Deep co-learning on transcription factors and their binding regions attains impeccable universality in plants

Sagar Gupta[1,2], Veerbhan Kesarwani[1,2], Umesh Bhati[1,2], Jyoti[1,2], Ravi Shankar*[1,2]

[1] Studio of Computational Biology & Bioinformatics,

The Himalayan Centre for High-throughput Computational Biology,

(HiCHiCoB, A BIC supported by DBT, India), Biotechnology Division,

CSIR-Institute of Himalayan Bioresource Technology (CSIR-IHBT),

Palampur (HP), 176061, India.

[2] Academy of Scientific and Innovative Research (AcSIR),

Ghaziabad, Uttar Pradesh- 201002

Authors' email addresses:

SG: sagargupta47321@gmail.com

VK: veerbhan8586@gmail.com

UB: bhatiumesh007@gmail.com

Jyoti: dshwljyoti@gmail.com

RS: ravish@ihbt.res.in

*To whom correspondence should be addressed- Email-ravish@ihbt.res.in

Abstract

Unlike animals, variability in transcription factors (TF) and their binding sites (TFBS) across the plants species is a major problem which most of the existing TFBS finding software fail to tackle, rendering them hardly of any use. This limitation has resulted into underdevelopment of plant regulatory research and rampant use of *Arabidopsis* like model species, generating misleading results. Here we report a ground-breaking transformers based deep-learning approach, PTFSpot, which learns from TF structures and their binding sites co-variability to bring a universal TF-DNA interaction model. During a series of extensive bench-marking studies, it not only outperformed the existing software by >30% lead, but also delivered consistently >90% accuracy even for those species and TF families which were never encountered during model building process. PTFSpot makes it possible now to accurately annotate TFBS across novel plant genomes even in the total lack of any TF information.

Introduction

Finding TF binding regions is central to understanding the transcriptional regulation across the genome. They work as cis-acting regulatory platforms where the transcription factors bind and control the expression of the dependent transcribing units. Realizing their importance along with other non-coding regions as the king makers of genome, an ambitious project ENCODE was launched in year 2003 soon after the completion of Human genome project (**ENCODE Project 2004**). Since then, functional annotations of large volume of human genome has been achieved through ENCODE like project, including identification of several TF binding regions and regulatory units. A big credit for such achievement goes to the rise of high-throughput technologies like protein binding microarrays (PBM) (**Berger and Bulyk 2006**), ChIP-Seq (**Johnson et al. 2007**) and its various variants like ChIP-exo (**Rhee and Pugh 2011**), and DAP-seq (**Bartlett et al. 2017**). This resulted into an explosion of data reporting binding regions having length from 36 to 100 bases for various transcription factors (**Nakato and Shirahige 2017**). To this date, there are approximately 1,28,467 ChIP-seq experiments reported at GEO/SRA for human alone. The binding regions identified from these techniques become a guide to define the precise binding sites and motif using various motif finding algorithms.

However, capturing all TF-DNA interactions through ChIP-seq and similar experiments in any organism itself is a costly and impractical affair. There is an infinitely large space of TF-DNA interactions covering various cellular, developmental, and environmental states, limiting the use of such techniques. Also, unlike animals where human has been the main focus, plants have enormous number of species which define an infinitely large search space for possible experiments to detect TF-DNA interactions. Therefore, it becomes essential to develop computational approaches which could model

TF-DNA interactions accurately. For that, one requires a substantial volume of high quality TF-DNA interactions data like the ones from ChIP-seq and related techniques mentioned above to build upon.

However, if one looks towards the status of developments in plants, an extremely sorry state emerges where hardly eight species of plants binding sites are sequenced, covering merely ~700 ChIP/DAP-seq experiments, majority related to *Arabidopsis thaliana and Zea mays*. A clear lack of appreciation for non-coding elements role in plants is evident by the fact that with almost 20 years since ENCODE project, there is no ENCODE like initiative for plants which have more information potential, more variability and numerous regulatory layers than human system, with an ecosystem enormously bigger than humans/animals. The lag of plant science in terms of data generation is not the only issue. This lag is equally reflected in terms of software resources and algorithms development for plant TFBS discovery. While for animal/human several software have been developed, there has been a very limited developments for plants. **Table 1** lists some software available for animals and plants where clear skew is visible (more information available in **Supplementary Table 1)**. Though, plants specific software too have evolved from motif/consensus based pattern search and resources like PLACE (**Higo et al. 1998**), they too have embraced the revolutions in machine and deep-learning. Yet, they still need to come with most revolutionary deep-learning algorithms like Transformers as well as need to approach more relevant biological principles specific to plants.

The existing software are overtly dependent upon the old school of motif discovery and user defined motifs, while recent reports suggest that TF binding is more about context and surroundings (**Sharma et al. 2021; Heikham and Shankar 2010; Zhou et al. 2013**). This step itself is heavily dependent upon binding experiments like PBM and ChIP-seq, on whose results, binding motif is define for any given TF. Many TFs share a similar binding motif but yet differ in their binding due to local surroundings and contexts (**Figure 1a & b**). Factors like DNA shape also play an important role in determining the binding sites for the TFs as shown very recently by some, including couple of software implementing shape information and observing performance improvement (**Mejía-Guerra and Buckler 2019; Ji et al. 2021; Yan et al. 2022; Sielemann et al. 2021**). Further to this, the choice of negative datasets with most of the software has been very relaxed as they randomly pick sequences and give too much weight to the consensus motif which can actually occur even in the non binding regions, creating weak datasets on which learnings have been done so far. Most alarming matter is that plant biologists are overtly using *Arabidopsis* models to report TFBS in other species and attempting to answer transcriptional regulation misdirected, generating potentially misleading results and information. Its root is the lack of studies, data for other species, and lack of reliable computational resources specific to plant systems. It requires an urgent attention of plant biologists to the fact that unlike issues with human and animal systems, the problems with plants is entirely different and more complex. Transcription factors and their binding sites display large degree of variability across the species, more pronounced across plants. Already it has been reported that the assumption of conservation of TFBS is flawed and inflated as high degree of variability exists across species (**Hombach et al. 2016; Hemberg and Kreiman 2011**). Thus, what may be working in *Arabidopsis* may not work in other plant species, and vice versa. Understanding and tackling the variability of plant transcription factors and their TFBS is the main challenge to develop plant specific TFBS models where most of the currently existing software fail miserably. Our group has performed a detailed review and assessment of plant based software which can be found elsewhere (**Jyoti et al. 2023**), while highlighting their issues and ways to improve further.

With this all foundations and understanding of the challenges, we here present an approach and software, PTFSpot, to detect transcription factors binding regions which is based on the following principles:

1) Instead of overtly relying on motifs, identify the most significant motifs specific to any transcription factor and use it as the seed to identify most significant flanking regions for additional information. The same theory we had successfully applied in discovering RNA binding proteins binding sites (**Sharma et al. 2021**).

2) Build upon more realistic datasets instead of general practice of using random datasets and randomized sequences with similar sequence composition. Use the significant motifs found in ChIP-seq and DAP-seq data (positive datasets) to locate them in unbound regions for the given experimental condition. This creates a highly confusing realistic negative dataset where overt importance of motifs is downgraded.

3) Using the motif seeds as anchors, represent the flanking regions with appropriate words of dimers (captures sequence composition as well as base stacking information), pentamers (reflects DNA shape), and heptamers (reflecting sequence motifs). Use them as words of characteristics of the elements which in our case are the various TFs.

4) Apply Transformers like state-of-the-art deep-learning algorithms which learn long distanced as well as local associations among the words, their co-occurrence while learning upon several hidden features, something not possible for traditional machine learning as well as other existing deep-learning methods.

5) Learn co-variability while assuming all transcription factors under one hypothetical transcription factor family which keeps changing its structure (the corresponding transcription factor) according to which the binding site also changes. Correlating the variability among the transcription factor structure and its binding regions from binding data even from a single species may bring a universal model which can work across any species even for never seen before transcription factors, as variability and interaction relationship is learned.

Based on these principles, we have developed a concurrent bi-modal Transformer and DenseNet based Deep-Learning universal system while learning from ChIP-seq and DAP-Seq binding data for 436 TFs, and their corresponding three dimensional structure. A highly extensive bench-marking study was carried out with three different experimentally validated datasets as well as never seen before species specific datasets to test the universality of the developed approach, PTFSpot. The results have been ground-breaking with performance lead of >30% over the existing software pool. Also, in terms of cross-species performance, PTFSpot has delivered an outstanding performance where it consistently scored above 90% accuracy for never encountered before species and transcription factors. This is something which has never been witnessed before. With this all, PTFSpot brings a new dawn in the area of plant regulatory research where one can accurately annotate the TFBS for any novel plant genome even in the total absence of knowledge and information about its transcription factors.

Materials and Methods

Data retrieval

For *Arabidopsis thaliana*, ChIP-seq and DAP-seq peak data were collected for 441 TFs from PlantPAN3.0 (**Chow et al. 2019**) and Plant Cistrome Database (**O'Malley et al. 2016**). While Plant Cistrome Database has 387 TFs covering 44 families with over 568 files of two different origins, i.e., i) DAP-seq and ii) ampDAP-seq, data from PlantPAN3.0 having 54 TFs covering 25 families with over 293 conditions. **Figure 1c** presents the number of collected TFs across 56 different families. All the data were collected in the form of coordinates. The peak data varied from 16 (HBI) to 52,997 (EPR1) peaks. Five TFs data were initially removed because there was not enough peak data availability. Therefore, only TFs with more than 400 distinct binding peaks were considered. HBI (16), TOC1 (191), GATA19 (202), DDF2 (326), and AP2 (329) were the five TFs with the fewest initial peaks. The remaining 436 TFs data contained a total of 5,753,198 distinct peaks. With all this, 5,753,198 peaks, ranging from FHY3 (444) to EPR1 (52,997), were accessible for the study. From the genome, initial coordinate information was extracted into sequences. The source and type of collected peak data are given in **Supplementary Table 2 Sheet 1**.

Identification of motif seed candidates

In the initial phase of the motif search, the peak regions of ChIP-seq and DAP-seq were scanned to identify the most prevalent 6-mer seeds with at least 70% identity. Reasoning behind this was the fact that several previously described TF motifs considered at least six bases (**Chow et al. 2019; O'Malley et al. 2016; Fornes et al. 2020; Jin et al. 2017**) while functional conservation is usually expected within the range of 70% identity. To satisfy a lower bound for such motifs, a k-mer with six bases and at least 70% identity would look for all 6-mers spectra that match each other. To determine the enrichment status of the k-mers (seeds), all possible k-mers were generated for each such sequence which were repeatedly and simultaneously searched through all the TF-bound regions indicated in the targets. Previously, we had introduced a similar motif and seed discovery algorithm to detect RBP-binding sites (**Sharma et al. 2021**)**.** We implemented the same motif and seed discovery algorithm to detect the motif seed candidates for the TFs.

Anchoring with the significant seeds

The peak data was assessed for the shared k-mer seeds and their relatives with at least 70% identity. Further analysis was done for the statistical importance of the seed candidates. For any TF, the best representative seeds were those which had a representation rate of at least 75% across the data. The occurrence probabilities of each k-mer and its relatives were computed using the random dataset to determine their null model distribution. A random dataset was generated from the genome by following peak data length distribution. Binomial test was done for significant enrichment analysis between binding and random data. Null model probabilities were computed using the randomly selected regions from DNA regions not reflected in the ChIP/DAP-seq data. All those k-mers which qualified the binomial test at 1% significance level were considered for the further analysis as the significant ones. These significantly enriched k-mers were the initial seeds for constructing the final motifs. Bidirectionally extension of the seeds was done iteratively by adding one nucleotide on each side at each step, followed by peak data search with at least 70% identity match. The final motifs were chosen based on satisfying of both criteria: the motif candidate had to exhibit at least 75% abundance in the peak data at 1% significance level as well as all matching instances displayed at least 70% identity. Such final motif which was most dominant tagged as the “prime motif”. In addition, a TFs could recognize both the strands of the DNA at a given position. Thus, the reverse complementary of the prime motif was also looked to strengthen the information. Positive and negative datasets for a specific

TF were built simultaneously using the input matrices (position weight matrices (PWM)) for the forward and reverse-complement motifs.

Comparison with experimentally reported motifs

We retrieved 197 TFs from JASPAR[32] that overlapped with our set of 441 TFs. A total of 201 motifs were reported for these 197 TFs. The length of the retrieved motifs ranged from 8-24 bases. To maintain similar motif search approach for fare comparison, these motifs were scanned similarly to those found using the PTFSpot approach. Experimentally verified motifs from Plant Cistrome Database were also considered in addition to the JASPAR motifs. This way a total of 589 motifs were retrieved. Similar searches were also done for these motifs. The relevance of the resemblance between the experimentally reported motif and the corresponding matching motif from PTFSpot was assessed using TOMTOM, a part of MEME suite (**Bailey et al. 2006**).

Datasets creation

The corresponding peak data sequences were transformed into instances of positive data sets once the motifs were anchored for each TF in the provided peaks data. The two ends of the found motifs were extended by -75 and +75 bases in both the directions to produce positive dataset instances for each TF while considering the genomic sequences after mapping the peak data. All sequences in any given sample and corresponding TF had the same length. Due to the motif's repeated occurrences inside a single sequence, the overall number of instances exceeded the peak data for most of the TFs.

Many genomic sites resemble interaction motifs found in the bound cases, but they do not interact in actual. In such scenario, taking random sequences into account does not significantly boost the aim of discrimination. Therefore, one needs to be extremely careful while creating a sound negative dataset.

Considering those genomic regions which display the regions similar to the prime motifs found enriched in ChIP-seq TF binding data but never appeared in ChIP-seq data, become very strong negative dataset candidates. Such negative instances capture the natural scenario of binding preferences where over importance of binding motif is downplayed.

Therefore, we removed all those regions from the genome which were reflected in the peak data, ensuring that the remaining regions never reflected in the peak data and were fit for negative instances selections. Next, they were searched for the prime and reverse complementary motifs of the TF similar to the positive data cases. Similarly, 75 bases flanks were considered while capturing the contextual information with more discrimination power (**Heikham and Shankar 2010**). Positive and negative instances were pooled for each TF in a 1:1 ratio without any overlap.

The dataset where the positive instances were derived from ChIP-seq studies has been called Dataset “A” throughout this study. On the other hand, the dataset whose the positive instances were derived from DAP-seq has been named as Dataset “B”. Additionally, a different dataset containing 387 TFs binding data was formed using data from the Plant Cistrome Database. They too were from DAP-seq studies. This dataset was also utilized by TSPTFBS (**Liu et al. 2021**) and TSPTFBS 2.0 (**Cheng et al. 2023**) for their study. The creation of positive data required the conversion of TFs peak data into FASTA. We used a large pool of 1,89,799 background sequences provided by TSPTFBS, randomly chosen from the intergenic regions of the TAIR10 *Arabidopsis thaliana* genome, to collect the negative data. The ratio of positive to negative data was kept as 1:1. This dataset has been named as Dataset “C” in the current study.

Further, two more datasets “D” and “E” were created from datasets “A” and “B”, respectively for the development of universal and generalized model for TF-DNA interactions and its testing. The only difference between datasets “A” and “D” and datasets “B” and “E”, was that 3D structures of the associated TFs were also considered in the development of these Datasets “D” and “E”. 3D structure of the associated TF were derived with the help of Alphafold2 (**Jumper et al. 2021**). For every positive and negative instance, 3D structure of that associated TF was taken as the input. For Dataset “D”, the total number of TFs was 51. For Dataset “E”, the total number of TFs was 325. Both the datasets contained TFs for *Arabidopsis thaliana*. Dataset “D” was kept aside purely for independent testing purpose only and Dataset “E” was taken into account for model training and testing.

Besides Dataset “D” and Dataset “E”, one more independent test dataset, known as Dataset “F” was also generated. This dataset was constructed to extend the examination of the generalization of the raised model for trans-species TFBS identification. This dataset contained 117 TFs from two species *Zea mays* (93 TF) and *Oryza sativa* (24 TF). This dataset has instance from GSE137972 (217 sample), GSE102920 (six sample), and ChIP-Hub (**Fu et al. 2022**) with over 60 conditions. A total of 2,278,620 ChIP-seq peaks were collected. For the positive binding instances of a given TF, we extended the peak region by mapping to its respective genome. For the negative instances of each TF, we randomly selected the sequences with the same sequence length and numbers as the positive instances after excluding the positive instances information for the considered species. It was ensured that any overlapping regions between negative and positive instances for each of the TFs were removed and finally these non-overlap regions were used as negative dataset instances for each TF. Dataset “F” is also another independent dataset kept aside in entirety and purely for testing purpose alone.

Word representations of sequence data

Each sequence can be considered as a collection of independent words. For instance, the dimeric words AT, TT, TG, GG, GC, CA, and AG represent the sequence ATTGGCAG. An alphabet of the four bases allows creating 16 unique 2-mer words, 1024 unique 5-mer words, and 16,384 unique 7-mer words. Every occurrence of the dinucleotide, pentamer, and heptamer sequences was transformed for the verbal representation of the sequences in an overlapping window. Dinucleotides, pentamers, and heptamers provide information on composition and base stacking, shape, and regional motif information, respectively (**Sharma et al. 2021; Zhou et al. 2013; Černý et al. 2020**). The input vector had a maximum length of 469 items and a sequence length of 159-162 bases. Word representations as dinucleotide, pentameric, and heptameric sequences were independent. Therefore, distinct integers were used as a token for each unique word. The tokenized version of the sequence were converted further into numeric vectors and matrices, known as embedding. The transformer encoder fed the tokenized and embedded sequences as input to train and assess the models. The TensorFlow (Keras) tokenizer class was used to implement the tokenization procedure.

Implementation of the Transformers

A deep learning (DL) strategy was used to construct TFBS discovery models in plants using tokenized and embedded vectors. The transformer-encoder's multi-headed attention mechanism generates reliable contextual correlations. They are placed into hidden space vectors, which the model's output layer categorizes.

The sequences from the previous step's tokenized representation were fed into the input layer of the Transformer system. To create an output matrix formatted as samples, sequence length, and embedding size, embedding converted each word's token into a word vector whose length equals the embedded size. There were "***n***" words in each sequence of length "***l***" (sentence size). The process of embedding words into vectors is represented as following:

Consider a text-based input sentence, **S = {** x_1 **,** x_2 **,** x_3 **……,** x_n **}**, where x_n represents the n^{th} word in the sentence. Each word in a sentence is transformed into a *d-dimensional* vector, where *d* (28) represents the dimension of the sequence element representations in the transformer model, commonly referred to as the model dimension, whose elements (I_d) carry optimized numeric weights, resulting in high level embedding vector, x_n **=[** $I_{n1}, I_{n2} \ldots\ldots, I_{nd}$ **]**.

Accordingly, the text-based sentence can be represented as an embedded words matrix, M ϵ $\mathbb{R}^{n \times d}$,

$$M = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \ldots \\ x_n \end{bmatrix} \begin{bmatrix} I_{11}, I_{12}, \ldots, I_{1d} \\ I_{21}, I_{22}, \ldots, I_{2d} \\ I_{31}, I_{32}, \ldots, I_{3d} \\ \ldots\ldots\ldots\ldots\ldots \\ I_{n1}, I_{n2}, \ldots, I_{nd} \end{bmatrix}$$

where each row corresponds to a word in "**S**". This matrix has dimensions of *n* x *d*, indicating the number of words in the sentence by the dimension used to represent each word in the embedding vector.

Each word in the matrix "***M***" is also combined with its corresponding positional encoding, "***P***". The positional encoding "***P***" shares the same dimension "***d***" as the word embedding vector, x_n , resulting in M' = M + P, where $\boldsymbol{P}$ = [$p_1, p_2, \ldots, p_d$]. The calculation of sinusoid positional encoding "***P***" are determined using the following equations:

$$P_{(i,2m)} = \sin\left(i/10000^{2m/d_{model}}\right)$$

$$P_{(i,2m+1)} = \cos\left(i/10000^{2m/d_{model}}\right)$$

where, "***i***" denotes the position of token and "***m***" stands for the encoding dimension.

Position encoding generates a matrix with a similar shape to be added to the embedding matrix. The shape of the matrix (samples, sequence length, embedding size) remains consistent throughout the Transformer and is eventually reshaped by the final output layers. The input embedding layer feeds its outputs into the subsequent layer, while the output embedding layer does the same for the next layer (encoder layer).

The encoder processes the input through a multi-head attention layer. The multi-head attention applied to M' ϵ $\mathbb{R}^d$ is computed in the following manner. The MultiHead result is obtained by concatenating h individual attention heads ($head_1$, $head_2$, . . . , $head_h$) and multiplying them by a weight matrix W^O .

$$MultiHead = Concat\ (head_1 , head_2 , \ldots , head_h)\ W^O$$

Each attention head ($head_i$) calculates the attention mechanism with input matrices Q, K, and V. These input matrices are derived from "***M'***" using optimizable weight matrices W^Q , W^K , and W^V .

$$head_i = Attention\ (Q, K, V)$$

where, Q = M'. W^Q ,

$$K = M'. W^{K} ,$$

$$V = M'. W^{V}$$

The attention operation itself is defined by;

$$\text{Attention} (Q, K, V) = \text{Softmax} \left(\frac{QK^{T}}{\sqrt{d_k}} \right) V$$

This entire process is applied to input matrices Q, K, and V. Finally, the h attention head outputs are concatenated and further transformed using an output weight matrix W^{O} .

The same steps were repeated and their individual attention vectors were concatenated and passed to the transformer-encoder's input feed-forward network block for additional processing. To lessen overfitting, the output of the multi-head attention layer was supplied to a dropout layer, followed by a normalization layer. After normalization, the output goes into two fully connected feed-forward layers which converges to the subsequent dropout layer, followed by the normalization layer. A GlobalAveragePooling1D layer was then implemented, which feeds its results to the third dropout layer of the model after the previous layer's input was collected via two fully connected feed-forward layer. Various activation functions from the available activation functions were examined for the layers. The final hidden layer passes its output to fourth dropout layer. A one-node classification layer using the sigmoid activation function with binary cross-entropy loss was implemented. The weights and learning rates were adjusted in this step with the help of the "Adam" optimizer. **Figure 2** depicts the operation of the implemented transformers system.

Optimization of the Transformer system

The transformer system was optimized using Bayesian optimization. The transformer encoders had a multi-head attention layer, where 14 self-attention heads were found to work best. The input sequence was padded if the length was shorter or longer than 160 bases to ensure a constant size of the input matrix. The output from the multi-head attention layer was passed into the dropout layer with a dropout fraction of 0.15.

Later, this result was normalized by another layer called the normalization layer. This layer was followed by a third layer called Feed Forward layer with 14 nodes, followed by another dropout layer with a dropout fraction of 0.15. A feed forward layer with 14 nodes, followed by another normalization layer, followed the above one. The result of the normalization layer was passed to the GlobalAveragePooling1D layer, followed by a third dropout layer with a dropout fraction of 0.25. Next to this, the pooled feature maps were passed to fully connected layers. The performance of the transformer was evaluated for several hidden layers where two hidden layers were found performing the best. The number of nodes in two hidden layers was set for the model, and different activation functions for the layers were explored among the available activation functions. Finally, two dense layers were configured, each with 40 and 16 hidden nodes with SELU and RELU activation functions, respectively. The last hidden layer was connected to the fourth dropout layer of the stack with a dropout fraction of 0.25, which finally passed its results to a single-node result classification layer based on a sigmoid activation function.

A binary cross-entropy loss function was used to calculate the loss and "Adam" optimizer was used to adjust the weights and learning rates. The learning rate of the optimizer was set to 0.001, and the model was trained using 25 epochs and batch size of 16. Since the problem in this study was not translation but classification, decoders were not needed. The number of encoders and performance was investigated, and it was found that increasing the encoder layers did not show a significantly large change in efficiency and only decreased after the sixth encoder layer. Thus, we went with one encoder layer to keep it lighter with the same performance.

The output of the transformers returns probability scores for each input sequence indicating the confidence of each case as the presence or absence of TFBS in the given sequence. The final hyperparameters of the output layer of the optimized model were: {"Activation function": sigmoid, "Loss function": binary cross-entropy, "Optimizer": Adam}. Related information from the optimization to the final model is listed in **Supplementary Table 2 Sheet 2** and illustrated in **Supplementary Figure 1**. The resulting final model was saved in Hierarchical Data Format 5 (HDF5). Since the whole system is implemented in TensorFlow, the HDF5 format provided the model graph definition and weights to the TensorFlow structure and saved the model for classification. **Figure 2** shows the detailed workflow and the implemented transformer architecture part in it.

Performance evaluation for the transformer part

According to standard practice, the datasets “A” and “B” were divided into train (70%) and test sets (30%). The developed model was tested on the 30% intact and completely untouched test portion. Four categories of performance confusion matrix namely true positives (TP), false negatives (FN), false positives (FP) and true negatives (TN) were evaluated. The performance of the built transformer models was evaluated using performance metrics such as sensitivity, specificity, accuracy, F1 score, and Mathews correlation coefficient (MCC) (**Chicco and Jurman, 2020**).

Performance measures were done using the following equations:

$$\text{Specificity}(\text{Sp}) = \frac{\text{TN}}{(\text{TN+FP})}$$

$$\text{Sensitivity}(\text{Sn}) = \frac{\text{TP}}{(\text{TP+FN})}$$

$$\text{Acc=}\frac{\text{TN+TP}}{(\text{TN+TP+FN+FP})}$$

$$F\,1-\text{Score=2}\times\left(\frac{\text{Precision}\times\text{Recall}}{\text{Precision+Recall}}\right)$$

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

$$\text{AUC=}\int_0^1 \text{Pr}[\text{TP}](v)\text{dv}$$

Where:

TP = True Positives, TN = True Negatives, FP = False Positives, FN = False Negatives, Acc = Accuracy, AUC = Area Under Curve.

In addition, 10 times independent random training and testing trials were used to assess the consistency of the model and its performance. Each time, the dataset was randomly split in the ratio of 90:10, with the first part used for training and the second one for testing. Each time a new model was built from the scratch. For each of these models, the above mentioned performance metrics were calculated. In addition, it was made sure that there was no overlap between the train and test sets to prevent any bias, and memory. This care has been taken for all the datasets taken in the present study.

Comparative benchmarking

The developed approach, PTFSpot, was comparatively benchmarked with seven different tools that reflect most advanced current approaches for TFBS detection. The compared tools were k-mer grammar (bag-of-*k*-mers and logistic regression), Wimtrap (XGBoost), SeqConv (Convolution Neural Nets (CNN)), TSPTFBS (CNN), TSPTFBS 2.0 (DenseNet), PlantBind (composite CNN, Long Short Term Memory (LSTM), and Attention mechanism) and DNABERT (Bidirectional Encoder Representations from Transformers (BERT)). Additionally, the benchmarking took into account two distinct datasets (Dataset "B", and "C") to conduct a completely unbiased evaluation of the performance of these tools. The reasoning to build Dataset “C” was to use the same set of data which was used by the compared tools to report their performance (**Shen et al. 2021b; Rivière et al. 2022**). Details on the datasets are already covered above.

Structure and binding motif analysis across species

We further studied the counterion binding of TFs to the target DNA sequences across the species. We downloaded the available common TFs (*Arabidopsis thaliana* and *Zea mays*) peak data from PCBase and Plant Cistrome Database. The 3D structure of the TFs was modeled using AlphaFold2 (**Jumper et al. 2021**). AlphaFold2 generated five models for each TF, but we selected the top-ranked model for our subsequent analyses.

ScanProsite (ExPASy) (http://prosite.expas.org) was used to confirm each modeled protein's functional domain and the number and name of amino acid residues found in the pocket of the active site. PTFSpot was applied to identify motifs for all these TFs across the species. Afterwards, we did a comparative investigation into 3D protein structure with their corresponding motif for all the common TFs taken. Comparative study of each TFs was done on the following bases: 1) based on the protein sequence, 2) based on the functional domain, 3) based on their respective 3D structure, and 4) the binding affinity of TF across the species.

Validation of prime motif and transcription factor interactions using molecular docking and simulation

For this study, we randomly selected four TFs (AT566940, CEJ1, WRKY75 and VRN1) and their identified prime binding sites. The prime binding sites identified through PTFSpot algorithm for above-mentioned four TFs were taken as flexible molecules. Each of them were studied for binding with the identified prime motif region and other regions around. Four more regions were taken for each TF. These regions were selected from sequences to evaluate the impact of flanking regions around and the prime motif region. In doing so, for each such region was removed and binding analysis was performed. Region ‘A’, represented the prime motif region, whereas region ‘B’ and ‘D’, were overlapping regions with the prime motif in both the directions. Region ‘C’ and ‘E’ were taken from the start and end of the sequence for every considered TFBS, respectively (**Figure 3c & 3d**). The pyDock server (**Jiménez-García et al. 2013**) was used to calculate the binding affinity and selectivity for these regions (A, B, C, D, E) with its TF protein. From each docked complex (with regions A, B, C, D, and E) with the lowest docking energy, the optimal complex structure was chosen for each. The strength of the binding region and target interaction was determined by the structural qualities of the target. The GROMACS package (**Berendsen et al. 1995**) was used to run the molecular dynamics simulation (MD) for a protein-DNA complex structure utilizing the amber99sb force field. Each complex was solvated in a cubic box with 1.2 nm distance between the complex and the solvated box edges. To neutralize the system's charge, sodium (Na^+) and chloride (Cl^-) ions were added, and the energy was minimized using the steepest descent algorithm. A 2fs time step was used in the simulation of each complex structure. All simulations of the complexes were undertaken for a 50 ns to verify the robustness of the results. Chimera (**Pettersen et al. 2004**), PyMOL (**L, D.W. 2002**) and Xmgrace software were used to analyze the results.

Transformer-DenseNet system for cross-species identification of the binding regions

Initially, only Transformer based model was implemented to identify TF binding regions for every individual TF. Any such model fails to perform well when run across other species for the same TF family. This happens due to variation in the TF structure and its binding regions with species. As discussed above, protein 3D structure plays an important role in determining binding motifs. We incorporated TF's structure information for every corresponding binding region from ChIP-seq/DAP-seq data. 3D protein structure based DenseNet (**Huang et al. 2017**) and the above developed Transformer model which takes sequences as input, were intertwined through a bi-modal architecture implementation to raise a universal model to identify TF-DNA interactions in global and trans-species manner.

The DenseNet part of PTFSpot takes atom-wise X, Y, Z coordinates of the 3D structure of a protein as the input. We normalized the coordinates by dividing it by the highest absolute value noted for the corresponding axis. This normalization ensured transformed and uniform values in range of -1 to 1. This coordinates based matrix was the input with 300×24×3 dimensions to a convolution layer. If the length of some protein was found shorter it was padded for the empty columns with value of zero. Size of 300 covers the amino acid positions in a window, for each of which corresponding 24 atoms exist with every atom mapped to three-dimensional space in terms of normalized X, Y, Z coordinates. Those amino acids which have less than 24 atoms in their conformation were padded for the empty columns with value of zero. The 24 elements vector size to represents the amino acids was considered after an analysis of more than 400 TFs where we found that the TF amino acids had maximum of 24 atoms in it. Thus, each vector position represented a particular atom number.

Building the DenseNet architecture

Input Layer: The input to the DenseNet is the 3D tensor representing the protein structure. The dimensions was 300×24×3 (300 amino acids, 24 atoms per amino acid, three X, Y, Z space coordinates for each atom).

DenseNet: We opted for DenseNet in our study due to its inherent advantages in mitigating the vanishing-gradient issue, enhancing the propagation of features, promoting feature reuse, and significantly minimizing the parameter count. Each layer receives a “collective knowledge” from all the preceding layers. Since each layer receives feature maps from all the preceding layers, this leads to higher computational and memory efficiency. The architecture of the DenseNet model employed in our study is composed of one convolution layer with 32 convolution filters (kernel size = 3), one batch normalization and 2D maxpooling pooling layer (stride = 2), ten dense blocks and nine transition layers (**Figure 2**). A total of 121 layers were implemented in DenseNet.

Within each dense block, ' m_0 ' represents the input to the convolutional network, consisting of ' l ' layers. Consequently, the ‘ l^{th} ' layer receives its input from the feature maps of all previous layers, which is represented as:

$$m_l = H_l\left(\left[m_0, m_1, \ldots\ldots, m_{l-1}\right]\right)$$

Here, $\left[m_0, m_1, \ldots\ldots, m_{l-1}\right]$ denotes the concatenation of feature maps generated in layers 0 to $l-1$. H_l (.) represents a non-linear transformation, which can be described as a composite operation consisting of batch normalization (BN), followed by a rectified linear unit (ReLU) function, and a 3x3 convolution (kernel size = 3). To facilitate down-sampling, the network architecture incorporates transition layers between each dense block. These transition layers include batch normalization, followed by a 3x3 convolution layer, maxpooling2D, batch normalization, 3x3 convolution layer, and conclude with a dropout layer. The growth rate, denoted by the hyperparameter '***k***' in DenseNet, plays a pivotal role in the architecture's remarkable performance. DenseNet can achieve impressive results even with a smaller growth rate, primarily due to its unique design that treats feature maps as a global state of the network. In each layer, '***k***' feature maps are contributed to this global state, and the total number of input feature maps (denoted as F_l) at the '***l-th***' layer is calculated as:

$$F_l = k_0 + k * (l - 1)$$

The term ' k_0 ' represents the number of channels in the input layer. After the final transition layer, the initial input, which was of shape 300x24x3, had been down-sampled to a final size of 1x1x64. Subsequently, the output of DenseNet was flattened and concatenated with that of transformers for the final classification steps. In the second part of the bi-modal architecture transformers were implemented with sequence as input constructed as described in above section on transformers. Afterwards, output passes on the GlobalAveragePooling1D layer which it passes to dropout and then to one hidden layer. From both parts (DenseNet and Transformers), outputs from the previous layer were concatenated and passed onto the batch normalization, dropout, dense, and dropout layers sequentially. A sigmoid activation function based single node classification layer was used with binary cross entropy loss function to calculate the loss. “Adam” optimizer was used at this point to adjust the weights and learning rates. The batch size was set to 64 and the number of epochs was set to 6. Further details of this module is provided in the **Figure 2**.

To train and test this bi-modal Transformer-DenseNet system, two different datasets were created from (Datasets “D” and “E” derived from the datasets “A” and “B”, respectively). Details are already covered in the datasets section above. Considering the TF protein structure and the corresponding binding region together while co-learning on them captures the relationship between protein structures and binding sites and variations in them. As mentioned in the methods section, Dataset “D” was kept for independent testing purpose only and Dataset “E” was taken into account for model training and testing. Data were split in the ratio of 70:30 on which the bi-modal system was trained and tested. Besides Dataset “D” and “E”, one more independent test dataset, known as Dataset “F” was also generated. This dataset contained 117 TFs from two species *Zea mays* (93 TF) and *Oryza sativa* (24 TF). Dataset “F” is considered as the independent test dataset to evaluate the performance of the developed Transformer-DenseNet based bi-modal deep learner for trans-species TF binding regions identification in plants. Details on all datasets is already covered in the datasets related section in the method sections.

Results and discussion

Most of the TF binding regions display a prime binding motif

It was observed that each of the TFs considered in this study had at least one such prime motif which covered at least 75% of the binding data. The prime motifs ranged from 9 to 12 bases. The consensus motif TC[CA]AAGCT[GT] (for SVP) was the least abundant among the prime motifs covering 75.96% of the bound data. Similarly, the consensus motif [CA][GT]G[GA]ATCC[CT]TT was the most abundant prime motif covering 97.22% of the bound data (for FHY1) in the ChIP-seq derived positive part of the Dataset “A”. Similarly, in the DAP-seq derived positive instances of Dataset “B”, the consensus motif [GA]GT[CT][GA]AC[CT][TA] was the least abundant among the prime motifs, covering 75.06% of WRKY24’s bound data. While, the consensus motif [CT][TA]CACGTG[TC] was the most abundant one with 98.3% data coverage for At1g78700 across its bound regions (**Supplementary Table 2 Sheet 3-4**).

The motif associations were intriguing, as distinct clusters emerged out of similarity among the identified prime motifs across different TFs. Such an observation was consistent with earlier works which had observed that TFs from the same family had somewhat similar binding motifs (**Fornes et al. 2020; Kielbasa et al. 2005**). To explore the overall relationships among prime motifs of different TF family, we clustered the PWMs of these 441 TFs. We identified 48 TF clusters having members closely related to one another with respect to their primary motifs (**Supplementary Figure 2 and 3**). We observed that closely related TFs tended to have similar prime motifs. Even the members from vast and functionally diverse bZIP and NAC families displayed close proximity and shared similar prime motifs in the respective clusters reflecting previous observations (**Corrêa et al. 2008; Olsen et al. 2005**). Although the 387 DAP-seq and 54 ChIP-seq based motifs comprehensively describe motif types, they were skewed towards more prominent and more studied families like AP2EREBP, bZIP, MYB, and WRKY while under representing other families like C2H2, G2like, Homeobox, ZFHD, and C3H (**Supplementary Figure 2 and 3**).

The motifs reported in the present study were compared with the experimentally reported motifs using TOMTOM (**Bailey et al. 2006**). Many of the motifs found in this study matched with the experimentally reported motifs (**Supplementary Table 2 Sheet 5-7**). It was also observed that several of these experimentally reported motifs were not the prime motif reported here but matched with other lower ranked motifs which either co-occurred with the prime motifs or were exclusively present, covering a comparatively lesser amount of ChIP-seq and DAP-seq data than the prime motifs reported in the present study. It may be noted that the motifs reported through ChIP-seq and DAP-seq were from much bigger experimental space and data than previously reported motifs which may be a subset of a larger space, influencing their relative ranking. Their occurrence in the data varied from 3.42% to 99% while the prime motifs reported in this study mostly covered at least 75% of the bound data (**Supplementary Table 2 Sheet 5-7**). **Figure 4** provides a snapshot of comparison between motifs reported in this study and their experimentally reported counterparts. Most of the time the experimentally validated motifs displayed similarity with the prime motif.

Structural and molecular dynamics analyses support the prime motifs as important one for TF binding

The docking results revealed that the identified prime motif regions were showing the highest binding affinity with their corresponding TFs than the other regions. The detailed information of docking score of all TFs is given in **Supplementary Table 2 Sheet 8.**

Comparative analysis of Root Mean Squared Deviation (RMSD) trajectories of five different TFs-DNA complexes from each sequence partner to the interacting TF were studied for stability. The trajectory was measured at 300K for 50-ns. The complexes with the prime motifs were found to be much more stable than the other regions. The method section provides the details on various regions considered. For example, in case of AT5G66940, comparative analysis of RMSD value for the Apo AT5G66940, the value ranged from 0.5 to 2.7 nm while for the prime motif consisting complex A, the value ranged from 0.5 to 2.4 nm and stabilized at 2 nm. Similarily, RMSD value for the complex B, the values ranged from 0.5 to 2.3 nm and stabilized at 2.2 nm, which was lesser stable. Complex C and D had RMSD values ranging from 0.1 to 4 nm and 0.5 to 2.2 nm and got stabilized at 3.9 nm and 2.1 nm, respectively. For the complex E, RMSD showed deviation from 0.1 to 3.8 nm and settled at 3.5 nm (**Figure 3j**). In all the five complexes, the sequence with the prime motif (complex A) was found to be more stable when compared to the others in the dynamic environment. Moreover, the second sharpest impact was observed for the overlapping and immedietely flanking regions around the prime motif. As one moves away from the prime motif region, the impact on binding diminishes. The identified prime motifs provided structural stability to the considered TF-DNA complexes (**Figure 3e-i**). A similar pattern was observed for other three TFs taken in this study (**Supplementary Figure 4**). The docked complex and its binding affinity values are provided in **Figure 3.**

Recognition of TF binding sites requires contextual information surrounding the anchoring motif

The motif discovery and anchoring steps helped in selecting the more appropriate contextual information and features. Motifs may occur significantly enriched in the binding data but by no means they are limited there only. Overtly relying on motifs alone may not help build a more complete model. The discovered motifs above worked as the point to zero upon to consider the potentially significant interaction spots across the DNA. Thus, it becomes imperative to assess their context and its contribution also. Therefore, 75 bases flanking regions from both ends were considered with the motif region as the central anchor. Previously, it was found that the flanking regions around the potential interaction sites in nucleic acids capture the local environment contributory information effectively (**Heikham and Shankar 2010; Schöne et al. 2016; Vr et al. 2018**). Also, the TFs which interact with the DNA through multiple domains use multiple interaction sites which are usually concentrated around a local region instead of being long-distanced interaction spots. The MD simulation and co-occurring motifs studies above very ably supported the same. The contextual information came in the form of other co-occurring motifs, sequence specific information, position specific information, and structural/shape information which could work as strong discriminators against the negative and positive instances for TF binding. Considering the flanking regions around the motifs, three different datasets were constructed, Datasets “A”, “B” and “C” as described in the methods section. **Figure 5** illustrates how these datasets were built.

For building the models for 436 TFs and their binding preferences, we used 10 different combinations of various sequence representations. They were evaluated for performance through the raised transformer encoder based deep learning model. An assessment was made for each representation considered where the Dataset “A” was split into 70:30 ratio to form the train and test datasets. This protocol worked as the ablation analysis to evaluate how each of these representations of the sequence was contributing towards the discrimination between the preferred binding and non-preferred binding regions. The observed accuracy for dimeric representation was just 80.78% on Dataset “A”. This was followed by introduction of pentameric and heptameric sequence representations which returned the accuracy values of 85.05% and 86.23% respectively, while covering a total of 156 and 154 words per sequence window, respectively. The ChIP/DAP-seq data doesn’t retain strand information and complementary strands are also present almost equally and in most of the cases they too contribute in the binding. Considering the anchor motif’s counterpart from the complementary strand for the same binding region may boost the discriminating power further. Therefore, both strands were were considered. By doing so, when we reevaluated the accuracy values for dimeric, pentameric, and heptameric representations, significant improvement by ~5% was noted for each of the individual representations. Yet, as can be seen here, individually all these representations displayed enough scope of improvement and needed information sharing with each other. Therefore, in the final stage the datasets were formed having all the three representations of the sequences which covered base stacking, sequence composition, shape and motif information together in terms of dimeric, pentameric, and heptameric words. On these the transformers learned contextually along with the prime motifs with much higher amount of information sharing across the representations.

Combination of the various representations of the sequences was done in a gradual manner in order to see their additive effect on the classification performance. These combinations of the words representations yielded a better result than using any single type words representations, as can be seen from **Figure 6a.** Dimers+heptamers (313 words), dimers+pentamers (315 words), and dimers+pentamer+heptamers (469 words) combinations yielded the accuracy values of 90.88%, 94.9%, and 95.57%, respectively, with the last one attaining even the best balance between the sensitivity and specificity values. Thus, combinations of different representations for the genomic sequences markedly improved the performance through the natural language processing approach of transformers. A remarkable jump in the overall performance was noted with an average accuracy of 95.57% (Avg F1-Score: 95.55%, MCC: 0.92). The accuracy values varied across the transcription factors from 89.67% (PIF4) to 99.9% (HB6) for Dataset "A" TFs. Though not a big gap, there was an average gap of 0.47% between sensitivity and specificity. It was found that an average accuracy of 96.18% (Avg F1-Score: 96.16%, MCC: 0.95) was obtained for Dataset "B" TFs, where the accuracy values varied from 79.64% (GT3A) to 99.71% (CRF10) with average specificity and sensitivity of 95.53% and 96.83%, respectively. Complete details about word representations and performance can be found in **Supplementary Table 3 Sheet 1-3**. Details of the implementation of the optimized transformer are already given in the methods section and **Figure 2**.

Overall, it was evident that the Transformer effectively grasps both distant and nearby word associations, while acquiring knowledge through multiple hidden features. On Dataset "A", the raised transformer models performance hardly fell below 90% accuracy. In overall, also the average performance of the models was significantly enhanced. As evident from the performance for Dataset "A" (**Supplementary Figure 5**), the developed approach demonstrated a consistently strong and dependable performance across the majority of the analyzed TFs.

Attention based deep-learning by transformers on the words delivers remarkably consistent performance

As mentioned in the Methods section and **Figure 5**, initially for performance testing two different datasets, "A" and "B" were created. Dataset "A" covered 51 TFs, totaling 2,299,667 instances of the binding regions while Dataset "B" covered 385 TFs, totaling 5,002,532 instances of the binding regions. For every TF from dataset "A" and "B", 70% of the dataset was used for the training purpose and 30% was kept aside as the totally unseen test set instances. To conduct unbiased performance testing without any potential recollection of data instances, it was ensured that the test data was kept aside in a mutually exclusive manner. Besides raising the trained models and testing it, as mentioned in the above section, 10 times random train-test trials have also been done to evaluate the consistency of the trained transformer system on the datasets "A" and "B".

This 10-fold random trials concurred with the above observed performance level and scored in the same range consistently. All of them achieved high quality ROC curves with high AUC values in the range of 0.9245 to 0.9561 (Dataset "A") and 0.9562 to 0.9869 (Dataset "B") while maintaining reasonable balance between specificity and sensitivity (**Figure 6b; Supplementary Table 3 Sheet 4-**

5). The remarkable performance consistency ensured about the robustness of the raised transformer models and reliability of its all future results.

PTFSpot transformer models consistently surpassed all other compared TF binding regions finding tools

A highly comprehensive series of benchmarking studies was performed, where initially two different datasets, "B", and "C", were used to evaluate the performance of the transformer models of PTFSpot with respect to seven different tools, representing most recent and different approaches of TF binding regions detection: k-mer grammar, Wimtrap, SeqConv, TSPTFBS, TSPTFBS 2.0, DNABERT, and PlantBind. The performance measure on the test set of Dataset "B" gave an idea how the compared algorithms in their existing form perform. The third dataset "C" was also used to carry out an objective comparative benchmarking, where each of the compared software was trained as well tested across a common dataset in order to fathom exactly how their learning algorithms differed in their comparative performance.

All these seven tools were trained and tested across Datasets "B" where PTFSpot outperformed almost all of them, for all the performance metrics considered (**Figure 6c**). TSPTFBS 2.0 came very close to the performance of PTFSpot transformers with an average accuracy of 96.22% whereas PTFSpot transformers scored an average accuracy of 96.18%. PTFSpot transformers achieved the MCC value of 0.955 whereas TSPTFBS 2.0 achieved MCC value of 0.9273. PTFSpot transformers displayed better performance balance. On the same Dataset "B", after TSPTFBS 2.0 and PTFSpot transformers, the next best performing tools were SeqConv and k-mer grammar which scored accuracy values of 88.30% and 87.34% and MCC values of 0.7978 and 0.7809, respectively. The values were significantly behind those observed for PTFSpot transformers by huge margins. Dataset “B” was developed to simulate the actual binding motif distribution scenario by selecting non-binding regions of the genome that contained the binding motif. This introduced significant complexity to the learning process while bringing it to more realistic data to learn from, ultimately preventing the motifs from being disproportionately overemphasized, as relevant contextual information became another important factor.

In the benchmarking study done across Dataset “C” PTFSpot transformers outperformed all the compared tools with significant margin with the similar level of performance (**Figure 6e**). It scored an average accuracy of 96.04% and MCC value of 0.9262. DenseNet based TSPTFBS 2.0 came close second one with an average accuracy of 95.91% and MCC value of 0.9222 but with a gap of ~0.8% between sensitivity and specificity scores. The third best performing tool was DNABERT which attained an average accuracy of 90.93% and MCC value of 0.8209. Such sharp fall in the MCC value was observed due to huge gap of ~4% between sensitivity and specificity values. PTFSpot transformers consistently demonstrated minimal variance and distribution in its performance, maintaining a strong balance in identifying both positive and negative instances with a high level of precision. This was evident through its high values across all the three performance metrics with least dispersion, affirming the algorithm's robustness (**Figure 6d & 6f**). The full details and data for this benchmarking study are given in **Supplementary Table 3 Sheet 6-7**.

It was observed from the above benchmarking study that the significant role of the negative dataset was eventually neglected by all these tools. These tools mostly considered random genomic sequences regions similar to the positive instances on the basis of GC percent. None of the tools tried to implement a more pragmatic approach to construct a robust negative dataset which could reflect the actual situation. That is why the tools performed very well on Dataset “C” where these tools achieved average accuracy values of >=90% when compared to the Dataset “B” where only three tools could achieve >=90% accuracy values. This could happen because all these tools never considered a realistic negative dataset and used just random sequences as the negative instances, a much simplistic representation than the actual. Most of these tools overtly rely on the binding motifs while completely discounting the fact that motifs similar to those found in the bound regions are present also in the binding free regions. Also, most of them did not learn from the context. All, resulting into marked inconsistent performance by the compared tools across the datasets. It needs to be reminded here that the dataset “C” was built using the same datasets used by the compared tools (**Liu et al. 2021; Cheng et al. 2023**). While dataset “B” was developed capturing the realistic situation by picking those non-binding regions of the genome which contained the binding motif and created a strong confusion to the learning process where eventually the motifs were not given excess weights as context also weighed in.

The transcription factors vary across species and so do their binding region motifs

One of the major calamities plant science has faced while exactly following the studies on humans is approaching TFs and DNA interactions in the similar fashion while overtly assuming binding sites conservation. This is why to find TFs and their binding sites they often turn to *Arabidopsis* models, ending up with largely misleading results. The transcription factors and their binding sites vary across the species (**Kielbasa et al. 2005; Shiu et al. 2005**). As the TF structure changes, the binding sites also changes, and so does the referred binding regions. Therefore, what works for one species, does not work for another. And this is why most of the existing tools fail to work in cross-species manner, making them almost of no use. We carried out a study here to verify the same. We selected peak data of the common TFs in *Arabidopsis thaliana* and *Zea mays* i.e., LHY1, MYB56, MYB62, MYB81, MYB88, and WRKY25. For both the species their corresponding prime motifs for associated similar

TFs were placed for comparison at the first. We observed that for the same TF different prime binding motifs existed. Therefore, we examined their corresponding TF protein structure to investigate possible structural changes in correspondence to the observed motif changes. We did a comparative study of each TF on the following bases: 1) based on the amino acid sequence, 2) based on the functional domain, 3) based on their respective 3D structure, and 4) we also analyzed the binding affinity of the TF across the species. The following observations were made:

1) We found that the amino acid sequence identity for each TFs was reasonably low between the two species. The sequence identities of LHY1, MYB56, MYB62, MYB81, MYB88 and WRKY25 between both species were 39%, 40.7%, 77.2%, 35.7%, 36.2% and 43.9%, respectively. The amino acids sequence-based comparative details of each TF are given in **Supplementary Table 3 Sheet 8.**

2) A functional domain is a protein's structural unit that is spatially distinct, compact, stable, and has the potential to fold and function alone. Based on domains study, we found that the binding domain class of the compared TFs across the species were same but their amino acid compositions were different. As an example, in the case of MYB88, both species contained two domains known as the HTH DNA-binding domain. The amino acid composition varied between these domains in the two species. Specifically, an identity of 33.3% was observed for the first domain, and an identity of 43.2% was noted for the second domain when comparing both species **(Figure 7bI & 7bII)**. The details of comparative study of other four TFs are given in **Supplementary Table 3 Sheet 9.**

3) Protein 3D structure comparison measures the structural similarity between two proteins. It is a foundational analysis for structure-based protein function prediction. 3D structure of the compared TFs were modelled with Alphafold2 (**Jumper et al. 2021**) as experimentally obtained structures of most of

the plant TFs is unavailable, while the remarkable accuracy of Alphafold2 has been widely acknowledged**.** Afterwards, we superimposed the 3D structures of the similar TFs of the compared species and noted the RMSD value between them. The RMSD value of each of the compared TF pair was above 0.6 A^o **(Figure 7)**. This strongly suggested that the same TF varies significantly in its 3D structure when one goes across the various plant species. The superimposed structures for WRKY25 and MYB88 are given in **Figure 7aIII and aIV,** respectively**.** The details of 3D structure comparison of other TFs is given in **Supplementary Figure 6.**

4) To assess the binding affinity between the transcription factor and its binding region motif, we studied the MYB88. Initially, we identified the prime motif for MYB88 in both species using PTFSpot. Subsequently, we conducted the docking analysis pairing the identified motifs with TF within the same species as well as the same for inter-species comparison. It was observed that when the TF of *Arabidopsis thaliana* was docked with its own motif, the binding affinity was -66.45 kcal/mol. However, when docked with the motif of *Zea mays*, the binding affinity went much lower with -46.70 kcal/mol **(Figure 7bIII & bIV)**. The same procedure was applied for *Zea mays* and similar pattern was observed there also **(Figure 7bV & bVI)**.

These important findings formed the foundation for the final form of PTFSpot as a universal TF-DNA interaction modeller which could work across the species and for even unseen TFs, while parallelly learning upon the structural variations and corresponding changes in binding region partner.

Deep co-learning on TF structure and corresponding binding regions brings impeccably accurate universal model of TF-DNA interaction spots

Performance over standard testing datasets may be claimed good, as has been done by most of the published software in the past. However, during the real world application of inter-species identification of TFBS, a huge performance gap exists, far below the acceptable limits. Some recent reports have highlighted the high degree of poor performance by a majority of the existing categories of TFBS discovery during the process of TFBS annotations where most of them end up reporting very high proportion of false positives (**Shen et al. 2021b; Liu et al. 2021; Ji et al. 2021; Jyoti et al. 2023**). Above we have showed how variability in transcription factor structure and corresponding binding sites happen across the plants. This becomes a major reason why the existing software pool does not work across the plant species. A possible solution for this appeared in learning the variation in the TF structure and corresponding binding sites. Such learning would not only enable us to detect TFBS in cross species manner but in fact would be able to detect any transcription factor's interaction across any genome, clearly bearing a deep reaching impact on the field.

As detailed in the methods sections about the architecture of PTFSpot implementation, a composite deep co-learning system was raised using Transformers and DenseNet which parallely learned upon the TFBS with sequence contexts and corresponding TF's 3D structure. This model was trained and tested on Dataset "E" (Derived from Dataset "B") training and testing components, in 70:30 split ratio with absolutely no overlap. This co-learning system was trained using 41 TFs and their corresponding DAP-seq data. Each TF represented a single TF family and had highest binding data available among all the TFs for the respective family. This limited consideration was done to overcome the burden of data handling and computational resources requirements arising due to that. Just considering this subset of TFs was able represent the whole of 325 TFs in Dataset "B" without posing much computational resources requirements. The performance of the raised model achieved an excellent accuracy of 98.3% with a balanced sensitivity and specificity values of 97.56% and 99.04%, respectively (**Figure 8a;**

Supplementary Table 4 Sheet 1). In order to assess how much important role this consideration of transcription factor’s structure played, we conducted the next part of this study to evaluate the model's performance without incorporating the 3D structure. Here, we exclusively utilized the prime motif anchored binding sequences to train and test a deep-learning transformer-based module using the same dataset of 41 TFs from *Arabidopsis thaliana*. The achieved accuracy for the developed model dropped to 83.6%, significantly lower than what was achieved above with co-learning on TF structure and corresponding binding data. This strongly supported the importance of co-learning on sequence and structural variability and the extent to which such system is capable to resolve long standing major issue of a sole universal model of TF DNA interactions. The above raised model, unlike the existing ones, assumed all TFs falling into one hypothetical family whose structure and corresponding binding regions varied from one member to another, and this co-variability was learned to correctly identify the binding sites for even those transcription factors on which it was not even trained.

The next important question was that how this co-learning system performed when introduced to different sets of transcription factors which had not even family representations in the above mentioned model which covered only 41 TF families. The Dataset “D”, derived from Dataset “A”, had included 51 TF dataset. In the first part, we considered 21 TFs datasets exclusively. Each of these TFs represented a TF family which was not included in the training of the above mentioned co-learning model. This entire dataset worked as another test set to evaluate how much capable the above raised co-learning model was for even those cases for which not even a single family member was fed into its training. The co-learning model achieved an astonishing accuracy of 93.7% with balanced sensitivity and specificity values of 93.38% and 94.02%, respectively (**Supplementary Table 4 Sheet 2)**. This clearly confirmed PTFSpot co-learning capability to detect the binding sites even for those transcription factors and families which were not even encountered during its training. Afterwards, we tested this model on the whole Dataset “D” where the total number of different TFs was 51. The average accuracy remained in the same range (93.6%) with very balanced sensitivity and specificity values of 93.2% and 94%, respectively (**Figure 8a; Supplementary Table 4 Sheet 2)**.

So far, we had assessed the PTFSpot co-learning model in the same species data i.e. *A. thaliana*. There it clearly showed its capability to accurately detect the binding regions for even those transcription factors which were never seen before as well as families whose not a single member was even seen before. This gave a good degree of confidence that such model would work even cross-species. The application value of any such software is mainly when it is applicable universally, across various transcription factors as well as across different species. More so when majority of plant species are not sequenced and now very frequently sequenced genomes are being released which essentially require TFBS annotations. Almost all of the existing software fail there. The reason for which is already discussed above.

The raised model's last assessment was set to determine how well it performed to identify trans-species TFBS. For this task a new dataset, Dataset “F”, was employed completely as another test set. This dataset contained 117 TFs from *Zea mays* (93 TFs) and *Oryza sativa* (24 TFs), retrieved from GSE137972 (217 sample), GSE102920 (six sample), and ChIP-Hub with over 60 conditions, elaborately explained in the methods section. It was created solely for trans-species validation purpose for the model's performance. For plants, hardly any tool has been able to show their practical applications to identify trans-species TFBS. In this study we included most recent and advance tools like *k-mer* grammar, TSPTFBS, TSPTFBS 2.0, SeqConv, and Wimtrap to clearly point this out. We looked for common TFs among different plant species to benchmark these tools for cross-species performance. In Dataset “F”, only six TFs from *Zea mays* (LHY1, MYB56, MYB62, MYB81, MYB88, WRKY25) were found for common ones which the existing software had any model developed (**Supplementary Table 4 Sheet 3)**. Thus, the comparative benchmarking was possible for only these six transcription factors for cross-species performance evaluation of the tools.

Two levels bench-marking was performed. In the first level, tools such as DNABERT, *k-mer* grammar, TSPTFBS, TSPTFBS 2.0, SeqConv, and Wimtrap were considered and the six common TFs data were taken for evaluation along with PTFSpot universal model. PTFSpot universal model achieved an outstanding average accuracy of 92.9% ranging from 91.76% to 94.9%, clearly underlining its capability to accurately identify TFBS in cross-species manner. The performance observed for all the other tools was extremely poor, with none surpassing an average accuracy value of 60% (**Figure 8b; Supplementary Table 4 Sheet 3)**, reiterating the fact that none of the existing tools are suitable for any practical application like cross-species TFBS identification and genomic annotations. PTFSpot has emerged as a breakthrough solution to this situation.

In the next step, the above analysis was carried forward for the entire Dataset “F”, which covered data for 117 TFs from the two species. Unlike the existing tools, PTFSpot does not depend on TF specific models and uses a single universal model as described above which can work for any transcription factor and genome, including never seen before ones. Here also, PTFSpot performed outstanding while attaining 93.58% average accuracy, MCC value of 0.87, and F1-score of 93.56%. **Figure 8c** provides further performance distribution illustration of PTFSpot for *O. sativa* and *Z. mays* transcription factors (**Supplementary Table 4 Sheet 4)**. All these series of validation and benchmarking studies proved that PTFSpot achieved a never seen before success in consistently and accurately identifying the binding sites for various families of transcription factor as well as across species due to its successful co-learning of the variability in structure and binding regions.

Web-server and stand-alone implementation

The source codes, accompanying supporting content, and tutorial are all hosted on the P-TFSpot server. Additionally, the PTFSpot server hosts the software in a user friendly interactive manner. To identify TF binding regions, users are required to upload their DNA sequences in FASTA format and its associated TF in pdb format, followed by clicking the submit button. The input data is then passed through Transformer-DenseNet engine in the back-end which identifies most potential binding regions to be looked upon for the given TFs. It is strongly recommended to execute the whole PTFSpot system by downloading the standalone versions available at Github for bulk study for more than 100 sequences or genome wide studies. Subsequently, a results page featuring relevant information in an interactive format is presented. Afterwards, users can also identify binding sites in the binding regions discovered through PTF-Vāc (**Gupta et al. 2024;** https://scbb.ihbt.res.in/PTF-Vac/index.php) and TF-MoDISco (**Shrikumar et al. 2018**). It is recommended to use PTF- Vāc which uses PTFSpot's universal model as guide to precisely identify the binding motifs and can identify the binding sites even for single sequence in *ab-initio* manner, which most of the existing software fail to achieve. Further details on PTF-Vāc can be found elsewhere (**Gupta et al. 2024**). Users can download results in a tabular format. The server implementation has been done using PHP, JavaScript, HTML5, and the Apache-Linux platform. The majority of codes were developed in Shell and Python for jobs related to data curation and back-end processing. The project was executed within the Ubuntu Linux platform's Open Source OS environment. The PTFSpot system is freely accessible at https://scbb.ihbt.res.in/PTFSpot/. An overview of the PTFSpot web server implementation is shown in **Figure 9**. It has been made freely available at https://scbb.ihbt.res.in/PTFSpot/ and GitHub (https://github.com/SCBB-LAB/PTFSpot).

Conclusion

Unlike animal genomes where everything is focused towards human with assumptions and understanding that conservation of transcription factors binding sites is high, plants pose totally different problems. They have large number of species with a majority without sequenced genomes and where variability is much higher. Amid this, plainly following the protocols set for animal TFBS discovery has only generated inaccurate findings in plants where similarity finding with *Arabidopsis* models has become a norm. This is very well reflected by the current pool of plant specific software to find TFBS which work with one by one TF specific models, mainly developed for *Arabidopsis*, and fail to correctly identify TFBS in other species. Nor they can even handle any transcription factor whose specific model they don’t have even within *Arabidopsis* or any species. The present study has brought these fact with highly extensive and layered investigations and also gave the reason why these software fail. Due to this all, there is a huge void in transcription regulatory research in plant science as there is no credible software which could be relied upon to discover TFBS in plants. This study learned from their pitfalls where the major reason were neglect of context, over dependence on motifs, and complete cluelessness about variability of TF structure and corresponding binding regions in plants. Using a novel deep co-learning architecture based on transformers and DenseNet, we have solved this problem and have developed a software, PTFSpot, which brings a universal model of TF and DNA binding regions interactions. It learns from the co-variability between binding protein structure and its binding regions without requiring to be specific for any particular transcription factor or its family. It can accurately identify the binding regions for any given transcription factor belonging to any family across any plant genome and can work for even any novel and never reported before transcription factors and genomes. With this, the present work is expected to drastically change the scenario of plant regulatory research as well as may cause extensive cutting of cost incurred on experiments to detect binding regions across a genome. PTFSpot will give very precise and accurate information to zero upon for any transcription factor’s most potential interacting sites across any plant genome. Also, the approach developed here may be used to other interactomics studies and develop intelligent systems in highly specific manner including unraveling of RBP RNA interactions in plants, which is another under developed but highly critical area of plant regulomics research. The next to this study may aim to bring an approach which could be time and condition specific in telling when a given spot would become active in regulation, a problem which remains poorly addressed and answering which may bring another level of revolution in regulatory research.

Acknowledgements

The work was carried out under the aegis of The Himalayan Centre for High-throughput Computational Biology (HiCHiCoB), a BIC supported by DBT, Govt. of India. SG and VK are thankful to DBT, India for financial support as project associateship. UB is thankful for DBT JRF fellowship. Jyoti is thankful for CSIR-UGC JRF fellowship. This MS has CSIR-IHBT MSID ##.

Author’s contributions

SG carried out the major parts of this study and developed the web-server and standalone versions of PTFSpot. VK carried out the structural analyses. Jyoti, UB, and VK carried out the benchmarking study. RS conceptualized, designed, analyzed, and supervised the entire study. SG, VK, UB, Jyoti, and RS wrote the MS.

Declaration of competing interest

The authors declare that they have no competing interests.

Funding

The work was carried out under the aegis of The Himalayan Centre for High-throughput Computational Biology (HiCHiCoB), a BIC supported by DBT, Govt. of India [BT/PR40122/BTIS/137/30/2021] and its National Network Project, S2S [Sanction order number: 102/IFD/SAN/397/2022-23 dated 26.08.2022].

Software and Data availability

All the secondary data used in the present study were publicly available and their due references and sources have been provided in **Supplementary Tables 1–4**. The software has also been made available at GitHub at https://github.com/SCBB-LAB/PTFSpot as well as a webserver at https://scbb.ihbt.res.in/PTFSpot/ (all related datasets in the study are hosted here).

References

Alipanahi, B., Delong, A., Weirauch, M.T., and Frey, B.J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat Biotechnol 33: 831–838.

Bailey, T.L., Williams, N., Misleh, C., and Li, W.W. (2006). MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic Acids Research 34: W369–W373.

Bartlett, A., O’Malley, R.C., Huang, S.C., Galli, M., Nery, J.R., Gallavotti, A., and Ecker, J.R. (2017). Mapping genome-wide transcription-factor binding sites using DAP-seq. Nat Protoc 12: 1659–1672.

Berendsen, H.J.C., van der Spoel, D., and van Drunen, R. (1995). GROMACS: A message-passing parallel molecular dynamics implementation. Computer Physics Communications 91: 43–56.

Berger, M.F. and Bulyk, M.L. (2006). Protein Binding Microarrays (PBMs) for Rapid, High-Throughput Characterization of the Sequence Specificities of DNA Binding Proteins. In Gene Mapping, Discovery, and Expression: Methods and Protocols, M. Bina, ed, Methods in Molecular Biology. (Humana Press: Totowa, NJ), pp. 245–260.

Černý, J., Božíková, P., Svoboda, J., and Schneider, B. (2020). A unified dinucleotide alphabet describing both RNA and DNA structures. Nucleic Acids Research 48: 6367–6381.

Cheng, H., Liu, L., Zhou, Y., Deng, K., Ge, Y., and Hu, X. (2023). TSPTFBS 2.0: trans-species prediction of transcription factor binding sites and identification of their core motifs in plants. Frontiers in Plant Science 14.

Chicco, D. and Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics 21: 6.

Chow, C.-N., Lee, T.-Y., Hung, Y.-C., Li, G.-Z., Tseng, K.-C., Liu, Y.-H., Kuo, P.-L., Zheng, H.-Q., and Chang, W.-C. (2019). PlantPAN3.0: a new and updated resource for reconstructing transcriptional regulatory networks from ChIP-seq experiments in plants. Nucleic Acids Research 47: D1155–D1163.

Corrêa, L.G.G., Riaño-Pachón, D.M., Schrago, C.G., Santos, R.V. dos, Mueller-Roeber, B., and Vincentz, M. (2008). The Role of bZIP Transcription Factors in Green Plant Evolution: Adaptive Features Emerging from Four Founder Genes. PLOS ONE 3: e2944.

Fornes, O. et al. (2020). JASPAR 2020: update of the open-access database of transcription factor binding profiles. Nucleic Acids Research 48: D87–D92.

Fu, L.-Y., Zhu, T., Zhou, X., Yu, R., He, Z., Zhang, P., Wu, Z., Chen, M., Kaufmann, K., and Chen, D. (2022). ChIP-Hub provides an integrative platform for exploring plant regulome. Nat Commun 13: 3413.

Ghandi, M., Mohammad-Noori, M., Ghareghani, N., Lee, D., Garraway, L., and Beer, M.A. (2016). gkmSVM: an R package for gapped-kmer SVM. Bioinformatics 32: 2205–2207.

Gupta, S., Jyoti, Bhati, U., Kesarwani, V., and Shankar, R. (2024). PTF-Vāc: Ab-initio discovery of plant transcription factors binding sites using deep co-learning encoders-decoders.: 2024.01.28.577608.

Heikham, R. and Shankar, R. (2010). Flanking region sequence information to refine microRNA target predictions. J Biosci 35: 105–118.

Hemberg, M. and Kreiman, G. (2011). Conservation of transcription factor binding events predicts gene expression across species. Nucleic Acids Research 39: 7092–7102.

Higo, K., Ugawa, Y., Iwamoto, M., and Higo, H. (1998). PLACE: A database of plant cis -acting regulatory DNA elements. Nucleic Acids Research 26: 358–359.

Hombach, D., Schwarz, J.M., Robinson, P.N., Schuelke, M., and Seelow, D. (2016). A systematic, large-scale comparison of transcription factor binding site models. BMC Genomics 17: 388.

Huang, G., Liu, Z., van der Maaten, L., and Weinberger, K.Q. (2017). Densely Connected Convolutional Networks. In, pp. 4700–4708.

Jha, A. and Shankar, R. (2014). MiRNAting control of DNA methylation. J Biosci 39: 365–380.

Ji, Y., Zhou, Z., Liu, H., and Davuluri, R.V. (2021). DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. Bioinformatics 37: 2112–2120.

Jiménez-García, B., Pons, C., and Fernández-Recio, J. (2013). pyDockWEB: a web server for rigid-body protein–protein docking using electrostatics and desolvation scoring. Bioinformatics 29: 1698–1699.

Jin, J., Tian, F., Yang, D.-C., Meng, Y.-Q., Kong, L., Luo, J., and Gao, G. (2017). PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. Nucleic Acids Research 45: D1040–D1045.

Johnson, D.S., Mortazavi, A., Myers, R.M., and Wold, B. (2007). Genome-Wide Mapping of in Vivo Protein-DNA Interactions. Science 316: 1497–1502.

Jumper, J. et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature 596: 583–589.

Jyoti, Ritu, Gupta, S., and Shankar, R. (2023). Comprehensive evaluation of plant transcription factors binding sites discovery tools.: 2023.11.07.566153.

Kielbasa, S.M., Gonze, D., and Herzel, H. (2005). Measuring similarities between transcription factor binding sites. BMC Bioinformatics 6: 237.

L, D.W. (2002). The PyMOL molecular graphics system. http://www.pymol.org/.

Liu, L., Zhang, G., He, S., and Hu, X. (2021). TSPTFBS: a Docker image for trans-species prediction of transcription factor binding sites in plants. Bioinformatics 37: 260–262.

Mejía-Guerra, M.K. and Buckler, E.S. (2019). A k-mer grammar analysis to uncover maize regulatory architecture. BMC Plant Biology 19: 103.

Nakato, R. and Shirahige, K. (2017). Recent advances in ChIP-seq analysis: from quality management to whole-genome annotation. Briefings in Bioinformatics 18: 279–290.

Olsen, A.N., Ernst, H.A., Leggio, L.L., and Skriver, K. (2005). NAC transcription factors: structurally distinct, functionally diverse. Trends in Plant Science 10: 79–87.

O'Malley, R.C., Huang, S.C., Song, L., Lewsey, M.G., Bartlett, A., Nery, J.R., Galli, M., Gallavotti, A., and Ecker, J.R. (2016). Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape. Cell 165: 1280–1292.

Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., and Ferrin, T.E. (2004). UCSF Chimera—A visualization system for exploratory research and analysis. Journal of Computational Chemistry 25: 1605–1612.

Rhee, H.S. and Pugh, B.F. (2011). Comprehensive Genome-wide Protein-DNA Interactions Detected at Single-Nucleotide Resolution. Cell 147: 1408–1419.

Rivière, Q., Corso, M., Ciortan, M., Noël, G., Verbruggen, N., and Defrance, M. (2022). Exploiting Genomic Features to Improve the Prediction of Transcription Factor-Binding Sites in Plants. Plant Cell Physiol 63: 1457–1473.

Schöne, S., Jurk, M., Helabad, M.B., Dror, I., Lebars, I., Kieffer, B., Imhof, P., Rohs, R., Vingron, M., Thomas-Chollier, M., and Meijsing, S.H. (2016). Sequences flanking the core-binding site modulate glucocorticoid receptor structure and activity. Nat Commun 7: 1–13.

Sharma, N.K., Gupta, S., Kumar, A., Kumar, P., Pradhan, U.K., and Shankar, R. (2021). RBPSpot: Learning on appropriate contextual information for RBP binding sites discovery. iScience 24: 103381.

Shen, L.-C., Liu, Y., Song, J., and Yu, D.-J. (2021a). SAResNet: self-attention residual network for predicting DNA-protein binding. Briefings in Bioinformatics 22: bbab101.

Shen, W., Pan, J., Wang, G., and Li, X. (2021b). Deep learning-based prediction of TFBSs in plants. Trends in Plant Science 26: 1301–1302.

Shen, Z., Bao, W., and Huang, D.-S. (2018). Recurrent Neural Network for Predicting Transcription Factor Binding Sites. Sci Rep 8: 15270.

Shiu, S.-H., Shih, M.-C., and Li, W.-H. (2005). Transcription Factor Families Have Much Higher Expansion Rates in Plants than in Animals. Plant Physiol 139: 18–26.

Shrikumar, A., Tian, K., Avsec, Ž., Shcherbina, A., Banerjee, A., Sharmin, M., Nair, S., and Kundaje, A. (2020). Technical Note on Transcription Factor Motif Discovery from Importance Scores (TF-MoDISco) version 0.5.6.5.

Sielemann, J., Wulf, D., Schmidt, R., and Bräutigam, A. (2021). Local DNA shape is a general principle of transcription factor binding specificity in Arabidopsis thaliana. Nat Commun 12: 6549.

The ENCODE (ENCyclopedia Of DNA Elements) Project (2004). Science 306: 636–640.

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

Vr, Y., D, B., D, G., Ja, R.-M., Az, A., and M, B. (2018). Flexibility and structure of flanking DNA impact transcription factor affinity for its core motif. Nucleic acids research 46.

Yan, W., Li, Z., Pian, C., and Wu, Y. (2022). PlantBind: an attention-based multi-label neural network for predicting plant transcription factor binding sites. Briefings in Bioinformatics 23: bbac425.

Yang, J., Ma, A., Hoppe, A.D., Wang, C., Li, Y., Zhang, C., Wang, Y., Liu, B., and Ma, Q. (2019). Prediction of regulatory motifs from human Chip-sequencing data using a deep learning framework. Nucleic Acids Research 47: 7809–7824.

Zhang, Q., Shen, Z., and Huang, D.-S. (2019). Modeling in-vivo protein-DNA binding by combining multiple-instance learning with a hybrid deep neural network. Sci Rep 9: 8484.

Zhou, T., Yang, L., Lu, Y., Dror, I., Dantas Machado, A.C., Ghane, T., Di Felice, R., and Rohs, R. (2013). DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale. Nucleic Acids Res 41: W56–W62.

Figure legends

Figure 1: Motif and data. (a) Clustering plot of the prime motifs for TF family AP2EREBP**.** The TFs share similarity in their prime motifs. **(b)** Different characteristics spectra for TFs belonging to same family, same cluster, and sharing similar prime motif. This highlights the importance of context. Despite having similar binding motifs, their binding preferences differ from each other due to context. **(c)** Lineplot for the data abundance for the collected 441 TFs (ChIP and DAP-Seq) across 56 different family.

Figure 2: Implementation of the PTFSpot Deep Co-learning system using Transformers and DenseNet to identify TFBS regions across plant genomes. The first part is a 14 heads attention transformers which learn from the dimeric, pentameric and heptameric words representations of any given sequence arising from anchoring prime motif's context. In the parallel, the bound TF's structure is learned by the DenseNet. Learning by both partners are joined finally together, which is passed on to the final fully connected layers to generate the probability score for existence of binding region in the center.

Figure 3: Molecular docking and simulation analysis between Transcription Factor (TF) and its binding motif. (a) 3D structure of AT5G66940 TF (*Arabidopsis thaliana*) with two domains (colored in Pale Red and Golden Red). **(b)** The arrangement of the loops and helices viewed from top after rotation at 90°, **(c-d)** Prime binding motif identification by PTFSpot for the TF AT5G66940, followed by selection of prime binding motif region as well as regions overlapping with it in its flanks and terminal end regions from both sides (A, B, D, C and E regions). Respective DNA 3D strcutures were generated for these regions as the central binding one. **(e-i)** The docked pose of each tiling-deletion variant DNA with the TF AT5G66940 and their binding affinity. Out of five docked complexes, the prime motif docked complex showed highest binding affinity than other complexes clearly supporting the importance in TF DNA interactions. **(j)** Molecular dynamics simulation of Apo TF and all five docked complex, the RMSD plot showed that the docked complex of AT5G66940 + region "A" exhibited lowest RMSD value among all, clearly reiterating the above observation that the prime motif region has importance in TF DNA interactions.

Figure 4: Comparison between experimentally reported motifs and motifs identified in the present study. a) Motifs derives from ChIP-seq TF data and **b)** Motifs derives from DAP-seq TF data. The similarity was determined using TOMTOM package. Several experimentally validated motifs were found similar to the identified prime motifs.

Figure 5: Flowchart representation of dataset formation. a) The protocol followed for Dataset "A" creation, **b)** Dataset "B" creation, **c)** Dataset "C" creation, **d)** Dataset "D" creation, **e)** Dataset "E" creation, and **f)** Dataset "F" creation. The datasets "A" and "B" contained the positive instances originating from ChIP-seq and DAP-seq, respectively, with sources PlantPAN3.0 and Plant Cistrome databases, respectively. Negative part of the datasets was formed by considering those genomic regions which display the regions similar to the prime motifs found enriched in DAP/ChIP-seq TF binding data but never appeared in the DAP/ChIP-seq data. Dataset "C" also contained the positive instances from the Plant Cistrome Database but for negative instances it contains random genomic regions. Most of the existing tools have used this dataset and its subsets. The datasets "D" and "E" were created from the datasets "A" and "B", respectively, while adding up the TF structural data also. Dataset "F" was constructed to be used completely as a test set to evaluate the universal model and its applicability in cross-species manner. This dataset covered 117 TFs from *Zea mays* (93 TF) and *Oryza sativa* (24 TF).

Figure 6: (a) Ablation analysis for three main properties for discriminating between the negative and positive instances. These words representations appeared highly additive and complementary to each other as the performance increased substantially as they were combined together. (**b)** Ten-fold independent training-testing random trials on Dataset "B" depicts consistent performance of PTFSpot transformers. **(c)** Objective comparative benchmarking result on Dataset "B". This datasets contained

the TF originating from DAP-seq from the plant cistrome database. Here all the compared tools were train and tested on the Dataset “B”. **(d)** Performance dispersion plot on Dataset “B”. PTFSpot transformers consistently demonstrated minimal variance and distribution in its performance, maintaining a strong balance in identifying both positive and negative instances with a high level of precision. **(e)** Objective comparative benchmarking on Dataset ”C”. Here, all the compared tools were first trained and then tested Dataset ‘C’ and evaluated for their performance. This gave a clear view on the performance of each of the compared algorithms. **(f)** Performance dispersion plot on Dataset “C”. PTFSpot transformers consistently demonstrated high values across all the three performance metrics with least dispersion, affirming the algorithm's robustness. From the plots it is clearly visible that for all these datasets and associated benchmarkings, PTFSpot consistently and significantly outperformed the compared tools for all the compared metrics (MCC values were converted to percentage representation for scaling purpose).

Figure 7. Co-variation in the structure of the transcription factor and the corresponding binding site across various species. (a) TF structure and its binding motif comparison across the species, **(aI)** prime binding motif of WRKY25 TF (*Arabidopsis thaliana*) and its 3D structure, **(aII)** prime binding motif of WRKY25 TF (*Zea mays*) and its 3D structure, **(aIII)** superimposed TF structures of *Arabidopsis thaliana* and *Zea mays*, with the structural differences measured in RMSD value. **(aIV)** prime binding motif of MYB88 TF (*Arabidopsis thaliana*) and its 3D structure, **(aV)** prime binding motif for WRKY25 TF (Zea mays) and its 3D structure, **(aVI)** superimposed TF structures for *Arabidopsis thaliana* and *Zea mays*, and corresponding structural difference in RMSD value. **(b)** Domain based comparison between *Arabidopsis thaliana* and *Zea mays* for MYB88, **(bI)** *Arabidopsis thaliana’s* MYB88’s has two domains. The first domain and its corresponding amino acids sequence are

in grey color. The second domain and its corresponding amino acids sequence is shown in the tan color, **(bII)** *Zea mays* MYB88 too has two domains. The first domain and its corresponding amino acids sequence are in purple color. The second domain and its corresponding amino acids sequence is shown in maroon color, **(bIII & bVI)** The docking analysis shows the stability of the complexes when a TF was docked to its binding motif within the same species and to the one from another species for the same TF. It is clearly evident that the same binding motif can't work across the species and it varies with species as well as the structure of the TF.

Figure 8: Performance and benchmarking of the universal model of PTFSpot. PTFSpot universal model was raised from 41 different TFs representing 41 different families, from *A. thaliana*. a) (i) The performance over the same 41 TF's test set (Dataset "E"). **(ii)** The performance over the remain TF's as test set from Dataset "E". **(iii)** The complete dataset "D" (containing ChIP-seq TFs) worked as another test set. Performance on all of them was in the same range with exceptional accuracy. **b)** Comparative benchmarking for the TFs whose models were available commonly among the compared tools. PTFSpot universal model outperformed them with huge leap for each TF. **c)** PTFSpot universal model was raised using *A. thaliana* TFs. It was tested in trans-species manner across rice and maize. For both, it returned outstanding results, reinforcing itself as the solution for reliable cross-species discovery of TFBS.

Figure 9: PTFSpot webserver implementation. a) Input data box where the user can either paste or load the input file with its associated TF structure, **b)** Input is DNA sequence in FASTA format and TF structure in PDB format. These sequences are chopped into sub-sequences of length 160 bases in an overlapping fashion, they alongwith the associated TF structure are passed into the Transformer-

DenseNet system which in turn provide position wise probability scores for the binding regions, **c)** The identified binding regions are highlighted by the box in the sequence. Download option for the result in the tabular format where the first column indicates sequence ID, second and third column represent start and end coordinates for binding region, and the fourth column represents the probability score for being a TF binding region or not. **d)** Also, the probability scores are represented in the form of interactive line plot depicting the distribution of probability score across position for the selected sequence. **e)** These probability scores are represented in the form of interactive bubble plot depicting the distribution of probability score on the sequences. One the PTFSpot's identified binding regions, user gets an option to run PTF-Vāc or TF-MoDISco to precisely identify the binding sites/motifs in the identified region.

Table 1. Brief list of some of the published tools for transcription factor binding site identification.

S.No.	Software	Algorithm	Encoding scheme	Biological relevance	Dataset	Species	Year	Webserver (W)/ Standalone (S)
1	**GkmSVM** **(Ghandi et al. 2016)**	SVM	gapped-kmer	Detection and modulation of functional sequence elements in regulatory DNA	ChIP-seq	Human	2014	S

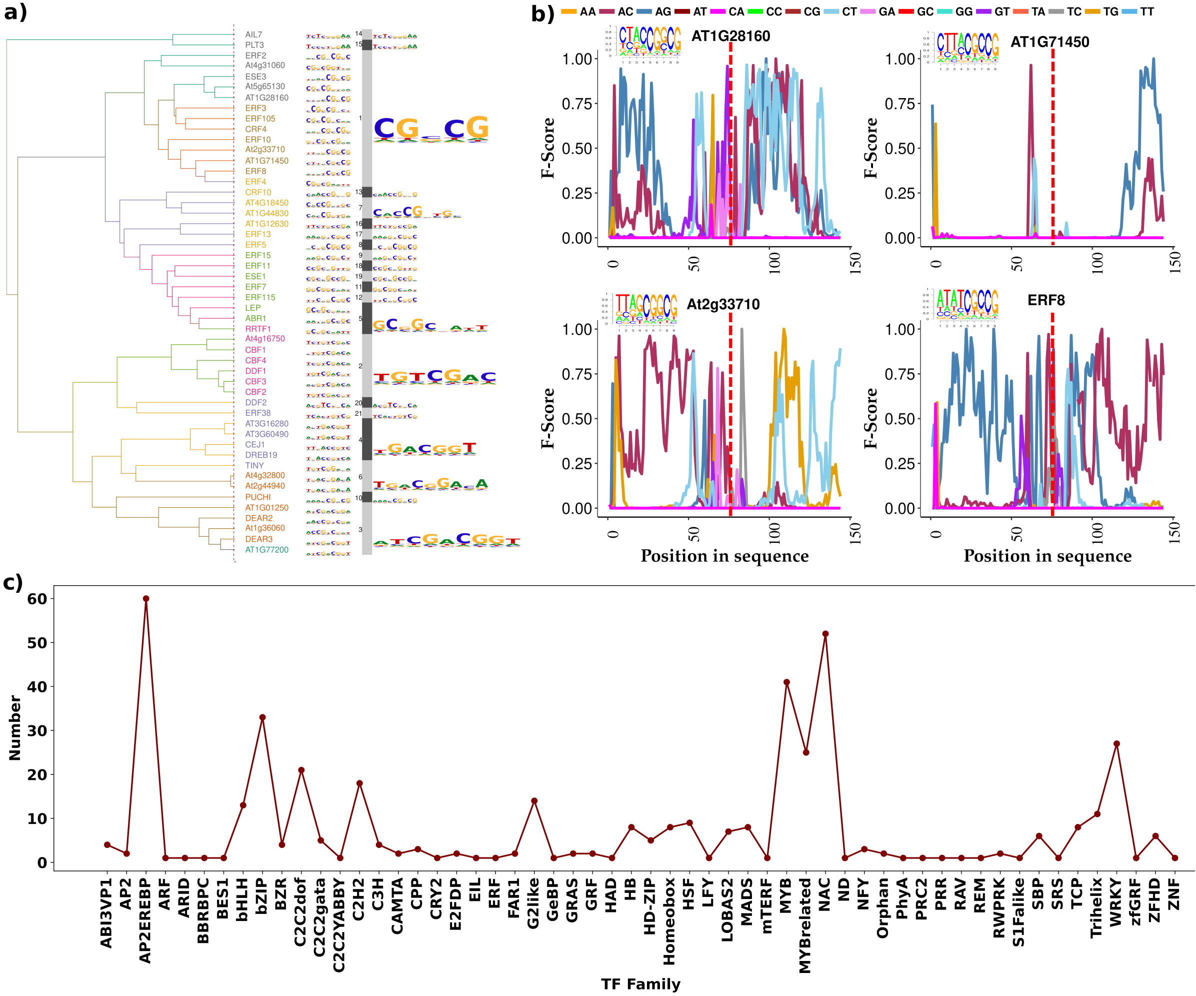
2	**pPromotif** **(Jha and Shankar 2014)**	Probabilistic modeling	position weight matrix, conservation index (Ci Value), and inter-nucleotide dependence	Plant transcription factor binding sites	AGRIS database and TBFS annotations in GenBank entries	*Arabidopsis thaliana*	2014	S
3	**DeepBind** **(Alipanahi et al. 2015)**	CNN	One-hot encode	Nucleic acid binding site prediction which can discover new patterns even when the locations of patterns within sequences are unknown	ChIP-seq	Human	2015	S
4	**KEGRU** **(Shen et al. 2018)**	Bi-GRU	*k*-mer embedding	Capture complex context information from the k-mer sequence	ChIP-seq	Human	2018	S
5	**k-mer grammar** **(Mejía-Guerra and Buckler**	Logistic regression	*k*-mers	Framework to exploit characteristic chromatin contexts and sequence	ChIP-seq, Mnase-seq	*Zea mays*	2019	S

	2019)			organization to classify regulatory regions based on sequence features - *k*-mers				
6	**DESSO (Yang et al. 2019)**	CNN	One-hot encoding	Identify motifs and identify TFBSs in both sequence and regional DNA shape features	ChIP-seq	Human	2019	S
7	**DeepRAM (Trabelsi et al. 2019)**	CNN/RNN	One-hot/*k*-mer embedding	Uses Different architectures using CNNs or RNNs to identify DNA/RNA sequence binding specificity	ChIP-seq	Human	2019	S
8	**WSCNNLSTM (Zhang et al. 2019)**	Multi-instance learning and hybrid neural network	*k*-mer embedding	Identify *in-vivo* protein-DNA binding	ChIP-seq	Human	2019	S

9	**SAResNet** **(Shen et al. 2021a)**	Self-attention mechanism +residual network	One-hot encoding	Identify DNA-protein binding and learning of the long-range dependencies from the DNA sequence	ChIP-seq	Human	2021	S
10	**SeqConv** **(Shen et al. 2021b)**	CNN	One-hot encoding	Identify more precise TF-DNA interaction regions in plants	ChIP-seq	*Zea mays*	2021	S
11	**TSPTFBS** **(Liu et al. 2021)**	CNN	One-hot encoding	TFBS prediction in plants	DAP-seq	*Arabidopsis thaliana*	2021	S
12	**DNABERT** **(Ji et al. 2021)**	BERT	*k*-mer encoding	Enables direct visualization of nucleotide-level importance and semantic relationship within input sequences for better interpretability and accurate identification of conserved sequence	ChIP-seq	Human	2021	S

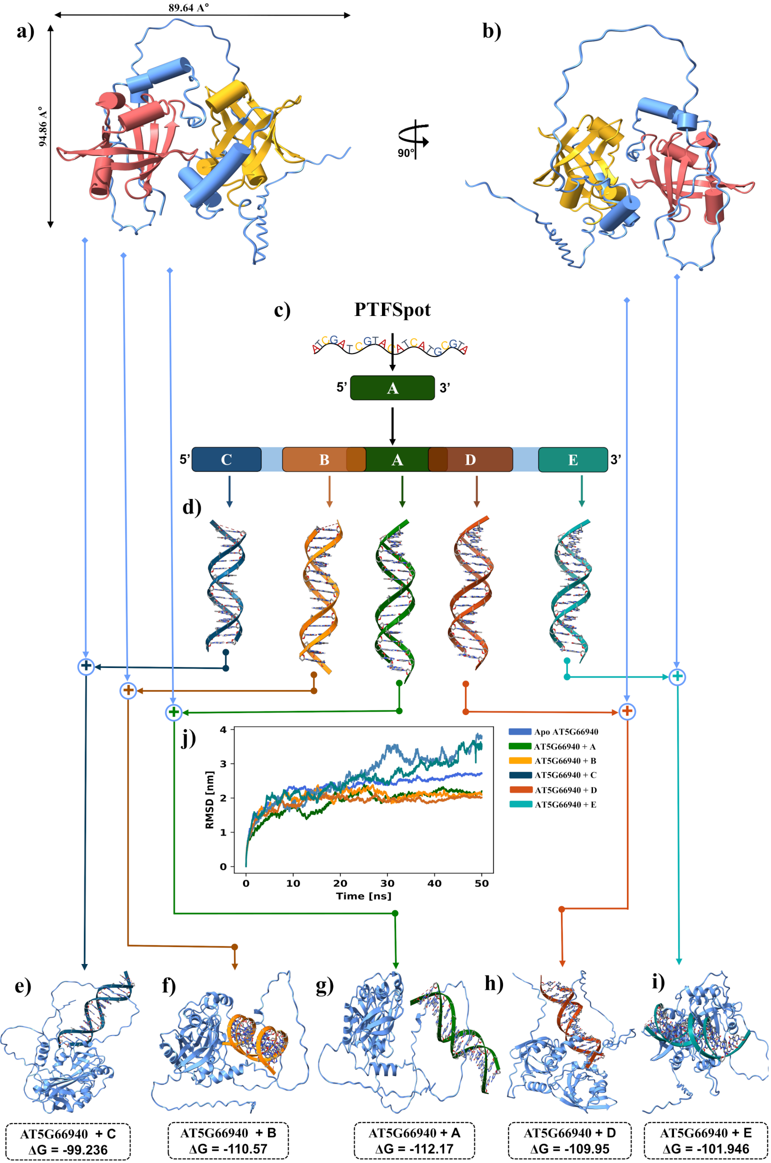
				motifs and functional genetic variant candidates.				
13	**Wimtrap** **(Rivière et al. 2022)**	XGBoost	Position weight matrices (PWMs)	Identify condition- or organ-specific *cis*-regulatory elements and TF gene targets, with a great flexibility regarding the input data	ChIP-seq	*Arabidopsis thaliana*	2022	S
14	**PlantBind** **(Yan et al. 2022)**	CNN+Bi-LSTM	One-hot encoding	Identify potential TFBSs of multiple TFs simultaneously	ChIP-seq	*Arabidopsis thaliana*	2022	S
15	**TSPTFBS 2.0** **(Cheng et al. 2023)**	DenseNet	One-hot encoding	TFBS prediction in plants	DAP-seq	*Arabidopsis thaliana*	2023	S/W

**HMM: Hidden Markov Model; SVM: Support Vector Machine; CNN: Convolutional neural network; LSTM: Long Short Term Memory; GRU: Gated Recurrent Unit; RNN: Recurrent Neural Network; BERT: Bidirectional Encoder Representations from Transformers

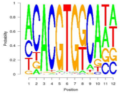
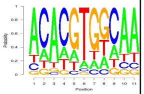
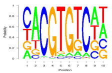
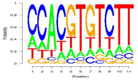
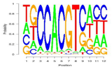
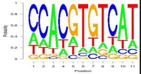
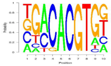
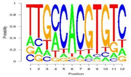
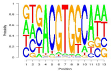
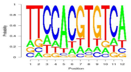
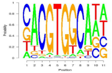
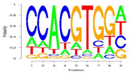
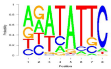
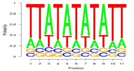
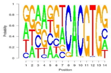
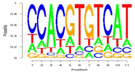
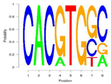
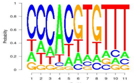
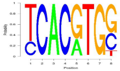
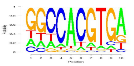
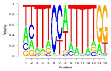
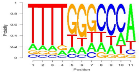

a)
AIL7
PLT3
ERF2
At4g31060
ESE3
At5g65130
AT1G28160
ERF3
ERF105
CRF4
ERF10
At2g33710
AT1G71450
ERF8
ERF4
CRF10
AT4G18450
AT1G44830
AT1G12630
ERF13
ERF5
ERF15
ERF11
ESE1
ERF7
ERF115
LEP
ABR1
RRTF1
At4g16750
CBF1
CBF4
DDF1
CBF3
CBF2
DDF2
ERF38
AT3G16280
AT3G60490
CEJ1
DREB19
TINY
At4g32800
At2g44940
PUCHI
AT1G01250
DEAR2
At1g36060
DEAR3
AT1G77200
14
15
1
13
7
16
17
8
9
18
19
11
12
5
2
20
21
4
6
10
3
b)
AA AC AG AT CA CC CG CT GA GC GG GT TA TC TG TT
AT1G28160
AT1G71450
At2g33710
ERF8
F-Score
1.00
0.75
0.50
0.25
0.00
0
50
100
150
Position in sequence
c)
Number
60
50
40
30
20
10
0
ABI3VP1
AP2
AP2EREBP
ARF
ARID
BBRBPC
BES1
bHLH
bZIP
BZR
C2C2dof
C2C2gata
C2C2YABBY
C2H2
C3H
CAMTA
CPP
CRY2
E2FDP
EIL
ERF
FAR1
G2like
GeBP
GRAS
GRF
HAD
HB
HD-ZIP
Homeobox
HSF
LFY
LOBAS2
MADS
mTERF
MYB
MYBrelated
NAC
ND
NFY
Orphan
PhyA
PRC2
PRR
RAV
REM
RWPRK
S1Falike
SBP
SRS
TCP
Trihelix
WRKY
zfGRF
ZFHD
ZNF
TF Family

Prime Motif
Binding Partner
ATHB23
REM19
BAM8
ABF2
At1g78700
At1g69690
FUS3
EIN3
AT3G09735
HSF3
GATA11
At1g25550
NLP7
ANAC004
SRS7
DEL2
AT3G12130
AGL15
Adof1
LBD13
ABR1
AT5G23930
ATHB13
MYB101
AT1G76880
AT3G42860
WRKY14
AT2G15740
bHLH10
SPL15
At1g19000
FRS9
ARF2
AT2G20110
AtGRF6
ANL2
BPC1
CRC
AT1G24250
RAV1
AT1G20910
3D Structures of TF
Atomic Coordinates
0.58 x
0.87 y
0.29 z
N C O H - - 24th atom
Amino Acid
Atoms Sequence
Atomic Coordinates
Input layer (300×24×3)
Conv2D layer with 3×3 kernels (300×24×32)
Conv2D + BN
MaxPooling2D
X10
BN + Relu + Conv2D + DO
Dense Block
Output
BN + Relu + Conv2D + BN
MaxPooling2D
Transition Block
Flatten
Concatenation
selu
Sigmoid
Adam
TFBS
0..............1
5′
-75
M
+75
3′
TCAATAGTATTGACACGTGTCAGTATTGACACGTATTGACAC
Word Generation
TC CA AA AT ... CA AC TCAAT CAATA ... GACAC TCAATAG CAATAGT ... TTGACAC
Tokenization and Embedding
Positional Embedding
P1 P2 P3 P4 P5 P6 ... P114 P115 P116 ... P468 P469
Query
Keys
Values
.A1 +
.A2+
.A3 +
.A14
MultiHeadAttention
Combined attention
Dropout Layer
Normalization Layer
relu
Dropout Layer
Normalization Layer
GlobalAveragePooling1D
Dropout Layer
relu

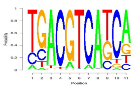
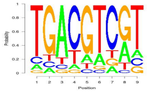
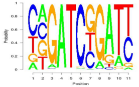
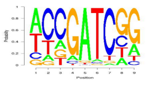
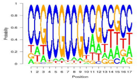
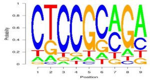
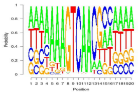
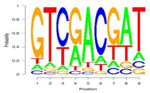
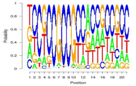
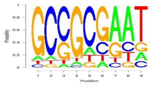
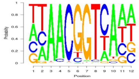
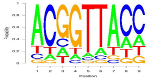
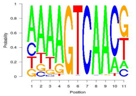
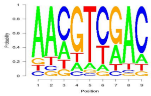
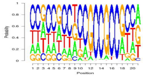
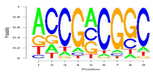
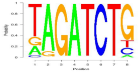
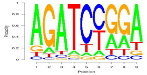
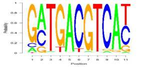
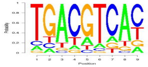
Ⓒ: Channel-wise Concatenation BN: BatchNormalization Layer DO: Dropout Layer

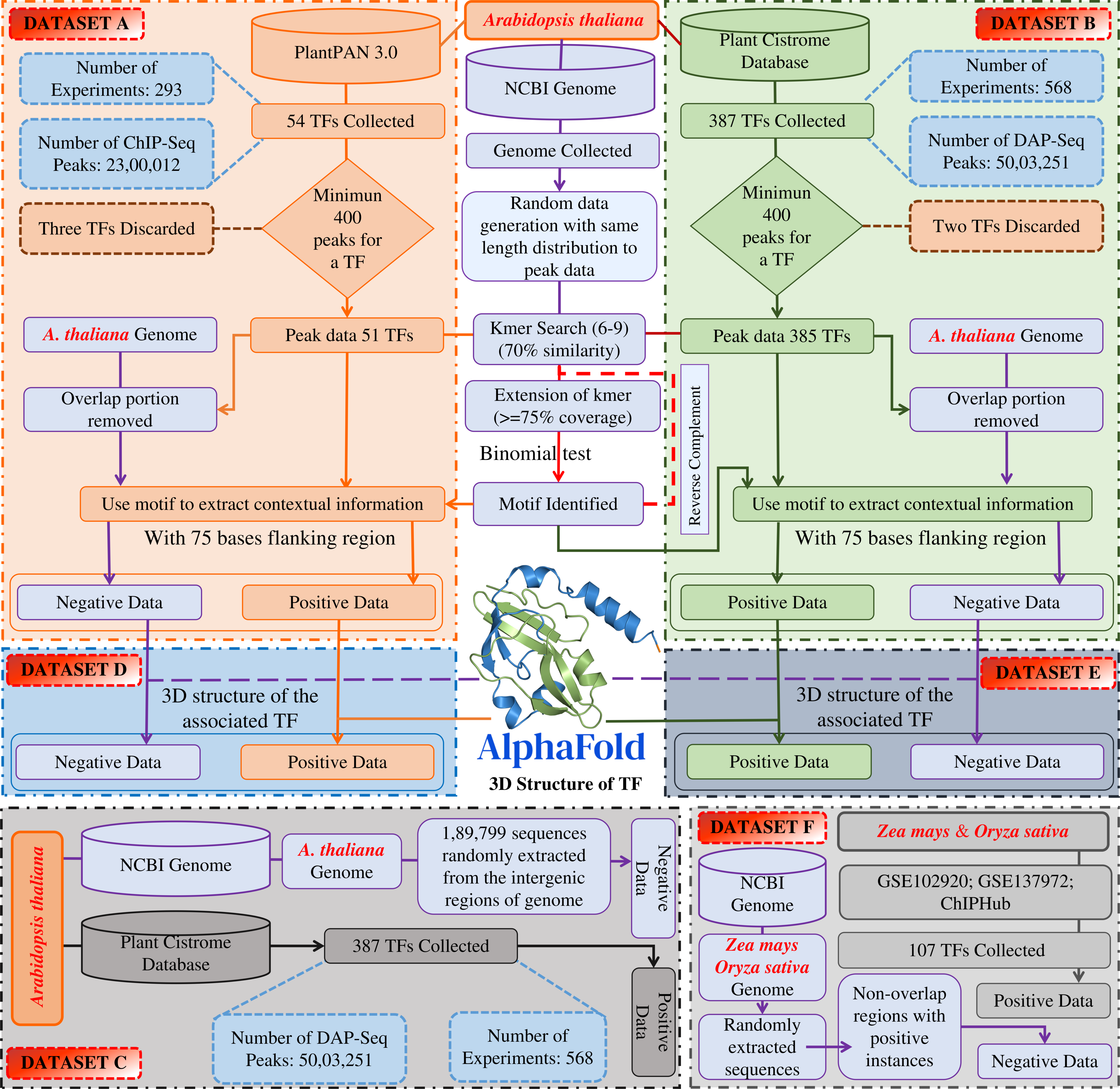

89.64 A°
a)
94.86 A°
b)
90°
c)
PTFSpot
ATCGATCGTACATCATGCGTA
5'
A
3'
5'
C
B
A
D
E
3'
d)
j)
RMSD [nm]
Time [ns]
0
1
2
3
4
0
10
20
30
40
50
Apo AT5G66940
AT5G66940 + A
AT5G66940 + B
AT5G66940 + C
AT5G66940 + D
AT5G66940 + E
e)
f)
g)
h)
i)
AT5G66940 + C
ΔG = -99.236
AT5G66940 + B
ΔG = -110.57
AT5G66940 + A
ΔG = -112.17
AT5G66940 + D
ΔG = -109.95
AT5G66940 + E
ΔG = -101.946

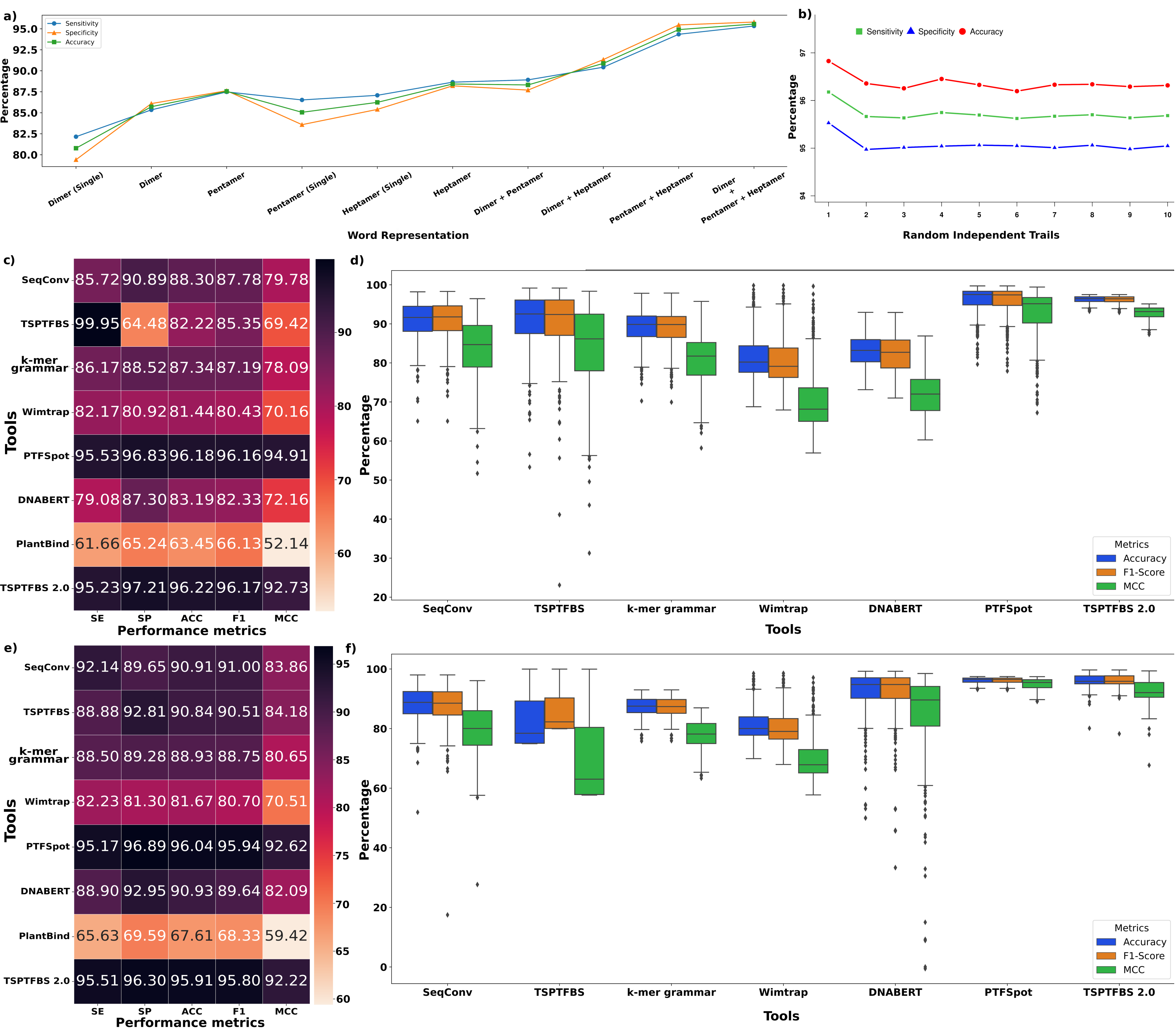
a)

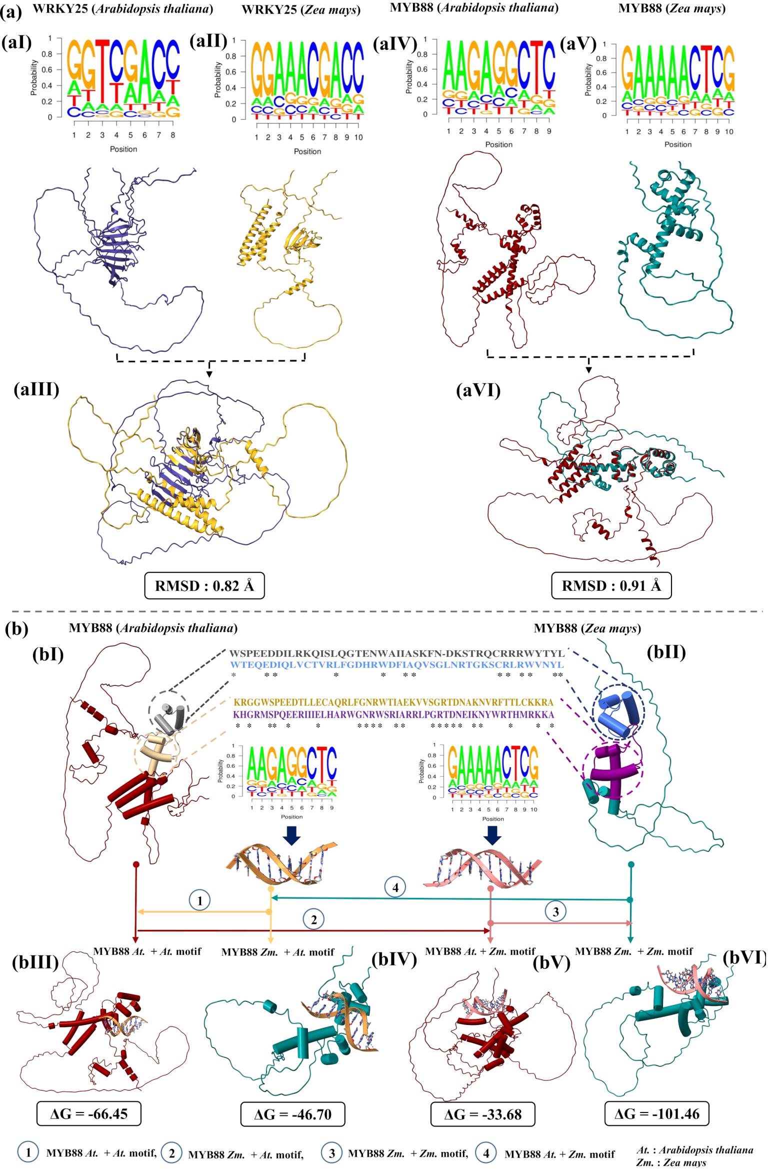
<table>
<tr><th>TF</th><th>JASPAR Motif</th><th>Prime Motif</th><th>Alignment</th><th>% of occurence for JASPAR Motif</th><th>% of occurence for Prime Motif</th></tr>
<tr><td>ABF1</td><td></td><td></td><td>ACACGTGGCA
||| |||| |||
ACACGTGGCA</td><td>72.9</td><td>96.5</td></tr>
<tr><td>ABF3</td><td></td><td></td><td>CACGTGTCTT
| ||| ||| |.|
CACGTGTCAT</td><td>75.7</td><td>96.1</td></tr>
<tr><td>ABF4</td><td></td><td></td><td>CCACGT
||||||
CCACGT</td><td>71.5</td><td>96.5</td></tr>
<tr><td>ABI5</td><td></td><td></td><td>TGCCACGTG
||. ||||||
TGACACGTG</td><td>75.9</td><td>87.2</td></tr>
<tr><td>GBF2</td><td></td><td></td><td>ACGTGTCA
| || ||.||
ACGTGGCA</td><td>73</td><td>95.2</td></tr>
<tr><td>GBF3</td><td></td><td></td><td>CCACGTG
| || ||||
CCACGTG</td><td>78</td><td>94.1</td></tr>
<tr><td>KAN1</td><td></td><td></td><td>ATATATT
|.| ||||
AGATATT</td><td>80</td><td>91.5</td></tr>
<tr><td>PIF1</td><td></td><td></td><td>CCACGTGTCAT
| || |||| .|.|
CCACGTGACCT</td><td>70.5</td><td>94.7</td></tr>
<tr><td>PIF4</td><td></td><td></td><td>CACGTGT
| | ||||.
CACGTGG</td><td>90</td><td>94.6</td></tr>
<tr><td>PIF5</td><td></td><td></td><td>CCACGTGA
| || ||| ||
CCACGTGA</td><td>90</td><td>96.3</td></tr>
<tr><td>SOC1</td><td></td><td></td><td>TTTTGG
||||||
TTTTGG</td><td>85</td><td>93.6</td></tr>
</table>

b)

<table>
<tr><th>TF</th><th>JASPAR Motif</th><th>Prime Motif</th><th>Alignment</th><th>% of occurence for JASPAR Motif</th><th>% of occurence for Prime Motif</th></tr>
<tr><td>TGA4</td><td></td><td></td><td>TGACGTCGT
|| |||||.|
TGACGTCAT</td><td>94.3</td><td>95.3</td></tr>
<tr><td>GATA19</td><td></td><td></td><td>CCGATCGG
|.||||||
CAGATCGG</td><td>95</td><td>96</td></tr>
<tr><td>LOB</td><td></td><td></td><td>CTCCGCAG
|.|| ||.|
CGCCGCCG</td><td>46.48</td><td>92.6</td></tr>
<tr><td>WRKY21</td><td></td><td></td><td>GTCGACG
|| |.|||
GTCAACG</td><td>10.41</td><td>85.7</td></tr>
<tr><td>LEP</td><td></td><td></td><td>GCCGCGA
| ||||.|
GCCGCCA</td><td>30.56</td><td>91.2</td></tr>
<tr><td>MYB44</td><td></td><td></td><td>ACGGTTA
| ||||.|
ACGGTCA</td><td>60.3</td><td>81.5</td></tr>
<tr><td>WRKY29</td><td></td><td></td><td>AACGTCGAC
| |.|||.||
AAAGTCAAC</td><td>87.35</td><td>85.3</td></tr>
<tr><td>ERF2</td><td></td><td></td><td>CCGACGGC
| || .||.|
CCGCCGCC</td><td>29.34</td><td>96.1</td></tr>
<tr><td>GATA11</td><td></td><td></td><td>AGATCCG
| ||||.|
AGATCTG</td><td>98.8</td><td>92.76</td></tr>
<tr><td>TGA9</td><td></td><td></td><td>TGACGTCA
|||| ||||
TGACGTCA</td><td>83.8</td><td>94.78</td></tr>
</table>


DATASET A
PlantPAN 3.0
Arabidopsis thaliana
NCBI Genome
Plant Cistrome Database
DATASET B
Number of Experiments: 293
54 TFs Collected
Number of ChIP-Seq Peaks: 23,00,012
Genome Collected
387 TFs Collected
Number of Experiments: 568
Number of DAP-Seq Peaks: 50,03,251
Three TFs Discarded
Minimun 400 peaks for a TF
Random data generation with same length distribution to peak data
Minimun 400 peaks for a TF
Two TFs Discarded
A. thaliana Genome
Peak data 51 TFs
Kmer Search (6-9) (70% similarity)
Peak data 385 TFs
A. thaliana Genome
Overlap portion removed
Extension of kmer (>=75% coverage)
Reverse Complement
Overlap portion removed
Binomial test
Use motif to extract contextual information
Motif Identified
Use motif to extract contextual information
With 75 bases flanking region
With 75 bases flanking region
Negative Data
Positive Data
Positive Data
Negative Data
DATASET D
DATASET E
3D structure of the associated TF
3D structure of the associated TF
Negative Data
Positive Data
AlphaFold
3D Structure of TF
Positive Data
Negative Data
Arabidopsis thaliana
NCBI Genome
A. thaliana Genome
1,89,799 sequences randomly extracted from the intergenic regions of genome
Negative Data
Plant Cistrome Database
387 TFs Collected
Positive Data
Number of DAP-Seq Peaks: 50,03,251
Number of Experiments: 568
DATASET C
DATASET F
Zea mays & Oryza sativa
NCBI Genome
GSE102920; GSE137972; ChIPHub
107 TFs Collected
Zea mays Oryza sativa Genome
Positive Data
Randomly extracted sequences
Non-overlap regions with positive instances
Negative Data


a)
Sensitivity
Specificity
Accuracy
Percentage
95.0
92.5
90.0
87.5
85.0
82.5
80.0
Dimer (Single)
Dimer
Pentamer
Pentamer (Single)
Heptamer (Single)
Heptamer
Dimer + Pentamer
Dimer + Heptamer
Pentamer + Heptamer
Dimer + Pentamer + Heptamer
Word Representation
b)
Sensitivity
Specificity
Accuracy
Percentage
97
96
95
94
1
2
3
4
5
6
7
8
9
10
Random Independent Trails
c)
Tools
SeqConv 85.72 90.89 88.30 87.78 79.78
TSPTFBS 99.95 64.48 82.22 85.35 69.42
k-mer grammar 86.17 88.52 87.34 87.19 78.09
Wimtrap 82.17 80.92 81.44 80.43 70.16
PTFSpot 95.53 96.83 96.18 96.16 94.91
DNABERT 79.08 87.30 83.19 82.33 72.16
PlantBind 61.66 65.24 63.45 66.13 52.14
TSPTFBS 2.0 95.23 97.21 96.22 96.17 92.73
SE
SP
ACC
F1
MCC
Performance metrics
90
80
70
60
d)
Percentage
100
90
80
70
60
50
40
30
20
SeqConv
TSPTFBS
k-mer grammar
Wimtrap
DNABERT
PTFSpot
TSPTFBS 2.0
Tools
Metrics
Accuracy
F1-Score
MCC
e)
Tools
SeqConv 92.14 89.65 90.91 91.00 83.86
TSPTFBS 88.88 92.81 90.84 90.51 84.18
k-mer grammar 88.50 89.28 88.93 88.75 80.65
Wimtrap 82.23 81.30 81.67 80.70 70.51
PTFSpot 95.17 96.89 96.04 95.94 92.62
DNABERT 88.90 92.95 90.93 89.64 82.09
PlantBind 65.63 69.59 67.61 68.33 59.42
TSPTFBS 2.0 95.51 96.30 95.91 95.80 92.22
SE
SP
ACC
F1
MCC
Performance metrics
95
90
85
80
75
70
65
60
f)
Percentage
100
80
60
40
20
0
SeqConv
TSPTFBS
k-mer grammar
Wimtrap
DNABERT
PTFSpot
TSPTFBS 2.0
Tools
Metrics
Accuracy
F1-Score
MCC


(a)
WRKY25 (*Arabidopsis thaliana*)
WRKY25 (*Zea mays*)
MYB88 (*Arabidopsis thaliana*)
MYB88 (*Zea mays*)
(aI)
(aII)
(aIV)
(aV)
Probability
Position
(aIII)
(aVI)
RMSD : 0.82 Å
RMSD : 0.91 Å
(b)
MYB88 (*Arabidopsis thaliana*)
MYB88 (*Zea mays*)
(bI)
(bII)
WSPEEDDILRKQISLQGTENWAIIASKFN-DKSTRQCRRRWYTYL
WTEQEDIQLVCTVRLFGDHRWDFIAQVSGLNRTGKSCRLRWVNYL
KRGGWSPEEDTLLECAQRLFGNRWTIAEKVVSGRTDNAKNVRFTTLCKKRA
KHGRMSPQEERIIIELHARWGNRWSRIARRLPGRTDNEIKNYWRTHMRKKA
1
2
3
4
MYB88 *At.* + *At.* motif
MYB88 *Zm.* + *At.* motif
MYB88 *At.* + *Zm.* motif
MYB88 *Zm.* + *Zm.* motif
(bIII)
(bIV)
(bV)
(bVI)
ΔG = -66.45
ΔG = -46.70
ΔG = -33.68
ΔG = -101.46
① MYB88 *At.* + *At.* motif, ② MYB88 *Zm.* + *At.* motif, ③ MYB88 *Zm.* + *Zm.* motif, ④ MYB88 *At.* + *Zm.* motif
At. : *Arabidopsis thaliana*
Zm. : *Zea mays*

a)
Percentage
100.0
97.5
95.0
92.5
90.0
87.5
85.0
82.5
Sensitivity
Specificity
Accuracy
F1-Score
MCC
i)
ii)
iii)
kmer grammar
SeqConv
TSPTFBS
DNABERT
PTFSpot
TSPTFBS 2.0
b)
Sensitivity
Specificity
Accuracy
MCC
Percentage
LHY1
MYB56
MYB62
MYB81
MYB88
WRKY25
Transcription Factor
c)
Sensitivity
Specificity
Accuracy
MCC
Density
Tool
Maize
Rice
Percentage

a
CSIR-INDIA
TFSpot
HIMCoB
The Computing DNA of Himalayas
Home
Download Dataset
Standalone Version
About us
paste an example
Browse...
No file selected.
Reset
Run
or
b
Transformer-DenseNet System
c
Download
show protein sequence
PTF-Vāc
TF-Modisco
d
seq1
Probability Score
Sequence Position
0
0.2
0.4
0.6
0.8
0
5
10
15
20
25
30
e
Score
Sequence Unique ID
0
0.2
0.4
0.6
0.8
1
1.2
1
2
3
4
5
6
7
8
9
10
Sequence
seq1
seq10
seq11
seq12
seq2
seq3
seq4
seq5
seq6
seq7
seq8
seq9

PTFSpot: A Deep Co-learning system for impeccably accurate transcription factor binding region discovery in plants

Enter a FASTA sequence **paste an example** (**Minimum sequence length required: 160 bases**)

About PTFSpot:
Unlike animals, variability in transcription factors (TF) and their binding sites (TFBS) across the plants species is a major problem which most of the existing TFBS finding software fail to tackle, rendering them hardly of any use. This limitation has resulted into underdevelopment of plant regulatory research and rampant misuse of Arabidopsis models, generating misleading results. Here we report a ground-breaking transformers based deep-learning approach, PTFSpot, which learns from TF structures and their binding sites to bring a universal TF-DNA interaction model. During a series of extensive bench-marking studies, it not only outperformed the existing software by >30% lead, but also delivered consistently >90% accuracy even for those species and TF families which were never encountered during model building process. PTFSpot makes it possible now to accurately annotate TFBS across novel plant genomes even in the total lack of any TF information. PTFSpot is freely available as stand-alone source codes at https://scbb.ihbt.res.in/PTFSpot/index.php.

```
>seq1
AAGAGAGGGAAAAATGTGACACGTGTGAAAACAGTTGTGCATTGTAAGAGACGAAAAAACAAAGGTTAAA
AGCGACAGGCTGGAAAAAAGAGGATGATGACGTGGCAGATATTGATTGGTGTTTGATTTTATATAGTGGC
TAAAAAGGCTATAATGAGCTTTCAGGGTGGTAGGCCATTTTTCGTCTCTCACGATTGATGA
>seq2
GATATAACGTAGCGAAGTGTTGCCACAGACGACGGAGAGACAAAAGATCGCTGACAAAGAGAGGATCTTT
AGAAACTGACGGCGAGATAAGATGTGCTGACGTGGCACGATTCCGGCGATATTGATTACGGCGGAGCTAT
AGGCGGCGCATGATAACATGTGTGTCATGGTGAGGAAGATTGGGTAACGGAAACCGTAGTA
>seq3
TATAAACGCGAGATAAAAAAAAAAAAAAAAGAGAGATTTTTGGGCTTAGAGTGAGGAGAGGAACGTTACTG
TGTGTGGTGTTACACGTAGGGAATTGCCACGTCATCATTTGTGTGATTCTGTTTTCGTTTTTGTTTATCT
AAAAATTAATTTAAGATCACGTCATAATTAGTGTTTTTATTTTAGATAACACTAATGATGG
>seq4
TTTTTAATCACAATATTGGATTGGGAAAGAAAAACTCTGGGTAGAATAAAATTAAATAAAAAAGCATTGA
CGTGGGTTGCAACCAGAGAAAAGTGATGACGTGTCATGTCGTGATTGGCTGAATATTATTGACGAAAGAA
GAAAAAAACGCTGTGTCTATCTCTCGACTCTCTCTCTCTCTGGCGATGGCTATCAAACCGG
>seq5
AAAAGGACATTTAATTATGAGTATGAATCTTTTATTATTACATAGATTTTTTAAAAAAGTGAAAAACAGA
AGAGAAAATAATAATGAAGAGAAATGCCACGTCATCCTAAGTTATACAAAACGCCCAATCAGCCTTCAAA
AACAACACCGTTGGGTCTTCAGGAGAGGGCTCTCTTTGTATCTTTATACTTTGTTTACAAA
```

Size limit: 5MB. For longer sequences kindly use the standalone version.

or

Please specify a file: Browse... No file selected.

Upload TF protein (PDB format): Browse... No file selected.

Click to download your results: Download

The user provided this protein sequence: show protein sequence

Click here to see motif discovered by PTF-Vāc: PTF-Vāc

Click here to see motif discovered by TF-Modisco: TF-Modisco

>seq1 [27-187]

AAGAGAGGGAAAAATGTGACACGTGTGAAAACAGTTGTGCATTGTAAGAGACGAAAAAACAAAGGTTAAAAGCGACAGGCTGGAAAAAAGAGGATGATGACGTGGCAGATATTGATTGGTGTTTGATTTTATATAGTGGCTAAAAAGGCTATAATGAGCTTTCAGGGTGGTAGGCCATTTTTCGTCTCTCACGATTGATGA

>seq10 [17-177]

GTTATGCTTTTTAAATCAAAACTCGTTTTGCTCTAAATTCTATAACTAAACTAAACACGTGGGTCTATTTGATTAGTATTATTATTTCTTTAAATGATGACGTGGACAGTTAAATGGATTTGGCAAATCCCCTGCAGATAACCAACATTATTCATTGGAATGCATAGTGATGGAATATAAAGGATGTGCCTAATTATTTTG

>seq11 TF Binding region not found

>seq12 TF Binding region not found

>seq2 [41-201]

GATATAACGTAGCGAAGTGTTGCCACAGACGACGGAGAGACAAAAGATCGCTGACAAAGAGAGGATCTTTAGAAACTGACGGCGAGATAAGATGTGCTGACGTGGCACGATTCCGGCGATATTGATTACGGCGGAGCTATAGGCGGCGCATGATAACATGTGTGTCATGGTGAGGAAGATTGGGTAACGGAAACCGTAGTA

>seq3 [3-163,15-175,16-176,18-178]

TATAAACGCGAGATAAAAAAAAAAAAAAGAGAGATTTTTGGGCTTAGAGTGAGGAGAGGAACGTTACTGTGTGTGGTGTTACACGTAGGGAATTGCCACGTCATCATTTGTGTGATTCTGTTTTCGTTTTTGTTTATCTAAAAATTAATTTAAGATCACGTCATAATTAGTGTTTTTATTTTAGATAACACTAATGATGG

>seq4 [30-190,31-191]

TTTTTAATCACAATATTGGATTGGGAAAGAAAAACTCTGGGTAGAATAAAATTAAATAAAAAAGCATTGACGTGGGTTGCAACCAGAGAAAAGTGATGACGTGTCATGTCGTGATTGGCTGAATATTATTGACGAAAGAAGAAAAAAACGCTGTGTCTATCTCTCGACTCTCTCTCTCTCTGGCGATGGCTATCAAACCGG

>seq5 TF Binding region not found

Select a sequence to see the scoring distribution across the sequence: seq1

Your selected sequence is : **seq1**

Probability score distribution across position for the selected sequence

Complete result plot for every sequence. Bubble size corresponds to the score.
Hovering the mouse over the bubble will show the sequence location and other related details.