Network Modeling Analysis in Health Informatics and Bioinformatics (2024) 13:25
https://doi.org/10.1007/s13721-024-00463-4

ORIGINAL ARTICLE

XDeMo: a novel deep learning framework for DNA motif mining using transformer models

Rajashree Chaurasia[1,2] · Udayan Ghose[2]

Received: 13 October 2023 / Revised: 26 April 2024 / Accepted: 26 April 2024

Abstract

Motivation: Recognizing and studying DNA patterns is crucial for improving knowledge of illnesses, cell function, and gene control. Motifs determine which transcription factor a protein may bind to, leading to a better unraveling of gene expression. Advancements in the fields of deep learning and high-throughput sequencing have made possible the exploration of motif discovery anew, with greater accuracy and performance. Methodology: In this paper, a novel deep learning framework (XDeMo – Transformer-based Deep Motifs) for DNA motif mining using Transformer models is proposed. Furthermore, a hybrid encoding scheme is also introduced, called 'blended' encoding specifically designed for use with deep learning transformer models that are trained using DNA sequences. Results: Our proposed transformer-based framework for DNA motif discovery augmented by blended encoding outperforms many state-of-the-art deep learning models on many baseline performance metrics when trained on the standard datasets. Our models demonstrated robust performance in predicting motifs with high discriminative power, precision, recall, and F1 score. Conclusion: The model's ability to capture intricate sequence patterns and long-range dependencies led to the discovery of biologically meaningful motifs that were verified from known transcription factor binding motif databases. This shows that our novel framework can be effectively used to find DNA motifs and therefore, aid in further downstream analyses for biomedical and biotechnological applications.

Significance

XDeMo's practical implications span the realms of gene regulation research, genomics tool development, molecular biology, and diagnostic applications. It offers a robust foundation for further advancements in genomic analysis, with the potential to accelerate discoveries in gene regulation and the development of novel therapeutic strategies.

Keywords Deep learning · DNA motif · Transcriptional factor · Transformer model · Blended encoding · ChIP-seq

1 Introduction

One of the biggest and most challenging tasks in the disciplines of bioinformatics and data science is finding patterns or motifs in the genomic DNA of organisms. The identification of DNA motifs is a critical first step in a wide range of biological applications (He et al. 2012). To better understand gene regulation, cell function, and diseases, it is essential to recognize and research DNA motifs, which are the brief, repeated sequence patterns of DNA nucleotides linked to a protein (Suter 2020). The existence of these motifs determines the specific Transcription Factor (TF) that each protein uses to bind to a corresponding region in the genome (Chaurasia and Ghose 2023). Both strands of DNA can include motifs. Additionally, TFs directly attach to the double-stranded DNA (Wang et al. 2014). To find such DNA motifs, one must look for them in sequences that have these Transcription Factor Binding Sites (TFBSs) (Lin et al. 2019). It is well known that TFs regulate gene activity in response to numerous environmental stimuli, which

✉ Rajashree Chaurasia
rajashree.14416490019@ipu.ac.in; rs.chaurasia87@gov.in; rajashree.chaurasia@gmail.com

[1] Department of Computer Engineering, Directorate of Training and Technical Education (Govt. of NCT of Delhi), Guru Nanak Dev DSEU Rohini Campus, Delhi, India

[2] University School of Information, Communication & Technology, Guru Gobind Singh Indraprastha University, Delhi, India

Published online: 13 May 2024

has a substantial influence on the development of disease (Zhang et al. 2021). By attaching to certain DNA or RNA sequences, TFs can modify the expression of genes (Pardiñas et al. 2018).

The quick discovery of multiple candidate motif positions throughout the genome has been made possible by the combination of Chromatin Immuno-precipitation sequencing (ChIP-seq), other advanced DNA sequencing technologies, and statistical analyses (Zambelli et al. 2012; Madrid et al. 2019). Using these high-throughput sequencing technologies along with computational approaches, the right TFBS based on motif sequence specificity may be identified. Studies by (Nutiu et al. 2011) and (Siggers and Gordân 2013) show that such strategies are dependable and reproducible. Deep Learning (DL), the most successful Machine Learning (ML) technique in bioinformatics, has been used in a variety of fields, including the categorization of DNA sequences (Xu et al. 2021). However, it has been discovered that there is not enough data to train the DL model for estimating motif length (Jin et al. 2020). An abundance of freely accessible DNA pattern datasets with particular motifs is made available by databases like ENCODE (Encyclopedia of DNA Elements), JASPAR (Just Another Scaffolds/Position-Weight Matrix (PWM) Database), TRANSFAC (TRANScription FACtor database), etc. to solve this problem (Alipanahi et al. 2015; Poliakov et al. 2014; Yang et al. 2019). These datasets may be used to provide enough training instances for effectively predicting DNA motifs, as well as verification of motifs discovered to evaluate a particular model.

1.1 Research gaps

In recent years, several DL methods have emerged as powerful tools for pattern recognition tasks in genomics. A comprehensive review of such DL models that are employed in the motif mining task is given in (Chaurasia and Ghose 2023; Trabelsi et al. 2019). The most popularly used DL strategies for motif mining have been Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and their hybrids. While CNNs can learn short-term relationships, RNNs help to learn the distant interdependencies among the motif attributes. Hybrid models attempt to realize both kinds of relationships among the motif features. Learning both contexts in the sequences is important to fully capture the complex patterns and relationships in the DNA sequence. Motifs that are critical for the regulation of genes could be missed by a model that can only capture short-range relationships. However, certain DNA motifs may consist of recurring patterns that happen at regular intervals, making it necessary for the model to acquire long-term relationships to recognize the motif. Nevertheless, recent advancements in DL technologies have given rise to several newer models like ResNet (Residual Neural Network) (He et al. 2016), U-Net (Falk et al. 2018), BERT (Bidirectional Encoder Representations from Transformers) (Kamath et al. 2022), GPT-3 (Generative Pre-trained Transformer 3) (Floridi and Chiriatti 2020), GPT-4 (Generative Pre-trained Transformer 4) (OpenAI 2023), etc. Specifically, the Transformer model, a type of neural network architecture originally proposed for natural language processing tasks, has shown promise for sequence analysis tasks in genomics.

1.2 Transformer models

Transformer models use self-attention procedures to process the complete input sequence at once, in contrast to conventional RNNs that process input sequences one by one or word by word. As a result, transformers are better able to do tasks that call for a grasp of context and semantic links with greater training speeds and parallelizability and also are able to record long-range dependencies between various portions of the sequence (Vaswani et al. 2017). Further, this increased parallelizability also enables these models to be trained on large datasets faster than any other DL models. Self-attention allows the model to concentrate on distinct elements of an input sequence by assigning weights to each element to determine its significance for producing a prediction (Otten 2023). DNABERT (Deoxyribonucleic Acid BERT) (Ji et al. 2021) and Enformer (Avsec et al. 2021) are based on Transformer models that can be used for the prediction of genomic elements from the DNA sequence. DNABERT is a pre-trained fine-tuned BERT model that uses k-mer tokenization to garner a comprehensive and transmissible interpretation of upstream and downstream nucleotide contexts in genomic DNA sequences. DNABERT has been used to predict TFBSs, promoter sequences, and splice sites and also has been compared to several state-of-the-art models like DeepBind (Alipanahi et al. 2015), DESSO (DEep Sequence and Shape mOtif) (Yang et al. 2019), DanQ (Quang and Xie 2016), etc. Enformer is another Transformer-based model that is used to predict the interactions between enhancer and promoter sequences. However, it has not directly been used to predict TFBSs or motifs in DNA sequences.

1.3 The proposed framework

In this paper, we propose a novel deep learning framework (XDeMo – Transformer-based Deep Motifs) for DNA motif mining using Transformer models. The proposed framework leverages the multi-head self-attention mechanism of the Transformer model to learn long-range dependencies and relationships of context between nucleotides in a DNA sequence. Further, we introduce a novel encoding

mechanism for DNA sequences especially tailored to capture both local and global sequence information that can be easily used with deep learning models. Our 'blended' encoding approach generates a feature representation that preserves information about local patterns (short-range dependencies) via one-hot encoding while also retaining information about k-mer frequencies over the whole sequence via raw embeddings (long-range dependencies). Machine learning models can employ this hybrid form to account for both short- and long-term dependencies when making predictions or classifications.

We establish the efficacy of our approach by evaluating it against current state-of-the-art DL methods on benchmark datasets through baseline performance metrics. The standard procedure for motif elicitation using computational analyses as well as the generalized DL framework for motif discovery are outlined in (Chaurasia and Ghose 2023), the latter of which we have applied in this paper.

The main research question addressed in this paper is whether the proposed deep learning framework can improve the accuracy and speed of DNA motif mining compared to existing methods. To answer this question, we conduct experiments to assess the performance of our framework and offer insights into the underlying mechanisms of the model. Our findings demonstrate that our framework performs at the cutting edge on several benchmark datasets. Overall, our proposed framework has the potential to advance the fields of DNA motif mining as well as machine learning, and enable the discovery of new regulatory motifs that are critical for understanding gene regulation and disease mechanisms. However, the model is bound by the quality and quantity of the labeled peak-called dataset used for training. The presence of motifs that align with known TF binding motifs is indicative of the model's potential but requires further investigation to establish their functional relevance.

The development of a novel framework like XDeMo is imperative due to the escalating complexity of genomics and computational biology. With the surge in data volume and diversity, there is a critical need for advanced tools capable of effectively handling intricate genomic datasets. XDeMo addresses this challenge by offering precise predictive capabilities, interpretability, and versatility across diverse research applications. It caters to the expanding knowledge of gene regulation, genetic variation, and disease mechanisms, ensuring researchers can navigate the evolving genomics landscape. Additionally, XDeMo integrates cutting-edge techniques and supports the quest for targeted therapies, positioning it as an indispensable asset for genomics and computational biology investigations in the modern era.

This paper is organized as follows: Sect. 2 elaborates upon the dataset employed, methodology, and experimental setup; Sect. 3 discusses the results obtained from training the model, and Sect. 4 concludes the paper with major findings, limitations, significance, and future scope of the work.

2 Materials and methods

An overview of our XDeMo framework is given in Fig. 1. We have utilized the ENCODE ChIP-seq TFBSs datasets which is a large collection of high-quality, curated, and normalized peak-called transcription factor ChIP-seq data, which has been used by various DL models to predict regulatory regions in DNA sequences, like DeepBind, DeepSEA (Zhou and Troyanskaya 2015), DESSO, DNABERT, etc., which have shown promising results in predicting TFBSs and other genomic features. In our previous work (Chaurasia and Ghose 2023), we presented the generic framework for the motif discovery process that uses DL technologies. In the following subsections, we detail the methodology and experimental setup corresponding to this framework.

2.1 The ENCODE ChIP-seq dataset

The ENCODE ChIP-seq TFBS dataset (ENCODE Project Consortium 2012; Luo et al. 2019) is recognized as a benchmark dataset for motif discovery because of its significant

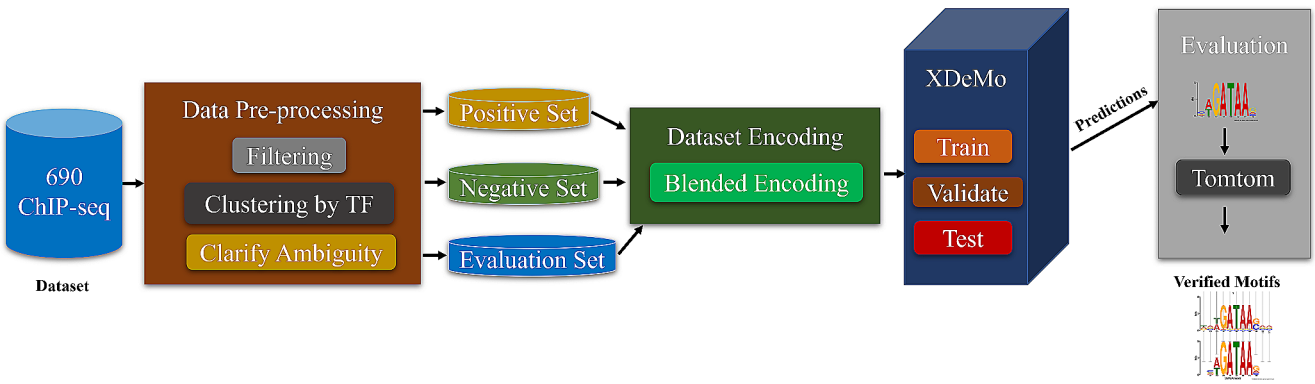

Fig. 1 XDeMo framework overview

contribution to expanding our comprehension of TFBSs and the associated regulatory processes in the human genome. It provides an expansive and representative sampling of TFBSs in distinct biological situations, with over 690 ChIP-seq datasets reflecting a wide range of regulatory factors, spanning numerous human cell types and embracing varying treatment conditions.

Furthermore, quality control procedures in the dataset, such as biological replicates and the use of IDR (Irreproducible Discovery Rate) analysis, guarantee a high level of data dependability. This thorough curation procedure yields a dataset enriched with easily replicated TFBSs, with the presence of false positives minimized. Additionally, because of its broad acceptance and widespread application in several motif-finding algorithms, such as DeepBind and DeepSEA, it has become a frequent indicator for measuring and evaluating the performance of these methods. Finally, because the dataset was generated by several ENCODE TFBS ChIP-seq production groups, it demonstrates a wide range of experimental methodologies and approaches.

As a result, the extensive scope of the ENCODE ChIP-seq TFBS dataset, restricted quality control, widespread acceptance, and different origins all establish it as a baseline dataset for motif discovery.

2.2 Data preprocessing

Preprocessing any dataset for quality, reduction of noise, and bias, is the first step in any machine-learning task. We downloaded the standard baseline 690 uniform TFBSs ChIP-seq datasets from the ENCODE website (available at https://hgdownload.soe.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeAwgTfbsUniform/) (ENCODE Project Consortium 2012; Luo et al. 2019). These datasets contain peak-called TF binding profile ChIP-seq experiments (Hitz et al. 2023) of numerous TFs for various human cell lines for the standard hg19 reference genome (i.e., GRCh37) (available at https://hgdownload.soe.ucsc.edu/goldenPath/hg19/bigZips/). We have filtered out any trials that have received any extra processing or whose quality is not deemed to be good, as was done in the DeepBind model. Subsequently, to examine and compare the model's performance with that of other models for each TF, we clustered the cumulative experiments by TF. We further filtered out any experiments that had ambiguous nucleotide bases (i.e., other than A, G, C, and T) and selected the top-ranked 14,300 (set A) and 7,300 (set B) ChIP-seq peaks for each TF. Sets A and B contain datasets for a total of 49 and 63 unique TFs, respectively. Next, a positively labeled set of DNA sequences was generated from sets A and B for each TF, centered at the peak and spanning a 201 bp (base pair) region of each ChIP-seq peak, while the negative set was constructed from shuffled positive sequences with matching dinucleotides composition (Alipanahi et al. 2015; Zeng et al. 2016). The negative set is generated so that it has a similar nucleotide composition, GC content, and an overall distribution to the positive set. From these sets, 300 lowest-ranked samples (non-overlapping and held out) were used for motif verification and evaluation (called the evaluation set).

2.3 Dataset encoding

DNA sequences must be translated into a numerical format that DL models can use as input to generate predictions. The two most popular encoding schemes used in DL models for DNA sequence analyses are one-hot encoding (Choong and Lee 2017) and k-mer encoding (Gunasekaran et al. 2021). One-hot scheme encodes each base of a sequence as a binary quadruple (one for each base – A, G, C, and T) vector where each base is represented by a '1' for its specific position in the encoding and the rest of the positions in the vector are '0'. For example, 'A' may be encoded as (1, 0, 0, 0), 'G' as (0, 1, 0, 0), 'C' as (0, 0, 1, 0), and 'D' as (0, 0, 0, 1). This is a relatively straightforward method that effectively captures low-level interactions but yields a very sparse representation that ignores the high-order characteristics of the underlying sequence. Despite the limitations, it is still a popular technique used in many DL models when the dataset is small. On the other hand, k-mer embedding captures higher-order sequence information by creating subsequences of length 'k' and representing them by a numerical low-dimensional vector based on aspects like the rate of occurrence. As a result, the model can discover intricate connections between the nucleotides in the sequences, potentially improving its performance in specific contexts. However, k-mer encoding also suffers from sparsity when datasets are large.

To overcome these limitations, we have introduced a novel encoding scheme (see Fig. 2). In this 'blended' encoding scheme, we first generate k-mer embedding for all k-mers of length 6 (decided upon by trial and error) for each sequence in a set to generate a fixed-length numerical vector. Each k-mer is represented as a base-4 number with k digits that denote the position of the k-mer in the embedding. The value of each digit is determined by the corresponding nucleotide's code in the k-mer (A = 0, C = 1, G = 2, T = 3). This creates a mapping of each k-mer to a unique position in the k-mer embedding in the range [0, 4^k-1]. This allows us to represent the sequences as fixed-length numerical vectors where each element corresponds to a specific k-mer and its value represents the frequency of occurrence. Next, we introduce a binary representation of the resulting embeddings where each unique k-mer is denoted by a vector of all zeros, except for a '1' at the position corresponding to that k-mer and the resulting one-hot

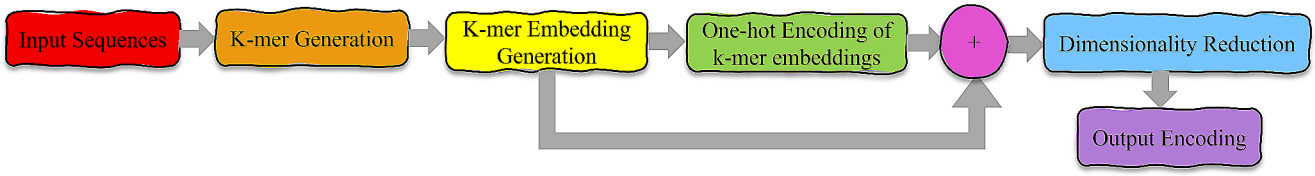

Fig. 2 Blended encoding scheme block diagram. Here, '+' denoted concatenation

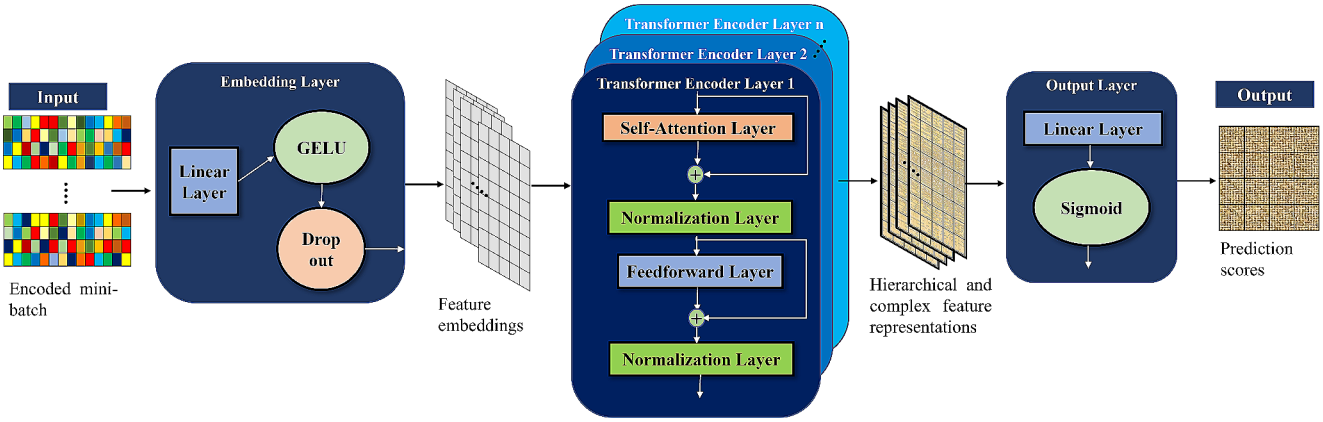

Fig. 3 XDeMo model architecture. There are 'n' transformer layers in which each layer has a multi-head self-attention mechanism followed by normalization and feedforward layers. These multi-head attention networks focus on different parts of the input feature embeddings coming from the embedding layer to generate complex representations

Table 1 Experimental setup

Set	Number of evaluation samples for each TF	Dataset size for each TF (number of samples)	Train, Test, Validation dataset split for each TF		
			Train (75%)	Validation (12.5%)	Test (12.5%)
A1	300	14,300	10,500	1,750	1,750
A2	300	14,300	21,000	3,500	3,500
B	300	7,300	10,500	1,750	1,750

encoded vector has a length equal to the number of possible k-mers i.e., 4^k. Now we concatenate the k-mer embedding vector with the binary (one-hot) vector to obtain a complete representation of the sequence information, both low-level and high-level features can be captured in this representation. However, since the dataset is large in our case, sparsity is still an issue that is dealt with by dimensionality reduction involving principal component analysis (PCA) to reduce the feature space to 201 dimensions. This is necessary to prevent overfitting, reduce any outliers or noisy data, and increase computational performance. Each of the positive, negative, and evaluation datasets for each TF is encoded using this blended encoding approach. The amount of variance retained is measured for each of the sets for each TF, separately (see Supplementary Table 1). It is worth noting that the overall average amount of variance retained after PCA is approximately 0.92 (i.e., 92%) which is more than satisfactory.

2.4 Model architecture and training

The transformer, a prominent design for sequence-to-sequence tasks, serves as the foundation for our model architecture. In the context of DNA motif identification, however, we simplify the transformer by excluding the decoder layers that are generally employed to generate output sequences in cases such as language translation tasks.

Since the emphasis of DNA motif discovery is on learning patterns and representations in input sequences rather than generating output sequences, disregarding the decoder layer is reasonable in this case.

Our transformer model (see Fig. 3) consists of an embedding layer, several transformer layers, and an output layer. The embedding layer takes the input data and maps it to a higher dimensional space using a linear transformation followed by a GELU (Gaussian Error Linear Unit) activation function (Hendrycks and Gimpel 2016) and dropout for regularization to prevent overfitting. The choice of GELU over the widely used ReLU (Rectified Linear Unit) activation (Fukushima 1975) is that GELU retains negative values instead of setting them to zero as ReLU does. The smoothness and continuous nature of GELU further contribute to more stable and predictable updates during training, improving the performance of the optimizer. The transformer layers are made up of several transformer encoders stacked on top of each other that form the core of our model. Each of these layers consists of a multi-head attention mechanism

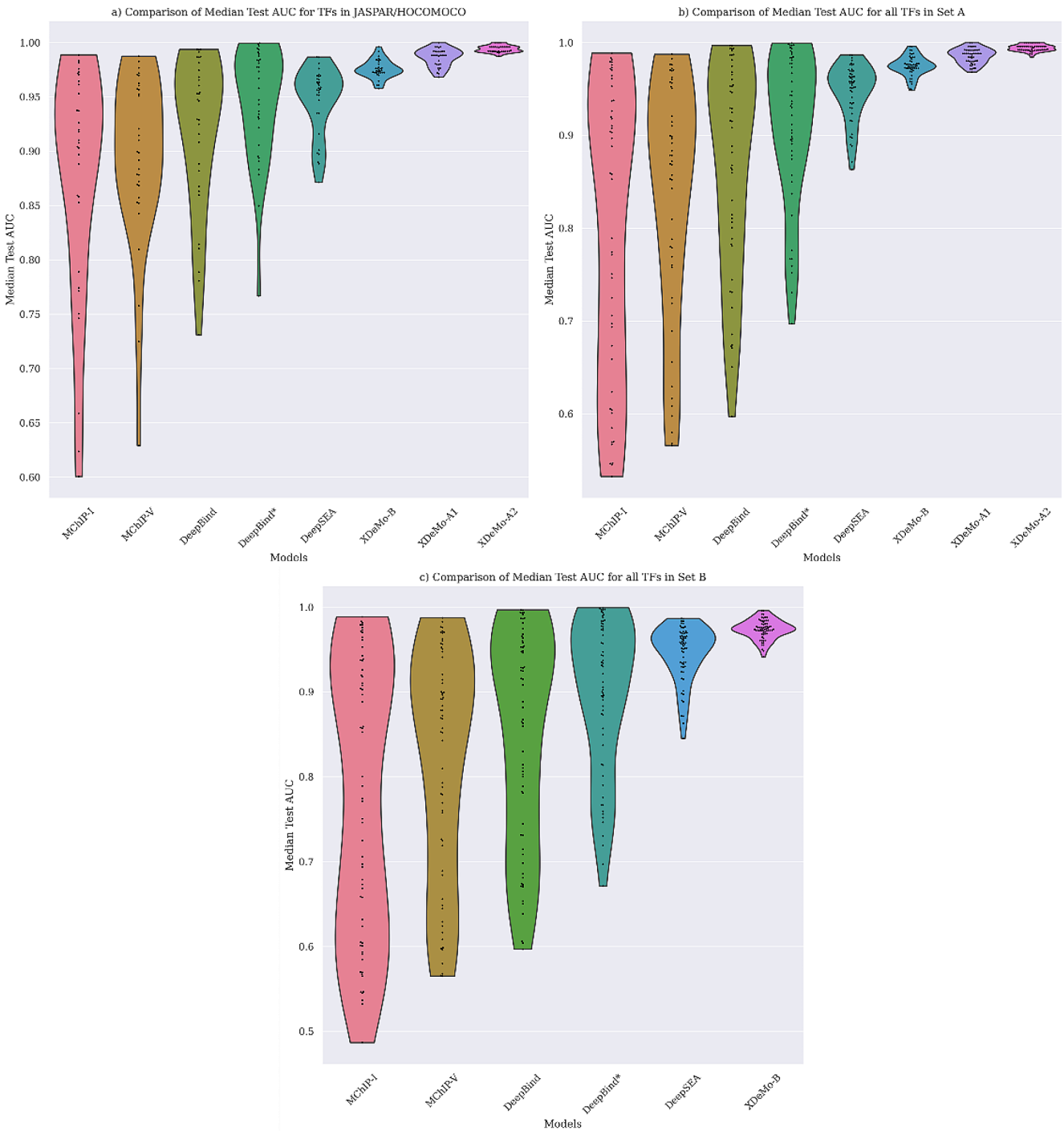

Fig. 4 Comparison of median test AUCs among the various DL models (a part of the data is based on (Alipanahi et al. 2015) and (Zhou and Troyanskaya 2015). All violin plots contain swarm plots within them to show the frequency of occurrence of the values. The individual plots have been scaled to this frequency. Part **(a)** shows the median test AUCs for only those TFs from set A that are in the standard JASPAR and HOCOMOCO databases. Parts **(b)** and **(c)** show the median test AUCs for all the TFs considered in their respective sets. XDeMo-A1, XDeMo-A2, and XDeMo-B refer to the XDeMo models for sets A1, A2, and B, respectively

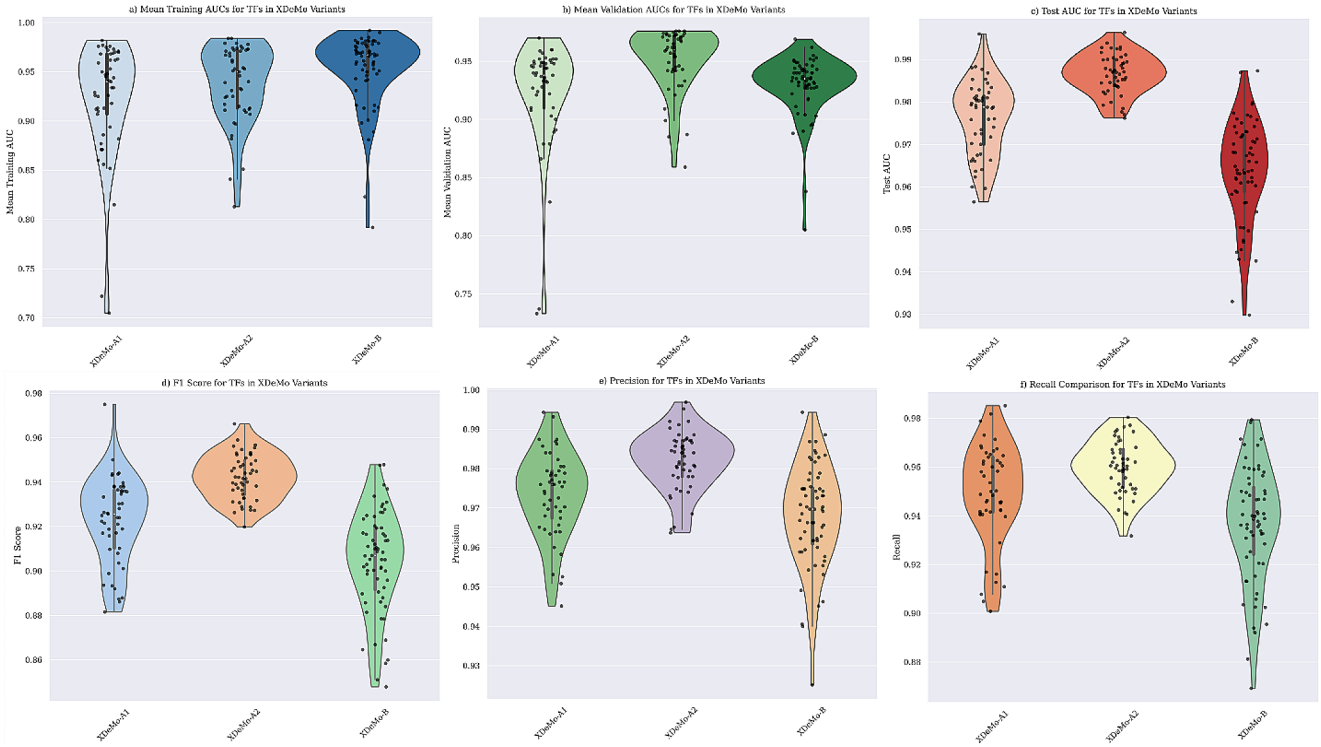

Fig. 5 Comparison of baseline performance metrics among the XDeMo model variants. Each subplot contains a swarm plot as well as a boxplot within it

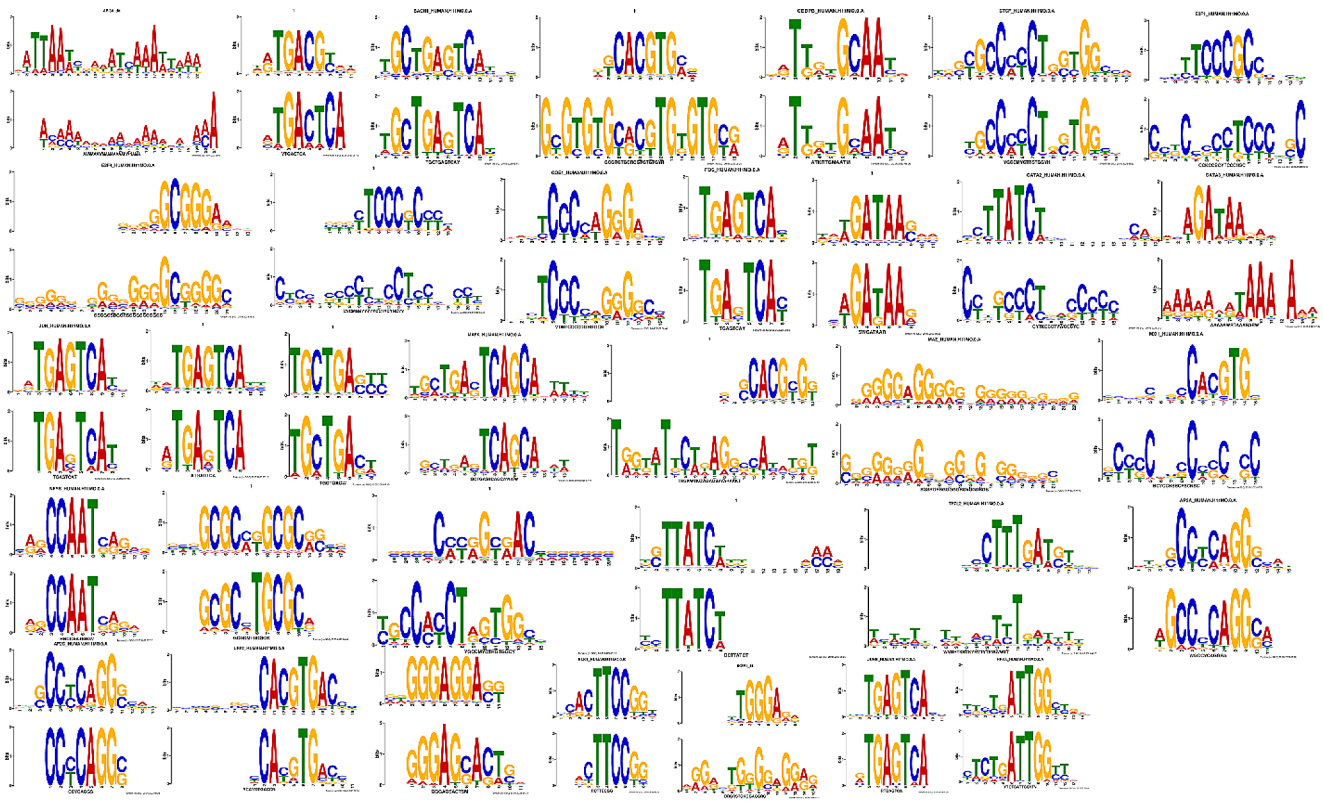

Fig. 6 Motifs discovered in XDeMo versus the motifs from known databases (HOCOMOCO). The upper part of each subplot displays the known annotated motif from HOCOMOCO and the lower part displays the corresponding motif as discovered by our model

followed by a feedforward neural network. The self-attention mechanism allows the model to record relationships between distinct points in the sequence, while the feed-forward networks furnish non-linear transformations.

Finally, the output layer takes the output from the transformer layers and applies a linear transformation followed by a sigmoid activation function to produce the final predictions. The model is trained using the binary cross-entropy loss function (Mannor et al. 2005), and optimized using a gradient-based method, Adam (Adaptive Moment Estimation) (Kingma and Ba 2014). The model's hyperparameters are tuned automatically using the grid search method (see Supplementary Table 2 for the choice of hyperparameters for each experimental setup). However, in some cases some of the hyperparameters are fine-tuned manually and early stopping is implemented in addition to dropout regularization to further manage overfitting to achieve a better performance.

Since Transformer models offer the advantage of shorter training times due to their high parallelizability, we can effectively train our models with large datasets. In a meta-analysis paper of various DL models by Trabelsi et al. (Trabelsi et al. 2019), some general guidelines are discussed for model selection and evaluation, one of them being that the training sample set should have at least 10,000 data points. We used set A to create two experimental setups viz. set A1 and set A2 with 10,500 and 21,000 training samples, respectively. In set A1, we kept only the even-numbered samples from the positive set and the odd-numbered samples from the negative set. For sets A2 and B, we considered the full dataset from both positive and negative sets. Table 1 shows the experimental setup of these sets along with the non-overlapping train, test, and validation splits.

2.5 Motif evaluation

We trained a separate transformer model for each TF in each experimental setup and obtained the prediction scores over the unlabeled evaluation dataset. From these prediction scores, we extracted 30 bp sub-sequences from the evaluation dataset from the center of the predicted position in the reference sequence (hg19) and used the MEME (Multiple Expectation-maximization for Motif Elicitation) programs (https://meme-suite.org/meme/tools/meme) to visualize the top three motif sequences ranging from 6 bp to 30 bp. The verification of the discovered motifs was accomplished by aligning them with recognized motifs in the HOCOMOCO (HOmo sapiens COmprehensive MOdel COllection available at https://hocomoco11.autosome.org/) database using Tomtom (available at https://meme-suite.org/meme/tools/tomtom) for the TFs present in the JASPAR and HOCOMOCO databases.

2.6 Performance evaluation metrics

Performance was compared with state-of-the-art DL models like DeepBind (Alipanahi et al. 2015), DeepBind* (DeepBind with the rest of the dataset that was not used for DeepBind) (Alipanahi et al. 2015), DeepSEA (Zhou and Troyanskaya 2015), DNABERT (Ji et al. 2021), Basset (Kelley et al. 2016), DanQ (Quang and Xie 2016), DeepSite (Zhang et al. 2019), and DESSO (Yang et al. 2019) using baseline performance metrics like AUC (area under the receiver operating characteristic curve), F1 score, precision, and recall. Each of these models plays a pivotal role in elucidating the intricacies of DNA motif identification, and their collective examination unveils the comprehensive landscape of deep learning methodologies in this domain. Further, since these models have been trained for the same standard datasets of ChIP-seq peak experiments, they constitute the ensemble of models under consideration. For a passive exploration of these models, their unique attributes and contributions, and their performance comparisons, see (Chaurasia and Ghose 2023).

All performance metric computations and analyses were performed on a Tesla T4 GPU machine (with an Intel Xeon CPU running at 2.20 GHz maximum speed).

3 Results and discussion

We calculated and compared the median test AUC for each TF in each experimental setup of our XDeMo model with other DL models as well as for two traditional computational models as given in DeepBind (Alipanahi et al. 2015). The non-DL models use the MEME-ChIP algorithm (https://meme-suite.org/meme/tools/meme-chip) according to the method described in the DeepBind paper. MChIP-I corresponds to the top motif discovered through MEME-ChIP and MChIP-V corresponds to the sum of scores of the top five motifs discovered. We also calculated the F1 score, precision, and recall for the held-out test dataset as well as for each epoch while training the models for each TF in each experimental set. Further, we computed the mean training AUC and mean validation AUC over each epoch, as well as their overall averages for individual TFs in sets A1, A2, and B. To reduce the impact of outliers and provide a more accurate approximation of the model's actual performance, we find that median AUCs are a better metric in comparison to mean AUCs. However, we report both metrics in our manuscript for completeness (see Supplementary Information for more details). Figure 4 presents the median test AUCs for different DL models considered here for the ENCODE datasets A and B. As is apparent from the plots, our model, XDeMo reports the highest median test

AUCs for almost all TFs, in each set (see Supplementary Table 3 for more details). The lowest median test AUC for any TF that is in JASPAR or HOCOMOCO is greater than 0.95 which is superior to the DeepBind versions as well as DeepSEA. Even when we consider the full set of TFs in sets A and B, the minimum value for median test AUC is approximately 0.94. Table 2 displays the comparison of the test AUC, mean precision, mean recall, and mean F1 scores among the various DL models over the ENCODE ChIP-seq datasets on the held-out set. All our XDeMo models report superior values for all the standard metrics considered here (shown in boldface).

The choice of a model can be aided by comparing the average AUCs for training and validation. Overfitting may be indicated if the training AUC is noticeably greater than the validation AUC. On the other hand, underfitting or poor generalization may be present if the validation AUC is significantly lower than the training AUC. Figure 5 represents various performance metrics compared across the three XDeMo model variants. The fact that all of our models have comparable mean training and validation AUCs shows that our model accurately captures the relationships and patterns seen in training sequences and generalizes to previously unseen sequences well. Further, the test set AUCs for all three models are also very high and comparable to both validation and training set AUCs, confirming that the model is indeed not suffering from the common problems of overfitting or underfitting in any way and can make accurate predictions on the evaluation set. XDeMo has higher precision (accurately predicted true positive rate) than any other model, meaning that it has the least number of false positives. Similarly, our model has the highest sensitivity or recall rates as well as the cumulative F1 score (harmonic mean of precision and sensitivity). However, it is worth noting that XDeMo trained on set B performs slightly worse than those models trained on sets A1 and A2. XDeMo-A2 has the best performance among the three variants due to the size of the dataset being doubled as compared to A1 and B. Set A2 contains 10,500 positive sequences whereas sets A1 and B contain only 5,250 positive samples. Between sets B and A1, XDeMo's performance is better on A1 for most of the TFs, mainly because this set has been derived using alternate sampling from a larger set and thus may contain a greater number of higher quality samples than set B in terms of signal values of the peak-called data. Additionally, set B contains 63 TFs while set A comprises 49 TFs making set A more robust in terms of signal attributes as these TFs have a greater ChIP-seq peak experiments across different human cell lines.

In contrast to all other models evaluated, XDeMo employs the GELU function rather than ReLU activation. GELU is a smooth and continuous function in contrast to ReLU, which is a piecewise linear function with a flat gradient for negative inputs. By offering a well-behaved gradient over the entire input space, GELU's smoothness aids the Adam optimization further. Training may become more reliable and effective as a result. Also, GELU helps to capture non-linear and complex patterns in the sequences which are important for eliciting high-level motif features. ReLU often suffers from the problem of 'dying neurons', especially for negative samples where many neurons become inactive due to zero gradients and their contribution to the network becomes void (Lu 2020). By translating negative inputs into non-zero values, GELU, on the other hand, maintains information from negative sequences that helps our model retain more information and perform better.

The baseline DL models use either a one-hot encoding scheme (DeepBind, DeepSEA, DanQ, Basset, DeepSite) or k-mer encoding (DNABERT) both of which have some limitations. The sparsity of one-hot and k-mer encoding

Model	Mean Test AUC	Mean F1 Score	Mean Precision	Mean Recall
Basset	0.86	0.685	0.799	0.729
DanQ	0.91	0.823	0.848	0.823
DeepBind	0.919	0.85	0.837	0.877
DeepSEA	0.919	0.836	0.84	0.858
DeepSite	0.88	0.795	0.817	0.822
DESSO	0.926	0.848	0.832	0.884
DNABERT-6	0.954	0.901	0.898	0.909
XDeMo-B	**0.964**	**0.905**	**0.968**	**0.937**
XDeMo-A1	**0.976**	**0.923**	**0.974**	**0.949**
XDeMo-A2	**0.987**	**0.942**	**0.982**	**0.96**

Table 2 Comparison of various baseline performance metrics across various DL models and XDeMo. A part of the data presented in this table is based on (Ji et al. 2021). DNABERT-6 refers to the DNABERT model that uses 6 bp wide k-mers. The table has been color-coded according to their values ranging from the lowest (towards the red spectrum) to the highest (towards the green spectrum)

restricts the non-transformer models to a small dataset size, thereby resulting in reduced model performance. The computation time for training these models is also high owing to a large number of convolution kernels in CNN or hybrid models. For instance, training a DanQ model on an NVIDIA Titan X GPU (Graphics Processing Unit) over 60 epochs roughly requires 360 h (i.e., 15 days) (Quang and Xie 2016). DNABERT which is a transformer-based model, reports a humongous pre-training time of 25 days for their models on eight NVIDIA 2080Ti GPUs (Ji et al. 2021). Our simplified transformer model architecture enhanced with blended encoding alleviates these limitations significantly. On the one hand, our encoding scheme alleviates the problem of sparsity while retaining the majority of low-level and high-level feature information including the positional weightage of these features (see Supplementary Table 1 for details about variance retention rates). On the other hand, the architecture of the transformer model enables it to efficiently learn these features and make accurate predictions on the potential binding sites and associated motifs quickly. For instance, XDeMo-B trained for a particular TF takes approximately 10 min over 50 epochs on a Tesla T4 GPU (Intel Xeon CPU @ 2.20 GHz). The approximate lower limit on the total time it takes to train all XDeMo model variants for all TFs in their respective set is 30 h, while the upper limit depends on the hyperparameter tuning procedure time for some motifs that are more complex than others (estimated at 4–5 days' total training time for all models on average). This immense speedup can be attributed to the advantage of the transformer model's high parallelizability in contrast to CNN, RNN, or hybrid models.

The multi-head self-attention network of the transformer layers focuses on individual parts of the input sequences simultaneously, allowing the model to capture long-term dependencies effectively. As motifs might appear at different junctures within the sequence, this can be advantageous for DNA motif detection. Transformers are also adaptable to varied dataset sizes since the number of layers and model dimensions may be simply increased or decreased during hyperparameter tuning (see Supplementary Table 2 for details on the choices of hyperparameters and Supplementary Table 4 for actual hyperparameters selected by grid search with early stopping and dropout regularization). Compared to DNABERT, our model has a maximum of 6 transformer layers with a maximum of 512 hidden units and 8 self-attention heads per layer. Since these parameters are fixed in DNABERT and flexible in our model, XDeMo can adapt well to different complexities of the various motifs and TFs. Moreover, DNABERT is more complex with double the number of transformer layers and a higher number of hidden units and self-attention heads per layer which is another reason for its much higher pre-training time. Since DNABERT uses the pre-trained BERT model which was initially built for language translation and other NLP (Natural Language Processing) tasks, it has to be fine-tuned to adapt to the DNA motif discovery problem. This fine-tuning further adds to the computation time and complexity of the network, making it less interpretable.

The predicted scores from our model were translated to 30 bp length sequences centered at the peaks of the scores with a threshold value ranging from 0.5 to 0.9. These sequences were stacked vertically and motif visualizations based on positional nucleotide probabilities (MEME) were compared with known motifs from the HOCOMOCO database (containing a large number of high-quality human TF binding models) using Tomtom (Pearson correlation coefficient and E-value less than unity). Figure 6 displays the Tomtom motif comparisons for all the annotated TFs in JASPAR and HOCOMOCO that were matched with their counterparts in these databases. Our model can predict 34 motifs out of the 39 annotated TF motifs with high statistical significance (p-value less than 0.05, E-value less than 1 for class A, B, and C motifs in HOCOMOCO version 11) as well as the rest of the unannotated ones. None of the earlier DL models have been able to predict as many DNA motifs as our model. Our trained models can accurately predict motifs in a small number of sequences with ease and are also scalable and flexible to accommodate a larger number of sequences.

4 Concluded comments

Our paper makes a significant contribution to the field of computer science, genomics, and information technology by proposing a novel deep learning-based approach for DNA motif prediction. In this paper, we used a Transformer-based model with a self-attention mechanism to solve the challenge of DNA motif mining. We also introduced a novel encoding mechanism and used it to train our models. Our approach was aimed at capturing both short-range and long-range relationships as well as complicated patterns in DNA sequences, allowing us to predict DNA motifs with high accuracy with shorter training times. Through extensive experimentation and evaluation on a benchmark dataset, we demonstrated the effectiveness of our approach. Our model achieved a high AUC of 0.987 (XDeMo-A2), indicating its ability to discriminate between binding and non-binding sequences. Additionally, it exhibited strong performance in terms of F1 score (0.942, XDeMo-A2), precision (0.982, XDeMo-A2), and recall (0.96, XDeMo-A2), highlighting its capability to accurately identify true TFBSs and motifs while minimizing false positives. We further discerned that our Transformer-based technique outperformed standard

RNNs, CNNs, their hybrids as well as other transformer models (DNABERT) in capturing complicated sequence patterns when compared to existing advanced algorithms. The Transformer model's multi-head self-attention mechanism enabled it to successfully understand relationships across distant nucleotide bases, resulting in better prediction accuracy. Moreover, our analysis revealed biologically meaningful motifs discovered by the model. A majority of these motifs aligned well with known TF binding motifs, validating the ability of our model to capture important regulatory sequences.

Nonetheless, it is crucial to acknowledge that the performance of the model is intrinsically linked to the quality and volume of the labeled peak-called dataset used during training. The existence of motifs that align with known TF binding motifs suggests potential in the model, but their functional importance requires additional examination and confirmation.

4.1 Practical applications

The practical ramifications of the proposed XDeMo framework are extensive and encompass diverse domains within genomics and computational biology. Firstly, the model's development and validation of an exceptionally precise motif prediction tool hold great potential for researchers involved in regulatory genomics. The accurate identification of transcription factor binding sites (TFBSs) and motifs is paramount for unraveling the intricacies of gene regulation. With its impressive precision and recall metrics, the model serves as a valuable asset for pinpointing regulatory sequences, thereby expediting focused experiments aimed at unraveling the mysteries of gene regulatory networks.

Secondly, the comparative analysis against existing algorithms underscores the supremacy of the Transformer-based approach. This revelation carries practical implications for researchers and practitioners in search of cutting-edge tools for sequence analysis. The Transformer's capacity to capture intricate sequence patterns and interrelationships across nucleotide bases can significantly benefit a wide array of applications, including variant detection, investigations into disease associations, and endeavors in the realm of functional genomics.

Furthermore, the model's discovery of biologically meaningful motifs offers a valuable resource for molecular biologists and geneticists. These motifs provide invaluable insights into potential regulatory elements embedded within genomic sequences. Experimental validation of these motifs has the potential to deepen our understanding of gene regulation, potentially unearthing novel targets for therapeutic interventions and drug discovery.

While the model's predictions exhibit a high degree of accuracy, the capacity to elucidate the underlying biological mechanisms and rationale behind specific predictions is of paramount importance. This interpretability factor can guide experimental design and hypothesis generation, empowering researchers to craft experiments that validate the model's predictions with precision.

The proposed encoding scheme, in itself, can be independently employed wherever genomic sequence patterns need to be analyzed using deep learning techniques. Machine learning models can easily incorporate this hybrid form of encoding to account for both short- and long-term dependencies. Because of this flexibility, researchers may be able to strike a balance between both of these kinds of dependencies based on the specific demands of their research. It is a versatile approach that may be used for a variety of genomic tasks, making it a valuable tool in genomics and computational biology.

4.2 Future initiatives

Future initiatives in the domain of motif mining offer exciting prospects for enhancing our understanding of DNA sequences. The integration of additional attention mechanisms represents a promising avenue for improving motif predictions. Sparse attention, in particular, holds the potential to enhance the model's ability to pinpoint crucial motifs amid complex genomic backgrounds. Moreover, a comprehensive assessment of the model's performance across an extensive array of benchmark datasets is paramount. Expanding the scope to encompass datasets characterized by varying motif attributes, complexities, and distributions would provide a holistic understanding of the model's strengths and limitations. Such an inclusive evaluation would ensure that the model's applicability extends across diverse genomic contexts, fortifying its utility in real-world applications.

In essence, the accurate prediction of regulatory sequences stands as a pivotal element in comprehending disease-associated variants and designing diagnostic assays for conditions with a genetic underpinning. The XDeMo framework, with its multifaceted contributions, emerges as a versatile tool poised to catalyze advancements across various facets of genomics and computational biology.

XDeMo is generalizable and scalable to different dataset sizes which enhances the robustness of our model. Our work paves the way for the use of sophisticated deep learning techniques in genomics research and has the potential to better knowledge of gene control processes, personalized drug discovery, and machine learning.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s13721-

024-00463-4.

Author contributions Rajashree Chaurasia and Udayan Ghose conceptualized the model architecture and methodology. Rajashree Chaurasia carried out the literature survey, data collection, preprocessing, and analysis, model construction, training, and evaluation, and wrote the manuscript. Udayan Ghose supervised and reviewed the manuscript preparation.

Data availability All the ChIP-Seq datasets that were used in this study were downloaded from the ENCODE (Encyclopedia of DNA Elements) database, which can be accessed and downloaded freely from the ENCODE website link (available at http://hgdownload.soe.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeAwgTfbsUniform/). The preprocessing steps that were performed on these datasets are detailed in the Methods section.

Declarations

Conflicts of interest The authors declare no conflicts of interest. No funding was received for conducting this study.

References

Alipanahi B, Delong A, Weirauch MT, Frey BJ (2015) Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat Biotechnol 33(8):831–838. https://doi.org/10.1038/nbt.3300

Avsec Ž et al (2021) Effective gene expression prediction from sequence by integrating long-range interactions. Nat Methods 18(10):1196–1203. https://doi.org/10.1038/s41592-021-01252-x

Chaurasia R, Ghose U (2023) Human DNA/RNA motif mining using deep-learning methods: a scoping review. Netw Model Anal Health Inf Bioinf 12(1). https://doi.org/10.1007/s13721-023-00414-5

Choong AC, Lee NK (2017) Evaluation of convolutional neural networks modeling of DNA sequences using ordinal versus one-hot encoding method. 2017 International Conference on Computer and Drone Applications (IConDA). https://doi.org/10.1109/iconda.2017.8270400

ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489(7414):57–74. https://doi.org/10.1038/nature11247

Falk T et al (2018) U-Net: deep learning for cell counting, detection, and morphometry. Nat Methods 16(1):67–70. https://doi.org/10.1038/s41592-018-0261-2

Floridi L, Chiriatti M (2020) GPT-3: its nature, scope, limits, and consequences. Mind Mach 30(4):681–694. https://doi.org/10.1007/s11023-020-09548-1

Fukushima K (1975) Cognitron: a self-organizing multilayered neural network. Biol Cybern 20(3–4):121–136. https://doi.org/10.1007/bf00342633

Gunasekaran H et al (2021) Analysis of DNA sequence classification using CNN and Hybrid models. Comput Math Methods Med 2021:1–12. https://doi.org/10.1155/2021/1835056

He M, Miyajima F, Roberts P et al (2012) Emergence and global spread of epidemic healthcare-associated Clostridium difficile. Nat Genet 45:109–113. https://doi.org/10.1038/ng.2478

He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/cvpr.2016.90

Hendrycks D, Gimpel K (2016) Gaussian Error Linear units (GELUs). arXiv e-prints. https://doi.org/10.48550/arXiv.1606.08415

Hitz BC et al (2023) Encode Unif Anal Pipelines. https://doi.org/10.1101/2023.04.04.535623

Ji Y, Zhou Z, Liu H, Davuluri RV (2021) DNABERT: pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. Bioinformatics 37(15):2112–2120. https://doi.org/10.1093/bioinformatics/btab083

Jin S, Zeng X, Xia F, Huang W, Liu X (2020) Application of deep learning methods. Biol Networks Briefings Bioinf 22(2):1902–1917. https://doi.org/10.1093/bib/bbaa043

Kamath U, Graham KL, Emara W (2022) Bidirectional encoder representations from Transformers (BERT). In Transformers for Machine Learning, pp. 43–70. https://doi.org/10.1201/9781003170082-3

Kelley DR, Snoek J, Rinn JL (2016) Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. Genome Res 26(7):990–999. https://doi.org/10.1101/gr.200535.115

Kingma DP, Ba J (2014) Adam: A Method for Stochastic Optimization. CoRR. doi: abs/1412.6980

Lin QXX, Thieffry D, Jha S, Benoukraf T (2019) TFregulomeR reveals transcription factors' context-specific features and functions. Nucleic Acids Res 48(2). https://doi.org/10.1093/nar/gkz1088

Lu L (2020) Dying ReLU and initialization: theory and numerical examples. Commun Comput Phys 28(5):1671–1706. https://doi.org/10.4208/cicp.oa-2020-0165

Luo Y et al (2019) New Developments on the encyclopedia of DNA elements (ENCODE) Data Portal. Nucleic Acids Res 48(D1). https://doi.org/10.1093/nar/gkz1062

Madrid F et al (2019) Matrix profile XX: Finding and visualizing time series motifs of all lengths using the matrix profile. 2019 IEEE International Conference on Big Knowledge (ICBK). https://doi.org/10.1109/icbk.2019.00031

Mannor S, Peleg D, Rubinstein R (2005) The cross entropy method for classification. Proc 22nd Int Conf Mach Learn - ICML '05. https://doi.org/10.1145/1102351.1102422

Nutiu R et al (2011) Direct measurement of DNA affinity landscapes on a high-throughput sequencing instrument. Nat Biotechnol 29(7):659–664. https://doi.org/10.1038/nbt.1882

OpenAI (2023) GPT-4 Technical Report. ArXiv. abs/2303.08774

Otten NV (2023) Self-attention made easy and how to implement it. Spot Intelligence. Accessed May 11, 2023. [URL: https://spotintelligence.com/2023/01/31/self-attention/]

Pardiñas AF et al (2018) Common schizophrenia alleles are enriched in mutation-intolerant genes and maintained by background selection. Nat Genet 50(3):381–389. https://doi.org/10.1038/s41588-018-0059-2

Poliakov A, Foong J, Brudno M, Dubchak I (2014) GenomeVISTA—an integrated software package for whole-genome alignment and visualization. Bioinformatics 30(18):2654–2655. https://doi.org/10.1093/bioinformatics/btu355

Quang D, Xie X (2016) DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. Nucleic Acids Res 44(11). https://doi.org/10.1093/nar/gkw226

Siggers T, Gordân R (2013) Protein–DNA binding: complexities and multi-protein codes. Nucleic Acids Res 42(4):2099–2111. https://doi.org/10.1093/nar/gkt1112

Suter DM (2020) Transcription factors and DNA play hide and seek. Trends Cell Biol 30(6):491–500. https://doi.org/10.1016/j.tcb.2020.03.003

Trabelsi A, Chaabane M, Ben-Hur A (2019) Comprehensive evaluation of deep learning architectures for prediction of DNA/RNA sequence binding specificities. Bioinformatics 35(14):i269–i277. https://doi.org/10.1093/bioinformatics/btz339

Vaswani A et al (2017) Attention Is All You Need. 31st Conference on Neural Information Processing Systems (NIPS 2017),

Long Beach, CA, USA, Dec. 2017. https://doi.org/10.48550/arXiv.1706.03762

Wang C et al (2014) The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance. Nat Biotechnol 32(9):926–932. https://doi.org/10.1038/nbt.3001

Xu H, Jia P, Zhao Z (2021) DeepVISP: deep learning for virus site integration prediction and motif discovery. Adv Sci 8(9):2004958. https://doi.org/10.1002/advs.202004958

Yang J et al (2019) Nucleic Acids Res 47(15):7809–7824. https://doi.org/10.1093/nar/gkz672. Prediction of regulatory motifs from human chip-sequencing data using a deep learning framework.

Zambelli F, Pesole G, Pavesi G (2012) Motif discovery and transcription factor binding sites before and after the next-generation sequencing era. Brief Bioinform 14(2):225–237. https://doi.org/10.1093/bib/bbs016

Zeng H, Edwards MD, Liu G, Gifford DK (2016) Convolutional neural network architectures for predicting DNA–protein binding. Bioinformatics 32(12):i121–i127. https://doi.org/10.1093/bioinformatics/btw255

Zhang Y, Qiao S, Ji S, Li Y (2019) DeepSite: bidirectional LSTM and CNN models for predicting DNA–protein binding. Int J Mach Learn Cybernet 11(4):841–851. https://doi.org/10.1007/s13042-019-00990-x

Zhang S et al (2021) Assessing deep learning methods in cis-regulatory motif finding based on genomic sequencing data. Brief Bioinform 23(1). https://doi.org/10.1093/bib/bbab374

Zhou J, Troyanskaya OG (2015) Predicting effects of noncoding variants with deep learning–based sequence model. Nat Methods 12(10):931–934. https://doi.org/10.1038/nmeth.3547