Interpretable Flavivirus Classification under Low-Coverage Conditions via ERF-Aware k-mer Attribution

Bahar İlgen[1,*], Afsheen Shahid[1] and Denise Kühnert[1]

[1]Centre for Artificial Intelligence in Public Health Research (ZKI-PH), Robert Koch Institute, Nordufer 20, Berlin, 13353, Germany

Abstract

Accurate classification of flaviviruses from genomic sequences is essential for outbreak surveillance and pathogen monitoring, particularly under real-world sequencing conditions where incomplete genomes and ambiguous nucleotide characters are common. While deep learning models have demonstrated strong performance in genomic classification tasks, their decision-making processes often remain difficult to interpret, limiting transparency and reliability in biomedical applications. In this work, we extend DiCNN-UniK, a dual-input convolutional neural network based on k-mer sequence embeddings, with an activation-based interpretability framework for identifying class-associated genomic sequence patterns. The proposed approach traces high-activation convolutional responses back to the original k-mer tokens contributing to classification decisions, enabling inspection of sequence patterns associated with individual flavivirus classes. In contrast to purely model-agnostic post-hoc explainability approaches, the framework leverages the internal architecture of the network to identify activation-associated ERF-window patterns linked to model predictions. Attribution analysis identified class-associated ERF-window signatures across ten flavivirus classes, with discriminative scoring revealing selectively activated patterns for individual classes. Anchor-level attribution signatures demonstrated greater robustness under simulated nucleotide ambiguity than exact ERF-window matching. Perturbation-based validation using logit-margin sensitivity demonstrated prediction-relevant attribution effects in all ten flavivirus classes. By combining flavivirus classification with architecture-aware attribution analysis, this study provides interpretable and perturbation-sensitive sequence-level explanations for flavivirus discrimination under low-coverage conditions.

Keywords

Explainable Artificial Intelligence, Flavivirus Classification, Alignment-Free Genomics, k-mer Representation, Effective Receptive Field, Model Interpretability, Attribution Methods, Low-Coverage Sequencing

CEUR-WS.org/Vol-4339/paper_8.pdf

1. Introduction

Flaviviruses are responsible for recurring vector-borne outbreaks, including dengue and Zika epidemics, creating an ongoing need for accurate genomic surveillance and rapid sequence-based classification methods [1, 2]. However, real-world sequencing data frequently contain incomplete genomic coverage, sequencing ambiguity, and fragmented reads, creating substantial challenges for reliable virus classification. Traditional alignment-based methods, such as multiple sequence alignment (MSA) and phylogenetic reconstruction, are computationally expensive and sensitive to incomplete or noisy sequence data [3, 4], limiting their scalability for high-throughput genomic surveillance applications.

Recent advances in deep learning and alignment-free genomic analysis [5] have enabled efficient classification directly from raw nucleotide sequences using k-mer representations and neural architectures. In particular, convolutional neural networks (CNNs) provide an effective framework for extracting local genomic sequence patterns from encoded k-mer embeddings. DiCNN-UniK is a dual-input CNN architecture utilizing 5-mer and 6-mer embeddings derived from universal k-mer libraries for robust flavivirus classification under low-coverage sequencing conditions, demonstrating high classification accuracy while maintaining strong performance with incomplete and ambiguous genomic sequences [6].

EXPLIMED 2026 - Third Workshop on Explainable Artificial Intelligence for the medical domain - 15-17 August 2026, Bremen, Germany

*Corresponding author.

ilgenb@rki.de (B. İlgen); kuehnertd@rki.de (D. Kühnert)

0000-0001-5725-0850 (B. İlgen); 0009-0007-3049-1751 (A. Shahid); 0000-0002-5657-018X (D. Kühnert)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2026-11-26

Despite these advances, the internal decision-making processes of genomic deep learning models remain difficult to interpret. Most existing explainability approaches in genomics rely on generic post-hoc techniques such as saliency maps, Integrated Gradients, SHAP, or perturbation-based analyses [7, 8, 9]. While these methods provide useful feature-level importance estimates, they are not explicitly designed to exploit the receptive field structure of k-mer-based convolutional architectures [10]. As a result, the sequence patterns associated with classification decisions often remain difficult to interpret, particularly under low-coverage conditions.

In k-mer-based convolutional models, attribution based on individual k-mers may become unstable under nucleotide ambiguity. ERF-window attribution instead provides a context-aware interpretation of locally distributed motif representations. In this work, we extend DiCNN-UniK with an ERF-aware activation attribution framework designed specifically for k-mer convolutional architectures [6]. Our approach traces high-activation convolutional responses back to their originating k-mer windows within the effective receptive field, enabling inspection of activation-associated sequence patterns linked to model predictions [11]. We further evaluate the stability of these attributed patterns under simulated nucleotide ambiguity and investigate their perturbation sensitivity through occlusion-based validation.

Unlike generic post-hoc attribution pipelines, the proposed framework directly associates convolutional activations selected by GlobalMaxPooling with ERF-aware k-mer windows, while evaluating attribution stability and prediction relevance under ambiguity-corrupted sequencing conditions. This enables interpretation of CNN-based flavivirus classifiers under degraded sequencing scenarios by linking high-activation ERF windows to class-enriched k-mer signatures. This transparent and perturbation-grounded interpretation of genomic deep learning models may support trustworthy AI-assisted surveillance workflows in outbreak monitoring and public health genomics. The contributions of this work are as follows:

- An ERF-aware attribution framework specifically designed for k-mer convolutional classifiers, providing attribution directly aligned with the model's internal convolutional selection mechanism.
- A discriminative scoring mechanism for identifying class-enriched k-mer signatures across flavivirus classes.
- A robustness analysis under simulated nucleotide ambiguity, evaluating attribution stability under degraded sequencing conditions.
- Perturbation-grounded validation using logit-margin sensitivity, providing evidence for the prediction relevance of identified signatures.

2. Background

2.1. Alignment-Free Genomic Classification

Accurate genomic classification is a fundamental task in pathogen surveillance, outbreak monitoring, and infectious disease research. Traditional genomic classification pipelines often rely on multiple sequence alignment (MSA) and phylogenetic reconstruction methods to identify evolutionary relationships between viral genomes [12, 13]. Although these approaches provide biologically interpretable comparisons, they become computationally expensive for large-scale genomic datasets and are highly sensitive to incomplete or noisy sequence data. In practical sequencing environments, partial genomes, sequencing ambiguities, and fragmented reads frequently reduce the reliability of alignment-based methods [14, 15].

To address these limitations, alignment-free approaches based on k-mer representations have become increasingly popular for genomic analysis and classification. In these methods, genomic sequences are decomposed into overlapping substrings of fixed length k, enabling direct numerical representation without explicit sequence alignment. k-mer-based approaches have demonstrated strong performance

in taxonomic classification, metagenomics, and viral detection tasks while substantially reducing computational complexity [5, 16].

Recent advances in deep learning have further expanded the utility of alignment-free genomic analysis. Convolutional neural networks (CNNs), recurrent architectures, and transformer-based foundation models have been applied to genomic sequence classification using nucleotide-level or k-mer-based tokenization strategies. CNN-based models are particularly effective for capturing local sequence motifs and hierarchical genomic patterns through convolutional filters operating on embedded sequence representations [17, 10]. However, despite their predictive success, many genomic deep learning models provide limited insight into the sequence patterns driving classification decisions, particularly under incomplete genomic coverage conditions.

2.2. Explainability in Genomic Deep Learning

The increasing adoption of deep learning methods in genomics has amplified the need for explainability and transparent model interpretation. In biomedical and public health applications, understanding the sequence patterns driving model predictions is essential for establishing trust, validating biological relevance, and identifying potential biases or shortcut learning behaviour.

Existing explainability approaches in genomic deep learning commonly rely on post-hoc feature attribution techniques such as saliency maps, Integrated Gradients, SHAP, or perturbation-based analyses [9, 8]. These methods aim to estimate the contribution of individual nucleotides, sequence regions, or learned embeddings to model predictions. In transformer-based genomic foundation models [18, 19], attention visualization has also been used to inspect long-range sequence dependencies and token interactions.

Despite their utility, many existing explainability methods remain largely model-agnostic and may not be specifically tailored to k-mer-based convolutional architectures [20]. As a result, the biological interpretation of attributed sequence regions often remains difficult, particularly when operating on incomplete genomes or ambiguous nucleotide sequences. Furthermore, attribution methods designed primarily for nucleotide-level transformer architectures may not fully capture the hierarchical motif representations learned by convolutional k-mer models [21].

These limitations are particularly relevant for alignment-free k-mer-based genomic classifiers, where convolutional filters learn distributed motif representations rather than explicit nucleotide-level dependencies. In such architectures, conventional attribution methods may not explicitly reveal biologically interpretable sequence signatures, especially in fragmented or mutation-prone viral genomes. These challenges motivate the development of architecture-aware interpretability frameworks capable of directly associating learned convolutional activations with conserved genomic motifs and taxonomically informative sequence patterns [21, 22].

2.3. Interpretable k-mer Representations

K-mers provide a natural intermediate representation between raw nucleotide sequences and higher-level genomic patterns. Because k-mers correspond to localized sequence units, they can serve as interpretable building blocks for identifying activation-associated sequence signatures linked to specific viral classes.

In convolutional genomic models, filters operating on embedded k-mer sequences can serve as local pattern detectors that respond strongly to recurring sequence motifs [23, 22]. When combined with GlobalMaxPooling operations, convolutional architectures effectively select the dominant activations corresponding to high-activation sequence regions. This architectural behaviour creates an opportunity to directly interpret the k-mer windows that contribute most strongly to model predictions.

Prior studies have suggested that rare or highly specific k-mer patterns may carry substantial class-associated information for fine-grained genomic classification tasks [24]. This observation is conceptually analogous to linguistic frequency distributions, where rare words can encode more specific

contextual information than highly frequent terms. Within genomic sequences, unique or low-frequency k-mers may similarly function as candidate signatures distinguishing closely related viral classes.

Building on this perspective, ERF-aware k-mer attribution frameworks provide a mechanism for identifying interpretable sequence-level patterns linked to model predictions. By tracing high-activation convolutional responses back to their originating k-mer windows within the effective receptive field [11], such frameworks enable inspection of activation-associated sequence patterns. They further enable assessment of the stability of these patterns under varying sequencing conditions.

3. Methods

3.1. Dataset and Preprocessing

The dataset and preprocessing pipeline follow the original DiCNN-UniK study [6]. Genomic sequence data for ten flavivirus classes were obtained from the NCBI Virus database, comprising four Dengue serotypes (DENV-1 through DENV-4) and six viruses known to circulate in the European region: Yellow Fever virus, Zika virus, West Nile virus, Usutu virus, Tick-Borne Encephalitis virus, and Japanese Encephalitis virus. Following the original DiCNN-UniK preprocessing protocol, sequences with a genomic coverage of at least 95% and composed exclusively of unambiguous nucleotide characters (A, T, G, C) were used for model development. The original study employed an 80/20 train-test split and reported 1,669 internal test samples. External validation datasets were curated separately at reduced genomic coverages of 20%, 50%, and 70% and included sequences containing IUPAC ambiguity codes. Near-duplicate sequences between training and validation sets were identified and removed using fuzzy string matching with a 95% similarity threshold to prevent data leakage.

3.2. DiCNN-UniK Architecture

DiCNN-UniK is a dual-input convolutional neural network for alignment-free flavivirus classification based on universal k-mer libraries. The model accepts two parallel input branches corresponding to k-mer sizes $k = 5$ and $k = 6$, selected based on Zipfian distribution analysis of genomic vocabulary to balance unique and common k-mer content. For each branch, k-mers are generated using a sliding window of stride 1 over the input sequence and encoded as integers using a pre-defined universal k-mer library covering all 4^k possible k-mers ($|\Sigma_5| = 1024$, $|\Sigma_6| = 4096$).

Each branch independently maps integer-encoded k-mer sequences through a 128-dimensional embedding layer, followed by two parallel one-dimensional convolutional layers with kernel sizes $F \in \{3, 5\}$ and 256 filters each. Global max pooling is applied to each convolutional output to extract the dominant activations across the full sequence length. The pooled feature vectors from both kernel sizes are concatenated within each branch, and the resulting representations from both k-mer branches are merged into a unified feature space. Two fully connected layers (512 and 256 neurons) with ReLU activations and dropout regularization ($p = 0.5$) process the merged features, followed by a softmax output layer for ten-class classification. The convolutional receptive field spans approximately 7–10 nucleotides through the combination of k-mer size and convolutional kernel width, enabling multi-resolution coverage of local genomic patterns.

3.3. K-mer Activation Attribution Framework

To interpret the classification decisions of DiCNN-UniK, we developed an activation-based attribution framework that traces convolutional responses back to their originating k-mer tokens. The framework operates directly on the internal structure of the trained model without requiring modification of the architecture or training procedure.

3.3.1. Filter-Level Attribution

For each input sample, four attribution sub-models are constructed corresponding to the four (k, kernel) branch combinations: $(k = 5, F = 3)$, $(k = 5, F = 5)$, $(k = 6, F = 3)$, and $(k = 6, F = 5)$. Each sub-model shares the weights of the original DiCNN-UniK model and outputs the Conv1D activation tensor before the GlobalMaxPooling operation, with shape $(L_{\text{conv}}, n_{\text{filters}})$, where L_{conv} is the sequence length after convolution and $n_{\text{filters}} = 256$.

For each convolutional filter f, the position with maximum activation is identified:

$$p_f^* = \arg\max_p \; A[p, f] \tag{1}$$

where $A[p, f]$ denotes the activation at position p for filter f.

Because convolutional filters with kernel size F operate over adjacent k-mer tokens, attribution was extended from single-token attribution to receptive-field-aware window attribution. For each maximal activation position p_f^*, the corresponding window of F adjacent k-mers was retrieved from the encoded sequence, representing the receptive field associated with the filter response. Because GlobalMaxPooling retains only the strongest activation for each filter, the corresponding receptive-field window was treated as the primary sequence pattern contributing to that filter's downstream representation. Retrieved k-mer windows were subsequently mapped back to their nucleotide string representations using the tokenizer inverse index.

3.3.2. Sample Selection and Aggregation

To ensure that attribution analysis reflects reliable model behaviour, only correctly classified samples with a softmax confidence score of at least $\tau = 0.9$ were included in downstream aggregation.

For each selected sample, activation values associated with attributed k-mer windows were aggregated by virus class and convolutional branch. Rather than using cumulative activation sums, mean activation scores were computed to reduce bias arising from class imbalance between flavivirus categories.

For each class c, branch b, and attributed k-mer window m, the mean activation score was defined as:

$$S(c, b, m) = \frac{1}{N_m} \sum_{i=1}^{N_m} A_i^* \tag{2}$$

where N_m denotes the number of filter-level occurrences of k-mer window m across all selected samples and convolutional filters, and A_i^* denotes the maximal activation value selected by GlobalMaxPooling for occurrence i.

To reduce the influence of rare outlier activations, only k-mer windows with at least five filter-level occurrences across the selected samples were retained for downstream analysis. For each class and branch, k-mer windows were ranked according to their mean activation scores, and the highest-ranking windows were retained for downstream analysis.

3.3.3. Discriminative Scoring

Mean activation alone does not distinguish class-enriched k-mer signatures from windows that activate strongly but occur rarely or are shared across multiple flavivirus classes. To account for both activation strength and within-class prevalence, we first computed a frequency-weighted activation score for each class c, branch b, and k-mer window m:

$$W(c, b, m) = S(c, b, m) \cdot F(c, b, m) \tag{3}$$

where $S(c, b, m)$ is the mean activation score and $F(c, b, m)$ is the sample frequency of window m, defined as the fraction of selected samples from class c in which the window was observed at least once.

A frequency-aware discriminative score was then computed as:

$$D_{\log}(c, b, m) = \log_2 \left(\frac{W(c, b, m) + \varepsilon}{\bar{W}(\neg c, b, m) + \varepsilon} \right) \tag{4}$$

where $\bar{W}(\neg c, b, m)$ denotes the mean frequency-weighted activation of window m across all classes other than c, and $\varepsilon = 10^{-8}$ prevents division by zero. Positive values indicate that the window receives relatively stronger and more prevalent activation in class c than in the remaining flavivirus classes. Because scores may become large when a window is absent or nearly absent from non-target classes, discriminative scores were interpreted together with sample frequency rather than as standalone evidence of class specificity.

3.4. Attribution Stability Under Simulated Nucleotide Ambiguity

To evaluate the stability of attributed k-mer signatures under increasing levels of nucleotide ambiguity, we performed a perturbation-based analysis by introducing simulated ambiguity into test sequences. For each selected sample, a fraction $r \in \{0.0, 0.3, 0.5, 0.7, 0.8\}$ of nucleotide positions was randomly replaced with the ambiguous character N, corresponding to increasing levels of nucleotide ambiguity.

For each ambiguity level, the top-10 attributed k-mer windows identified from the clean sequence were compared to those from the ambiguity-masked sequence using Jaccard similarity:

$$J = \frac{|K_{\text{clean}} \cap K_{\text{masked}}|}{|K_{\text{clean}} \cup K_{\text{masked}}|} \tag{5}$$

A Jaccard similarity close to 1.0 indicates a high degree of overlap between attribution signatures extracted from the original and ambiguity-masked sequences, whereas lower values indicate increasing divergence between the resulting attribution sets.

Stability analysis was performed using stratified sampling across flavivirus classes. For each class, up to 50 correctly classified high-confidence samples ($\tau \geq 0.9$) were randomly selected to ensure balanced class representation during attribution stability evaluation.

3.5. Occlusion-Based Perturbation Validation

To evaluate whether highly attributed k-mer windows were prediction-relevant to model outputs, an occlusion-based perturbation analysis was performed. For each correctly classified high-confidence sample, the top attributed ERF-aware k-mer windows identified by the attribution framework were selectively masked at the nucleotide level by replacing the corresponding nucleotide positions with the ambiguous character N, followed by re-encoding of both k-mer branches. The resulting change in softmax confidence for the predicted class was then measured.

As a baseline comparison, random nucleotide spans of identical length were masked using the same procedure. Because softmax probabilities were highly saturated for correctly classified high-confidence samples, perturbation sensitivity was evaluated using logit-margin changes rather than raw probability differences. The logit margin was defined as the log-probability of the true class minus the maximum log-probability across all other classes. Attribution quality was evaluated by comparing the logit-margin reduction produced by attribution-guided masking against matched random masking. A positive attribution effect was defined as a larger logit-margin reduction following attribution-guided masking relative to random masking. Consistent attribution effects across flavivirus classes would indicate that the identified sequence regions contribute more strongly to model predictions than randomly selected sequence regions.

3.6. Implementation Details

The attribution framework was implemented in Python using TensorFlow 2.x and Keras. Attribution sub-models were constructed by extracting intermediate layer outputs from the trained DiCNN-UniK

model without modifying the original architecture or weights. Input sequence lengths for each k-mer branch were derived directly from the model input shapes to ensure consistency with the training configuration. To assess whether attribution scores might be influenced by extremely rare sequence patterns, a hapax legomenon analysis was performed. No hapax k-mers were observed for either $k = 5$ or $k = 6$, indicating that all possible k-mers within the corresponding vocabulary spaces ($4^5 = 1024$, $4^6 = 4096$) occurred at least once in the training corpus.

4. Results

4.1. Classification Performance

As the primary objective of this study is model interpretability rather than classifier development, the previously reported classification performance of DiCNN-UniK is reproduced to provide context for the attribution analyses. The original DiCNN-UniK study reported 99% classification accuracy and an AUC of 1.0 on an internal test dataset of 1,669 samples. External validation on datasets with reduced genomic coverage demonstrated strong generalisation performance, with accuracy remaining above 98% at 70% coverage and above 97% at 50% coverage. Performance declined at lower coverage levels but remained high across most flavivirus classes. In contrast, the HyenaDNA transfer model achieved comparable internal test performance but substantially lower accuracy on external validation datasets. The previously reported classification results are summarized in Table 1 for reference [6].

Table 1
Classification performance comparison of DiCNN-UniK and HyenaDNA transfer model on internal test and external validation datasets. Baseline classification results reproduced from the original DiCNN-UniK study [6].

Model	Test Acc.	AUC	Val. Set 1	Val. Set 2	Val. Set 3	Inf. Time (ms)
DiCNN-UniK	99%	1.0	1.00	0.97	0.98	4.19
HyenaDNA-TM	99%	1.0	0.13	0.26	0.41	64.46

4.2. ERF-Window Activation Attribution

Attribution analysis was performed on the full available dataset of 8,342 sequences (training and test combined), of which 8,310 correctly classified samples with softmax confidence ≥ 0.9 were retained for aggregation, covering all ten flavivirus classes. Sample counts per class ranged from 96 (Tick-Borne Encephalitis) to 2,676 (Dengue_2), reflecting the class imbalance present in the original dataset. For each sample, filter-level activation attribution was performed across four convolutional branches $(k = 5, F = 3)$, $(k = 5, F = 5)$, $(k = 6, F = 3)$, and $(k = 6, F = 5)$, yielding ERF-aware k-mer window signatures spanning effective receptive fields of 7–10 nucleotides. Mean activation scores were aggregated per class after applying a minimum filter-level occurrence threshold of five to exclude rare outlier activations. Among the evaluated branches, the $(k = 6, F = 5)$ configuration produced the most prevalent and class-enriched ERF-window signatures. Table 2 therefore reports the highest-ranking ERF-window signatures per flavivirus class for this branch, ranked by their frequency-aware discriminative score ($\log_2$ Disc. Score).

The activation-associated k-mer heatmap (Figure 1) reveals distinct class-level activation profiles across flavivirus groups. Several highly activated signatures show strong class enrichment, while others exhibit partial activation across multiple related classes, reflecting both conserved and discriminative sequence patterns captured by the attribution framework.

4.3. Discriminative Scoring

To identify sequence signatures selectively associated with individual flavivirus classes rather than broadly activated across multiple classes, discriminative scores were computed as the $\log_2$ ratio of class-specific activation to the corresponding activation observed across all remaining classes. Table 3

Table 2
Top activation-associated ERF-window signatures per flavivirus class (k=6, kernel=5). Windows represent sequences of five adjacent 6-mers within the effective receptive field. Scores reflect mean filter-level activation across correctly classified high-confidence samples. **N** denotes the number of retained high-confidence sequences per class ($\tau \geq 0.9$) from the full dataset used for attribution aggregation.

Class	Top ERF-window Signature	N	Mean	$\log_2$ Disc.
Dengue_1	AGATAG | GATAGT | ATAGTT | TAGTTG | AGTTGA	1514	4.379	28.55
Dengue_2	TACGAG | ACGAGC | CGAGCC | GAGCCA | AGCCAG	2676	4.204	27.74
Dengue_3	TCAAAT | CAAATA | AAATAC | AATACA | ATACAC	1909	2.570	27.92
Dengue_4	TGGGCG | GGGCGT | GGCGTA | GCGTAT | CGTATC	676	4.381	28.65
Japanese Encephalitis	CATGAC | ATGACC | TGACCG | GACCGT | ACCGTG	111	3.458	28.14
Tick-Borne Encephalitis	AGATAG | GATAGT | ATAGTC | TAGTCT | AGTCTG	96	2.847	27.50
Usutu Virus	CAAGCG | AAGCGA | AGCGAA | GCGAAG | CGAAGA	188	2.599	27.95
West Nile Virus	GATCGC | ATCGCC | TCGCCG | CGCCGG | GCCGGA	350	3.085	27.54
Yellow Fever	AACCGG | ACCGGG | CCGGGA | CGGGAT | GGGATA	125	5.722	29.02
Zika Virus	AACGAG | ACGAGC | CGAGCC | GAGCCA | AGCCAA	665	4.320	28.56

reports the highest-scoring anchor k-mer for each flavivirus class in the $(k = 6, F = 5)$ branch. All classes exhibited highly discriminative anchor signatures, with $\log_2$ discriminative scores ranging from 24.99 to 26.29. The identified anchor k-mers provide a localized summary of the class-enriched ERF-window signatures reported in Table 2, highlighting sequence elements that contribute most strongly to class-specific attribution patterns.

Table 3
Top discriminative anchor k-mers per flavivirus class ($k = 6$, kernel=5), ranked by $\log_2$ discriminative score. Mean denotes the average activation magnitude across correctly classified high-confidence samples ($\tau \geq 0.9$). Anchor k-mers provide a localized summary of the most discriminative ERF-window signatures.

Class	Top Anchor k-mer	N	Mean	$\log_2$ Disc.
Dengue_1	AGCTGA	1514	0.846	26.29
Dengue_2	GAAGAG	2676	0.838	25.07
Dengue_3	AGGCGA	1909	2.852	25.78
Dengue_4	TAGGTG	676	0.862	25.53
Japanese Encephalitis	AAATGG	111	0.960	25.34
Tick-Borne Encephalitis	GGATGG	96	0.857	25.94
Usutu Virus	AGCAAG	188	0.526	25.50
West Nile Virus	CGAAGC	350	0.733	25.38
Yellow Fever	GTGCTA	125	1.917	25.30
Zika Virus	TCTGGT	665	0.365	24.99

4.4. Attribution Stability Under Simulated Nucleotide Ambiguity

Attribution stability was evaluated by introducing simulated nucleotide ambiguity at five levels ($r \in \{0.0, 0.3, 0.5, 0.7, 0.8\}$) and measuring Jaccard similarity between attributed k-mer sets from clean and ambiguity-masked sequences. Anchor-level attribution similarity remained consistently higher than exact ERF-window similarity under increasing ambiguity perturbation (Figure 2), suggesting that local attribution anchors are relatively more robust to symbolic corruption than full receptive-field window signatures.

At a retained nucleotide fraction of 0.7, mean anchor-level Jaccard similarity across branches reached 0.103 compared to 0.031 for ERF-window matching. Both measures progressively decreased under increasing ambiguity perturbation, although anchor-based attribution remained consistently more stable throughout the evaluated range.

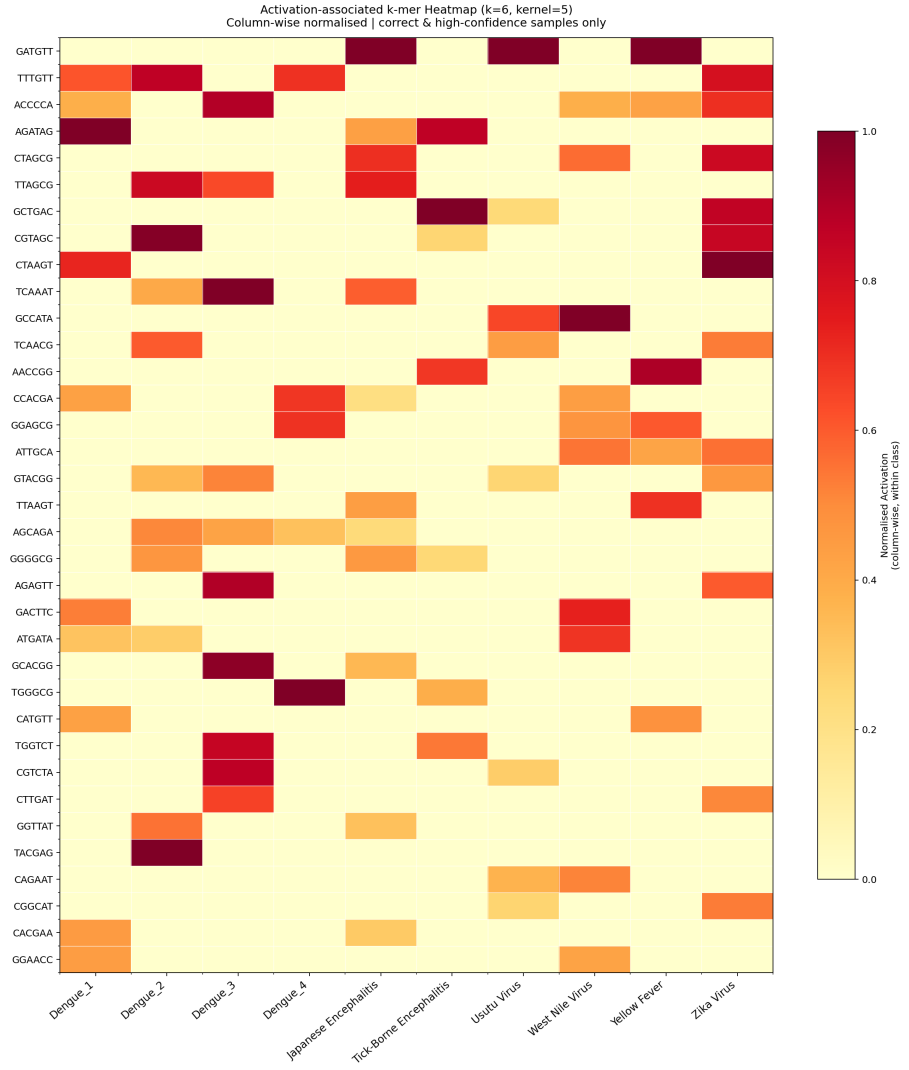


Figure 1: Activation-associated k-mer heatmap for the $(k = 6, \text{kernel} = 5)$ branch. Each cell represents the column-wise normalised mean activation score per flavivirus class. Rows are sorted by their total activation across all classes. Only k-mers with at least five filter-level occurrences are shown. Attribution was performed on correctly classified high-confidence samples only ($\tau \geq 0.9$).

4.5. Occlusion-Based Perturbation Validation

To evaluate whether identified ERF-window signatures were prediction-relevant to model outputs, an occlusion-based perturbation analysis was performed on a stratified subset of correctly classified high-confidence samples (50 per flavivirus class). Because softmax probabilities were highly saturated for correctly classified samples (mean confidence ≥ 0.99), perturbation sensitivity was evaluated using logit-margin changes rather than raw probability differences.

For each sample, the top-15 most activated ERF-window spans identified via per-sample attribution were masked at the nucleotide level by replacing the corresponding nucleotide positions with the ambiguous character N, after which both k-mer branches were re-encoded. The attribution spans used for occlusion were derived from the $(k = 6, F = 5)$ attribution branch, which was selected as the representative configuration for downstream attribution analyses and validation experiments. To provide a comparable perturbation baseline, random control masking was performed using local random nucleotide blocks matched to the total masked nucleotide length of the attributed ERF regions. This local random-block baseline was selected to ensure comparable perturbation geometry between attribution-guided and random masking. Random masking was repeated across 15 trials and averaged

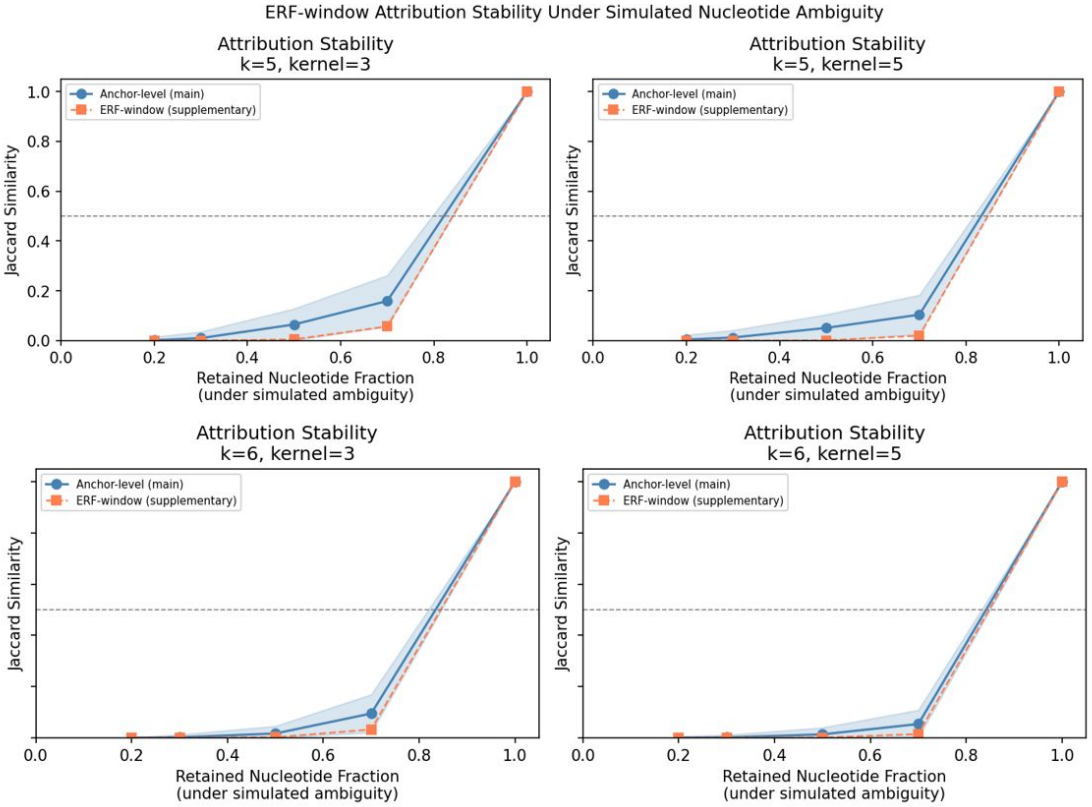


Figure 2: Attribution stability under simulated nucleotide ambiguity. Anchor-level Jaccard similarity (blue) remained higher than exact ERF-window similarity (orange) across all four convolutional branches under increasing ambiguity perturbation.

per sample.

Attribution-guided masking produced larger logit-margin reductions than matched local random-block masking across all ten flavivirus classes (10/10; Figure 3), indicating that the identified ERF-window signatures consistently captured prediction-relevant sequence evidence. Effect magnitudes nevertheless varied substantially across classes. Effects were particularly pronounced for Zika Virus, Usutu Virus, Japanese Encephalitis, and Dengue-3, while Tick-Borne Encephalitis exhibited the smallest but still consistently positive attribution effect. These findings suggest that the identified attribution signatures capture prediction-relevant sequence patterns beyond local random perturbation effects, with heterogeneous effect magnitudes reflecting varying degrees of localized sequence evidence across flavivirus classes.

5. Discussion

This study presented an ERF-aware activation attribution framework for interpreting the classification decisions of DiCNN-UniK, a dual-input convolutional neural network for alignment-free flavivirus classification. By tracing high-activation convolutional responses back to their originating k-mer windows within the effective receptive field, the framework enables attribution of predictions to sequence regions that directly correspond to the evidence selected by the model during classification, rather than to post-hoc approximations of model behaviour. The attribution analysis identified class-associated ERF-window signatures across all ten flavivirus classes, with discriminative scoring revealing patterns more selectively activated for individual classes relative to others. Stability analysis revealed a progressive decline in attribution overlap with increasing simulated nucleotide ambiguity. While classification performance remained relatively robust under reduced genomic coverage, anchor-level attribution overlap decreased substantially earlier, indicating that prediction robustness and attribution

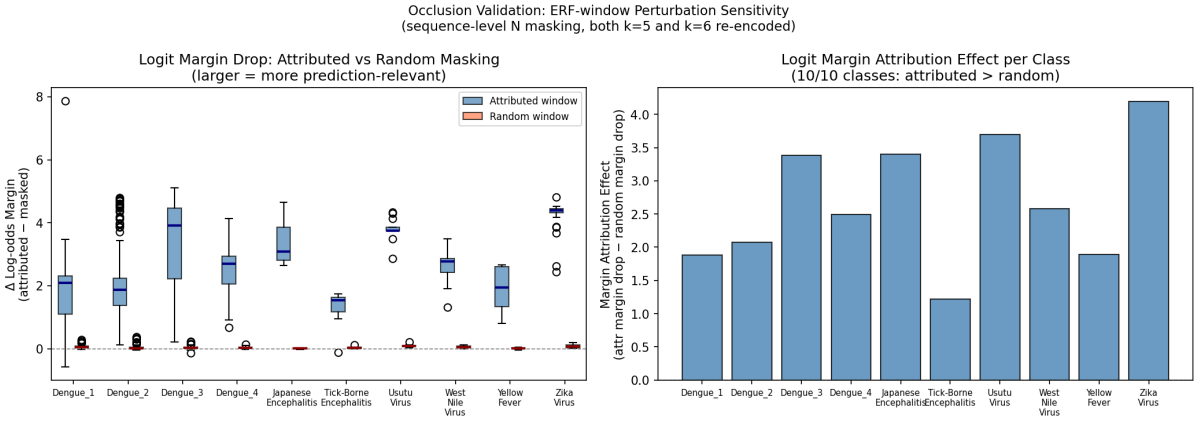


Figure 3: Occlusion-based perturbation validation using logit-margin sensitivity. Left: distribution of logit-margin reductions following attribution-guided ERF masking (blue) versus local random-block masking (orange) per flavivirus class. Right: mean margin attribution effect per class (attributed minus random). Positive values indicate that attribution-guided sequence regions contribute more strongly to model predictions than matched local random regions.

robustness represent distinct properties of the model. Occlusion-based validation further supported the prediction relevance of identified windows, with attributed-window masking producing larger confidence reductions than random masking in the majority of classes. Highly activated ERF windows were not always identical to the most discriminative anchor k-mers, highlighting that activation strength alone does not necessarily correspond to class specificity.

The proposed attribution approach differs from generic post-hoc explainability methods such as SHAP, Integrated Gradients, saliency maps, and TF-MoDISco in that it directly leverages the internal architectural structure of the model, specifically the GlobalMaxPooling operation and ERF-aware convolutional windows, rather than treating the model as a black box. As a result, attributed regions reflect the effective sequence context responsible for filter activation, providing explanations that are more closely aligned with the model's actual decision process. While these methods estimate feature importance through gradient-based, perturbation-based, or motif-discovery strategies, the proposed framework attributes predictions to ERF-aware sequence windows selected through the model's convolutional and pooling operations. A systematic empirical comparison with alternative explainability approaches remains an important direction for future work.

Table 4
Preliminary reference-genome localisation of selected top discriminative anchor k-mers based on the top five exact matches identified in the corresponding reference genome.

Class	Anchor k-mer	Reference	Mapped region(s)
Dengue_1	AGCTGA	NC_001477.1	E, NS1, NS2A, NS3, NS4B, NS5
Dengue_2	GAAGAG	NC_001474.2	E, NS1, NS2A, NS3, NS5
Dengue_3	AGGCGA	NC_001475.2	E, NS2A, NS3, NS4B, NS5
Dengue_4	TAGGTG	NC_002640.1	E, NS3, NS5
Japanese Encephalitis	AAATGG	NC_001437.1	E, NS1, NS2A, NS3, NS4B
Tick-Borne Encephalitis	GGATGG	NC_001672.1	NS1, NS3, NS5
Usutu Virus	AGCAAG	NC_006551.1	E, NS1, NS2B, NS3, NS4A, NS4B, NS5
West Nile Virus	CGAAGC	NC_009942.1	NS1, NS2A, NS3, NS5
Yellow Fever	GTGCTA	NC_002031.1	NS3, NS5
Zika Virus	TCTGGT	NC_035889.1	NS2B, NS3, NS5

Preliminary manual inspection of the top five exact reference-genome matches revealed localisation of selected high-scoring attributed k-mers to annotated coding regions including the Envelope protein (E) as well as NS1, NS2A, NS2B, NS3, NS4A, NS4B, and NS5-associated sequence regions (Table 4).

Several of these regions are associated with flaviviral replication processes and contain conserved sequence elements across flavivirus species. These observations suggest that the attribution framework captures biologically plausible sequence regions rather than uniformly distributed sequence noise. Although these findings are qualitative and not intended as formal biological validation, they indicate that the proposed attribution framework may capture biologically meaningful sequence patterns. In practical surveillance settings, attribution-associated k-mer windows may assist virologists in inspecting sequence regions linked to classification decisions, particularly under ambiguity-corrupted sequencing conditions where alignment-based inspection becomes difficult.

5.1. Limitations

Several important limitations of the proposed framework should be noted. First, activation magnitude does not necessarily imply causal importance. The framework characterizes activation-associated and perturbation-sensitive sequence regions rather than definitive biological mechanisms. Identified k-mer windows should be interpreted as candidate signatures associated with model predictions, not as validated biological motifs. Second, the stability analysis performed in this study simulates nucleotide ambiguity through random character replacement and does not replicate the distributional characteristics of real low-coverage sequencing data, which may exhibit amplicon bias, terminal truncation, or assembly gaps. Validation on real low-coverage genomic data would strengthen the generalisability of the stability findings. Third, the occlusion-based validation measures prediction sensitivity to window masking but does not establish direct biological causality. Integration with biological annotation data, such as protein region mapping or known conserved motif databases, would be required to establish the biological relevance of the identified signatures. Fourth, attribution aggregation was performed on correctly classified high-confidence samples from the full available dataset, including both training and test sequences. While this strategy was intended to maximize attribution coverage across flavivirus classes, the resulting attribution signatures may partially reflect patterns observed during model development. Future work should evaluate attribution stability and discriminative signatures on strictly held-out datasets to assess their generalisability. Finally, the attribution framework was evaluated on a single convolutional architecture and dataset. Generalisation to other k-mer-based genomic models or broader pathogen classification tasks remains future work.

5.2. Future Work

Several directions may further extend the proposed framework. Combining ERF-aware activation attribution with gradient-based explainability methods may provide complementary perspectives on sequence-level relevance and attribution faithfulness. More systematic mapping of attributed k-mer windows to annotated flaviviral genomic regions and functional elements could help establish a stronger biological interpretation of the identified signatures. In addition, evaluation on empirical low-coverage sequencing datasets, including fragmented and assembly-incomplete genomes, would provide a more realistic assessment of attribution stability under practical surveillance conditions. Future studies may also investigate the applicability of the framework to broader pathogen classification settings and alternative k-mer-based neural architectures.

6. Conclusion

This study introduced an ERF-aware activation attribution framework for interpretable flavivirus classification using DiCNN-UniK. By tracing filter-level convolutional activations to their originating k-mer windows within the effective receptive field, the framework provides architecture-aware sequence-level explanations aligned with the model's internal convolutional selection mechanism. Attribution stability analysis and occlusion-based perturbation validation demonstrated that anchor-level attribution signatures remained relatively more stable than exact ERF-window matching under simulated nucleotide

ambiguity and were associated with prediction-relevant sequence regions. Preliminary genomic inspection further suggested correspondence between selected attributed signatures and annotated flaviviral coding regions. Overall, the proposed framework represents a step toward more transparent and interpretable alignment-free genomic deep learning models for pathogen surveillance applications.

Acknowledgments

This work has been financially supported by the German Federal Ministry of Health (BMG) under grant No. 2523DAT400 through the project "AI-assisted analysis and visualization of pandemic situations" (AIDAVis-PANDEMICS).

Online Resources

The project repository is available at: https://github.com/baharilgen/erf-kmer-attribution

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] F. S. Madere, A. V. Andrade da Silva, E. Okeze, E. Tilley, A. Grinev, K. Konduru, M. García, M. Rios, Flavivirus infections and diagnostic challenges for dengue, west nile and zika viruses, npj Viruses 3 (2025) 36. doi:10.1038/s44298-025-00114-z.

[2] K. B. Park, H. H. Patnaik, T.-Y. Kim, Y. H. Jo, N.-Y. Kim, S.-C. Yang, W.-G. Lee, H.-I. Lee, S.-H. Cho, Y. S. Han, Current trends in large-scale viral surveillance methods in mosquitoes, Entomological Research 50 (2020) 292–308. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/1748-5967.12439. doi:https://doi.org/10.1111/1748-5967.12439.

[3] D. J. van Zyl, M. Dunaiski, H. Tegally, C. Baxter, T. de Oliveira, J. S. Xavier, T. I. A. research study group, Alignment-free viral sequence classification at scale, BMC Genomics 26 (2025) 389. URL: https://doi.org/10.1186/s12864-025-11554-5. doi:10.1186/s12864-025-11554-5.

[4] K. E. Wade, L. Chen, C. Deng, G. Zhou, P. Hu, Investigating alignment-free machine learning methods for hiv-1 subtype classification, Bioinformatics Advances 4 (2024) vbae108. URL: https://doi.org/10.1093/bioadv/vbae108. doi:10.1093/bioadv/vbae108.

[5] J. Ren, X. Bai, Y. Y. Lu, K. Tang, Y. Wang, G. Reinert, F. Sun, Alignment-free sequence analysis and applications, Annual Review of Biomedical Data Science 1 (2018) 93–114. URL: https://doi.org/10.1146/annurev-biodatasci-080917-013431. doi:10.1146/annurev-biodatasci-080917-013431.

[6] A. Shahid, J.-U. Ulrich, D. Kühnert, Fast and alignment-free flavivirus classification from low-coverage genomes, bioRxiv (2026). URL: https://doi.org/10.64898/2026.02.20.706982. doi:10.64898/2026.02.20.706982, preprint.

[7] K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps, in: ICLR Workshop, 2014.

[8] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: D. Precup, Y. W. Teh (Eds.), Proceedings of the 34th International Conference on Machine Learning, volume 70 of *Proceedings of Machine Learning Research*, PMLR, 2017, pp. 3319–3328. URL: https://proceedings.mlr.press/v70/sundararajan17a.html.

[9] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural

Information Processing Systems, volume 30, Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

[10] A. Talukder, C. Barham, X. Li, H. Hu, Interpretation of deep learning in genomics and epigenomics, Briefings in Bioinformatics 22 (2021) bbaa177. doi:10.1093/bib/bbaa177.

[11] W. Luo, Y. Li, R. Urtasun, R. Zemel, Understanding the effective receptive field in deep convolutional neural networks, in: D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 29, Curran Associates, Inc., 2016. URL: https://proceedings.neurips.cc/paper_files/paper/2016/file/c8067ad1937f728f51288b3eb986afaa-Paper.pdf.

[12] P. Simmonds, Methods for virus classification and the challenge of incorporating metagenomic sequence data, Journal of General Virology 96 (2015). URL: https://doi.org/10.1099/jgv.0.000016. doi:10.1099/jgv.0.000016.

[13] P. Simmonds, P. Aiewsakun, Virus classification – where do you draw the line?, Archives of Virology 163 (2018) 2037–2046. URL: https://doi.org/10.1007/s00705-018-3938-z. doi:10.1007/s00705-018-3938-z.

[14] A. Zielezinski, H. Z. Girgis, G. Bernard, C.-A. Leimeister, K. Tang, T. Dencker, A. K. Lau, S. Röhling, J. J. Choi, M. S. Waterman, M. Comin, S.-H. Kim, S. Vinga, J. Almeida, C.-K. Chan, B. T. James, F. Sun, B. Morgenstern, Benchmarking of alignment-free sequence comparison methods, Genome Biology 20 (2019) 144. doi:10.1186/s13059-019-1755-7.

[15] Q. Zhang, S.-R. Jun, M. Leuze, D. Ussery, I. Nookaew, Viral phylogenomics using an alignment-free method: A three-step approach to determine optimal length of k-mer, Scientific Reports 7 (2017) 40712. URL: https://doi.org/10.1038/srep40712. doi:10.1038/srep40712.

[16] O. Bonham-Carter, J. Steele, D. Bastola, Alignment-free genetic sequence comparisons: a review of recent approaches by word analysis, Briefings in Bioinformatics 15 (2014) 890–905. URL: https://doi.org/10.1093/bib/bbt052. doi:10.1093/bib/bbt052.

[17] P. K. Koo, S. R. Eddy, Representation learning of genomic sequence motifs with convolutional neural networks, PLoS Computational Biology 15 (2019) e1007560. doi:10.1371/journal.pcbi.1007560.

[18] 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. URL: https://doi.org/10.1093/bioinformatics/btab083. doi:10.1093/bioinformatics/btab083.

[19] H. Dalla-Torre, L. Gonzalez, J. Mendoza-Revilla, N. Lopez Carranza, A. H. Grzywaczewski, F. Oteri, C. Dallago, E. Trop, B. P. de Almeida, H. Sirelkhatim, G. Richard, M. Skwark, K. Beguir, M. Lopez, T. Pierrot, Nucleotide transformer: building and evaluating robust foundation models for human genomics, Nature Methods 22 (2025) 287–297. URL: https://doi.org/10.1038/s41592-024-02523-z. doi:10.1038/s41592-024-02523-z.

[20] G. Novakovsky, N. Dexter, M. W. Libbrecht, W. W. Wasserman, S. Mostafavi, Obtaining genetics insights from deep learning via explainable artificial intelligence, Nature Reviews Genetics 24 (2023) 125–137. URL: https://doi.org/10.1038/s41576-022-00532-2. doi:10.1038/s41576-022-00532-2.

[21] A. Shrikumar, K. Tian, Žiga Avsec, A. Shcherbina, A. Banerjee, M. Sharmin, S. Nair, A. Kundaje, Technical note on transcription factor motif discovery from importance scores (tf-modisco) version 0.5.6.5, 2020. URL: https://arxiv.org/abs/1811.00416. arXiv:1811.00416.

[22] H. Zeng, M. D. Edwards, G. Liu, D. K. Gifford, Convolutional neural network architectures for predicting dna–protein binding, Bioinformatics 32 (2016) i121–i127. URL: https://doi.org/10.1093/bioinformatics/btw255. doi:10.1093/bioinformatics/btw255.

[23] J. Zhou, O. G. Troyanskaya, Predicting effects of noncoding variants with deep learning–based sequence model, Nature Methods 12 (2015) 931–934. URL: https://doi.org/10.1038/nmeth.3547. doi:10.1038/nmeth.3547.

[24] K. Vervier, P. Mahé, M. Tournoud, J.-B. Veyrieras, J.-P. Vert, Large-scale machine learning for metagenomics sequence classification, Bioinformatics 32 (2016) 1023–1032. URL: https://doi.org/10.1093/bioinformatics/btv683. doi:10.1093/bioinformatics/btv683.