

Helixpr: a helix-aware periodic representation of DNA sequence for enhancer identification

Received: 27 April 2026

Accepted: 14 July 2026

Published online: 12 August 2026

Cite this article as: Shi X., Kang J., Sun N. *et al.* Helixpr: a helix-aware periodic representation of DNA sequence for enhancer identification. *BMC Bioinformatics* (2026). <https://doi.org/10.1186/s12859-026-06563-4>

Xiang Shi, Jiayi Kang, Nan Sun & S. S.-T. Yau

We are providing an unedited version of this manuscript to give early access to its findings. Before final publication, the manuscript will undergo further editing. Please note there may be errors present which affect the content, and all legal disclaimers apply.

If this paper is publishing under a Transparent Peer Review model then Peer Review reports will publish with the final article.

ARTICLE IN PRESS

HelixPR: a helix-aware periodic representation of DNA sequence for enhancer identification

Xiang Shi¹, Jiayi Kang², Nan Sun³, S.S.-T. Yau^{1*}

¹Department of Mathematical Sciences, Tsinghua University, 100084, Beijing, P.R. China.

²Hetao Institute of Mathematics and Interdisciplinary Sciences (HIMIS), 518000, Guangdong, P.R. China.

³Beijing Institute of Mathematical Sciences and Applications (BIMSA), 101408, Beijing, P. R. China.

*Corresponding author(s). E-mail(s): yau@uic.edu;

Abstract

Background: Enhancers are distal cis-regulatory elements that play essential roles in gene regulation, development, and disease. Although numerous computational methods have been developed for enhancer identification, commonly used DNA encoding strategies, such as one-hot and k -mer representations, typically treat DNA as a linear symbolic sequence. This simplification ignores the intrinsic geometry of the DNA double helix, where nucleotides are organized along a periodic three-dimensional structure with a helical repeat of approximately **10–11** base pairs. To address this limitation, we propose *HelixPR* (*Helix-aware Periodic Representation*), a biologically motivated DNA sequence representation for enhancer identification.

Results: HelixPR augments per-nucleotide one-hot encoding with a sinusoidal periodic encoding using a period of $T = 10$ base pairs. It further incorporates a global nucleotide-composition token to capture sequence-level compositional information complementary to local nucleotide patterns. Together, these components form a compact, interpretable, and information-preserving representation of DNA sequences. Combined with a lightweight one-dimensional convolutional neural network, HelixPR achieved competitive performance across benchmark datasets. On Liu’s benchmark, HelixPR achieved an accuracy of 0.8434 and a Matthews correlation coefficient of 0.6874 for enhancer identification. For enhancer strength classification, HelixPR achieved an accuracy of 0.8400 and an AUC of 0.9346, outperforming baseline methods. On Basith’s benchmark, HelixPR obtained a mean balanced accuracy of 0.9778, substantially outperforming existing methods. Further analysis showed that HelixPR supports closed-form decoding and that HelixPR-highlighted regions are enriched for

known transcription-factor binding motifs, supporting both the interpretability and biological relevance of the proposed representation.

Conclusions: These results suggest that incorporating the periodic structure of the DNA double helix can improve enhancer modeling. By combining nucleotide identity, helical periodicity, and global compositional information, HelixPR offers a biologically grounded framework for enhancer identification and related regulatory sequence prediction tasks.

Keywords: Enhancer identification, DNA sequence representation, Positional encoding, DNA periodicity, Convolutional neural network

1 Introduction

Enhancers are cis-regulatory DNA elements, typically ranging from 50 to 1500 bp in length, that activate the transcription of target genes through the binding of transcription factors (TFs) [1]. They are found in both eukaryotes and prokaryotes and are particularly abundant in the human genome [2]. A typical mammalian cell contains thousands of active enhancers, and more than one million enhancers are estimated to be active across all human cell types [3]. Enhancers are involved in gene regulation during development, evolution, and disease [4, 5] including cancer and heart disease [6]. As a result, accurate enhancer identification remains a central problem in genomics. Unlike proximal regulatory elements such as promoters, enhancers are distal and orientation-independent. They can be located in introns, exons, or intergenic regions, may lie up to one million bp away from target genes, and can even regulate transcription across chromosomes [7]. The positional flexibility makes enhancer identification particularly challenging.

Experimental approaches, such as chromatin immunoprecipitation followed by sequencing (ChIP-seq) and DNase I sequencing (DNase-seq), can identify enhancer elements by profiling transcription-factor binding, chromatin accessibility, or other regulatory activities. However, these approaches are often time-consuming and expensive. To address these limitations, computational methods have been developed to predict enhancers directly from sequence data. For example, iEnhancer-2L [8] employs Pseudo k-tuple [9] embeddings together with a support vector machine; Enhancer-IF [10], iEnhancer-EL [11], and iEnhancer-XG [12] integrate multiple feature encoding schemes (including EIIP [13] and k-mer [14]) with ensemble learners such as XGBoost, SVM, and random forest to capture different aspects of sequence information; iEnhancer-ELM [15] combines the k-mer method with an attention mechanism for enhancer identification; CapsEnhancer [16] and Caps-ac4C [17] use chaos game representation to encode nucleotide sequences as two-dimensional images, and then apply capsule networks to learn sequence patterns. These studies enable rapid and effective identification of candidate enhancer regions prior to experimental validation.

However, most existing methods still rely on sequence representations such as one-hot encoding, k-mer encoding, and EIIP. These representations generally model

DNA as a one-dimensional ordered string of nucleotides, and therefore do not explicitly characterize the spatial continuity or periodic geometry of the DNA backbone. Better capturing such structural properties may provide additional information for distinguishing enhancer from non-enhancer sequences. In particular, one-hot and EIIP encodings preserve nucleotide identity or physicochemical values at each position, but lack explicit structural priors. Frequency-based k-mer features can capture local compositional patterns, but they discard global nucleotide ordering and are generally non-invertible. Moreover, the k-mer feature space grows exponentially with k , leading to the curse of dimensionality.

A particularly important structural property missing from these representations is the approximately 10–11, bp helical periodicity observed across eukaryotic and prokaryotic genomes [18]. This periodicity is closely associated with nucleosomal positioning and DNA supercoiling [19], and may provide informative signals for distinguishing regulatory from non-regulatory regions. Recent work has introduced a multi-perspective representation [20] to capture topological properties of DNA in its native three-dimensional conformation, demonstrating that structural encoding can improve downstream classification. Motivated by these findings, we propose the *Helix-aware periodic DNA sequence representation* (HelixPR), a novel sequence representation that explicitly integrates the three-dimensional helical geometry of the DNA double helix into the feature space for enhancer identification. Unlike conventional encoding schemes that mainly treat DNA as a linear symbolic sequence, HelixPR introduces a biologically motivated periodic prior that reflects the characteristic helical repeat of DNA. The main contributions of this work are summarized as follows.

1. We introduce HelixPR, a sequence representation that explicitly encodes the intrinsic periodic property of the DNA double helix for enhancer classification.
2. HelixPR integrates one-hot encoding, a helix-aware periodic Fourier positional encoding, and a prepended global composition token analogous to the [CLS] token in BERT-based models, thereby combining local nucleotide ordering, helical phase information, and sequence-level compositional statistics within a unified representation.
3. Coupled with a lightweight four-layer one-dimensional CNN, HelixPR achieves strong performance on two benchmark datasets, outperforming existing enhancer identification methods, including methods based on large pre-trained language models.
4. HelixPR is injective and allows exact recovery of the original DNA sequence, thereby preserving sequence information while incorporating biologically meaningful structural priors.

2 Materials and Methods

2.1 Data Source

2.1.1 Liu’s Dataset

Enhancers belonging to different subgroups, such as strong versus weak enhancers and poised versus inactive enhancers, exhibit distinct biological activities and regulatory effects on their target genes [4, 8]. Based on chromatin-state information from nine cell lines—H1ES, K562, GM12878, HepG2, HUVEC, HSMM, NHLF, NHEK, and HMEC—the enhancer dataset proposed in [8] consists of DNA sequences segmented into 200 bp fragments. To reduce sequence redundancy, DNA sequences with high similarity ($\geq 80\%$) were removed using CD-HIT [21]. The resulting benchmark contains 1,484 enhancers, comprising 742 strong enhancers and 742 weak enhancers, together with 1,484 non-enhancers. For fair comparison, an independent test set was further constructed in [11] following the same protocol; it includes 200 enhancers (100 strong and 100 weak) and 200 non-enhancers. Together, the training set and the independent test set constitute the Liu benchmark.

Following prior works [11], enhancer classification on Liu’s dataset is evaluated as two sequential binary tasks: (i) distinguishing enhancers from non-enhancers, and (ii) classifying identified enhancers as strong or weak. We train the model for the first task, and then transfer the results to the second task through fine-tuning.

2.1.2 Basith’s Dataset

The dataset proposed by [22] consists of eight subsets derived from eight cell lines: epithelial (HEK293), normal human epidermal keratinocyte (NHEK), lymphoblast (K562), B lymphocyte (GM12878), human mammary epithelial cell (HMEC), human skeletal muscle myoblasts (HSMM), normal human lung fibroblasts (NHLF), and human umbilical vein endothelial cells (HUVEC). Since each cell type exhibits a distinct gene expression signature, which is largely driven by cell type-specific enhancer usage [1], this dataset provides a biologically meaningful benchmark for enhancer identification.

In the training sets, the numbers of enhancer (positive) and non-enhancer (negative) sequences are balanced, whereas in the test sets, the number of non-enhancer sequences is twice that of enhancer sequences. Compared with a fully balanced setting, Basith’s dataset therefore better reflects real-world classification scenarios with class imbalance. Another major difference between the two benchmarks lies in sequence length. In Liu’s dataset, all sequences are fixed at 200 bp, whereas in Basith’s dataset, sequence lengths range from 204 bp to 2000 bp across different cell lines. Detailed sample counts and further analyses of dataset characteristics, including sequence length, are provided in the Supplementary Material.

2.2 Baseline Methods

Feature extraction is a critical step in sequence-based machine learning, as it determines how biological information is encoded for downstream classifiers [23]. In this

section, we briefly review three widely used DNA encoding methods that serve as baselines in our experiments: one-hot, k-mer, and EIIP encoding. Our proposed method, HelixPR, is introduced in Section 2.3.

2.2.1 One-hot Encoding

For a DNA sequence $S = s_1 s_2 \dots s_L$ of length L , where $s_i \in \{A, T, C, G\}$, one-hot encoding maps each nucleotide to a four-dimensional binary vector.

$$\mathbf{v}_k = \begin{cases} (1, 0, 0, 0), & s_k = A, \\ (0, 1, 0, 0), & s_k = T, \\ (0, 0, 1, 0), & s_k = C, \\ (0, 0, 0, 1), & s_k = G. \end{cases} \quad (1)$$

Concatenating $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_L$ produces a binary matrix of size $(L \times 4)$.

One-hot encoding preserves nucleotide identity and sequence order, but it provides a sparse representation with limited explicit structural information. In particular, it represents positional information only through the discrete order of nucleotides in the sequence, without directly encoding the periodic geometry of the three-dimensional DNA double helix.

2.2.2 K-mer Encoding

The k -mer method represents DNA sequences by counting the occurrence frequencies of all possible contiguous subsequences of length k . Given a sequence of length L , it can be tokenized into $(L - k + 1)$ overlapping k-mers. For example, for the sequence $S = \text{ATCGAT}$ with $k = 3$, the vocabulary size is $4^3 = 64$, and only the subsequences $\{\text{ATC}, \text{TCG}, \text{CGA}, \text{GAT}\}$ have non-zero frequencies.

As k increases, k-mer encoding can capture longer-range local dependencies; however, it also leads to exponential growth of the feature space, resulting in the curse of dimensionality. Conversely, when k is small, multiple distinct sequences may map to identical k-mer embeddings [20]. Formally, k-mer encoding is a many-to-one projection that maps a sequence of length L to a fixed-dimensional frequency vector, thereby irreversibly discarding ordering information. As a result, the original sequence generally cannot be recovered from its k-mer representation, meaning that the encoding is *not* injective. Moreover, like one-hot encoding, k-mer methods treat DNA as a purely linear string and capture limited information about the three-dimensional structure of the double helix and its periodic property.

2.2.3 Electron–Ion Interaction Pseudopotentials(EIIP)

EIIP encoding represents a DNA sequence as a single numerical signal by assigning each nucleotide its Electron–Ion Interaction Pseudopotential value. Specifically, EIIP maps each nucleotide to a scalar corresponding to the energy of its free electrons: $A \mapsto 0.1260$, $T \mapsto 0.1335$, $C \mapsto 0.1340$ and $G \mapsto 0.0806$.

Compared with one-hot encoding, EIIP provides a more compact representation and introduces physicochemical information into the sequence encoding. However, it

still reduces each nucleotide to a single scalar and does not explicitly preserve richer structural relationships or helical geometric information.

2.3 HelixPR: a Helix-aware Periodic Representation

The preceding methods treat DNA as a linear sequence of characters, thereby neglecting its three-dimensional physical structure. In reality, a single strand of the DNA double helix traces a regular cylindrical helix in three-dimensional space, which can be parameterized as

$$\begin{cases} x(t) = r \cos(t), \\ y(t) = r \sin(t), \\ z(t) = ct, \end{cases} \quad (2)$$

where r denotes the helical radius and c is the axial rise per unit increase in t . This geometry naturally induces periodic patterns along the sequence axis. Studies in both eukaryotic and prokaryotic genomes have consistently reported a characteristic periodicity of approximately 10–11 bp [18], which has been associated with nucleosomal positioning and DNA supercoiling [19].

To explicitly encode this periodic three-dimensional structure, we propose HelixPR, a hybrid representation that integrates three complementary components:

1. a global nucleotide composition token (CLS),
2. per-nucleotide one-hot content,
3. helix-aware periodic Fourier positional encoding.

2.3.1 Global Composition Encoding (CLS Token)

Let n_A, n_T, n_C and n_G denote the counts of nucleotides A, T, C and G in a sequence S of length L . Their normalized frequencies are defined as

$$f_A = \frac{n_A}{L}, \quad f_T = \frac{n_T}{L}, \quad f_C = \frac{n_C}{L}, \quad f_G = \frac{n_G}{L}.$$

The global composition token is then constructed as

$$\mathbf{c} = [f_A, f_T, f_C, f_G] \in \mathbb{R}^4. \quad (3)$$

This token summarizes the global nucleotide composition of the entire sequence and is prepended to the per-nucleotide representations. In this sense, it plays a role analogous to the [CLS] token in BERT-based models [15]; therefore, we refer to it as the CLS token throughout this paper. This token is also related to the 1-mer embedding and the count vector used in the natural vector method [20]. By introducing this CLS token, HelixPR embeds sequence-level nucleotide statistics directly into the representation. Consequently, the classifying model receives global compositional information at the input stage, rather than relying solely on deeper layers to progressively aggregate such information from local nucleotide features.

2.3.2 Fourier Positional Encoding

As described in Equation 2, a single strand of the DNA double helix can be approximated by a cylindrical helical trajectory in three-dimensional space. Conventional sequence representations can be viewed as considering only the projection of this trajectory onto the helical axis, namely the linear order along the z -axis, while omitting the periodic phase information revealed by its projection onto the xOy plane. Under the projection onto the xOy plane, the helical phase of a nucleotide can be represented by the trigonometric components $(\cos \theta, \sin \theta)$, which describe its periodic position around the helical axis. Assuming that one helical turn corresponds to T base pairs, the phase of the nucleotide at position k is defined as

$$\theta_k = \frac{2\pi k}{T},$$

where T denotes the helical period in base pairs.

In this work, we set $T = 10$ to approximate the well-established 10~11 bp helical repeat of the B-form DNA double helix. This formulation allows the positional encoding to reflect the helical phase of nucleotides, rather than merely serving as an abstract index-based position embedding. Biologically, nucleotides or regulatory motifs separated by approximately one helical turn may have similar spatial orientations on the DNA helix, which may influence transcription factor binding and enhancer regulatory organization. Unlike generic positional encodings that mainly provide position indices, our objective is not to construct a general-purpose position representation. Instead, we aim to inject a biologically motivated periodic prior into the feature space and concentrate the encoding on the characteristic structural scale of the DNA double helix. The effect of this design choice is empirically validated in Section 3.7.

The corresponding four-dimensional Fourier positional vector is defined as

$$\mathbf{p}_k = [\sin(\theta_k), \cos(\theta_k), \cos(\theta_k), \sin(\theta_k)] \in \mathbb{R}^4. \quad (4)$$

Remark 1 The four-dimensional form of $\mathbf{p}_k$ is chosen to ensure dimensional consistency with the one-hot embedding $\mathbf{v}_k \in \mathbb{R}^4$, allowing direct channel-wise addition without introducing a projection layer. Among possible single-period trigonometric constructions, the replicated sine-cosine pattern $[\sin \theta_k, \cos \theta_k, \cos \theta_k, \sin \theta_k]$ provides a simple dimensionally aligned design. Under the nucleotide ordering (A, T, C, G) , the first and fourth channels correspond to A and G , which are purines, and receive the $\sin \theta_k$ component. The second and third channels correspond to T and C , which are pyrimidines, and receive the $\cos \theta_k$ component. This layout introduces a fixed $\pi/2$ phase offset between the purine and pyrimidine channels. We emphasize that this correspondence is a consequence of the channel arrangement rather than an explicitly imposed biological constraint.

2.3.3 Feature Fusion

The final embedding of the nucleotide at position k is obtained by linearly combining the one-hot content and the helix-aware Fourier positional encoding:

$$\mathbf{e}_k = \mathbf{v}_k + \alpha \mathbf{p}_k, \quad (5)$$

where $\mathbf{v}_k$ denotes the one-hot vector defined in Equation 1, $\mathbf{p}_k$ denotes the Fourier positional encoding and α is a scaling coefficient that controls the relative contribution of the positional term.

In this work, α is set to a small value ($\alpha = 0.05$; see Section 3.7), so the Fourier positional encoding acts as a mild modulation of the original one-hot representation rather than replacing nucleotide identity. Since each element of $\mathbf{p}_k$ is bounded within $[-1, 1]$, each channel of $\mathbf{v}_k$ is perturbed by at most $|\alpha| = 0.05$. Therefore, the dominant binary pattern of the one-hot vector is preserved, while position-dependent helical phase information is incorporated into the input representation.

2.3.4 Final Sequence Representation

The complete HelixPR representation is given by

$$\mathbf{E} = (\mathbf{c}, \mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_L) \in \mathbb{R}^{(L+1) \times 4}.$$

where $\mathbf{c}$ is the global composition token defined in Equation 3 and $\mathbf{e}_k$ denotes the fused embedding of the nucleotide at position k . The parameter choices $T = 10$ and $\alpha = 0.05$ are justified in Section 3.7. In particular, $T = 10$ is aimed to align with the known 10 ~ 11 bp helical repeat of the DNA double helix. Fig. 1 illustrates the overall framework of HelixPR and the downstream classification architecture.

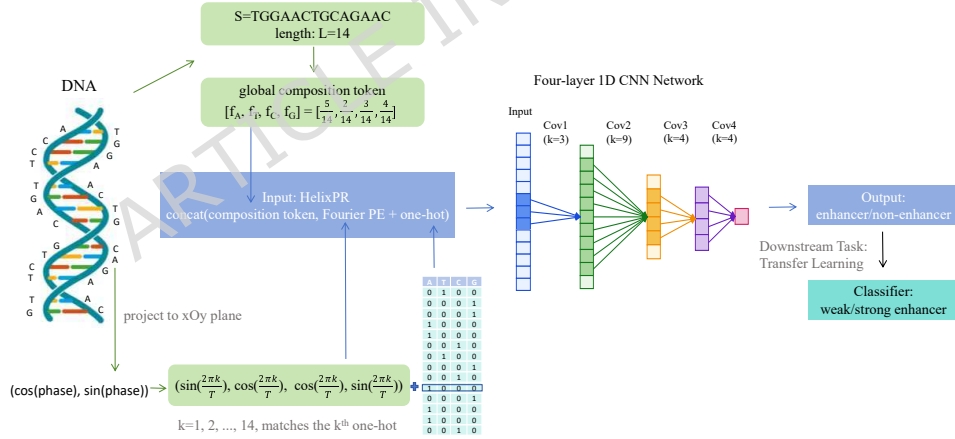


Fig. 1: The framework of HelixPR and the network architecture

2.4 Theoretical Analysis: Invertibility

Proposition 1 *Let $\alpha \in (0, 0.5)$. For any nucleotide sequence S of length L , the HelixPR encoding $\mathbf{E}$ uniquely determines S . Consequently, HelixPR is injective (invertible on its image).*

Proof Fix any position $k \in \{1, \dots, L\}$. The one-hot vector $\mathbf{v}_k$ has exactly one channel equal to 1 and the remaining three channels equal to 0. For the positional perturbation, note that for every channel $j \in \{1, 2, 3, 4\}$,

$$|\alpha \cdot p_{k,j}| = \alpha |p_{k,j}| \leq \alpha \cdot \max(|\sin \theta_k|, |\cos \theta_k|) \leq \alpha,$$

since $|\sin \theta_k| \leq 1$ and $|\cos \theta_k| \leq 1$. Hence, for the active channel j^* (where $v_{k,j^*} = 1$):

$$e_{k,j^*} = 1 + \alpha \cdot p_{k,j^*} \geq 1 - \alpha > 0.5,$$

and for every inactive channel $j \neq j^*$ (where $v_{k,j} = 0$):

$$e_{k,j} = \alpha \cdot p_{k,j}, \quad |e_{k,j}| \leq \alpha < 0.5.$$

Therefore, applying the threshold $\tau = 0.5$ to $\mathbf{e}_k$ uniquely identifies $j^* = \arg \max_j e_{k,j}$, which recovers $\mathbf{v}_k$ and hence s_k . The composition token $\mathbf{c}$ is decoded separately as global frequency statistics and does not affect recovery of s_k . Since this recovery holds for every position k , the entire sequence S is uniquely recovered from $\mathbf{E}$, establishing injectivity. $\square$

The proposition shows that HelixPR retains the invertibility property of one-hot encoding, in contrast to frequency-based k -mer representations that may collapse different sequences into the same feature representation.

2.5 Prediction Framework for Enhancer Identification

2.5.1 Data Pre-processing

Since sequence lengths vary across different cell lines in Basith’s dataset, all sequences were zero-padded to a unified length $L_{\max}$, defined as the maximum sequence length observed in the corresponding training set. Padding was appended to the end of each sequence to preserve the original order of the valid positions. In contrast, no padding was required for Liu’s dataset, since all sequences in that benchmark have the same fixed length.

Feature-wise standardization was then applied. Specifically, the channel-wise mean and standard deviation were computed over all valid, non-padded positions in the training set. The resulting statistics were then used to normalize the training, validation, and test sets, thereby avoiding data leakage from the test sets.

2.5.2 CNN-based Sequence Modeling

We implemented a one-dimensional convolutional neural network (CNN) in PyTorch to classify enhancer sequences from their HelixPR representations. The model is designed to capture local sequential patterns from the input feature and progressively transform them into a compact global representation for downstream binary classification.

The proposed CNN consists of four convolutional feature extraction blocks, followed by a global pooling layer and a fully connected classifier. This design enables

the model to learn hierarchical local sequence patterns and aggregate them into a fixed-dimensional representation for binary classification. Each block contains three components: a one-dimensional convolution layer, a rectified linear unit (ReLU) activation function, and a batch normalization layer. Given an input feature map $\mathbf{X}^{(l-1)}$, the output of each block can be formulated as

$$\mathbf{X}^{(l)} = \text{BN}^{(l)} \left(\text{ReLU} \left(\text{Conv1D}^{(l)}(\mathbf{X}^{(l-1)}) \right) \right),$$

All convolutional layers use padding that preserves the sequence length, so the output length remains the same after each convolution. This design maintains positional alignment across layers and facilitates subsequent global feature aggregation.

The four blocks are configured as follows, where d is the input feature dimension and $h = 128$. For HelixPR, one-hot and EIIP, $d = 4$; for k-mer encodings, $d = 4^k$.

1. $d \rightarrow h$ channels with kernel size 3
2. $h \rightarrow 2h$ channels with kernel size 9
3. $2h \rightarrow h$ channels with kernel size 4
4. $h \rightarrow h/2$ channels with kernel size 4

The kernel sizes are chosen to capture patterns at multiple receptive fields. The small kernel in the first block captures short-range nucleotide dependencies, which is also compatible with the three-base periodic characteristics reported in eukaryotic genomic sequences. The larger kernel in Block 2 captures broader local motifs over a scale close to one helical turn of DNA, and the medium-sized kernels in Blocks 3–4 aggregate these local features hierarchically.

After the multi-scale convolutional blocks, adaptive global average pooling aggregates position-wise features into a fixed-length vector:

$$\mathbf{z} = \frac{1}{L} \sum_{i=1}^L \mathbf{X}_i^{(4)},$$

where L is the sequence length and $\mathbf{X}_i^{(4)}$ denotes the feature vector at position i after the fourth convolutional block.

This operation transforms the output of Block 4, whose channel dimension is $h/2 = 64$, into a unified 64-dimensional embedding $\mathbf{z} \in \mathbb{R}^{64}$, regardless of the input sequence length.

The pooled representation is passed to a two-layer fully connected classifier:

$$\hat{\mathbf{y}} = W_2 \text{ReLU}(W_1 \mathbf{z} + b_1) + b_2,$$

where $W_1 \in \mathbb{R}^{32 \times 64}$, $b_1 \in \mathbb{R}^{32}$, $W_2 \in \mathbb{R}^{2 \times 32}$, and $b_2 \in \mathbb{R}^2$. The output $\hat{\mathbf{y}} \in \mathbb{R}^2$ denotes the predicted logits for the binary classification task.

The overall architecture of the proposed model is illustrated in Fig. 1, showing the sequential flow from the HelixPR-encoded input through convolutional blocks to the final binary output.

2.5.3 Training Strategy and Optimization

The model was optimized using the AdamW optimizer with a learning rate of 1×10^{-4} , and cross-entropy loss was adopted as the objective function. Considering

the differences in dataset size, sequence length between the two benchmark datasets, dataset-specific training strategies were adopted.

For Liu’s benchmark dataset, all models were trained for 40 epochs. This relatively short training schedule is appropriate because Liu’s dataset is small, containing only 1,484 training sequences, and all sequences have the same fixed length. These properties make the optimization process less challenging, while limiting the number of epochs helps reduce the risk of overfitting.

For Basith’s benchmark dataset, the original training set of each cell line was further split into a training set and a validation set at a ratio of 9:1. All models were trained for 100 epochs, and the final model for each of the eight cell lines was selected according to its performance on the corresponding validation set. Stratified sampling was used for the validation split to preserve the positive-to-negative ratio.

To assess the stability of the proposed representation under the same model architecture, we repeated the CNN experiments with different embeddings using five random seeds and reported the results as the mean $\pm$ standard deviation. The same set of random seeds was used for all embeddings to ensure a fair comparison.

2.6 Performance Assessment

In this study, six metrics were adopted to evaluate the performance of enhancer prediction: sensitivity (Sn), specificity (Sp), accuracy (Acc), balanced accuracy (Bacc), Matthews correlation coefficient (MCC) and Area Under the Receiver Operating Characteristic Curve (AUC). These metrics provide a comprehensive evaluation from different perspectives.

Let TP denotes the true positive; TN, the true negative; FP, the false positive; FN, the false negative. The metrics are calculated as follows.

$$\begin{aligned} Sn &= \frac{TP}{TP + FN} \\ Sp &= \frac{TN}{TN + FP} \\ Acc &= \frac{TP + TN}{TP + TN + FP + FN} \\ Bacc &= \frac{Sn + Sp}{2} \\ MCC &= \frac{(TP * TN) - (FP * FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \end{aligned}$$

MCC ranges in value from -1 to 1 and other metrics range from 0 to 1. In realistic classification settings, an AUC value greater than 0.5 indicates that the model performs better than random guessing. For all metrics, higher values indicate better performances.

3 Results and Discussions

3.1 Performance on Standard Benchmark Dataset (Liu)

To evaluate the effect of different sequence representations, we first trained the same CNN architecture on Liu’s training set and then tested it on the independent test set

(200 enhancers and 200 non-enhancers) described in Section 2.1.1. All encoding methods share an identical network configuration and training schedule, so performance differences can be attributed primarily to the quality of the sequence representation rather than to differences in model design. The performance is reported as the mean $\pm$ standard deviation across five random seeds in Table 1, with the best result in each column highlighted in bold.

Table 1: Performance of CNN with different embeddings on Liu’s test dataset.

Embedding	ACC	SN	SP	MCC	AUC
HelixPR	0.8434 $\pm$ 0.0038	0.8250 $\pm$ 0.0050	0.8619 $\pm$ 0.0105	0.6874 $\pm$ 0.0080	0.8667 $\pm$ 0.0003
One-hot	0.8149 $\pm$ 0.0077	0.8148 $\pm$ 0.0127	0.8150 $\pm$ 0.0132	0.6299 $\pm$ 0.0154	0.8602 $\pm$ 0.0031
4-mer	0.7496 $\pm$ 0.0124	0.7155 $\pm$ 0.0210	0.7833 $\pm$ 0.0293	0.5004 $\pm$ 0.0252	0.8285 $\pm$ 0.0060
5-mer	0.7504 $\pm$ 0.0119	0.6970 $\pm$ 0.0182	0.8033 $\pm$ 0.0115	0.5033 $\pm$ 0.0235	0.8291 $\pm$ 0.0037
EIIP	0.7663 $\pm$ 0.0176	0.7525 $\pm$ 0.0556	0.7800 $\pm$ 0.0260	0.5338 $\pm$ 0.0345	0.8053 $\pm$ 0.0140

Among all evaluated representations, HelixPR achieved the best overall performance across the reported metrics. Compared with one-hot encoding, the most direct baseline under the same CNN architecture, HelixPR improved the mean ACC from 0.8149 to 0.8434 and the mean MCC from 0.6299 to 0.6803. Improvements were also observed in both SN and SP, indicating that HelixPR led to more balanced classification performance. In addition, the mean AUC increased from 0.8602 to 0.8667, suggesting a modest but consistent gain in overall discriminative ability.

One-hot encoding clearly outperformed the k -mer and EIIP representations, suggesting that preserving explicit nucleotide identity and positional order is important for CNN-based enhancer classification. Its multi-channel sequence representation provides a strong baseline for convolutional feature extraction.

The k -mer representations, including 4-mer and 5-mer encodings, performed worse than both HelixPR and one-hot encoding. Increasing k from 4 to 5 led to slight improvements in most metrics except SN, but the overall performance of the two k -mer settings remained similar. This observation is consistent with the limitation discussed in Section 2.2.2: frequency-based k -mer encodings summarize sequence composition but do not explicitly preserve the positional arrangement of nucleotides.

EIIP achieved slightly higher ACC, SN, and MCC than the k -mer representations, but obtained a lower AUC. This suggests that EIIP can provide useful nucleotide-level physicochemical information, whereas its scalar one-channel representation is less effective than the multi-channel representations of one-hot encoding and HelixPR in the CNN-based setting.

Statistical significance was assessed using two-sided paired t -tests across the same random seeds. Compared with one-hot encoding, HelixPR significantly improved ACC and MCC ($p < 0.05$), while the AUC improvement was not statistically significant. HelixPR significantly outperformed both 4-mer and 5-mer in ACC, MCC, and AUC ($p < 0.01$), and also significantly improved these metrics over EIIP ($p < 0.05$).

Table 2: The performance of different models on Liu’s test dataset

Method	Acc	Sn	Sp	MCC	AUC
HelixPR+CNN	0.8434	0.8250	0.8619	0.6874	0.8667
iEnhancer-GAN	0.7840	0.8110	0.7580	0.5670	-
iEnhancer-5Step	0.7900	0.8200	0.7600	0.5800	-
iEnhancer-ECNN	0.7690	0.7850	0.7520	0.5370	0.8320
iEnhancer-ELM	0.8300	0.8000	0.8600	0.6612	0.8560

Table 2 compares HelixPR+CNN with previously reported methods on Liu’s independent test set. Since some compared methods could not be directly reproduced, their performance values were taken from the corresponding publications. For HelixPR+CNN, we report the mean results over multiple random seeds. Overall, HelixPR+CNN achieves the best performance across all reported metrics. Compared with the strongest baseline, iEnhancer-ELM, HelixPR+CNN improves ACC from 0.8300 to 0.8434, SN from 0.8000 to 0.8250, SP from 0.8600 to 0.8619, MCC from 0.6612 to 0.6874, and AUC from 0.8560 to 0.8667. These results indicate that HelixPR+CNN provides a favorable balance between positive and negative predictions and achieves strong overall discriminative performance on this benchmark.

Instead of relying on a highly complex architecture, HelixPR+CNN adopts a lightweight four-layer CNN, suggesting that an appropriate sequence representation can provide informative features for enhancer classification. It is also worth noting that iEnhancer-ELM incorporates a pre-trained enhancer language model and a multi-scale attention mechanism, whereas HelixPR+CNN does not rely on pre-training. Therefore, the competitive performance of HelixPR+CNN highlights the value of introducing biologically motivated representation design into a relatively simple classification framework.

3.2 Performance on Cell Type-specific Dataset (Basith)

Basith’s dataset provides a more challenging and realistic setting: test sets are class-imbalanced (2:1 ratio of non-enhancers to enhancers), and sequence lengths are not fixed in each cell lines. We therefore report balanced accuracy (Bacc) as the primary metric, together with Sn, Sp, MCC and AUC. Table 3 summarizes the performance of different sequence representations on the eight test subsets of Basith’s independent test dataset, where all encodings were evaluated under the same CNN architecture and training schedule. The performance is reported as the mean $\pm$ standard deviation across five random seeds.

Across all cell-lines, HelixPR achieved the best result on all reported metrics. For example, on GM12878, HEK293, HUVEC, K562, and NHEK, HelixPR reached nearly saturated performance, with BACC values above 0.99 and AUC values close to 1.0. Statistical significance was assessed using two-sided paired t -tests across the same five random seeds, and HelixPR highly significantly outperformed all baseline representations on all eight Basith subsets in BACC, MCC, and AUC ($p < 0.001$). On the relatively more difficult HMEC and NHLF datasets, HelixPR also remained the

Table 3: Performance of CNN with different embeddings on Basith’s test dataset

Dataset	Embedding	BACC	SN	SP	MCC	AUC
GM12878	HelixPR	0.9922 ± 0.0091	0.9939 ± 0.0085	0.9905 ± 0.0121	0.9814 ± 0.0219	0.9993 ± 0.0011
	One-hot	0.7823 ± 0.0182	0.7158 ± 0.1274	0.8489 ± 0.1352	0.5869 ± 0.0636	0.8920 ± 0.0050
	4-mer	0.7939 ± 0.0036	0.6944 ± 0.0155	0.8934 ± 0.0137	0.6035 ± 0.0100	0.8840 ± 0.0046
	5-mer	0.7895 ± 0.0053	0.7134 ± 0.0272	0.8655 ± 0.0277	0.5827 ± 0.0182	0.8742 ± 0.0073
	EIIP	0.7773 ± 0.0310	0.6798 ± 0.1521	0.8747 ± 0.2016	0.6117 ± 0.0944	0.8902 ± 0.0036
HEK293	HelixPR	0.9970 ± 0.0019	0.9988 ± 0.0011	0.9953 ± 0.0042	0.9921 ± 0.0059	0.9999 ± 0.0001
	One-hot	0.7927 ± 0.0174	0.7909 ± 0.1033	0.7945 ± 0.0731	0.5705 ± 0.0156	0.8793 ± 0.0061
	4-mer	0.7590 ± 0.0053	0.7019 ± 0.0324	0.8161 ± 0.0231	0.5104 ± 0.0056	0.8438 ± 0.0034
	5-mer	0.7647 ± 0.0055	0.7125 ± 0.0172	0.8169 ± 0.0188	0.5205 ± 0.0136	0.8452 ± 0.0072
	EIIP	0.7800 ± 0.0109	0.7577 ± 0.0788	0.8024 ± 0.0787	0.5492 ± 0.0289	0.8666 ± 0.0134
HMEC	HelixPR	0.8931 ± 0.0175	0.9411 ± 0.0708	0.8451 ± 0.0734	0.7587 ± 0.0441	0.9758 ± 0.0175
	One-hot	0.7515 ± 0.0161	0.7763 ± 0.1406	0.7267 ± 0.1396	0.4933 ± 0.0202	0.8504 ± 0.0033
	4-mer	0.7176 ± 0.0061	0.7307 ± 0.0205	0.7045 ± 0.0271	0.4136 ± 0.0130	0.7953 ± 0.0093
	5-mer	0.7269 ± 0.0069	0.7178 ± 0.0315	0.7361 ± 0.0315	0.4350 ± 0.0140	0.8063 ± 0.0077
	EIIP	0.7075 ± 0.0297	0.7415 ± 0.2125	0.6734 ± 0.1986	0.4198 ± 0.0310	0.8098 ± 0.0066
HSMM	HelixPR	0.9840 ± 0.0052	0.9878 ± 0.0113	0.9803 ± 0.0087	0.9618 ± 0.0113	0.9990 ± 0.0009
	One-hot	0.7024 ± 0.0071	0.7124 ± 0.0658	0.6925 ± 0.0556	0.3864 ± 0.0127	0.7783 ± 0.0075
	4-mer	0.6772 ± 0.0101	0.6926 ± 0.0214	0.6618 ± 0.0254	0.3357 ± 0.0196	0.7486 ± 0.0075
	5-mer	0.6762 ± 0.0109	0.7312 ± 0.0446	0.6213 ± 0.0287	0.3329 ± 0.0210	0.7432 ± 0.0115
	EIIP	0.6883 ± 0.0170	0.7517 ± 0.1548	0.6249 ± 0.1415	0.3674 ± 0.0242	0.7715 ± 0.0045
HUVEC	HelixPR	0.9980 ± 0.0026	0.9991 ± 0.0010	0.9969 ± 0.0052	0.9948 ± 0.0076	1.0000 ± 0.0000
	One-hot	0.7152 ± 0.0103	0.7765 ± 0.0911	0.6540 ± 0.0741	0.4101 ± 0.0199	0.7910 ± 0.0085
	4-mer	0.6919 ± 0.0042	0.7274 ± 0.0203	0.6564 ± 0.0223	0.3627 ± 0.0083	0.7627 ± 0.0023
	5-mer	0.6939 ± 0.0045	0.7254 ± 0.0148	0.6624 ± 0.0128	0.3666 ± 0.0086	0.7635 ± 0.0050
	EIIP	0.6963 ± 0.0307	0.7257 ± 0.1717	0.6670 ± 0.1126	0.3807 ± 0.0401	0.7780 ± 0.0042
K562	HelixPR	0.9950 ± 0.0044	0.9966 ± 0.0072	0.9934 ± 0.0087	0.9878 ± 0.0117	1.0000 ± 0.0000
	One-hot	0.7707 ± 0.0096	0.8304 ± 0.0529	0.7110 ± 0.0463	0.5133 ± 0.0168	0.8444 ± 0.0053
	4-mer	0.7471 ± 0.0028	0.7632 ± 0.0264	0.7311 ± 0.0259	0.4707 ± 0.0063	0.8254 ± 0.0026
	5-mer	0.7439 ± 0.0106	0.7574 ± 0.0291	0.7305 ± 0.0275	0.4648 ± 0.0214	0.8206 ± 0.0118
	EIIP	0.7282 ± 0.0196	0.7566 ± 0.1383	0.6999 ± 0.1010	0.4412 ± 0.0245	0.8155 ± 0.0084
NHEK	HelixPR	0.9966 ± 0.0010	0.9956 ± 0.0017	0.9975 ± 0.0007	0.9930 ± 0.0019	0.9998 ± 0.0002
	One-hot	0.6912 ± 0.0567	0.5602 ± 0.2253	0.8222 ± 0.1231	0.4080 ± 0.0587	0.8154 ± 0.0068
	4-mer	0.6885 ± 0.0057	0.6901 ± 0.0177	0.6809 ± 0.0119	0.3581 ± 0.0105	0.7534 ± 0.0078
	5-mer	0.6874 ± 0.0084	0.7171 ± 0.0374	0.6577 ± 0.0346	0.3549 ± 0.0162	0.7521 ± 0.0103
	EIIP	0.6838 ± 0.0219	0.7196 ± 0.1638	0.6481 ± 0.1359	0.3587 ± 0.0306	0.7629 ± 0.0147
NHLF	HelixPR	0.9664 ± 0.0176	0.9497 ± 0.0365	0.9830 ± 0.0213	0.9376 ± 0.0327	0.9972 ± 0.0028
	One-hot	0.7502 ± 0.0152	0.7924 ± 0.1106	0.7080 ± 0.1225	0.4844 ± 0.0265	0.8451 ± 0.0056
	4-mer	0.7416 ± 0.0108	0.7624 ± 0.0276	0.7208 ± 0.0076	0.4590 ± 0.0190	0.8197 ± 0.0080
	5-mer	0.7430 ± 0.0045	0.7569 ± 0.0352	0.7291 ± 0.0354	0.4633 ± 0.0099	0.8171 ± 0.0057
	EIIP	0.6967 ± 0.0324	0.7492 ± 0.2353	0.6442 ± 0.1847	0.3973 ± 0.0309	0.8008 ± 0.0155

best-performing representation, although the absolute BACC values were lower than those observed in the easier datasets.

Among other representations, one-hot encoding generally provided a strong baseline, especially in terms of BACC, MCC, and AUC. However, it was not uniformly better than k -mer or EIIP on every metric. For instance, on GM12878, 4-mer achieved a slightly higher BACC than one-hot, and EIIP achieved a higher MCC. On HSMM and K562, EIIP showed higher sensitivity than one-hot, although this was accompanied by lower specificity. Similarly, k -mer representations sometimes achieved higher specificity, such as on GM12878, K562, and NHLF. These results indicate that different encodings produce different sensitivity–specificity trade-offs, while HelixPR consistently improves the overall balance.

The performance of HelixPR+CNN was not uniform across all Basith subsets, suggesting that enhancer prediction difficulty varies among cell-line-specific datasets. Some subsets, such as GM12878, HSMM, and HUVEC, reached very high BACC and AUC values, whereas HMEC and NHLF remained relatively more challenging. For example, on HMEC, HelixPR improved BACC from 0.7515 with one-hot encoding to 0.8931 and increased AUC from 0.8504 to 0.9758, but its final performance was still lower than that observed in the easier subsets. These results indicate that HelixPR consistently improves the baseline representation, while the achievable performance still depends on the intrinsic separability of each cell-line-specific dataset.

To better understand the striking results on several Basith subsets, we added analyses of train–test redundancy, sequence length, GC content, CpG frequency, and nucleotide/dinucleotide composition in Supplementary Material. The results did not reveal direct train–test leakage, and simple sequence features alone were insufficient to explain the CNN performance. Length showed relatively strong discriminative power in some subsets, such as GM12878 and HEK293, but was much weaker in harder subsets such as HMEC. GC content, CpG frequency, and nucleotide/dinucleotide frequencies generally showed lower discriminative power. These findings suggest that sequence length, composition bias, and cell-line-specific sequence properties may affect the difficulty of individual Basith subsets, but they do not fully explain the performance gains. The consistent improvement over one-hot encoding supports the effectiveness of HelixPR in providing additional helix-aware sequence information for enhancer prediction.

Table 4 compares HelixPR+CNN with previously reported methods on the same benchmark dataset. The performance values of baseline methods were taken from the corresponding publications. For HelixPR+CNN, we report the mean values over five random seeds.

HelixPR+CNN consistently outperformed both iEnhancer-ELM and Enhancer-IF across all eight cell lines and all reported metrics (Table 4). Averaged over the eight subsets, HelixPR+CNN achieved a mean Bacc of 0.9778, improving over iEnhancer-ELM (0.7844) by 0.1934 and over Enhancer-IF (0.7693) by 0.2085. The gains in MCC were substantial: HelixPR+CNN obtained a mean MCC of 0.9509, compared with 0.5504 for iEnhancer-ELM and 0.5210 for Enhancer-IF. These results demonstrate the strong and consistent advantage of HelixPR+CNN on Basith’s benchmark datasets.

Notably, the advantage of HelixPR+CNN is particularly evident on the HSMM and HUVEC subsets, which appear to be relatively challenging for the baseline methods. On both subsets, iEnhancer-ELM and Enhancer-IF showed comparatively low performance, with Bacc values around 0.72–0.74 and MCC values below 0.46. In contrast, HelixPR+CNN maintained very high performance, achieving Bacc values above 0.98 and MCC values above 0.96. This contrast suggests that HelixPR+CNN remains effective even on subsets where existing methods exhibit limited classification performance, highlighting the robustness of the proposed representation.

Overall, these results consistently demonstrate that HelixPR provides a more effective sequence representation than conventional encodings and, when coupled with a lightweight CNN, achieves outperforming performance across both Liu’s and Basith’s benchmark datasets.

Table 4: The performance of different models on Basith’s test dataset

Dataset	Method	Bacc	Sn	Sp	MCC	AUC
GM12878	HelixPR+CNN	0.9922	0.9939	0.9905	0.9814	0.9993
	iEnhancer-ELM	0.8222	0.7564	0.8879	0.6475	0.9179
	Enhancer-IF	0.8140	0.7550	0.8730	0.6260	0.9010
HEK293	HelixPR+CNN	0.9970	0.9988	0.9953	0.9921	0.9999
	iEnhancer-ELM	0.8732	0.8666	0.8798	0.7283	0.9443
	Enhancer-IF	0.8170	0.7750	0.8580	0.6250	0.8930
HMEC	HelixPR+CNN	0.8931	0.9411	0.8451	0.7587	0.9758
	iEnhancer-ELM	0.7645	0.7638	0.7652	0.5068	0.8631
	Enhancer-IF	0.7530	0.7400	0.7650	0.4850	0.8390
HSMM	HelixPR+CNN	0.9840	0.9878	0.9803	0.9618	0.9990
	iEnhancer-ELM	0.7193	0.7191	0.7194	0.4179	0.7948
	Enhancer-IF	0.7180	0.7090	0.7270	0.4170	0.7940
HUVCEC	HelixPR+CNN	0.9980	0.9991	0.9969	0.9948	1
	iEnhancer-ELM	0.7334	0.7691	0.6977	0.4417	0.8045
	Enhancer-IF	0.7370	0.7520	0.7220	0.4500	0.8110
K562	HelixPR+CNN	0.9950	0.9966	0.9934	0.9878	1
	iEnhancer-ELM	0.7974	0.8180	0.7767	0.5679	0.8712
	Enhancer-IF	0.7730	0.7730	0.7720	0.5230	0.8390
NHEK	HelixPR+CNN	0.9966	0.9956	0.9975	0.9930	0.9998
	iEnhancer-ELM	0.7766	0.7229	0.8303	0.5453	0.8716
	Enhancer-IF	0.7480	0.7340	0.7620	0.4770	0.8100
NHLF	HelixPR+CNN	0.9664	0.9497	0.9830	0.9376	0.9972
	iEnhancer-ELM	0.7884	0.8236	0.7532	0.5479	0.8623
	Enhancer-IF	0.7940	0.7890	0.7980	0.5650	0.8650

3.3 Comparison between results on Liu’s dataset and Basith’s dataset.

Notably, the improvement of HelixPR over baseline methods is moderate on Liu’s dataset but more pronounced on Basith’s dataset. The two benchmarks differ in several aspects. Liu’s dataset is constructed as a general enhancer prediction benchmark by combining enhancer sequences from multiple cell lines, and all sequences have a fixed length of 200 bp. In contrast, the Basith benchmark evaluates enhancer prediction separately for each cell line, with larger training sets and variable sequence lengths ranging from 204 bp to 2000 bp.

These differences may affect the relative contribution of different sequence representations. On Liu’s dataset, one-hot encoding with CNN already provides competitive performance, and HelixPR mainly offers an additional improvement. On Basith’s dataset, the longer sequences and cell-line-specific subsets provide a setting in which positional and periodic sequence information can contribute more substantially.

Overall, the results on the two datasets indicate that HelixPR is effective across different benchmark settings. The comparison further suggests that helical positional information is a useful complement to conventional nucleotide representations, especially in datasets with longer sequences, larger sample sizes, and cell-line-specific enhancer prediction tasks.

3.4 Analysis of Feature Space: Feature Visualisation

To further examine the structure of the feature space learned by the model, we visualized the feature representations extracted from the model on both benchmark datasets, as shown in Fig. 2 and Fig. 3. Specifically, for each sample, we extracted

the 64-dimensional output of the global pooling layer after the multi-scale convolutional blocks as its learned embedding. These features were then projected into a two-dimensional space using principal component analysis (PCA). For each dataset, the PCA model was fitted on training data and subsequently applied to the corresponding test data, ensuring that both training and test samples were visualized in a shared feature space. For visualization, nonenhancer and enhancer samples are shown in blue and red, respectively. Within each class, training samples are plotted as circles in lighter shades of the same color, whereas test samples are represented by darker triangles to facilitate visual distinction from the training distribution.

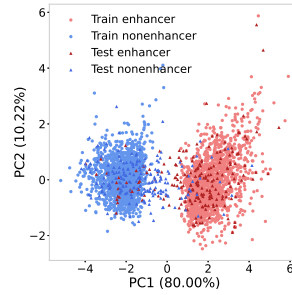


Fig. 2: PCA visualization for Liu's dataset

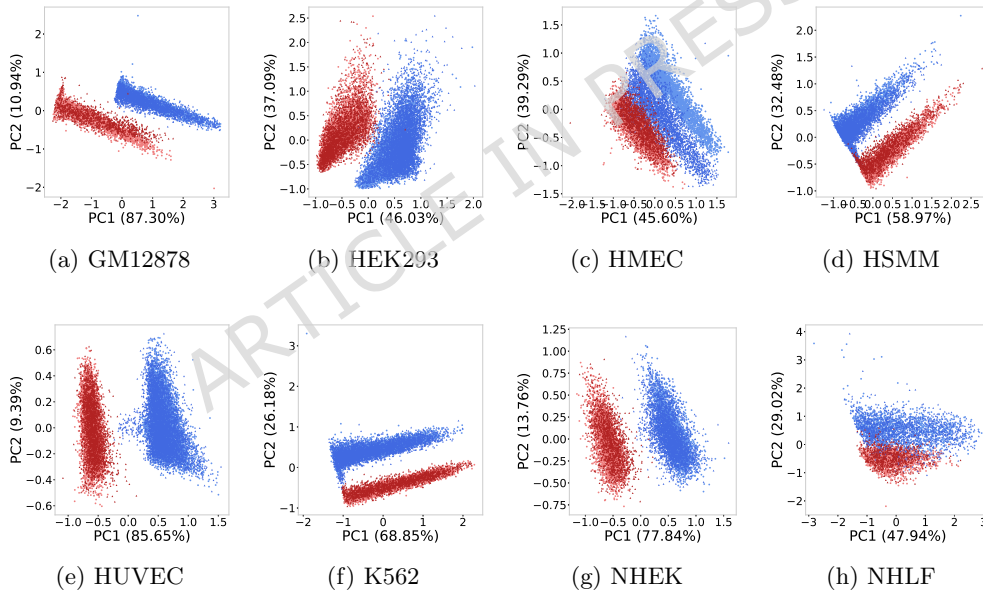


Fig. 3: PCA visualization for Basith's dataset

On Liu’s dataset and across all eight subsets of Basith’s dataset, the PCA visualizations show a consistent pattern in the learned feature space. Enhancer and non-enhancer samples from the training set are clearly separated, and most test samples are projected to the corresponding regions defined by the training distribution. This indicates that the model learns discriminative representations that are largely preserved across training and test data.

Interestingly, the geometric structure of the projected features differs between the two benchmarks. The subsets of Basith’s dataset generally exhibit elongated, ellipse-like or cone-shaped distributions, whereas Liu’s dataset shows two more compact and approximately circular class clusters. Despite these differences, the class-wise separation remains evident across datasets, suggesting that the proposed method can adapt to different data distributions.

Overall, these PCA results provide qualitative evidence that the proposed model learns robust and discriminative representations for enhancer classification. The alignment between training and test samples in the projected feature space further supports the generalization ability of the learned HelixPR-based features.

3.5 Transfer Learning to Enhancer Strength Classification

Building on the enhancer identification results, we further evaluate whether the HelixPR-based representation learned for enhancer identification can be transferred to enhancer strength classification, which aims to distinguish strong enhancers from weak enhancers. This task is more challenging than binary enhancer identification because both strong and weak enhancers belong to the enhancer class and therefore share many common sequence characteristics. As a result, the discriminative patterns between them are more subtle and harder to capture. In addition, the available enhancer-strength training set is relatively small, containing 742 strong and 742 weak enhancer sequences. Training a model from scratch on this limited dataset may not fully exploit the general enhancer-related features learned from the enhancer identification task.

To address this issue, we adopt a transfer learning strategy. The convolutional backbone pretrained on enhancer identification is used to initialize the strength classification model, while the original output layer is replaced with a new task-specific classification head for distinguishing strong and weak enhancers. A two-phase fine-tuning procedure is then applied. In the first phase, the pretrained backbone is frozen, and only the newly initialized classification head is optimized to adapt to the strength classification task. In the second phase, all parameters are unfrozen and jointly optimized with a smaller learning rate, allowing the model to refine the transferred feature representations in an end-to-end manner. The model is trained for 40 epochs in total, and the test performance is reported in Table 5.

The proposed fine-tuning framework achieved substantial performance improvements over existing baseline methods on the strong/weak enhancer classification task. Compared with the best baseline, our model improves MCC from 0.2800 to 0.6913, yielding a gain of over 0.4. This marked increase in MCC indicates that the proposed model provides a more reliable and balanced separation between strong and weak enhancers, rather than improving performance by favoring only one class. The proposed model also raises accuracy from 0.6350 to 0.8400 and AUC from 0.6801 to

Table 5: The performance of different models on enhancer strength test dataset

Method	Acc	Sn	Sp	MCC	AUC
HelixPR+CNN	0.8400	0.9300	0.7500	0.6913	0.9346
iEnhancer-5step	0.6350	0.7400	0.5300	0.2800	-
iEnhancer-EL	0.6100	0.5400	0.6800	0.2222	0.6801
iEnhancer-2L	0.6050	0.4700	0.7400	0.2181	0.6678

0.9346, suggesting a much stronger ability to rank enhancer sequences according to their strength-related discriminative patterns. In addition, the model achieves higher sensitivity while also obtaining higher specificity. These results demonstrate that the fine-tuning strategy effectively transfers general enhancer-related knowledge learned from enhancer identification to the more fine-grained enhancer strength classification task.

3.6 Motif enrichment in HelixPR-highlighted regions

To examine whether HelixPR-highlighted regions contain biologically meaningful regulatory patterns, we performed attribution-guided motif enrichment analysis. Specifically, we selected correctly predicted enhancer sequences from the independent test set and used integrated gradients to calculate nucleotide-level importance scores. Fixed-length windows centered on high-importance positions were extracted as important regions. For each important window, a matched background window of the same length was sampled from the same sequence while avoiding overlap with the important region. We then performed motif enrichment analysis using SEA [24] with the JASPAR CORE vertebrate non-redundant motif database [25]. Motifs with q-value < 0.05 were considered significant.

Using the HEK293 subset of Basith’s dataset as an example, SEA identified 212 significantly enriched motifs in HelixPR-highlighted regions. The enriched motifs showed clear differences between important and matched background windows. For example, the top-ranked motif was detected in 68.67% of important windows and 51.67% of background windows, with an enrichment ratio of 1.33 and a q-value of 6.42×10^{-11} . Representative enriched motifs are summarized in Table 6. These results suggest that the regions emphasized by the HelixPR-based CNN are enriched for known transcription-factor binding motifs, supporting the biological interpretability of the learned sequence patterns.

3.7 Parameter Analysis and Control Experiments

To determine the two key parameters used in HelixPR, namely the scaling factor α and the periodicity T , we conducted a hyperparameter sensitivity analysis on Liu’s dataset. A controlled one-factor-at-a-time strategy was adopted. Specifically, when the effect of α was analyzed, T was fixed at 10; when the effect of T was analyzed, α was fixed at 0.05. This setting allows the influence of each hyperparameter to be evaluated independently.

Table 6: Representative motif enrichment in HelixPR-highlighted regions on the HEK293 subset. Important and Background denote the proportions of attribution-important and matched background windows containing each motif, respectively.

Motif ID	TF	Consensus	Important (%)	Background (%)	Q-value
MA1150.2	RORB	AWWTRGGTCA	68.67	51.67	6.42×10^{-11}
MA0891.2	GSC2	TAATCC	78.89	63.67	1.67×10^{-10}
MA0117.3	Mafb	AWWNTGCTGAC	36.56	23.67	3.11×10^{-7}
MA1521.2	MAFA	TGCTGASTCAGCA	30.78	19.33	1.95×10^{-6}
MA1520.2	MAF	TGCTGASTCAGCA	34.44	22.78	3.23×10^{-6}
MA0722.2	VAX1	YTAATTA	70.78	59.00	8.33×10^{-6}
MA0648.2	GSC	TAATCC	35.00	23.78	8.33×10^{-6}
MA0888.2	EVX2	TAATTA	64.00	51.89	8.33×10^{-6}
MA0883.2	Dmbx1	RWNMGGATTA	25.44	15.56	8.33×10^{-6}
MA0036.4	GATA2	CTTATCT	44.56	32.78	1.08×10^{-5}

The corresponding performance on the Liu’s test set is shown in Fig. 4. For the scaling factor α , HelixPR achieves the most favorable overall performance at $\alpha = 0.05$. Although larger values of α , such as 0.1 and 0.5, lead to slightly higher AUC values, their ACC and MCC decrease, indicating less balanced classification performance.

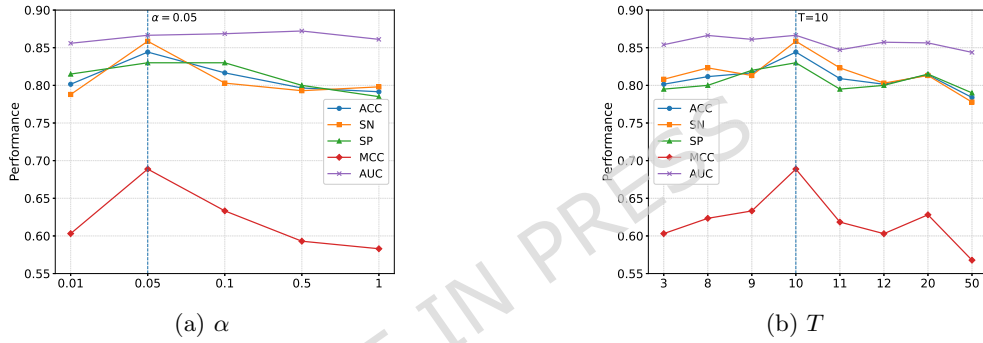


Fig. 4: Parameter comparison on Liu’s dataset

For the periodicity parameter T , we further evaluated several values around and beyond the DNA helical periodicity, including $T = 3, 8, 9, 10, 11, 12, 20$, and 50 . The results show that $T = 10$ achieves the best overall performance, with the highest ACC, SN, MCC, and AUC among the tested periods. In comparison, nearby periods such as $T = 8$ and $T = 9$ also produce competitive AUC values, but their ACC and MCC are lower than those obtained with $T = 10$. Periods farther from 10, such as $T = 12, 20$, and 50 , generally lead to weaker performance. These observations are consistent with the biological prior that one DNA helical turn is approximately 10 base pairs, supporting the use of $T = 10$ in HelixPR. Based on the above analysis, we set $\alpha = 0.05$ and $T = 10$ in all experiments.

To further examine whether the improvement of HelixPR comes from DNA helical periodicity rather than simply introducing additional positional information, we

compared HelixPR with several control representations. As summarized in Table 7, all positional encoding controls were constrained to the same four-dimensional input space and were added to the one-hot vectors before being fed into the same CNN backbone. This design avoids confounding effects from increased input dimension or model capacity.

Table 7: Comparison of HelixPR with different control representations on Liu’s dataset.

Method	ACC	SN	SP	MCC	AUC
One-hot	0.8166	0.8050	0.8283	0.6334	0.8568
One-hot plus CLS token	0.8166	0.8150	0.8182	0.6332	0.8587
CLS token only	0.7513	0.7250	0.7778	0.5034	0.8057
Standard positional encoding	0.7814	0.8000	0.7626	0.5631	0.8480
Random-period encoding	0.7990	0.8050	0.7929	0.5980	0.8434
Learned positional encoding	0.7990	0.8150	0.7828	0.5982	0.8592
HelixPR without CLS token	0.8241	0.8050	0.8434	0.6488	0.8603
HelixPR	0.8434	0.8250	0.8619	0.6874	0.8667

The composition-token-only model uses only global A/C/G/T frequencies as input, reducing the sequence-level CNN to an MLP classifier. Its relatively low performance indicates that global nucleotide composition provides only limited discriminative information when used alone. Combining the composition token with one-hot encoding achieved a substantial performance improvement over the composition-token-only model, showing that sequence-level nucleotide patterns are essential for this task. Moreover, the one-hot plus composition-token model also outperformed one-hot encoding alone and produced more balanced sensitivity and specificity, suggesting that the global composition token provides complementary information and helps the model better capture differences between positive and negative samples. Nevertheless, this representation still did not reach the performance of the full HelixPR model, indicating that the helix-aware positional information further contributes to the effectiveness of HelixPR.

We then compared HelixPR with several positional encoding controls. The standard sinusoidal positional encoding introduces generic absolute positional information, while the random-period encoding provides periodic signals that are not tied to DNA helical structure. Both controls produced lower overall performance than HelixPR, suggesting that the gain is not simply due to adding positional information or arbitrary periodicity. The learnable positional encoding achieved relatively competitive AUC, but its ACC and MCC remained lower than those of HelixPR, implying that unconstrained learnable position encoding may not provide the same balanced classification ability as the biologically motivated helical prior.

In addition, HelixPR without the CLS token was evaluated to isolate the contribution of the helical positional representation itself. This variant outperformed one-hot encoding and other generic positional encoding controls on most metrics, indicating that the helical positional component already provides useful sequence-level information. The full HelixPR model further improved the performance after incorporating the

CLS token, suggesting that global composition information and DNA helical periodicity are complementary. Overall, these control experiments show that HelixPR benefits from the joint modeling of nucleotide identity, global composition information, and biologically meaningful helical periodicity.

4 Conclusions

In this study, we proposed HelixPR, a helix-aware periodic DNA sequence representation for enhancer identification. Unlike conventional encoding methods that mainly treat DNA as a linear symbolic sequence, HelixPR explicitly incorporates the periodic geometry of the DNA double helix into the input representation. By combining per-nucleotide one-hot content, helix-aware Fourier positional encoding, and a global nucleotide-composition token, HelixPR provides a compact, interpretable, and biologically motivated representation for regulatory sequence modeling.

Experimental results on two standard benchmark datasets demonstrate that these design advantages translate into practical predictive gains. Under the same CNN architecture, HelixPR consistently outperformed one-hot, EIIP, and frequency-based k -mer representations, highlighting the benefit of incorporating helical periodic information into DNA sequence features. Moreover, HelixPR combined with a lightweight four-layer CNN achieved better overall performance than previously reported methods, including models based on large pre-trained language models. On Liu’s benchmark, the HelixPR-based CNN also showed strong transferability when applied to enhancer strength classification. These results suggest that accurate enhancer prediction and downstream strength classification can be achieved without relying on complex architectures, emphasizing the importance of biologically meaningful representation design in genomic prediction tasks. PCA-based visualizations further showed that HelixPR helps the CNN learn more discriminative feature spaces across different benchmarks. In addition, attribution-guided motif enrichment analysis showed that HelixPR-highlighted regions were enriched for known transcription-factor binding motifs, supporting the biological interpretability of the learned sequence patterns.

The ablation and parameter analyses further confirmed the effectiveness of the main components of HelixPR. The Fourier positional encoding improved the one-hot baseline, and the global composition token provided complementary sequence-level information. Notably, the best performance was obtained when the helical period was set to $T = 10$, consistent with the known 10–11 bp periodicity of the DNA double helix. HelixPR also outperformed alternative positional encoding strategies, suggesting that its advantage comes from the biologically motivated helical prior rather than from simply adding positional information.

Another important property of HelixPR is its injectivity. In contrast to frequency-based k -mer representations, which cannot establish a one-to-one correspondence with the original sequence, HelixPR retains the invertibility of one-hot encoding and supports exact sequence recovery. Beyond one-hot encoding, HelixPR further incorporates biologically meaningful periodic information while preserving nucleotide identity.

Future work may explore adaptive periodic encoding schemes that can account for variations in DNA helical structure across different genomic contexts. In addition, integrating HelixPR with other regulatory genomic features, such as chromatin accessibility and transcription factor binding profiles, may further improve enhancer modeling. Overall, this study suggests that incorporating the periodic geometry of the DNA double helix into sequence representations is a promising direction for computational enhancer analysis and broader regulatory sequence modeling.

Declarations.

Funding

This work is supported in part by funds from the National Science Foundation of China (Grant No. # 12171275), Tsinghua University Education Foundation and China Postdoctoral Science Foundation (Grant No. 2025M783075).

Conflicts of Interest

The authors declare that they have no competing interests.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Data and Code Availability

The datasets analysed in this work were obtained from publicly available benchmark datasets. Liu's benchmark dataset was obtained from the Supplementary Information S1 of iEnhancer-2L [8] and the Supplementary Data of iEnhancer-EL [11]. Basith's dataset was obtained from the Enhancer-IF [22] data download page: <http://thegleelab.org/Enhancer-IF/EnhancerData.html>.

The processed datasets and source code used in this study are publicly available at <https://github.com/xiangShi22/HelixPR>.

Author contribution

S.S.-T.Y. conceived and designed the study. X.S. conducted the experiments, and J.K. analyzed the results. X.S. wrote the original draft, J.K., N.S. and S.S.-T.Y. reviewed the manuscript.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used AI-based editing tools in order to improve language and readability. After using this tool, the authors reviewed and edited the content as needed and took full responsibility for the content of the publication.

References

- [1] Long, H.K., Prescott, S.L., Wysocka, J.: Ever-changing landscapes: Transcriptional enhancers in development and evolution. *Cell* **167**(5), 1170–1187 (2016)
- [2] Kulaeva, O.I., Nizovtseva, E.V., Polikanov, Y.S., Ulianov, S.V., Studitsky, V.M.: Distant activation of transcription: Mechanisms of enhancer action. *Molecular and Cellular Biology* **32**(24), 4892–4897 (2012). PMID: 23045397
- [3] Hnisz, D., Abraham, B., Lee, T., Lau, A., Saint-André, V., Sigova, A., Hoke, H., Young, R.: Super-enhancers in the control of cell identity and disease. *Cell* **155**(4), 934–947 (2013)
- [4] Shlyueva, D., Stampfel, G., Stark, A.: Transcriptional enhancers: from properties to genome-wide predictions. *Nature Reviews Genetics* **15**(4), 272–286 (2014)
- [5] Visel A, P.L. Rubin EM: Genomic views of distant-acting enhancers. *Nature* **461**(7261), 199–205 (2009)
- [6] Kvon, E.Z., Waymack, R., Gad, M., Wunderlich, Z.: Enhancer redundancy in development and disease. *Nature Reviews Genetics* **22**, 324–336 (2021)
- [7] Lomvardas, S., Barnea, G., Pisapia, D.J., Mendelsohn, M., Kirkland, J., Axel, R.: Interchromosomal interactions and olfactory receptor choice. *Cell* **126**(2), 403–413 (2006)
- [8] Liu, B., Fang, L., Long, R., Lan, X., Chou, K.-C.: ienhancer-2l: a two-layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition. *Bioinformatics* **32**(3), 362–369 (2016)
- [9] Chen, W., Lei, T.-Y., Jin, D.-C., Lin, H., Chou, K.-C.: Pseknc: A flexible web server for generating pseudo k-tuple nucleotide composition. *Analytical Biochemistry* **456**, 53–60 (2014)
- [10] Erwin, G., Truty, R., Kostka, D., Pollard, K., Capra, J.: Integrating diverse datasets improves developmental enhancer prediction. *PLoS computational biology* **10** (2013)
- [11] Liu, B., Li, K., Huang, D.-S., Chou, K.-C.: ienhancer-el: identifying enhancers and their strength with ensemble learning approach. *Bioinformatics* **34**(22), 3835–3842 (2018)
- [12] Cai, L., Ren, X., Fu, X., Peng, L., Gao, M., Zeng, X.: ienhancer-xg: interpretable sequence-based enhancers and their strength predictor. *Bioinformatics* **37**(8), 1060–1067 (2020)
- [13] Nair, A., Pillai, S.: A coding measure scheme employing electron-ion interaction pseudo potential (eiip). *Bioinformation* **1**, 197–202 (2006)

- [14] Ioshikhes, I., Bolshoy, A., Derenshteyn, K., Borodovsky, M., Trifonov, E.N.: Nucleosome dna sequence pattern revealed by multiple alignment of experimentally mapped sequences. *Journal of Molecular Biology* **262**(2), 129–139 (1996)
- [15] Li, J., Wu, Z., Lin, W., Luo, J., Zhang, J., Chen, Q., Chen, J.: ienhancer-elm: improve enhancer identification by extracting position-related multiscale contextual information based on enhancer language models. *Bioinformatics Advances* **3**(1), 043 (2023)
- [16] Yao, L., Xie, P., Guan, J., Chung, C.-R., Huang, Y., Pang, Y., Wu, H., Chiang, Y.-C., Lee, T.-Y.: Capsenhancer: an effective computational framework for identifying enhancers based on chaos game representation and capsule network. *Journal of chemical information and modeling* **64**(14), 5725–5736 (2024)
- [17] Yao, L., Xie, P., Dong, D., Guo, Y., Guan, J., Zhang, W., Chung, C.-R., Zhao, Z., Chiang, Y.-C., Lee, T.-Y.: Caps-ac4c: an effective computational framework for identifying n4-acetylcytidine sites in human mrna based on deep learning. *Journal of Molecular Biology* **437**(6), 168961 (2025)
- [18] Zhurkin, V.B.: Periodicity in dna primary structure is defined by secondary structure of the coded protein. *Nucleic acids research* **9**(8), 1963–1971 (1981)
- [19] Herzel, H., Weiss, O., Trifonov, E.N.: 10-11 bp periodicities in complete genomes reflect protein structure and dna folding. *Bioinformatics (Oxford, England)* **15**(3), 187–193 (1999)
- [20] Shi, X., Kang, J., Sun, N., Zhao, X., Yau, S.S.-T.: Multi-perspective natural vector: A novel method for viral sequence feature extraction. *Journal of Computational Biology* (2025). PMID: 41355777
- [21] Fu, L., Niu, B., Zhu, Z., Wu, S., Li, W.: Cd-hit: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**(23), 3150–3152 (2012)
- [22] Basith, S., Hasan, M.M., Lee, G., Wei, L., Manavalan, B.: Integrative machine learning framework for the identification of cell-specific enhancers from the human genome. *Briefings in Bioinformatics* **22**(6), 252 (2021)
- [23] Asgari, E., Mofrad, M.R.: Continuous distributed representation of biological sequences for deep proteomics and genomics. *PloS one* **10**(11), 0141287 (2015)
- [24] Bailey, T.L., Grant, C.E.: SEA: Simple enrichment analysis of motifs. *bioRxiv* (2021)
- [25] Ovek Baydar, D., Rauluseviciute, I., Aronsen, D.R., Blanc-Mathieu, R., Bonthuis, I., de Beukelaer, H., Ferenc, K., Jegou, A., Kumar, V., Lemma, R.B., Lucas, J., Pochon, M., Yun, C.M., Ramalingam, V., Deshpande, S.S., Patel, A., Marinov,

G., Wang, A.T., Aguirre, A., Castro-Mondragon, J., Baranasic, D., Chèneby, J., Gundersen, S., Johansen, M., Khan, A., Kuijjer, M.L., Hovig, E., Lenhard, B., Sandelin, A., Vandepoele, K., Wasserman, W., Parcy, F., Kundaje, A., Mathelier, A.: JaspAr 2026: expansion of transcription factor binding profiles and integration of deep learning models. *Nucleic Acids Research* **54**(D1), 184–193 (2026)

ARTICLE IN PRESS