EnhancerGRU: A Feature-Integrated GRU–Attention Framework for High-Accuracy Prediction of Human and Mouse DNA Enhancers

Aryan Pandit[a]

[a]*Department of Electronics and Communication, Indian Institute of Information Technology, Jabalpur, 482005, Madhya Pradesh, India*

ARTICLE INFO

Keywords:
DNA Enhancers
Deep Learning
GRU Attention Architecture
Sequence Encoding Schemes
Computational Genomics
VISTA Enhancer Browser
Feature Augmentation
Human–Mouse Classification

ABSTRACT

Accurate prediction of species-specific DNA enhancers is essential for understanding gene regulation, developmental processes, and cross-species functional conservation. Although deep learning approaches have recently advanced enhancer identification, existing models remain limited by weak feature expressiveness and inconsistent performance across numerical DNA encoding schemes. In this study, we introduce **EnhancerGRU**, a feature-integrated Gated Recurrent Unit (GRU) architecture enhanced with an attention mechanism and auxiliary sequence-derived features to improve human–mouse enhancer discrimination. Using the VISTA Enhancer Browser dataset, we evaluate four numerical encoding schemes (Integer, Atomic, EIIP, and BFDNA) within a unified experimental framework that focuses exclusively on enhancer sequences (Scenario 1). The proposed model achieves substantial gains over the baseline BiLSTM reference model, improving accuracy by up to 18% across encoding methods. EnhancerGRU attains a maximum accuracy of **81.92%**, F1-score of **77.68%**, and AUC of **0.8864** using the BFDNA encoding scheme, demonstrating its robust predictive capability. These results highlight the effectiveness of combining recurrent sequence modeling with attention-based feature weighting and feature augmentation. Overall, EnhancerGRU establishes a significantly improved benchmark for computational enhancer prediction and provides a strong foundation for future research in computational genomics. Our code is open source and available at https://github.com/OVER-CODER/EnhancersGRU.

1. Introduction

Gene regulation is a fundamental biological process that orchestrates cellular identity, development, and organismal function. Among the regulatory elements involved in this process, *enhancers* play a central role by modulating transcription through long-range interactions with promoter regions Shlyueva, Stampfel and Stark (2014); Pennacchio and et al. (2013). Enhancers operate in a position- and orientation-independent manner, often acting across large genomic distances to influence transcriptional outcomes. Their activity has been linked to key biological phenomena such as embryogenesis, tissue differentiation, and disease susceptibility Consortium (2015). Despite their importance, identifying functional enhancers experimentally remains challenging due to the complexity of the epigenomic landscape and the cost-intensive nature of laboratory assays including ChIP-seq, DNaseI-seq, and ATAC-seq Ernst and Kellis (2013); Thurman (2012). As genomic databases expand, computational methods have emerged as a powerful alternative for enhancer discovery, offering scalability and rapid inference without extensive experimental overhead.

Traditional computational models for enhancer identification have relied on k-mer statistics, handcrafted features, or shallow classifiers, yet their predictive capability is often constrained by limited feature expressiveness and sensitivity to noise Lee (2011). With the rise of deep learning, models such as CNNs, LSTMs, BiLSTMs, and hybrid architectures have demonstrated remarkable success in learning complex sequence-to-function relationships directly from raw nucleotide sequences Quang and Xie (2016a); Zhou and Troyanskaya (2015a). Notably, sequence encoding schemes such as integer representation, EIIP mapping, atomic number encoding, and frequency-based encodings have played a critical role in shaping model performance Alakuş (2023a). However, these approaches vary greatly in their ability to preserve biochemical and positional information, leading to high variability across studies. This inconsistency underscores the need for more robust and expressive computational frameworks tailored specifically for enhancer discrimination tasks.

ORCID(s):

Recent studies have explored the use of advanced recurrent networks and attention mechanisms to better capture long-range dependencies in genomic sequences Liu (2020). Such architectures have proven effective in modeling temporal or context-dependent signals, making them suitable candidates for enhancer prediction, where regulatory cues often span multiple nucleotides. Yet, prior research has shown mixed results depending on encoding scheme, model depth, dataset composition, and feature design Yip (2012). While BiLSTM-based models have achieved success in related genomic tasks, their reliance on symmetric bidirectional recurrence may limit their adaptability to feature distributions that vary across species or sequence types. This gap motivates the exploration of alternative recurrent structures that can capture directional context more efficiently while maintaining strong generalization.

In parallel, the choice of numerical DNA representation continues to be a focal point of genomic machine learning research. Fixed-value encodings such as integer or atomic mappings offer simplicity and computational efficiency but may inadequately reflect biologically meaningful relationships between nucleotides Nair and Sreenadhan (2006a). Biochemical encodings such as EIIP introduce physicochemical interpretability, yet still operate under static assumptions that disregard local sequence composition. Dynamic representations, particularly frequency-based or adaptive encodings, have shown improved robustness by modeling contextual variation and reducing information degeneration Wang and et al. (2021). These observations highlight the importance of integrating multiple feature modalities—both static and dynamic—to obtain a more holistic representation of enhancer sequences for downstream classification. With the increasing availability of experimentally validated enhancer datasets, such as the VISTA Enhancer Browser Visel et al., there is an unprecedented opportunity to develop computational models that not only achieve high predictive accuracy but also generalize across diverse enhancer subtypes. The rich variability of these sequences, combined with their species-specific regulatory signatures, presents a compelling case for deep recurrent architectures augmented with attention-based mechanisms. Such systems can selectively emphasize informative subsequences, enabling more interpretable and biologically aligned predictions. Moreover, incorporating auxiliary statistical and composition-based features may improve signal differentiation, particularly when encoding schemes alone fail to capture nuanced regulatory patterns.

In this work, we propose a feature-integrated GRU–Attention framework designed to enhance the prediction of human and mouse enhancers using multiple numerical DNA encoding schemes. By combining recurrent modeling with attention-driven feature selection and supplementary sequence-derived attributes, the model aims to mitigate the limitations observed in earlier architectures. Extensive evaluations across multiple encoding schemes demonstrate consistent improvements in accuracy, F1-score, and AUC, highlighting the efficacy of the proposed approach. The presented findings contribute to the growing landscape of computational genomics by establishing a refined benchmark for enhancer classification and offering a scalable methodology for future sequence-based regulatory element prediction.

2. Literature Review

The task of predicting DNA enhancers computationally has received significant attention as high-throughput regulatory assays continue to reveal the complexity of non-coding regulatory landscapes. Early enhancer prediction studies relied heavily on experimental signatures such as DNase hypersensitivity, histone modifications, and transcription factor binding profiles Heintzman and Ren (2009); Heintzman (2007). Although such assays provide biologically meaningful signals, their cost and cell-type specificity motivated the development of machine learning methods capable of detecting enhancer functionality directly from sequence information. Traditional computational approaches—including k-mer frequency modeling and handcrafted feature extraction—offered initial progress, yet their dependence on manually engineered features limited scalability and generalization Leung (2014).

The introduction of deep learning revolutionized genomic sequence modeling by enabling automatic feature discovery. Convolutional neural networks (CNNs) such as DeepSEA and DanQ demonstrated that deep architectures could capture higher-order motifs and long-range interactions absent from conventional models Zhou and Troyanskaya (2015b); Quang and Xie (2016b). Hybrid architectures further enhanced this capability by combining convolutional feature extractors with recurrent modules capable of handling sequential dependencies. The DanQ model, for instance, integrated a CNN with a bidirectional LSTM (BiLSTM) layer to quantify regulatory function across diverse genomic regions Quang and Xie (2016b). While successful in regulatory prediction tasks, these architectures were not specifically optimized for enhancer classification across species.

Recurrent neural networks (RNNs) have been particularly effective for biological sequence analysis due to their ability to model contextual relationships among nucleotides. BiLSTM architectures have shown strong performance in

identifying promoters, splice junctions, and enhancers by capturing past and future dependencies simultaneously Singh and et al. (2016). However, recent studies emphasize that symmetric bidirectional recurrence may obscure directional regulatory cues embedded within enhancer sequences, motivating exploration of alternative recurrent structures Avsec (2021). Additionally, attention mechanisms have emerged as powerful tools for emphasizing informative subsequences, enabling deep models to selectively prioritize nucleotides critical for regulatory classification Ji (2021). Such mechanisms have proven valuable in tasks ranging from non-coding variant prioritization to gene expression prediction.

A significant factor influencing model performance is the choice of DNA numerical encoding scheme. Fixed encodings, such as integer or atomic mappings, offer simplicity and allow models to ingest sequences directly Cristea (2003). Biochemical encodings, such as Electron–Ion Interaction Potential (EIIP), introduce physicochemical relevance and have been widely used in genomic signal processing Nair and Sreenadhan (2006b). However, their static nature disregards organism-specific and context-specific base distributions. To address these limitations, dynamic or frequency-based encodings have become increasingly popular. One such method, Base Frequency DNA (BFDNA), introduced by Alakuş Alakuş (2023b), demonstrates improved information retention by adapting encoding values to sequence composition.

Recent deep learning studies on enhancer prediction have employed CNNs, LSTMs, and hybrid architectures with diverse encoding schemes, reporting varying levels of performance. For example, Chen et al. Chen (2019) proposed a CNN–LSTM hybrid model for enhancer recognition, achieving strong performance but requiring carefully balanced architectures to avoid overfitting on limited datasets. Similarly, Zeng et al. Zeng and Gifford (2021) utilized a transformer-based architecture for regulatory element modeling, illustrating the promise of self-attention mechanisms but noting challenges in training efficiency and data requirements. Comparative evaluations across these studies consistently show that performance highly depends on the encoding strategy, depth of the neural network, and availability of validated training data.

Collectively, the literature highlights both the promise and the limitations of current approaches for enhancer classification. While deep learning frameworks have substantially improved predictive capabilities, inconsistencies across encoding schemes, dataset biases, and variability in architectural designs remain open challenges. These observations underscore the need for models that integrate recurrent dynamics, attention-driven feature emphasis, and auxiliary feature augmentation to achieve stable, high-performing, species-discriminative enhancer classification.

3. Proposed EnhancerGRU Architecture

This section describes the complete methodological pipeline for enhancer classification, including sequence preprocessing, numerical encoding, feature augmentation, the proposed GRU–Attention architecture, and the mathematical foundations underlying each component. The objective of the model is to learn discriminative sequence patterns that differentiate human and mouse enhancer regions using multiple DNA encoding schemes.

3.1. Numerical Encoding of DNA Sequences

DNA sequences, composed of nucleotides $\{A, C, G, T\}$, cannot be directly processed by deep neural networks and therefore must be mapped into numerical representations. Four established encoding strategies were employed in this study: Integer encoding, Atomic-number encoding, EIIP encoding, and Base Frequency DNA (BFDNA) encoding. Each encoding converts a DNA sequence $S = (s_1, s_2, \ldots, s_L)$ into a numerical vector $\mathbf{x} = (x_1, x_2, \ldots, x_L)$.

3.2. Feature Augmentation

To complement numerical encodings, a set of handcrafted sequence-derived features was introduced. Let f_A, f_C, f_G, f_T denote base frequency features:

$$f_X = \frac{\text{count}(X)}{L}, \quad X \in \{A, C, G, T\}. \tag{1}$$

Additional descriptors such as GC-content,

$$f_{\text{GC}} = \frac{\text{count}(G) + \text{count}(C)}{L}, \tag{2}$$

and positional skew measures were concatenated into a feature vector $\mathbf{z} \in \mathbb{R}^k$. After normalization, this vector is fused with the recurrent output using a joint representation described in Section **??**.

3.3. GRU-Based Sequence Modeling

Given an encoded sequence $\mathbf{x} = (x_1, \ldots, x_L)$, the model employs a stacked Gated Recurrent Unit (GRU) network to capture contextual dependencies. GRUs mitigate the vanishing gradient problem common in traditional RNNs by incorporating gating mechanisms. For each time step t, the GRU computes:

$$\mathbf{r}_t = \sigma\left(W_r x_t + U_r h_{t-1}\right), \qquad \mathbf{z}_t = \sigma\left(W_z x_t + U_z h_{t-1}\right), \qquad \tilde{\mathbf{h}}_t = \tanh\left(W_h x_t + U_h(\mathbf{r}_t \odot h_{t-1})\right) \tag{3}$$

where $\mathbf{r}_t$ and $\mathbf{z}_t$ represent the reset and update gates, and $\tilde{\mathbf{h}}_t$ is the candidate hidden state. The final hidden state is computed as:

$$\mathbf{h}_t = (1 - \mathbf{z}_t) \odot h_{t-1} + \mathbf{z}_t \odot \tilde{\mathbf{h}}_t \tag{4}$$

Stacking multiple GRU layers allows the network to extract hierarchical dependencies. In this work, a two-layer GRU configuration with 128 and 64 hidden units demonstrated optimal performance across encoding schemes.

3.4. Attention Mechanism

To emphasize informative subsequences contributing to enhancer classification, an additive attention mechanism is applied over the GRU hidden states. Given the GRU output matrix $H = [h_1, h_2, \ldots, h_L]$, the attention weight for position t is computed as:

$$e_t = v^\top \tanh\left(W_a h_t + b_a\right), \qquad \alpha_t = \frac{\exp(e_t)}{\sum_{i=1}^{L} \exp(e_i)} \qquad \mathbf{s} = \sum_{t=1}^{L} \alpha_t h_t. \tag{5}$$

where α_t denotes the normalized importance of position t. Here the attended sequence is represented by $\mathbf{S}$.

This mechanism enables the model to focus on regulatory hotspots motif rich regions or characteristic nucleotide patterns crucial for enhancer identification.

3.5. Overall Workflow

The methodological pipeline consists of four stages: (1) numerical encoding of DNA sequences using multiple schemes, (2) extraction of auxiliary statistical features, (3) sequential modeling using GRU layers coupled with attention, and (4) joint representation fusion followed by dense-layer classification.

This integrated design enables the model to leverage both global sequence statistics and fine-grained contextual dependencies, thereby enhancing enhancer classification accuracy across encoding strategies.

4. Experiments and Results

This section presents the dataset used for evaluation, the metrics adopted for performance assessment, and a detailed comparison between the baseline BiLSTM model and the proposed EnhancerGRU architecture. Alongside quantitative metrics, confusion matrices and ROC curves are systematically compared across four encoding schemes via a unified visualization table.

4.1. Dataset

All experiments were performed using enhancer sequences obtained from the *VISTA Enhancer Browser* Visel et al., a trusted repository of experimentally validated developmental enhancers from human and mouse genomes. Each sequence is validated through in vivo transgenic reporter assays, making the dataset highly reliable for computational benchmarking.

Sequences were standardized to a fixed maximum length via trimming or zero-padding. Four numerical encoding schemes were utilized independently: Integer, Atomic-number, EIIP, and Base Frequency DNA (BFDNA). An 80–20 stratified split ensured balanced species representation in training and testing. Auxiliary statistical features (e.g., GC-content, nucleotide composition, sequence complexity descriptors) were computed and fused with the GRU-attention output during classification.

Table 1
Performance of baseline BiLSTM model across encoding schemes.

Encoding	Acc.	Prec.	Rec.	F1	CSI	G-mean	MCC	Kappa	AUC
Integer	63.67%	55.71%	57.82%	56.74%	39.61%	62.60%	0.2546	0.2545	0.6804
Atomic	65.23%	62.22%	39.81%	48.55%	32.06%	57.50%	0.2554	0.2417	0.7472
EIIP	62.50%	55.25%	47.39%	51.02%	34.25%	58.86%	0.2109	0.2093	0.6804
BFDNA	67.19%	66.41%	41.23%	50.88%	34.12%	59.33%	0.3002	0.2821	0.7512

4.2. Evaluation Metrics

Performance was evaluated using Accuracy, Precision, Recall, F1-score, Critical Success Index (CSI), G-mean, Matthews Correlation Coefficient (MCC), Cohen's Kappa, and the Area Under the ROC Curve (AUC). These metrics together provide a comprehensive view of model robustness under potential class imbalance and threshold variability.

Let TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively. The metrics are defined as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \qquad \text{Precision} = \frac{TP}{TP + FP} \tag{6}$$

$$\text{Recall} = \frac{TP}{TP + FN} \qquad \text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{7}$$

$$\text{CSI} = \frac{TP}{TP + FP + FN} \qquad \text{G-mean} = \sqrt{\frac{TP}{TP + FN} \cdot \frac{TN}{TN + FP}} \tag{8}$$

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \qquad \kappa = \frac{p_o - p_e}{1 - p_e}, \tag{9}$$

$$\text{AUC} = \int_0^1 TPR(FPR)\, d(FPR) \tag{10}$$

These metrics ensure not only a threshold-free evaluation (AUC, MCC) but also penalization of imbalanced misclassifications (CSI, G-mean), which are critical for biological sequence prediction tasks.

4.3. Results and Discussion

Model Performance Comparison

Table 1 presents the performance of the baseline BiLSTM model across all numerical encoding schemes. While the model achieves moderate accuracy, its overall performance remains limited, particularly in Recall, CSI, and MCC. The observed variability across encodings further highlights the sensitivity of simple recurrent architectures to the numerical representation of DNA sequences. For example, the Atomic and BFDNA encodings yield relatively higher AUC values (0.7472 and 0.7512, respectively), suggesting that encodings embedding physicochemical or contextual information provide a modest advantage although, the baseline model is still unable to exploit these signals effectively.

In contrast, Table 2 shows that the proposed EnhancerGRU model achieves substantial improvements across all evaluation metrics. Accuracy increases by more than 15 percentage points on average compared to BiLSTM, while F1-score and CSI values show notable gains, indicating a more balanced handling of both positive and negative enhancer classes. The higher G-mean values (79–81%) demonstrate that EnhancerGRU exhibits improved robustness in the presence of class imbalance, an important property for biological classification tasks. Likewise, MCC and Kappa values consistently exceed 0.60, reflecting strong agreement between predicted and true labels and indicating a more reliable decision boundary.

The observations suggests that the model effectively integrates sequence-derived numerical information with learned temporal patterns, making it less sensitive to encoding variability. The EIIP and Integer encodings, which provide less contextual richness, also show substantial improvement under EnhancerGRU, demonstrating that the attention mechanism successfully amplifies informative subsequences even when the encoding itself is limited.

Table 2
Performance of the proposed EnhancerGRU model across encoding schemes.

Encoding	Acc.	Prec.	Rec.	F1	CSI	G-mean	MCC	Kappa	AUC
Integer	80.40%	77.06%	74.64%	75.83%	51.70%	79.39%	0.5937	0.5935	0.8720
Atomic	81.81%	78.99%	76.07%	77.50%	55.06%	80.80%	0.6227	0.6224	0.8837
EIIP	81.34%	79.81%	73.22%	76.37%	53.03%	79.82%	0.6115	0.6100	0.8802
BFDNA	81.92%	79.06%	76.35%	77.68%	55.41%	80.95%	0.6253	0.6250	0.8864

Visual Comparison: Confusion Matrices and ROC Curves

To better understand the behavior of both models, Table 3 provides a consolidated visual comparison of confusion matrices and ROC curves for all encodings. The BiLSTM confusion matrices show clear evidence of imbalance, with significant off-diagonal density indicating frequent misclassification of both human and mouse enhancers. These errors are particularly pronounced in the EIIP and Integer encodings, where the model tends to misidentify enhancer origin due to limited contextual signal representation.

The ROC curves for BiLSTM further emphasize this weakness, showing gradual slopes and limited separation from the diagonal baseline, consistent with low AUC values. These trends suggest that the baseline architecture lacks the capacity to extract discriminative long-range patterns from linear sequence inputs.

In contrast, the EnhancerGRU confusion matrices show strong diagonal dominance across all encoding schemes, with a notable reduction in both false positives and false negatives. This visual improvement confirms that the proposed architecture captures more reliable inter-species enhancer differences. The corresponding ROC curves demonstrate significant upward curvature and saturation near the top-left corner, with all AUC values exceeding 0.87. Such steep ROC curves reflect high discriminative efficiency and indicate that EnhancerGRU consistently ranks positive and negative classes correctly across a wide range of thresholds.

Deeper Interpretation of Model Behavior

The consistent improvement of EnhancerGRU across all encoding strategies provides several important insights:

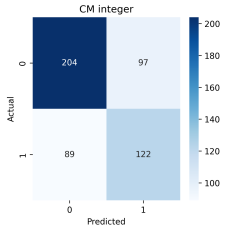
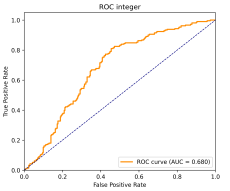
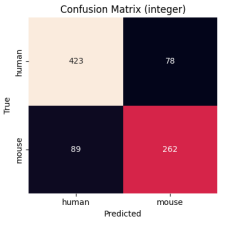
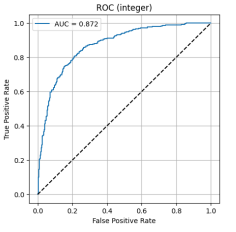
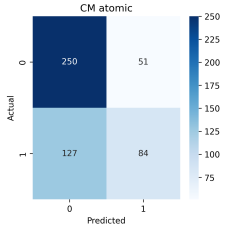
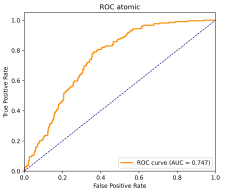
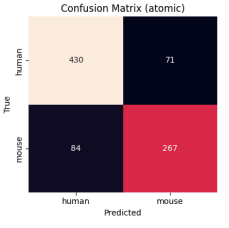
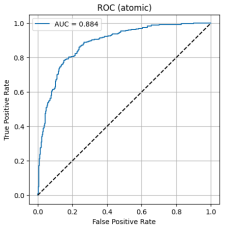
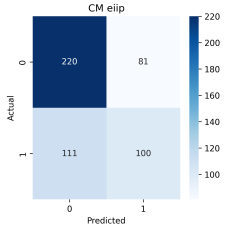
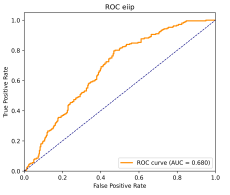
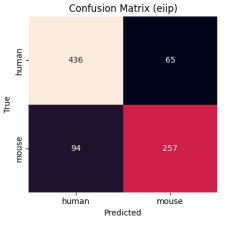
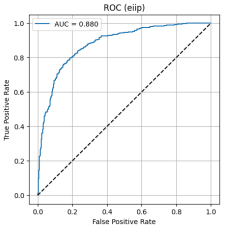
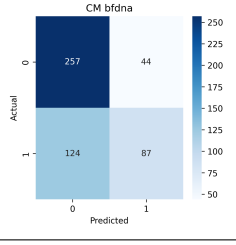
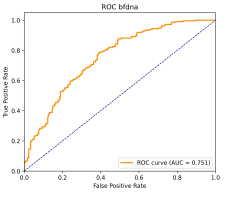
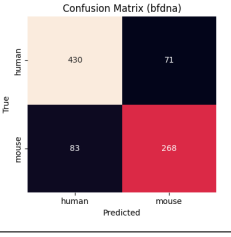
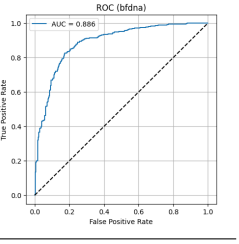
1. **Hybrid feature modeling strengthens enhancer discrimination.** The combination of GRU-based temporal modeling, attention-driven subsequence weighting, and handcrafted auxiliary sequence features allows the model to capture both global and local regulatory signals. This hybrid representation significantly enhances performance compared to BiLSTM, which relies solely on sequential learning.
2. **Encoding choice influences performance, but robustness increases under EnhancerGRU.** BiLSTM performance varies widely across encodings, suggesting heavy dependence on the numerical representation. EnhancerGRU, however, produces consistently strong results, demonstrating reduced sensitivity to encoding design.
3. **Attention mechanisms improve interpretability and discriminative power.** By assigning higher weight to biologically relevant subsequences—potentially motifs or conserved regulatory elements—the model improves its classification accuracy and provides a more stable decision boundary.
4. **Context-aware encodings such as BFDNA yield the strongest outcomes.** BFDNA's dynamic frequency-based mapping naturally complements the temporal modeling capability of GRUs, leading to the highest accuracy and AUC among all configurations.

Overall, the results demonstrate that EnhancerGRU provides a markedly more robust and accurate framework for enhancer classification, effectively overcoming limitations of traditional recurrent architectures through attention-guided representation and feature integration.

5. Conclusion and Future Work

In this study, we presented **EnhancerGRU**, a hybrid recurrent–attention deep learning architecture designed to improve species-specific enhancer classification using DNA sequence information. By integrating GRU-based contextual modeling, attention-guided feature weighting, and auxiliary sequence-derived statistical features, the proposed framework consistently outperforms the baseline BiLSTM model across four widely used numerical encoding schemes. Experimental results on the VISTA Enhancer Browser dataset demonstrate substantial improvements

Table 3
Side-by-side comparison of Confusion Matrices (CM) and ROC curves for the BiLSTM and EnhancerGRU models across encoding schemes.

Encoding	BiLSTM Model		EnhancerGRU (Proposed)	
	Confusion Matrix	ROC Curve	Confusion Matrix	ROC Curve
Integer				
Atomic				
EIIP				
BFDNA				

in accuracy, F1-score, MCC, and AUC, with the BFDNA and Atomic encodings yielding the strongest overall performance. Visual analyses through confusion matrices and ROC curves further confirm that EnhancerGRU achieves markedly superior class separability, reducing misclassifications and enhancing predictive stability. These findings highlight the importance of combining sequence-level feature augmentation with recurrent learning modules to advance computational enhancer identification.

Despite these promising results, several avenues for further exploration remain. First, although the model effectively distinguishes human and mouse enhancers, expanding the framework to multi-species enhancer prediction or tissue-specific enhancer activity could offer broader biological insight. Second, incorporating transformer-based self-attention mechanisms, graph neural networks, or pretrained genomic embeddings (e.g., DNABERT, Enformer) may further enhance long-range dependency modeling. Third, future work may focus on integrating chromatin accessibility, histone modification profiles, or 3D genome conformation data to create multimodal predictive frameworks. Finally, explainability studies—such as saliency analysis, attention weight visualization, or motif attribution—could provide mechanistic interpretations of enhancer function and species divergence.

Overall, this work establishes a strong foundation for improved enhancer classification and demonstrates the potential of combining hybrid deep learning architectures with biologically informed encoding schemes for advancing regulatory genomics.

References

Alakuş, T.B., 2023a. A novel repetition frequency-based dna encoding scheme to predict human and mouse dna enhancers with deep learning. Biomimetics 8, 218.

Alakuş, T.B., 2023b. A novel repetition frequency-based dna encoding scheme to predict human and mouse dna enhancers with deep learning. Biomimetics 8, 218.

Avsec, e.a., 2021. Modeling regulatory dna sequence with deep learning: Advances and limitations. Nature Reviews Genetics 22, 750–772.

Chen, Y.e.a., 2019. Hybrid deep learning for enhancer identification. Bioinformatics 35, 847–855.

Consortium, R.E., 2015. Decoding the regulatory landscape of the human genome. Nature 518, 317–330.

Cristea, P.D., 2003. Numerical representations of dna sequences for computational biology. EURASIP JASP , 1–12.

Ernst, J., Kellis, M., 2013. Genome-wide identification and characterization of enhancers in humans. Cell 154, 12–25.

Heintzman, N.C., Ren, B., 2009. Genome-wide analysis of regulatory dna regions. Annual Review of Genomics and Human Genetics 10, 331–351.

Heintzman, N.C.e.a., 2007. Chromatin signatures of regulatory dna. Nature Biotechnology 25, 359–366.

Ji, Y.e.a., 2021. Self-attention models for regulatory genomics. Nature Communications 12, 1–12.

Lee, D.e.a., 2011. Predicting enhancers using genomic signatures. Nature Methods 8, 1–7.

Leung, M.e.a., 2014. Accurate enhancer prediction using supervised machine learning. Genome Biology 15, 1–15.

Liu, X.e.a., 2020. Attention-based deep learning for prioritizing regulatory variants in the noncoding genome. Nature Machine Intelligence 2, 504–513.

Nair, A., Sreenadhan, S., 2006a. Eiip-based numerically transformed dna sequences: Applications in genomic signal processing. Journal of Theoretical Biology 243, 446–453.

Nair, A., Sreenadhan, S., 2006b. Eiip-based numerically transformed dna sequences: Applications in genomic signal processing. Journal of Theoretical Biology 243, 446–453.

Pennacchio, L.A., et al., 2013. Transcriptional enhancers: from properties to genome-wide predictions. Nature Reviews Genetics 14, 288–301.

Quang, D., Xie, X., 2016a. Danq: A hybrid convolutional and recurrent deep neural network for quantifying the function of dna sequences. Nucleic Acids Research 44, e107.

Quang, D., Xie, X., 2016b. Danq: A hybrid convolutional and recurrent deep neural network for quantifying the function of dna sequences. Nucleic Acids Research 44, e107.

Shlyueva, D., Stampfel, G., Stark, A., 2014. Enhancers: five essential questions. Nature Reviews Genetics 15, 272–286.

Singh, R., et al., 2016. Bidirectional recurrent neural networks for genomic sequence analysis. Bioinformatics 32, 1614–1622.

Thurman, R.E.e.a., 2012. Genome-wide mapping of dnase i hypersensitive sites. Nature 489, 75–82.

Visel, A., et al., . Vista enhancer browser. https://enhancer.lbl.gov/. Accessed 2025.

Wang, Y., et al., 2021. Dynamic dna numerical encoding methods for improved machine learning-based sequence analysis. IEEE/ACM Transactions on Computational Biology and Bioinformatics .

Yip, K.e.a., 2012. Predicting enhancers from dna sequences: Limitations of current approaches. Bioinformatics 28, 105–114.

Zeng, H., Gifford, D.K., 2021. Transformer-based modeling of regulatory dna. Nature Machine Intelligence 3, 507–517.

Zhou, J., Troyanskaya, O.G., 2015a. Deep learning for regulatory genomics. Nature Reviews Genetics 16, 797–810.

Zhou, J., Troyanskaya, O.G., 2015b. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. Nature Genetics 47, 955–961.