency maps are a common choice of attribution analysis in genomics, but are limited to qualitative analyses [17]. This further emphasizes the need for a method to quantify reliability.

The existing method [18] employs a metric that quantifies pattern consistency within a population of attribution maps generated from different models. It utilizes Kullback-Leibler divergence (KLD) to assess the reliability of the features captured by a model. The idea behind this approach is that highly attributed regions will show a sparse distribution by consistently identifying certain k-mers as important, which would be more likely to contain biologically significance. On the other hand, the k-mer distribution across the entire sequence will be more diffuse, showing roughly similar frequencies of all k-mers. KLD calculates the difference between these distributions, so a high KLD indicates a more reliable model that better identifies important k-mers. A lower KLD indicates that a model is more likely focusing on spurious noise.

The arbitrary choices involved in this metric, however, limit its effectiveness. The cutoff to determine highly-attributed regions neglects binding sites, and the selection of a fixed k-mer length assumes that motifs are a uniform size, which does not accurately reflect biological sequences. Here, this method is expanded to eliminate the need for arbitrary choices. Synthetic data, which contained ground truth, was used to develop the enhanced metric, which was then shown by saliency maps to generalize well to real-world data. The resulting method offers a more comprehensive way to quantitatively assess model reliability, allowing improved selection of models that are both effective and human interpretable.

Methods

Attribution analysis

Convolutional neural networks (CNNs) have gained recognition in many fields, especially for tasks like image recognition [19]. These models use filters to detect local patterns across larger inputs; for instance, in genomics, these filters slide across a sequence to identify motifs. CNNs offer a more robust approach to motif discovery compared to traditional computational methods due to their ability to detect patterns regardless of their position in the input. However, increasingly complex models also become more difficult to interpret. This raises concerns about the ability of models to capture biologically meaningful features, as deep learning models have been observed to capture spurious, data-specific relationships that provide no real insight. In order to effectively leverage the computational power offered by CNNs, it is crucial to turn the black box transparent.

Attribution analysis works to resolve this challenge of interpretability. Several types of attribution methods have been developed, including gradient-based, perturbation-based, and model-agnostic [20, 21, 22]. By identifying the parts of an input that have the greatest impact on the output, these methods provide insight into the decision-making within a complex model. Saliency maps, which use backpropagation to compute the gradients of each nucleotide at each position, are a popular choice in genomics. These maps are visualized as heatmaps, which show nucleotide-level importance in influencing a model's decision.

Synthetic & real-world tasks

Synthetic binary classification task

The synthetic data [23] consisted of 20,000 sequences of 200 nts each, embedded with known motifs. The task consisted of a binary classification problem: sequences in the positive class contained at least 3 motifs from a set of "core motifs," and those in the negative class contained at least 3 from a different set of "background motifs." The sequences were pooled and randomly split into training, validation, and test sets along a 0.7, 0.1, and 0.2 split, respectively.

The baseline model was a deep neural network with 5 hidden layers. By changing various aspects on top of the base architecture and using different random initializations, a total of 390 models were trained, of which 327 passed the performance threshold of 0.97. Models were trained for 100 epochs using Adam optimizer [24], using binary cross-entropy loss to monitor learning rate decay. The varied aspects included batch normalization, activation function (ReLU or exponential), and regularization strategies (input mixup [25], manifold mixup [26], input noise [27], manifold noise, adversarial training, or spectral norm regularization [28]). In this analysis, only predetermined model weights were used.

Since this data was simulated, there was ground truth regarding which motifs were embedded and their precise locations. Signal-to-noise ratio (SNR) was used to quantify how well models captured ground truth motifs. Signal was defined as the average attribution scores at the positions of embedded motifs, and noise as the average of the top 10 false positive attribution scores. Noise therefore highlights the regions that would most likely be inaccurately identified as motifs in practical applications. High values of SNR indicate that a model is able to identify more meaningful information. While it is able to quantify model reliability, calculation of SNR requires "pixel-level" ground truth of motif positions and is thus limited to synthetic data.

Experimental chromatin accessibility task

The experimental data [29] was sourced from the ENCODE database [30] and consisted of ATAC-seq data for 15 human cell lines (analysis here focused on the GM12878 cell line). This data was used for 2 different tasks: (1) binary classification, which determined whether a region was accessible or not based on the statistical significance of ATAC-seq peaks, and (2) quantitative regression, which predicted coverage values to indicate degree of accessibility. Again, the data was split. Chromosome 8 was held out for testing, chromosome 9 was held out for validation, and the rest (excluding the Y chromosome) was used for training.

A total of 26 models were trained, 8 on the binary classification task and 18 on quantitative regression. Multiple base architectures were used, including forms of Basenji, BPNet, and a baseline CNN [31, 32]. As with the synthetic data, activation function (ReLU or exponential) was varied. Models were trained for up to 100 epochs (employing early stopping) using Adam optimizer with default parameters. Predictive performance for the test set was measured by Pearson's correlation coefficient. Again, only predetermined model weights were used here.

Metric improvements

In the existing method, three steps were taken to calculate KLD. First, a k-mer size was selected, followed by selection of a score cutoff to define attributed positions, and lastly calculation of global k-mer frequencies (which served as the uninformative prior). The first two steps were used to calculate the frequencies of attributed k-mers, and KLD quantified the difference in these two distributions. Reliable models were expected to produce a more sparse distribution, resulting in a greater value of KLD. Both these steps, however, are dependent on arbitrary choices. The original analysis used 6-mers with attributed positions defined as those with scores in the 90th percentile of each sequence (including a buffer of 2 nts added to one end).

Two major additions removed the need for arbitrary choices. To remove the cutoff for defining attributed positions, a method for weighting by attribution scores was developed to allow all instances of each k-mer to be counted. For each possible k-mer of a given length, a weight was calculated by summing the absolute attribution scores of its constituent nucleotides and aggregating these sums across each instance of the k-mer within the sequence. To normalize the weights, these aggregate sums for each unique k-mer were divided by the sum of the absolute attribution scores in the entire sequence. These were used to create a weighted frequency in which each k-mer in the global frequency was multiplied by its corresponding weight. This weighted frequency was then used as the second distribution in the KLD calculation.

To remove the need for selecting a value of k, a method for weighting KLD based on length was developed. These length weights were derived from the variance ratios in KLD when evaluated against SNR, which was used as a measure of ground truth reliability. For a value of k, a higher ratio indicated that the variance in KLD was able to better explain the corresponding variance in SNR and thus better distinguish between different models based on interpretability. On the other hand, a lower variance indicated poorer model differentiation. Values of k that demonstrated higher variance ratios were therefore considered more informative.

KLD for each value of k demonstrated an exponential relationship with SNR, so a linear model was used to evaluate the weights based on the logarithm of SNR. To ensure generalizability of the scaling weights, the synthetic models were split into 10 unique subsets, and one subset was held out each time for testing across 10 trials. The training process additionally incorporated k-fold cross-validation (k=5) to prevent overfitting.

This complete method was then validated on a real-world task. The attribution weighting and length scaling were both applied to the chromatin accessibility data, which, unlike the

synthetic dataset, did not contain ground truth. Qualitative analysis of attribution maps was used to assess the results of the new metric on real-world data.

Results

The attribution weighting method was applied to the k-mer frequencies from each sequence. As the scores required for this were obtained through attribution analysis, ground truth interpretability was not required, so these weights could be applied to both the synthetic and experimental sequences directly. Fig. 2 displays an example comparison of the weighted frequency and prior from one model (a baseline CNN trained on the quantitative regression task).

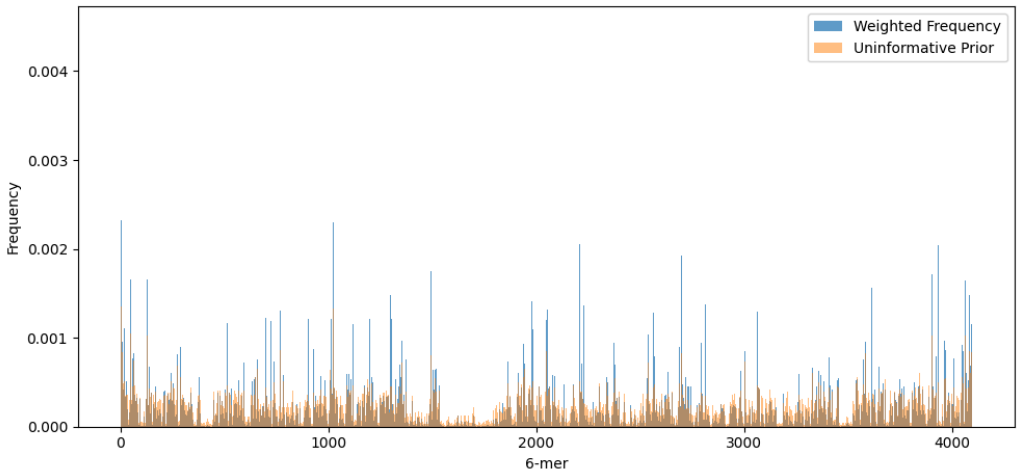

Figure 2. Comparison of frequency distributions between the attribution-weighted frequency and uninformative prior of 6-mers for one of the models trained on experimental data. Applying attribution scores as weights identifies sequence features that were not captured in the prior.

Since the real-world data lacked ground truth, the synthetic data was used to develop the length weights (Fig. 3a). To ensure that this final metric (weighted by attribution scores and scaled by length) related to ground truth interpretability, SNR served as a calibration. Plotting SNR against weighted KLD confirmed that this metric did indeed correlate strongly with ground truth (Fig. 3b). A key advantage of KLD is that it can be used to indicate reliability on any task,

unlike SNR which is limited to synthetic data. This correlation demonstrated the validity of this approach to quantify reliability on any task, including those where ground truth is not available.

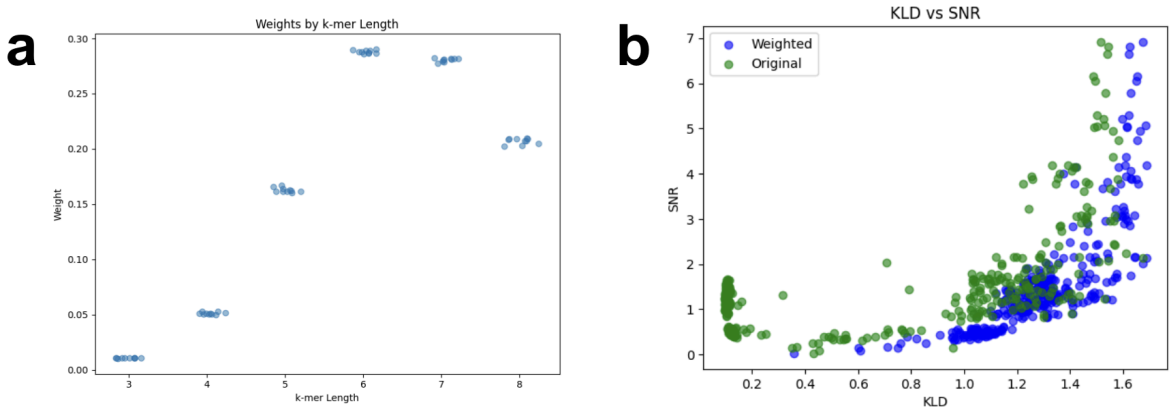

Figure 3. Length scaling incorporates results from 6 k-mer lengths into a single correlation. a) The weights generated by each subset are very consistent, validating the scaling approach. b) The scaling weights were used to generalize all the data from k=3-8 into a single correlation. This metric showed a significantly stronger correlation (r_s = .868) when compared to the original method (r_s = .587).

To assess the generalizability of this metric, it was applied to the chromatin accessibility task. Since this data does not contain ground truth, attribution analysis was used to validate the results. The saliency maps, which show the nucleotide-level importance at each position in a sequence, qualitatively demonstrate the levels of signal and noise in a model's output. Attribution weighting and the previously calculated length weights were applied to this data.

Pearson correlation of these models was plotted against their weighted KLD. The resulting plot was not monotonic, demonstrating that performance and interpretability are indeed not necessarily correlated. This indicates that predictive performance cannot be accurately used to assess model reliability, supporting the need for the KLD metric.

To demonstrate the validity of this metric on real-world data, pairs of models with similar predictive performance but different KLD values were compared at three levels of performance

(Fig. 4a). The corresponding saliency maps for these pairs showed that models with higher KLD indeed generated attribution maps with clearer motifs and far less noise (Fig. 4b). This highlights the issues with model selection based on performance, as it shows that interpretability can vary significantly among models with matched predictive performance. In order to employ such models in practical applications, model selection must incorporate interpretability metrics to ensure reliable downstream analysis.

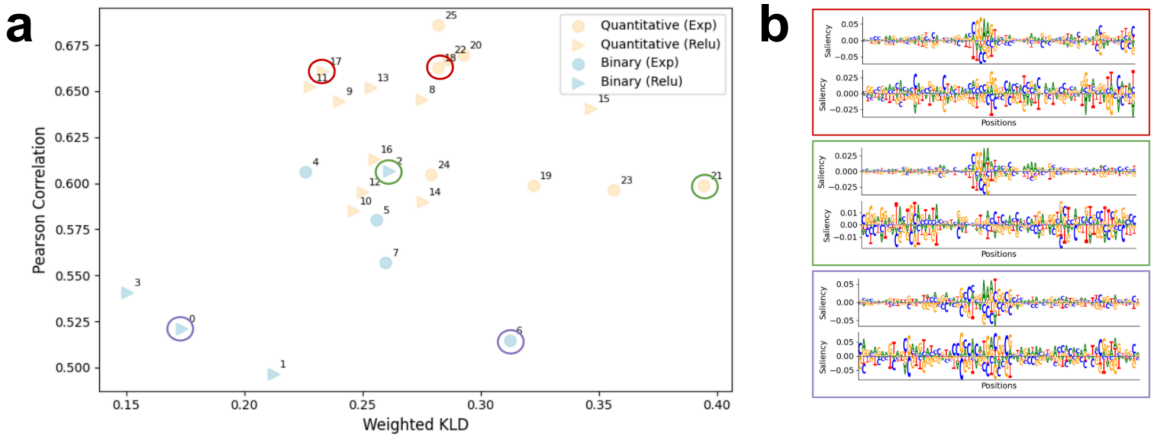

Figure 4. Attribution maps support the generalization of the metric to experimental data. a) Pairs of models evaluated on chromatin accessibility data after applying the new method show similar generalization performance but a spread in KLD. b) The efficacy of the metric is examined through qualitative assessment of attribution maps. All pairs indicate that models with a greater weighted KLD are better able to identify motifs and contain less noise.

Discussion

The practice of model selection based purely on performance overlooks the crucial aspect of reliability. This is particularly relevant in domains such as genomics, where downstream interpretability is just as important as predictive accuracy. In selecting models that have strong predictive power while producing human-interpretable outputs, an efficient method for quantifying reliability is needed. The analysis here supports the idea that performance does not

necessarily correlate with interpretability, nor is there a clear trade-off. The demonstrated metric, on the other hand, shows a strong correlation with ground truth interpretability (SNR).

This method offers several advantages over the original. By incorporating attribution-weighted k-mers, all k-mers are considered without the need for a threshold, ensuring no potential motifs are neglected. Additionally, the multi-scale approach introduced here simultaneously considers k-mers of multiple lengths, which is a more accurate representation of biological truth, as motifs are not uniform in length. The enhanced metric demonstrates a significantly greater correlation with SNR; in particular, a greater spread in KLD is seen among the models with lower SNR, which better distinguishes them by interpretability.

This method is limited by its reliance on attribution analysis, as these maps and the motifs they indicate are here viewed as standalone structures, which can be misinformative. Higher-order motif interactions, which are important in actual biological systems, are not accounted for [33]. To improve this metric, future work could integrate interaction modeling techniques to provide a more comprehensive view of motifs and their interactions. Validating the method across a wider variety of models and datasets would also be beneficial so as to improve understanding of its generalizability.

Code:
https://colab.research.google.com/drive/1BfNFsKkm4wiGWFQIs4RXSdIBQDsUMpKK?usp=sharing

References

[1] Nam, J., Dong, P., Tarpine, R., Istrail, S., & Davidson, E. H. (2010). Functional *cis* -regulatory genomics for systems biology. *Proceedings of the National Academy of Sciences*, *107*(8), 3930–3935. https://doi.org/10.1073/pnas.1000147107

[2] Park, P. J. (2009). ChIP–seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics*, *10*(10), 669–680. https://doi.org/10.1038/nrg2641

[3] Grandi, F. C., Modi, H., Kampman, L., & Corces, M. R. (2022). Chromatin accessibility profiling by ATAC-seq. *Nature Protocols*, *17*(6). https://doi.org/10.1038/s41596-022-00692-9

[4] Alcántara-Silva, R., Alvarado-Hermida, M., Díaz-Contreras, G., Sánchez-Barrios, M., Carrera, S., & Galván, S. C. (2017). PISMA: A Visual Representation of Motif Distribution in DNA Sequences. *Bioinformatics and Biology Insights*, *11*, 117793221770090. https://doi.org/10.1177/1177932217700907

[5] Bock, C., Datlinger, P., Chardon, F., Coelho, M. A., Dong, M. B., Lawson, K. A., Lu, T., Maroc, L., Norman, T. M., Song, B., Stanley, G., Chen, S., Garnett, M., Li, W., Moffat, J., Qi, L. S., Shapiro, R. S., Shendure, J., Weissman, J. S., & Zhuang, X. (2022). High-content CRISPR screening. *Nature Reviews Methods Primers*, *2*(1). https://doi.org/10.1038/s43586-021-00093-4

[6] Georgescu, R., Bandara, G., & Sun, L. (2003). Saturation mutagenesis. *Methods in Molecular Biology (Clifton, N.J.)*, *231*, 75–83. https://doi.org/10.1385/1-59259-395-X:75

[7] Bailey, T. L., & Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings. International Conference on Intelligent Systems for Molecular Biology*, *2*, 28–36. https://pubmed.ncbi.nlm.nih.gov/7584402/

[8] Ghandi, M., Lee, D., Mohammad-Noori, M., & Beer, M. A. (2014). Enhanced Regulatory Sequence Prediction Using Gapped k-mer Features. *PLoS Computational Biology*, *10*(7), e1003711. https://doi.org/10.1371/journal.pcbi.1003711

[9] Liu, J., Li, J., Wang, H., & Yan, J. (2020). Application of deep learning in genomics. *Science China. Life Sciences*, *63*(12), 1860–1878. https://doi.org/10.1007/s11427-020-1804-5

[10] Zou, J., Huss, M., Abid, A., Mohammadi, P., Torkamani, A., & Telenti, A. (2018). A primer on deep learning in genomics. *Nature Genetics*, *51*(1), 12–18. https://doi.org/10.1038/s41588-018-0295-5

[11] Liu, Q., Xia, F., Yin, Q., & Jiang, R. (2018). Chromatin accessibility prediction via a hybrid deep convolutional neural network. *Bioinformatics*, *34*(5), 732–738. https://doi.org/10.1093/bioinformatics/btx679

[12] Yang, J., Ma, A., Hoppe, A. D., Wang, C., Li, Y., Zhang, C., Wang, Y., Liu, B., & Ma, Q. (2019). Prediction of regulatory motifs from human Chip-sequencing data using a deep learning framework. *Nucleic Acids Research*, *47*(15), 7809–7824. https://doi.org/10.1093/nar/gkz672

[13] Li, X., Xiong, H., Li, X., Wu, X., Zhang, X., Liu, J., Bian, J., & Dou, D. (2022). Interpretable deep learning: interpretation, interpretability, trustworthiness, and beyond. *Knowledge and Information Systems*, *64*(12), 3197–3234. https://doi.org/10.1007/s10115-022-01756-8

[14] Talukder, A., Barham, C., Li, X., & Hu, H. (2020). Interpretation of deep learning in genomics and epigenomics. *Briefings in Bioinformatics*, *22*(3). https://doi.org/10.1093/bib/bbaa177

[15] Jyoti, A., Ganesh, K. B., Gayala, M., Tunuguntla, Nandita Lakshmi, Kamath, S., & Balasubramanian, V. N. (2022). On the Robustness of Explanations of Deep Neural Network Models: A Survey. *arXiv:2211.04780.* https://doi.org/10.48550/arXiv.2211.04780

[16] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

[17] Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *arXiv:1312.6034.* https://doi.org/10.48550/arXiv.1312.6034

[18] Majdandzic, A., Rajesh, C., Tang, A., Toneyan, S., Labelson, E., Tripathy, R., & Koo, P. K. (2022). Selecting deep neural networks that yield consistent attribution-based interpretations for genomics. *Proceedings of Machine Learning Research*, *200*, 131–149. https://proceedings.mlr.press/v200/majdandzic22a.html

[19] Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J., & Chen, T. (2018). Recent advances in convolutional neural networks. *Pattern Recognition*, *77*, 354–377. https://doi.org/10.1016/j.patcog.2017.10.013

[20] Ivanovs, M., Kadikis, R., & Ozols, K. (2021). Perturbation-based methods for explaining deep neural networks: A survey. *Pattern Recognition Letters*, *150*, 228–234. https://doi.org/10.1016/j.patrec.2021.06.030

[21] Ancona, M., Ceolini, E., Öztireli, C., & Gross, M. (2018). Towards better understanding of gradient-based attribution methods for Deep Neural Networks. *arXiv:1711.06104.* https://doi.org/10.48550/arXiv.1711.06104

[22] Chowdhury, T., Rahimi, R., & Allan, J. (2023). Rank-LIME: Local Model-Agnostic Feature Attribution for Learning to Rank. *arXiv (Cornell University)*, 33–37. https://doi.org/10.1145/3578337.3605138

[23] 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. https://doi.org/10.1038/s42256-020-00291-x

[24] Kingma, D. P., & Ba, J. (2014, December 22). Adam: A Method for Stochastic Optimization. *arXiv:1412.6980*. https://doi.org/10.48550/arXiv.1412.6980

[25]Zhang, H., Cisse, M., Dauphin, Y. N., & Lopez-Paz, D. (2018). mixup: Beyond Empirical Risk Minimization. *arXiv:1710.09412*. https://doi.org/10.48550/arXiv.1710.09412

[26] Verma, V., Lamb, A., Beckham, C., Najafi, A., Mitliagkas, I., Courville, A., López-Paz, D., & Bengio, Y. (2018). Manifold Mixup: Better Representations by Interpolating Hidden States. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1806.05236

[27] Cohen, J. M., Rosenfeld, E., & Kolter, J. Z. (2019). Certified Adversarial Robustness via Randomized Smoothing. *arXiv:1902.02918*. https://doi.org/10.48550/arXiv.1902.02918

[28] Yoshida, Y., & Miyato, T. (2017, May 31). Spectral Norm Regularization for Improving the Generalizability of Deep Learning. *arXiv:1705.10941* https://doi.org/10.48550/arXiv.1705.10941

[29] Toneyan, S., Tang, Z., & Koo, P. K. (2022). Evaluating deep learning for predicting epigenomic profiles. *Nature Machine Intelligence*, *4*(12), 1088–1100. https://doi.org/10.1038/s42256-022-00570-9

[30] ENCODE Project Consortium. (2004). The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*, *306*(5696), 636–640. https://doi.org/10.1126/science.1105136

[31] Kelley, D. R., Reshef, Y. A., Bileschi, M., Belanger, D., McLean, C. Y., & Snoek, J. (2018). Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Research*, *28*(5), 739–750. https://doi.org/10.1101/gr.227819.117

[32] Koo, P. K., Ploenzke, M., Anand, P., Paul, S., & Majdandzic, A. (2023). ResidualBind: Uncovering Sequence-Structure Preferences of RNA-Binding Proteins with Deep Neural Networks. *Methods in Molecular Biology (Clifton, N.J.)*, *2586*, 197–215. https://doi.org/10.1007/978-1-0716-2768-6_12

[33] Zrimec, J., Börlin, C. S., Buric, F., Muhammad, A. S., Chen, R., Siewers, V., Verendel, V., Nielsen, J., Töpel, M., & Zelezniak, A. (2020). Deep learning suggests that gene expression is encoded in all parts of a co-evolving interacting gene regulatory structure. *Nature Communications*, *11*(1), 6141. https://doi.org/10.1038/s41467-020-19921-4