



OPEN DNA promoter task-oriented dictionary mining and prediction model based on natural language technology

Ruolei Zeng¹, Zihan Li²✉, Jialu Li² & Qingchuan Zhang²✉

Promoters are essential DNA sequences that initiate transcription and regulate gene expression. Precisely identifying promoter sites is crucial for deciphering gene expression patterns and the roles of gene regulatory networks. Recent advancements in bioinformatics have leveraged deep learning and natural language processing (NLP) to enhance promoter prediction accuracy. Techniques such as convolutional neural networks (CNNs), long short-term memory (LSTM) networks, and BERT models have been particularly impactful. However, current approaches often rely on arbitrary DNA sequence segmentation during BERT pre-training, which may not yield optimal results. To overcome this limitation, this article introduces a novel DNA sequence segmentation method. This approach develops a more refined dictionary for DNA sequences, utilizes it for BERT pre-training, and employs an Inception neural network as the foundational model. This BERT-Inception architecture captures information across multiple granularities. Experimental results show that the model improves the performance of several downstream tasks and introduces deep learning interpretability, providing new perspectives for interpreting and understanding DNA sequence information. The detailed source code is available at https://github.com/katouMegumiH/Promoter_BERT.

Promoters are DNA sequences that initiate transcription and regulate gene expression in organisms¹. The TATA box is a common promoter sequence in many eukaryotic genes, recognized by transcription factors that facilitate the formation of the RNA polymerase transcription complex, thereby promoting transcription². Given the crucial role promoters play in gene transcription, accurately predicting their locations is essential for understanding gene expression patterns and the function of genetic regulatory networks.

The human genome has often been referred to as the 'book of life'³, with DNA described as the 'language of the cell'. Due to the similarities between natural language and biological sequences, deep learning methods, particularly those incorporating natural language processing, have become increasingly popular in bioinformatics. For instance, Zhu et al.⁴ introduced Depicter, a deep learning approach utilizing CNNs and capsule networks to predict promoters, demonstrating high performance. Similarly, Oubounyt et al.⁵ employed a combination of convolutional neural networks (CNNs) and Long Short-Term Memory (LSTM) networks to effectively characterize and identify promoter sequences in human and mouse genomes. In addition, BERT is increasingly used in DNA prediction analysis studies due to its excellent performance in NLP^{6,7}.

However, most existing studies that apply BERT to DNA sequence prediction, such as DNABERT⁸, rely on the K-mer method to construct the model's dictionary, treating DNA fragments of length k as "words." While the K-mer approach is widely adopted, it suffers from a key limitation: it arbitrarily segments DNA sequences without incorporating prior biological knowledge about the functional significance of specific sequences. This method treats DNA fragments as simple strings, without understanding their roles in biological processes, which can lead to the loss of important context and patterns necessary for accurate promoter identification. By relying on K-mers, these methods may miss subtle biological interactions and sequence dependencies that are critical for predicting functional regions like promoters. In contrast, our study avoids arbitrary segmentation by applying a more refined strategy to construct a dictionary for DNA sequences. Instead of relying on predefined K-mer lengths, we propose a segmentation approach inspired by natural language processing algorithms that better captures the complex structure of DNA sequences.

¹Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN 55455, USA. ²National Engineering Research Centre for Agri-Product Quality Traceability, Beijing Technology and Business University, No.11 Fucheng Road, Beijing 100048, China. ✉email: 18811325239@163.com; zhangqingchuan@btbu.edu.cn

Specifically, This article introduces a segmentation method based on words in DNA and constructed a dictionary with reference to new word discovery algorithms in natural language processing, and then pre-trained the BERT model. Specifically, we used the constructed dictionary as the basis for word segmentation in DNA sequences and pre-trained the BERT model by using the processed DNA fragments as a large-scale pre-defined input language for BERT, which outputs a digital vector that captures the required information such as context, meaning and syntax equivalent to human language, as shown in Fig. 1 (a). For the splicing of downstream models, we borrowed model structures from^{6–8}, etc., and improved on the BERT-CNN to implement a BERT-Inception structure, which increases its depth and width compared to the traditional convolutional neural network (CNN) in order to receive data of different granularity and learn features in parallel, with final classification performed by a fully connected (FC) layer, as shown in Fig. 1 (b). Experiments demonstrate the model improved performance in several downstream tasks and introduce deep learning interpretability, providing new insights into the interpretation and decipherment of DNA sequence information, as shown in Fig. 1 (c).

Related work

Recent studies have shown an increasing interest in biological sequence feature representation methods in bioinformatics, particularly for DNA sequences⁹. An effective DNA sequence feature representation method is essential for deciphering non-coding DNA sequences, a key challenge in biological sequence analysis. Deep learning-based methods have demonstrated outstanding performance in various tasks, such as predicting protein functions¹⁰, identifying electron transport proteins¹¹, and detecting antiviral peptides of varying lengths¹². These methods leverage neural networks like LSTM and CNN to achieve significant advantages.

Biological sequences like DNA and protein sequences can be viewed as textual data, showing parallels with human language. This perspective has led to the development of models like ProtVec and GeneVec for protein and gene sequences, respectively¹³. Additionally, the adaptation of the FastText model for DNA sequences, including enhancers and promoters, has shown promising results^{1,14}. Building on this idea, the application of pre-trained language models like BERT in bioinformatics classification tasks has become increasingly popular¹⁵. By processing DNA sequences as text, segmenting them, and using them to pre-train BERT, researchers have achieved excellent results in tasks such as promoter sequence identification¹⁶, N6-methyladenine prediction¹⁷, and enhancer sequence classification¹⁸. The "piRNA (2L)-PseKNC" model by Khan et al., which combines hybrid pseudo-K-tuple nucleotide composition (PseKNC) and deep neural networks (DNN) for the accurate prediction of piRNA sequences and their functions¹⁹. The development of iProm-Zea by Kim et al., which employs a two-layer convolutional neural network (CNN) for identifying TATA and TATA-less promoters in Zea mays, demonstrates the effective use of deep learning for promoter classification²⁰. In addition, the surge in deep learning's popularity is driven by increased computing power and the availability of extensive sequencing data. These advancements have enabled the application of deep learning techniques to genome functional annotation recognition. For instance, LSTM models have been utilized to predict nucleosome dynamic intervals

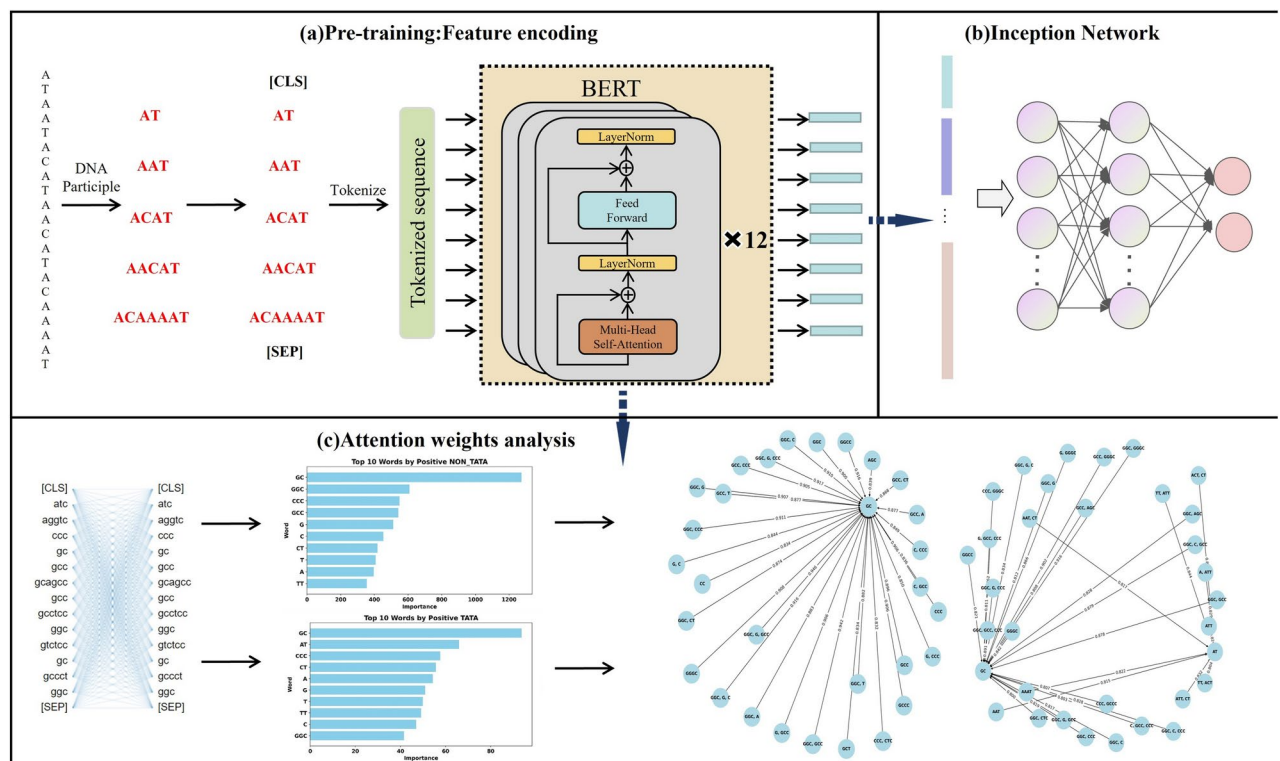


Figure 1. Overview of the framework.

(NDIs) from time-series data with high accuracy²¹. Similarly, CNNs have proven effective in transcription factor identification²². Furthermore, RNN models have successfully predicted the fluorescence characteristics of DNA-templated silver nanoclusters (AgNCs)²³. Transformer models have likewise achieved state-of-the-art results in predicting DNA methylation patterns²⁴.

These deep learning approaches have significantly advanced the understanding of gene regulation, genome organization, and mutation effects. For example, models like scBERT have improved cell type annotation and gene significance interpretability from single-cell RNA sequencing data²⁵. The ExplaiNN model employs interpretable neural networks for genomics, successfully identifying transcription factor binding patterns²⁶. Additionally, the GenNet framework has demonstrated deep learning's potential in predicting phenotypes from genetic data²⁷. Knowledge-primed neural networks (KPNNs) combine biological knowledge with deep learning to enhance prediction accuracy and biological interpretability²⁸. Furthermore, Helixer, a deep-learning-based framework, has improved eukaryotic genome annotation²⁹. Interpretive studies of Transformer models like DNABERT and Transformer-XL have shown improved model interpretability through attention graph analysis^{8,30,31}. In cancer research, biologically guided deep learning models have significantly improved prognosis and response predictions to cancer immunotherapy by integrating tumor biological parameters³².

By integrating these developments, it is evident that deep learning has profoundly impacted bioinformatics, offering powerful tools for analyzing and interpreting complex biological data.

Material and methods

This section introduces a methodology for segmenting DNA sequences into meaningful 'words' and a subsequent algorithm for disambiguating these segments. These techniques are crucial for analyzing and interpreting genomic data more effectively.

Dataset

Large-scale domain corpora are frequently utilized in NLP tasks to train models that can produce compressed feature representations that allow downstream models to capture sequence information more effectively. The reference genome hg19 dataset (FASTA file) and the annotations of the corresponding sequences (BED file) were downloaded from UCSC, and the final dataset for training the DNA language model was constructed.

Predicting gene promoters is a significant challenge in bioinformatics. For our study, we selected the TATA, non-TATA, and combinatorial datasets from the experiment conducted by Yanrong et al.⁸, which utilized human TATA and non-TATA promoters, each 10,000 bp in length, sourced from the Eukaryotic Promoter Database (EPDnew)³³.

DNA dictionary construction for sequence segmentation

Constructing biologically relevant dictionaries for DNA sequences involves two key steps: segmenting DNA sequences into meaningful "words" and disambiguating these segments to filter out erroneous candidates to construct the DNA dictionary, as shown in Fig. 2.

DNA-based word segmentation method

In order to be able to rationally divide the "words" in a DNA sequence, a DNA dictionary must first be constructed, as shown in Fig. 2 (a). First, the DNA sequence is genetically cleaved, with each gene considered as a sentence and each base as a word; second, any two neighboring words in the sentence are studied. There are two possibilities for two neighboring words, either they belong to a common word or they are boundaries of two words. Since the correlation of two neighboring words belonging to the same word must be stronger than the correlation of two neighboring words belonging to different words, that is to say, if the probability of two words appearing together is greater than the probability of them being randomly aligned together, we consider that these two words belong to the same word and vice versa. Based on the above idea, the algorithm is as follows: the sentence is scanned word by word, and if two neighboring words satisfy the formula (1), they are considered to be a part of the word until the two neighboring words do not satisfy the formula, and then the two words are disconnected. In this way, the sentence can be divided into different words and the set of different words is the set of candidate words.

$$P(W_i)(W_{i+1}) > P(W_i) * P(W_{i+1}) \quad (1)$$

where $P(W_i)$ is the probability of occurrence of a single word; $P(W_i)(W_{i+1})$ is the probability of co-occurrence of two words adjacent to each other, $i \in \{1, 2, 3 \dots, n\}$, n is the sentence length.

Since there may be 'false words' in the above generated candidate word set, in order to make the generated dictionary more reasonable, the interference of 'false words' in the candidate word set can be excluded by introducing the time dimension based on the generated candidate word set, and the algorithm is named as forgetting. The algorithm is named as forgetting algorithm. In order to count the frequency of the candidate words entered into the database at each moment, the decay indicator R is set, based on the fact that the frequency of 'true words' grows faster than the decay, while for 'false words', due to the randomness of their appearance, the frequency growth will be lower than the decay, i.e., they will be slowly 'forgotten'. The decay indicator R is shown in Eq. (2).

$$R = R_0 e^{-\alpha(t-t_0)} \quad (2)$$

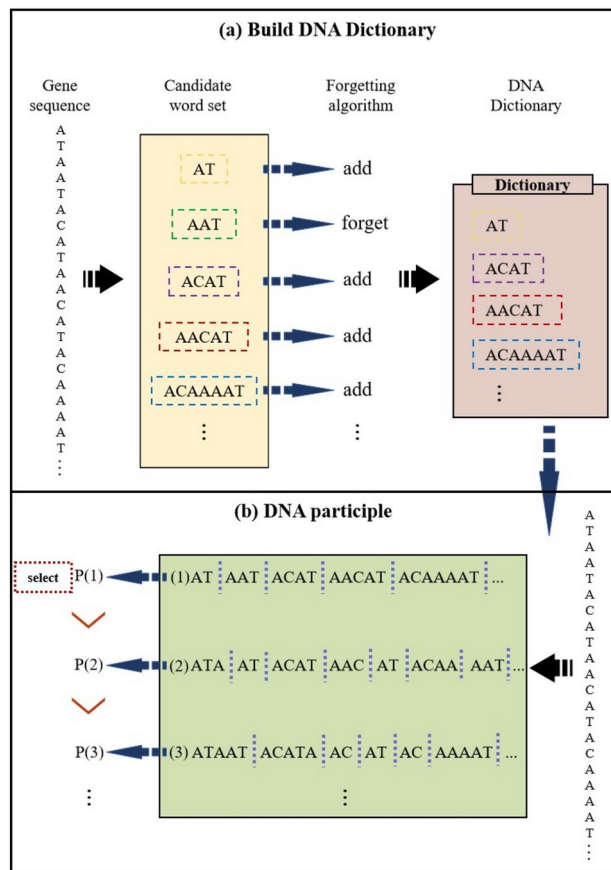


Figure 2. Overview of the DNA-based word segmentation and disambiguation process.

where R_0 is the frequency of a word when it was last added to the candidate word set; t is the current time; t_0 is the time when a word was last added to the candidate word set; and α is a constant.

Therefore, based on the above idea, the frequency of each word is calculated at each moment when it is added to the set of candidate words, and when the frequency is greater than the decay metric R , the word is considered as a "true word", and vice versa for a "false word". The set of all "true words" is the DNA dictionary.

DNA disambiguation algorithm

When splitting DNA sequences into words, a single sequence (sentence S) can have multiple segmentation schemes. The DNA disambiguation algorithm aims to identify the optimal segmentation scheme by maximizing the probability of the sequence occurrence based on the constructed DNA dictionary.

Specifically, The DNA word splitting algorithm is based on the DNA dictionary and treats the gene sequences as sentences and splits the sentences, however, a sentence S can have n splitting methods, then the best splitting method is to ensure that the probability of occurrence of the split sentence S is maximum, that is, the probability of all words in the sentence S occurring at the same time is maximum, as shown in Fig. 2b. Therefore, the algorithm objective is to find a segmentation scheme for each sentence S that maximizes $P(S)$ in Eq. (3).

$$P(S) = P(W_1, W_2, \dots, W_n) \quad (3)$$

Where $P(S)$ is the probability of occurrence of sentence S ; $P(W_1, W_2, \dots, W_n)$ is the joint probability of each word in the word separation scheme.

To achieve this, the algorithm iterates through potential segmentation schemes and computes their probabilities using the DNA dictionary. The scheme with the highest joint probability is selected as the optimal segmentation, as shown in Fig. 2 (b).

BERT pre-training model

As shown in Fig. 3, this step involves tokenizing the segmented DNA words generated from the constructed dictionary, followed by pretraining a BERT model. The pretraining process uses a Masked Language Model

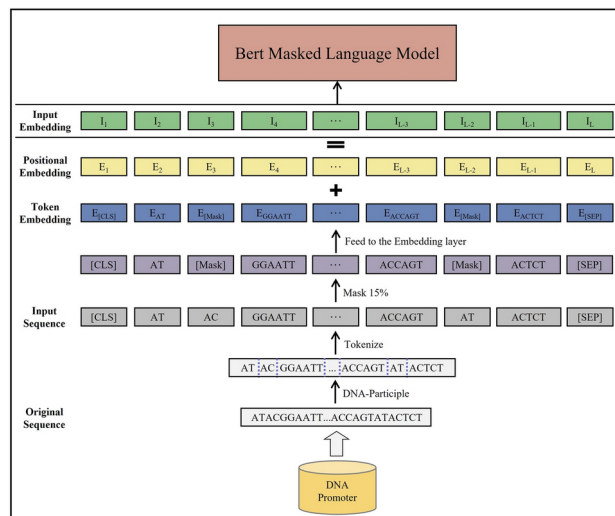


Figure 3. Overview of the BERT pre-training framework for DNA sequences.

(MLM) objective, where 15% of the tokens are randomly masked, allowing the model to learn contextual information within DNA sequences.

BERT is a transformer-based model for contextualized language representation that has outperformed humans in many NLP tasks. The core strategy of BERT involves a combination of pre-training and fine-tuning, where a broad linguistic model is first trained on extensive unlabeled data, making BERT capable of general understanding of text. Then, using task-specific labeled data, the parameters of BERT are fine-tuned to solve various application problems.

The encoder part of BERT consists of multiple transformer layers, each containing a multi-headed self-attentive mechanism and a feedforward network. These layers encode the input sequence, allowing the model to gain an understanding of the relationships between the elements in the sequence. Unlike traditional language models that learn dependencies in a unidirectional manner, BERT's transformer layers are bidirectional, meaning they can learn dependencies from both the left and right contexts of a token. This bidirectional learning is particularly valuable for biological sequences like DNA, where the functional meaning of a subsequence can depend on both the preceding and succeeding bases.

In addition, BERT uses the Masked Language Model (MLM) approach during pretraining, where 15% of the tokens are randomly masked in the input sequence. The model's task is to predict the masked tokens based on the surrounding context. This process forces the model to learn not just the immediate local context of a token but also long-range dependencies across the sequence. By learning to predict masked segments in the DNA sequence, BERT is trained to understand both short-term and long-term contextual relationships within the sequence, allowing it to capture complex biological patterns that may span across distant regions of the sequence.

By leveraging tokenized DNA words and a masked language modeling objective, the pretrained BERT model serves as a robust foundation for downstream tasks, enabling deeper understanding and analysis of DNA sequences.

Inception network

As shown in Fig. 4, this step employs the Inception architecture to classify promoter sequences using embeddings generated by the pretrained BERT model. The Inception network is designed to process features at multiple scales in parallel, enabling the extraction of rich and diverse representations from promoter-specific DNA data.

The inception architecture marks a significant advancement in CNN development³⁴. The core idea of the inception architecture is to use convolutional kernels of different sizes to process the input data and to perform parallel computation in the same layer as shown in Fig. 3. Specifically, it computes the feature map in parallel using a set of convolutional kernels, which includes multiple convolutional kernels of different sizes. Furthermore, the Inception structure utilizes a pooling layer to downsample the feature map, and uses 1×1 convolutional kernels for degradation and scaling to reduce the parameters and computation. With this parallel computation, this architecture is more effective at extracting features from input data, thus improving the accuracy of the classification task. Also, since the Inception structure uses multiple convolutional kernel sizes with 1×1 convolutional kernels for downscaling and scaling, it has a relatively small number of parameters and computation to better avoid overfitting problems. The proposed classifier in this article is built based on the V1 convolutional kernel. It performs a one-dimensional convolution operation on the BERT embedding vectors with several filters of different sizes (kernel size $\in [1,6]$). In addition, maximum pooling is performed. The obtained outputs are then concatenated and sent to the next layer. By performing the convolution in the same layer, the network can obtain a wider view.

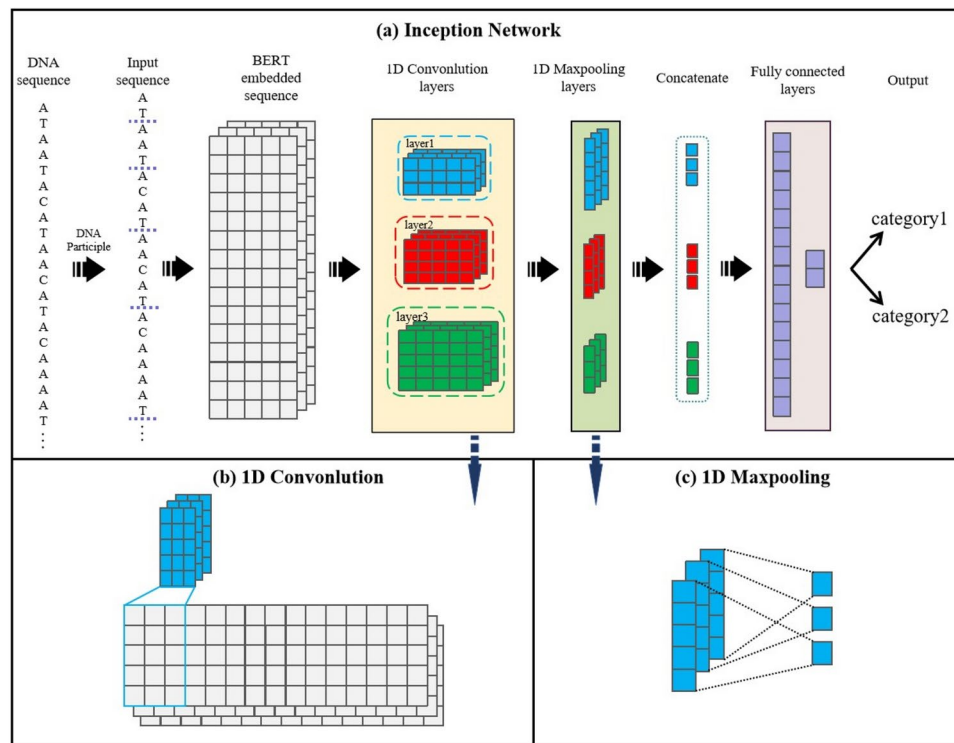


Figure 4. Overview of the Inception network.

Experimental analysis and results

Dataset and settings

Predicting gene promoters remains a highly challenging problem in bioinformatics. This article builds on the datasets used in the experiments by YanRong et al.⁸, specifically utilizing the TATA, non-TATA, and combined datasets sourced from the Eukaryotic Promoter Database (EPDnew)³³. These datasets include human TATA and non-TATA promoter sequences, each 10,000 bp in length.

This article employs multiple metrics to thoroughly assess the model's performance, including Accuracy (ACC), Sensitivity (Sn), Specificity (Sp), Matthews Correlation Coefficient (MCC), F1-Score, and Area Under the Receiver Operating Characteristic Curve (AUC). The formulas for these metrics are detailed below. Accuracy (ACC) is calculated by dividing the total number of correct predictions by the total number of predictions, providing a measure of the model's overall accuracy.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

Sensitivity (Sn), which reflects the ability of the model to recognize positive samples. The formula is shown below:

$$Sn = \frac{TP}{TP + FN} \quad (5)$$

Specificity (Sp) refers to the proportion of negative samples correctly recognized by the model. The formula is shown below:

$$Sp = \frac{TN}{TN + FP} \quad (6)$$

MCC provides a balanced measure that can give a more truthful evaluation of the performance of the model, especially when the dataset is imbalanced. While accuracy may be high in imbalanced datasets due to the model's preference for the majority class, MCC gives a more comprehensive view by considering both classes. Its value ranges from -1 to $+1$, where $+1$ indicates a perfect prediction, 0 indicates no predictive power, and -1 indicates a completely incorrect prediction. Thus, MCC is a more reliable metric when dealing with imbalanced data as it accounts for both false positives and false negatives, helping to ensure that the model is performing well across all classes. The formula is shown below:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{7}$$

F1-Score is the reconciled average of Precision and Recall. It combines the precision and sensitivity of the model to provide a balanced performance metric. The formula is shown below:

$$F1\text{-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{8}$$

where precision is calculated as the number of true positive samples (TP) divided by the total number of predicted positive samples (TP + FP), while recall, also referred to as sensitivity or TPR, is the number of true positive samples (TP) divided by the total number of actual positive samples (TP + FN). The F1-Score provides a single metric to evaluate the model’s predictive performance for the positive class, particularly in situations where the class distribution is highly imbalanced.

AUC, representing the area under the ROC curve, is a statistical measure used to evaluate the strengths and weaknesses of a binary classification model. It compares the true positive rate with the false positive rate, with a value closer to 1 indicating better model performance; an AUC near 0.5 suggests the model’s performance is akin to random chance.

Comparative experiment

In order to scientifically assess the effectiveness of the BERT-Inception model, we selected the DNABERT⁸ model, which is currently one of the best-performing models in the field of promoter recognition tasks, as well as the iProm-Zea²⁰ and DeePromoter⁵ models for comparison experiments. These models represent different approaches to promoter recognition and classification. Additionally, due to the large dataset, 10-fold cross-validation was used for all experiments in order to validate the performance of the model. The results of the comparison experiments are shown in Table 1.

From the comparison results, we can see that the BERT-Inception model outperforms the iProm-Zea, DNABERT, and DeePromoter models in terms of accuracy, F1 score, sensitivity (Sn), specificity (Sp), Matthews correlation coefficient (MCC), and area under the curve (AUC) across all datasets. In particular, for the non-TATA dataset, the BERT-Inception model achieves an accuracy of 0.982, an F1 score of 0.983, and an AUC value of 0.996, which is significantly better than the performance of the other models.

The comparison results show that BERT-Inception outperforms iProm-Zea, DNABERT, and DeePromoter in all evaluation metrics, including accuracy, F1 score, sensitivity, specificity, Matthews correlation coefficient (MCC), and AUC across all datasets. In particular, for the non-TATA dataset, BERT-Inception achieves an accuracy of 0.982, an F1 score of 0.983, and an AUC of 0.996, significantly outperforming all other models. For the TATA dataset, BERT-Inception also outperforms iProm-Zea and DNABERT in terms of accuracy and AUC. Specifically, BERT-Inception achieves an accuracy of 0.976, an F1 score of 0.942, and an AUC of 0.990, while iProm-Zea reaches an accuracy of 0.965, an F1 score of 0.926, and an AUC of 0.980. DNABERT performs similarly to iProm-Zea on this dataset, with an accuracy of 0.965 and an AUC of 0.981. Although iProm-Zea achieves the highest sensitivity (0.972) for identifying TATA promoters, its overall performance still lags slightly behind BERT-Inception in terms of accuracy and AUC. On the Group dataset, BERT-Inception shows its superiority with an accuracy of 0.957, an F1 score of 0.964, and an AUC of 0.993, outperforming both iProm-Zea and DNABERT.

The results highlight that BERT-Inception, when pre-trained on DNA sequence divergences, is able to capture more complex biological patterns and relationships within DNA sequences compared to DNABERT,

Model	Dataset	ACC	F1	Sn	Sp	MCC	AUC
DeePromoter* ⁵	NON-TATA	0.956	0.954	0.950	0.960	0.913	0.926
	TATA	0.585	0.678	0.550	0.620	0.285	0.650
	Group	0.917	0.917	0.910	0.920	0.835	0.951
iProm-Zea* ²⁰	TATA	0.965	0.926	0.972	0.916	0.872	0.980
	NON-TATA	–	–	–	–	–	–
	Group	–	–	–	–	–	–
DNABERT* ⁸	NON-TATA	0.970	0.975	0.974	0.980	0.891	0.935
	TATA	0.965	0.932	0.969	0.971	0.865	0.981
	Group	0.940	0.953	0.950	0.964	0.879	0.968
BERT-Inception*	NON-TATA	0.982	0.983	0.983	0.988	0.925	0.996
	TATA	0.976	0.942	0.978	0.982	0.882	0.990
	Group	0.957	0.964	0.967	0.982	0.901	0.993

Table 1. Performance comparison of DNABERT, BERT-Inception, DeePromoter, and iProm-Zea models across different datasets. All asterisks (*) indicate results run on the local computer.

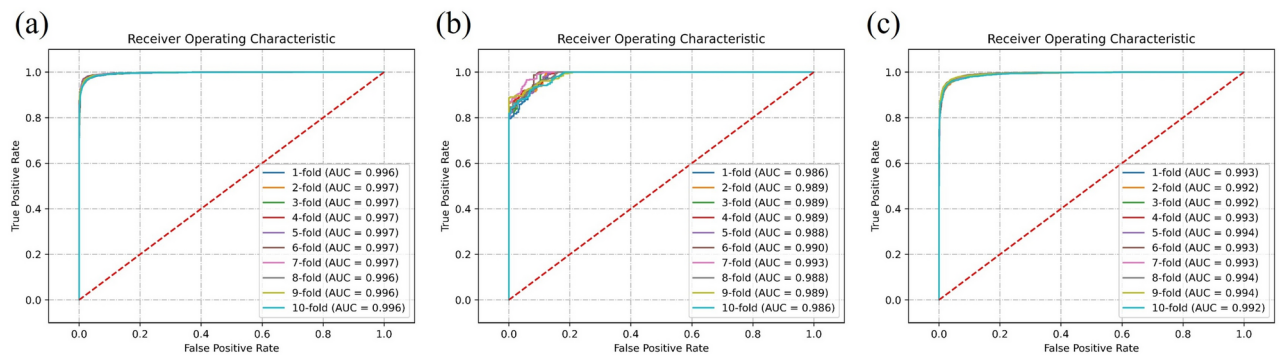


Figure 5. ROC curves for BERT-Inception. (a–c) ROC ((a) NON-TATA, (b) TATA, (c) Group).

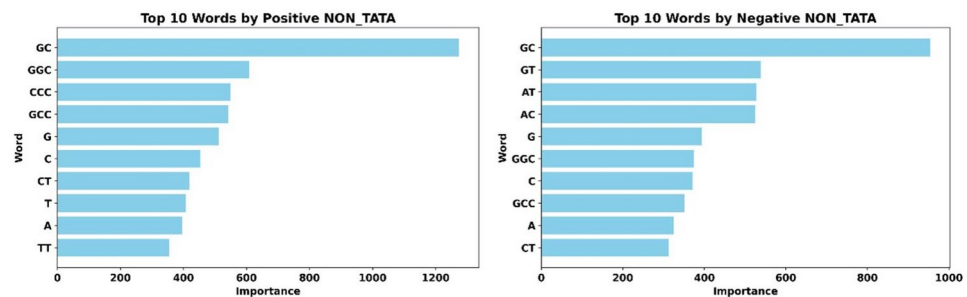


Figure 6. Top 10 words for non-TATA promoter importance.

which was trained using k-mers. This richer context helps BERT-Inception achieve better performance in distinguishing between promoters and non-promoters, especially for non-TATA promoters. The ROC curves for each model across the non-TATA, TATA, and Group datasets, shown in Figure 5, further demonstrate the superior performance of BERT-Inception.

DNA word importance ranking

This article aimed to evaluate the significance of each DNA fragment in the sequence by analyzing the overall attention values and frequencies within the model. Promoters were divided into two categories: TATA and non-TATA promoters, with separate in-depth interpretability analyses conducted for each type. Specifically, we first assessed and ranked the key terms in the TATA and non-TATA promoter datasets by calculating the average attention value for each term across the 12 heads in the final layer of the BERT model. This approach is justified by the fact that the high-level features in the last layer of BERT synthesize the information processed throughout the entire network, which can provide rich and precise information for downstream tasks. Immediately after that, we ranked the importance of all terms and selected the top 10 DNA terms based on their importance. On this basis, the Apriori algorithm was used to analyze the association rules on the positive sample data in the two types of promoters, and only the rules with a confidence level higher than 0.8 were retained, and an in-depth analysis was performed based on the dataset containing the top 10 DNA terms ranked in terms of importance in the visual display of the association rules.

As shown in Fig. 6, The left side displays the important vocabulary for the positive samples, while the right side shows it for the negative samples. Notably, in the non-TATA promoter dataset, the top 10 important words exhibit clear differences between the positive and negative samples in both word composition and importance distribution. In the positive sample, most of the top 10 words in terms of importance are two-letter combinations, such as "GC", "GGC", and "CCC", among which "GC" ranks first with its significant frequency of occurrence and attention score, showing its prominent importance in the positive sample. These GC-rich sequences, such as "GGC" and "CCC", are particularly relevant because they often serve as binding sites for transcription factors like Sp1 and p53, which are critical for the regulation of gene expression. For example, p53, a well-known tumor suppressor, is involved in the activation of genes related to cell cycle control and apoptosis, often binding to GC-rich sequences. Sp1, another transcription factor, also binds to GC-box motifs, which are frequently found in promoter regions. The attention weights captured by DNABERT⁸ in these sequences might reflect their role as key regulatory elements in non-TATA promoters. In contrast, the top ten terms in terms of importance in the negative sample are also dominated by two-letter combinations, but with slightly different terms, such as "GT" and "AT" showing high importance, while "GC GC" is still important, but with a change in importance. The importance of terms in the positive samples reflects the key role of these sequences in recognizing non-TATA promoters, especially the "GC"-rich regions, which may be associated with specific regulatory mechanisms of

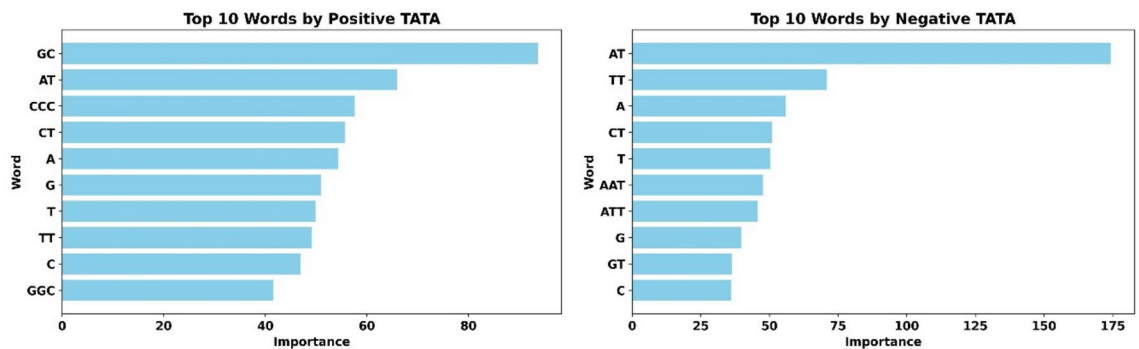


Figure 7. Top 10 words for TATA promoter importance.

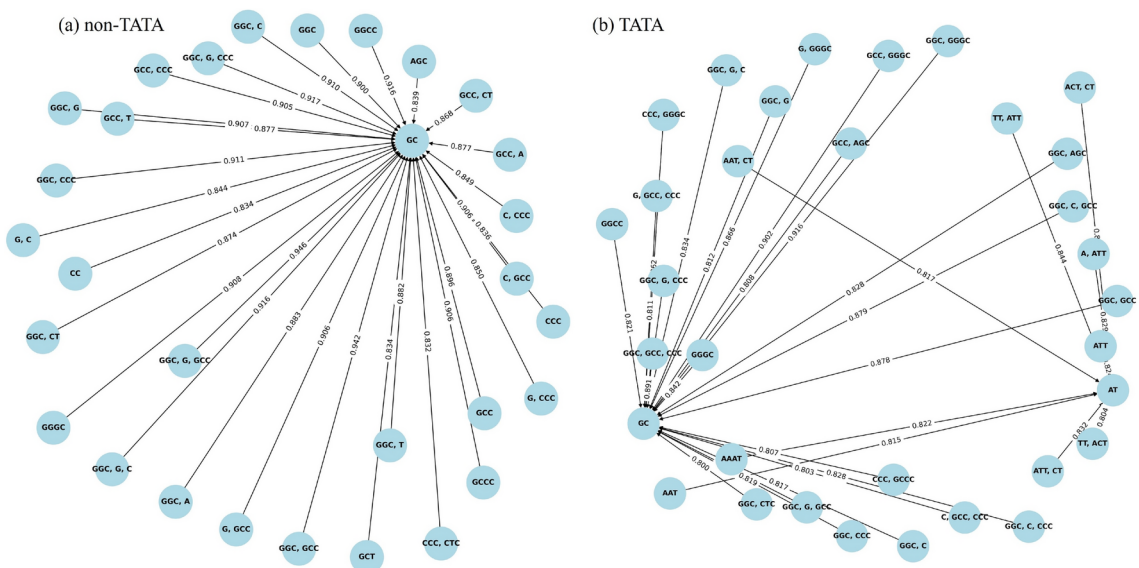


Figure 8. Visualization of lexical association rules for the importance of the top 10 promoters.

gene expression. The distribution of terms in the negative samples may reflect the presence of other sequence features outside the non-TATA promoter regions, suggesting different functions of these regions in the genome.

As shown in Fig. 7, the positive samples usually consisted of single and double words, which highlighted the prominence of sequences such as "GC" and "AT" in promoter recognition, while "CCC" ranked third in the proportion of sequences, the frequent occurrence of which may point to specific regulatory mechanisms related to CpG islands or histone modification sites. In contrast, the negative samples showed a greater diversity in word composition, especially a significant increase in the proportion of three-letter words such as "AAT" and "ATT", reflecting different patterns of sequence composition in the negative samples. This variability could indicate the presence of other regulatory elements or less conserved motifs that play different roles in gene regulation. This variability was evident in the distribution of significant words between the positive and negative samples, revealing a possible unique role of the TATA promoter in the gene regulatory network.

As shown in Fig. 8a, there are some obvious association rules that reveal the close association between different DNA terms in non-TATA promoter sequences. Most of them are highly correlated with "GC", e.g., the "GGC, G, CCC" → "GC" rule suggests that when "AGC" is present in the sequence, "GC" is also likely to be present, and the confidence level of "GC" is about 0.92. Similarly, the "C" → "GC", "CC" → "GC", "GCCC" → "GC" and "CCC" → "GC" rules show a high correlation between the different DNA terms in the non-TATA promoter sequences. These rules underscore the functional importance of GC-rich regions, which are associated with critical transcription factor binding sites like those for p53 and Sp1. These transcription factors are essential for transcriptional regulation, particularly in response to stress and during the regulation of cell growth and survival. "GC" and "CCC" → "GC" rules all indicate that single or consecutive "C" patterns are highly correlated with "GC" occurrences. In addition, more complex rules such as "GCC, CCC" → "GC" show that there is a high correlation between "GCC" and "CCC" when they occur together. These complex patterns further suggest that the presence of GC-rich motifs is not random but likely plays a coordinated role in the activation of transcription, especially in the context of non-TATA promoters. In contrast, "GCC" → "CCC" suggests that when

"GCC" and "CCC" occur together, "GC" is also likely to occur. These rules reveal that the occurrence of specific word combinations in non-TATA promoter sequences is not random, but has clear biological significance and functional relevance.

As shown in Fig. 8b, a series of high-confidence rules are displayed, revealing close connections between specific word combinations in the sequence. Similar to the non-TATA promoter, most of the DNA words are related to "GC", for example, there are several rules showing the close association between "GGC" and "GC", such as "GGC", "C" → "GC" and "GGC", "GCC" → "GC", and the confidence levels of these rules are 0.816 and 0.878, respectively. 0.816 and 0.878, suggesting that "GGC" is likely to occur when "GGC", as well as other specific sequence combinations, are present. The consistency of this pattern highlights the prevalence of 'GC-rich regions in TATA promoters and their critical role in regulating gene expression. The consistent occurrence of these GC-rich motifs, including "GGC" and "GCC", aligns with the findings of DNABERT⁸, where these sequences were shown to be essential for the binding of transcription factors such as Sp1 and p53, which are involved in key gene regulatory processes. The confidence level of the "GCC", "GGGC" → "GC" rule is 0.902, which further demonstrates that "GC" is most likely to appear when "GCC" and "GGGC" are present in the TATA promoter. The probability of "GC" appearing is very high when "GCC" and "GGGC" coexist. In addition, the association rule is not limited to a single sequence combination; more complex combinations such as "GGC", "G", "GCC", and "GC" also show a high confidence level with a value of 0.891, highlighting the relevance of complex sequence patterns in TATA promoter function. Unlike the non-TATA promoters, some rules showed a high correlation with "AT", e.g. "AAT", and "CT". "AAAT" → "AT", and the confidence level of these rules is also above 0.8, which suggests that when "AAT", "CT" or "AAAT" exist, "AT" is also likely to exist.

Conclusion and future works

This article introduces a novel method for segmenting DNA sequences into meaningful words and constructs a DNA-specific dictionary for pre-training the BERT model. This approach enhances the BERT-CNN model, forming a BERT-Inception structure capable of capturing information at various granularities. Experimental results demonstrate improvements in promoter prediction tasks, particularly in distinguishing TATA and non-TATA promoters. The BERT-Inception model exhibits high accuracy and reliability, outperforming existing models such as DNABERT across multiple metrics. The interpretability analysis provides new insights into gene expression regulatory mechanisms by identifying key DNA fragments and their associations.

Future work will focus on refining the DNA dictionary construction by incorporating more sophisticated algorithms and considering additional biological factors. Extending the BERT-Inception model to other genomic tasks, such as enhancer prediction and splice site identification, will help evaluate its generalizability and robustness.

In conclusion, the proposed method utilizes a deep learning approach to provide a new perspective on DNA sequence analysis.

Data availability

The data used to support the findings of this study are available from the corresponding author upon request.

Received: 7 September 2024; Accepted: 19 December 2024

Published online: 02 January 2025

References

1. Le, N. Q. K., Yapp, E. K. Y., Nagasundaram, N. & Yeh, H.-Y. Classifying promoters by interpreting the hidden information of DNA sequences via deep learning and combination of continuous fasttext n-grams. *Front. Bioeng. Biotechnol.* **7**, 305 (2019).
2. Oubounyt, M., Louadi, Z., Tayara, H. & Chong, K. T. Deepromoter: Robust promoter predictor using deep learning. *Front. Genet.* **10** (2013).
3. Eisenhaber, F. A decade after the first full human genome sequencing: When will we understand our own genome?. *J. Bioinform. Comput. Biol.* **10**, 1271001 (2012).
4. Zhu, Y. et al. Computational identification of eukaryotic promoters based on cascaded deep capsule neural networks. *Brief. Bioinform.* **22**, bbaa299 (2021).
5. Oubounyt, M., Louadi, Z., Tayara, H. & Chong, K. T. Deepromoter: Robust promoter predictor using deep learning. *Front. Genet.* **10**, 286 (2019).
6. Chen, X., Cong, P. & Lv, S. A long-text classification method of Chinese news based on Bert and CNN. *IEEE Access* **10**, 34046–34057 (2022).
7. Abas, A. R., Elhenawy, I., Zidan, M. & Othman, M. Bert-cnn: A deep learning model for detecting emotions from text. *Comput. Mater. Contin.* **71** (2022).
8. Ji, Y., Zhou, Z., Liu, H. & Davuluri, R. V. Dnabert: Pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. *Bioinformatics* **37**, 2112–2120 (2021).
9. Routhier, E. & Mozziconacci, J. Genomics enters the deep learning era. *PeerJ* **10**, e13613 (2022).
10. Guo, Y., Wu, J., Ma, H., Wang, S. & Huang, J. Deep ensemble learning with atrous spatial pyramid networks for protein secondary structure prediction. *Biomolecules* **12**, 774 (2022).
11. Le, N. Q. K., Ho, Q.-T., Yapp, E. K. Y., Ou, Y.-Y. & Yeh, H.-Y. Deepet: A deep convolutional neural network architecture for investigating and classifying electron transport chain's complexes. *Neurocomputing* **375**, 71–79 (2020).
12. Li, J., Pu, Y., Tang, J., Zou, Q. & Guo, F. Deepavp: A dual-channel deep neural network for identifying variable-length antiviral peptides. *IEEE J. Biomed. Health Inform.* **24**, 3012–3019 (2020).
13. Asgari, E. & Mofrad, M. R. Continuous distributed representation of biological sequences for deep proteomics and genomics. *PLoS one* **10**, e0141287 (2015).
14. Le, N. Q. K. et al. Enhancer-5step: Identifying enhancers using hidden information of DNA sequences via Chou's 5-step rule and word embedding. *Anal. Biochem.* **571**, 53–61 (2019).
15. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint [arXiv:1810.04805](https://arxiv.org/abs/1810.04805) (2018).

16. Le, N. Q. K., Ho, Q.-T., Nguyen, V.-N. & Chang, J.-S. Bert-promoter: An improved sequence-based predictor of DNA promoter using Bert pre-trained model and Shap feature selection. *Comput. Biol. Chem.* **99**, 107732 (2022).
17. Le, N. Q. K. & Ho, Q.-T. Deep transformers and convolutional neural network in identifying DNA N6-methyladenine sites in cross-species genomes. *Methods* **204**, 199–206 (2022).
18. Le, N. Q. K., Ho, Q.-T., Nguyen, T.-T.-D. & Ou, Y.-Y. A transformer architecture based on Bert and 2D convolutional neural network to identify DNA enhancers from sequence information. *Brief. Bioinform.* **22**, bbab005 (2021).
19. Khan, S., Khan, M., Iqbal, N., Khan, S. A. & Chou, K.-C. Prediction of pirnas and their function based on discriminative intelligent model using hybrid features into Chou's psekcnc. *Chemom. Intell. Lab. Syst.* **203**, 104056 (2020).
20. Kim, J., Shujaat, M. & Tayara, H. iprom-zea: A two-layer model to identify plant promoters and their types using convolutional neural network. *Genomics* **114**, 110384 (2022).
21. Liu, J., Zhou, D. & Jin, W. Prediction of nucleosome dynamic interval based on long-short-term memory network (lstm). *J. Bioinform. Comput. Biol.* **20**, 2250009 (2022).
22. Zeng, H., Edwards, M. D., Liu, G. & Gifford, D. K. Convolutional neural network architectures for predicting DNA-protein binding. *Bioinformatics* **32**, i121–i127 (2016).
23. Zhai, F., Guan, Y., Li, Y., Chen, S. & He, R. Predicting the fluorescence properties of hairpin-DNA-templated silver nanoclusters via deep learning. *ACS Appl. Nano Mater.* **5**, 9615–9624 (2022).
24. De Waele, G., Clauwaert, J., Menschaert, G. & Waegeman, W. CPG transformer for imputation of single-cell methylomes. *Bioinformatics* **38**, 597–603 (2022).
25. Yang, F. et al. scbert as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-SEQ data. *Nat. Mach. Intell.* **4**, 852–866 (2022).
26. Novakovsky, G., Fornes, O., Saraswat, M., Mostafavi, S. & Wasserman, W. W. Explainn: Interpretable and transparent neural networks for genomics. *Genome Biol.* **24**, 154 (2023).
27. van Hilten, A. et al. Gennet framework: Interpretable deep learning for predicting phenotypes from genetic data. *Commun. Biol.* **4**, 1094 (2021).
28. Fortelny, N. & Bock, C. Knowledge-primed neural networks enable biologically interpretable deep learning on single-cell sequencing data. *Genome Biol.* **21**, 1–36 (2020).
29. Stiehler, F. et al. Helixer: Cross-species gene annotation of large eukaryotic genomes using deep learning. *Bioinformatics* **36**, 5291–5298 (2020).
30. Zhang, S. et al. Applications of transformer-based language models in bioinformatics: A survey. *Bioinform. Adv.* **3**, vbab001 (2023).
31. Clauwaert, J., Menschaert, G. & Waegeman, W. Explainability in transformer models for functional genomics. *Brief. Bioinform.* **22**, bbab060 (2021).
32. Jiang, Y. et al. Biology-guided deep learning predicts prognosis and cancer immunotherapy response. *Nat. Commun.* **14**, 5135 (2023).
33. Dreos, R., Ambrosini, G., Cavin Périer, R. & Bucher, P. Epd and epdnew, high-quality promoter resources in the next-generation sequencing era. *Nucleic Acids Res.* **41**, D157–D164 (2013).
34. Szegedy, C. et al. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1–9 (2015).

Acknowledgements

This work was supported by the National Key Technology R&D Program of China (Grant No.2019YFC1606401); Project of Beijing Municipal University Teacher Team Construction Support Plan (Grant No.BPHR20220104), and Project of Beijing Scholars Program (Grant No.099).

Author contributions

Conceptualization, R.Z. and Q.Z.; methodology, R.Z. and Z.L.; software, J.L. and Q.Z.; validation, Z.L. and Q.Z.; All authors have read and agreed to the published version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Z.L. or Q.Z.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2024