TRAFICA: Improving Transcription Factor Binding Affinity Prediction using Deep Language Model on ATAC-seq Data

Yu Xu[1], Chonghao Wang[1], Ke Xu[1], Yi Ding[1], Aiping Lyu[2,*], and Lu Zhang[1,3,*]

[1]Department of Computer Science, Hong Kong Baptist University
[2]School of Chinese Medicine, Hong Kong Baptist University
[3]Institute for Research and Continuing Education, Hong Kong Baptist University
[*]Corresponding author: aipinglu@hkbu.edu.hk, ericluzhang@hkbu.edu.hk

ABSTRACT

In silico prediction of transcription factor (TF) binding affinity plays a vital role in determining TF binding preferences and understanding gene regulation. Many existing tools utilize known TF-DNA binding data to train conventional machine learning and deep learning models to predict TF-DNA binding affinities. The majority of these tools do not consider the natural preferences of DNA sequences binding to TFs, which could be influenced by chromatin accessibility. In this study, we developed TRAFICA, a deep language model to predict TF-DNA binding affinities by integrating chromatin accessibility from ATAC-seq and known TF-DNA binding data. We pre-trained TRAFICA on an integrated ATAC-seq dataset with over 13 million nucleotide sequences in a self-supervised manner to learn potential TF-DNA binding preferences and contextual relationships within DNA sequences. TRAFICA was fine-tuned using PBM and HT-SELEX datasets to predict *in vitro* TF-DNA binding affinities. We also implemented AdapterFusion to enhance its ability to predict *in vivo* TF-DNA binding affinities. We observed TRAFICA significantly outperformed the six existing tools in predicting *in vitro* TF-DNA binding affinities and was comparable with the best tool on *in vivo* prediction.

***K*eywords** transcription factor binding affinity · chromatin accessibility · *in vitro* prediction

1 Introduction

The binding of transcription factors (TFs) to DNA is a crucial step in gene regulation. Typically, TFs recognize and interact with specific promoter sequences in genomes to regulate the expression of nearby downstream genes. The binding preferences of TFs *in vivo* are more intricate than that observed *in vitro*, as *in vivo* TF binding mechanism could be influenced by cellular dynamics, such as the positions of nucleosomes and co-binding with other factors [1, 2]. Nucleosome positions can restrict access to binding sites, preventing TFs from binding to certain DNA sequences, and the co-binding mechanism refers to the cooperative binding of multiple TFs and co-factors, such as the transcriptional corepressor TUP1-SSN6 complex in yeast [3]. In contrast, *in vitro* TF binding preferences are primarily attributed to the intrinsic binding of TFs to DNA sequences, and can be assayed through *in vitro* binding experiments in the absence of intricate cellular environments [4, 5]. These experiments can provide valuable insights into investigating *in vivo* TF binding preferences.

Several technologies have been developed to measure TF binding preferences, both *in vivo* and *in vitro*. For example, Chromatin ImmunoPrecipitation followed by sequencing (ChIP-seq) [6] is a popular *in vivo* sequencing technology used to identify the binding regions of a specific TF in a genome-wide manner. This technology has a significant limitation, as the antibodies for chromatin immunoprecipitation are not available for many TFs [1]. Protein Binding Microarray (PBM) [7] is an *in vitro* sequencing technology that measures the binding intensities of a specific protein to a set of artificial DNA sequences, in which all k-mers can be represented. High-Throughput Systematic Evolution of Ligands by EXponential enrichment (HT-SELEX) [8] is a well-established *in vitro* technology, which involves a multi-cycle selection procedure to screen sequences binding to a target protein. This technology could yield abundant sequences with high affinities to a TF of interest. Moreover, Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq) [9] can be utilized to assess chromatin accessibility (**Figure 1A**) that influences the binding of TFs. The rapid development of these technologies has contributed to large-scale publicly available sequencing data that have yet to be fully utilized [10, 11, 12].

A focus of interest lies in the development of computational approaches based on *in vitro* experimental data to ascertain the intrinsic binding preferences of a specific TF and infer its binding affinity to a new sequence (i.e., inferring the affinity to the promoter of a gene). Previous studies [4, 13, 14, 15, 16, 17] have trained the conventional machine learning and deep learning models to predict relative TF-DNA binding affinities *in vitro* using sequence data from PBM and HT-SELEX experiments. For instance, DNAshapeR [13] employs a multiple linear regression model with $L2$ penalty and utilizes 13 DNA shape features (including inter-base pair types, intra-base pair types, and minor groove width (MGW)) derived from sequences as inputs to predict binding affinities. DNAffinity [14] represents a sequence as a feature vector composed of 4 classes of DNA properties (base pair parameters, stiffness, sequence pattern, and electrostatics), and employs a random forest regression model for binding affinity prediction. However, conventional machine learning models typically necessitate feature engineering on DNA sequences to obtain fixed-length DNA descriptors. In contrast, deep learning models possess the capability to learn hidden patterns directly from sequences of varying lengths without feature engineering. DeepBind [15] applies a convolutional neural network (CNN) to capture binding motifs by CNN kernels and predict sequence preferences. DLBSS [16] integrates 4 DNA shape features with sequences as inputs to the CNN model for inferring binding affinities. CRPTS [17] combines a long short-term memory network (a variant of recurrent neural network) with CNN as a hybrid model to further enhance predictive performance. DeepSELEX [4] is specifically designed to learn the enrichment pattern of sequences from HT-SELEX data, employing a multi-class CNN as its framework.

Although these models have demonstrated advancements in predicting *in vitro* TF-DNA binding affinities, they have yet to incorporate chromatin accessibility that affects the natural preferences of DNA sequences binding to TFs. In this study, we developed a deep language model named TRAFICA (**TR**anscription factor binding **AF**finity prediction **I**ntegrating **C**hromatin **A**ccessibility) that integrated chromatin accessibility and known TF-DNA binding data to predict TF-DNA binding affinities. The framework of TRAFICA is based on the transformer-encoder architecture [18], which has emerged as the most popular framework in the field of natural language processing (NLP) in recent years. Additionally, pre-training is widely used in large language models to facilitate learning hidden patterns from large-scale unlabeled data, benefiting related downstream tasks. Successful language models such as BERT [19], T5 [20], and the GPT series (including GPT-3 [21], InstructGPT (a prototype of ChatGPT) [22], and GPT-4 [23]) present the effectiveness of the models pre-trained with large volumes of data. To take advantage of pre-training, we constructed an ATAC-seq dataset (**Supplementary Figure S1**)) with over 13 million sequences to pre-train TRAFICA in order to learn contextual relationships within DNA sequences and potential TF binding preferences (**Figure 1B**). We utilized 506 TF-DNA binding datasets from PBM and HT-SELEX experiments to fine-tune the pre-trained model and to evaluate the ability of TRAFICA in predicting *in vitro* binding affinities (**Figure 1C(1)**). TRAFICA outperforms the contemporary *in vitro* binding affinity prediction tools including DeepBind [15], DLBSS [16], CRPT/CRPTS [17], DNAshapeR [13] and DNAffinity [14]. Specifically, TRAFICA achieves the highest average Pearson correlation

coefficient (PCC) of 0.649 on 66 PBM datasets (**Figure 2**), and the best average coefficient of determination (R^2) of 0.977 on 440 HT-SELEX datasets (**Figure 3**). We also employed a fine-tuning approach AdapterFusion [24] to improve the performance of TRAFICA in predicting *in vivo* binding affinities for a particular TF by incorporating binding information from other TF-DNA binding datasets (**Figure 1C(2)** and **(3)**). The result shows that TRAFICA reaches state-of-the-art (SOTA) performance on *in vivo* prediction (**Figure 4**).

2 Methods

2.1 Data collection

We collected ATAC-seq data from the human ATAC-seq experiments in the Encyclopedia of DNA Elements (ENCODE) project [10, 11]. Specifically, we chose the ATAC-seq experiments that were conducted on three biosample types: cell lines, tissues, and primary cells (defined by the ENCODE consortium), and downloaded all available "bed narrowPeak" files of these experiments. As a quality control step, peaks with low statistical significance ($-log_{10}(p\text{-value}) \leq 100$) were removed. The p-values for each peak are provided in the "bed narrowPeak" files, computed by the peak calling pipeline of the ENCODE project. For each of the remaining peaks, we identified a 200-bp genomic region (99-bp upstream and 100-bp downstream) centered at the peak summit (the right panel of **Figure 1A**). We then utilized Bedtools [25] to retrieve the corresponding sequence in this region from the GRCh38 reference genome (version: GCF_000001405.40), which was downloaded from the National Center for Biotechnology Information (NCBI) database. After executing the aforementioned procedures, we obtained a total of 13,013,152 ATAC-seq sequences. These sequences are derived from 422 individual experiments (**Supplementary Table S1** and **Supplementary Figure S1**), and utilized to pre-train TRAFICA to improve its predictive capacity.

We utilized PBM data from the DREAM5 PBM study [26] to fine-tune the pre-trained model to predict *in vitro* binding affinities. The DREAM5 PBM experiments employ two distinct microarray designs called 'HK' and 'ME' on 66 mouse TFs, totaling 132 experimental datasets. Each dataset contains a set of 35-bp sequences with corresponding binding intensities. The limited availability of high-affinity sequences is an inherent disadvantage of PBM data. To address this issue, the authors in the DREAM5 PBM study implemented a combination of eight data pre-processing methods for enhancing predictive performance [26]. However, in this study, we opted not to employ any data pre-processing methods in order to assess the efficacy of TRAFICA in capturing binding patterns from less informative data.

Additionally, we utilized public HT-SELEX experimental data from three previous studies [27, 28, 29] to fine-tune TRAFICA. These experimental data are available at the European Nucleotide Archive (ENA) [12]. We downloaded the following data files from the ENA: (1) Jolma et al. [27]: 547 HT-SELEX experiments covering 461 unique human and mouse TFs (ENA accession: PRJEB3289); (2) Yang et al. [28]: 521 HT-SELEX experiments covering 445 unique human and mouse TFs (ENA accession: PRJEB14744); (3) Yin et al. [29]: 2265 HT-SELEX experiments covering 550 unique human TFs (ENA accessions: PRJEB20112 for initial selection cycles and PRJEB9797 for other selection cycles). For the HT-SELEX experiments from Jolma et al. [27], we only retained the experimental data used in DeepBind [15], where the clone type was indicated (full-length TF or DNA-binding domain only). We also filtered out the experiments that lacked any of the following information: Ensembl ID, structural domains, and amino acid sequences. The filtered results are listed in **Supplementary Table S2**. To ensure the reliability of HT-SELEX experiments, we performed rigorous quality control (**Supplementary Note 1**) to filter out low-quality experiments that exhibited low levels of sequence enrichment and diversity. After quality control, 440 high-quality experiments remained (**Supplementary Table S3**), consisting of 339 unique TFs (297 human TFs and 42 mouse TFs). For each high-quality HT-SELEX experiment, we utilized the R programming package "SELEX" [30] to calculate relative k-mer binding affinities and merged the k-mer affinities for each sequence to obtain sequence affinities (**Supplementary Note 2**). The sequences and corresponding affinities from the last selection cycles of HT-SELEX are used to fine-tune the pre-trained model, as the last cycle typically contains a higher ratio of high-affinity sequences compared to the early cycles in an HT-SELEX experiment. Furthermore, the sequencing experiments in Yang et al. [28] have a notably higher read depth compared to the others, with an almost 10-fold increase in the average number of reads per sequencing file (from 168,000 to 1,656,000) compared to Jolma et al. [27]. This results in a large number of sequences, which significantly increases the time spent on fine-tuning and could potentially destroy the knowledge learned during pre-training. To alleviate this issue, we limited our analysis to the first 0.5 million sequences ordered by relative affinities in each dataset from Yang et al. [28], as we observed that most datasets from the three studies contained fewer than 0.5 million sequences (**Supplementary Figure S4**).

Moreover, we incorporated the ChIP-seq data used in DeepBind [15] to test the model's predictive capability for *in vivo* binding affinities, including 507 datasets representing 138 unique TFs. We identified TFs that were presented in both HT-SELEX and ChIP-seq datasets to establish a list of *in vitro* - *in vivo* dataset pairs for *in vivo* evaluation. This yielded 108 pairs consisting of 42 HT-SELEX and 73 ChIP-seq datasets, covering 28 unique TFs (**Supplementary**

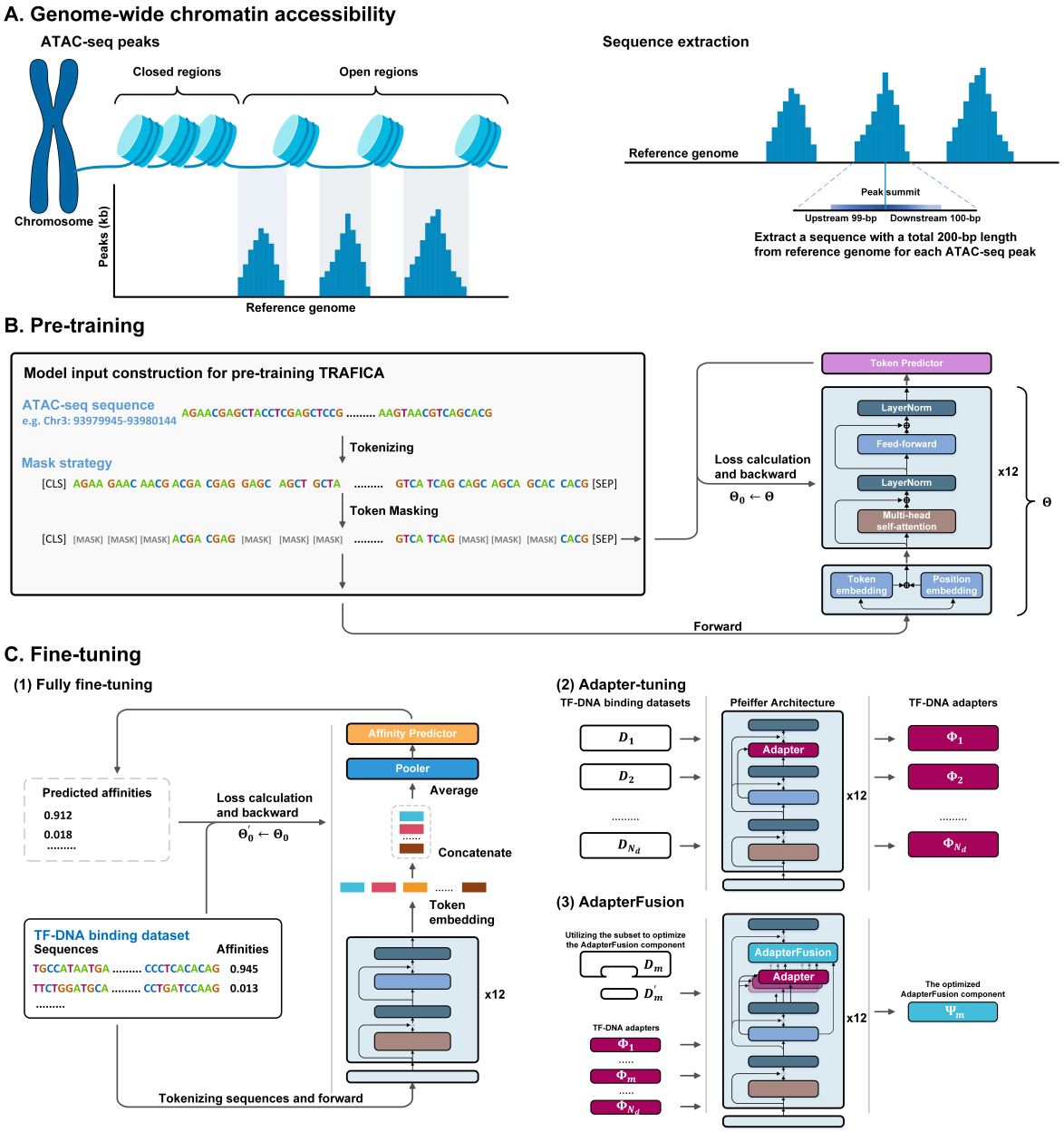

Figure 1. Overview of TRAFICA. **(A)** Left panel: Diagram of chromatin accessibility. Chromatin-accessible regions in a genome can be detected with ATAC-seq technology. Right panel: Sequence extraction from the reference genome. For each ATAC-seq peak, a 200-bp sequence centered at its peak summit is extracted. **(B)** Pre-training TRAFICA. Model inputs are constructed by tokenizing ATAC-seq sequences and masking multiple contiguous regions within each sequence. A token predictor is incorporated to perform self-supervised learning by predicting the masked regions. Θ and Θ_0 represent the parameters of the initial and pre-trained models, respectively. **(C)** Fine-tuning TRAFICA. (1) The pre-trained model is fully fine-tuned on a specific TF-DNA binding dataset. A pooler module and an affinity predictor are employed in order to perform binding affinity prediction. $\Theta_0^{'}$ represents the parameters of the fully fine-tuned model. (2) An adapter is incorporated into the pre-trained model and fine-tuned on a TF-DNA binding dataset. Θ_0 is frozen during this phase. For a set of datasets $\{D_1, ..., D_{N_d}\}$, the TF-DNA adapters $\{\Phi_1, ..., \Phi_{N_d}\}$ can be fine-tuned. (3) To enhance the predictive performance of TRAFICA for a target TF m, an AdapterFusion component is employed to combine the TF binding information carried by the TF-DNA adapters $\{\Phi_1, ..., \Phi_m, ..., \Phi_{N_d}\}$. The adapter Φ_m is fine-tuned on a dataset D_m that is experimentally assayed from the TF m. The AdapterFusion component (Ψ_m) is optimized on a subset $D_m^{'}$ that is sampled from D_m, learning to weight the TF-DNA adapters for TF-DNA binding affinity prediction. Θ_0 and $\{\Phi_1, ..., \Phi_m, ..., \Phi_{N_d}\}$ were frozen during this phase.

Table 1. Model configuration

Model components	Size
Dimension of token embeddings	768
Maximum length of input tokens	512
Number of attention heads	12
Number of transformer-encoder blocks	12
Dimension of feed-forward embeddings	3072

Table S4). Furthermore, we selected 138 ChIP-seq datasets out of 507 to implement AdapterFusion for enhancing the *in vivo* prediction. Considering a particular TF had multiple corresponding experimental data, we randomly selected one of them to ensure the selected datasets involved all TFs in the ChIP-seq datasets (**Supplementary Table S5**).

2.2 Overview of TRAFICA

To process nucleic acid sequences, we tokenize each sequence into an ordered list of 4-mers. In this case, each token is represented by four consecutive nucleotides, such as "AGAA" and "GAAC" shown in the left panel of **Figure 1B**. We generate a token vocabulary comprising all possible 4-mers, encompassing four symbols representing DNA nucleotides and the symbol 'N' denoting unidentified nucleotides within sequences. These unidentified nucleotides may arise due to technical limitations and sequencing errors. Following previous studies [19, 31], we also introduce five special tokens into the vocabulary, including "[CLS]," "[SEP]," "[MASK]," "[PAD]," and "[UNK]." Specifically, "[CLS]" is added at the beginning of sequences, and is typically used for sequence classification; "[SEP]" indicates the end of sequences; "[MASK]" is applied for token prediction during the pre-training phase; "[PAD]" is used to pad sequences of varying lengths to a fixed length, which enables efficient processing of batches of sequences (note that the padding token "[PAD]" does not contribute to the calculation of attention scores); "[UNK]" is assigned to unrecognized tokens that are absent in the vocabulary. The size of the vocabulary is 630 ($5^4 + 5$).

TRAFICA framework is based on the vanilla transformer-encoder structure [18, 19], which only utilizes the self-attention mechanism to capture contextual relationships in sequential data. As shown in the right panel of **Figure 1B**, the model structure consists of a token embedding layer, a position embedding layer, and twelve transformer-encoder blocks. The two embedding layers are fully connected networks designed to encode the semantic and positional information of tokens. Each transformer-encoder block consists of four components: a multi-head self-attention module, a feed-forward module, layer normalization, and residual connections. Within the multi-head self-attention module, the self-attention mechanism enables the model to capture the complex relationships of input sequences by computing attention scores of each token pair, and the multi-head mechanism allows the model to attend to the information of input sequences from different aspects. The feed-forward module is a stack of two fully connected layers with a non-linear activation function called Gaussian Error Linear Units (GELU) [32], enabling the model to learn intricate dependencies between tokens. Layer normalization is applied to normalize the hidden representation, and residual connections are used to preserve information from earlier layers. The model configuration is presented in **Table 1**.

Moreover, in the pre-training phase (**Figure 1B**), a token predictor (**Supplementary Figure S2A**) is incorporated to facilitate self-supervised learning. The token predictor is composed of two fully connected layers, a GELU activation function and layer normalization, which is discarded after pre-training. Three fine-tuning approaches are employed in this study: fully fine-tuning, adapter-tuning, and AdapterFusion. In the fully fine-tuning phase (**Figure 1C(1)**), a pooler module and an affinity predictor (**Supplementary Figure S2B, C**) are utilized for binding affinity prediction. The pooler module is implemented as a fully connected layer with a Tanh activation function, and the affinity predictor consists of two fully connected layers. In the adapter-tuning phase (**Figure 1C(2)**), an adapter is incorporated into the pre-trained model using the Pfeiffer architecture [24], and is fine-tuned to learn TF binding patterns from a specific TF-DNA binding dataset. The structure of adapters (**Supplementary Figure S3A**) consists of two fully connected layers and a ReLU activation function [33]. Adapter-tuning is the preparatory phase for AdapterFusion. In the AdapterFusion phase (**Figure 1C(3)**), an AdapterFusion component (**Supplementary Figure S3B**) is incorporated into the pre-trained model, allowing the model to weight each adapter based on the attention mechanism [24]. More details of the model structure are presented in **Supplementary Note 3**.

2.3 Pre-training

As shown in **Figure 1B**, we utilize the collected ATAC-seq sequences to pre-train TRAFICA. Masking individual tokens will simplify the model pre-training and obscure the model from learning deep semantic relationships, as a single token can be determined by its preceding and succeeding tokens [31]. For example, in a tokenized sequence

"AGAA [MASK] AACG", the masked token can be ascertained as "GAAC" by its neighbor tokens. To address this issue, we mask contiguous tokens instead of individual ones. In practice, we mask multiple contiguous regions of a sequence rather than a single contiguous region (the right panel of **Figure 1B**). This allows the model to learn more nuanced relationships among tokens, thereby improving the accuracy of the token prediction (**Supplementary Note 4**). The proportion of masked tokens in each sequence is 15% (by default) in this study. The cross-entropy of masked token prediction for a sequence is formulated as follows:

$$L_{mask} = \frac{1}{M}\sum_{i=1}^{M} -\log(\frac{\exp(\hat{T}_i[v(t_i)])}{\sum_{j=1}^{|v|}\exp(\hat{T}_i[j])}) \tag{1}$$

where M represents the number of masked tokens in the sequence. $v(\cdot)$ denotes the index of a specific token in the token vocabulary, and $|v|$ represents the size of the token vocabulary. $\hat{T}_i \in \mathbb{R}^{|v|}$ is the output of the token predictor for i-th masked token. Given an ATAC-seq dataset denoted as D_0, we minimize the cross-entropy of masked tokens in the pre-training phase:

$$\Theta_0 \leftarrow \arg\min_{\Theta} L_{mask}(D_0;\Theta) \tag{2}$$

where Θ and Θ_0 represent the parameters of the initial and the pre-trained model, respectively.

TRAFICA is pre-trained over a total number of 110,000 steps using a batch size of 384. During the pre-training phase, we employ a linear warm-up strategy for the learning rate, increasing it from 0 to 0.0001 over the first 10,000 steps. Subsequently, we decay the learning rate to 0 over the remaining steps. We apply the AdamW optimizer [34] with the hyperparameters $\beta_1 = 0.9$, $\beta_2 = 0.98$, and $\epsilon = 10^{-7}$ to calculate the gradient and update the model parameters. The pre-training process is executed on an Nvidia Tesla A100 GPU.

2.4 Fully fine-tuning

We fully fine-tune the pre-trained model on the TF-DNA binding datasets obtained from PBM and HT-SELEX experiments, enabling TRAFICA to predict binding affinities for specific TFs. In this phase, all of the model parameters are trainable. As shown in **Figure 1C(1)**, the token embeddings out of the last transformer-encoder block are concatenated and averaged to form a joint embedding. This embedding is then passed to the pooler module (**Supplementary Figure S2B**) to be transformed as a sequence embedding that serves as input to the affinity predictor (**Supplementary Figure S2C**). Given a specific TF-DNA binding dataset (denoted as $D^{'}$) with N sequences and corresponding affinities $\{y_1, y_2, ..., y_N\}$, we minimize the mean squared error (MSE) in the fully fine-tuning phase as follows:

$$\Theta_0^{'} \leftarrow \arg\min_{\Theta_0} L_{MSE}(D^{'};\Theta_0) \tag{3}$$

$$L_{MSE} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2 \tag{4}$$

where $\Theta_0^{'}$ represents the parameters of the model fine-tuned on $D^{'}$. $\hat{y}_i$ is the output of the affinity predictor for i-th sequence. For each TF-DNA binding dataset, we randomly select 1/5 of the data as a held-out test set and fine-tune the pre-trained model on the remaining data. During the fine-tuning phase, we employ 1/10 of the remaining data as a validation set for model selection and early stopping, and 9/10 as a training set for gradients calculation to update model parameters. TRAFICA is fine-tuned using the AdamW optimizer with a constant learning rate of 0.00002 and a batch size of 128.

2.5 AdapterFusion

We implement AdapterFusion [24] to enhance the predictive ability of TRAFICA by integrating other TF binding information. We first utilize TF-DNA binding datasets to fine-tune adapters during the adapter-tuning phase (**Figure 1C(2)**), resulting in a set of TF-DNA adapters that learned diverse TF binding information. The fine-tuning configuration of adapter-tuning is largely similar to those for fully fine-tuning, including the loss calculation (**Equation 4**), dataset split (test, training, and validation sets), optimizer, learning rate, and batch size. The only difference is that the parameters of the pre-trained model (Θ_0) are frozen in the adapter-tuning phase. Considering that we have a set of TF-DNA binding datasets $\{D_1, D_2, ..., D_{N_d}\}$ for N_d unique TFs, we minimize the MSE in the adapter-tuning phase as follows:

$$\Phi_i \leftarrow \arg\min_{\Phi} L_{MSE}(D_i;\Theta_0,\Phi) \tag{5}$$

where Φ and Φ_i ($i \in \{1, 2, ..., N_d\}$) represent the parameters of the initial and the fine-tuned adapters, respectively. Subsequently, we combine the set of TF-DNA adapters to improve the prediction of binding affinities for a target TF (corresponding to the dataset D_m and the adapter Φ_m) by optimizing the AdapterFusion component (**Figure 1C(3)** and **Supplementary Figure S3B**) as follows:

$$\Psi_m \leftarrow \arg\min_{\Psi} L_f(D_m^{'}; \Theta_0, \Phi_1, ..., \Phi_m, ..., \Phi_{N_d}, \Psi) \tag{6}$$

where Ψ and Ψ_m represent the parameters of the initial and the optimized AdapterFusion component, respectively. The dataset D_m is employed again to optimize the AdapterFusion component. If the size of the dataset D_m is large (e.g., over 100,000 sequences in most HT-SELEX datasets (**Supplementary Figure S4**)), it may lead to the model overfitting to the target TF-DNA adapter Φ_m, thereby disregarding the effects of other TF-DNA adapters. To mitigate this issue, we randomly sample a subset $D_m^{'}$ from D_m for the optimization of the AdapterFusion component. Specifically, the test sets in $D_m^{'}$ and D_m are identical, while the training and validation sets in $D_m^{'}$ consist of 2,000 and 1,000 sequences (by default) sampled from the corresponding sets in D_m, respectively. The utilization of subsets is justified, as the TF-DNA adapters have learned the TF binding patterns during the adapter-tuning phase. Moreover, the parameters of the pooler module and the affinity predictor are randomly initialized at the beginning of each adapter-tuning and AdapterFusion procedure.

2.6 Evaluation metrics

We employ several evaluation metrics to assess the performance of TRAFICA. Specifically, we utilize PCC and R^2 to evaluate the performance in predicting *in vitro* affinities. Given a set of input sequences $X = \{x_1, x_2, ..., x_n\}$ with the label $Y = \{y_1, y_2, ..., y_n\}$ and the predicted affinities $f(X) = \{f(x_1), f(x_2), ..., f(x_n)\}$, the computation of these two metrics is formulated as follows:

$$PCC(Y, f(X)) = \frac{\sum_{i=1}^{n}((y_i - \bar{y})(f(x_i) - \overline{f(x)}))}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(f(x_i) - \overline{f(x)})^2}} \tag{7}$$

$$R^2(Y, f(X)) = 1 - (\frac{\sum_{i=1}^{n}(y_i - f(x_i))^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}) \tag{8}$$

where f and n represent the prediction model and the number of sequences in this set, respectively. $\overline{f(x)} = \frac{1}{n}\sum_{i=1}^{n} f(x_i)$ is the mean value of predicted affinities. Additionally, we employ the area under the receiver operating characteristic curve (AUROC) as the metric to evaluate the performance of predicting *in vivo* binding affinities on ChIP-seq datasets, considering the binary nature of the labels in ChIP-seq data.

3 Results

3.1 Performance of predicting TF-DNA relative binding affinities on PBM data

We adopted the DREAM5 PBM evaluation protocol to assess the performance of TRAFICA and compared it with existing tools, including DeepBind [15], DLBSS [16], CRPT/CRPTS [17], DNAshapeR [13], and DNAffinity [14]. CRPTS employed an identical framework as CRPT and incorporated four DNA shape features as an extra input. The evaluation protocol involves fine-tuning models using TFs 1 to 33 on 'HK' sequences and testing on the corresponding 'ME' sequences, as well as fine-tuning models using TFs 34 to 66 on 'ME' sequences and testing on the corresponding 'HK' sequences. We employed PCC as the evaluation metric, as the fine-tuning and testing datasets for a specific TF were generated from two distinct microarray designs ('HK' and 'ME') that led to the incompatibility of relative affinities between the datasets. As shown in **Figure 2A** and **Supplementary Figure S5**, TRAFICA achieved the best performance with an average PCC of 0.649, significantly outperformed the second-best tool DNAffinity with an average PCC of 0.433 (p<1e-12). We further analyzed our predicted results for each TF family within the DREAM5 PBM datasets. As shown in **Figure 2B**, TRAFICA had the highest PCC of 0.838 (50% PI=[0.799, 0.891], PI: percentile interval) for the TF family with *Myb/SANT* protein domain.

We also utilized PBM data to investigate the influence of using two distinct pooling methods to generate a joint embedding: (1) mean tokens: averaging all token embeddings out of the last transformer-encoder block; (2) [CLS] token: utilizing the embedding of "[CLS]" as the joint embedding. As shown in **Figure 2C**, applying the mean tokens method yielded slightly better predictive performance than utilizing the [CLS] token method. To assess the effectiveness of the pre-trained model, we directly trained prediction models from scratch and compared their performance against the

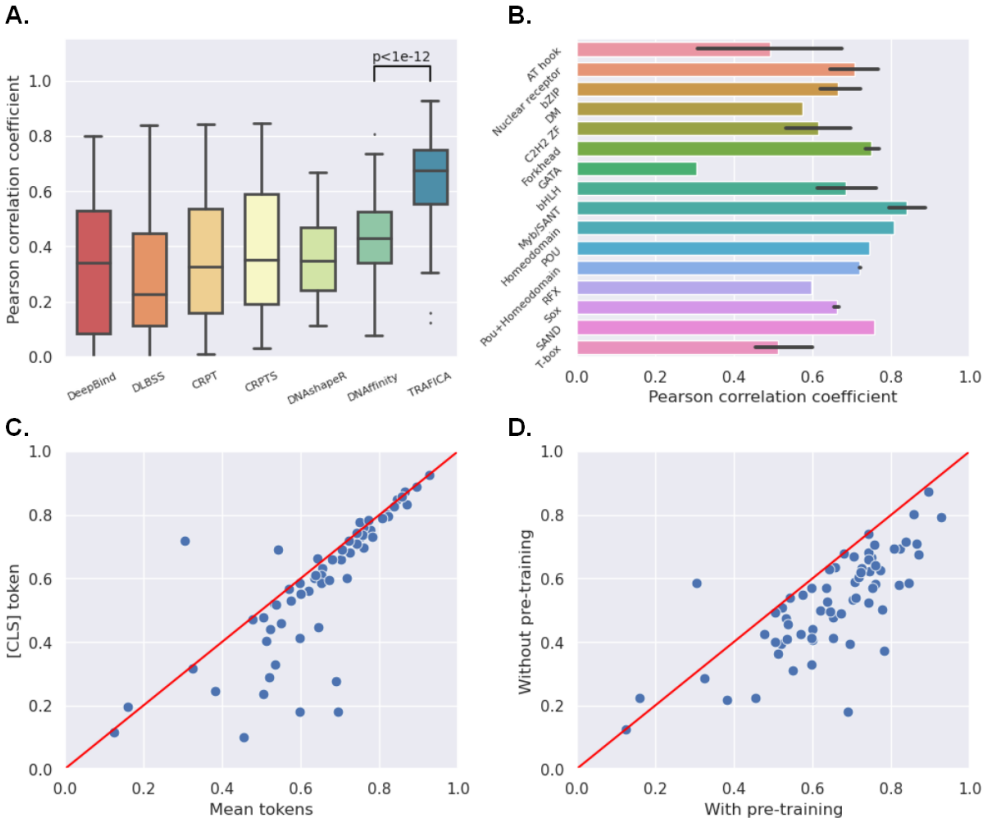

Figure 2. Performance of *in vitro* binding affinity prediction on PBM data. **(A)** Predictive performance comparison of TRAFICA and six existing tools on the PBM datasets. TRAFICA significantly outperforms the other tools, with a p<1e-12 as compared to the second-best tool. The black line inside each box represents the median of each tool's performance. P-values were calculated using Wilcoxon rank-sum test. **(B)** Predictive performance of TRAFICA for different TF families. The error bars indicate the 50th percentile intervals. **(C)** Impact of utilizing two distinct pooling methods. The use of the mean tokens method is slightly better than the method of [CLS] token in predicting binding affinities. **(D)** Effectiveness of the pre-trained model. The utilization of the pre-trained model leads to a notable improvement in most cases.

models fine-tuned from the pre-trained model on the same datasets. We observed a significant increase in the average PCC from 0.526 to 0.649 (p<1e-5) by utilizing the pre-trained model, which demonstrated the superiority of employing the pre-trained model over training models from scratch (**Figure 2D**).

3.2 Performance of predicting TF-DNA relative binding affinities on HT-SELEX data

We evaluated the performance of TRAFICA on the high-quality HT-SELEX datasets, which exhibited significant sequence enrichment (**Supplementary Table S3**) in contrast to the low enrichment of PBM datasets. Apart from the tools discussed in the previous section, we also compared the performance of TRAFICA with DeepSELEX [4], which was specifically developed for modeling HT-SELEX data. R^2 and PCC were used as the evaluation metrics. As shown in **Figure 3**, TRAFICA significantly outperformed other tools. In particular, when evaluated on the datasets from Jolma et al. [27] ($n = 40$), Yang et al. [28] ($n = 274$), and Yin et al. [29] ($n = 126$), TRAFICA achieved an average R^2 (PCC) of 0.973 (0.995), 0.987 (0.997), and 0.957 (0.992), respectively. n represents the number of datasets. More detailed analyses of the performances between TRAFICA and each respective tool are shown in **Supplementary Figure S6** and **S7**. We also conducted a TF family-based analysis and observed no significant difference in performance for different TF families (**Supplementary Figure S8**). This observation might be attributed to the high sequence enrichment of the HT-SELEX datasets utilized in this study.

Moreover, we observed that deep-learning-based tools effectively learned TF binding patterns solely from sequences, even in the absence of structural features. As shown in **Figure 3A**, the difference in performance between CRPT (taking

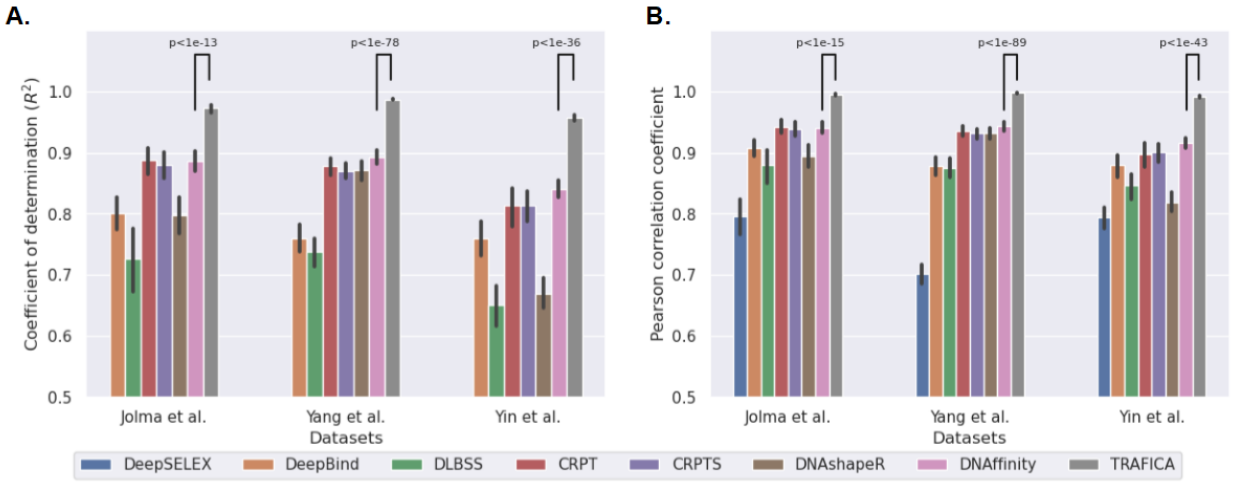

Figure 3. Performance of *in vitro* binding affinity prediction on HT-SELEX data. Predictive performance comparison of TRAFICA and existing tools on the HT-SELEX datasets employing two evaluation metrics (**A**) Coefficient of determination (R^2) and (**B**) Pearson correlation coefficients (PCC). We only employed PCC to evaluate DeepSELEX, as it was trained without the utilization of binding affinities. TRAFICA significantly outperforms the existing tools. Error bars represent 95% confidence intervals. P-values were calculated using Wilcoxon rank-sum test.

sequences as inputs) and CRPTS (taking sequences and structural features as inputs) was statistically insignificant (p=0.641 for Jolma et al. [27]; p=0.159 for Yang et al. [28]; p=0.367 for Yin et al. [29]; P-values were calculated based on R^2). Additionally, DeepSELEX demonstrated the poorest performance, primarily due to the lack of incorporating binding affinities as labels during training.

3.3 Integration of *in vitro* and *in vivo* TF binding information improving prediction on ChIP-seq data

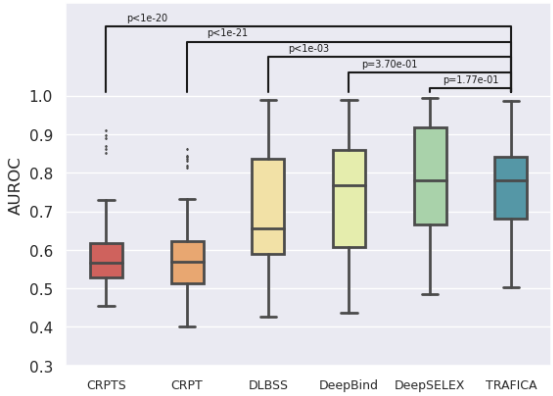

Figure 4. Performance of *in vivo* binding affinity prediction on ChIP-seq data. TRAFICA demonstrates a competitive performance compared to contemporary tools. P-values were calculated using Wilcoxon rank-sum test. If the difference in performance between TRAFICA and a particular tool is statistically insignificant ($p \geq 0.05$), the exact p-value is reported.

Several studies have utilized *in vitro* models to predict *in vivo* binding affinities, investigating the potential value of *in vitro* intrinsic binding preferences in understanding complex *in vivo* binding mechanisms [15, 4]. Thus, we assessed the *in vivo* prediction ability of TRAFICA by fine-tuning the model on an HT-SELEX dataset and then testing it on the corresponding ChIP-seq datasets with identical TF. This assessment was implemented on 108 HT-SELEX - ChIP-seq dataset pairs. Besides, we assumed that if two TFs were related (e.g., TFs involved in an identical biological pathway), their *in vivo* binding preferences would be similar. Based on this assumption, given a target TF of interest, we applied AdapterFusion to integrate its *in vitro* binding information with the *in vivo* binding information from other TFs (excluding the target TF) to enhance the ability of TRAFICA in predicting *in vivo* binding affinities

(**Figure 1C(2), C(3)**). More details of the AdapterFusion implementation are illustrated in **Supplementary Note 5**. We compared the *in vivo* predictive performance of TRAFICA with other tools including DeepBind, DLBSS, CRPT/CRPTS, and DeepSELEX. To ensure a fair comparison, we did not adopt the tools based on conventional machine learning models, as they typically required inputs of a fixed length. This is incompatible with our *in vivo* evaluation protocol that employs the sequences with varying lengths during fine-tuning (HT-SELEX: 20, 30, or 40-bp) and testing (ChIP-seq: 101-bp). The area under the receiver operating characteristic curve (AUROC) was used as the evaluation metric since the labels of ChIP-seq datasets were binary. As shown in **Figure 4**, the differences in performance between TRAFICA and two SOTA tools (DeepSELEX and DeepBind) were statistically insignificant, as evidenced by the p-values of 0.177 and 0.370, respectively. Additionally, we analyzed the influence of three fine-tuning approaches (fully fine-tuning, adapter-tuning, and AdapterFusion) on the prediction of *in vivo* binding affinities. The result indicated that the AdapterFusion approach significantly outperformed the others (p =1.53e-2 against fully fine-tuning; p =2.56e-2 against adapter-tuning; **Supplementary Figure S9**).

Moreover, we observed that the models performed well on HT-SELEX datasets tended to have a lower predictive performance on ChIP-seq datasets, as evidenced by the results shown in **Figure 3** and **Figure 4**. This observation could be attributed to the discrepancies in TF binding patterns between *in vivo* and *in vitro* conditions due to the presence of cellular environments [2, 1]. Even two *in vivo* ChIP-seq experiments in different cell lines could have distinct binding patterns for an identical TF [35]. Remarkably, TRAFICA implemented with AdapterFusion yielded a superior performance to DeepSELEX and DeepBind in predicting *in vitro* binding affinities (**Supplementary Figure S10**). This indicated that TRAFICA achieved SOTA performance in predicting both *in vitro* and *in vivo* binding affinities.

3.4 Analysis of TF binding motifs

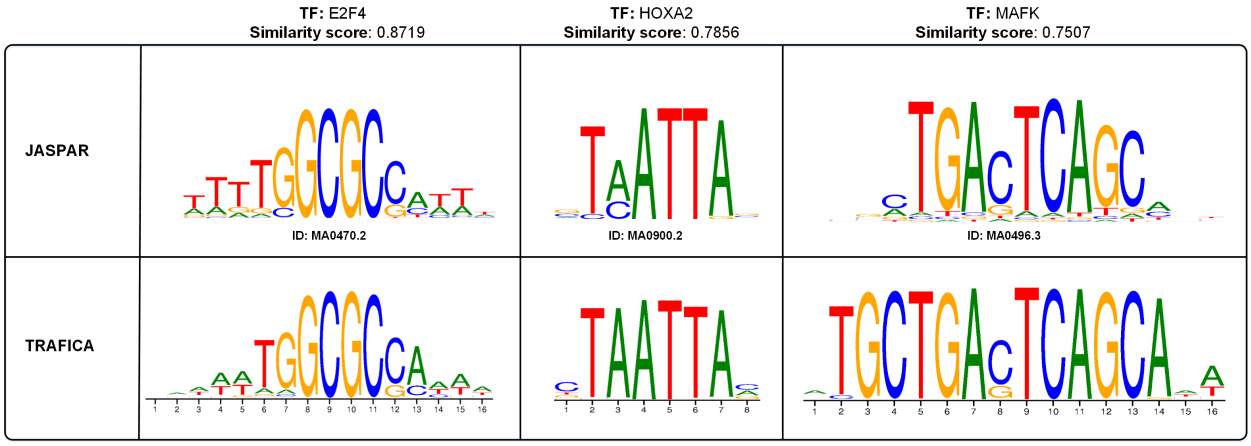

Figure 5. Motif logos of three TFs: E2F4, HOXA2, and MAFK. The top panel displays the validated TF binding motifs from the JASPAR2022 database, whereas the bottom panel presents the TF binding motifs derived from the attention scores computed by TRAFICA. The similarities between each two motifs were measured by employing MoSBAT-e scores.

Sequence motifs are commonly represented as position frequency matrices (PFMs) or sequence logos that provide an intuitive visualization of TF binding preferences. We employed an attention-based method (modified from DNABERT-viz [31], **Supplementary Note 6**) to derive TF binding motifs. In particular, we employed fully fine-tuned models on the HT-SELEX datasets with corresponding top-1000 high-affinity sequences from those datasets as inputs. During the forward propagation, we collected the attention scores of each sequence from the multi-head self-attention modules for motif discovery. The results were compared with the validated binding motifs from the JASPAR2022 database. We identified the TFs that were presented in both derived and validated motifs to construct comparable motif pairs, leading to 362 pairs (**Supplementary Table S5**). Two motifs in a pair represent the binding preferences of an identical TF. We employed MoSBAT energy (MoSBAT-e) scores [36] to quantify the similarity of each motif pair, and utilized "WebLogo" [37] to generate motif logos for visualization. The logos of the top-3 motif pairs ranked by similarities are shown in **Figure 5**, and their similarities are 0.8719, 0.7856, and 0.7507, corresponding to the TFs of E2F4 (JASPAR ID: MA0470.2), HOXA2 (JASPAR ID: MA0900.2), and MAFK (JASPAR ID: MA0496.3), respectively. The high concordance of these three pairs provides a possible explanation for the advanced performance of TRAFICA in predicting binding affinities on HT-SELEX data, implying the high effectiveness of the model in

learning biologically relevant TF-DNA binding patterns. Furthermore, we represented the distribution of similarities in **Supplementary Figure S11** and displayed the motif logos in **S12-S19** for 362 pairs.

4 Discussion

The abundance of sequencing data from PBM, HT-SELEX, ChIP-seq, and ATAC-seq technologies provides opportunities to uncover the binding preferences of TFs in different aspects. Although ChIP-seq technology allows for direct measurement of genome-wide binding patterns for specific TFs, it does have certain limitations [1]. For example, the antibodies required for ChIP-seq experiments are not available for many TFs, restricting the investigation of those TFs. Additionally, the results of ChIP-seq experiments are influenced by cell types and epigenetic modifications (e.g., methylation), posing challenges in discovering intrinsic TF binding preferences. PBM technology is commonly utilized to detect the binding intensities of designed sequences through fluorescence signals [38] for a particular TF of interest. Most designed sequences have low binding affinities with the TF. In contrast, HT-SELEX experiments can yield sequence data with a higher ratio of high-affinity sequences, thereby making it suitable to explore the binding affinities of protein-nucleic acid complexes. Several studies have utilized HT-SELEX technology to investigate the binding preferences of TFs [8, 27, 28, 29], RNA-binding proteins [39, 40], and drug-target proteins (e.g., aptamer-binding proteins [41]). Unlike the previously mentioned sequencing technologies that assess the binding preferences of specific TFs, ATAC-seq technology measures chromatin accessibility that impacts the natural preferences of TF-binding. In this study, we pre-trained TRAFICA to learn these natural preferences from ATAC-seq data and fine-tuned the pre-trained model on TF-DNA binding datasets to learn the binding preferences of specific TFs. Our results demonstrated that the integration of ATAC-seq data significantly improves the predictive performance of TF-DNA binding affinities.

We observed a general discrepancy in the prediction between *in vitro* and *in vivo* binding affinities. The models that achieved the superior predictive performance of *in vitro* binding affinities commonly perform inferiorly in predicting *in vivo* binding affinities (**Figure 3** and **4**). This discrepancy can be attributed to the difference between *in vivo* and *in vitro* TF-binding, as *in vivo* TF-binding is not only determined by intrinsic binding preferences but also influenced by dynamic cellular environments [1, 42]. To alleviate this issue, we employed the fine-tuning approach AdapterFusion to incorporate the binding information of multiple TFs to improve the *in vivo* prediction for a target TF. Specifically, we combined the binding information from the *in vitro* data of the target TF and the available *in vivo* data of other TFs, which allowed TRAFICA to achieve *in vivo* SOTA predictive performance for the target TF without its *in vivo* binding data. Moreover, the AdapterFusion approach presents two potential valuable advantages: 1. It can handle different types of labels, such as continuous labels in PBM and HT-SELEX datasets, and binary labels in ChIP-seq datasets. 2. It can address the incompatibility of relative affinities between TF-DNA binding datasets.

Although TRAFICA performed well in predicting relative TF-DNA binding affinities, there were several limitations. For instance, if we integrate a large number of adapters in the procedure of AdapterFusion, the inference time will increase significantly. The utilization of AdapterFusion in TRAFICA might be unable to enhance performance for a particular TF, if the integrated adapters are learned from the datasets completely irrelevant to the TF. Moreover, the relative binding affinities of HT-SELEX sequences were estimated by sub-sequence enrichment among experimental selection cycles, which might introduce estimation biases. We are actively developing TRAFICA to address these issues and to extend the model to the prediction of general protein-nucleic acid binding affinities (e.g., protein-aptamer complex) in future work.

Data availability

The code of TRAFICA is available at `https://github.com/ericcombiolab/TRAFICA`. The pre-trained model and the fine-tuned TF-DNA adapters (HT-SELEX and ChIP-seq) are available at `https://huggingface.co/Allanxu/TRAFICA`. The processed sequencing data of ATAC-seq, HT-SELEX, PBM, and ChIP-seq can be accessed from `https://zenodo.org/doi/10.5281/zenodo.8248339`.

Author contributions statement

LZ and YX conceived the study; YX designed TRAFICA; YX implemented the algorithm; YX conducted the experiments and analyzed the results; YX wrote the article; CHW, KX, YD, APL, and LZ reviewed the paper. All authors read and approved the final manuscript.

References

[1] Samuel A Lambert et al. "The human transcription factors". In: *Cell* 172.4 (2018), pp. 650–665.

[2] Matthew Slattery et al. "Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins". In: *Cell* 147.6 (2011), pp. 1270–1282.

[3] Alexander D Johnson. "Molecular mechanisms of cell-type determination in budding yeast". In: *Current opinion in genetics & development* 5.5 (1995), pp. 552–558.

[4] Maor Asif and Yaron Orenstein. "DeepSELEX: inferring DNA-binding preferences from HT-SELEX data using multi-class CNNs". In: *Bioinformatics* 36.Supplement_2 (2020), pp. i634–i642.

[5] Yaron Orenstein and Ron Shamir. "A comparative analysis of transcription factor binding models learned from PBM, HT-SELEX and ChIP data". In: *Nucleic acids research* 42.8 (2014), e63–e63.

[6] David S Johnson et al. "Genome-wide mapping of in vivo protein-DNA interactions". In: *Science* 316.5830 (2007), pp. 1497–1502.

[7] Michael F Berger et al. "Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities". In: *Nature biotechnology* 24.11 (2006), pp. 1429–1435.

[8] Arttu Jolma et al. "Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities". In: *Genome research* 20.6 (2010), pp. 861–873.

[9] Jason D Buenrostro et al. "ATAC-seq: a method for assaying chromatin accessibility genome-wide". In: *Current protocols in molecular biology* 109.1 (2015), pp. 21–29.

[10] ENCODE Project Consortium et al. "An integrated encyclopedia of DNA elements in the human genome". In: *Nature* 489.7414 (2012), p. 57.

[11] Yunhai Luo et al. "New developments on the Encyclopedia of DNA Elements (ENCODE) data portal". In: *Nucleic acids research* 48.D1 (2020), pp. D882–D889.

[12] Rasko Leinonen et al. "The European nucleotide archive". In: *Nucleic acids research* 39.suppl_1 (2010), pp. D28–D31.

[13] Jinsen Li et al. "Expanding the repertoire of DNA shape features for genome-scale studies of transcription factor binding". In: *Nucleic Acids Research* 45.22 (2017), pp. 12877–12887.

[14] Sandro Barissi et al. "DNAffinity: a machine-learning approach to predict DNA binding affinities of transcription factors". In: *Nucleic Acids Research* 50.16 (2022), pp. 9105–9114.

[15] Babak Alipanahi et al. "Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning". In: *Nature biotechnology* 33.8 (2015), pp. 831–838.

[16] Qinhu Zhang, Zhen Shen, and De-Shuang Huang. "Predicting in-vitro transcription factor binding sites using DNA sequence+ shape". In: *IEEE/ACM transactions on computational biology and bioinformatics* 18.2 (2019), pp. 667–676.

[17] Siguo Wang et al. "Predicting transcription factor binding sites using DNA shape features based on shared hybrid deep learning architecture". In: *Molecular Therapy-Nucleic Acids* 24 (2021), pp. 154–163.

[18] Ashish Vaswani et al. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).

[19] Jacob Devlin et al. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805* (2018).

[20] Colin Raffel et al. "Exploring the limits of transfer learning with a unified text-to-text transformer". In: *The Journal of Machine Learning Research* 21.1 (2020), pp. 5485–5551.

[21] Tom Brown et al. "Language models are few-shot learners". In: *Advances in neural information processing systems* 33 (2020), pp. 1877–1901.

[22] Long Ouyang et al. "Training language models to follow instructions with human feedback". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 27730–27744.

[23] OpenAI. "GPT-4 Technical Report". In: *ArXiv* abs/2303.08774 (2023).

[24] Jonas Pfeiffer et al. "AdapterFusion: Non-destructive task composition for transfer learning". In: *arXiv preprint arXiv:2005.00247* (2020).

[25] Aaron R Quinlan and Ira M Hall. "BEDTools: a flexible suite of utilities for comparing genomic features". In: *Bioinformatics* 26.6 (2010), pp. 841–842.

[26] Matthew T Weirauch et al. "Evaluation of methods for modeling transcription factor sequence specificity". In: *Nature biotechnology* 31.2 (2013), pp. 126–134.

[27] Arttu Jolma et al. "DNA-binding specificities of human transcription factors". In: *Cell* 152.1-2 (2013), pp. 327–339.

[28] Lin Yang et al. “Transcription factor family-specific DNA shape readout revealed by quantitative specificity models”. In: *Molecular systems biology* 13.2 (2017), p. 910.

[29] Yimeng Yin et al. “Impact of cytosine methylation on DNA binding specificities of human transcription factors”. In: *Science* 356.6337 (2017), eaaj2239.

[30] C Rastogi et al. “SELEX: Functions for analyzing SELEX-seq data. R package version 1.30.0”. In: *Bioconductor* (2022).

[31] Yanrong Ji et al. “DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome”. In: *Bioinformatics* 37.15 (2021), pp. 2112–2120.

[32] Dan Hendrycks and Kevin Gimpel. “Bridging Nonlinearities and Stochastic Regularizers with Gaussian Error Linear Units”. In: *CoRR* abs/1606.08415 (2016).

[33] Jonas Pfeiffer et al. “Adapterhub: A framework for adapting transformers”. In: *arXiv preprint arXiv:2007.07779* (2020).

[34] Ilya Loshchilov and Frank Hutter. “Decoupled weight decay regularization”. In: *arXiv preprint arXiv:1711.05101* (2017).

[35] Jason Gertz et al. “Genistein and bisphenol A exposure cause estrogen receptor 1 to bind thousands of sites in a cell type-specific manner”. In: *Genome research* 22.11 (2012), pp. 2153–2162.

[36] Samuel A Lambert et al. “Motif comparison based on similarity of binding affinity profiles”. In: *Bioinformatics* 32.22 (2016), pp. 3504–3506.

[37] Gavin E Crooks et al. “WebLogo: a sequence logo generator”. In: *Genome research* 14.6 (2004), pp. 1188–1190.

[38] Michael F Berger and Martha L Bulyk. “Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors”. In: *Nature protocols* 4.3 (2009), pp. 393–411.

[39] Arttu Jolma et al. “Binding specificities of human RNA-binding proteins toward structured and linear RNA sequences”. In: *Genome research* 30.7 (2020), pp. 962–973.

[40] Kaitlin U Laverty et al. “PRIESSTESS: interpretable, high-performing models of the sequence and structure preferences of RNA-binding proteins”. In: *Nucleic Acids Research* 50.19 (2022), e111–e111.

[41] Jiehua Zhou and John Rossi. “Aptamers as targeted therapeutics: current potential and challenges”. In: *Nature reviews Drug discovery* 16.3 (2017), pp. 181–202.

[42] Werner Pieter Veldsman. “GRID-seq assisted prediction of transcription factor binding motifs”. In: *bioRxiv* (2018), p. 429332.

Supplementary Notes

Supplementary Note 1: Quality control for HT-SELEX experiments

Given an HT-SELEX experiment with N experimental selection cycles (excluding the initial cycle), we implement quality control as follows:

1. We apply a fifth-order Markov model to estimate the expected frequency of 8-mers (denoted as E_0) in the initial cycle. We also calculate the observed frequencies of 8-mers (O_i, $i \in \{1, ..., N\}$) in other selection cycles, respectively. The enrichment score of 8-mers in the i-th cycle (denoted as ES_i) is defined as the ratio of O_i to E_0 [43].
2. Following the previous studies [44, 45], we exclude this experiment based on the following criteria: (a) the enrichment patterns are uncorrelated in neighboring cycles, as evidenced by the Spearman's rank correlation coefficient of the enrichment scores ES_3 and ES_4 is less than 0.8; (b) the final cycle exhibits a low level of sequence enrichment and diversity, indicated by the high-frequent sequences (counts>10) less than 500; (c) no 8-mer count is greater than 100 in the initial cycle, which could result in inaccurate estimates of E_0.
3. We remove the low-complexity sequences that present a DUST [46] score larger than 2 in the final cycle, as low-complexity sequences may consist of the repeated k-mer with a high relative affinity. Given a sequence a with a length $n > 2$, its DUST score $S(a)$ is formulated as follows:

$$S(a) = \frac{\sum_{t \in R} c_t(a)(c_t(a) - 1)/2}{l - 1} \tag{9}$$

   where R is the set of 64 possible triplets of DNA base combinations (e.g., AAA, AAT, ATT,...). $l = n - 2$ is the number of triplets in the sequence a. $c_t(a)$ represents the frequency of occurance of the triplet t within the sequence a.

Supplementary Note 2: Relative binding affinity estimation of HT-SELEX data

We utilize the R language package "SELEX" to estimate relative binding affinities of k-mers on HT-SELEX data. The estimation is based on the cycle-to-cycle enrichment of k-mers [43, 47]. The optimal hyper-parameter k is determined by computing the information gain (Kullback-Leibler divergence of selection cycle 2 to cycle 0). We employ a searching range $k \in \{3, 4, ..., 13, 14\}$, as the minimum sequence length of the HT-SELEX datasets used in this study is 14. Given a sequence a, we then merge the k-mer affinities A_{kmer} into the sequence affinity $A_{seq}(a)$ by the following:

$$A_{seq}(a) = \frac{\sum_{i=1}^{K} A_{kmer_i}}{\sqrt{K}} \tag{10}$$

where K represents the number of k-mers in the sequence a. This approach (Eq. 10) is introduced in the previous study [48] that integrates the statistics of k-mers to generate the relative binding affinity of a sequence. To demonstrate the validity and reasonableness of estimated binding affinities, we present the distribution of sequence affinities on the HT-SELEX datasets from Jolma et al. [49]. As shown in **Figure S20**, in most datasets, the proportion of high-affinity sequences is very small.

Supplementary Note 3: Model structure of TRAFICA

S 3.1: Embedding layers

The token embedding and position embedding layers are two fully connected networks, denoted as $W_t \in \mathbb{R}^{d_{vocab} \times d}$ and $W_p \in \mathbb{R}^{l_{max} \times d}$, where d_{vocab} and l_{max} represent the size of the token vocabulary and the maximum number of input tokens. Given a sequence with a length $l_s \geq 4$, it is tokenized as a list of 4-mer tokens and two specific tokens '[CLS]' and '[SEP]' at both ends. The length of this token list is $l = l_s - 3 + 2$. Each token t and the corresponding position index p are converted as two one-hot encoding vectors $x_t \in \mathbb{R}^{d_{vocab}}$ and $x_p \in \mathbb{R}^{l_{max}}$, respectively. The process of the embedding layers can be formulated as follows:

$$h_t = x_t^T W_t \oplus x_p^T W_p \tag{11}$$

$$H^0 = Concat(h_1, h_2, ...h_l) \tag{12}$$

where $h_i \in \mathbb{R}^d$ is the hidden embedding capturing the semantic and positional information of the token t_i, and d denotes the dimension of the hidden embedding. $Concat(\cdot)$ is the operation of concatenation. The matrix $H^0 \in \mathbb{R}^{l \times d}$ of the sequence is the input for the first transformer-encoder block.

S 3.2: Transformer-encoder

The structure of the transformer-encoder is composed of four components, a multi-head self-attention module, a feed-forward module, layer normalization, and residual connections. Each transformer-encoder block takes as input the output from the previous block, except the first block. Given the transformer-encoder block with n_{head} attention heads, the dimension of hidden representation $d_k = d/n_{head}$ in each head, and an input matrix $X \in \mathbb{R}^{l \times d}$, the process of the multi-head self-attention module can be described as follows:

$$MultiHead(X) = Concat(AttHead_1, AttHead_2, ...AttHead_{n_{head}})W^O \tag{13}$$

where $W^O \in \mathbb{R}^{d \times d}$ is the learnable parameters used to aggregate the outputs of attention heads. The self-attention mechanism for each head is defined as follows:

$$AttHead_i = Softmax(\frac{XW_i^Q \otimes (XW_i^K)^T}{\sqrt{d_k}})XW_i^V \tag{14}$$

where $\{W_i^Q \in \mathbb{R}^{d \times d_k}, W_i^K \in \mathbb{R}^{d \times d_k}, W_i^V \in \mathbb{R}^{d \times d_k}\}_{i=1}^{n_{head}}$ are the learnable parameters of the linear projection for each attention head, and $\sqrt{d_k}$ is the scaling factor. $\otimes$ denotes the operation of dot-production. Given an input $X \in \mathbb{R}^{l \times d}$, the feed-forward module is formulated as follows:

$$FNN(X) = GELU(XW_1^{FNN} + b_1^{FNN})W_2^{FNN} + b_2^{FNN} \tag{15}$$

where $\{W_1^{FNN} \in \mathbb{R}^{d \times d_{FNN}}, b_1^{FNN} \in \mathbb{R}^{d_{FNN}}, W_2^{FNN} \in \mathbb{R}^{d_{FNN} \times d}, b_2^{FNN} \in \mathbb{R}^d\}$ are the learnable parameters of the first and second fully connected layers in this module. d_{FNN} represents the dimension of immediate embeddings in the feed-forward module. $GELU(\cdot)$ is the non-linear activation function. Based on **Equation 13** and **15**, the forward propagation of the transformer-encoder blocks is depicted as follows:

$$H_1 = LayerNorm(MultiHead(H) \oplus H) \tag{16}$$

$$H_2 = LayerNorm(FNN(H_1) \oplus H_1) \tag{17}$$

where $H \in \mathbb{R}^{l \times d}$ represents the output from the previous block ($H = H^0$ for the first block and $H = H_2^{i-1}$ for the i-th block). $\oplus$ denotes the operation of the element-wise addition.

S 3.3: Token predictor

The token predictor consists of two fully connected layers with a GELU activation function and layer normalization for performing masked token prediction in the pre-training phase. Given the output of the last transformer-encoder block H^N, the operation of the token predictor is defined as follows:

$$\hat{T} = LayerNorm(GELU(H^{n_{blocks}}W_1^t + b_1^t))W_2^t + b_2^t \tag{18}$$

where $\{W_1^t \in \mathbb{R}^{d \times d}, b_1^t \in \mathbb{R}^d, W_2^t \in \mathbb{R}^{d \times |v|}, b_2^t \in \mathbb{R}^{|v|}\}$ are the parameters of the predictor. $H^{n_{blocks}} \in \mathbb{R}^{l \times d}$ represents the output of the last transformer block. The predicted scores are denoted as $\hat{T} \in \mathbb{R}^{l \times |v|}$, where $|v|$ is the size of the token vocabulary.

S 3.4: Pooler module

The pooler module is composed of a fully connected layer with a Tanh activation function to transform the token embeddings $H^{n_{blocks}} = \{h_1^{n_{blocks}}, h_2^{n_{blocks}}, ...h_l^{n_{blocks}}\}$ into a sequence embedding $h_{seq} \in \mathbb{R}^{1 \times d}$. The process of obtaining sequence embeddings is formulated as follows:

$$h_{seq} = Tanh(h_{mean}^{n_{blocks}}W^{pool} + b^{pool}) \tag{19}$$

where

$$h_{mean}^{n_{blocks}} = \frac{1}{l-2}Concat(h_2^{n_{blocks}}, h_3^{n_{blocks}}, ..., h_{l-1}^{n_{blocks}}) \tag{20}$$

where $\{W^{pool} \in \mathbb{R}^{d \times d}, b^{pool} \in \mathbb{R}^d\}$ is the parameters of the pooler module. We remove the embeddings of the tokens '[CLS]' and '[SEP]' in generating the sequence embedding.

S 3.5: Affinity predictor

We employ two fully connected layers as a regression predictor to transform the sequence embedding h_{seq} into a scalar value of the predicted affinity $\hat{y}$. The process of this predictor is shown as follows:

$$\hat{y} = (h_{seq}W_1^{reg} + b_1^{reg})W_2^{reg} + b_2^{reg} \tag{21}$$

where $\{W_1^{reg} \in \mathbb{R}^{d \times \lfloor d/2 \rfloor}, b_1^{reg} \in \mathbb{R}^{\lfloor d/2 \rfloor}, W_2^{reg} \in \mathbb{R}^{\lfloor d/2 \rfloor \times 1}, b_2^{reg} \in \mathbb{R}^1\}$ are the parameters of the predictor.

S 3.6: Adapter

The structure of adapters consists of two fully connected layers that are used to learn the binding patterns from TF-DNA binding datasets. The process of adapter is illustrated as follows:

$$H_3 = (((FNN(H_1)W_1^{adapter} + b_1^{adapter})W_2^{adapter} + b_2^{adapter}) \oplus H_2) \tag{22}$$
$$H_4 = LayerNorm(H_3 \oplus H_1) \tag{23}$$

where $\{W_1^{adapter} \in \mathbb{R}^{d \times d/r}, b_1^{adapter} \in \mathbb{R}^{d/r}\}$ and $\{W_2^{adapter} \in \mathbb{R}^{d/r \times d}, b_2^{adapter} \in \mathbb{R}^{d}\}$ are the parameters of the fully connected layers, and r represents the size of the reduction factor ($r = 16$ by default). $FNN(\cdot)$ is defined in **Equation 15**. H_1 and H_2 are the output of **Equation 16** and **17**, respectively. $H_3 \in \mathbb{R}^{l \times d}$ is the output of the adapter. $H_4 \in \mathbb{R}^{l \times d}$ is taken as the input for the next transformer block ($H = H^0$ for the first transformer-encoder block and $H = H_4^{i-1}$ for the i-th block).

S 3.7: AdapterFusion

The AdapterFusion component is based on the attention mechanism, calculating attention scores to weight each adapter in each transformer-encoder block. The forward propagation of the AdapterFusion component is formulated as follows:

$$s_t = Softmax(g_t^T W_{AF}^Q \otimes z_{i,t}^T W_{AF}^K), i \in \{1, 2, ..., n_{adapter}\} \tag{24}$$
$$z_{i,t}^{'} = z_{i,t}^T W_{AF}^V, i \in \{1, 2, ..., n_{adapter}\} \tag{25}$$
$$Z_t^{'} = Concat(z_{1,t}^{'}, z_{2,t}^{'}, ..., z_{n_{adapter},t}^{'}) \tag{26}$$
$$H_5 = Concat(s_1^T Z_1^{'}, s_2^T Z_2^{'}, ..., s_l^T Z_l^{'}) \tag{27}$$
$$H_6 = LayerNorm(H_5 \oplus H_1) \tag{28}$$

where $n_{adapter}$ represents the number of TF-DNA adapters we manage to integrate. $W_{AF}^K \in \mathbb{R}^{d \times d}$, $W_{AF}^V \in \mathbb{R}^{d \times d}$, and $W_{AF}^Q \in \mathbb{R}^{d \times d}$ are the projection layers for Key, Value, and Query in the attention mechanism. $g_t \in \mathbb{R}^d$ is the output of the feed-forward module for t-th token. $z_{i,t} \in \mathbb{R}^d$ represents the output of i-th TF-DNA adapter for t-th token, which is derived from **Equation 22**. $s_t \in \mathbb{R}^{n_{adapter}}$ represents the attention scores of the adapters for t-th token. H_5 is the output of the AdapterFusion component. $H_6 \in \mathbb{R}^{l \times d}$ is taken as the input for the next transformer block ($H = H^0$ for the first transformer-encoder block and $H = H_6^{i-1}$ for the i-th block).

Supplementary Note 4: Token masking for pre-training

The length of each masked contiguous region is fixed and determined by the number of regions. For example, considering a 200-bp sequence with a masked ratio of 15%, if we masked two distinct regions, the length of each masked region would be calculated as $200 \times 15\% \div 2 = 15$.

We assessed the impact of varying numbers of masked regions during the pre-training phase. As shown in **Figure S21**, the optimizer successfully identified the appropriate gradient direction and escaped gradient saddle points (local optimal status) within the interval of 1,000 to 10,000 pre-training steps. This resulted in a significant decrease in pre-training loss and a corresponding increase in pre-training accuracy.

We selected the model parameters that were pre-trained employing the masked contiguous regions of 10 in this study, as it achieved the highest pre-training accuracy of approximately 80% throughout the entire pre-training procedure (the right panel of **Figure S21**).

Supplementary Note 5: Implementation of AdapterFusion for *in vivo* evaluation

The *in vivo* evaluation was implemented on 108 *in vivo* - *in vitro* dataset pairs (consisting of 73 ChIP-seq datasets (denoted as D^{ChIP1}) and 42 HT-SELEX datasets (denoted as D^{SELEX})), involving fine-tuning the pre-trained model on an HT-SELEX dataset and testing on corresponding ChIP-seq datasets with an identical target TF. We fine-tuned the pre-trained model (Θ_0) on D^{SELEX} to obtain 42 TF-DNA adapters $\{\Phi_1, \Phi_2, ..., \Phi_{42}\}$. D^{ChIP1} were utilized for evaluation. We also fine-tuned the pre-trained model on 138 ChIP-seq datasets (denoted as D_2^{ChIP2}) covering 138 unique TFs to obtain *in vivo* binding information for diverse TFs. This resulted in a set of *in vivo* TF-DNA adapters $\{\Phi_1^{'}, \Phi_2^{'}, ..., \Phi_{138}^{'}\}$. Note that the sequences in D^{ChIP1} are not existing in D^{ChIP2}, while some of the datasets in D^{ChIP1} and D^{ChIP2} may correspond to an identical target TF. We implement AdapterFusion for each $\Phi_i \in \{\Phi_1, \Phi_2, ..., \Phi_{42}\}$ to improve *in vivo* prediction as follows:

1. We introduce the AdapterFusion component denoted as Ψ_i in the pre-trained model.
2. We then filter out the *in vivo* TF-DNA adapter $\Phi_j^{'} \in \{\Phi_1^{'}, \Phi_2^{'}, ..., \Phi_{138}^{'}\}$ that corresponds to the target TF same as Φ_i. The remaining 137 adapters are denoted as $\Phi^{'}$.
3. We froze the parameters of Θ_0, $\{\Phi_i, \Phi^{'}\}$, and optimize Ψ_i.

In this evaluation protocol, we do not incorporate the *in vivo* binding information of the target TF into the model as we remove $\Phi_j^{'}$ in step 2. We integrate the *in vitro* binding information of the target TF (Φ_i) and the *in vivo* binding information of other TFs ($\Phi^{'}$) to improve the prediction of the target TF. Note that we apply the classification predictor (**Figure S2D**) rather than the regression predictor in fine-tuning the model on ChIP-seq datasets using binary cross-entropy.

Supplementary Note 6: Motif discovery based on attention scores

S 6.1: Calculation of token attention scores for a sequence by TRAFICA

We employ a modified version of DNABERT-viz [50] to derive binding motifs from high-affinity sequences based on attention scores. The attention scores of the tokens in a sequence can be computed as follows:

1. In each attention head, we calculate an attention matrix through the dot-production of query and key embeddings in the multi-head self-attention module, and discard the elements of corresponding "[CLS]" and "[SEP]" tokens in the attention matrix.
2. We then summarize the attention matrix by rows to obtain the attention vector representing the average attention of each token received from other tokens (including itself).
3. As we apply 12 transformer-encoder blocks and 12 attention heads in each block, we have 12×12 attention matrices for an input sequence. We derive the attention vectors from these attention matrices by conducting the above steps 1 and 2. The resulting 12×12 attention vectors are then collected and averaged into one vector as the representation of token attention scores.

S 6.2: Motif discovery using high-affinity HT-SELEX sequences

We select the top-1000 sequences ranked by affinities from an HT-SELEX dataset for motif discovery. We define high-attention tokens as those with an attention score larger than the mean score (averaging all attention scores) in a sequence. Subsequently, we identify contiguous high-attention tokens with a minimum size $l_{min} = 5$ in the high-affinity sequences as high-attention regions. These regions are treated as motif candidates. It is possible that multiple motif candidates are presented in one sequence. The following steps are conducted to derive motifs from such motif candidates:

1. Given that the majority of motif candidates obtained from the high-affinity HT-SELEX sequences exhibited significant similarity, some of them being identical, we perform pairwise alignment between all pairs of candidates to merge them. The alignment of two candidates is successful if the align score exceeds $\max(l_{min} - 1, 0.5 \times \min(l_{motif1}, l_{motif2}))$, where l_{motif1} and l_{motif2} represent the length of the candidate, respectively.
2. The merged candidates are updated with new positions in sequences. We use the new positions of a candidate to scan its corresponding sequences and extract the sub-sequences with a window size l_{sub} (l_{sub}=8 for 20-bp HT-SELEX sequences; l_{sub}=16 for 30 and 40-bp HT-SELEX sequences).
3. For each merged candidate, We then utilize the fixed-length sub-sequences to construct the position frequency matrix (PFM) that represents a binding motif of the TF.

Supplementary References

[43] Matthew Slattery et al. "Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins". In: *Cell* 147.6 (2011), pp. 1270–1282.

[44] Han Yuan et al. "BindSpace decodes transcription factor binding signals by large-scale sequence embedding". In: *Nature methods* 16.9 (2019), pp. 858–861.

[45] Sandro Barissi et al. "DNAffinity: a machine-learning approach to predict DNA binding affinities of transcription factors". In: *Nucleic Acids Research* 50.16 (2022), pp. 9105–9114.

[46] Aleksandr Morgulis et al. "A fast and symmetric DUST implementation to mask low-complexity DNA sequences". In: *Journal of Computational Biology* 13.5 (2006), pp. 1028–1040.

[47] Todd R Riley et al. "SELEX-seq: a method for characterizing the complete repertoire of binding site preferences for transcription factor complexes". In: *Hox Genes: Methods and Protocols* (2014), pp. 255–278.

[48] Peng Jiang et al. "MPBind: a Meta-motif-based statistical framework and pipeline to Predict Binding potential of SELEX-derived aptamers". In: *Bioinformatics* 30.18 (2014), pp. 2665–2667.

[49] Arttu Jolma et al. "DNA-binding specificities of human transcription factors". In: *Cell* 152.1-2 (2013), pp. 327–339.

[50] Yanrong Ji et al. "DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome". In: *Bioinformatics* 37.15 (2021), pp. 2112–2120.

[51] Jonas Pfeiffer et al. "AdapterFusion: Non-destructive task composition for transfer learning". In: *arXiv preprint arXiv:2005.00247* (2020).

Supplementary Tables

- Supplementary Table S1: Details of the collected ENCODE ATAC-seq experiments
- Supplementary Table S2: Details of the collected HT-SELEX experiments
- Supplementary Table S3: Results of quality control on the HT-SELEX experiments
- Supplementary Table S4: The *in vivo* - *in vitro* evaluation pairs
- Supplementary Table S5: ChIP-seq datasets used for fine-tuning TF-DNA adapters
- Supplementary Table S6: Motif comparisons

These tables are available at `https://github.com/ericcombiolab/TRAFICA`

Supplementary Figures

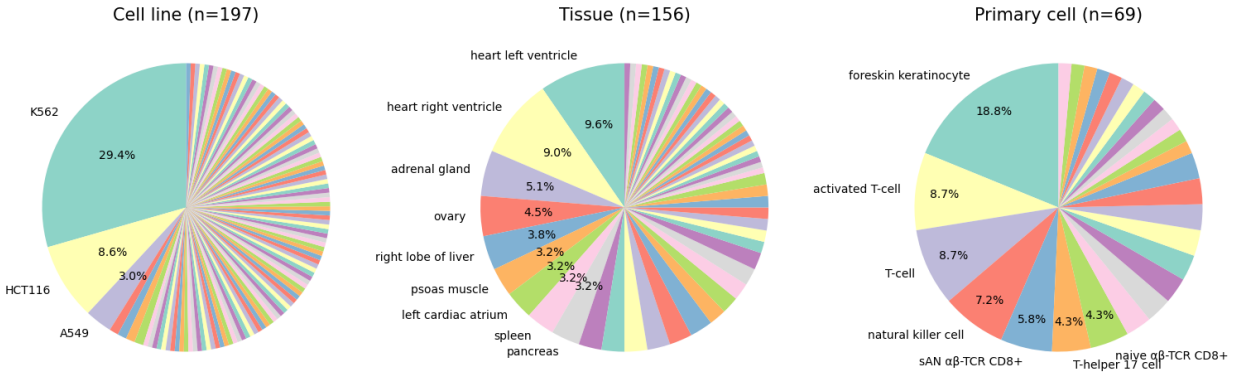

Figure S1. The provenance of ATAC-seq data. A total of 422 human ATAC-seq experimental datasets from the ENCODE project were collected, comprising 197 datasets from distinct cell lines, 156 from various tissues, and 69 from diverse primary cells. The pie plots do not include the biosample types with a proportion of less than 3%. Abbreviations: sAN $\alpha\beta$-TCR CD8+ (stimulated activated naive CD8-positive, alpha-beta T cell), naive $\alpha\beta$-TCR CD8+ (naive thymus-derived CD8-positive, alpha-beta T cell).

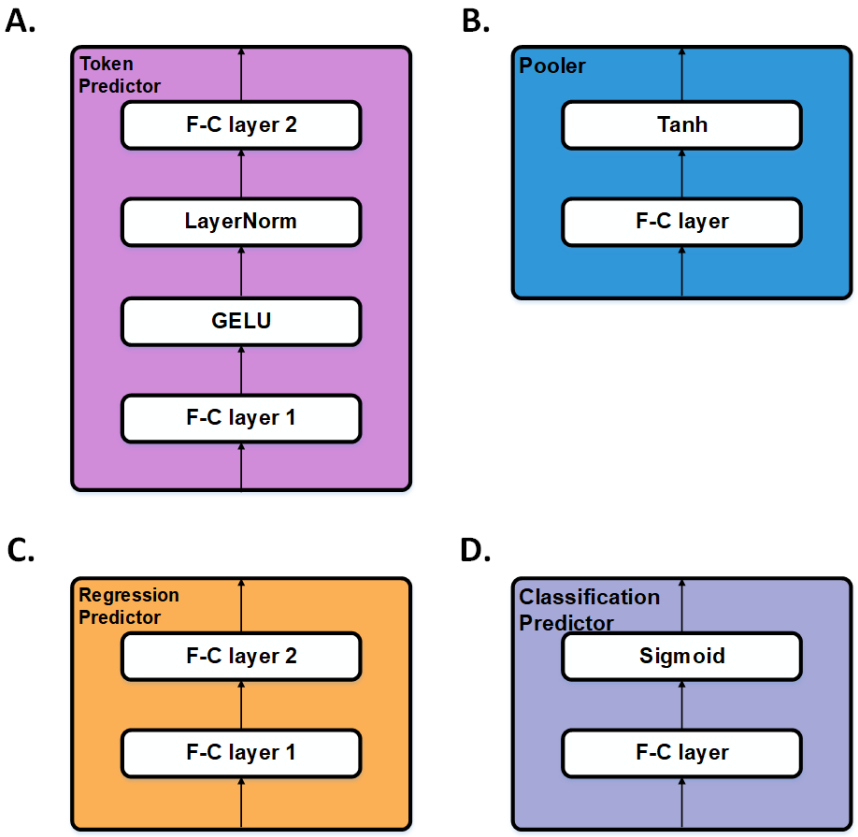

Figure S2. The module schematics of three predictors and the pooler. **(A)** Token predictor consists of 4 sub-components, including two fully connected (F-C) layers, a GELU activation, and LayerNorm. Given an input $h \in \mathbb{R}^d$ and the size $|v|$ of the token vocabulary, the output embedding size of F-C layers 1 and 2 are d and $|v|$, respectively. **(B)** Pooler module comprises an F-C layer for input transformation into the same dimension and a nonlinear activation function Tanh. **(C)** The F-C layers in the regression predictor are for dimension reduction. Given an input $h \in \mathbb{R}^d$, the first layer reduces the dimension of input into $\lfloor d/2 \rfloor$, and the second layer transforms $\lfloor d/2 \rfloor$ into 1. **(D)** Given an input $h \in \mathbb{R}^d$, the F-C layer in the classification predictor converts an input vector as a scalar and then normalizes this scalar in a range from 0 to 1 by the Sigmoid function.

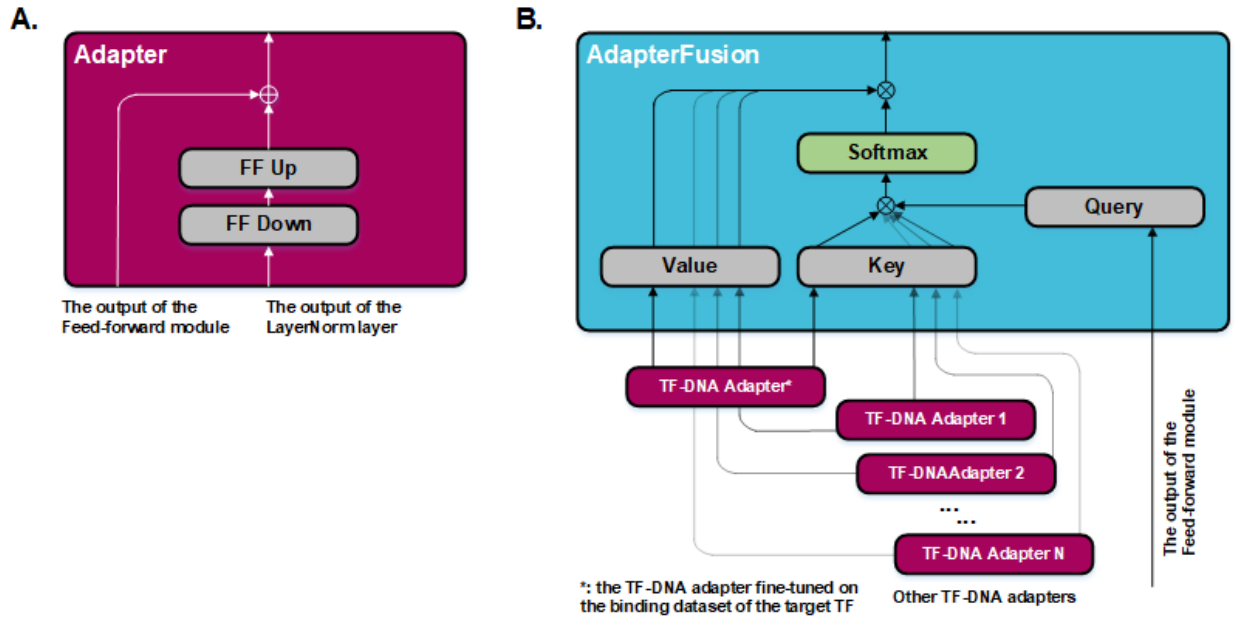

Figure S3. The schematics of Adapter and AdapterFusion. **(A)** Adapter architecture for adapter-tuning. It involves two ordered fully connected layers that serve as the feed-forward downsample (FF Down) and the feed-forward upsample (FF Up) layers. Given a reduction factor r and an input vector $h \in \mathbb{R}^d$, the dimension of the bottleneck vector interposed between two feed-forward layers is d/r. The activation function of ReLU followed by each fully connected layer is hidden for simplification and consistency with the AdapterFusion paper [51]. The symbol of $\oplus$ refers to tensor addition. **(B)** AdapterFusion architecture for integrating TF-DNA adapters. The softmax function is applied to the dot-product of the output of Query and Key projection layers in the attention mechanism, which learns to weight the TF-DNA adapters for mixing the binding information carried by these adapters to improve the affinity prediction of the target TF.

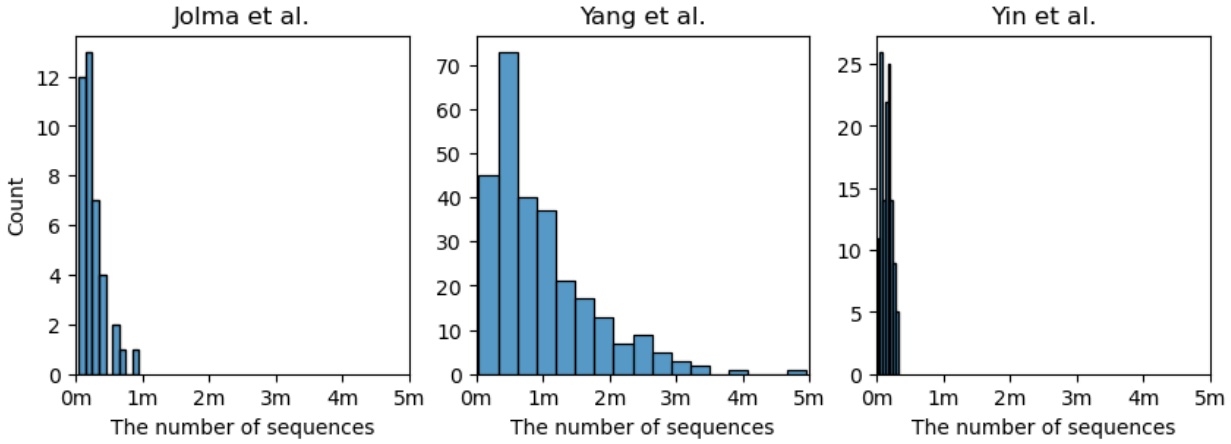

Figure S4. The data size of HT-SELEX datasets from three studies. The range of the x-axis is up to 5 million sequences.

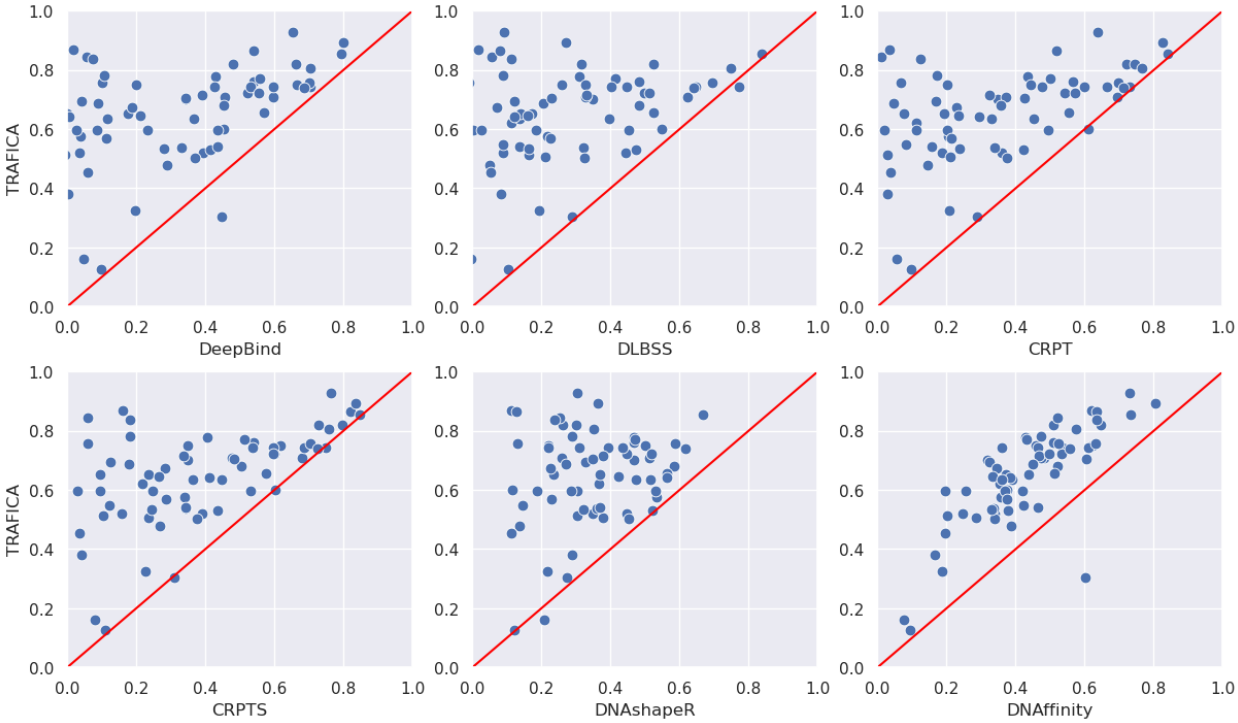

Figure S5. The comparison of TRAFICA predictive performance with each existing tool evaluating on PBM datasets. The evaluation metric is Pearson correlations between experimental data and predicted results.

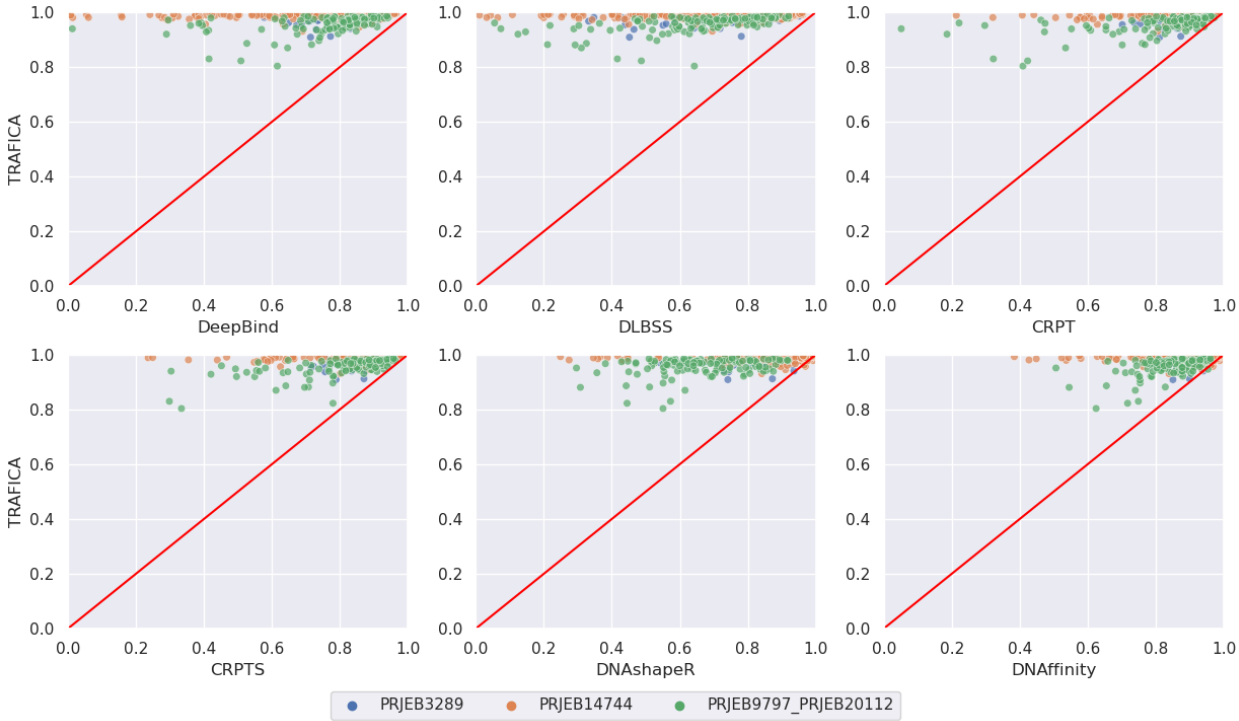

Figure S6. The comparison of TRAFICA predictive performance with each existing tool evaluated on HT-SELEX datasets. The evaluation metric is the coefficient of determination between experimental data and predicted results.

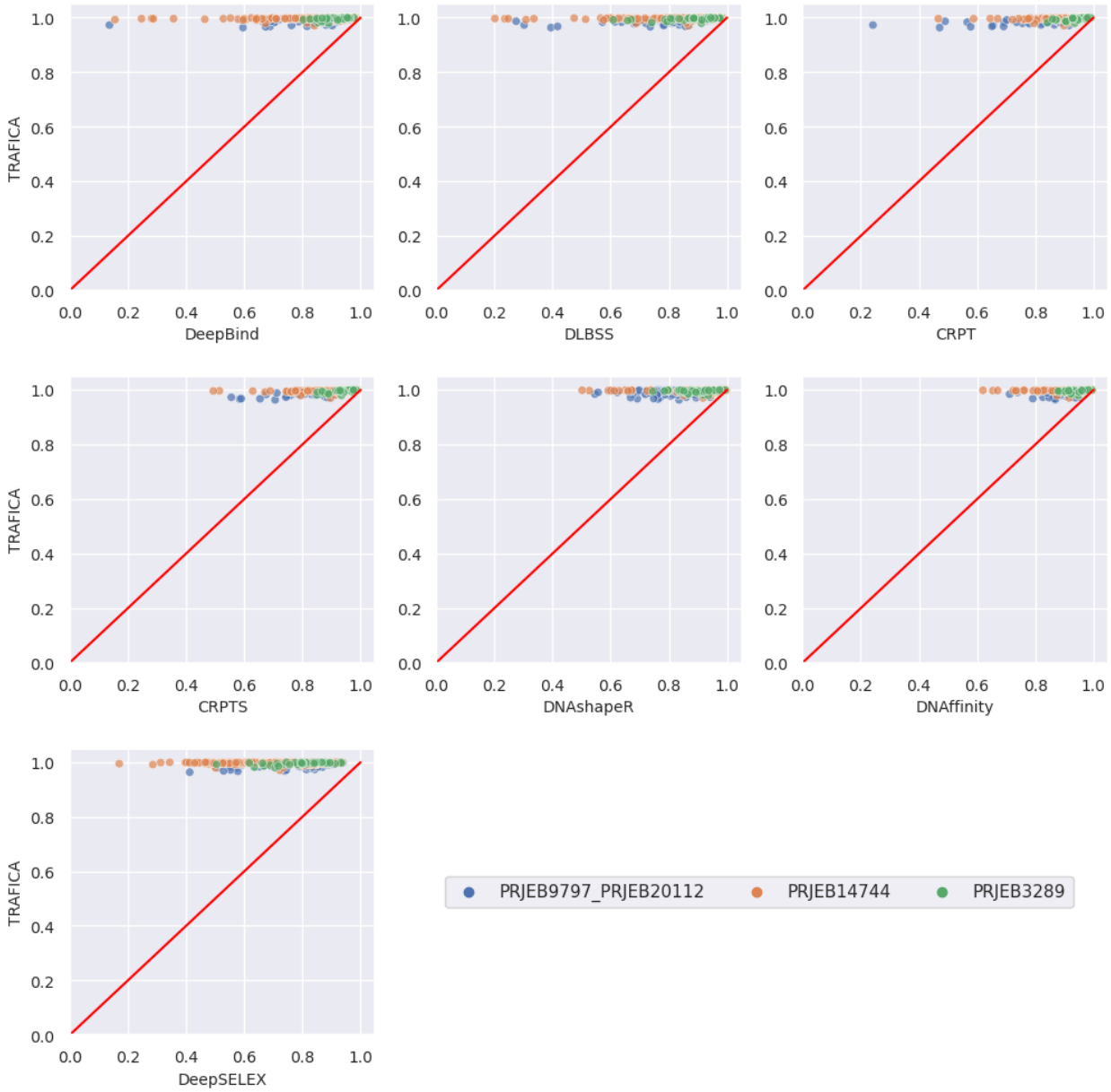

Figure S7. The comparison of TRAFICA predictive performance with each existing tool evaluated on HT-SELEX datasets. The evaluation metric is Pearson correlation coefficients between experimental data and predicted results.

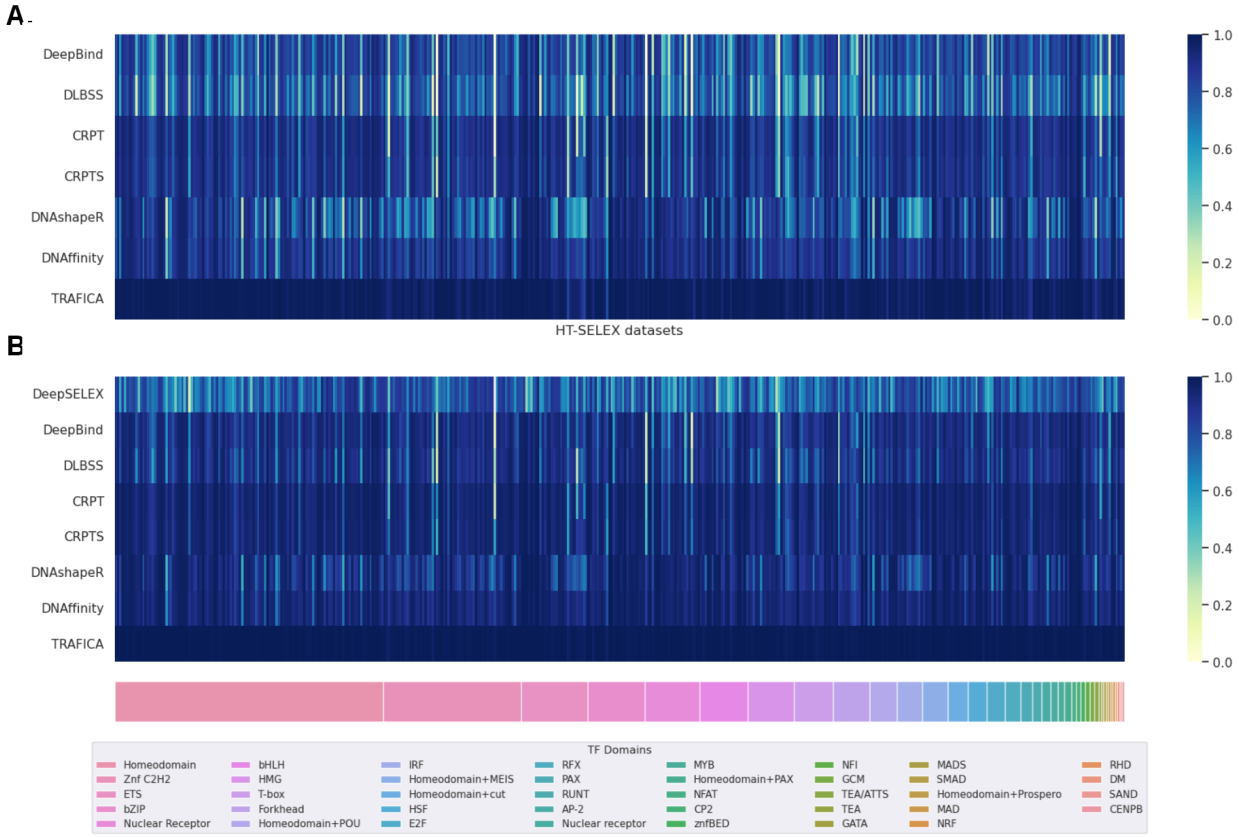

Figure S8. The performance of predicting *in vitro* binding affinities for different TF families on HT-SELEX data. **(A)** Coefficient of determination as the metric. **(B)** Pearson correlation coefficients as the metric.

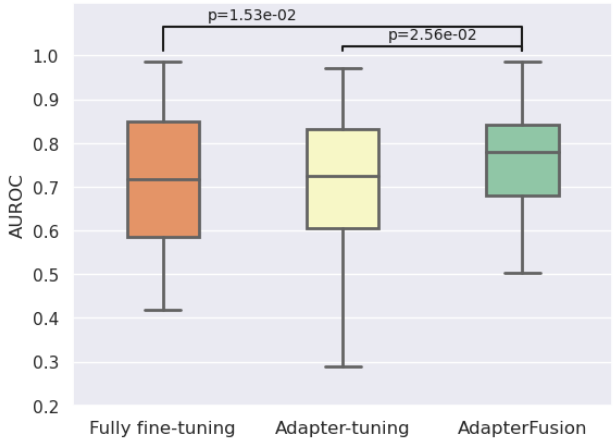

Figure S9. The comparison of *in vivo* affinity predictive performance using different fine-tuning approaches. AdapterFusion significantly outperformed adapter-tuning (without integrating other ChIP-seq adapters) and fully fine-tuning. P-values were calculated by using Wilcoxon rank-sum test.

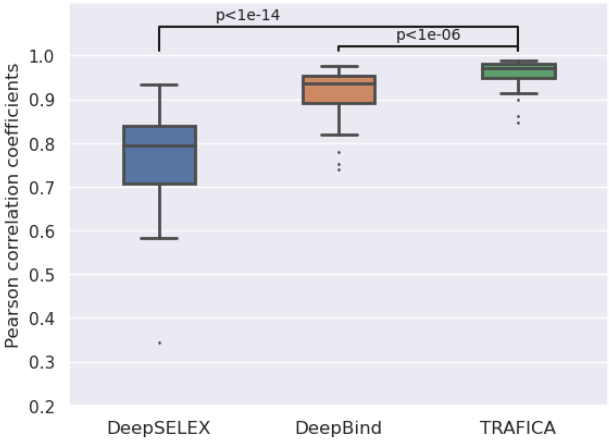

Figure S10. Performance comparison of TRAFICA (implemented with AdapterFusion), DeepSELEX, and DeepBind in predicting *in vitro* binding affinities. P-values were calculated by using Wilcoxon rank-sum test.

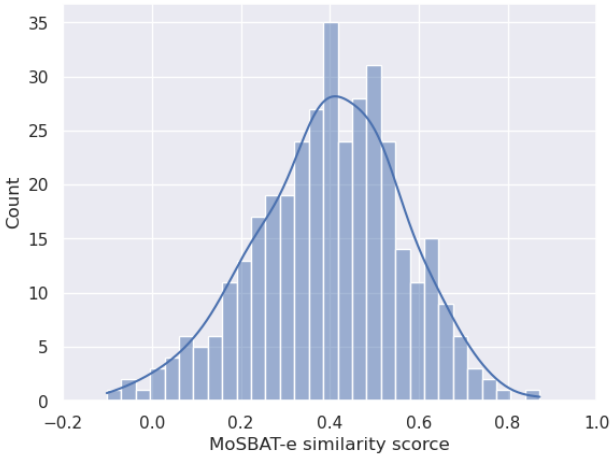

Figure S11. The distribution of MoSBAT-e similarity scores for the motif pairs of JASPAR motifs and the motifs discovered from TRAFICA attention scores. There are 362 motif pairs in total.

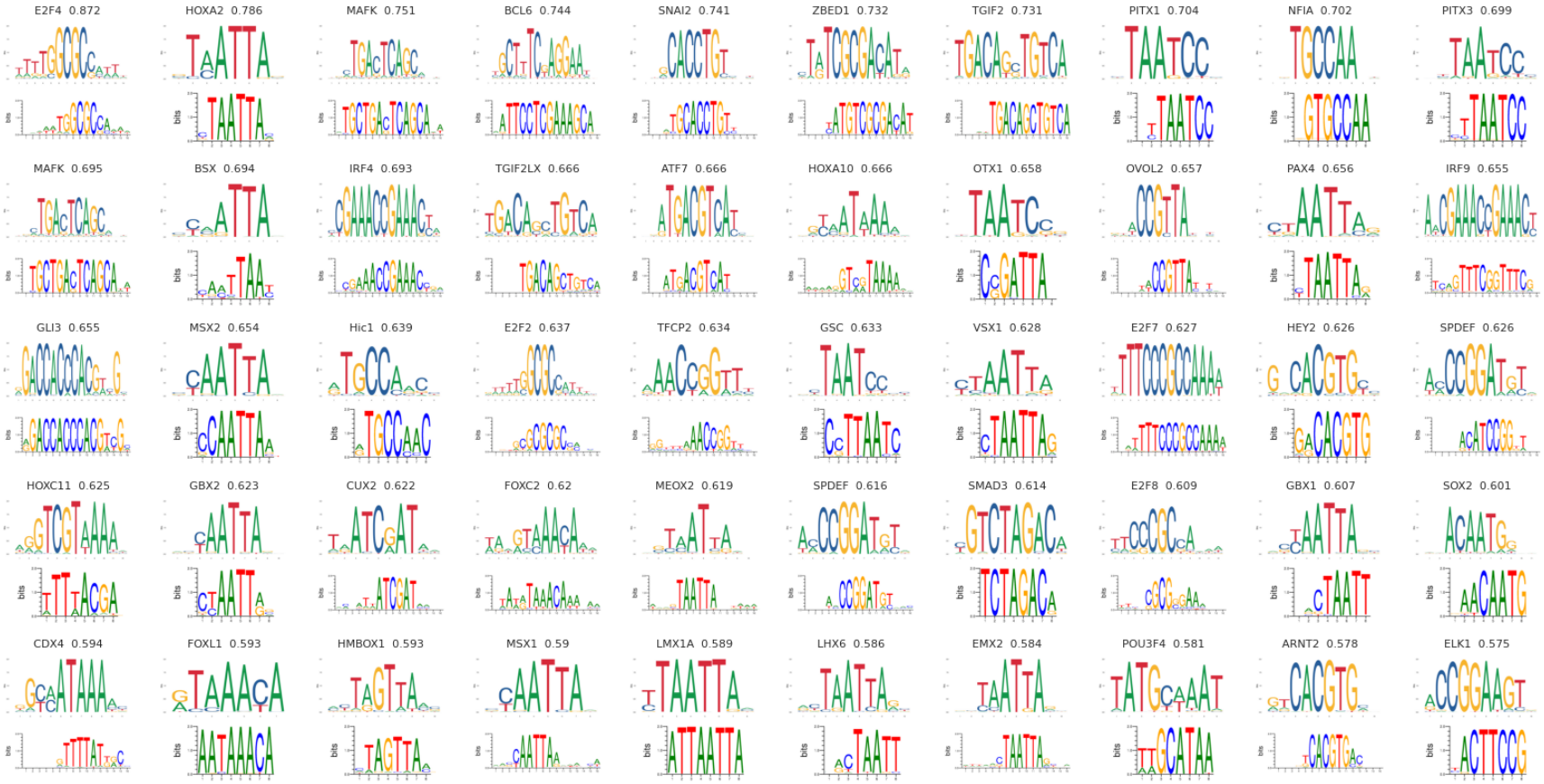

Figure S12. The top-50 motif logos ordered by the MoSBAT-e similarity between the motifs derived from attention scores and JASPAR motifs. The titles of each subfigure (a pair motif logos) indicate TF names and MoSBAT-e similarity scores.

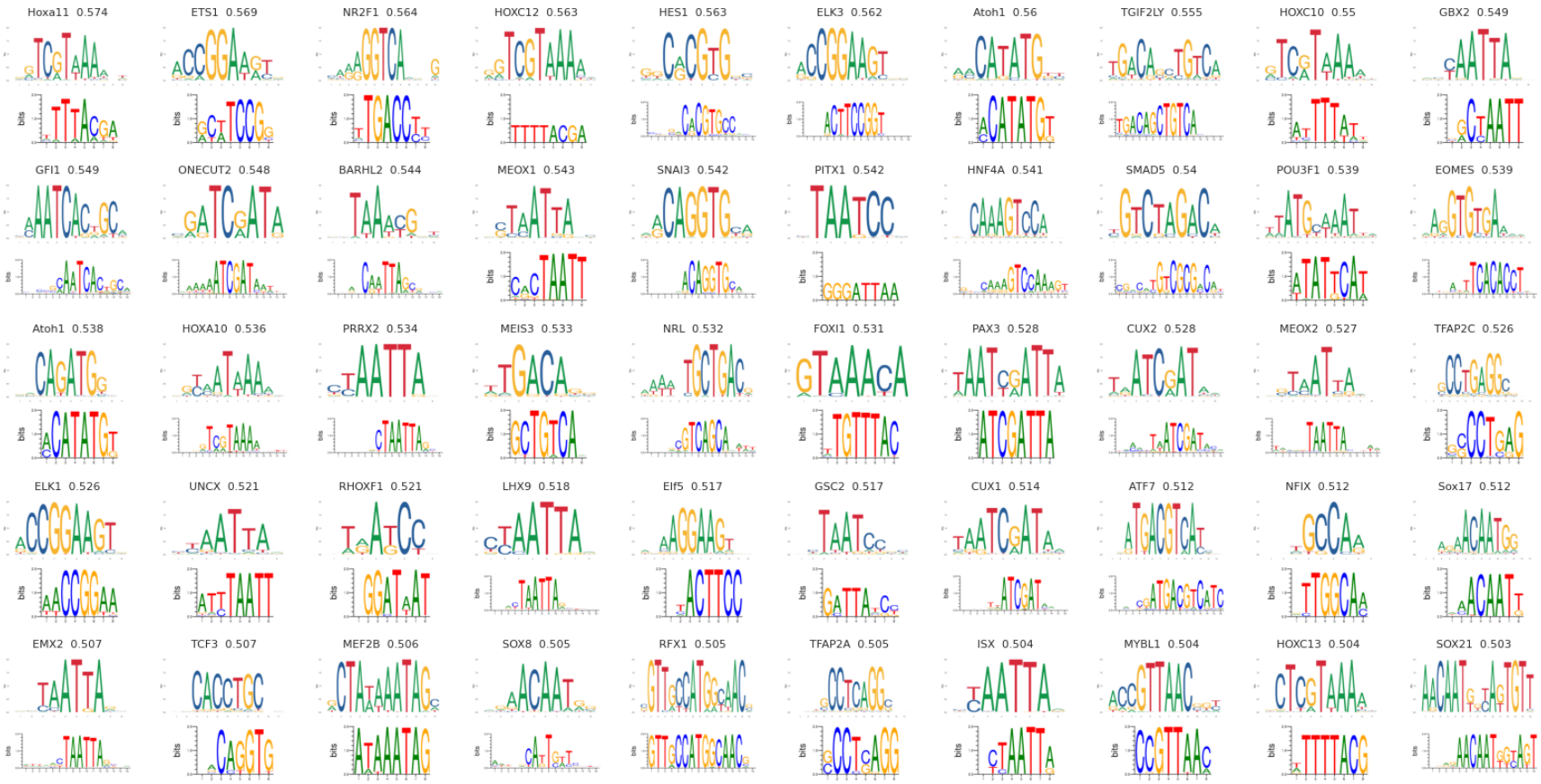

Figure S13. The motif logos, ranked from 51 to 100, are presented in order of the MoSBAT-e similarity between the motifs derived from attention scores and those from JASPAR motifs.

Figure S14. The motif logos, ranked from 101 to 151, are presented in order of the MoSBAT-e similarity between the motifs derived from attention scores and those from JASPAR motifs.

Figure S15. The motif logos, ranked from 151 to 200, are presented in order of the MoSBAT-e similarity between the motifs derived from attention scores and those from JASPAR motifs.

Figure S16. The motif logos, ranked from 201 to 250, are presented in order of the MoSBAT-e similarity between the motifs derived from attention scores and those from JASPAR motifs.

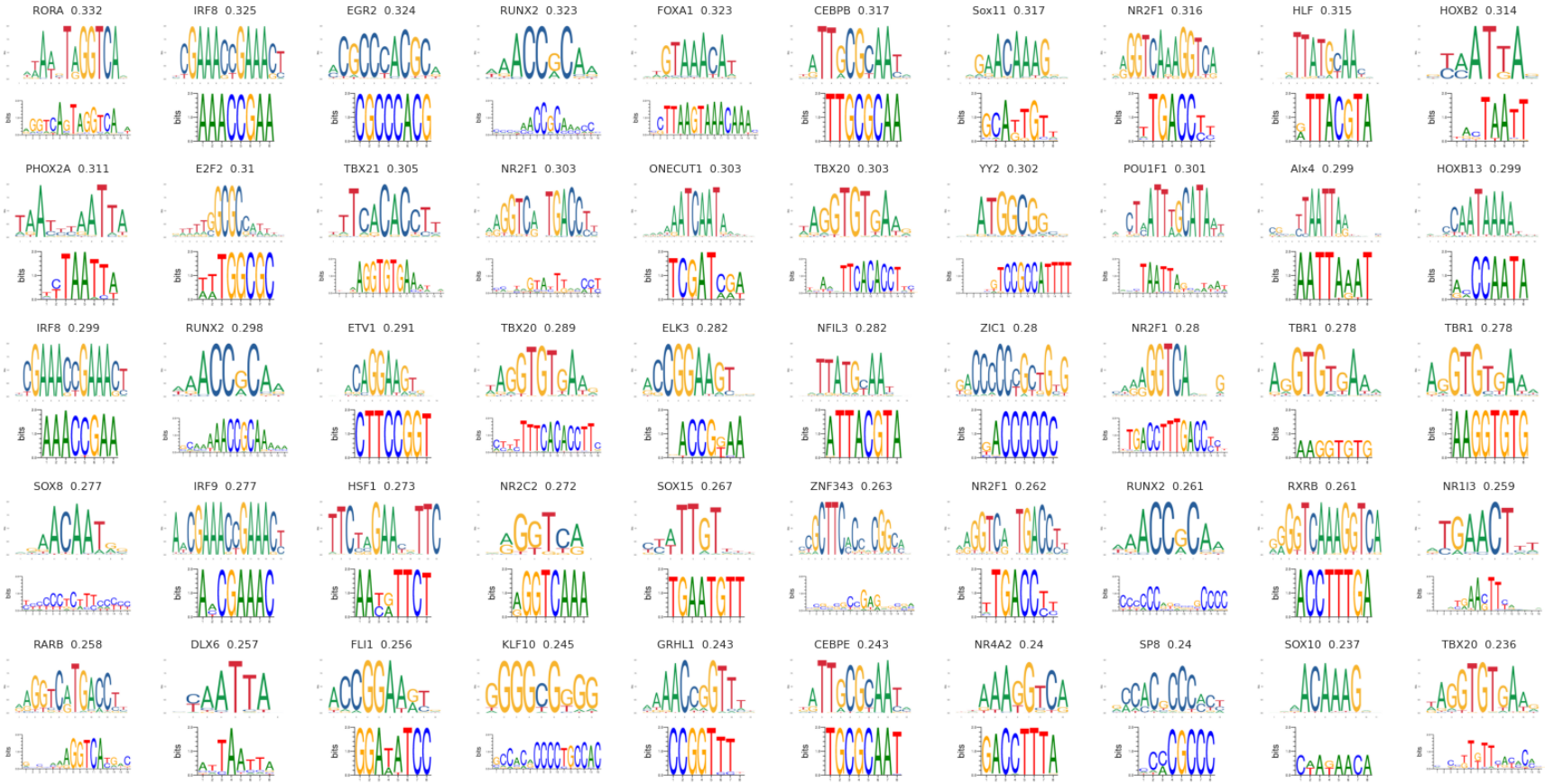

Figure S17. The motif logos, ranked from 251 to 300, are presented in order of the MoSBAT-e similarity between the motifs derived from attention scores and those from JASPAR motifs.

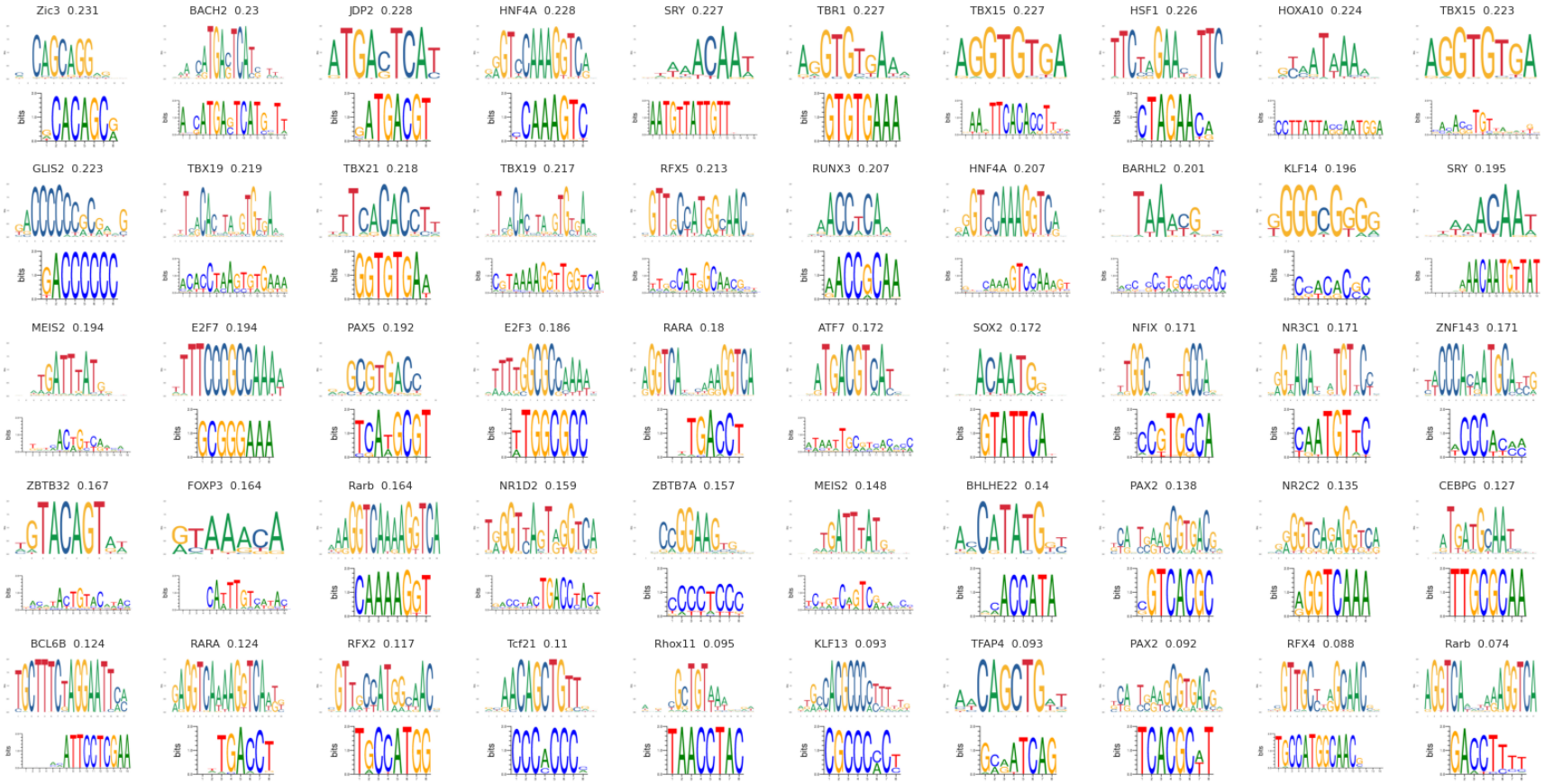

Figure S18. The motif logos, ranked from 301 to 350, are presented in order of the MoSBAT-e similarity between the motifs derived from attention scores and those from JASPAR motifs.

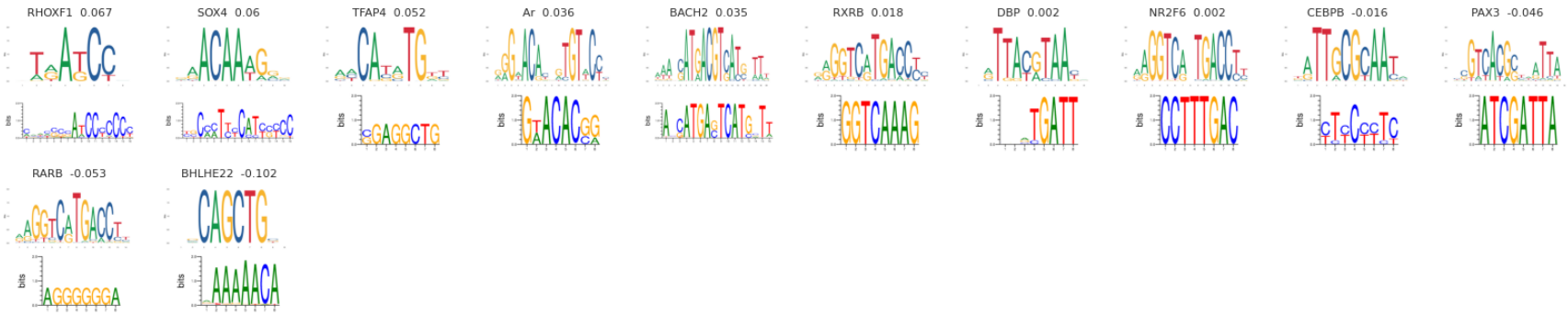

Figure S19. The bottom-12 motif logos ordered by the MoSBAT-e similarity between the motifs derived from attention scores and JASPAR motifs.

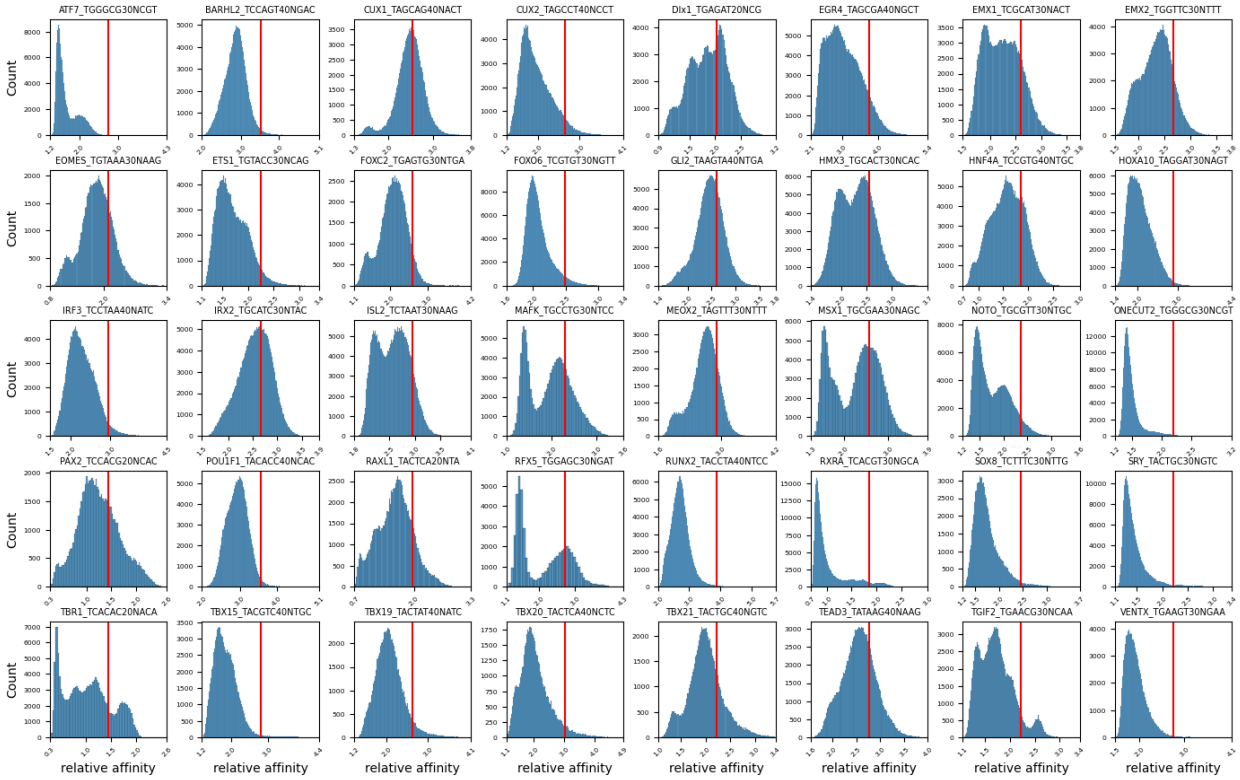

Figure S20. The distribution of estimated binding affinities on the datasets from Jolma et al. [49]. The titles of each subfigure represent the identifiers of HT-SELEX experiments (format: "TF name" + "ligand name"). The x-axis values corresponding to the vertical red lines are half the values of the affinity range in each dataset.

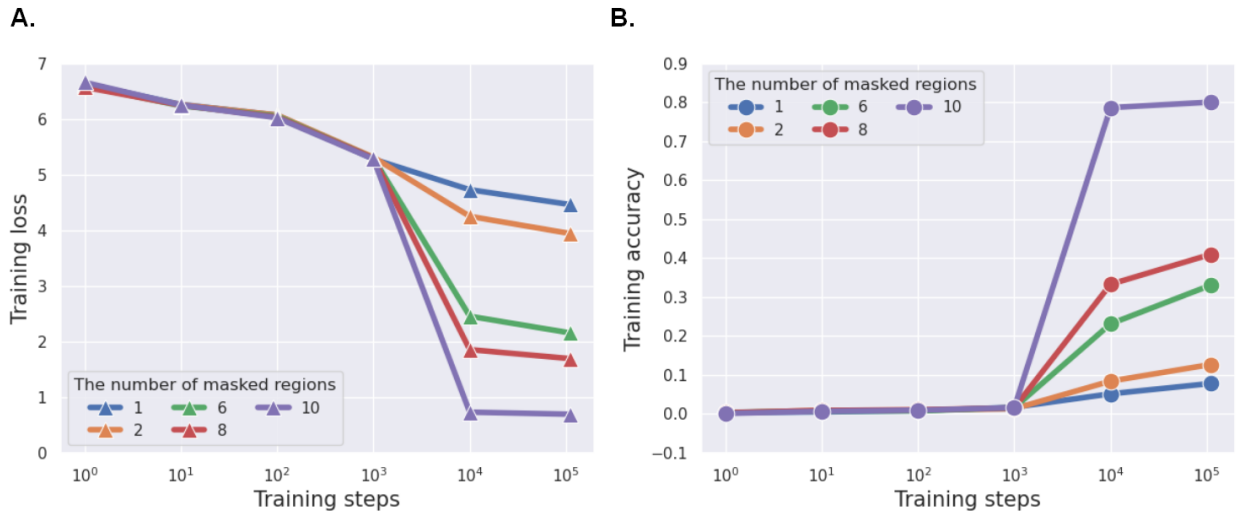

Figure S21. The performance of token prediction in the pre-training phase. **(A)** The prediction loss of masked tokens in each training step. **(B)** The prediction accuracy of masked tokens in each training step.