OXFORD

Briefings in Bioinformatics, 2024, **25(4)**, bbae324
https://doi.org/10.1093/bib/bbae324
Problem Solving Protocol

PTFSpot: 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, Himachal Pradesh 176061, India
[2]Academy of Scientific and Innovative Research (AcSIR), Ghaziabad, Uttar Pradesh 201002, India

*Corresponding author. E-mail: Email-ravish@ihbt.res.in

Abstract

Unlike animals, variability in transcription factors (TFs) and their binding regions (TFBRs) across the plants species is a major problem that most of the existing TFBR 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 revolutionary transformers-based deep-learning approach, PTFSpot, which learns from TF structures and their binding regions' co-variability to bring a universal TF-DNA interaction model to detect TFBR with complete freedom from TF and species-specific models' limitations. During a series of extensive benchmarking studies over multiple experimentally validated data, it not only outperformed the existing software by >30% lead but also delivered consistently >90% accuracy even for those species and TF families that were never encountered during the model-building process. PTFSpot makes it possible now to accurately annotate TFBRs across any plant genome even in the total lack of any TF information, completely free from the bottlenecks of species and TF-specific models.

Keywords: transcription factor; transcriptional regulation; deep learning; DenseNet; protein modelling; transformers

Introduction

TF-binding regions is central to understanding the transcriptional regulation across the genome. The rise of high-throughput technologies to detect such TF-deoxyribonucleic acid (DNA) interactions like protein binding microarrays (PBMs) [1], ChIP-Seq [2], and its various variants like ChIP-exo [3] and DAP-seq [4] has resulted into an explosion of DNA-binding region data for various TFs [5]. To this date, there are ∼128,467 ChIP-seq experiments reported at Gene Expression Omnibus (GEO)/Sequence Read Archive (SRA) for human alone. However, capturing all TF-DNA interactions through such experiments in any organism itself is a costly and impractical affair. One essentially requires some able computational approach to identify such TFBRs.

Unlike animals where human has been the main focus, plants have enormous number of species that define an extremely huge search space for possible experiments to detect TF-DNA interactions. If one compares the status of developments in plants with respect to animals, a huge gap is evident with hardly eight species of plants sequenced for their TFs' binding regions, covering merely ∼700 ChIP/DAP-seq experiments for selected few TFs, mostly related to *Arabidopsis thaliana* and *Zea mays*. This lag in experimental data is equally reflected in terms of software resources and algorithms development for plant TFBR discovery. While for animal/human, several software have been developed, there has been very limited development for plants. Table 1 lists some software available for animals and plants where clear skew is visible (more information available in Supplementary Table 1). Therefore, it becomes urgent to develop computational approaches that could model TF-DNA interactions accurately for plants system, which may also reduce the dependence on costly binding experiments to a great extent.

The existing software tools are overtly dependent upon the old school of motif discovery and user-defined motifs, while reports suggest that TF binding is more about context and surroundings [17, 22–24]. The context and surroundings around active binding motifs are defined by local sequence and shape preferences in highly specific manner. The motif-finding step itself is heavily dependent upon binding experiments like PBM and DAP/ChIP-seq, on whose results, binding motif is defined for any given TF. Many TFs share a similar binding motif but yet differ in their binding due to local surroundings, shape, and contexts (Fig. 1a and b) [10, 18, 20, 24]. Binding of a TF to its DNA target involves a prior step of local scanning of the target region in the range of around 90–150 bases. The sequence compositions, degenerate consensus sequences, and cooperative motifs in the entire region were found contributing to the final halting of the TF around its target gene [25, 26]. Further to this, the choice of negative datasets with most of the software have been very relaxed as they randomly pick

Received: March 19, 2024. **Revised:** June 7, 2024. **Accepted:** June 19, 2024

Table 1. Brief list of some of the published tools for TF-binding site identification

S.No.	Software	Algorithm	Encoding scheme	Biological relevance	Dataset	Species	Year	Webserver (W)/ Standalone (S)
1	GkmSVM [6]	SVM	gapped-kmer	Detection and modulation of functional sequence elements in regulatory DNA	ChIP-seq	Human	2014	S
2	pPromotif [7]	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 [8]	CNN	One-hot encode	Nucleic acid binding site prediction that can discover new patterns even when the locations of patterns within sequences are unknown	ChIP-seq	Human	2015	S
4	KEGRU [9]	Bi-GRU	*k*-mer embedding	Capture complex context information from the *k*-mer sequence	ChIP-seq	Human	2018	S
5	*k*-mer grammar [10]	Logistic regression	*k*-mers	Framework to exploit characteristic chromatin contexts and sequence organization to classify regulatory regions based on sequence features—*k*-mers	ChIP-seq, Mnase-seq	*Zea mays*	2019	S
6	DESSO [11]	CNN	One-hot encoding	Identify motifs and identify TFBSs in both sequence and regional DNA shape features	ChIP-seq	Human	2019	S
7	DeepRAM [12]	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 [13]	Multi-instance learning and hybrid neural network	*k*-mer embedding	Identify *in vivo* protein-DNA binding	ChIP-seq	Human	2019	S
9	SAResNet [14]	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	AgentBind [15]	CNN + BiLSTM	One-hot encoding	Score the importance of context sequences	ChIP-seq	Human	2021	S
10	SeqConv [16]	CNN	One-hot encoding	Identify more precise TF-DNA interaction regions in plants	ChIP-seq	*Z. mays*	2021	S
11	TSPTFBS [17]	CNN	One-hot encoding	TFBS prediction in plants	DAP-seq	*A. thaliana*	2021	S
12	DNABERT [18]	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 motifs and functional genetic variant candidates.	ChIP-seq	Human	2021	S
13	Wimtrap [19]	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	*A. thaliana*	2022	S
14	PlantBind [20]	CNN + Bi-LSTM	One-hot encoding	Identify potential TFBSs of multiple TFs simultaneously	ChIP-seq	*A. thaliana*	2022	S
15	TSPTFBS 2.0 [21]	DenseNet	One-hot encoding	TFBS prediction in plants	DAP-seq	*A. 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.

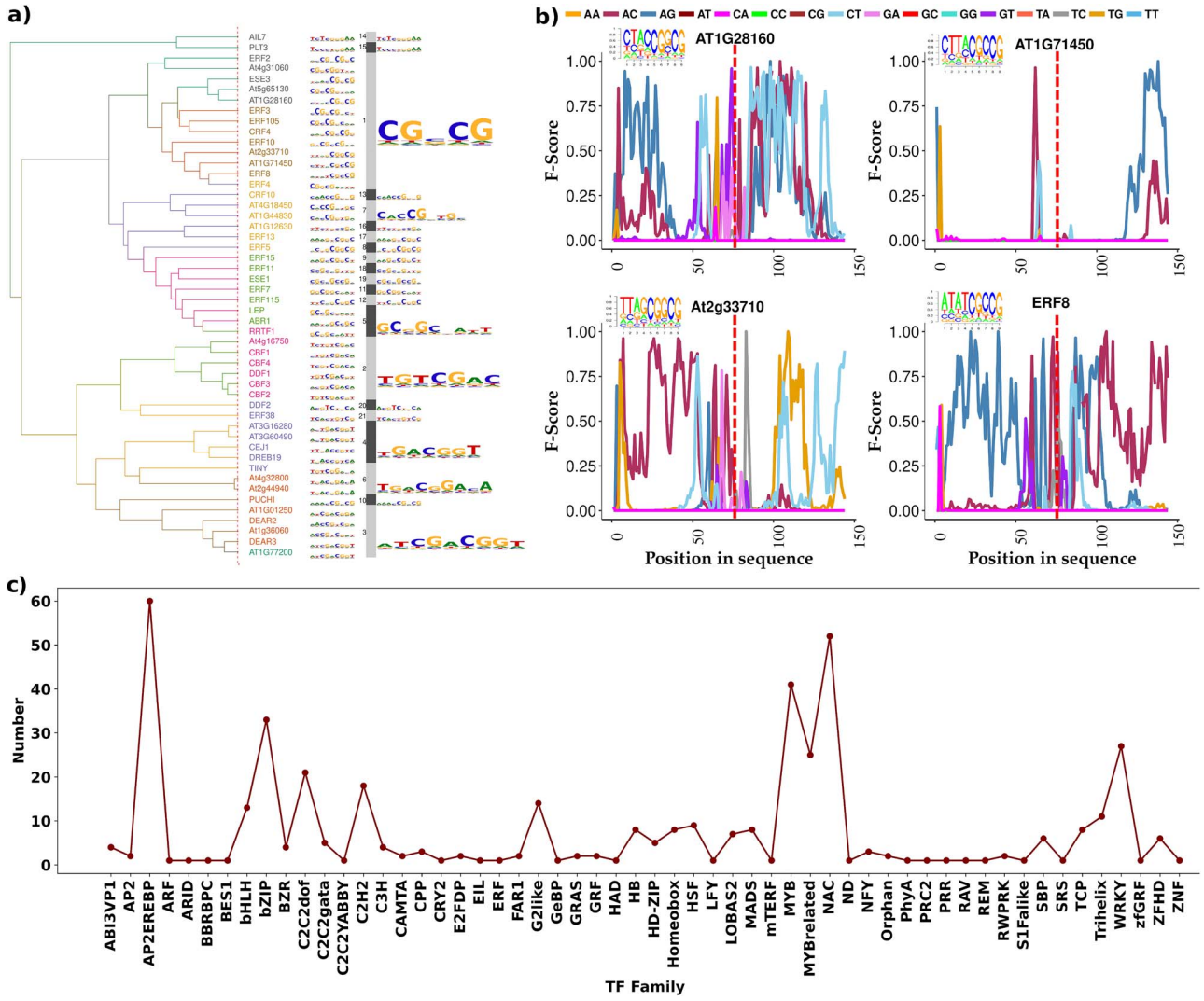

Figure 1. Motif and data. (a) Clustering plot of the prime motifs for TF family AP2EREBP. The TFs share similarity in their prime motifs. (b) Different characteristics spectra for TFs belonging to same family, same cluster, and sharing similar prime motif. This highlights the importance of context. Despite 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 families.

sequences and give too much weight to the consensus motif, which can actually occur even in the nonbinding regions, creating weak datasets on which learning have been done so far. Thus, giving so much weight to some motif alone to determine binding of a TF to DNA is itself a misplaced practice.

The most alarming matter is the fact that the plant biologists are overtly using TF and species-specific model like those developed for *Arabidopsis*-specific TFs to report TF DNA binding in other species, 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. Plant genomes, in general, exhibit enormous variability [27, 28], and TFs and their binding regions display large degree of variability across the plant species [29–32]. Thus, what may be working in *A. thaliana* may not work in other plant species, and vice versa. Understanding and tackling the variability of plant TFs and their TFBRs is the main challenge to develop plant-specific TFBR models where most of the currently existing software almost fail.

With all this foundation and understanding of the challenges, we here present a unique and universal approach, PTFSpot, to detect TFBRs across plant genomes based on the following principles:

1) Instead of overtly relying upon motifs, identify the most significant motifs specific to any TF and use it as the seed/anchor to identify most significant flanking regions for additional information. This is because the TF scans a whole local region before halting at any given location, a process that depends a lot upon the flanking regions' environment [25].

2) As already mentioned above, the enriched motifs are just one important feature. However, not all such motifs are bound by TF, as the flanking regions' environment is also important. This information could become strongly discriminating if the negative set considers such unbound region's motifs and flanking region information for them also. Therefore, use the significant motifs found in ChIP/DAP-seq data (positive datasets) to locate them in the unbound regions for the given experimental condition. This creates a highly confusing realistic negative dataset where overt importance of motifs is downplayed.

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). Doing this may boost the discriminating power through sequence and structural information of the flanking regions.

4) Apply Transformers-like state-of-the-art deep-learning algorithms that 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) Finding TFBR in cross-species manner failed so far because the existing approaches wrongly assume that binding preferences of a TF remains static across the species while overtly relying on some enriched motif found in one or few species. While the fact is that the TF sequence, structure, and its binding preferences vary across the species and even with their splice variants. Therefore, for reliable and universal TFBR discovery, one needs to learn the covariability between the TF itself and its binding preferences. If the available TFs and their binding preferences are learned this way together, they may bring a universal single model for TF:DNA interaction where covariability between structures and sequences could answer binding preferences for any TF. Thus, learn covariability between TF structure and binding regions while assuming all TFs under one hypothetical TF that keeps changing its composition and structure (the corresponding TF) according to which the preferred binding region also changes. This learning even from a single species may bring a universal model, which can work across any species even for never-seen-before TFs, as variability and interaction relationship are learned. Thus, learning this way even on *A. thaliana* data alone, which is also the most abundant one, would turn to be a boon.

Based on these principles, we have developed a Transformer-DenseNet Deep-Learning universal system while learning from ChIP/DAP-Seq binding data for 436 TFs and their corresponding 3D structural details. A highly extensive benchmarking study was carried out with three different experimentally validated datasets as well as never-seen-before species-specific experimental binding datasets to test the universality of PTFSpot. The results have been groundbreaking 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 plant species and TFs. This is something that has never been witnessed before and stands revolutionary as it will empower to detect the TFBRs across any plant species for any TF with impeccable accuracy, and may even bypass the need of costly experiments like DAP-seq to detect the TFBRs.

Materials and methods

Data retrieval

ChIP/DAP-seq peak data for 436 TFs spanning 5 753 198 distinct peaks were retrieved from PlantPAN3.0 (54 TFs) and Plant Cistrome Database (387 TFs) [33, 34]. Genomic sequences were extracted from the peak coordinates (Supplementary Table 2 Sheet 1). All details on data retrieval are illustrated in Fig. 2.

Approach to identifying motif seeds candidates and anchoring significant seeds

For the 436 TFs from *A. thaliana*, ChIP/DAP-seq peak regions were scanned to identify prevalent 6-mer seed candidates with ≥70% identity, based on previous observations [33–36]. Every sequence was initially represented in the form for overlapping hexameric seeds, which were used to scan across the sequences for the most similar seed regions among themselves and were accordingly piled up against each other. Enrichment of *k*-mers was determined by computing their occurrence probabilities in peak data relative to a random genomic background model following the peak length distribution. The seeds represented in ≥75% of peaks with significant enrichment (binomial test, $P < .01$) were selected and iteratively extended bidirectionally while retaining instances with ≥70% identity. For every extension step, statistical significance, the identity, and coverage criteria were repeatedly evaluated. The final motifs obtained after these steps were called as the anchoring seeds, which satisfied the two criteria: ≥75% abundance ($P < 0.01$) in the peak data and ≥70% identity. Most enriched ones among them were called the primes. They were used as the anchors to derive the flanking region contexts for the datasets creation. The process was done for both the strands. This motif discovery approach has been introduced by us previously [37], while further details are given in the Supplementary Information Materials and Methods section.

Dataset creation

To generate positive datasets, peak data sequences were transformed into instances once the motifs were anchored for each TF in the provided peaks data. ±75 bases in both directions from the ends of the prime motif regions were taken to produce the instances.

To form the negative datasets, regions reflected in the peak data were removed, leaving only unrepresented regions for selection. These regions were scanned for prime and reverse complementary motifs similar to the positive instances, flanked by 75 bases for context [38]. Positive and negative instances were pooled at a 1:1 ratio, without overlap.

Datasets "A" and "B" comprised positive instances derived from ChIP-seq and DAP-seq experiments, respectively, on *A. thaliana* TFs. Dataset "C" leveraged DAP-seq data for 387 TFs from the Plant Cistrome Database [17, 21]. To develop universal models capturing TF-DNA interactions, Datasets "D" and "E" further incorporated the 3D structures of the corresponding TFs from AlphaFold2 [39]. Dataset "E", with 325 *Arabidopsis* TFs, was used for training and evaluation. An independent test Dataset "F" was curated to assess cross-species generalization, comprising 117 TFs from *Z. mays* (93) and *Oryza sativa* (24) with over 2.2 million ChIP-seq peaks from public data repositories [40]. For all the datasets, every possible overlap and redundancies (full or partial) were screened out. Further details on dataset construction are provided in the Supplementary Information Materials and Methods section and Fig. 2.

Word representations and tokenization for sequence data

Utilizing a four-bases DNA alphabet yielded 16 unique 2-mer words, 1024 unique 5-mer words, and 16 384 unique 7-mer words. Dinucleotides, pentamers, and heptamers encode composition, base stacking, shape, and regional motif information [22, 37, 41]. The resulting word representations were tokenized by assigning a distinct integer to each unique word, creating an input vector of 469 tokenized words. These numeric token embeddings served as the inputs to the Transformer encoder part. The TensorFlow (Keras) tokenizer class was employed to implement the tokenization procedure.

Implementation of the transformers

Each tokenized sequence was converted into a 2D word embedding matrix, with rows determined by the encoding vector

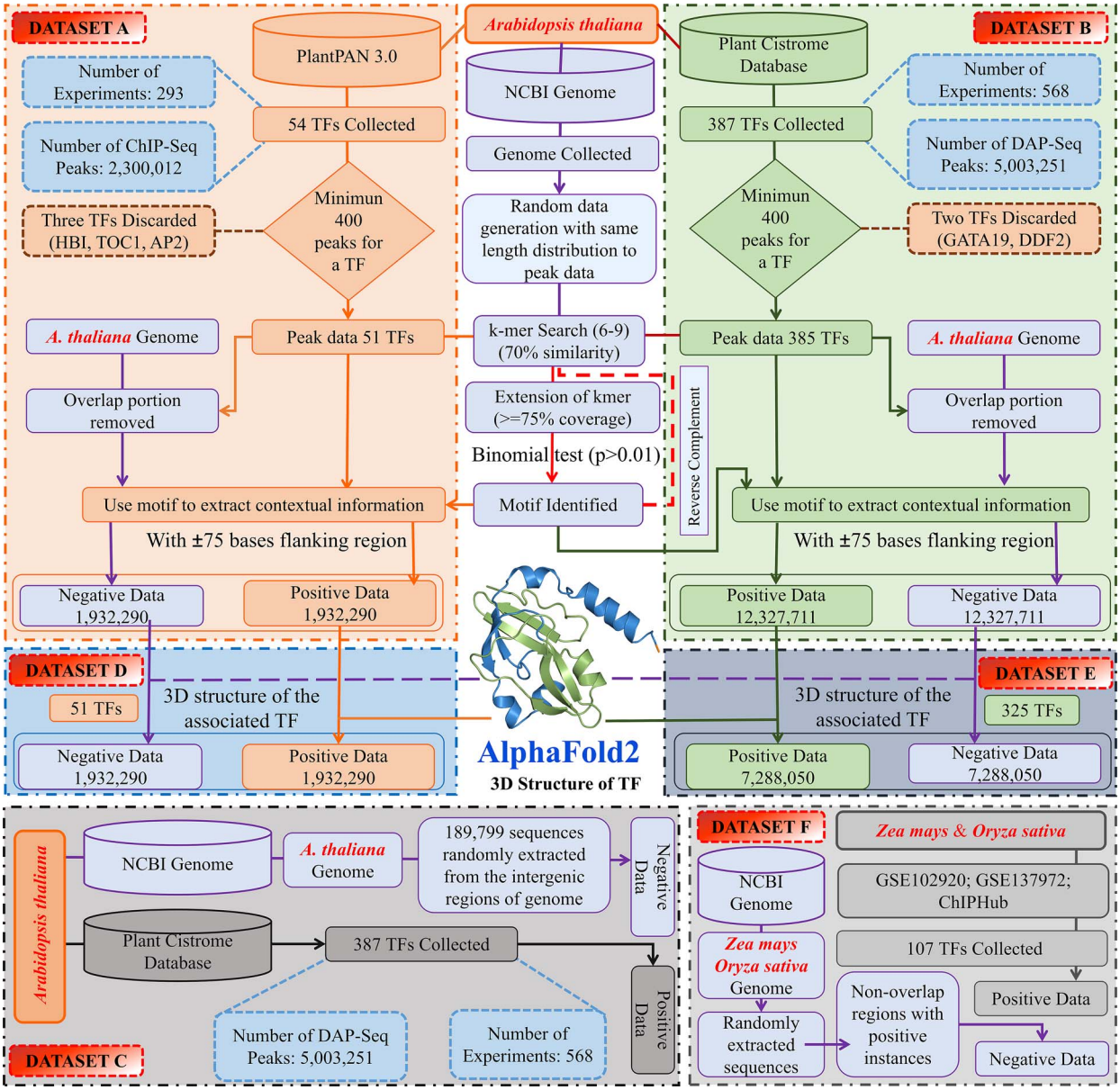

Figure 2. Flowchart representation of dataset formation. (a) The protocol followed for dataset "A" creation, (b) Dataset "B" creation, (c) Dataset "C" creation, (d) Dataset "D" creation, (E) Dataset "E" creation, and (f) Dataset "F" creation. The datasets "A" and "B" contained the positive instances originating from ChIP-seq and DAP-seq, respectively, with sources PlantPAN3.0 and Plant Cistrome databases, respectively. The negative part of the datasets was formed by considering those genomic regions that 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 in order 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 *Z. mays* (93 TF) and *O. sativa* (24 TF).

size ($d = 28$) and columns by the number of tokenized words (n). Positional encodings of dimension "d" were computed in parallel using sinusoidal functions [42] and combined with the word embeddings, forming the input M' to the Transformer encoder. Within the multi-headed self-attention mechanism, the embedded sequence M' was projected onto query (Q), key (K), and value (V) matrices using learnable weight matrices W^Q, W^K, W^V:

$$Q = M'.W^Q, K = M'.W^K, V = M'.W^V$$

The attention scores were then computed through scaled dot-product attention between Q and K, followed by softmax normalization and multiplication with V:

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

This computes pairwise attention weights between all words, capturing their contextual associations.

A multi-headed transformer featuring 14 attention heads was executed in parallel, with their outputs concatenated to capture different relational perspectives. The combined contextual representations underwent feed-forward processing, normalization, dropout regularization and global average pooling, and a final classification layer with a sigmoid activation function for binary

prediction of TF-binding region. The Adam optimizer [43] was employed for model training using binary cross-entropy loss. This multiheaded self-attention architecture enabled effective learning of long-range dependencies and contextual TF-DNA binding preferences from the tokenized, embedded DNA sequence inputs across 436 *A. thaliana* TFs. Hyperparameter optimization was conducted employing Bayesian optimization. Comprehensive implementation and optimization details of the transformer system are given in the Supplementary Information Materials and Methods section.

Structural and molecular dynamics studies

The 3D structures of TFs were modeled using AlphaFold2 [39], with the top-ranked models selected for the study. ScanProsite (http://prosite.expas.org) confirmed the functional domain and amino acid residues in the active site pocket. Comparative studies were conducted based on protein sequence, functional domain, 3D structure, and binding affinity. Additional details regarding the validation of prime motif and TF interactions through molecular docking [44] and simulation are given in the Supplementary Materials and Methods section.

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

The TF and binding regions' covariability was learned through a hybrid Transformer-DenseNet system. DenseNet is a very-high-depth convolutional neural network (CNN)–based Deep-Learning architecture that learns the spatial patterns much efficiently than CNNs due to its higher depth and capability to keep the learning from previous layers afresh while effectively mitigating the vanishing gradient issue. 3D structure information and the corresponding binding region from ChIP/DAP-seq data for each TF were considered for its training. The DenseNet [45] part processes atom-wise coordinates, while the Transformer part processes the sequence-based input as described in the section Implementation of the Transformers. Normalized coordinates are inputted into a convolution layer with the dimensions of $300 \times 24 \times 3$, accommodating the TF's amino acid positions and atoms. Zero-padding maintains consistent matrix dimensions for shorter sequences. This approach stems from an analysis of over 400 TFs, revealing a maximum of 24 atoms per amino acid. Full implementation and optimization details are provided in the Supplementary Materials and Methods section.

Building the DenseNet architecture

Each layer in DenseNet receives information from all preceding layers, facilitating more efficient feature learning. The DenseNet model employed in our study consists of one convolution layer with 32 filters (kernel size = 3), followed by batch normalization and 2D maxpooling (stride = 2). It comprises 10 dense blocks and nine transition layers, totaling 121 layers.

Within each dense block, the input X_l is concatenated with the feature maps of all preceding layers, denoted as $(m_0, m_1, \ldots, m_{l\text{-}1})$. Each layer within the dense block consists of batch normalization, ReLU activation, and 3×3 convolution. Transition layers facilitate down-sampling and include batch normalization, 3×3 convolution, maxpooling2D, batch normalization, another 3×3 convolution, and a dropout layer. The growth rate "k" determines the number of feature maps contributed to the global state.

After down-sampling, the output is flattened and concatenated with the transformer output for classification. This concatenated output undergoes batch normalization, dropout, dense, and dropout layers before passing through a sigmoid activation-based single-node classification layer. We utilized the 'Adam' optimizer for weight adjustment, with a batch size of 64 and six epochs. Further details of this module are provided in Fig. 3.

Performance assessment

According to standard practice, every dataset involved in training–testing was divided into train (70%) and test datasets (30%). The developed Transformer-DenseNet model was tested on the 30% intact and completely untouched test portion. Four categories of performance confusion matrix, namely, true positives (TPs), false negatives (FNs), false positives (FPs), and true negatives (TNs), were evaluated. The performance of the built Transformer-DenseNet model was evaluated using performance metrics such as sensitivity, specificity, accuracy, F1 score, and Mathews correlation coefficient (MCC) [46].

Performance measures were done using the following equations:

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

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

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

$$F1 - Score = 2 \times \left(\frac{\mathrm{Pr}ecision \times \mathrm{Re}call}{\mathrm{Pr}ecision + \mathrm{Re}call}\right)$$

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

where:

TP = true positives, TN = true negatives, FP = false positives, FN = false negatives, Acc = accuracy.

A test was also conducted to determine if there was any significant over-fitting occurring in the Transformer-DenseNet final model. The gold-standard method for detecting such over-fitting is through 10 times independent random training and testing trials, which compares the mean absolute error (MAE) between the training and testing performances. 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 scratch and evaluated on the corresponding test set. In addition, it was made sure that there was no overlap between any of 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.

Full methods details are provided in the Supplementary Materials and Methods section. We strongly recommend readers to refer to that for a comprehensive understanding of the methodological details employed in this study.

Results and discussion

Learning on cumulative contextual information surrounding the anchoring prime motif helps identify transcription factor–binding regions with high accuracy

The prime motif discovery (detailed methods and results about which are given in the Supplementary Information Results section) helped in selecting the more appropriate contextual information and features. The motifs may occur significantly enriched in the binding data, but by no means they are limited there only; they also appear in the nonbinding regions. The discovered motifs above worked as the point to zero upon to consider the

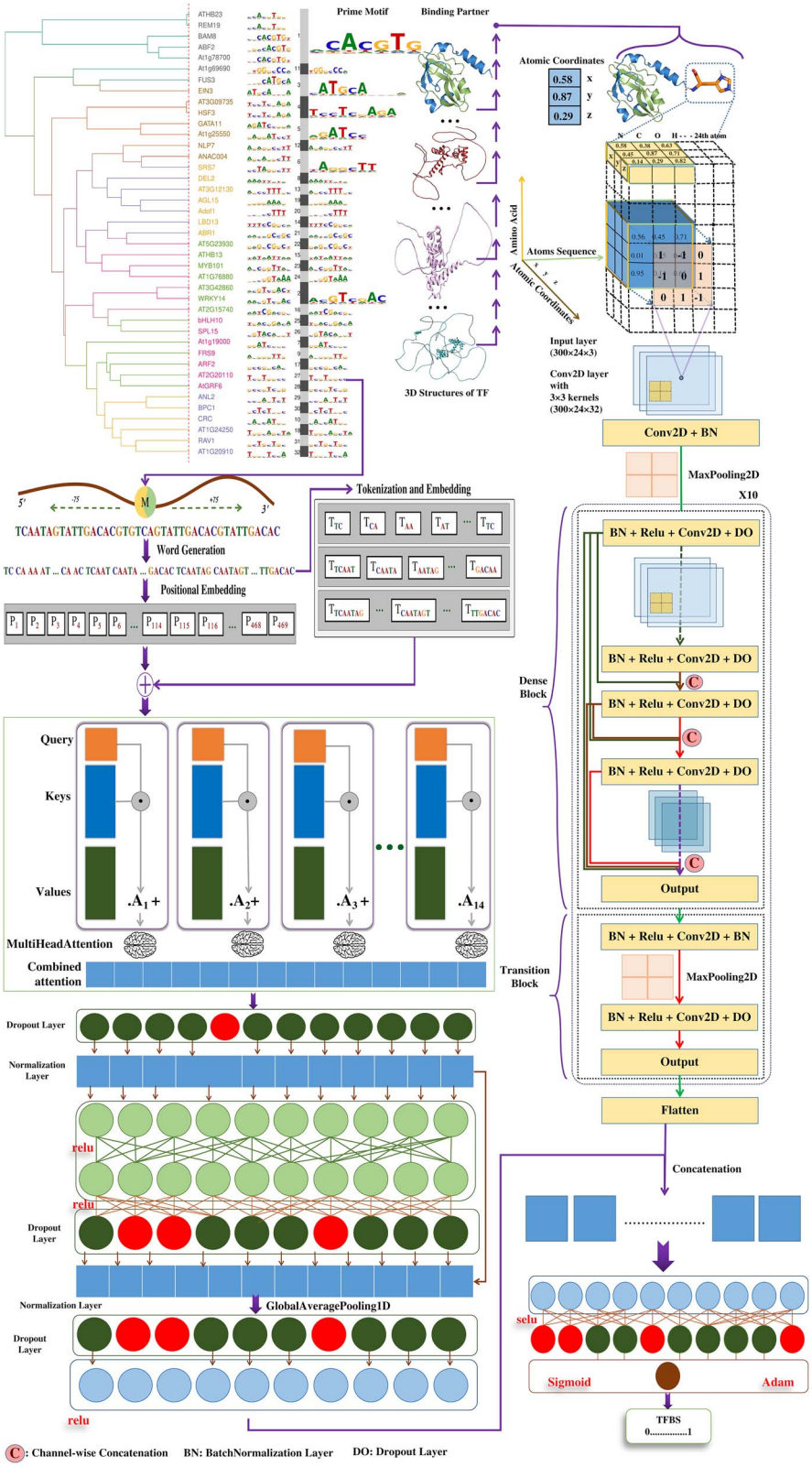

Figure 3. Implementation of the PTFSpot Deep Co-learning system using Transformers and DenseNet to identify TFBR across plant genomes. The first part is a 14 heads attention transformers which learn from the dimeric, pentameric, and heptameric word representations of any given sequence arising from anchoring 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 a binding region in the considered sequence.

potentially significant interaction spots across the DNA. Thus, it was imperative to assess their context and its contribution. Therefore, 75 bases-flanking regions from both the ends of the motif were considered. Previously, it was found that such extent of the flanking regions around the potential interaction sites in nucleic acids captures the local environment contributory

information effectively [37, 38, 47–50]. Also, such regions were found important in determining the stationing of TFs through a localized search for right points to halt at [25]. The contextual information may come in the form of other co-occurring motifs, sequence and position-specific information, and structural/shape information that 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 Materials and Methods and related supplementary section. Figure 2 illustrates how these datasets were built.

For building the models for 436 TFs and their binding preferences, 10 different combinations of various sequence representations were used. An assessment was made for each representation considered where the Dataset "A" was split into 70:30 ratio to form the train and test sets. This protocol worked as the ablation analysis to evaluate how each of these representations of the sequence was contributing toward the discrimination between the preferred binding and nonpreferred binding regions through the transformer encoders (Fig. 4a). The observed accuracy for dimeric representation was just 80.78% on Dataset "A." This was followed by introduction of pentameric and heptameric sequence representations that 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 do not 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 considered. By doing so, a 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 together. On these, the transformers learned contextually along with the prime motifs with much higher amount of information sharing across the representations.

Combination of 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 word representations yielded a better result than using any single-type word representations, as can be seen from Fig. 4b. 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 Materials and Methods and associated supplementary section and Fig. 4a.

Ten-fold random trials performance concurred with the above-observed performance level and scored in the same range consistently. All of them achieved high-quality Receiver Operating Characteristic (ROC) curves with high Area Under the Curve (AUC) values in the range of 0.9245–0.9561 (Dataset "A") and 0.9562–0.9869 (Dataset "B") while maintaining reasonable balance between specificity and sensitivity (Fig. 4c; Supplementary Fig. 1; Supplementary Table 3 Sheet 4–5). To conduct an unbiased performance testing without any potential recollection of data instances, it was ensured that no overlap and redundancy existed across the data. The remarkable performance consistency ensured about the robustness of the raised transformer models and reliability of its all future results. It was evident that the Transformer effectively grasped both distant and nearby words associations, while acquiring knowledge through multiple hidden features.

PTFSpot transformer models consistently surpassed all other compared transcription factor–binding region-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 nine different tools, representing most recent and different approaches of TFBR detection: AgentBind (DanQ, LSTM based), AgentBind (DeepSea, CNN based), *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-Transformers outperformed almost all of them, for all the performance metrics considered (Fig. 4d). TSPTFBS 2.0 came very close to the performance of PTFSpot-Transformers with an average accuracy of 96.02% and MCC of 0.9185 (PTFSpot-Transformers average accuracy: 95.98%, average MCC: 0.9231). On the same Dataset 'B', the next-best-performing tool was AgentBind-DanQ (Avg accuracy: 87.77% and MCC: 0.7807), a very distant one in performance.

On Dataset "C," PTFSpot transformers outperformed all the compared tools with significant margin with a similar level of performance (Fig. 4f). PTFSpot transformers consistently demonstrated minimal variance in its performance, maintaining a strong balance in accurately identifying both positive and negative instances. This was evident through its high values across all the three performance metrics with least dispersion, affirming the algorithm's robustness (Fig. 4e and g). The full details and data for this benchmarking study are given in Supplementary Table 3 Sheet 6–7.

The transcription factors vary across species and so do their binding regions' preferences

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 [31]. This is why in order to find TFs and their binding sites, *Arabidopsis* and TF-specific models have been rampantly used, ending up with largely misleading results. All of them assume that the TFs and their binding preferences remain static and give no weight age to their co-variability. While in actual, the TFs and their binding regions vary across the plant species [29, 30, 32]. As the TF structure changes, the binding sites also changes, and so does the preferred binding regions. Therefore, what works for one species, does not work for another until co-variability between TFs and their binding preferences is learned. And this is why most of the existing tools fail to work in cross-species manner, making them almost of no use. A study was done here to verify the same. Peak data of the common TFs were selected in *A. thaliana* and *Z. mays* i.e., LHY1, MYB56, MYB62, MYB81, MYB88, and WRKY25. For both the species their corresponding prime motifs for the same TF were compared. It was observed that for the same TF different prime binding motifs existed. On comparing the

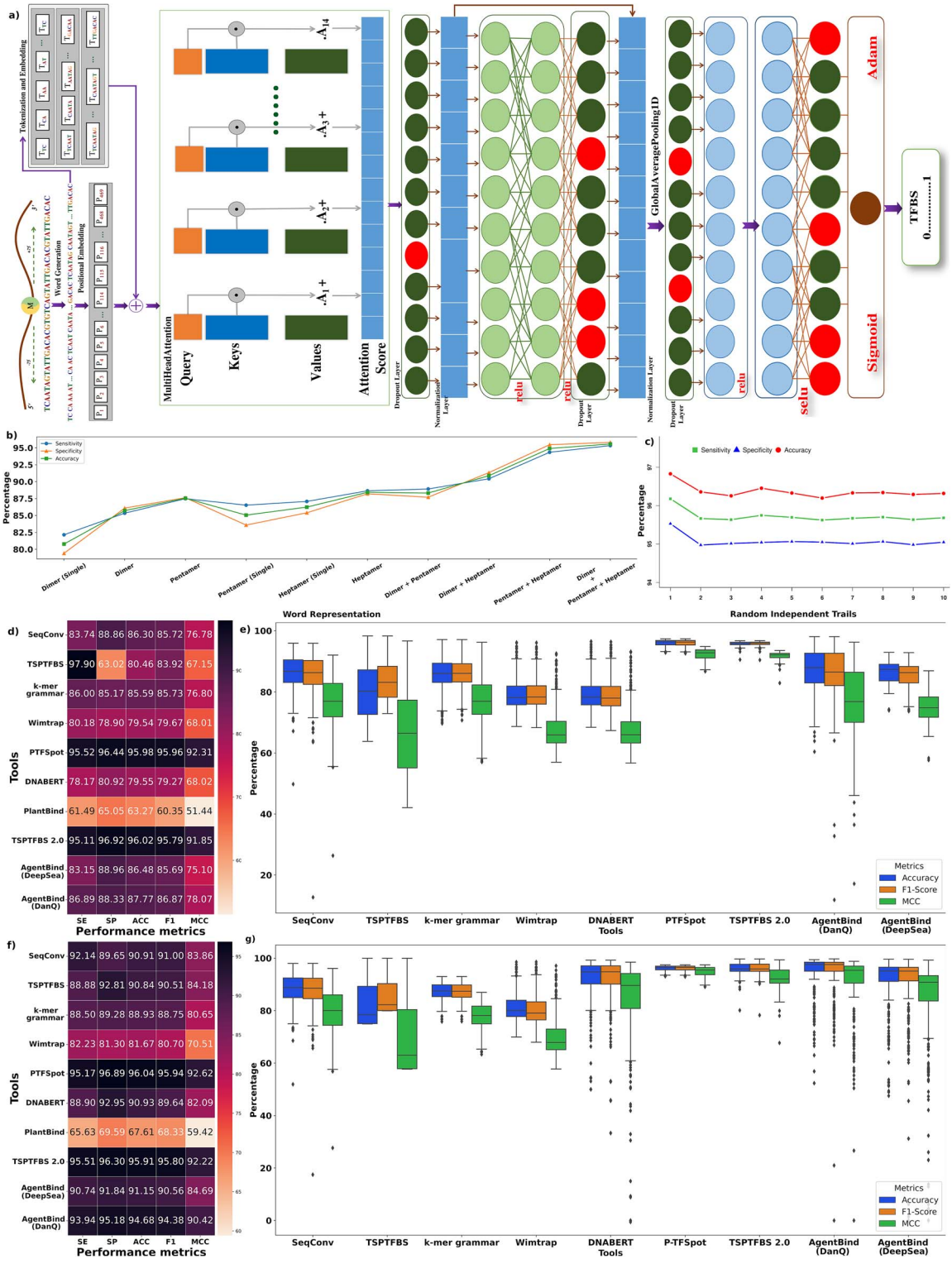

Figure 4. (a) Transformer-only model's implementation details. This part utilizes DNA sequences–based information in terms of 7-mer, 5-mers, and dimeric words while basing around (+ −75 bp) the prime motif to detect the contextual information. (b) Ablation analysis for three main properties for discriminating between the negative and positive instances. These word representations appeared highly additive and complementary to each other as the performance increased substantially as they were combined together. (c) Ten-fold independent training–testing random trials on Dataset "B" depicts consistent performance of PTFSpot transformers. (d) Objective comparative benchmarking result on Dataset "B." These 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." (e) 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. (f) 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. (g) 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).

same TF's sequences and structures between the two species the following was observed:

1) The amino acid sequence identity for each TF was reasonably low between the two species, with lowest going up to 39% (LHY1). The amino acids sequence-based comparative details of each TF are given in Supplementary Table 3 Sheet 8.

2) The binding domain class of the compared TFs across the species were same but their amino acid compositions were different. For example, in the case of MYB88, both species contained two Helix-turn-helix (HTH) DNA-binding domain, but an identity of only 33.3% was observed for the first domain and an identity of 43.2% was noted for the second domain when comparing both species (Fig. 5bI & II).

3) The 3D structures of the same TF between the two species were compared and the Root Mean Square Deviation (RMSD) difference between them was found above 0.6 Å (Fig. 5) [51]. 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 Fig. 5aIII and IV, respectively. The details of 3D structure comparison of other TFs is given in Supplementary Fig. 2.

4) A TF's binding affinity was higher to the prime motif within the same species than other species. As in Fig. 5, taking the case of MYB 88, it was observed that when the TF of *A. thaliana* was docked with its own prime motif, the binding affinity was −66.45 kcal/-mol. However, when the *A. thaliana* MYB88 was docked with the prime motif for MYB88 of *Z. mays*, the binding affinity went much lower with −46.70 kcal/mol (Fig. 5bIII & IV). The same procedure was applied for *Z. mays* MYB88 for its binding to its own prime motif and that of *A. thaliana,* and a similar pattern was observed there also (Fig. 5bV & VI), clearly underlining that there is cross-species variability in TF binding preferences that is grossly neglected and leading to wrong study designs.

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

Deep co-learning on transcription factor sequence, structure, and corresponding DNA-binding regions brings impeccably accurate universal model of transcription factor-DNA interaction spots

During the real-world application of cross-species identification of the TF-binding regions, 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 software tools for TFBR discovery during the process of their annotations where most of them end up reporting a very high proportion of false positives [17, 18, 52]. Above, we have showed how variability in TF structure and corresponding binding regions happen across the plants, which none of the existing tools has attempted to learn. This becomes a major reason why the existing software pool does not work across the plant species.

As detailed in the Materials and Methods sections about the architecture of PTFSpot implementation, a composite deep co-learning system was raised using Transformers and DenseNet, which parallelly learned upon the TF-binding regions in DNA with sequence contexts and the corresponding TF's 3D structure and sequence. This model was trained and tested on Dataset "E" training and testing components, in 70:30 split ratio with absolutely no overlap and all redundancies removed. This co-learning system was trained using 41 TFs and their corresponding DAP-seq data. Each selected TF represented a single TF family and had the highest binding data available among all the TFs for the respective family. 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 (Fig. 6a; Supplementary Table 4 Sheet 1), almost perfectly capturing the binding regions for every TF considered. The information sharing between the TF structure and binding region covariability was so strong that when the model was assessed by removing the structural part, the accuracy dropped drastically to just 83.6%, significantly lower (*P*-value: 4.70e-24) than what was achieved above with co-learning on TF structure and corresponding binding data. 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 covariability was learned to correctly identify the binding preferences for even those TFs 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 TFs that had not even family representations in the above mentioned model, which covered 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 that was never included in the training of the above mentioned co-learning model. Here, 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). A structural comparison-based cluster analysis across all the 41 TFs in training and 21 TFs in the testing sets, both representing totally different TF family sets, was done. It was observed that performance observed for the most distant TF family in the test set was at par with the one with the closest distance with the members of the training set, and the overall performance across all the families was at the same level (Supplementary Fig. 3). Separate 10-fold training–testing runs were made to measure the performance consistency where all the runs scored at the similar level consistently. Also, the MAE for training was found to be 0.0348, while for testing. it was 0.0362, resulting in a very small difference of only 0.0014. A *t*-test comparing the MAE values for the training and test sets yielded a highly insignificant result of ∼45%, much above the significance threshold of 5% or lower, further confirming the absence of any possibility of any significant over-fitting (Supplementary Table 2 Sheet 2). This all concurred again that the model truly learned the covariability between TF structure and its binding preferences with remarkable consistency and robustness.

Afterward, 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 (Fig. 6a; Supplementary Table 4 Sheet 2).

The application value of any such software is mainly when it is applicable universally, across various TFs (which PTFSpot qualified above) as well as across different species, more so when very frequently sequenced genomes of various plants are being released, which essentially require TFBR annotations. Almost all of the existing software fail there.

The raised model's last assessment was set to determine how well it performed to identify cross-species TF-binding regions. For this task, a new dataset, Dataset "F," was employed completely as another test set. This dataset contained 117 TFs from *Z. mays*

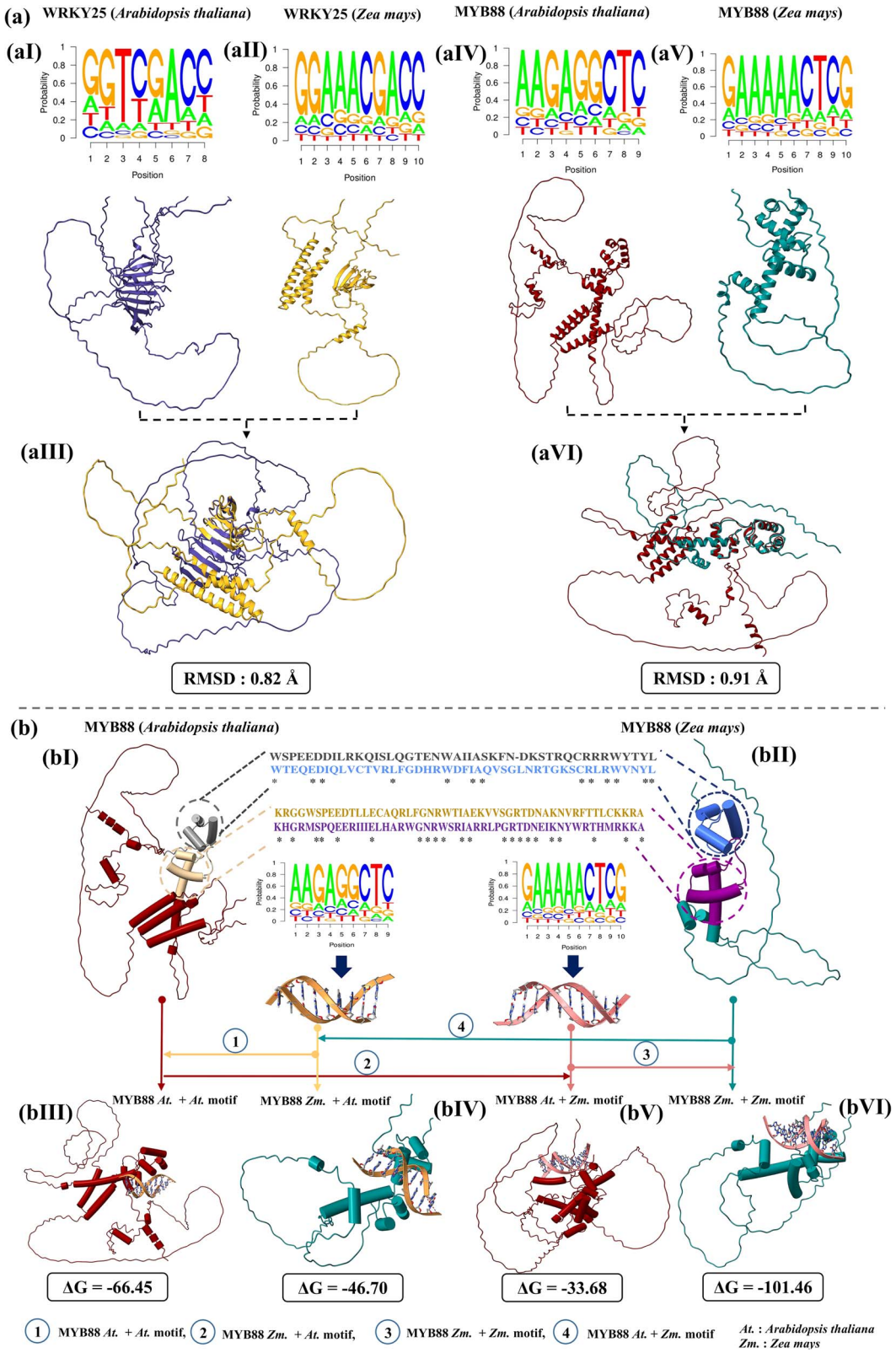

Figure 5. Covariation in the structure of the transcription factor and the corresponding binding site across various species. (a) TF structure and its binding motif comparison across the species, (aI) prime binding motif of WRKY25 TF (*A. thaliana*) and its 3D structure, (aII) prime binding motif of WRKY25 TF (*Z. mays*) and its 3D structure, (aIII) superimposed TF structures of *A. thaliana* and *Z. mays*, with the structural differences measured in RMSD value. (aIV) prime binding motif of MYB88 TF (*A. thaliana*) and its 3D structure, (aV) prime binding motif for WRKY25 TF (*Z. mays*) and its 3D structure, (aVI) superimposed TF structures for *A. thaliana* and *Z. mays*, and corresponding structural difference in RMSD value. (b) Domain-based comparison between *A. thaliana* and *Z. mays* for MYB88, (bI) *A. thaliana*'s MYB88 has two domains. The first domain and its corresponding amino acids sequence are in gray color. The second domain and its corresponding amino acids sequence is shown in the tan color, (bII) *Z. 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 and 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 don't work across the species and it varies with species as well as the structure of the TF.

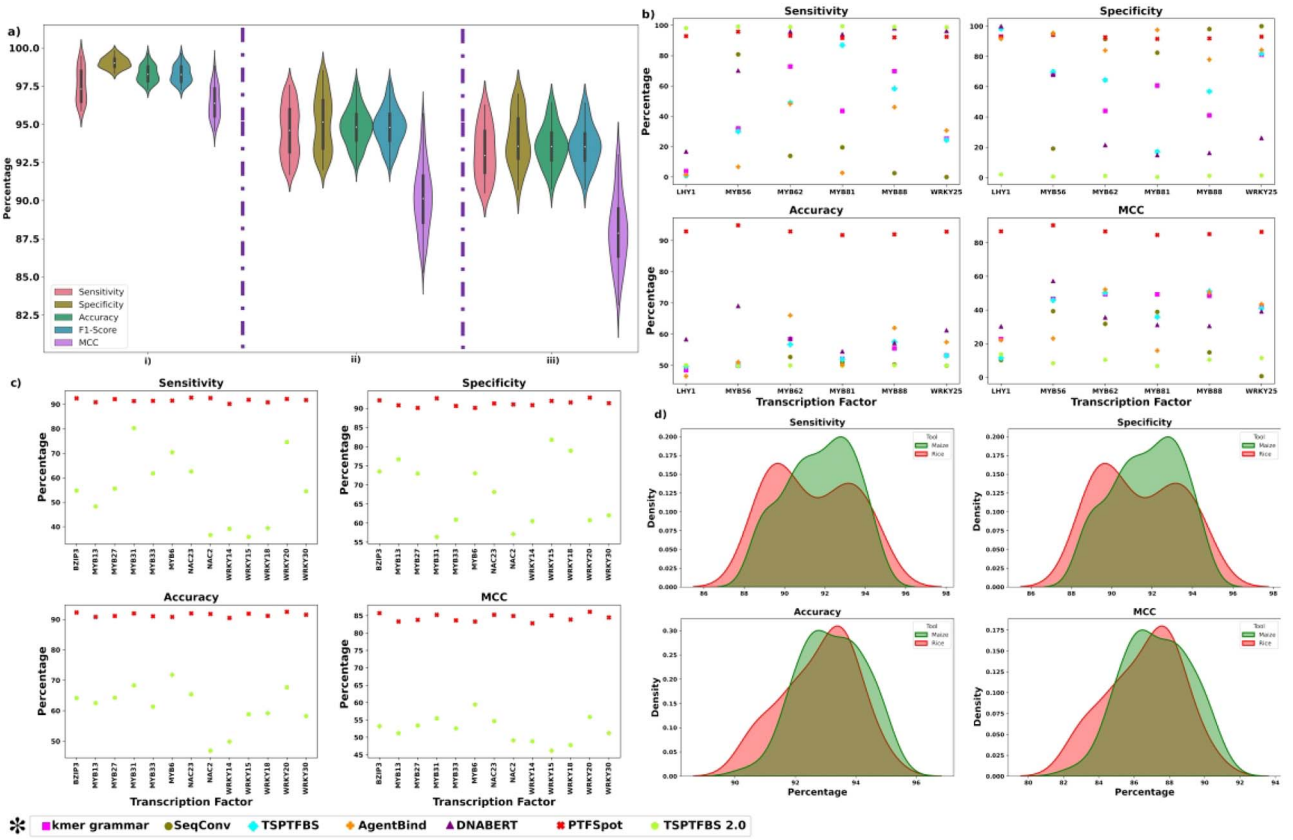

Figure 6. Performance and benchmarking of the universal model of PTFSpot. PTFSpot universal model was raised from 41 different TFs representing 41 different families, from *A. thaliana*. (i) (a) The performance over the same 41 TF's test set (Dataset "E"). (b) The performance over the remaining TFs as test set from Dataset "E." (c) 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. (ii) 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. (iii) Comparative benchmarking for the 13 wheat TFs (371 066 peak regions, equal number of negative instances following the same protocol as dataset "F") whose models were available in TSPTFBS2.0 for comparison. PTFSpot universal model outperformed TSPTFBS2.0 for every compared TF by huge margins (*P*-value: 1.46e-05; Kruskal–Wallis test). Reason to select TSPTFBS2.0 for this comparison was that among the existing software tools, TSPTFBS2.0 was found as the best performer. (iv) 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 TFBRs.

(93 TFs) and *O. sativa* (24 TFs), retrieved from GSE137972 (217 samples), GSE102920 (6 samples), and ChIP-Hub with over 60 conditions. It was created solely for cross-species validation purpose for the model's performance. In this study, we included the most recent and advanced tools like *k-mer* grammar, TSPTFBS, TSPTFBS 2.0, SeqConv, and Wimtrap for comparative benchmarking here. We also included a novel approach, AgentBind, based on context learning from the flanking regions, which utilizes CNN/LSTM to learn from the sequence patterns [15]. We looked for common TFs among different plant species to benchmark these tools for cross-species performance. In Dataset "F," only six TFs from *Z. mays* (LHY1, MYB56, MYB62, MYB81, MYB88, WRKY25) were found for common ones that the existing software tools had any model developed for (Supplementary Table 4 Sheet 3). Thus, the comparative benchmarking was possible for only these six TFs for cross-species performance evaluation of the tools. For AgentBind, models for these TFs were raised using TF-specific data used for PTFSpot while using AgentBind's protocol.

Two levels of benchmarking was performed. In the first level, seven tools, DNABERT, *k-mer* grammar, TSPTFBS, TSPTFBS 2.0, SeqConv, AgentBind, and Wimtrap, were considered and the six common TFs data were taken for evaluation along with the PTFSpot universal model. The 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 TFBR in cross-species manner. The performance observed for all the other tools was extremely poor, with none surpassing an average accuracy value of 60% (Fig. 6b; Supplementary Table 4 Sheet 3), reiterating the fact that none of the existing tools are suitable for any practical application like cross-species TFBR identification and genomic annotations. PTFSpot has emerged as a breakthrough solution to this situation.

An additional benchmarking exercise was done where TSPTFBS 2.0 performance was compared with PTFSpot for the DAP-seq binding data for 13 TFs (371 066 peak regions, equal number of negative instances following the same protocol as dataset "F") from wheat [53]. These TFs were selected only because TSPTFBS 2.0 had models for only these TFs. TFPTFBS 2.0 was selected here because it was found as the best-performing one among the existing pool of software. For this comparison also, PTFSpot significantly outperformed TSPTFBS 2.0 (*P*-value: 1.46e-05) with 30% performance lead over TSPTFBS 2.0 in terms of accuracy (31% lead in MCC) while attaining 91.48.% accuracy (84.43% MCC), clearly reiterating its impeccable performance in plant TF-binding region discovery without any bounding to TF and species-specific models, which none of the existing tools has attained so far.

In the next step, the above analysis was carried forward for the entire Dataset "F," which covered data for 117 TFs from rice and maize. Here also, PTFSpot performed outstanding while attaining 93.58% average accuracy, an MCC value of 0.87, and an F1-score of 93.56%. Figure 6d provides further performance distribution illustration of PTFSpot for *O. sativa* and *Z. mays* TFs (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 TF as well as across species due to its successful co-learning of the variability in structure and binding regions.

To give a short glimpse of the kind of impact PTFSpot could have due to its capabilities to detect covariability between TF structure and its binding preferences, we extended our above-described case of MYB88 example. MYB88 is reported to influence PIN7 gene involved in auxin efflux and transport associated with plant development [54] by binding at five locations in the promoter of PIN7. When TSPTFBS2.0 was run to detect the same in *A. thaliana,* it could not report any binding site for MYB88. However, PTFSpot detected all of them there. The homologous gene for PIN7 in Maize is zmPIN1c [55] whose promoter was also scanned for MYB88 binding. This gene displays ~60% identity between Maize and *A. thaliana*, and exhibits a remarkable variability despite retaining its function. As already showed in the section above, even the structure of MYB88 has drifted a lot from *Arabidopsis* to Maize. When the *Arabidopsis*-specific Transformer-only PTFSpot model of MYB88 was run to scan for its binding regions in the promoter of zmPIN1c in maize, nothing was found. TSPTFBS2.0 also reported nothing for MYB88 there. However, when the universal model–based PTFSpot was run, it detected two binding locations for MYB88 there. To validate any possible regulatory role of MYB88 in transcriptional regulation of zmPIN1c in maize, we performed gene expression correlation analysis between them, utilizing data from nine experimental conditions available at Maize Expression Atlas database (European Bioinformatics Institute: [https://www.ebi.ac.uk/gxa/home]. Remarkably, a very strong Pearson correlation coefficient of 0.97 was observed, indicating a very high possibility of regulation of zmPIN1c by MYB88 in maize. Further details of this analysis are presented in Supplementary Tables 5 and 6. This small demonstration highlights the kind of deep impact PTFSpot may have in unraveling the regulatory systems of plants while breaking several age long bottlenecks.

Conclusion

The present work brings a revolutionary new approach, PTFSpot, which learns from the covariability between binding protein structure and its binding regions without requiring to be specific for any particular TF or its family-specific model. It can accurately identify the binding regions for any given TF belonging to any family across any plant genome, and can work for even any novel and never-reported-before TFs and genomes with the same level of accuracy. 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 TF-binding regions across a genome.

Key Points

- Plant genomes are highly variable, which is transferred to all of its component elements, including transcription factors (TFs) and their binding region (TFBR) preferences. Despite that, this fact has been grossly ignored and incorrect practice of using TF:DNA interaction models for one species to another is dominant, creating misleading and incorrect reports and findings.
- None of the existing software tools to detect TFBRs in plants is equipped to understand and capture this variability. This is why most of them fail to perform well during cross-species applications and perform well only for those species and TFs for which they have specific models built, making them hardly of any practical value like annotation TFBRs for newly sequenced plant genomes.
- For the first time here, relationship between a TF's sequence and structure variability and its binding region preferences has been learned through a Transformer-DenseNet Deep-Learning system, delivering a single universal model of TF:DNA interactions that can work for any TF and any plant species, seen or unseen, with equal performance. This has liberated the TF regulatory research in plants from the abovementioned bottlenecks and makes it feasible to detect TFBRs with utmost accuracy and reliability for any TF and any plant genome.
- The developed tool, PTFSpot, has been tested across a huge volume of experimental data where it breached the accuracy of 98% and always scored >90% in every validation test, while maintaining an average lead of >30% on the compared tools. In cross-species tests done on rice, wheat, and maize, where none of the existing software were found able to attain even 60% average accuracy, PTFSpot attained consistently >90% accuracy values for completely unseen-before TFs and genomes.
- PTFSpot is expected to revolutionize the plant TF regulatory research as it will empower to mimic costly high-throughput experiments like DAP-seq with its highly accurate TFBR identification for any TF and plant genome, known or novel. This philosophy can also be extended to animal systems as well as other interaction studies.

Supplementary data

Supplementary data are available online at *Briefings in Bioinformatics* online.

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. S.G. and V.K. are thankful to DBT, India for financial support as project associateship. U.B. is thankful for DBT JRF fellowship. Jyoti is thankful for CSIR-UGC SRF fellowship. The authors are thankful to Ritu for her inputs on DenseNet. This MS has CSIR-IHBT MSID 5451.

Conflicts of interest: The authors declare no conflicts of interest.

Funding

The work was funded under National Network Project, S2S [BT/PR40177/BTIS/137/49/2022].

Author contributions

Sagar Gupta (Data curation, Formal analysis, Methodology, Software, Visualization, Resources, Writing—original draft), Veerbhan Kesarwani (Methodology, Formal analysis, Validation, Writing—original draft), Umesh Bhati (Formal analysis, Validation, Writing—original draft), Jyoti (Formal analysis, Validation, Writing—original draft), Ravi Shankar (Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Writing—original draft, Writing—review & editing).

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 1–6. 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 also).

References

1. Berger MF, Bulyk ML. Protein binding microarrays (PBMs) for the rapid, high-throughput characterization of the sequence specificities of DNA binding proteins. *Methods Mol Biol* 2006;**338**:245–60.
2. Johnson DS, Mortazavi A, Myers RM. *et al.* Genome-wide mapping of in vivo protein-DNA interactions. *Science* 2007;**316**:1497–502. https://doi.org/10.1126/science.1141319.
3. Rhee HS, Pugh BF. Comprehensive genome-wide protein-DNA interactions detected at single nucleotide resolution. *Cell* 2011;**147**:1408–19. https://doi.org/10.1016/j.cell.2011.11.013.
4. Bartlett A, O'Malley RC, Huang SC. *et al.* Mapping genome-wide transcription factor binding sites using DAP-seq. *Nat Protoc* 2017;**12**:1659–72. https://doi.org/10.1038/nprot.2017.055.
5. Nakato R, Shirahige K. Recent advances in ChIP-seq analysis: from quality management to whole-genome annotation. *Brief Bioinform* 2017;**18**:279–90. https://doi.org/10.1093/bib/bbw023.
6. Ghandi M, Lee D, Mohammad-Noori M. *et al.* Enhanced regulatory sequence prediction using gapped k-mer features. *PLoS Comput Biol* 2014;**10**:e1003711. https://doi.org/10.1371/journal.pcbi.1003711.
7. Jha A, Shankar R. MiRNAting control of DNA methylation. *J Biosci* 2014;**39**:365–80. https://doi.org/10.1007/s12038-014-9437-9.
8. Alipanahi B, Delong A, Weirauch MT. *et al.* Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotechnol* 2015;**33**:831–8. https://doi.org/10.1038/nbt.3300.
9. Shen Z, Bao W, Huang D-S. Recurrent neural network for predicting transcription factor binding sites. *Sci Rep* 2018;**8**:15270. https://doi.org/10.1038/s41598-018-33321-1.
10. Mejía-Guerra MK, Buckler ES. A k-mer grammar analysis to uncover maize regulatory architecture. *BMC Plant Biol* 2019;**19**:103. https://doi.org/10.1186/s12870-019-1693-2.
11. Yang J, Ma A, Hoppe AD. *et al.* Prediction of regulatory motifs from human Chip-sequencing data using a deep learning framework. *Nucleic Acids Res* 2019;**47**:7809–24. https://doi.org/10.1093/nar/gkz672.
12. Trabelsi A, Chaabane M, Ben-Hur A. Comprehensive evaluation of deep learning architectures for prediction of DNA/RNA sequence binding specificities. *Bioinformatics* 2019;**35**:i269–77. https://doi.org/10.1093/bioinformatics/btz339.
13. Zhang Q, Shen Z, Huang D-S. Modeling in-vivo protein-DNA binding by combining multiple-instance learning with a hybrid deep neural network. *Sci Rep* 2019;**9**:8484. https://doi.org/10.1038/s41598-019-44966-x.
14. Shen L-C, Liu Y, Song J. *et al.* SAResNet: self-attention residual network for predicting DNA-protein binding. *Brief Bioinform* 2021;**22**:bbab101. https://doi.org/10.1093/bib/bbab101.
15. Zheng A, Lamkin M, Zhao H. *et al.* Deep neural networks identify sequence context features predictive of transcription factor binding. *Nat Mach Intell* 2021;**3**:172–80. https://doi.org/10.1038/s42256-020-00282-y.
16. Shen W, Pan J, Wang G. *et al.* Deep learning-based prediction of TFBSs in plants. *Trends Plant Sci* 2021;**26**:1301–2. https://doi.org/10.1016/j.tplants.2021.06.016.
17. Liu L, Zhang G, He S. *et al.* TSPTFBS: a Docker image for trans-species prediction of transcription factor binding sites in plants. *Bioinformatics* 2021;**37**:260–2. https://doi.org/10.1093/bioinformatics/btaa1100.
18. Ji Y, Zhou Z, Liu H. *et al.* DNABERT: pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. *Bioinformatics* 2021;**37**:2112–20. https://doi.org/10.1093/bioinformatics/btab083.
19. Rivière Q, Corso M, Ciortan M. *et al.* Exploiting genomic features to improve the prediction of transcription factor-binding sites in plants. *Plant Cell Physiol* 2022;**63**:1457–73. https://doi.org/10.1093/pcp/pcac095.
20. Yan W, Li Z, Pian C. *et al.* PlantBind: an attention-based multi-label neural network for predicting plant transcription factor binding sites. *Brief Bioinform* 2022;**23**:bbac425. https://doi.org/10.1093/bib/bbac425.
21. Cheng H, Liu L, Zhou Y. *et al.* TSPTFBS 2.0: trans-species prediction of transcription factor binding sites and identification of their core motifs in plants. Front. *Plant Sci* 2023;**14**:1175837. https://doi.org/10.3389/fpls.2023.1175837.
22. Zhou T, Yang L, Lu Y. *et al.* DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale. *Nucleic Acids Res* 2013;**41**:W56–62. https://doi.org/10.1093/nar/gkt437.
23. Chaudhari HG, Cohen BA. Local sequence features that influence AP-1 cis-regulatory activity. *Genome Res* 2018;**28**:171–81. https://doi.org/10.1101/gr.226530.117.
24. Sielemann J, Wulf D, Schmidt R. *et al.* Local DNA shape is a general principle of transcription factor binding specificity in Arabidopsis thaliana. *Nat Commun* 2021;**12**:6549. https://doi.org/10.1038/s41467-021-26819-2.
25. Castellanos M, Mothi N, Muñoz V. Eukaryotic transcription factors can track and control their target genes using DNA antennas. *Nat Commun* 2020;**11**:540. https://doi.org/10.1038/s41467-019-14217-8.
26. Suter DM. Transcription factors and DNA play Hide and Seek. *Trends Cell Biol* 2020;**30**:491–500. https://doi.org/10.1016/j.tcb.2020.03.003.
27. Panchy NL, Azodi CB, Winship EF. *et al.* Expression and regulatory asymmetry of retained Arabidopsis thaliana transcription factor genes derived from whole genome duplication. *BMC Evol Biol* 2019;**19**:77. https://doi.org/10.1186/s12862-019-1398-z.
28. Bennetzen JL, Ma J, Devos KM. Mechanisms of recent genome size variation in flowering plants. *Ann Bot* 2005;**95**:127–32. https://doi.org/10.1093/aob/mci008.
29. Bao Y, Hu G, Grover CE. *et al.* Unraveling cis and trans regulatory evolution during cotton domestication. *Nat Commun* 2019;**10**:5399. https://doi.org/10.1038/s41467-019-13386-w.

30. Shiu S-H, Shih M-C, Li W-H. Transcription factor families have much higher expansion rates in plants than in animals. *Plant Physiol* 2005;**139**:18–26. https://doi.org/10.1104/pp.105.065110.
31. Lambert SA, Yang AWH, Sasse A. *et al.* Similarity regression predicts evolution of transcription factor sequence specificity. *Nat Genet* 2019;**51**:981–9. https://doi.org/10.1038/s41588-019-0411-1.
32. Lehti-Shiu MD, Panchy N, Wang P. *et al.* Diversity, expansion, and evolutionary novelty of plant DNA-binding transcription factor families. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* 2017;**1860**:3–20. https://doi.org/10.1016/j.bbagrm.2016.08.005.
33. Chow C-N, Lee T-Y, Hung Y-C. *et al.* PlantPAN3.0: a new and updated resource for reconstructing transcriptional regulatory networks from ChIP-seq experiments in plants. *Nucleic Acids Res* 2019;**47**:D1155–63. https://doi.org/10.1093/nar/gky1081.
34. O'Malley RC, Huang SC, Song L. *et al.* Cistrome and Epicistrome features shape the regulatory DNA landscape. *Cell* 2016;**165**:1280–92. https://doi.org/10.1016/j.cell.2016.04.038.
35. Fornes O, Castro-Mondragon JA, Khan A. *et al.* JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res* 2020;**48**:D87–92. https://doi.org/10.1093/nar/gkz1001.
36. Jin J, Tian F, Yang D-C. *et al.* PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. *Nucleic Acids Res* 2017;**45**:D1040–5. https://doi.org/10.1093/nar/gkw982.
37. Sharma NK, Gupta S, Kumar A. *et al.* RBPSpot: learning on appropriate contextual information for RBP binding sites discovery. *iScience* 2021;**24**:103381. https://doi.org/10.1016/j.isci.2021.103381.
38. Heikham R, Shankar R. Flanking region sequence information to refine microRNA target predictions. *J Biosci* 2010;**35**:105–18. https://doi.org/10.1007/s12038-010-0013-7.
39. Jumper J, Evans R, Pritzel A. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* 2021;**596**:583–9. https://doi.org/10.1038/s41586-021-03819-2.
40. Fu L-Y, Zhu T, Zhou X. *et al.* ChIP-hub provides an integrative platform for exploring plant regulome. *Nat Commun* 2022;**13**:3413. https://doi.org/10.1038/s41467-022-30770-1.
41. Černý J, Božíková P, Svoboda J. *et al.* A unified dinucleotide alphabet describing both RNA and DNA structures. *Nucleic Acids Res* 2020;**48**:6367–81. https://doi.org/10.1093/nar/gkaa383.
42. Vaswani A, Shazeer N, Parmar N. *et al.* Attention is all you need. *Advances in Neural Information Processing Systems* 2017;**30**:5998–6008.
43. Kingma DP, Ba J. *Adam: A Method for Stochastic Optimization*. arXiv preprint, 2014; 1412.6980.
44. Liu Z, Guo J-T, Li T. *et al.* Structure-based prediction of transcription factor binding sites using a protein-DNA docking approach. *Proteins* 2008;**72**:1114–24. https://doi.org/10.1002/prot.22002.
45. Huang G, Liu Z, van der Maaten L. *et al.* Densely connected convolutional networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017; 4700–8.
46. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* 2020;**21**:6. https://doi.org/10.1186/s12864-019-6413-7.
47. Schöne S, Jurk M, Helabad MB. *et al.* Sequences flanking the core-binding site modulate glucocorticoid receptor structure and activity. *Nat Commun* 2016;**7**:12621. https://doi.org/10.1038/ncomms12621.
48. Yella VR, Bhimsaria D, Ghoshdastidar D. *et al.* Flexibility and structure of flanking DNA impact transcription factor affinity for its core motif. *Nucleic Acids Res* 2018;**46**:11883–97. https://doi.org/10.1093/nar/gky1057.
49. Zambelli F, Pesole G, Pavesi G. PscanChIP: finding over-represented transcription factor-binding site motifs and their correlations in sequences from ChIP-Seq experiments. *Nucleic Acids Res* 2013;**41**:W535–43. https://doi.org/10.1093/nar/gkt448.
50. Grossman SR, Zhang X, Wang L. *et al.* Systematic dissection of genomic features determining transcription factor binding and enhancer function. *Proc Natl Acad Sci* 2017;**114**:E1291–300. https://doi.org/10.1073/pnas.1621150114.
51. Carugo O, Pongor S. A normalized root-mean-square distance for comparing protein three-dimensional structures. *Protein Sci* 2001;**10**:1470–3. https://doi.org/10.1110/ps.690101.
52. Jyoti R, Gupta S, Shankar R. Comprehensive evaluation of plant transcription factors binding sites discovery tools. *bioRxiv* 2023;2023–11. https://doi.org/10.1101/2023.11.07.566153.
53. Zhang Y, Li Z, Zhang Y. *et al.* Evolutionary rewiring of the wheat transcriptional regulatory network by lineage-specific transposable elements. *Genome Res* 2021;**31**:2276–89. https://doi.org/10.1101/gr.275658.121.
54. Wang H-Z, Yang K-Z, Zou J-J. *et al.* Transcriptional regulation of PIN genes by FOUR LIPS and MYB88 during Arabidopsis root gravitropism. *Nat Commun* 2015;**6**:8822. https://doi.org/10.1038/ncomms9822.
55. Forestan C, Varotto S. The role of PIN auxin efflux carriers in polar auxin transport and accumulation and their effect on shaping maize development. *Mol Plant* 2012;**5**:787–98. https://doi.org/10.1093/mp/ssr103.