nature communications

Article | Article in Press

https://doi.org/10.1038/s41467-026-79328-5

Genomic language models decode bidirectional promoter activity and reveal conserved regulatory patterns across plant species

Received: 30 December 2025

Accepted: 15 October 2026

Published online: 05 November 2026

Check for updates

Shang Gao[1,2], Tingxi Yu[1,2], Jinran Wu[3], Shoukun Chen[1,2], Muhammad A. Farooq[1,2], Takele W. Gebrewahid[1,2,4], Jiahui Geng[1,2], Yuanyuan Jiang[1,2], Zhangping Huang[1,2], Jiankang Wang[1,2], Geoffrey J. McLachlan[3] & Huihui Li[1,2,5]

Bidirectional promoters (BDPs), intergenic regions that coordinate transcription of oppositely oriented genes, pervade eukaryotic genomes, yet sequence signatures remain poorly defined. Existing approaches to BDP identification rely on distance thresholds and cannot infer activity from DNA. Here, we use Evo2, a genomic language model, to predict BDP activity and uncover sequence-encoded regulatory principles across four plant species (Arabidopsis, maize, rice, and wheat). We stratify BDPs by co-expression strength and compare Evo2-derived embeddings against *k*-mer profiles, transcription factor binding site matrices, and convolutional neural networks trained on sequence. Evo2 embeddings outperform all baselines, demonstrating that genomic language models capture regulatory information underlying BDP activity. Models trained on one or multiple species transfer to held-out species, and *in planta* validation in *Spartina alterniflora* confirms predicted high-activity BDPs. Sparse autoencoder analysis links Evo2 embeddings to nucleosome positioning and species-specific transcription factor landscapes, providing insight into BDP activity. This work establishes how genomic language models can extract bidirectional promoter logic from sequence and identify conserved regulatory rules across plants.

Bidirectional promoters (BDPs) are intergenic regions initiating transcription of oppositely oriented gene pairs[1–3]. Found across eukaryotes[2–5], they are thought to coordinate expression and maximize use of accessible chromatin. Although BDPs have been identified via transcriptomic and genomic profiling, the sequence determinants underlying their function remain unclear—specifically, whether intrinsic DNA organization can distinguish genuinely active BDPs from merely co-located gene pairs.

Previous studies of promoter architecture show that quantitative sequence representations capture transcriptional activity[6,7]. Recent advances in genomic language modeling have revealed that latent features extracted from raw DNA correspond to functionally meaningful regulatory elements[8–11]. Supervised classification in high-

[1]State Key Laboratory of Crop Gene Resources and Breeding, Institute of Crop Sciences, Chinese Academy of Agricultural Sciences (CAAS), Beijing, China. [2]Nanfan Research Institute, CAAS, Sanya, China. [3]School of Mathematics and Physics, University of Queensland, Brisbane, QLD, Australia. [4]College of Agriculture, Aksum University, Tigray, Ethiopia. [5]Agricultural Information Institute, CAAS, Beijing, China. e-mail: lihuihui@caas.cn

dimensional sequence space has further extended promoter activity prediction[7,12]. These developments suggest that bidirectional promoter function may be encoded in the statistical structure of genomic DNA itself. Combining computational approaches with empirical understanding of BDP architecture thus provides a basis for asking whether sequence-driven models can recover the organizational rules governing bidirectional transcription.

A major obstacle is the lack of a sequence-based, interpretable method to distinguish truly active BDPs from adjacent genes. Current computational frameworks rely on promoter length cutoffs and cannot directly assess transcriptional competence from DNA[13–17], limiting cross-species applicability and restricting analysis to well-annotated genomes. We hypothesized that if a genomic language model encodes the latent regulatory grammar embedded in nucleotide sequences, then embeddings from its learned representations would allow classifiers to predict BDP activity directly. Here, regulatory syntax describes the local organizational principles of bidirectional transcription; cross-species transfer of these patterns would reflect a higher-order regulatory grammar shared across plant genomes.

In this work, we assemble candidate BDP datasets from four plant genomes and quantify transcriptional coupling via co-expression correlation. Using balanced high- and low-activity BDP cohorts, we train supervised classifiers on Evo2 embeddings and benchmark against *k*-mer frequencies, transcription factor binding matrices, and convolutional networks using raw DNA. We evaluate cross-species generalization through multi-species training and leave-one-species-out testing, validating predicted high-activity BDPs *in planta* in the non-model grass *Spartina alterniflora*. Mechanistic analysis shows that Evo2-derived features capture nucleosome occupancy and transcription factor binding patterns, explaining the model's predictive performance. These findings demonstrate that genomic language models extract functional and evolutionary signatures of bidirectional promoters directly from sequence and identify shared regulatory principles across plant species.

Results

Predicting BDP activity from DNA sequence using supervised machine learning

We hypothesized that BDP activity, defined here as coordinated transcription from both flanking genes, could be predicted directly from DNA sequence. To test this, we developed a workflow to identify candidate BDPs, quantify their transcriptional coupling, generate sequence features via multiple representations, and train classifiers capable of discriminating high- from low-activity bidirectional promoters. The pipeline incorporated gLM-derived embeddings, *k*-mer frequencies, transcription factor binding matrices, and raw sequence encoding, with interpretability analysis of learned features via sparse autoencoders (Fig. 1a–f).

Identifying and characterizing candidate BDPs

We identified candidate BDPs from four plant genomes, namely *Arabidopsis thaliana*, maize (*Zea mays*), rice (*Oryza sativa*), and wheat (*Triticum aestivum*), using proximity criteria. Previous studies employed ≤1 kb spacing between transcription start sites (TSS) as an identification threshold[3], with exceptions in cotton (1500 bp)[16] and yeast (1750 bp)[17], although this criterion requires evaluation against BDP activity. We therefore applied a flexible distance cutoff: ≤1 kb for overlapping genes and ≤20 kb for non-overlapping pairs. This yielded 15,600 candidates from Arabidopsis, 14,182 from rice, 3500 from maize, and 6520 from wheat. BDP candidate length distributions were highly right-skewed across all species, with a prominent peak near 200 bp (Fig. 2a, b).

To quantify BDP activity, we used Pearson correlation coefficients (PCC) between expression profiles of divergently transcribed gene pairs, restricting analysis to pairs expressed in >1000 samples across diverse tissues and developmental stages (Fig. 1b). This metric is biologically justified: functional BDPs should drive coordinated transcription, yielding positive expression correlation, whereas independent regulation or shared promoter silence produces uncorrelated or negative correlations.

Expression data came from the Plant Public RNA-seq Database (PPRD)[18], which provides genome-wide transcriptomes spanning tissue types, developmental stages, and conditions for all four species. To evaluate whether tissue-specific sampling introduced bias, we assessed four complementary metrics, including standard deviation, interquartile range, network stability, and sign consistency, across both global datasets and tissue-restricted subsets (Supplementary Data 1). Global datasets comprised 28,164, 19,664, 11,726, and 5816 samples for *Arabidopsis*, maize, rice, and wheat, respectively, with standard deviations ranging from 0.252 to 0.355 and interquartile ranges from 0.293 to 0.563. Tissue-specific subsets showed markedly higher variance (SD: 0.379–0.518; IQR: 0.581–0.917), confirming that partial sampling would introduce systematic bias. Using the full datasets thus mitigated this confound. Restricting to gene pairs with expression data in >1000 samples further ensured that PCC values reliably captured BDP activity.

We next examined whether BDP length predicts activity. Plotting PCC against candidate length revealed bimodal distributions across species, with Gaussian mixture modeling ($G = 2$) identifying ~1000 bp as a transition point (Fig. 2a, b, and Supplementary Fig. 1a–d). This boundary suggested that BDPs shorter versus longer than 1 kb exhibited distinct regulatory properties. Applying the ≤1 kb criterion thus yielded 4524 high-confidence BDP candidates (Supplementary Data 2) with PCC values ranging from −0.5 to +0.9 and a median near 0.1 (Fig. 2c–f), a broad distribution reflecting both strong bidirectional promoters and genomic regions with proximity but no functional coupling.

Gene Ontology enrichment analysis of genes flanking high-activity BDPs (top 5%, 10%, and 15% by PCC) revealed consistent enrichment in transcriptional regulation, stimulus response, and cellular organization across all species, with both conserved and species-specific functional signatures (Supplementary Fig. 2).

Constructing balanced training datasets

Binary classification requires discrete labels, but PCC values are continuous. Arbitrary thresholds would be subjective and noisy, and soft labeling would blur the model's ability to distinguish extreme regulatory states. We therefore employed extremal sampling: for each species, we ranked candidates by PCC and selected the top x% as high-activity and bottom x% as low-activity, with x denoting the label ratio (LR) parameter (Fig. 1d).

Balanced datasets, containing equal numbers of positive and negative examples, yield performance metrics (accuracy, F_1-score) that faithfully reflect discriminative ability rather than bias toward the majority class, a common problem in imbalanced genomic classification[19]. The LR parameter determines training set size: at LR = 0.01, the top and bottom 1% form a balanced set of 2% of all candidates; at LR = 0.10, the top and bottom 10% comprise 20% of the total. We tested 14 LR values (0.01–0.14, step = 0.01) to assess robustness across training sizes and avoid overfitting to arbitrary cutoffs. This strategy intentionally excludes intermediate-PCC samples, which likely represent ambiguous or context-dependent regulation, thereby focusing on clearly functional and non-functional BDP extremes.

gLM-derived embeddings enable informative BDP activity prediction from DNA sequences

We examine whether DNA sequence alone can reliably predict BDP activity. We constructed classification pipelines using our curated set of 4524 high-confidence BDP candidates (Fig. 1c), extracting features

https://doi.org/10.1038/s41467-026-79328-5

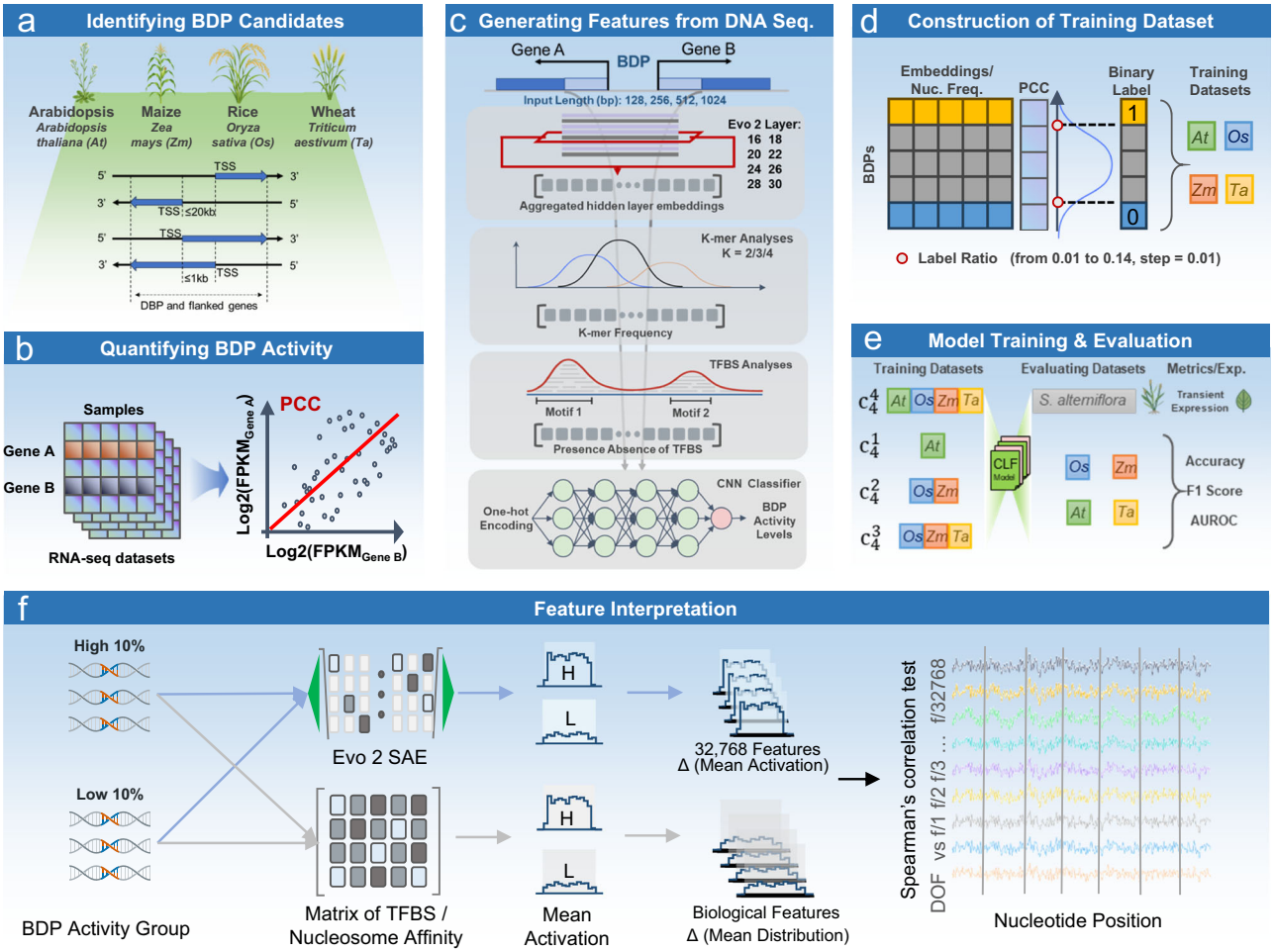


Fig. 1 | Workflow for predicting bidirectional promoter (BDP) activity from DNA sequence. a Identification of candidate BDPs from *Arabidopsis thaliana*, rice (*Oryza sativa*), maize (*Zea mays*), and wheat (*Triticum aestivum*). Divergent gene pairs were selected where transcription start sites (TSSs) were <1 kb apart for overlapping genes or ≤20 kb for non-overlapping pairs. **b** Quantification of BDP activity via Pearson correlation coefficient (PCC) of flanking gene expression profiles. **c** Multiple sequence representations for training: Evo2 gLM embeddings (layers 24, 26, 28, 30; lengths 128, 256, 512, 1,024 bp), *k*-mer frequency profiles (di-, tri-, tetranucleotide), transcription factor binding site matrices, and raw sequences for CNN analysis. **d** Balanced dataset construction using rank-based extremal sampling. BDP candidates were ranked by PCC, with the top *x*% designated high-activity and bottom *x*% designated low-activity, where x is the label ratio (LR). Balanced datasets (50:50 class prevalence) were generated across 14 LR values (0.01–0.14, step = 0.01). **e** Model training and evaluation. Supervised LightGBM classifiers trained on one to three species were tested via leave-one-species-out validation using accuracy, F_1-score, and AUROC. The final four-species model was applied to *Spartina alterniflora* predictions and validated *in planta*. **f** Sparse auto-encoder interpretability workflow. High-activity (top 10%) and low-activity (bottom 10%) BDP groups were stratified by transcriptional coupling strength. Per-nucleotide means of 32,768 Evo2-SAE latent features, nucleosome affinity, and TFBS occupancy were computed within each group to generate differential (Δ) profiles. Spearman rank correlation quantified covariation between SAE features and regulatory signatures.

via four complementary strategies. For each BDP, we obtained DNA sequences of 128, 256, 512, and 1024 bp centered at the locus midpoint. We generated latent embeddings from eight hidden layers (16th, 18th, 20th, 22nd, 24th, 26th, 28th, and 30th) of Evo2. We computed *k*-mer frequencies at scales $k = 2, 3, 4$, and concatenated $k2$–$k4$ profiles. We scanned sequences against 805 plant transcription factor binding site (TFBS) motifs from JASPAR 2024, identified 703 matches, and constructed binary presence–absence matrices for each motif. Finally, we trained convolutional neural networks (CNNs) using one-hot-encoded sequences as end-to-end baselines. The first three feature sets were paired with LightGBM classifiers, while CNNs served as an independent comparison.

We systematically optimized Evo2 inputs by evaluating all combinations of eight layers and four sequence lengths across 14 LR thresholds (448 configurations total). Across metrics (AUROC, F_1-score, accuracy), performance improved monotonically with sequence length, with 1024 bp consistently outperforming shorter windows (Fig. 3a, and Supplementary Fig. 3). Among layers, the 16th through 26th exhibited comparable predictive power, while performance declined at layers 28 and 30. We selected the 24th layer with 1024 bp centered sequences for all subsequent analyses, achieving maximum $F_1 = 0.93$ and sustaining $F_1 > 0.79$ across all LR values (Fig. 3b).

The optimized Evo2 configuration substantially outperformed alternatives across the complete LR spectrum (Fig. 3b, and Supplementary Figs. 4–6). Evo2-derived embeddings achieved higher AUROC at every LR than *k*-mer frequencies (average cross-LR AUROC > 0.7), TFBS matrices (comparable but slightly weaker), or CNNs on raw sequences (consistently weakest across all lengths and label ratios). These results demonstrate that BDP activity is predictable from DNA sequence and that Evo2 embeddings maintain robust performance across varying training-set compositions.

Cross-species generalizability of BDP activity predictions

To assess whether models trained on one species could transfer to others, we employed leave-one-species-out validation (Figs. 1e and 3c–f). Each species served once as a held-out test set while we

https://doi.org/10.1038/s41467-026-79328-5

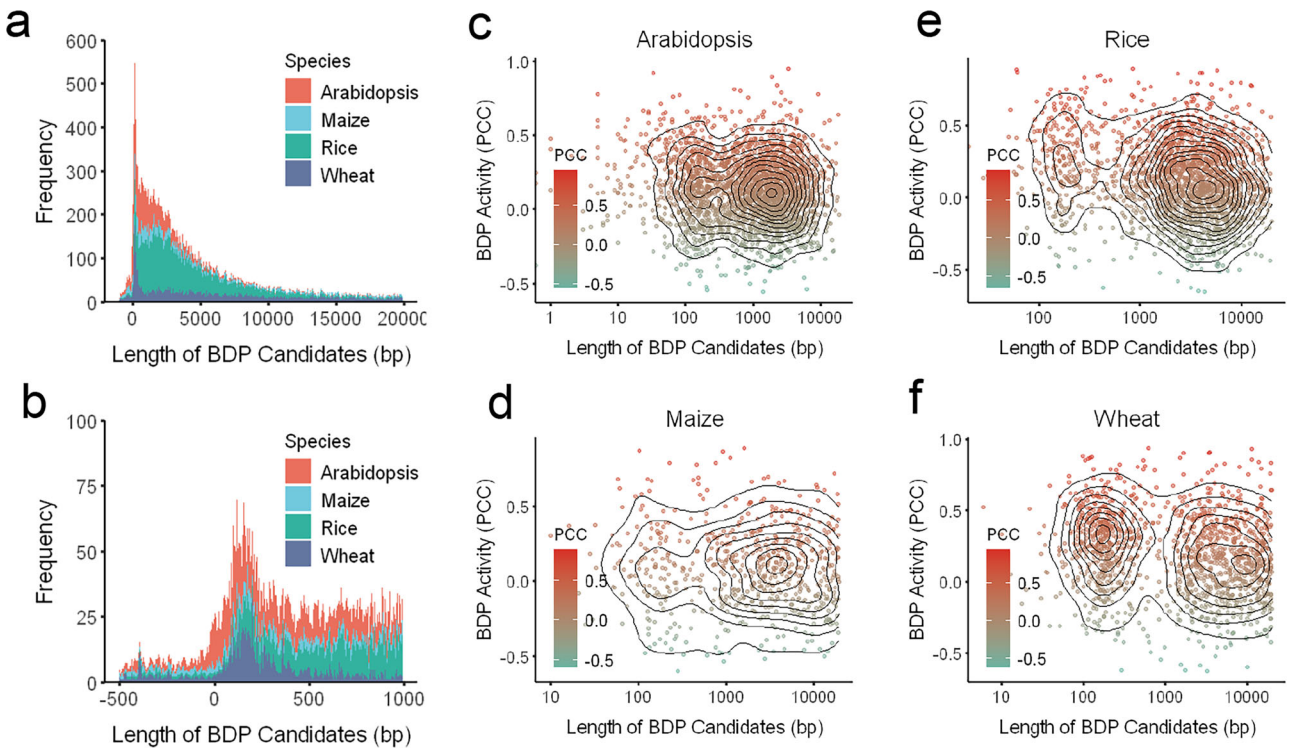


Fig. 2 | BDP candidate identification by sequence length and activity. **a**, **b** Length distributions of candidate BDPs identified using inter-TSS criteria (<1 kb overlapping, ≤20 kb non-overlapping genes). Distributions peak near 200 bp with extended right tails across all species. Expression-based BDP activity (PCC) distributions for *Arabidopsis* (**c**), maize (**d**), rice (**e**), and wheat (**f**) with log10-transformed *x*-axis. Source data are provided as a Source Data file.

trained on different combinations from the remaining three: single-species sets (three configurations), two-species sets (three configurations), and three-species sets (one configuration). All models were evaluated across LR values using AUROC.

Cross-species performance revealed two striking patterns. First, training set size stabilized predictions: at low LR, sparse labeled samples caused substantial AUROC fluctuations; as LR increased, nearly all cross-species models maintained AUROC > 0.5. Second, predictive power improved with additional training species, with multi-species models often outperforming single-species counterparts across all LR thresholds. Phylogenetic distance constrained transferability. Arabidopsis, the sole eudicot, showed the poorest generalization, achieving AUROC < 0.7 even under three-species training at maximum LR (Fig. 3c). Conversely, the three Poaceae monocots transferred robustly among themselves. Wheat as a test species reached peak AUROC ≈ 0.84 under three-species training (Fig. 3f), rice sustained high accuracy across all LR values (Fig. 3e), and maize showed intermediate performance (Fig. 3d). These results establish that Evo2-derived models generalize across plant species with fidelity proportional to phylogenetic relatedness.

In planta evaluation of predicted activity of BDP candidates in *Spartina alterniflora*

Given strong cross-species performance among closely related monocots, we tested predictions in *S. alterniflora*, a halophytic grass in the Poaceae family. We identified 1163 candidate BDPs using the ≤1000 bp criterion, converted their 1024 bp centered sequences to Evo2 24th-layer embeddings, and applied the three-species classifier (maize, rice, wheat) to predict 136 high-activity and 158 low-activity candidates. To validate predictions experimentally, we randomly selected 20 high-activity and 20 low-activity candidates, cloned them into bidirectional dual-fluorescent reporter constructs (BDP-eGFP/mCherry) to drive opposite-strand expression, and included unilateral control constructs for single-strand assessment (Fig. 4a, and Supplementary Fig. 7a). CaMV35S served as positive control[20,21]. We transiently expressed constructs in *Nicotiana benthamiana* leaves and monitored reporter fluorescence.

Unilateral controls confirmed detectable transcriptional activity on both strands for 16 high-activity and 13 low-activity candidates. Bidirectional analysis revealed that among the 16 validated high-activity BDPs, only locus SA_29G013100–SA_29G013200 failed to drive simultaneous reporter expression. Of the remaining 15, three showed substantially different fluorescence between strands. At the same time, 12 exhibited comparable bidirectional expression (Fig. 4). Low-activity candidates performed as predicted: three exhibited coordinated bidirectional expression, while ten showed little or no evidence of dual-strand regulation, consistent with their predicted low activity (Supplementary Fig. 7). These results confirm that models trained on Poaceae genomic data can predict BDP activity in non-model grasses.

Sparse autoencoder reveals nucleosome and transcription factor signatures in Evo2 embeddings

Nucleosome-free chromatin and shared transcription factor binding are recognized determinants of BDP activity across eukaryotes[3]. To establish whether Evo2 embeddings encode these canonical regulatory features, we applied sparse autoencoder (SAE) analysis to map latent representations onto nucleosome affinity and TFBS occupancy profiles (Fig. 1f). We extracted 32,786 SAE-derived latent features from the 24th layer for each BDP candidate, defined high- and low-activity groups as the top and bottom 10% by transcriptional coupling strength, and calculated differential profiles for both SAE features and regulatory signatures. Spearman rank correlation between individual SAE features and regulatory profiles quantified the biological relevance of Evo2 representations.

https://doi.org/10.1038/s41467-026-79328-5

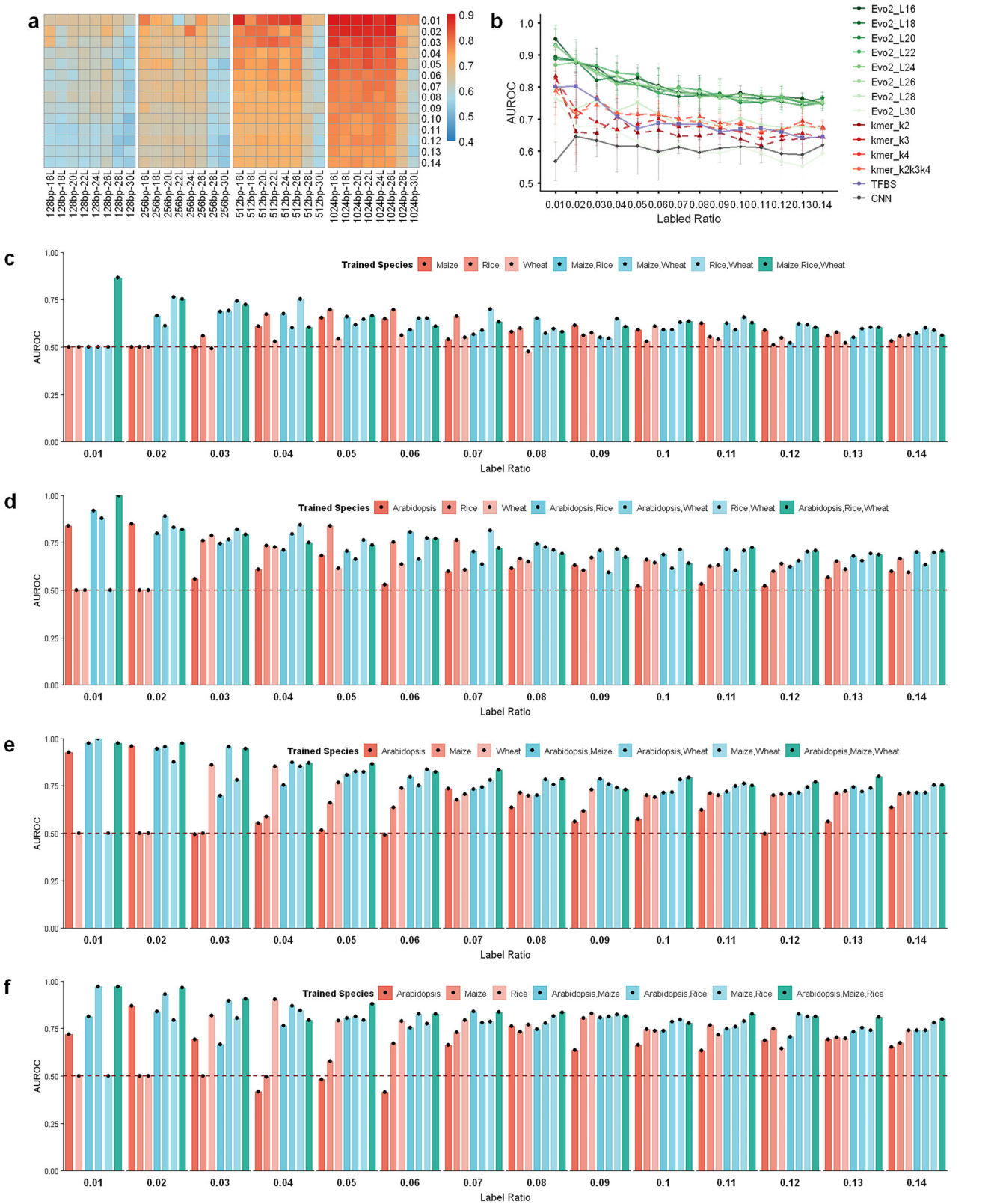


Fig. 3 | gLM embeddings predict BDP activity across plant species. a AUROC heatmap comparing Evo2 hidden layers (16th, 18th, 20th, 22nd, 24th, 26th, 28th, 30th) and input lengths (128, 256, 512, 1024 bp) across 14 label ratios (0.01–0.14, step = 0.01). **b** AUROC comparison of LightGBM classifiers using Evo2 embeddings from layers 16–30, *k*-mer features (*k* = 2, 3, 4 or concatenated 2/3/4-mers), or transcription factor binding site (TFBS) features, and CNN models across label ratios. Points show mean AUROC ± standard deviation (SD) across five cross-validation folds (*n* = 5 folds). The folds are computational evaluation splits, not biological replicates. Cross-species generalization of gLM-embedding classifiers trained on one, two, or three species and tested on the held-out species: *Arabidopsis* (**c**), maize (**d**), rice (**e**), and wheat (**f**). Each bar and overlaid point shows a single held-out-species test-set evaluation (*n* = 1 evaluation per training-species combination and label ratio), using the full labelled training set for that configuration. The dashed horizontal line indicates AUROC = 0.5. Source data are provided as a Source Data file.

https://doi.org/10.1038/s41467-026-79328-5

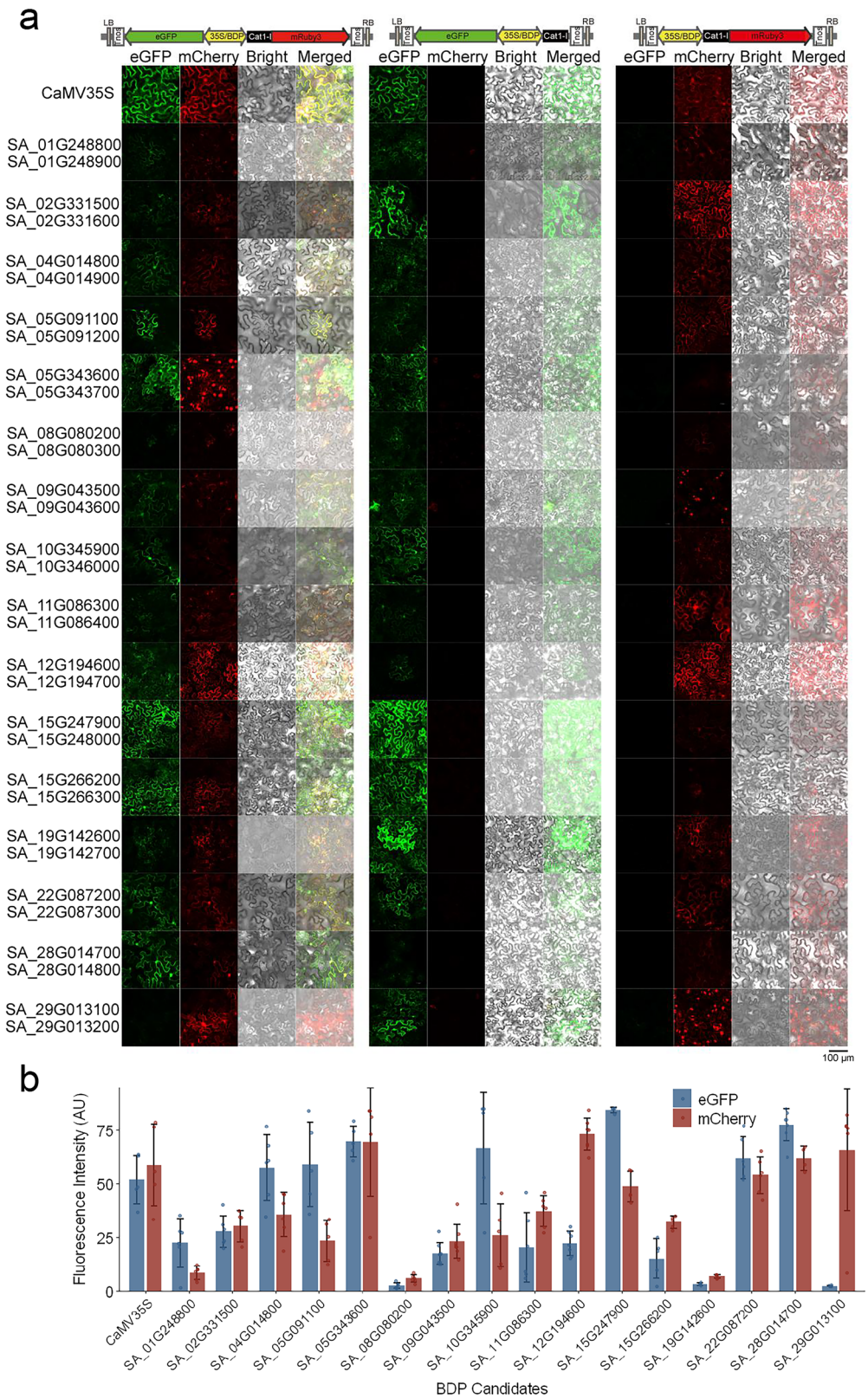


Fig. 4 | Experimental validation of predicted high-activity BDPs in *Spartina alterniflora*. **a** Top: schematic of three reporter constructs, bidirectional vector with candidate BDP driving eGFP and mCherry in opposite orientations, eGFP-only unilateral control, and mCherry-only unilateral control. Bottom: representative confocal images of 16 BDP candidates and a separate CaMV 35S positive control (17 constructs in total) in four channels (eGFP, mCherry, bright field, merged). Scale bar, 100 µm. **b** Fluorescence intensity (arbitrary units). Blue/red bars show mean eGFP/mCherry intensity ± SD; dots represent single-cell measurements from independent infiltration experiments (one cell per independent infiltration region). At least 4 independent experiments were analysed per construct and reporter. Source data are provided as a Source Data file.

gLM-derived latent features capture nucleosome affinity signatures of BDP activity

Nucleosome affinity mapping across 1024 bp BDP-centered sequences revealed conserved patterns differentiated by activity state (Fig. 5a). All four species displayed a valley-shaped nucleosome affinity profile with pronounced minima at the BDP midpoint, reflecting nucleosome-depleted chromatin permissive for bidirectional initiation. High-activity BDPs consistently exhibited lower central nucleosome occupancy than low-activity counterparts, an effect most pronounced in wheat, where low-activity BDPs showed substantially elevated nucleosome enrichment at the core region. This nucleosome landscape divergence indicates that nucleosome-free chromatin facilitates bidirectional transcription.

We linked these chromatin patterns to gLM latent space by computing Spearman rank correlations between differential nucleosome affinity (Δ) and all Evo2-SAE features (Fig. 5b–e). Multiple features correlated significantly with nucleosome landscapes, revealing both conserved and species-specific associations. Feature 27089 exemplified cross-species variation: positively correlated with Δ affinity in *Arabidopsis* ($\rho = 0.361$) but strongly negatively correlated in maize and wheat ($\rho = -0.67$ and -0.638, respectively) (Fig. 5b). Despite opposite correlation directions, this signal indicates the feature encodes sequence properties governing nucleosome occupancy. Feature 15680, by contrast, showed marked species specificity, achieving strong correlation with Δ affinity only in wheat ($\rho = 0.809$) with weak associations elsewhere (Fig. 5e). These findings demonstrate that Evo2 latent features effectively capture nucleosome affinity information from sequence alone, suggesting that nucleosome dynamics underlie bidirectional promoter recognition.

gLM embeddings resolve transcription factor binding landscapes associated with BDP activity

TFBS enrichment analysis identified 123 transcription factors over-represented in high-activity BDP sequences (>50% prevalence per species) across multiple superfamilies, AP2/ERF, MIKC-type MADS, DOF, TCP, MYB, and ARF, with AP2/ERF comprising the largest proportion (Fig. 6a). The three Poaceae species shared 74 enriched TFs dominated by ERF/DREB members, while *Arabidopsis* harbored 27 distinct factors mainly from MADS-box and DOF families, reflecting lineage-specific BDP regulatory architecture between eudicots and monocots (Fig. 6b, and Supplementary Fig. 8). TFBS positioning differed markedly: *Arabidopsis* showed uniform distribution across the 1024 bp interval, whereas maize, rice, and wheat displayed prominent bilateral peaks flanking the BDP center (Fig. 6c).

We assessed whether Evo2 embeddings encode these TFBS signatures by computing differential positional TFBS abundance (Δ) profiles between high- and low-activity groups and correlating them with Evo2-SAE features via Spearman rank correlation (Fig. 6d). Four representative TFs from distinct superfamilies, ABR1, ARF14, DOF3.6, and FLC, revealed complex cross-species patterns. ABR1 showed consistent moderate positive correlation across all species ($\rho = 0.325$–0.424), while DOF3.6 was strongly negatively correlated in *Arabidopsis* ($\rho = -0.532$) but moderately positive in grasses ($\rho = 0.378$–0.420). FLC associations were positive in *Arabidopsis* and maize, negative in rice and wheat. This lineage-specific covariation indicates that Evo2 embeddings capture sequence-encoded TF binding landscapes across divergent plant lineages, providing mechanistic insight into robust cross-species BDP activity prediction.

Discussion

This work demonstrates that genomic language models capture biologically meaningful regulatory principles underlying plant bidirectional promoters and that these principles generalize across species. LightGBM classifiers trained on Evo2 representations captured transcriptional coupling of divergent gene pairs across four plant genomes, indicating that latent features encode functionally relevant regulatory information. Cross-species validation in *S. alterniflora*, confirmed by *in planta* assays, established generalizability beyond training species. Sparse autoencoder analysis further revealed that gLM embeddings capture conserved nucleosome affinity signatures and species-specific transcription factor landscapes governing bidirectional transcription, suggesting these models formalize the regulatory grammar linking sequence organization to promoter function across plant lineages.

Sparse autoencoders (SAEs) have proven powerful for interpreting genomic language models by decomposing dense latent representations into discrete, biologically interpretable features. Accumulating evidence confirms that SAEs recover meaningful sequence signatures, from compositional biases to transcription factor motifs and chromatin accessibility patterns, across multiple gLM architectures, including HyenaDNA and Evo2[9,22–24]. By projecting hidden-layer activations into an overcomplete sparse feature space, SAEs disentangle superimposed regulatory signals within embeddings, providing an interpretable bridge between predictions and molecular mechanisms.

The framework extends naturally to practical applications. By inferring BDP activity from sequence alone, we enable identification of candidate bidirectional promoters in newly assembled genomes where transcriptomic data are scarce. Evo2-SAE feature analysis further permits rational design of synthetic bidirectional promoters by highlighting potentially informative sequence elements. The same workflow adapts readily to other bidirectional or paired regulatory systems, antisense promoters, and enhancer–promoter interfaces[6,9], and could incorporate complementary modalities such as chromatin accessibility or methylation embeddings to enhance resolution without altering the core principle that sequence encodes regulatory logic.

More broadly, this phenotype-guided approach to decoding *cis*-regulatory grammar, stratifying sequences by quantitative phenotypes, decomposing gLM embeddings via SAEs, and correlating differential feature activation with functional signals, offers a generalizable strategy for regulatory genomics. Unlike conventional motif enrichment relying on predefined databases, this data-driven workflow identifies functional sequence features directly from primary DNA and quantitative phenotypic readouts. This strength proves particularly valuable for non-model organisms with limited epigenomic resources, as demonstrated by cross-species prediction and validation in *S. alterniflora*. The framework readily extends to unidirectional promoters, enhancers, silencers, and other regulatory elements, providing a versatile paradigm for comparative regulatory genomics across plant diversity.

Methods

Genomic data and bidirectional promoter (BDP) identification

Genome assemblies and gene annotations (GFF3 format) for *A. thaliana*, *O. sativa*, *Z. mays*, and *T. aestivum* were retrieved from Ensembl Plants (release 55), with human data from Ensembl (release 115). Bidirectional promoter candidates were identified using an in-house R pipeline. Gene coordinates, IDs, chromosomal positions, and strand information were parsed from GFF3 files, sorted by chromosome and position, then screened for oppositely oriented pairs. Candidate BDPs required transcription start site (TSS) distances ≤1 kb for overlapping genes or ≤20 kb for non-overlapping pairs, thresholds chosen to encompass empirically established BDP cutoffs while permitting evaluation of longer intergenic spacing. Genomic sequences for all qualifying candidates were extracted for downstream analysis.

Expression data and BDP activity quantification

Transcriptional coupling between BDP-flanking genes was assessed using PCC between normalized expression profiles. Expression data came from the PPRD[18], comprising 28,164 *Arabidopsis*, 11,726 rice,

https://doi.org/10.1038/s41467-026-79328-5

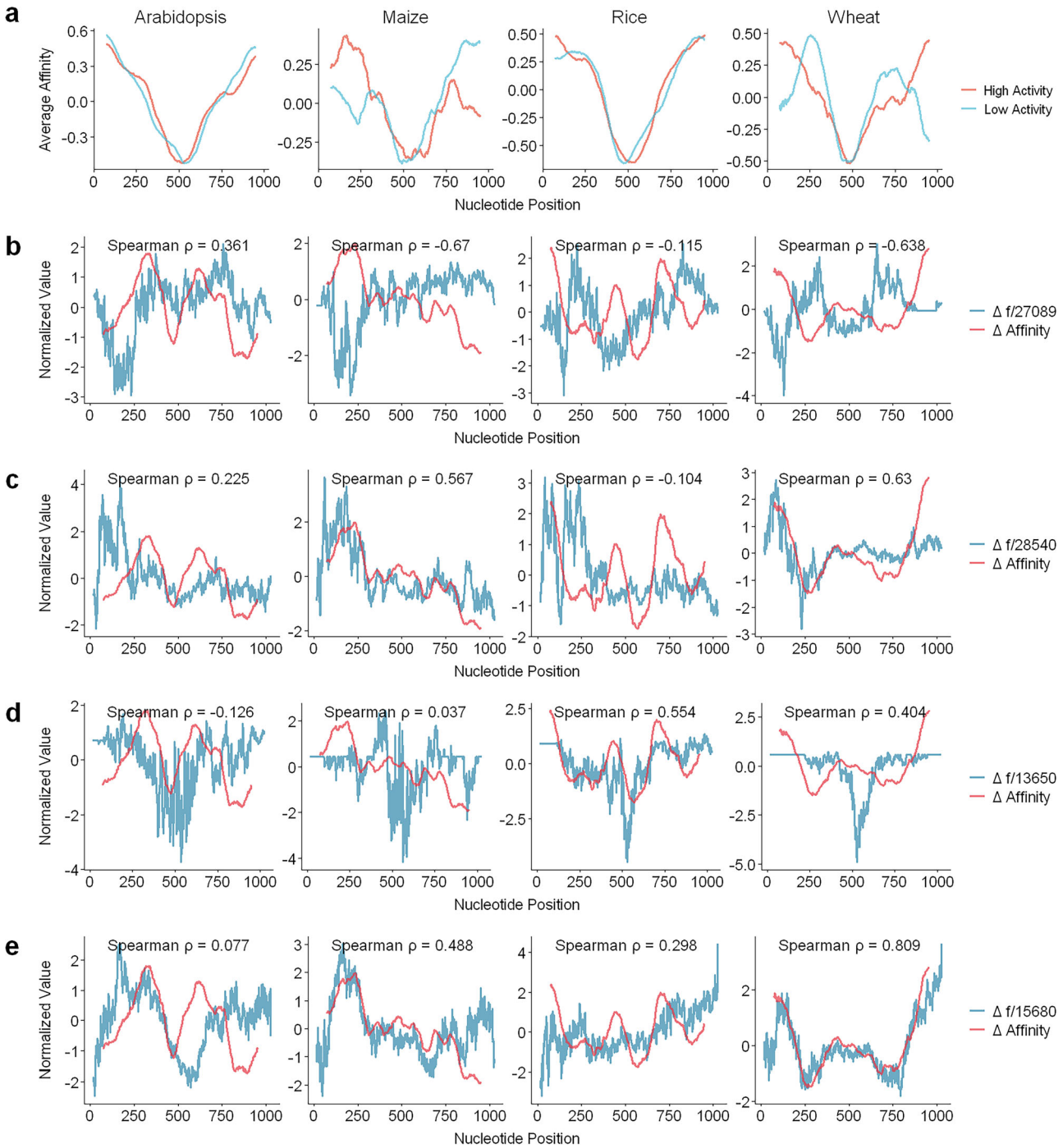


Fig. 5 | Nucleosome affinity signatures of BDP activity are encoded in Evo2 embeddings. a Positional nucleosome affinity across 1024 bp BDP-centered sequences comparing high-activity (red) and low-activity (blue) groups in all four species. **b**–**e** Spearman rank correlation between differential nucleosome affinity (Δ Affinity, red) and representative Evo2-SAE latent features (blue) along sequence positions. Each panel shows one feature with highest absolute correlation coefficients (ρ) for tested species. Source data are provided as a Source Data file.

19,664 maize, and 5816 wheat libraries spanning diverse tissues, developmental stages, and treatments. For each candidate, FPKM values for both flanking genes were log10-transformed and filtered to exclude missing observations. Candidate pairs with fewer than 1000 samples were discarded. Resulting PCC values served as continuous BDP activity metrics: positive correlations indicate coordinated expression consistent with bidirectional promoter function.

Assessing bias in PPRD-derived BDP activity estimates

To evaluate whether tissue-stratified subsets of PPRD introduce systematic bias into PCC-based activity estimates, we compared statistical metrics from tissue-restricted and global datasets using four complementary measures: standard deviation (SD)[25], Interquartile Range (IQR)[25], Network Stability[26], and Sign Consistency[27].

https://doi.org/10.1038/s41467-026-79328-5

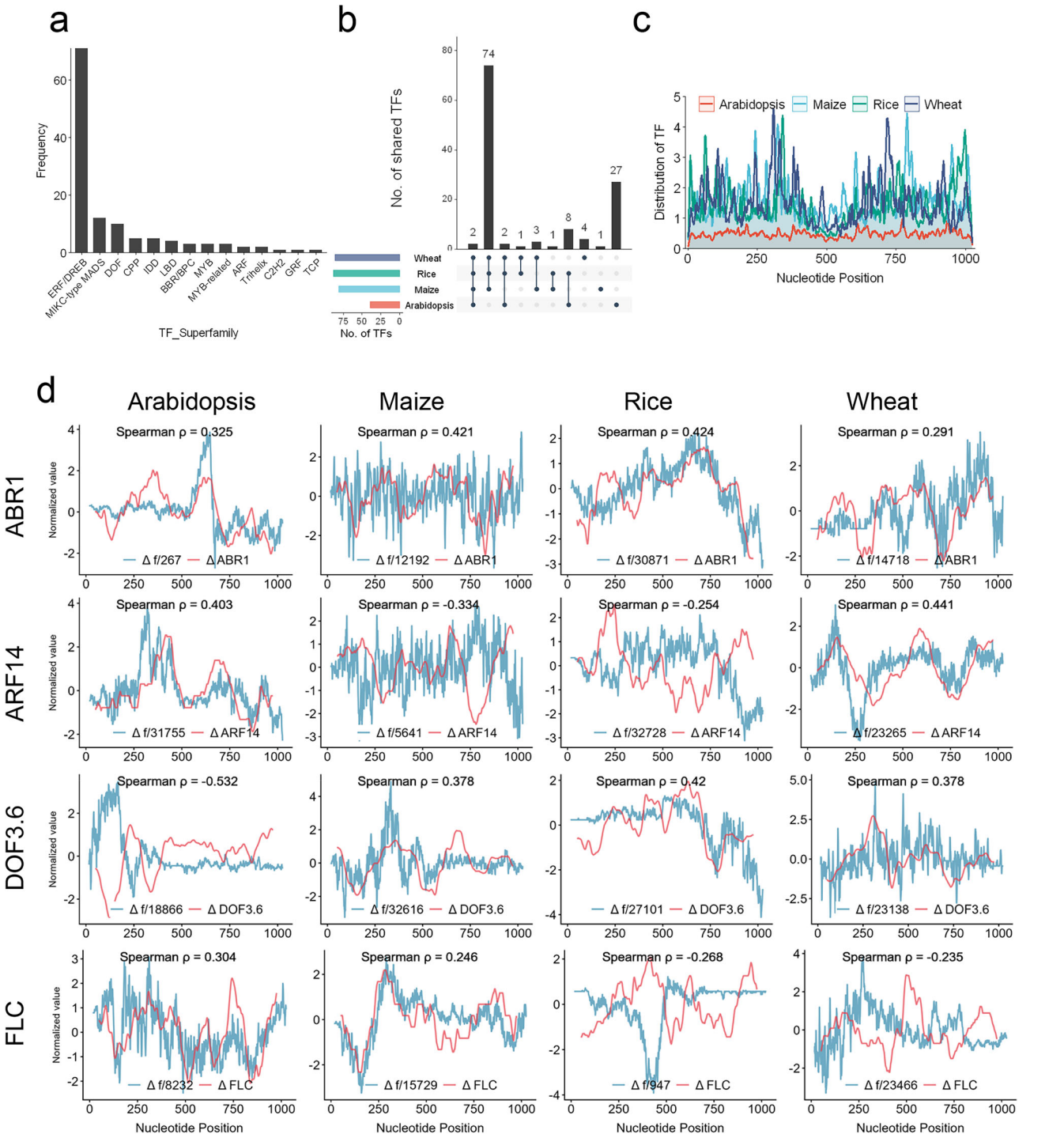


Fig. 6 | TFBS landscapes of BDP activity are encoded in Evo2 embeddings. **a** Superfamily classification and frequency of transcription factors enriched in high-activity BDPs. **b** Cross-species overlap of enriched transcription factors among the four species. **c** Positional TFBS density distribution across 1024 bp BDP-centered sequences for all species. **d** Spearman rank correlation between differential TFBS abundance (red) of four representative transcription factors and Evo2-SAE latent features (blue) along positions, with highest absolute correlation coefficients (ρ) labeled per species. Source data are provided as a Source Data file.

Standard Deviation: SD quantified the overall dispersion of PCC values:

$$SD = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(r_i - barr)^2} \quad (1)$$

where r_i is the correlation coefficient of each gene pair and $\bar{r}$ is the mean value. Smaller SD indicates concentrated, robust correlation estimates with reduced estimation bias.

Interquartile Range: IQR measures the central distribution spread:

$$\mathrm{IQR} = Q_3 - Q_1 \quad (2)$$

where Q_1 and Q_3 are first and third quartiles. IQR is less sensitive to outliers than SD and better reflects central concentration; a smaller IQR suggests reduced tissue-induced bias.

Network Stability: To assess co-expression network robustness under sampling variation via bootstrap resampling. For each tissue and global dataset, 20 samples were randomly selected without replacement (30 replicates), and PCC values were recalculated for all gene pairs. Network stability was defined as mean correlation between vectorized bootstrap correlation matrices:

$$\text{Network Stability} = \frac{1}{N}\sum_{i\neq j} cor\left(C_i, C_j\right) \tag{3}$$

where C_i and C_j are vectors from bootstrap runs i and j, respectively. Higher stability indicates consistent co-expression structure resistant to sampling bias.

Sign Consistency: To evaluate directional robustness of co-expression associations across bootstrap replicates. For each gene pair, the proportion of bootstrap runs with positive correlation was:

$$p = \frac{1}{B}\sum_{b=1}^{B} I(r_{ij}^{b} > 0) \tag{4}$$

where B is the total number of bootstrap runs and $I(\cdot)$ is the indicator function. Sign consistency was summarized as:

$$\text{Sign Consistency} = \text{mean}(2|p-0.5|) \tag{5}$$

This metric ranges from 0 to 1, with higher values indicating stable correlation direction reflecting stronger biological reliability and lower sampling bias. All analyses used R (v4.2.1).

Sequence features for model training were generated as described below. Computational pipelines used Python (v3.14.7), NumPy (v2.5.3), pandas (3.0.6), scikit-learn (v1.9.1) and Biopython (v1.88).

gLM-derived embeddings

Evo2 (v0.6.0), a genomic language model[9], generated latent sequence embeddings for each BDP candidate. DNA sequences centered on the intergenic region were extracted at 128, 256, 512, and 1024 bp lengths, tokenized using the Evo2 tokenizer, and processed through the model to collect hidden-layer activations from layers 16, 18, 20, 22, 24, 26, 28, and 30. Output tensors were aggregated into 4,096-dimensional vectors representing BDP sequences. Grid search over sequence lengths and layers identified optimal configurations for prediction.

K-mer frequency features

We computed *k*-mer frequency profiles from the same centered intergenic sequences as baselines against gLM embeddings[28]. Motif frequencies were independently calculated for $k = 2, 3$, and 4, with each *k*-mer size tested as a standalone feature set and concatenated profiles tested for combined performance.

Transcription factor binding site (TFBS) prediction

TFBS identification used FIMO from the MEME Suite (v5.5.9)[29] to scan BDP sequences against plant core position weight matrices from JASPAR 2024[30], at a significance threshold of $P < 1 \times 10^{-4}$. A frequency matrix documented TFBS occurrences across all candidates, serving dual purposes: as a sequence feature set for BDP activity prediction and as input for positional distribution and enrichment analyses comparing high- and low-activity BDP groups.

Nucleosome affinity prediction

Nucleosome affinity across all 1024 bp sequences was predicted using NuPoP (v2.20.0)[31] with model 4 (nucleosome positioning and kinetic unwrapping parameters). The species parameter was set to 10 for *Arabidopsis* (eudicots) and 11 for monocots. Per-base affinity scores were extracted and averaged within high- and low-activity subgroups to generate positional affinity profiles.

Balanced training datasets

BDP candidates were ranked by PCC within each species, with top *x*% and bottom *x*% defined as high- and low-activity classes respectively, where *x* is the label ratio (LR). Balanced datasets were generated across fourteen LR values (0.01–0.14, step = 0.01), ensuring equal positive and negative samples. This strategy emphasized confident examples while minimizing ambiguity from intermediate PCC values.

Supervised model training and cross-species evaluation

LightGBM (v4.7.0) classifiers[32] were trained on Evo2 embeddings or *k*-mer features across four configurations: (1) single-species training and testing; (2) one-to-one cross-species; (3) two-to-two cross-species; (4) three-to-one leave-one-out. Within-species model evaluation used fivefold cross-validation. Models used fixed hyperparameters: num_leaves = 127, max_depth = 9, min_child_samples = 30, λL1 = λL2 = 1.0, feature_fraction = 0.7, bagging_fraction = 0.8, bagging_freq = 3. Grid searches optimized learning rates (0.01–0.10, step = 0.02) and LR (0.01–0.14, step = 0.01). For cross-species generalization in Fig. 3c–f, each model was fitted once to the full labelled training set for the specified species combination and label ratio, and evaluated once on the held-out species test set to maximize use of the available data.

Performance was evaluated using Accuracy, F_1-score, and Area Under the Receiver Operating Characteristic Curve (AUROC). Accuracy was defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

F_1-score was determined as:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{7}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{8}$$

$$F_1\text{score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \tag{9}$$

AUROC quantified the classifier's capacity to distinguish high- and low-activity BDPs across decision thresholds.

CNN-based BDP activity prediction

CNNs predicted BDP activity from raw sequences as a deep learning baseline. Each 1024 bp sequence was one-hot encoded as a 4×1024 matrix. The CNN followed canonical regulatory element prediction architecture[33]: convolutional layers served as de novo motif detectors; max-pooling reduced dimensionality while preserving key signals; fully connected layers integrated features for binary classification. Models were trained with binary cross-entropy loss optimized by Adam, evaluated via AUROC, F_1-score, and accuracy under fivefold stratified cross-validation matching the LightGBM framework, implemented in PyTorch (v2.14.0).

Experimental validation in *Spartina alterniflora*

S. alterniflora candidates were identified using the ≤1 kb inter-TSS criterion and processed with the optimized gLM configuration (24th layer, 1024 bp). The three-species classifier (maize, rice, and wheat) identified high- and low-activity candidates, from which 20 of each were randomly selected for experimental validation. We constructed the bidirectional dual-fluorescent reporter vectors (BDP-eGFP/

mCherry) and unilateral control vectors using a modified pCAMBIA1301 backbone. In the bidirectional constructs, each candidate BDP region drove eGFP and mCherry expression in opposite orientations. Unilateral controls independently assessed transcriptional activity at each strand: one construct tested eGFP expression alone with no reporter on the opposite side, while another tested mCherry alone. Constructs were transformed into *Agrobacterium* GV3101 for transient expression in *N. benthamiana* leaves, which were cultured at 22 °C under a 16:8-h light:dark photoperiod. Fluorescence was imaged via confocal microscopy (eGFP: 488 nm excitation, 505–530 nm emission; mCherry: 552 nm excitation, 580–620 nm emission). Strand-specific transcriptional strength was quantified using ImageJ (v1.54p) to measure mean gray intensity from manually segmented fluorescent regions within each cell, with background autofluorescence from empty negative controls subtracted from all readings. Each fluorescence data point represented one cell from an independent infiltration region in an independently performed experiment. All usable measurements were included; measurements affected by tissue damage or detection failure were not included. The number of independent experiments therefore varied between constructs and reporter channels; fluorescence intensity is summarized as mean ± SD.

Sparse autoencoder (SAE) analysis of gLM latent features

To interpret learned representations, we applied the Evo2 Sparse Autoencoder (BatchTopK SAE) trained on residual stream activations from layer 24 of the Evo2 7B model, projecting 4096-dimensional activation vectors into a 32,768-dimensional sparse feature space ($k = 64$)[9]. For each species, Evo2 embeddings of high- and low-activity BDPs were processed to extract sparse feature vectors. Features with activation density below 1×10^{-5} were considered inactive.

BDP sequences were partitioned into high-activity (top 10%) and low-activity (bottom 10%) groups by transcriptional coupling strength. Per-nucleotide mean profiles were computed for SAE latent features and two regulatory signatures, nucleosome occupancy affinity and TFBS occupancy, within each group. Differential (Δ) profiles were generated by subtracting low-activity means from high-activity means. Spearman rank correlation between individual SAE features and each regulatory profile quantified the biological relevance of learned representations.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All transcriptome datasets supporting the calculation of bidirectional promoter activity were retrieved from the public Plant Public RNA-seq Database (https://plantrnadb.com/): *Arabidopsis* (https://plantrnadb.com//athrdb/download/gene_FPKM_200501.csv.gz), maize (https://plantrnadb.com//zmrna/download/fpkmAllData.csv.gz), rice (https://plantrnadb.com//ricerna/download/rice.FPKM.csv.gz), wheat (https://plantrnadb.com//wheatrna/download/fpkmAllData.csv.gz). Genomic assembly and gene annotation files for *Arabidopsis thaliana*, *Oryza sativa*, *Zea mays*, and *Triticum aestivum* were downloaded from Ensembl Plants release 55 (https://ftp.ensemblgenomes.ebi.ac.uk/pub/plants/release-55/). Processed high-confidence bidirectional promoter candidate datasets generated in this study are deposited alongside the source code repository [https://github.com/AIBreeding/BDP-identifier]. Source data are provided with this paper.

Code availability

Genomic information of BDP candidates and code developed in this study is available at Github [https://github.com/AIBreeding/BDP-identifier][34]. The repository includes Python pipelines for extracting BDP flanking sequences, generating Evo2 genomic language model embeddings, *k*-mer feature calculation, LightGBM classification model training, cross-species prediction evaluation, and sparse autoencoder interpretability analysis, together with a readme file and preprocessed candidate BDP tables.

References

1. Adachi, N. & Lieber, M. R. Bidirectional gene organization: a common architectural feature of the human genome. *Cell* **109**, 807–809 (2002).
2. Duttke, S. H. et al. Human promoters are intrinsically directional. *Mol. Cell* **57**, 674–684 (2015).
3. Nemsick, S. & Hansen, A. S. Molecular models of bidirectional promoter regulation. *Curr. Opin. Struct. Biol.* **87**, 102865 (2024).
4. Xu, Z. et al. Bidirectional promoters generate pervasive transcription in yeast. *Nature* **457**, 1033–1037 (2009).
5. Wei, W., Pelechano, V., Järvelin, A. I. & Steinmetz, L. M. Functional consequences of bidirectional promoters. *Trends Genet.* **27**, 267–276 (2011).
6. Barbadilla-Martínez, L., Klaassen, N., van Steensel, B. & de Ridder, J. Predicting gene expression from DNA sequence using deep learning models. *Nat. Rev. Genet.* **26**, 666–680 (2025).
7. Benegas, G., Ye, C., Albors, C., Li, J. C. & Song, Y. S. Genomic language models: opportunities and challenges. *Trends Genet.* **41**, 286–302 (2025).
8. Yu, H. et al. An interpretable RNA foundation model for exploring functional RNA motifs in plants. *Nat. Mach. Intell.* **6**, 1616–1625 (2024).
9. Brixi, G. et al. Genome modelling and design across all domains of life with Evo 2. *Nature* **652**, 1349–1361 (2026).
10. Avsec, Ž et al. Advancing regulatory variant effect prediction with AlphaGenome. *Nature* **649**, 1206–1218 (2026).
11. Zhai, J. et al. Cross-species modeling of plant genomes at single-nucleotide resolution using a pretrained DNA language model. *Proc. Natl. Acad. Sci. USA* **122**, e2421738122 (2025).
12. Wang, X. et al. Accelerating promoter identification and design by deep learning. *Trends Biotechnol.* **43**, 3071–3087 (2025).
13. Liu, S. J., Yue, Q. J. & Zhang, W. Structural and functional analysis of an asymmetric bidirectional promoter in *Arabidopsis thaliana*. *J. Integr. Plant Biol.* **57**, 162–170 (2015).
14. Vogl, T. et al. Engineered bidirectional promoters enable rapid multi-gene co-expression optimization. *Nat. Commun.* **9**, 3589 (2018).
15. Chen, Y., Li, H., Li, Y.-Y. & Li, Y. Pan-cancer analysis of head-to-head gene pairs in terms of transcriptional activity, co-expression and regulation. *Front. Genet.* **11**, 560997 (2021).
16. Yang, J. et al. Identification and functional characterization of bidirectional gene pairs and their intergenic regions in cotton. *BMC Plant Biol.* **24**, 829 (2024).
17. Jin, Z. et al. Unraveling the regulatory dynamics of bidirectional promoters for modulating gene co-expression and metabolic flux in Saccharomyces cerevisiae. *Nucleic Acids Res.* **53**, gkaf511 (2025).
18. Yu, Y., Zhang, H., Long, Y., Shu, Y. & Zhai, J. Plant public RNA-seq database: a comprehensive online database for expression analysis of~ 45 000 plant public RNA-seq libraries. *Plant Biotechnol. J.* **20**, 806 (2022).
19. Johnson, J. M. & Khoshgoftaar, T. M. Survey on deep learning with class imbalance. *J. Big Data* **6**, 27 (2019).
20. Lv, X. et al. Construction vascular-specific expression bi-directional promoters in plants. *J. Biotechnol.* **141**, 104–108 (2009).
21. Xie, M., He, Y. & Gan, S. Bidirectionalization of polar promoters in plants. *Nat. Biotechnol.* **19**, 677–679 (2001).
22. Nguyen, E. et al. Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. *Adv. Neural Inf. Process. Syst.* **36**, 43177–43201 (2023).

23. Guan, H., He, J. & Zhang, J. Sparse autoencoders reveal interpretable structure in small gene language models. Preprint at https://arxiv.org/abs/2507.07486 (2025).
24. Kmiec, E., O'Brien, S. & McCoy, M. Hidden state genomics: Graph-based analysis of sparse auto-encoder feature activity in genomic language models. Preprint at *bioRxiv* https://www.biorxiv.org/content/10.64898/2026.05.13.725007v1 (2026).
25. Pearson, K. VII. Mathematical contributions to the theory of evolution.—III. Regression, heredity, and panmixia. in *Philosophical Transactions of the Royal Society of London Series A, Containing Papers of a Mathematical or Physical Character*. 253–318 (1896).
26. Shannon, C. P. et al. SABRE: A method for assessing the stability of gene modules in complex tissues and subject populations. *BMC Bioinform.* **17**, 460 (2016).
27. Patil, A., Nakai, K. & Kinoshita, K. Assessing the utility of gene co-expression stability in combination with correlation in the analysis of protein-protein interaction networks. *BMC Genomics* **12**, S19 (2011).
28. Meng, X. et al. Predicting transcriptional responses to cold stress across plant species. *Proc. Natl. Acad. Sci. USA* **118**, e2026330118 (2021).
29. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: Scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018 (2011).
30. Rauluseviciute, I. et al. JASPAR 2024: 20th anniversary of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **52**, D174–D182 (2024).
31. Xi, L. et al. Predicting nucleosome positioning using a duration Hidden Markov Model. *BMC Bioinform.* **11**, 346 (2010).
32. Ke, G. et al. LightGBM: a highly efficient gradient boosting decision tree. *Adv. Neural Inf. Process. Syst.* **30** (2017).
33. Alipanahi, B., Delong, A., Weirauch, M. T. & Frey, B. J. Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. *Nat. Biotechnol.* **33**, 831–838 (2015).
34. Gao, S. et al. *Genomic Language Models Decode Bidirectional Promoter Activity and Reveal Conserved Regulatory Patterns Across Plant Species*. BDP-identifier, https://doi.org/10.5281/zenodo.23056184, (2026).

Author contributions

H.L. conceived the original idea and designed the work. S.G. and H.L. wrote the manuscript. S.G., T.Y., M.A.F., J.R.W. and T.W.G. conducted data analyses. S.C., Z.H., J.G. and Y.J. conducted experiments. J.K.W. and G.J.M. provided constructive advice on the experimental design and the interpretation of the results. All authors read and approved the final manuscript.

Funding

This work was supported by the National Natural Science Foundation of China (32361143514), the Hainan Provincial Natural Science Foundation of China (624MS119), the Key R&D Programs of Hainan Province (ZDYF2024XDNY210), and the Basic Research Center, Innovation Program of Chinese Academy of Agricultural Sciences (CAAS-C-SIAF-202303 and CAAS-ZDRW202503).

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-026-79328-5.

Correspondence and requests for materials should be addressed to Huihui Li.

Peer review information *Nature Communications* thanks Marco Herde, Rui Wang and the other, anonymous, reviewer for their contribution to the peer review of this work. A peer review file is available.

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

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

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

© The Author(s) 2026

This Article in Press is shared early to give you faster access to new research. It is citable and carries a permanent DOI. The final edited version will replace it automatically.