ISAAC: Intervention-Based Structural Auditing of Deep Models for Transcription Factor Binding

Barbara Tarantino[1,*], Paolo Giudici[1]

[1]University of Pavia, Pavia, Italy

Abstract

Transcription factor (TF) binding prediction is the problem of determining, from raw DNA sequence, whether a genomic region is recognised and bound by a specific regulatory protein. Deep learning models achieve high predictive accuracy on this task, yet their behavioural response to biologically motivated perturbations remains largely unvalidated. Predictive accuracy alone does not establish whether model outputs reflect genuine regulatory signals or arise from spurious sequence correlations, which limits the reliability of current explainable AI (XAI) approaches in scientific domains.

We introduce *ISAAC* (Intervention-based Structural Auditing for Computational Models), a post-hoc framework that evaluates model behaviour through controlled, sequence-level interventions grounded in position weight matrices (PWMs). PWMs are probabilistic representations of TF binding preferences derived from experimental data; high affinity of a sequence window to a PWM indicates compatibility with the known binding preference of the factor. ISAAC generates matched structure-aligned and structure-violating perturbations and summarises the resulting response distributions along three complementary axes: distributional separability (Entanglement, $\mathcal{E}$), effect magnitude (Collapse, $\mathcal{C}$), and contrast stability (Instability, $\mathcal{I}$).

We apply ISAAC to DeepBind, DeepSEA, and BPNet on ENCODE TF binding datasets. In predictive regimes where accuracy is sufficient, the audited models exhibit response patterns suggesting systematic misalignment with the regulatory reference. These findings support structural auditing as a necessary complement to accuracy and attribution in XAI for genomics.

Keywords

Explainable AI, Structural Auditing, Behavioural Evaluation, Genomic Deep Learning, Transcription Factor Binding

1. Introduction

Transcription factor (TF) binding prediction is the problem of classifying short DNA sequences as bound or unbound by a specific regulatory protein, based on the presence of characteristic sequence patterns known as motifs. Motifs are commonly represented as position weight matrices (PWMs), which encode the nucleotide frequencies observed at each position across experimentally characterised binding sites as log-odds scores. High affinity of a sequence window to a PWM therefore indicates compatibility with the known binding preference of the factor.

Deep learning models have become the dominant approach for TF binding prediction and achieve strong performance across a wide range of benchmarks [1, 2, 3]. These architectures map raw DNA sequences to binding signals and frequently outperform classical motif-based methods. Nonetheless, strong predictive performance does not imply that a model relies on biologically meaningful regulatory mechanisms to generate its predictions.

Predictive evaluation provides only an observational characterisation of model behaviour and does not reveal how predictions are produced. Models with similar accuracy may rely on different input dependencies or internal strategies, a point formalised in veridical data science [4] and extensively documented in shortcut learning research [5]. This concern is particularly critical in regulatory genomics, where TF binding is governed by sequence context, combinatorial interactions, and partially stochastic

Late-breaking work, Demos and Doctoral Consortium, colocated with the 4th World Conference on eXplainable Artificial Intelligence: July 01–03, 2026, Fortaleza, Brazil

*Corresponding author.

barbara.tarantino@unipv.it (B. Tarantino)

0000-0001-6384-2561 (B. Tarantino); 0000-0002-4198-0127 (P. Giudici)

mechanisms [6]. Biological knowledge constrains how perturbations of a DNA sequence should influence binding predictions without uniquely determining model outputs, which motivates evaluating whether model responses are consistent with that knowledge rather than merely well-calibrated on a held-out test set.

Large-scale genomic datasets such as ENCODE intensify this concern. These resources contain highly correlated sequence data in which spurious patterns and experimental artifacts can remain predictive in-distribution [7, 8], and standard train-test evaluation is insufficient to expose them [9].

Most explainability approaches for genomic deep learning rely on post-hoc attribution methods that assign relevance scores to individual input positions [10, 11]. While attribution maps are useful for qualitative inspection, they do not directly test whether model behaviour is consistent with external biological knowledge under structured perturbations [12]. Recent work in XAI has accordingly emphasised the value of behavioural evaluation under controlled interventions as a complementary paradigm [13, 14].

We introduce *ISAAC*, a post-hoc, model-agnostic auditing framework that evaluates whether model responses to controlled perturbations are consistent with an explicit regulatory reference, operating exclusively on observable input-output behaviour without requiring retraining.

2. Related Work

Explainability research in genomic deep learning has largely focused on post-hoc attribution methods such as saliency maps, integrated gradients, and DeepLIFT [10, 11, 15]. These methods are sensitive to implementation choices and may fail sanity checks that test their dependence on learned model parameters [12]. Evaluation frameworks such as Quantus [16] assess attribution properties including faithfulness, robustness, and randomisation stability; ISAAC addresses related questions, but at the level of interventional response distributions rather than attribution scores.

Beyond attribution, perturbation-based methods evaluate models by directly modifying input sequences. In genomics, *in silico* mutagenesis (ISM) and its efficient implementation [17] systematically alter individual nucleotide positions to assess local sensitivity, and BPNet [18] extended this approach with co-mutation analyses to infer motif syntax. These methods are widely used but typically without a structural prior that partitions perturbations into reference-aligned and reference-violating classes for distributional comparison [5, 8].

ISAAC builds on both lines of work while addressing their limitations. Attribution methods characterise local feature relevance without testing consistency with an external reference; ISM probes sensitivity without organising perturbations around a structural prior. A closer analogue is TCAV [19], which evaluates model behaviour against user-defined concepts through directional derivatives in internal activation space. ISAAC follows the same principle but operates exclusively on observable input-output behaviour, replacing abstract concepts with a biologically grounded intervention space derived from an explicit regulatory prior. This makes the auditing procedure model-agnostic and directly interpretable in terms of domain knowledge, without requiring access to internal model components.

3. ISAAC Framework

3.1. Problem Setting

Let $x = (x_1, \ldots, x_L) \in \mathcal{X}$ be a DNA sequence of length L over the alphabet $\mathcal{A} = \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$, and let $f_\theta : \mathcal{X} \to \mathbb{R}$ be a trained model producing a real-valued binding score, with parameters θ held fixed throughout. Domain knowledge is encoded as a PWM from a curated database such as JASPAR [20], which assigns to each candidate binding window an affinity score

$$A(x, s) = \sum_{j=0}^{L_m - 1} \mathrm{PWM}[x_{s+j},\, j],$$

where s is the start position and L_m the motif length. High values of $A(x, s)$ indicate compatibility with the known binding preference of the factor.

3.2. Intervention Design

ISAAC evaluates model behaviour by comparing responses to two complementary classes of sequence perturbations. Interventions $\delta : \mathcal{X} \to \mathcal{X}$ preserve sequence length and alphabet and modify only a local window. Structure-aligned interventions target positions with high PWM affinity; non-structural interventions target positions outside the regulatory reference.

Formally, structurally relevant positions are $\mathcal{S}_{\text{struct}}(x) = \{s \mid A(x, s) \geq q_\alpha(x)\}$, where $q_\alpha(x)$ is the $(1 - \alpha)$ empirical quantile with $\alpha = 0.05$. Structure-aligned interventions act on the highest-affinity site $s^\star = \arg\max_s A(x, s)$ through three operators of increasing severity: Precision Weakening (PW), which replaces the highest-contributing nucleotide with its least compatible alternative; Structural Scrambling (SS), which permutes the most affinity-contributing positions; and Complete Knockout (CK), which replaces all motif positions with their anti-consensus bases. Each operator guarantees $A(\delta(x), s^\star) - A(x, s^\star) < 0$ by construction. Non-structural interventions act on windows of identical length outside $\mathcal{S}_{\text{struct}}(x)$ through matched random mutations, random scrambling, and GC-content-preserving substitutions, with positions selected independently of the regulatory prior.

For each intervention δ, the model response is $\Delta f_\theta(x, \delta) = f_\theta(\delta(x)) - f_\theta(x)$, and auditing is based on the empirical distributions $\mathcal{D}_\theta(\mathcal{J}) = \{\Delta f_\theta(x, \delta) : x \sim P_\mathcal{X},\ \delta \in \mathcal{J}\}$ across the two intervention classes.

3.3. Behavioral Diagnostics

The intervention design yields two empirical response distributions,

$$\mathcal{D}_\theta^{\text{struct}} = \{\Delta f_\theta(x, \delta) : \delta \in \mathcal{I}_{\text{struct}}\}, \qquad \mathcal{D}_\theta^{\text{nonstruct}} = \{\Delta f_\theta(x, \delta) : \delta \in \mathcal{I}_{\text{nonstruct}}\}.$$

ISAAC summarises their relationship through three diagnostics, each with an explicit, bounded formulation (Table 1).

Entanglement ($\mathcal{E}$). Entanglement quantifies distributional overlap between the two response distributions using the Overlapping Coefficient [21]:

$$\mathcal{E} = \text{OVL}\big(\mathcal{D}_\theta^{\text{struct}},\, \mathcal{D}_\theta^{\text{nonstruct}}\big) = \int \min(p(z),\, q(z))\, dz,$$

where p and q denote the density functions of $\mathcal{D}_\theta^{\text{struct}}$ and $\mathcal{D}_\theta^{\text{nonstruct}}$ respectively, estimated empirically via adaptive histogram binning. Since $\mathcal{E}$ is the integral of the pointwise minimum of two probability densities, it lies in $[0, 1]$ by construction. Values near zero indicate well-separated distributions; values near one indicate near-complete overlap.

Collapse ($\mathcal{C}$). Collapse summarises the magnitude of the mean response contrast through an exponential transformation of Cohen's d [22]:

$$d = \frac{|\mu_{\text{struct}} - \mu_{\text{nonstruct}}|}{\sigma_{\text{pooled}}}, \qquad \mathcal{C} = e^{-d},$$

where μ_{struct} and $\mu_{\text{nonstruct}}$ are the empirical means and σ_{pooled} the pooled standard deviation. Since $d \geq 0$, $\mathcal{C} \in (0, 1]$, with values near zero indicating large standardised separation and values near one reflecting negligible average contrast.

Table 1
ISAAC behavioral diagnostics. Lower values indicate better structural alignment.

Axis	*Metric*	*Formula*	*Range*
Separability	Entanglement $\mathcal{E}$	$\mathrm{OVL}(p, q)$	$[0, 1]$
Effect size	Collapse $\mathcal{C}$	e^{-d}	$(0, 1]$
Stability	Instability $\mathcal{I}$	$\mathrm{CV}^2/(1 + \mathrm{CV}^2)$	$[0, 1)$

Instability ($\mathcal{I}$). Instability measures the relative variability of paired intervention contrasts, defined for each input as $\Delta(x) = \Delta f_\theta(x, \delta_{\text{struct}}) - \Delta f_\theta(x, \delta_{\text{nonstruct}})$, via a bounded transformation of the coefficient of variation [23]:

$$\mathrm{CV} = \frac{\sigma_\Delta}{\mathbb{E}|\Delta|}, \qquad \mathcal{I} = \frac{\mathrm{CV}^2}{1 + \mathrm{CV}^2},$$

where σ_Δ is the standard deviation of Δ and $\mathbb{E}|\Delta|$ its mean absolute value. Since $t \mapsto t/(1 + t)$ is strictly increasing on $[0, +\infty)$ and bounded above by one, $\mathcal{I} \in [0, 1)$ by construction. All three metrics share the convention that lower values correspond to better structural alignment with the regulatory reference.

4. Experimental Design

Datasets. We apply ISAAC to TF binding prediction using ENCODE ChIP-seq datasets [7] from three human cell lines: A549, GM12878, and HepG2. Data construction and preprocessing follow the protocol described in [24]. Each dataset comprises approximately 5×10^5 DNA sequences of length $L = 576$ bp with binary labels. Sequences are partitioned into training (80%) and held-out sets (20%) by stratified sampling to preserve class balance, in accordance with the benchmark protocol. This split is used solely to establish a predictive validity gate; the structural diagnostics are computed post-hoc on frozen models and are therefore independent of how training data were partitioned. From the training split, 10^5 sequences are sampled for model training, fixing the size to ensure a comparable predictive setting across architectures.

Models and Training. Three representative architectures are audited: DeepBind [1], DeepSEA [2], and BPNet [18]. Each is implemented in PyTorch following an architecture-faithful configuration that preserves the defining structural properties of the original design: convolutional motif detectors in DeepBind, deep convolutional feature hierarchies in DeepSEA, and dilated convolutional residual blocks in BPNet. Each architecture is adapted to the binary classification setting and produces a single real-valued output score per input sequence. All models are trained under a shared protocol: Adam optimiser with learning rate 10^{-3}, batch size 64, binary cross-entropy loss, and 20 epochs, repeated over five independent random seeds. This shared protocol is a deliberate choice to ensure that differences in structural diagnostics are attributable to architectural properties rather than to training configuration disparities. After training, all models are frozen and treated as deterministic operators during auditing.

DeepSEA is included in the training protocol under the same controlled configuration but does not reach a stable predictive threshold ($\mathrm{AUROC} > 0.6$) in any of the three cell lines. It is therefore excluded from structural auditing and reported separately for completeness. This exclusion reflects the behaviour of DeepSEA under this specific controlled configuration, which may place architectures with a larger parameter budget at a disadvantage, and should not be interpreted as a general statement about the architecture.

Regulatory Reference. Structure-aligned interventions are defined using the CTCF PWM from JASPAR (ID: MA0139.1) [20]. CTCF is one of the most thoroughly characterised human transcription factors, with a well-defined canonical binding motif that shows relatively limited cell-type specificity,

Table 2
Predictive performance (AUROC), reported as mean ± standard deviation over five independent training seeds. DeepSEA is excluded from structural auditing under the controlled training regime described in the text.

Cell line	*Model*	*AUROC*
A549	BPNet	0.912 ± 0.001
	DeepBind	0.702 ± 0.021
	DeepSEA	0.576 ± 0.170
GM12878	BPNet	0.898 ± 0.002
	DeepBind	0.721 ± 0.017
	DeepSEA	0.573 ± 0.164
HepG2	BPNet	0.919 ± 0.001
	DeepBind	0.702 ± 0.017
	DeepSEA	0.500 ± 0.000

making it a methodologically conservative and appropriate reference for validating the auditing procedure. The PWM is used exclusively to guide intervention placement and does not enter the training objective.

Auditing Protocol. Structural auditing is conducted post-hoc on frozen models through repeated balanced subsampling of the held-out set. For each combination of dataset, architecture, and training seed, ten balanced subsamples of 5,000 sequences are drawn, yielding 5×10 auditing configurations per dataset-architecture pair. ISAAC diagnostics are computed for each configuration and summarised as means with 95% percentile confidence intervals across subsamples, controlling for variability arising from training stochasticity and data subsampling.

Intervention Validation. The effectiveness and specificity of the intervention design are verified independently of the predictive models across $n = 50{,}000$ intervention instances. Structural interventions induce a consistent and substantial reduction in PWM affinity ($\Delta A = -3.97 \pm 0.11$ log-odds), whereas non-structural interventions produce negligible changes (-0.07 ± 0.02). Non-structural interventions produce large Hamming distances from the original sequence (mean 8.10 ± 4.32 nt over 20 bp windows), and only 1.7% of instances inadvertently create new motif-compatible sites while 3.6% disrupt existing ones, confirming that the two intervention families are well-controlled and structurally distinct.

5. Results

5.1. Predictive Validity

BPNet achieves strong and stable predictive performance across all three cell lines (AUROC $= 0.912 \pm 0.001$ in A549, 0.898 ± 0.002 in GM12878, and 0.919 ± 0.001 in HepG2, aggregated over five seeds). DeepBind operates in a lower but reproducible regime across the same datasets (AUROC from 0.702 ± 0.021 to 0.721 ± 0.017). DeepSEA does not meet the predictive threshold under the present training configuration and is consequently excluded from structural auditing (Table 2).

5.2. Structural Auditing

Structural auditing results are reported in Table 3 and Fig. 1. All metrics are interpreted on bounded $[0, 1]$ scales (see Table 1 for exact ranges), in which lower values indicate better alignment with the regulatory reference.

DeepBind exhibits consistently high entanglement ($\mathcal{E} = 0.925$–0.932) and collapse ($\mathcal{C} = 0.933$–0.953) across all three cell lines, indicating that model responses to structure-aligned and non-structural perturbations are substantially overlapping and only weakly differentiated in magnitude. Instability

Table 3
ISAAC auditing diagnostics, reported as mean with 95% confidence intervals aggregated over training seeds and balanced subsamples. Lower values indicate better structural alignment for all three metrics.

Cell line	*Model*	*Entanglement*	*Collapse*	*Instability*
A549	BPNet	0.881 [0.877, 0.884]	0.867 [0.853, 0.881]	0.695 [0.688, 0.702]
	DeepBind	0.925 [0.910, 0.941]	0.953 [0.945, 0.961]	0.853 [0.846, 0.860]
GM12878	BPNet	0.887 [0.884, 0.889]	0.941 [0.917, 0.965]	0.667 [0.662, 0.673]
	DeepBind	0.929 [0.920, 0.939]	0.933 [0.910, 0.956]	0.853 [0.836, 0.870]
HepG2	BPNet	0.884 [0.881, 0.887]	0.884 [0.872, 0.896]	0.679 [0.676, 0.681]
	DeepBind	0.932 [0.913, 0.952]	0.941 [0.912, 0.971]	0.871 [0.848, 0.894]

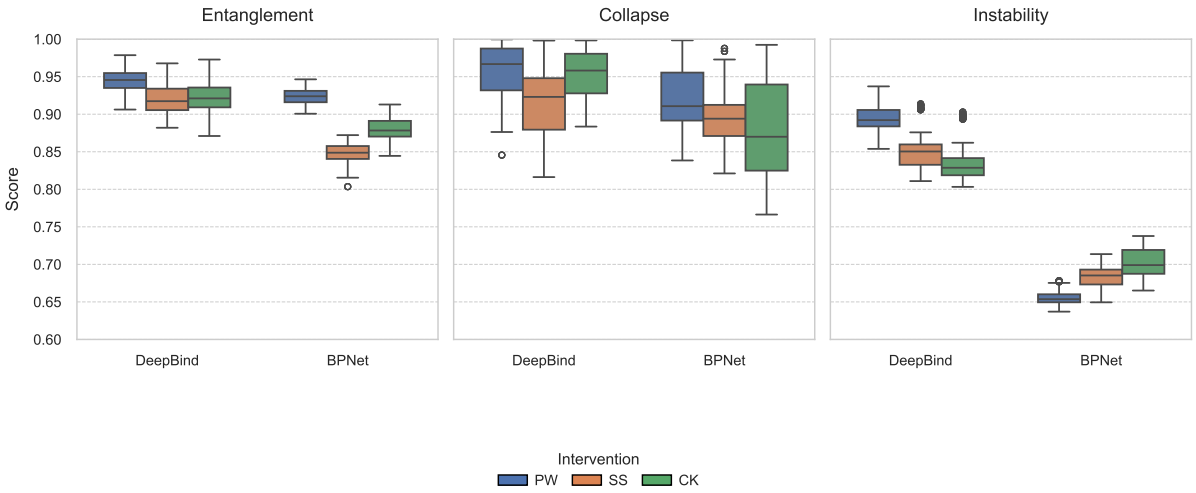

Figure 1: Distribution of ISAAC diagnostics across models, aggregated over datasets and training seeds for architectures operating in a valid predictive regime (AUROC > 0.6). Lower values indicate better structural alignment. Both audited architectures exhibit systematic failure profiles despite achieving adequate predictive performance.

values are similarly elevated ($\mathcal{I} = 0.853$–0.871), reflecting strongly input-dependent expression of any residual regulatory sensitivity.

BPNet achieves moderately lower entanglement ($\mathcal{E} = 0.881$–0.887) and instability ($\mathcal{I} = 0.667$–0.695) relative to DeepBind, suggesting somewhat more regularised responses across inputs. However, collapse remains high in GM12878 ($\mathcal{C} = 0.941$), and all three diagnostics remain well above the low-failure regime for both architectures.

These patterns persist across all three intervention operators (PW, SS, CK), including the most severe perturbations produced by complete motif knockouts. Differences between architectures emerge primarily in response dispersion, while distributional overlap and average contrast remain broadly similar across operators. The consistency of these patterns suggests that the observed failure profiles are structural in nature rather than specific to any particular perturbation type.

6. Implications and Limitations

The results suggest that models achieving adequate predictive accuracy may nonetheless produce response distributions in which structure-aligned and non-structural perturbations are only weakly distinguished, indicating that predictive performance does not certify behavioural alignment with the specified regulatory reference.

From a methodological perspective, ISAAC differs from attribution-based approaches in the question it addresses: rather than assigning relevance to individual input positions, it evaluates how model

responses are organised across matched structural and non-structural intervention classes. The two perspectives address distinct aspects of model behaviour and are naturally complementary: locally plausible attributions may coexist with globally weak behavioural alignment to the regulatory reference, precisely the gap between plausibility and faithfulness identified in the XAI literature [25].

Because ISAAC operates solely on observable input-output behaviour and requires only an externally specified reference structure, it is applicable in principle to any scientific domain where structural knowledge can be stated independently of the learning algorithm.

The analysis uses CTCF as a well-characterised regulatory reference in a controlled experimental setting; results should therefore be interpreted as evidence of behavioural alignment with respect to this specific prior, rather than as general claims about mechanistic correctness. ISAAC does not ask whether a model recovers the PWM, but whether its learned behaviour is consistent with the structural constraints the PWM encodes. The validity of the diagnostics is conditioned on the quality of the external regulatory prior, and extensions to additional transcription factors would further assess the generality of the observed failure profiles. Consistency across intervention operators and training seeds is adopted as the primary indicator of structural alignment because it provides a model-agnostic, distribution-level signal that requires no access to internal model components and is directly interpretable in terms of domain knowledge; the reported confidence intervals provide an empirical characterisation of estimate stability. Extensions to additional TF priors, Transformer-based architectures, and regression settings represent natural directions for follow-up work.

Overall, these findings support intervention-based behavioural auditing as a principled complement to predictive evaluation and attribution analyses, advancing the goals of trustworthy AI in scientific domains where explicit prior knowledge makes structured behavioural evaluation both feasible and necessary. Code and processed datasets are publicly available online.[1][2].

Acknowledgments

We acknowledge funding from the Italian PRIN2020 project FIN4GREEN - Finance for a sustainable, green and resilient society.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] B. Alipanahi, A. Delong, M. T. Weirauch, B. J. Frey, Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning, Nature Biotechnology 33 (2015) 831–838.

[2] J. Zhou, O. Troyanskaya, Predicting effects of noncoding variants with deep learning–based sequence model, Nature Methods 12 (2015) 931–934.

[3] Y. Ji, Z. Zhou, H. Liu, R. V. Davuluri, DNABERT: pre-trained bidirectional encoder representations from transformers model for DNA-language in genome, Bioinformatics 37 (2021) 2112–2120.

[4] B. Yu, K. Kumbier, Veridical data science, Proceedings of the National Academy of Sciences of the United States of America 117 (2020) 3920–3929.

[5] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. S. Zemel, W. Brendel, M. Bethge, F. A. Wichmann, Shortcut learning in deep neural networks, Nature Machine Intelligence 2 (2020) 665–673.

[6] S. A. Lambert, A. Jolma, L. F. Campitelli, et al., The human transcription factors, Cell 172 (2018) 650–665.

[1] https://github.com/BarbaraTarantino/isaac-xai-tf

[2] https://osf.io/k73td/

[7] ENCODE Project Consortium, An integrated encyclopedia of DNA elements in the human genome, Nature 489 (2012) 57–74.
[8] A. J. DeGrave, J. D. Janizek, S.-I. Lee, AI for radiographic COVID-19 detection selects shortcuts over signal, Nature Machine Intelligence 3 (2021) 610–619.
[9] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, K.-R. Müller, Unmasking clever hans predictors and assessing what machines really learn, Nature Communications 10 (2019) 1096.
[10] K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps, in: Workshop at the 2nd International Conference on Learning Representations (ICLR), 2014.
[11] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: Proceedings of the 34th International Conference on Machine Learning (ICML), volume 70, 2017, pp. 3319–3328.
[12] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, B. Kim, Sanity checks for saliency maps, in: Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS), 2018, pp. 9525–9536.
[13] R. C. Fong, A. Vedaldi, Interpretable explanations of black boxes by meaningful perturbation, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3449–3457.
[14] A. Dhurandhar, S. Haldar, D. Wei, K. N. Ramamurthy, Trust regions for explanations via black-box probabilistic certification, in: Proceedings of the 41st International Conference on Machine Learning (ICML), 2024, pp. 10736–10764.
[15] A. Shrikumar, P. Greenside, A. Kundaje, Learning important features through propagating activation differences, in: Proceedings of the 34th International Conference on Machine Learning (ICML), volume 70, 2017, pp. 3145–3153.
[16] A. Hedström, L. Weber, D. Krauss, D. Bareeva, F. Motzkus, W. Samek, S. Lapuschkin, M. Hägele, Quantus: An explainable ai toolkit for responsible evaluation of neural network explanations and beyond, Journal of Machine Learning Research 24 (2023) 1–11.
[17] S. Nair, A. Shrikumar, J. Schreiber, A. Kundaje, fastism: performant in silico saturation mutagenesis for convolutional neural networks, Bioinformatics 38 (2022) 2397–2403.
[18] Ž. Avsec, M. Weilert, A. Shrikumar, S. Krueger, A. Alexandari, K. Dalal, R. Fropf, C. McAnany, J. Gagneur, A. Kundaje, J. Zeitlinger, Base-resolution models of transcription-factor binding reveal soft motif syntax, Nature Genetics 53 (2021) 354–366.
[19] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, R. Sayres, Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV), in: Proceedings of the 35th International Conference on Machine Learning (ICML), 2018, pp. 2668–2677.
[20] O. Fornes, J. A. Castro-Mondragon, A. Khan, R. van der Lee, X. Zhang, P. A. Richmond, B. P. Modi, S. Correard, M. Gheorghe, D. Baranašić, et al., JASPAR 2020: update of the open-access database of transcription factor binding profiles, Nucleic Acids Research 48 (2020) D87–D92.
[21] H. F. Inman, E. L. Bradley, The overlapping coefficient as a measure of agreement between probability distributions and point estimation of the overlap of two normal densities, Communications in Statistics – Theory and Methods 18 (1989) 3851–3874.
[22] J. Cohen, Statistical Power Analysis for the Behavioral Sciences, 2 ed., Routledge, New York, 1988.
[23] B. S. Everitt, The Cambridge Dictionary of Statistics, 2 ed., Cambridge University Press, Cambridge, 2005.
[24] N. Ghosh, D. Santoni, I. Saha, G. Felici, Predicting transcription factor binding sites with deep learning, International Journal of Molecular Sciences 25 (2024) 4990.
[25] A. Jacovi, Y. Goldberg, Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness?, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020, pp. 4198–4205.