PlantCAD2: A Long-Context DNA Language Model for Cross-Species Functional Annotation in Angiosperms

Jingjing Zhai[1*+], Aaron Gokaslan[2*], Sheng-Kai Hsu[1], Szu-Ping Chen[3], Zong-Yan Liu[3], Edgar Marroquin[2], Eric Czech[5], Betsy Cannon[5], Ana Berthel[1], M. Cinta Romay[1,3], Matt Pennell[4], Volodymyr Kuleshov[2+], Edward S. Buckler[1,3,6]

1 Institute for Genomic Diversity, Cornell University, Ithaca, NY USA 14853
2 Department of Computer Science, Cornell University, Ithaca, NY, USA 14853
3 Section of Plant Breeding and Genetics, Cornell University, Ithaca, NY USA 14853
4 Department of Computational Biology, Cornell University, Ithaca, NY, USA 14853
5 Open Athena AI Foundation, New York, NY, USA 10001
6 USDA-ARS; Ithaca, NY, USA 14853

* These authors contributed equally to this work

+ To whom correspondence may be addressed. Email: jz963@cornell.edu and vk379@cornell.edu

Abstract

Understanding how DNA sequence encodes biological function remains a fundamental challenge in biology. Flowering plants (angiosperms), the dominant terrestrial clade, exhibit maximal biochemical complexity, extraordinary species diversity (over 100,000 species), relatively recent origins (~160 million years), ~200-fold variation in genome size and relative compact coding regions compared with other eukaryotes. These features present both a unique challenge and opportunity for pre-training DNA language models to understand plant-specific evolutionary conservation, regulatory architectures and genomic functions. Here, we introduce PlantCAD2, a long-context, plant-specific DNA language model with single-nucleotide resolution, pre-trained on 65 angiosperm genomes, together with a series of public benchmarks for evaluation. Comprehensive zero-shot testing shows that PlantCAD2 (676 million parameters) efficiently captures evolutionary conservation, surpassing the 7-billion-parameter Evo2 model in 10 of 12 tasks. With parameter-efficient fine-tuning, PlantCAD2 also outperforms the 1-billion-parameter AgroNT across seven cross-species tasks. Moreover, its 8 kb context window substantially improves accessible chromatin prediction in large genomes such as maize (AUPRC increasing from 0.587 to 0.711), underscoring the importance of long-range context for modeling distal regulation. Together, these results establish PlantCAD2 as a powerful, efficient, and versatile foundation model for plant genomics, enabling accurate genome annotation across diverse species.

Introduction

Deciphering how DNA sequence encodes molecular functions, phenotypes and fitness remains a fundamental goal in biology. The rapid decline in sequencing costs has enabled large-scale initiatives such as the Darwin Tree of Life project [1], the Earth BioGenome Project [2], the Vertebrate Genomes Project [3], and the 10KP Plant Genome Project [4], which collectively aim to sequence tens of thousands of species across the tree of life, with plants alone contributing over a thousand assembled genomes [5]. While genomic data accumulates exponentially, functional annotations lag far behind, particularly in plants where labeled data exists for only a few model species and crops [6]. This gap highlights the critical need for computational models that can learn from raw sequences alone and transfer knowledge to other plant species.

Recent advances in foundation models pre-trained with self-supervised strategy have opened new possibilities for interpreting genomic sequences at scale [7]. Unlike traditional supervised machine learning approaches, which typically require large amounts of labeled data, foundation models are pre-trained on vast collections of unlabeled sequence data. This is particularly advantageous in biology, especially plant biology, where high-quality labeled datasets are often limited. Foundation models can then be fine-tuned on specific downstream tasks using only a small number of labeled examples. This approach has achieved significant success in protein science, where models such as ESM [8–10], ProtTrans [11], and ProBERT [12] have demonstrated strong performance in predicting protein function [13], structure [14], and variant effect [15].

In contrast, genomic LMs are still rapidly evolving, with recent developments spanning DNA [16–25], RNA [26–28] and transcriptomes [29–31]. Among DNA LMs, early work such as the DNABERT [16] pre-trained BERT [32] model on the human genome showed improved performance in predicting regulatory sequence elements compared to supervised models such as DeepSEA [33], DanQ [34] and Basset [35]. Subsequently, more general-purpose DNA LMs have emerged: Evo (pre-trained on all prokaryotic and phage genomes) [18], Evo2 (pre-trained on genomes across tree of life) [19], and Nucleotide Transformer (pre-trained on 850 genomes excluding plants) [25] have demonstrated success across a wide range of tasks, from regulatory element discovery to evolutionary constraint prediction. Notably, multi-species pretraining has proven particularly important for learning evolutionary conservation [17,20,25,36]. Plant-specific models have also emerged. GPN [20], pre-trained on eight Brassicales genomes using a 25-layer convolutional neural network, demonstrated strong performance in variant effect prediction. While AgroNT [22], pre-trained on 48 genomes and modeled regulatory sequences using longer context windows with a non-overlapping k-mer encoding method. To further improve plant-specific genome modeling, we previously developed PlantCaduceus [23] (PlantCAD), a DNA LM pre-trained on 16 divergent angiosperm genomes. It uses the Caduceus [24] architecture, a Mamba-based [37] design that efficiently models both DNA strands simultaneously. PlantCAD achieved up to a 7-fold improvement over the next-best model in cross-species gene annotation tasks and variant effect prediction tasks.

However, PlantCAD is limited by its context window of 512 base pairs, restricting its ability to model many biological processes that depend on long-range sequence information [38]. Many regulatory elements can influence gene expression over tens to hundreds of kilobases and they are key contributors to phenotypic variation [39–41], yet remain challenging to capture with short-context models. Therefore, we hypothesize that PlantCAD would struggle to model chromatin accessibility and gene expression patterns across diverse genomic contexts due to the lack of comprehension in long-range regulatory interactions. While AgroNT extends the context window to 6,000 bp, its non-overlapping k-mer tokenization strategy sacrifices single-nucleotide resolution, making it unsuitable for tasks requiring base-level precision such as variant effect prediction [20,22,23]. In contrast, Evo2 is a general-purpose DNA language model with single-nucleotide resolution, but its massive size (7-40 billion parameters) limits accessibility and, more importantly, its pre-training across the entire tree of life makes it less-suited to capture the plant-specific regulatory patterns. Unlike animal genomes which rely on complex long-range regulation, plants have larger and more variable genomes with expanded gene families and diverse metabolic repertoires. Flowering plants (angiosperms) [42], the dominant terrestrial clade [43], exhibit maximal biochemical complexity, extraordinary species diversity, and wide genome size variation, yet maintain relatively compact coding regions, making them both a challenge and an ideal testbed for plant-specific DNA language models.

In this paper, we introduce PlantCAD2, an improved DNA LM tailored to angiosperm genomes. PlantCAD2 is pre-trained using a masked language modeling objective on 65 curated flowering plant genomes. PlantCAD2 is built on the efficient Mamba2 architecture [45], which scales linearly with sequence length instead of quadratically such as transformers [46]. It supports 8,192-bp input windows and reverse-complement equivariance, allowing the model to capture long-range, strand-invariant regulatory features. To reduce pretraining bias, we applied sampling strategies that both down-weight repetitive sequences and emphasize coding and regulatory regions.

Subsequently**,** we first evaluated PlantCAD2 on 12 comprehensive benchmarks using a zero-shot strategy, demonstrating its efficiency and capacity to capture evolutionary conservation (**Table 1**). We then fine-tuned the model on seven functional genomics tasks including chromatin accessibility, gene expression, and protein abundance to further demonstrate its state-of-the-art cross-species predictive ability (**Table 2**). Together, these results highlight PlantCAD2's ability to generalize across species and tasks, and to serve as a versatile foundation model for plant genome interpretation.

Table 1. Zero-shot evaluation summary compared with the best-performing benchmark models. For each task, the bold and underscored value indicates the highest score.

Category	Task	Description	Metric	PlantCAD2 vs best benchmark
Cross-species evolutionary conservation (Figure 2)	Conservation within Andropogoneae (Genome-wide)	Predict conserved vs non-conserved sites using alignments within 35 Andropogoneae genomes (n = 19,030 vs 19,030)	AUROC	**0.725** vs 0.691
	Conservation within Poaceae (non-TIS)	Predict conserved vs non-conserved coding sites (excluding TIS) within Poaceae (n=103,368 vs 80,317)		0.713 vs **0.822**
	Conservation within Poaceae (TIS)	Predict conserved vs non-conserved TIS sites (n=26,650 vs 10,012)		**0.670** vs 0.551
Key junction recovery (Figure 3)	Translation initiation site (maize)	Recover masked ATG start codon (n = 39,035)	Accuracy	**0.657** vs 0.447
	Translation termination site (maize)	Recover masked TAG/TAA/TGA stop codon (n = 39,035)		**0.410** vs 0.256
	Splice donor (maize)	Recover masked GT motif (n = 153,869)		**0.910** vs 0.741
	Splice acceptor (maize)	Recover masked AG motif (n = 153,869)		**0.900** vs 0.738
Within-species conservation (Figure S3)	Translation initiation site (maize)	Predict core TIS vs non-core TIS in maize (n = 28,291 vs 8,118)	AUROC	**0.710** vs 0.624
	Translation termination site (maize)	Predict core TTS vs non-core TTS in maize (n = 28,291 vs 8,118)		0.618 vs **0.628**
	Splice donor (maize)	Predict core splice donor vs non-core splice donor in maize (n = 123,183 vs 21,367)		**0.808** vs 0.754
	Splice acceptor (maize)	Predict core splice acceptor vs non-core splice acceptor in maize (n = 123,183 vs 21,367)		**0.836** vs 0.761
Structural variant effect (Figure 4)	Structural variant effect prediction	Predict conserved deletions vs non-conserved deletions (n = 7,662 vs 10,413)	AUPRC	**0.841** vs 0.771

Note: For key junction recovery and within-species conservation, comparable results were observed in tomato.

Table 2. Fine-tuning evaluation summary compared with the best-performing benchmark models. For each task, the bold and underscored value indicates the highest score.

Category	**Task**	**Dataset (train→test)**	**Task Type**	**Metric**	**PlantCAD2 vs best benchmark**
Chromatin accessibility (Figure 5)	Cross-species accessible regions	Arabidopsis→ 10 species	Binary classification	AUPRC (mean across 10)	**0.409** vs 0.340
	Cross-species accessible regions (multi-species)	9 species→ 2 species	Binary classification	AUPRC (mean across 2)	**0.570** vs 0.499
	Cell-type-specific accessible regions	Maize (hold-out chr10)	Multi-label classification	AUPRC (mean across 92 cell types)	**0.662** vs 0.650
Gene Expression (Figure 6)	Cross-species leaf gene on/off prediction	15 Andropogoneae → 26 NAM genomes	Binary classification	AUROC	**0.854** vs 0.819
	Cross-species leaf absolute gene expression	15 Andropogoneae → 26 NAM genomes	Regression	Spearman correlation	**0.633** vs 0.616
Protein Translation (Figure 6)	Cross-species leaf translation on/off prediction	Arabidopsis → Maize	Binary classification	AUROC	**0.692** vs 0.597
	Cross-species leaf absolute translation abundance	Arabidopsis → Maize	Regression	Spearman correlation	**0.321** vs 0.181

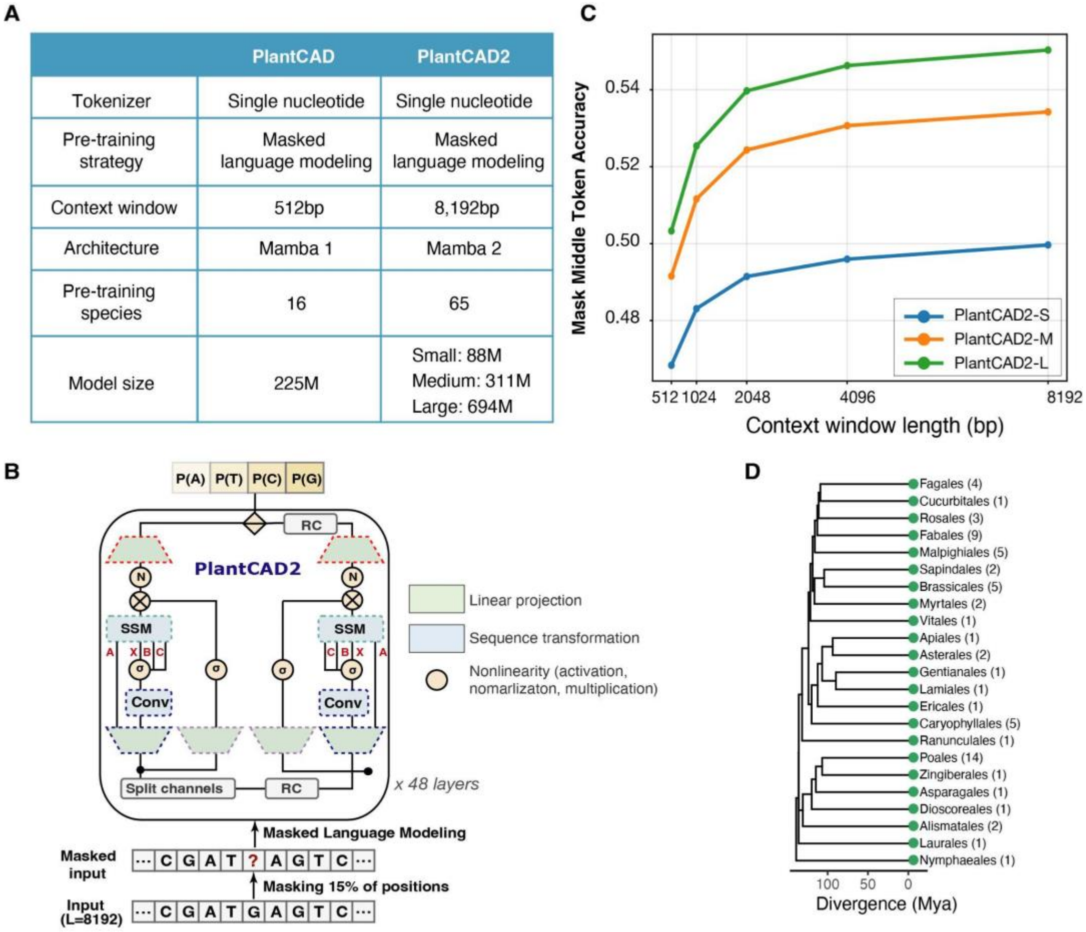
Note: sample sizes (n) for training and test sets are provided in Supplementary Tables.

Results

PlantCAD2: a long-context DNA language model for angiosperms

PlantCAD2 builds on the original PlantCAD [23] DNA language model, preserving its single-nucleotide tokenization and masked language modeling objective, while introducing four major improvements: architectural efficiency, context length, parameter scale, and phylogenetic breadth (**Figure 1A**).First, PlantCAD2 retains the Caduceus [24] architecture with its bidirectional, reverse-complement-equivariant design, but replaces the original Mamba [37] blocks with Mamba2 blocks [45]. Mamba2 introduces substantial improvements over Mamba1, leveraging structured state space duality for more efficient parallel training and simplifying recurrence computations to reduce memory usage (see Methods). Compared to traditional transformer architectures [46,47], PlantCAD2 model architecture shows a much slower increase in inference time than modernBERT models under the same input and output dimensions (**Figure S1A**), due to the inherent efficiency of state space models in handling long sequences [37,48]. Exploiting this efficiency, PlantCAD2 takes 8,192 base pair (bp) windows, which is a 16-fold increase over the 512-bp windows used in PlantCAD. Second, to evaluate the effect of model sizes on performance, we trained a series of depth-scaled PlantCAD2 models of 88M, 311M and 694M parameters (**Fig. 1A-1B**), which we named PlantCAD2-S, PlantCAD2-M, and PlantCAD2-L respectively. As expected, following pre-training, the largest model (PlantCAD2-L) demonstrated the best masked token prediction accuracy (0.657), followed by PlantCAD2-M (0.641) and PlantCAD2-S (0.598), when evaluated on hold-out test set by randomly masking 15% of nucleotides per sequence (**Figure S1B**). However, the largest model also shows slowest inference speed, reflecting the typical trade-off between accuracy and computational efficiency (**Figure S1C**). Despite differences in model size, the three models showed high correlation in their per-species prediction accuracies (r > 0.97), suggesting consistent learning patterns across scales (**Figure S1D; Supplemental Table 1**). Third, to assess the effect of input length on pretraining accuracy, we varied the context window size from 512bp to 8,192bp and evaluated performance by masking the central token. All three models showed improved masked token prediction accuracy with longer contexts, underscoring the importance of extended context for modeling kilobase-scale genomic dependencies (**Figure 1C**). Lastly, we expanded the evolutionary diversity of the training dataset from 16 to 65 angiosperm genomes (**Figure 1D; Supplemental Table 1**), selecting one representative species per genus to maximize phylogenetic breadth. When analyzing pre-training performance across species, we found a weak positive correlation between genome size and masked token accuracy (r = 0.525; **Figure S1E,**

Supplemental Table 1). This relationship is likely driven by the fact that larger genomes tend to contain more repetitive sequences [49]. Since the masked language modeling objective can predict repetitive elements more easily than non-repetitive elements even after applying down-sampling and down-weighting (see Methods), the amount of repeats in the test set could inflate accuracy [20,23,36]. Consistent with our expectation, we also detected a positive correlation between the number of repeats in the test set and masked language modeling accuracy (**Figure S1F**), this also highlights the importance of down-sampling and down-weighting repetitive sequences [36] in pre-training DNA language models.



	PlantCAD	PlantCAD2
Tokenizer	Single nucleotide	Single nucleotide
Pre-training strategy	Masked language modeling	Masked language modeling
Context window	512bp	8,192bp
Architecture	Mamba 1	Mamba 2
Pre-training species	16	65
Model size	225M	Small: 88M Medium: 311M Large: 694M

Figure 1. Overview of the PlantCAD2 model. (A) Comparison of PlantCAD1 and PlantCAD2 model configurations. PlantCAD2 introduces a longer context window, upgraded architecture (Mamba2), expanded pre-training species set, and scaled model sizes (small: 88M, medium: 311M, large: 694M parameters), while maintaining single-nucleotide tokenization. **(B)** Schematic of the PlantCAD2 architecture based on Mamba2 with reverse-complement (RC) equivariance, convolutional and state space modules (SSM), and a masked language modeling objective applied to 8,192 bp input sequences. **(C)** Effect of context window length on model performance. The y-axis shows the prediction accuracy of three models when masking the single central token in the held-out test set. **(D)** Phylogenetic distribution of the 65 angiosperm genomes across flowering plant orders. Numbers in parentheses indicate the number of species included from each order.

PlantCAD2 accurately predicts evolutionary conservation with zero-shot strategy

Evolutionary conservation, commonly estimated through multiple sequence alignment (MSA), is widely used to identify deleterious mutations that may reduce organismal fitness [50–53]. However, genome-wide MSA is particularly challenging in plants due to extensive transposable element (TE) insertions and their high turnover rate, which obscure orthologous relationships outside of conserved coding regions [54]. This limitation highlights the need for alignment-free approaches to assess conservation across diverse plant genomes. Given that PlantCAD2 is pre-trained on 65 evolutionary distant species, we hypothesize that PlantCAD2 can be used to predict evolutionary conservation without multiple sequence alignment. We first evaluated how accurate PlantCAD2 is to distinguish highly conserved sites versus less conserved sites using a zero-shot strategy. As illustrated in **Figure 2A,** we used the masked nucleotide/token prediction accuracy from the frozen model to represent per-base conservation, which means highly conserved bases would receive higher predicted probabilities for the reference allele, whereas less conserved bases would yield lower confidence scores. We benchmarked the performance of PlantCAD2 against three baselines: its predecessor PlantCAD, GPN (a plant specific DNA LM trained on Brassicales genomes), and Evo2, a general-purpose DNA language model pre-trained using a causal language modeling (CLM) objective, also known as next-token prediction. Unlike masked language modeling, which enables access to both upstream and downstream context, CLM imposes a strict left-to-right constraint, rendering Evo2 inherently unidirectional. Therefore, we input the entire sequences without masking for Evo2 and use the likelihood of the model to represent conservation. Notably, Evo2 was trained at a substantially greater scale, with 7 billion parameters and 9.3 trillion nucleotides, which is over 310× more training data than used for PlantCAD2, therefore providing a rigorous benchmark for assessing the efficiency and representational power of our models. We excluded AgroNT from zero-shot evaluation as its non-overlapping k-mer tokenization strategy prevents single-nucleotide resolution tasks, and we previously demonstrated its limited zero-shot capabilities [23].

We assessed this strategy in two independent tasks. First, we performed cross-species alignments of 34 Andropogoneae genomes [23] to the sorghum reference genome, and identified highly conserved and less conserved sites based on alignment coverage and identity (see Methods). PlantCAD2 consistently outperformed PlantCAD in distinguishing highly conserved from less conserved sites in the sorghum genome, with the largest PlantCAD2 achieving the highest AUROC (**Figure 2B; Supplemental Table 2**). Notably, PlantCAD2-M achieved slightly better performance than Evo2 (AUROC 0.708 vs 0.691) despite being ~22-fold smaller (311M vs 7B parameters), while PlantCAD2-L, being ~11-fold smaller (694M parameters), further improved to 0.73. This demonstrates that our PlantCAD2 models can match or exceed Evo2's performance with substantially fewer parameters. Given that PlantCAD2 is pre-trained with a context window of 8192 bp, we also examined the effect of context length on conservation prediction. AUROC scores increased with longer sequence contexts, plateauing at 4096 bp for all PlantCAD2 models (**Figure S2; Supplemental Table 2**). These findings indicate that evolutionary constraint signals benefit from broader sequence context and that larger models with extended receptive fields are better suited to capture these dependencies.

In the second task, we used multiple sequence alignments from coding sequences of 325 Poaceae genomes to calculate phyloP scores and define highly conserved sites (phyloP > 5) and less conserved sites (phyloP < 1.5). While the relationship between selection and phyloP scores can be nuanced [55], restricting phyloP calculation to coding regions helps mitigate alignment noise caused by the very high transposable element turnover rate in plant genomes [56], providing a more reliable benchmark for conservation prediction. Given Evo2's unidirectional nature from its autoregressive architecture, we hypothesized it might struggle with features requiring bidirectional context, particularly translation initiation sites (TIS), where both upstream regulatory motifs in the 5′ UTR and downstream coding sequence context critically influence start codon recognition and conservation [57]. To test this hypothesis, we separately evaluated performance on TIS versus non-TIS positions within coding sequences. For non-TIS sites, PlantCAD2 models showed lower performance compared to Evo2 (**Figure 2C; Supplemental Table 2**), potentially because Evo2's training data included mature mRNA sequences while PlantCAD2 was trained exclusively on genomic DNA, giving Evo2 an advantage in coding sequence conservation tasks. Interestingly, when evaluating TIS conservation, we observed a strong bias in Evo2: its AUROC dropped drastically to 0.534, barely above random. In contrast, PlantCAD2 maintained robust performance (AUROC: 0.632–0.670; **Figure 2D; Supplemental Table 2**). Even the 65M-parameter GPN outperformed the 7B-parameter Evo2 on this task (AUROC of 0.551), further highlighting Evo2's architectural limitations for TIS prediction (**Figure 2D; Supplemental Table 2**). This TIS-specific weakness in Evo2 validates our hypothesis: without access to coding sequence contexts that are more evolutionary constrained, Evo2 cannot properly assess the conservation patterns at translation start sites. Overall, these results demonstrate that PlantCAD2 provides more consistent and unbiased conservation predictions across different genomic contexts.

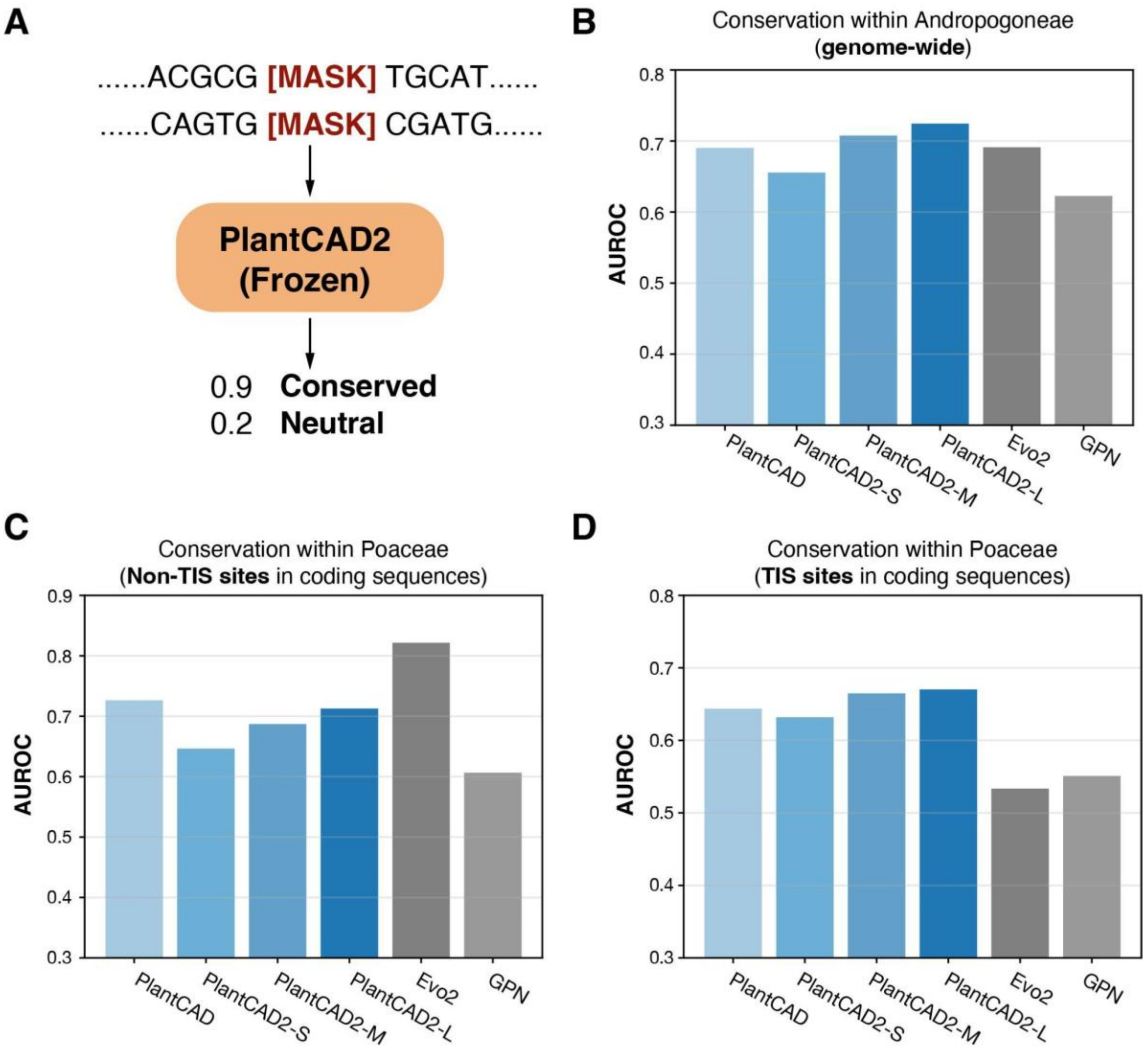


Figure 2. PlantCAD2 accurately predicts evolutionary conservation using zero-shot strategy. (A) Zero-shot conservation prediction approach using masked token probabilities. **(B)** AUROC of conservation of the Sorghum genome within the Andropogoneae tribe. **(C)** AUROC of conservation within Poaceae for non-TIS sites in coding sequences. **(D)** AUROC of conservation within Poaceae for TIS sites in coding sequences.

PlantCAD2 accurately predicts within-species conserved transcriptional and translational junction sites with zero-shot strategy

We next quantified how well PlantCAD2 captured the sequence context that defines key transcriptional and translational junctions. Using a similar zero-shot strategy, we designed four tasks to recapitulate motifs (**Figure 3**). Instead of masking one base pair, for each annotated junction, we replaced the canonical motif with consecutive [MASK] tokens: ATG for the translation initiation site (TIS), TAG/TGA/TAA for the translation termination site (TTS), GT for the splice donor, and AG for the splice acceptor. We then extracted a fixed 8,192-bp window centered on the masked motif and presented the entire masked sequence to the model without fine-tuning. A prediction was considered correct if the model's top-1 reconstruction exactly matched the canonical motif.

As above, we benchmarked the performance of PlantCAD2 against PlantCAD, GPN and Evo2. Given what we observed in **Figure 2D** that Evo2 is limited with its poor TIS conservation prediction, we evaluated it using two configurations: (1) forward sequences (Evo2-fwd), where the model uses upstream context to predict the junction, and (2) reverse-complement sequences (Evo2-rc), where the model uses downstream context (reverse complemented) to predict the junction. For GPN and PlantCAD, which are both limited to context windows of 512 bp, we used 512-bp windows centered on the junctions for evaluation.

When evaluated on both maize (*Zea mays*), which was included in pre-training and tomato (*Solanum lycopersicum*), which was not included in pre-training, PlantCAD2 consistently outperformed PlantCAD1 across all junction types (**Figure 3; Supplemental Table 3**). Notably, even the smallest PlantCAD2 model (88M parameters) outperformed the original PlantCAD (311M parameters), demonstrating that architectural improvements and expanded phylogenetic diversity (65 vs. 16 genomes) provide substantial benefits beyond parameter scaling alone. Accuracy increased with model size, following expected scaling law, with the largest model achieving the highest masked-motif prediction accuracy across both species.

As expected, Evo2 showed strong directional effects due to its unidirectional architecture. For junctions where downstream context is more informative (TIS and splice acceptor), Evo2-rc performed better, as the reverse-complement orientation allows the model to 'see' the downstream coding sequences that provide stronger signals. Conversely, for junctions where upstream context matters more (TTS and splice donor), Evo2-fwd showed superior performance. The sharp contrast in performance between forward sequences and reverse complemented sequences reflects a fundamental limitation of causal language models: their unidirectional nature prevents them from accessing both upstream and downstream signals simultaneously. In contrast, PlantCAD2's bidirectional and reverse-complement equivariant design achieved robust performance regardless of sequence orientation, consistently leveraging both upstream and downstream context for all junction types.

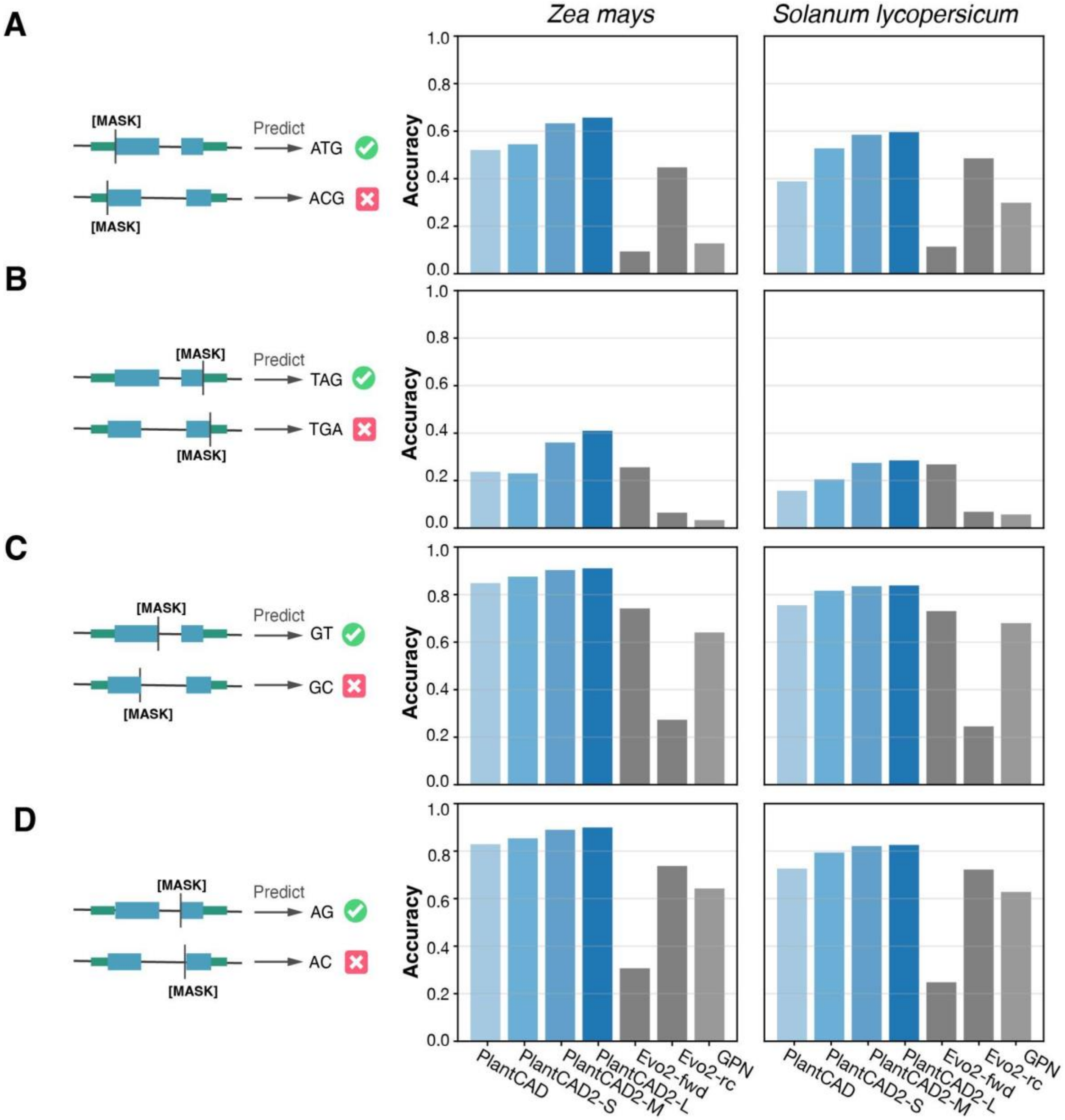


Figure 3. PlantCAD2 accurately predicts transcriptional and translational junction sites using zero-shot masked motif prediction. Left panels show the masking strategy where canonical motifs are replaced with [MASK] tokens and models predict the correct sequence. Right panels show prediction accuracy for each model on maize (left, included in training) and tomato (right, excluded from training). **(A)** Translation initiation sites (ATG masking). **(B)** Translation termination sites (TAG/TGA/TAA masking). **(C)** Splice donor sites (GT masking). **(D)** Splice acceptor sites (AG masking).

While recovering canonical junction motifs demonstrates basic sequence understanding, we next tested whether PlantCAD2 captures deeper evolutionary signals that distinguish core genes (evolutionarily constrained and present across taxa) from non-core genes (rapidly evolving and taxa-specific). We extracted each model's log-likelihood of the canonical motif as a conservation score and evaluated binary classification performance using AUROC (**Figure S3**). For Evo2, we selected the optimal orientation based on junction type (forward for TTS/donor, reverse-complement for TIS/acceptor) as determined above. Remarkably, even though the tomato genome was excluded from PlantCAD2's 65 pre-training genomes, PlantCAD2 consistently outperformed Evo2—which did include tomato during pre-training. This demonstrates strong cross-species generalization: PlantCAD2 learned transferable conservation patterns from other angiosperms that effectively predict functional constraints in unseen species. These results highlight PlantCAD2's ability to capture fundamental evolutionary principles instead of just recognizing simple motif recognition.

PlantCAD2 predicts functional structural variants with zero-shot strategy

In addition to single-nucleotide conservation, we investigated whether PlantCAD2 can generalize to predicting the functional impact of structural variants, such as small deletions, using a zero-shot approach. To do this, we simulated a set of deletions in the Arabidopsis genome and computed the Δlog P score, defined as the log-likelihood ratio between the reference and mutated sequences surrounding the deletion site, averaged across the deletion window (**Figure 4A**). To assess how well Δlog P reflects the functional deletions, we used phyloP scores derived from multiple sequence alignments from 63 genomes [58] to classify deletions as either highly conserved or less conserved based on their average phyloP values.

PlantCAD2's zero-shot Δlog P scores showed strong positive correlation with phyloP-based constraint scores (**Figure 4B**), with the model assigning higher likelihoods to mutations in evolutionarily constrained regions. To quantify this relationship, we binarized deletions into "highly conserved" and "less conserved" categories based on phyloP scores (**Figure 4A**) and evaluated whether Δlog P could discriminate between them. PlantCAD2 achieved robust classification performance (**Figure 4C; Supplemental Table 4**), with even the 88M-parameter PlantCAD2-S outperforming the 7B-parameter Evo2, highlighting the advantage of plant-specific training over general-purpose models again. Additionally, we also observed classification performance is saturated with just 20 bp of flanking sequence on each side (**Figure S4**), indicating that local sequence context sufficiently captures the functional impact of small deletions. This strong performance is particularly impressive given that PlantCAD2 was never explicitly trained on structural variants, suggesting it learned generalizable sequence constraint patterns during pre-training. These results suggest that PlantCAD2's learned representations generalize beyond single-nucleotide changes, capturing broader sequence dependencies relevant to noncoding structural variation. This provides a scalable alternative to traditional alignment-based conservation methods. Notably, this also represents one of the first efforts to use DNA LMs for estimating indel effects in plant genomes, underscoring the potential of foundation models in addressing complex variant interpretation challenges.

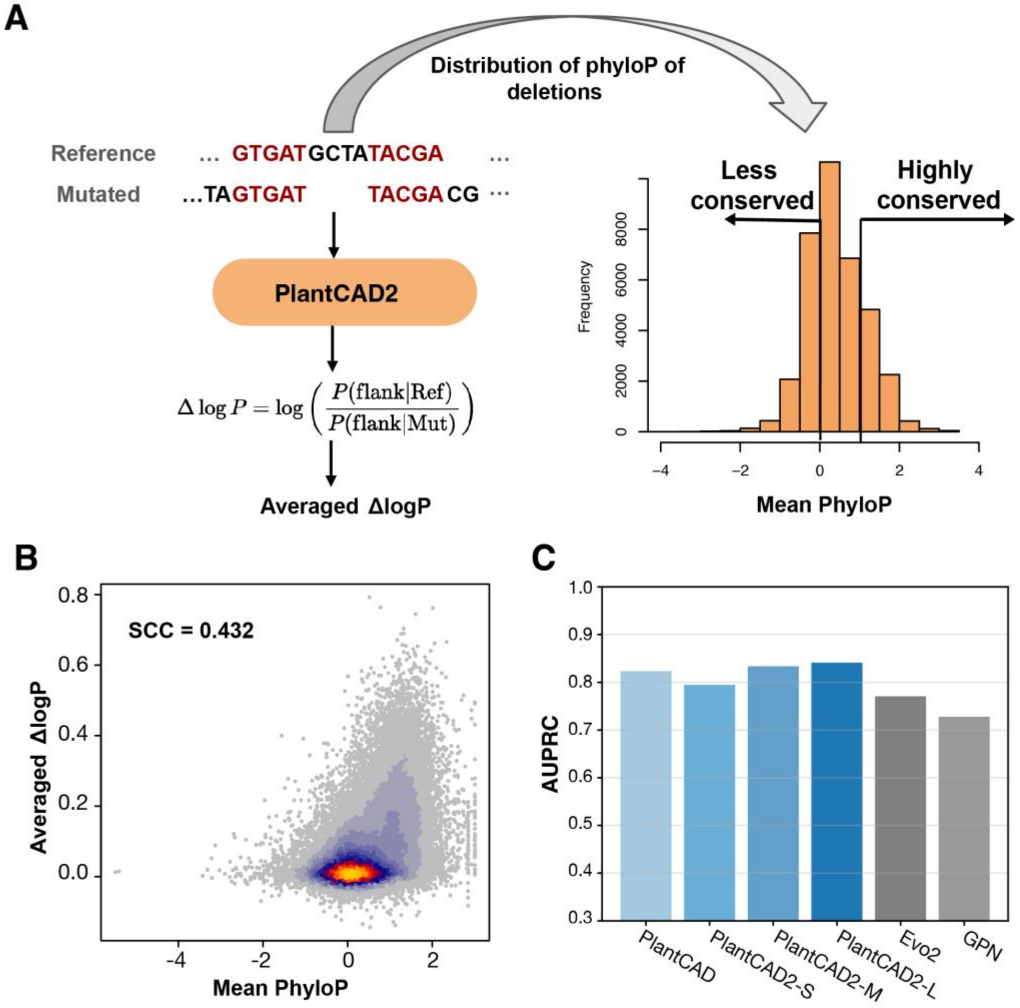


Figure 4. PlantCAD2 predicts functional impact of structural variants using zero-shot strategy. (A) ΔlogP calculation approach for deletion variants and phyloP score distribution for classification. **(B)** Scatter plot showing the positive correlation between PlantCAD2's ΔlogP scores and phyloP-based conservation scores. **(C)** AUROC performance distinguishes highly conserved from less conserved deletions.

Fine-tuning PlantCAD2 accurately predicts cross-species chromatin accessible regions and cell-type-specific accessible regions

We next investigated whether PlantCAD2 learned chromatin states by assessing its performance in predicting genome-wide chromatin accessibility across multiple plant species. We formulated this as a binary classification task, in which the model predicts whether a given genomic region corresponds to an accessible chromatin region, as defined by ATAC-seq (**Methods**). We used recently published ATAC-seq data including 11 diverse plant species, including both dicots and monocots [59]. In this task, positive examples correspond to accessible peaks from ATAC-Seq, while negative examples were from genomic background regions. We used 600-bp genomic windows for all models, as this resolution captures the typical size of ATAC-seq peaks while providing sufficient sequence context for regulatory element prediction. Due to the biological reality of accessible regions comprising only a small fraction of the entire genome, this task is highly imbalanced (**Supplementary Table 5**). For example, less than 1% of regions in the maize genome are labeled as positive.

To effectively leverage the pre-trained foundation model, we used a Low-Rank Adaptation (LoRA) [60] fine-tuning strategy for PlantCAD2, which inserts small trainable rank-decomposition matrices into the feedforward layers while keeping the rest of the model frozen (**Figure 5A**). This approach updates only a small fraction of parameters, enabling efficient task-specific adaptation with minimal risk of overfitting or forgetting the pre-trained knowledge. To assess the contribution of pre-training, we compared this approach to two supervised baselines: (1) a fully supervised version of PlantCAD2-S from scratch, where all model parameters were randomly initialized and updated during training; and (2) a commonly used CNN+LSTM [33] architecture trained from scratch. We also benchmarked against AgroNT [22], another plant-specific DNA LM. We excluded GPN and PlantCAD due to their limited 512-bp context window, and Evo2 due to both its consistently lower zero-shot performance compared to PlantCAD2 and the computational infeasibility of fine-tuning a 7B-parameter model. AgroNT, with its transformer architecture and intermediate size (1 billion parameters), provides a more practical and fair comparison point that supports efficient LoRA adaptation. All models were trained using Arabidopsis and validated on hold-out chromosomes within Arabidopsis as well as on 10 additional test species spanning a broad phylogenetic range.

Given the strong class imbalance of this task, we measured model performance using the area under the precision–recall curve (AUPRC), which is more informative than AUROC in imbalanced classification settings. LoRA fine-tuned PlantCAD2 consistently achieved the best performance in both within-species evaluation and cross-species generalization, outperforming supervised baselines and AgroNT across all test species (**Figure 5B; Supplementary Table 5**). And we observed a strong negative correlation between genome size of AUPRC, which reflects the increasing difficulty of distinguishing sparse regulatory elements in large intergenic regions (**Figure S5**). Comparing fine-tuned DNA LMs to supervised models (whether using CNN+LSTM or Supervised PlantCAD2-S), it's obvious fine-tuned DNA LMs consistently outperformed supervised models trained from scratch, indicating that pre-training enables better learning of chromatin states, particularly when transferring knowledge across species (**Figure 5B**). Specifically, the supervised models (whether using CNN+LSTM or Supervised PlantCAD2-S) trained on Arabidopsis generalized reasonably well to closely related dicots, but their performance declined substantially when applied to evolutionarily distant monocots such as maize and barley. In contrast, fine-tuned PlantCAD2 retained strong predictive accuracy across both dicots and monocots, demonstrating its ability to capture regulatory features conserved across deep evolutionary divergence. These results underscore the power of combining self-supervised pre-training with parameter-efficient fine-tuning for plant regulatory genomics.

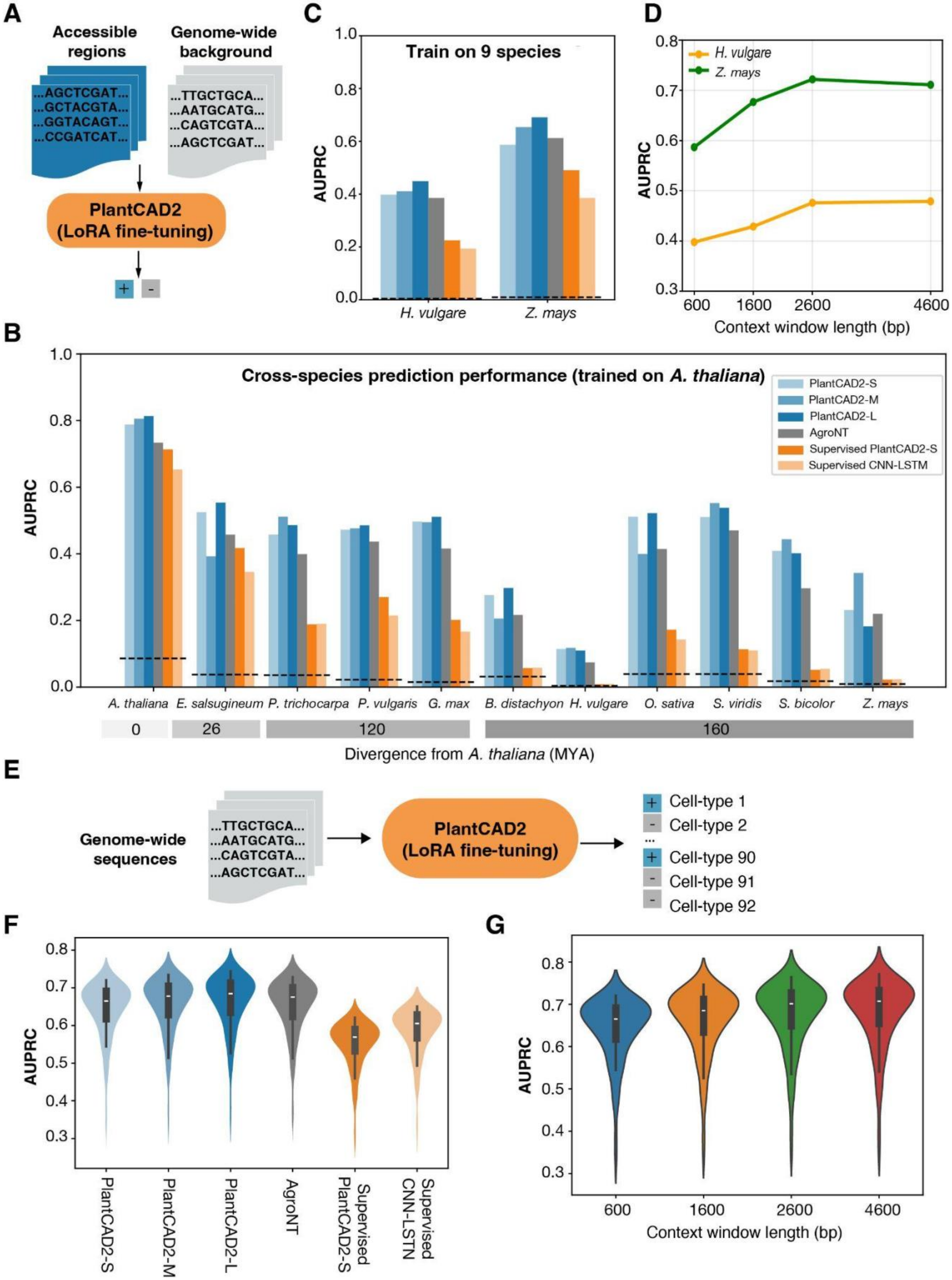

A
Accessible regions
...AGCTCGAT...
...GCTACGTA...
...GGTACAGT...
...CCGATCAT...
Genome-wide background
...TTGCTGCA...
...AATGCATG...
...CAGTCGTA...
...AGCTCGAT...
PlantCAD2 (LoRA fine-tuning)
+
-
C
Train on 9 species
AUPRC
0.0
0.2
0.4
0.6
0.8
1.0
H. vulgare
Z. mays
D
H. vulgare
Z. mays
AUPRC
0.3
0.4
0.5
0.6
0.7
0.8
600
1600
2600
4600
Context window length (bp)
B
Cross-species prediction performance (trained on A. thaliana)
PlantCAD2-S
PlantCAD2-M
PlantCAD2-L
AgroNT
Supervised PlantCAD2-S
Supervised CNN-LSTM
AUPRC
0.0
0.2
0.4
0.6
0.8
1.0
A. thaliana
E. salsugineum
P. trichocarpa
P. vulgaris
G. max
B. distachyon
H. vulgare
O. sativa
S. viridis
S. bicolor
Z. mays
0
26
120
160
Divergence from A. thaliana (MYA)
E
Genome-wide sequences
...TTGCTGCA...
...AATGCATG...
...CAGTCGTA...
...AGCTCGAT...
PlantCAD2 (LoRA fine-tuning)
+ Cell-type 1
- Cell-type 2
...
+ Cell-type 90
- Cell-type 91
- Cell-type 92
F
AUPRC
0.3
0.4
0.5
0.6
0.7
0.8
PlantCAD2-S
PlantCAD2-M
PlantCAD2-L
AgroNT
Supervised PlantCAD2-S
Supervised CNN-LSTN
G
AUPRC
0.3
0.4
0.5
0.6
0.7
0.8
600
1600
2600
4600
Context window length (bp)

Figure 5. PlantCAD2 predicts chromatin accessibility across species and cell types. (A) LoRA fine-tuning approach for binary accessibility prediction using ATAC-seq peaks versus genomic background. **(B)** Cross-species AUPRC performance when trained on Arabidopsis, showing superior generalization of PlantCAD2 models compared to supervised baselines across evolutionary distances. **(C)** Multi-species training performance on held-out barley and maize. **(D)** Effect of context window length on accessibility prediction accuracy for PlantCAD2-S. **(E)** Multi-label classification approach for cell-type-specific accessibility prediction. **(F)** Performance comparison across models for 92 cell types . **(G)** Context window effects on cell-type-specific prediction accuracy for PlantCAD2-S.

While the foundation model fine-tuned on Arabidopsis already demonstrated clear advantages in cross-species prediction, we further fine-tuned a multi-species version of PlantCAD2 using accessible chromatin regions from multiple plant genomes to enhance its robustness across diverse lineages. With maize and barley held out as test species, this multi-species model achieved impressive AUPRC scores of 0.691 for maize and 0.449 for barley (**Figure 5C; Supplemental Table 6**). To investigate whether extended sequence context could further improve prediction accuracy, we maintained the same 600-bp labels but varied the input window size by including different amounts of flanking sequence. For computational efficiency, we conducted this analysis using PlantCAD2-S. Performance consistently improved with longer context windows, with AUPRC increasing from 0.587 to 0.711 for maize and from 0.398 to 0.479 for barley when extending from 600 bp to 4,600 bp (**Figure 5D**). This substantial improvement suggests that distal regulatory elements and broader chromatin context beyond the immediate peak boundaries contribute to accessibility prediction, highlighting the advantage of PlantCAD2's long-context architecture. These fine-tuned models serve as a robust predictor of chromatin accessibility across flowering plants and are publicly available to the community as a ready-to-use resource for regulatory annotation in non-model species.

To further assess whether PlantCAD2 can resolve cell-type–specific regulatory landscapes, we tested its ability to predict accessible chromatin regions identified through single-cell ATAC-seq (scATAC-seq) in maize [61]. In contrast to the binary classification task used for genome-wide accessibility, this task was framed as a multi-label classification problem, where each genomic window could be accessible in one or more cell types (**Figure 5E**). We curated high-confidence cell-type–specific peaks across major maize tissues from published scATAC-seq datasets (**Methods**), using them as labels for multi-label fine-tuning and evaluation.

We applied the same LoRA fine-tuning strategy used in prior experiments, adapting PlantCAD2 to predict cell-type–specific accessibility using only a small number of trainable parameters. As in previous sections, we compared performance against two supervised baselines: a CNN+LSTM model trained from scratch and a fully supervised version of PlantCAD2. All models were trained on all maize cell types and evaluated on held-out chromosomes, with performance measured using micro-averaged precision-recall curves across cell types. Despite the complexity and subtlety of cell-type–specific regulatory signatures, LoRA fine-tuned PlantCAD2 can still achieve very high accuracy and outperformed other baselines (**Figure 5F; Supplemental Table 7**). Similar to our genome-wide accessibility results, extending the input context window beyond the core 600-bp peak region further improved cell-type specificity, with AUPRC increasing from 0.665 to 0.707 when using 4,600-bp windows (**Figure 5G**). This suggests that cell-type–specific regulatory programs are influenced by broader chromatin context and distal regulatory interactions. The model captured both shared and lineage-specific accessibility patterns, demonstrating that pre-trained DNA representations can be effectively adapted to fine-grained regulatory annotations. These results suggest that PlantCAD2 is not only effective at modeling general chromatin accessibility across species but is also capable of distinguishing cell-type–specific regulatory programs within a single genome.

Fine-tuning PlantCAD2 predict cross-species gene expression and protein abundance

To evaluate PlantCAD2's ability to capture gene regulatory signals, we fine-tuned PlantCAD2 models using LoRA for two complementary tasks: leaf gene expression and leaf translation (**Figure 6A, 6D**). Each task involved both classification (on/off status) and regression (absolute expression/translation level) objectives. For gene expression, we used promoter and terminator sequences (1024 bp each) as input; for translation, we used 500 bp upstream sequences. Following the same strategy as the accessible chromatin prediction task, we fine-tuned PlantCAD2 with LoRA and compared its performance with supervised PlantCAD2-S, CNN+LSTM, and AgroNT.

For cross-species gene expression modeling, we fine-tuned PlantCAD2 on a diverse panel of 15 plant species and evaluated predictions in the maize Nested Association Mapping (NAM) population [62]. Across both binary leaf expression prediction and absolute expression level prediction, PlantCAD2 consistently outperformed established baselines such as AgroNT and supervised CNN+LSTM. In the maize NAM population, even the smallest PlantCAD2 (PlantCAD2-S, 88M parameters) outperformed the performance of the much larger AgroNT model

(1B parameters) (**Figure 6B-C; Supplemental Table 8**), demonstrating the efficiency of our foundation model framework. The largest model, PlantCAD2-L, achieved the best AUROC for binary leaf expression and the highest Spearman correlation for absolute expression prediction.

Given that regulatory information also lies outside the proximal promoter [38], we evaluated the effect of varying input window sizes on gene expression prediction. Increasing the window from 1 kb to 4 kb both upstream of the transcription start site and downstream of the transcription stop site resulted in measurable improvements, raising the AUROC from 0.8221 to 0.8455 for binary leaf expression task and Spearman correlation from 0.6296 to 0.6455 for absolute expression prediction task on the NAM test set (**Figure S6**). These improvements highlight the role of distal enhancers and long-range motifs in shaping expression. However, previous studies in both humans and plants have shown that current deep learning models lack the resolution to capture allele-specific effects [62–64]. We therefore tested whether fine-tuning a foundation model could mitigate this limitation by evaluating per-orthogroup correlations within the NAM population, a comparison sensitive to allelic differences. Consistent with prior findings [62–64], only marginal improvements were observed for leaf absolute expression prediction, with the median Spearman correlation increasing from 0.112 (supervised CNN+LSTM) to 0.140 (PlantCAD2-S) (**Figure S7**). These results suggest that achieving allele-specific resolution will likely require specialized training strategies, such as explicitly modeling cis-regulatory variants [65,66].

Translation prediction was based on ribosome profiling (ribo-seq), a sequencing-based approach that estimates translation activity by mapping ribosome-protected mRNA fragments. Because ribo-seq data are scarce in plants, we restricted training to Arabidopsis and tested both within-species performance and cross-species transfer to maize. Interestingly, although supervised PlantCAD2-S (88M) is much larger than the CNN-LSTM (~1.7M), the latter performed better, suggesting that large supervised models are prone to overfitting when trained on limited data. By contrast, fine-tuned PlantCAD2 with parameter-efficient LoRA maintained robust performance without signs of overfitting. However, cross-species regression of absolute translation levels was less effective (**Figure 6F**), suggesting that the model may have captured noise inherent in ribo-seq–based quantitative estimates. Encouraged by the strong transfer observed for binary classification of leaf translation, we next tested whether gene expression could similarly be transferred from Arabidopsis to maize. Using a separate Arabidopsis gene expression dataset [67] with 1,024 bp upstream and downstream sequences, we found that direct transfer performed poorly (AUROC: 0.786 in Arabidopsis vs. 0.631 in maize) compared to protein abundance prediction (0.790 in Arabidopsis vs. 0.692 in maize) (**Figure S8**). This contrast may highlight a fundamental difference between regulatory layers: translational control appears to be more evolutionarily conserved than transcriptional regulation. Consistent with evidence that protein abundance is under stronger selective constraint than transcript levels [63], these results explain why translation prediction transfers effectively across species, whereas accurate gene expression prediction requires phylogenetically diverse training data.

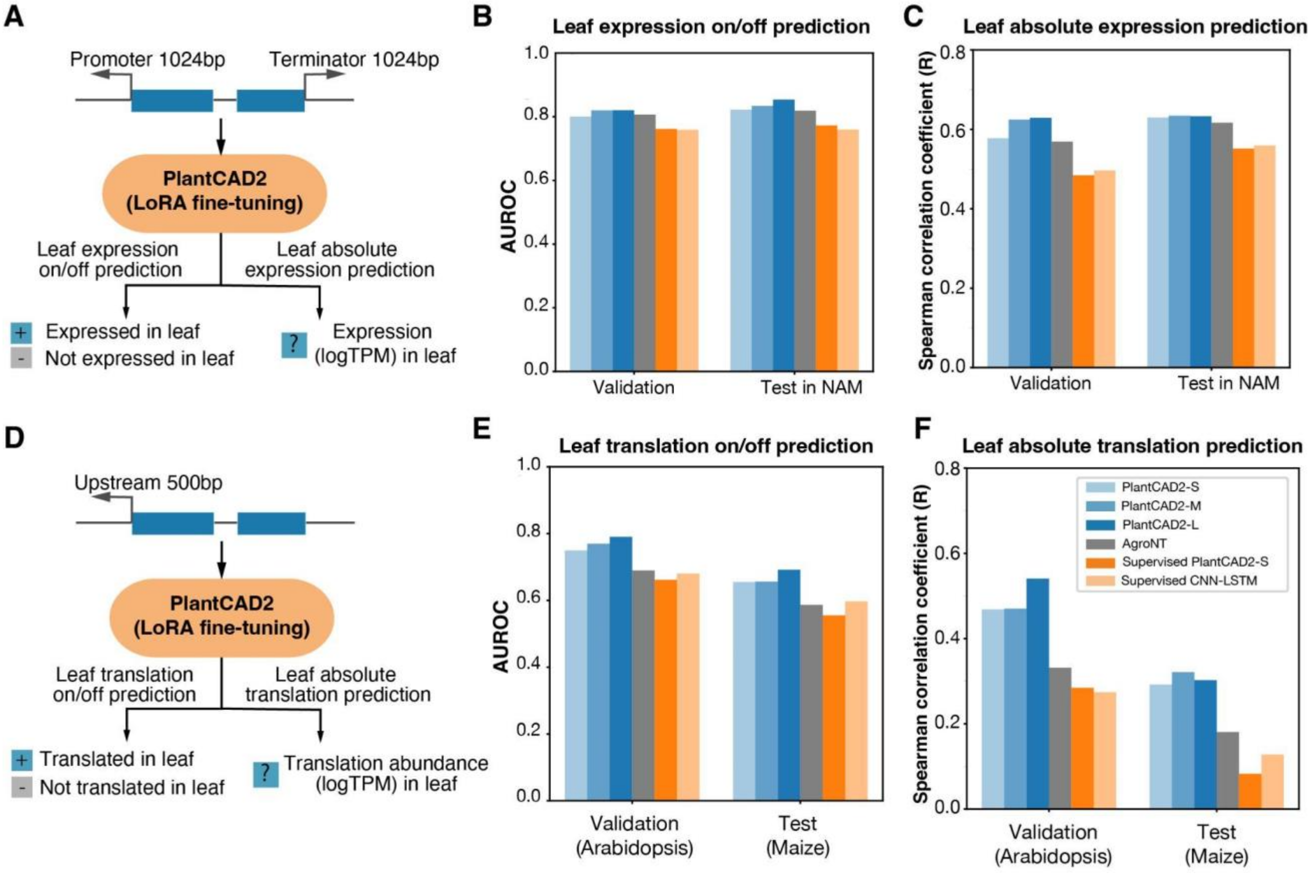


Figure 6. PlantCAD2 predicts gene expression and translation across species. (A) Gene expression prediction pipeline using promoter and terminator sequences (1024 bp each) for binary classification and regression tasks. **(B-C)** Cross-species gene expression performance on maize NAM population for binary on/off prediction (B) and absolute expression levels (C). **(D)** Translation prediction pipeline using 500 bp upstream sequences. **(E-F)** Translation prediction performance trained on Arabidopsis and tested cross-species on maize for binary on/off prediction (E) and absolute translation levels (F).

Discussions

In this work, we present PlantCAD2, a long-context window DNA language model that substantially advances the sequence-to-function modeling in plant genomics. Building on the foundation laid by PlantCAD, PlantCAD2 features a model architecture that is three times larger, a 16-fold longer context window (8,192bp vs. 512bp), and a pre-training dataset that is evenly distributed across angiosperm orders to better capture phylogenetic diversity. Through comprehensive zero-shot and fine-tuned evaluations, we demonstrate that PlantCAD2 not only exhibits strong cross-species generalization, but also achieves superior performance across a wide range of sequence-to-function tasks, including evolutionary conservation prediction, functional important junction sites prediction involved in both transcription and translation, variant (including indels) effect estimation, cis-regulatory activity, gene expression, and protein translation.

PlantCAD2 represents a significant step toward a foundational model for plant genomics. Rather than building task-specific models for each application, it enables unified modeling of sequence-to-function relationships that can be efficiently adapted across cell types, tissues, and species. This paradigm shift opens new opportunities to integrate deep learning into practical breeding applications. For example, PlantCAD2 could assist in prioritizing causal variants in GWAS studies, interpreting SVs in noncoding regions, or guiding sequence design for synthetic promoters with desired expression patterns. Its ability to transfer knowledge across evolutionarily distant species further enhances its utility for crop improvement, particularly in non-model organisms where high-quality training data are limited but genomic sequences are available.

Despite its advances, PlantCAD2 also presents new challenges. First, its large model size may limit deployment in GPU-constrained environments. Developing distilled or compressed versions that retain high performance while reducing compute demands is an important next step. Second, accessibility to wet-lab biologists or breeders remains limited by technical barriers. Building intuitive interfaces, pretrained APIs, and end-to-end pipelines will be crucial to broaden the use of PlantCAD2 in broader plant science communities. Third, while the 8,192-bp context window allows PlantCAD2 to model distal regulatory elements, further extending this capability would be valuable for capturing long-range interactions such as enhancer–promoter loops. For example, in maize, the teosinte branched 1 (*tb1*) enhancer [68,69] and Vegetative to generative transition 1 (*Vgt1*) [70] are located approximately 70kb and 60kb upstream of their target genes, respectively.

However, capturing such interactions will likely require novel tokenization or compression strategies that can represent long, repetitive sequences without sacrificing resolution.

Looking forward, future directions include combining PlantCAD2 with multi-modal data such as DNA methylation and chromatin states will provide trans-factors to the genome. In addition, diffusion-based sequence generation models could also be promising coupled with synthetic biology. Ultimately, we envision PlantCAD2 and its successors as key building blocks for a sequence-to-function foundation model capable of enabling predictive genomics and rational genome design in diverse plant species.

Methods

Preparing pre-training genomes

A total of 65 genomes were selected for pre-training from the Phytozome database. To ensure taxonomic relevance and minimize redundancy, we applied a series of manual filtering steps. First, non-angiosperm species were excluded. For each remaining species, we retained only the most recent genome assembly version. In cases where two haplotypes were available for a species, we selected the haplotype with the higher N50 value; if N50 values were comparable, we retained the assembly with fewer scaffolds to prioritize less fragmented genomes. Taxonomic information, including order, family, and genus, was appended to each genome to facilitate downstream analyses, and their relationships were visualized using a published time-callibrated phylogeny [71]. For each selected genome, we extracted genomic sequences centered on each annotated gene, extending 5 kilobases (kb) upstream and 5 kb downstream from the gene center. These ±5 kb gene-centered regions were then segmented into overlapping windows of 8,192 bp with a step size of 4,096 bp, ensuring comprehensive coverage of regulatory and genic features while maintaining continuity across sequence boundaries. These windows served as input sequences for model pre-training. We then employed a de-novo pipeline to annotate highly repetitive sequences [72] (https://github.com/baoxingsong/dCNS), then repetitive sequences are down-weighted during pre-training as demonstrated important in previous studies [20]

PlantCAD2 model architecture and pre-training

PlantCAD2 builds upon the Caduceus architecture [24] used in PlantCAD [23], retaining its key design principles while incorporating architectural improvements. Like PlantCAD1, PlantCAD2 maintains three core features: (1) bidirectional sequence processing, where sequences are processed both forward and reverse with outputs summed together; (2) reverse-complement (RC) equivariance, ensuring the model commutes with RC operations; and (3) parameter-efficient bidirectional implementation through shared linear projections between forward and reverse passes.

The primary architectural improvement in PlantCAD2 is the replacement of Mamba1 blocks [37] with Mamba2 blocks [45]. Mamba2 introduces a structured state space duality that recasts the selective state space computation into an equivalent convolutional form using structured matrices, improving parallelism and hardware efficiency. This dual representation enables significantly faster training (up to 2–4× in some scenarios) while retaining the input-dependent selection mechanism that allows the model to dynamically modulate state updates based on sequence content. These advances allow PlantCAD2 to efficiently handle 8,192 bp sequences with linear computational complexity.

For the pre-training of PlantCAD2, each model was trained for 240,000 steps using a Decoupled AdamW optimizer [73] with the global batch size of 2,048. The learning rate is 2E-4 with a cosine decay scheduler, and 6% of the training duration was dedicated to warm up. The learning rate decayed to 4E-6 by the end of training. The default BERT [32] masking recipe was used with a masking probability of 15%. For each masked token: i) there is an 80% probability it will be replaced by a special token ([MASK]), ii) a 10% probability it will be replaced by a random token, and iii) a 10% probability it will remain unchanged. Unless otherwise specified, all models were trained using a sequence length of 8192 base pairs. A weight decay of 1E-5 was applied throughout the training process.

Evolutionary constraint prediction using the zero-shot strategy

To evaluate the extent to which PlantCAD2 captures evolutionary conservation signals, a zero-shot strategy was applied to predict constrained genomic regions. Two independent tasks were used. The first task focused on Sorghum bicolor, using conservation estimates from the Andropogoneae tribe, a large clade of approximately 1,200 grass species that descended from a common ancestor approximately 18 million years ago [74]. To generate conservation labels, 34 high-quality genomes [75] were aligned to the Sorghum bicolor reference genome using AnchorWave [76]. Per-base conservation was quantified using alignment identity scores across all species. Sites with high-quality coverage (i.e., aligned in at least 34 out of 35 species) were retained for analysis. Among these, positions with an identity score ≥34 were labeled as conserved, while those with identity scores <15 were labeled as neutral. Sites with intermediate identity scores or insufficient coverage were excluded from evaluation to ensure high-confidence labels [23].

The second task utilized conservation scores derived from multiple sequence alignments (MSAs) of orthologous coding sequences from 325 Poaceae genomes, a high-quality subset of the recently published set of 727 genomes. Using *Pharus latifolius* as an outgroup species, gap columns were removed prior to conservation estimates. Using PHAST [77], PhyloP scores were calculated per site based on a neutral model derived from fourfold degenerate sites and "LRT" methods with the mode "CONACC". Sites with phyloP scores above 5 were classified as conserved, while those below 1.5 were considered neutral. Sites with intermediate scores were excluded to maintain label clarity. For TIS sites, we retained all 36,668 (26,653 conserved vs 10,015 less conserved) sites given their biological importance. For non-TIS sites, we downsampled to 183,687 sites (103,369 conserved versus 80,318 neutral) for computational efficiency while maintaining the conserved-to-neutral ratio.

For both tasks, the evaluated site was centered (4096th) within a 8,192 bp input sequence, and the reference base at that position was masked. The model's predicted likelihood of the reference allele was extracted and used as the zero-shot conservation score. Higher likelihoods were hypothesized to reflect stronger conservation. Model performance was assessed using AUROC, comparing scores between conserved and neutral sites.

Core and non-core gene classification using the zero-shot strategy

To assess the ability of PlantCAD2 to distinguish between core and non-core genes in a population, a zero-shot strategy was applied to classify within species conservation of genes in maize and tomato. For maize, the pangene table derived from 26 Nested Association Mapping genomes [78] was used. Core genes were defined as those present in all 26 NAM genomes, whereas non-core genes included both dispensable genes (present in 2-23 genomes) and private genes (present in only one genome). For genes with multiple transcripts, the canonical transcript specified in the annotation was used. For tomato, the pan-genome dataset assembled from 586 high-quality genomes [79] was used. Genes present in all 586 accessions were defined as core genes, and non-core genes consist of dispensable (present in 6-580 accessions) and private (present in less than 5 accessions). The longest transcript was selected to represent each gene across all analyses.

To quantify the model's prediction at each functional junction, a masked motif accuracy score was defined. For example, to evaluate translation initiation sites, the canonical ATG start codon was masked, and the model's predicted likelihoods for the three masked nucleotides were extracted and averaged. A similar approach was applied to other junction types, including translation termination sites (TAA, TAG, TGA), splice donor sites (GT), and splice acceptor sites (AG), by masking the corresponding motifs and calculating average token likelihoods.

To evaluate performance, core genes were treated as positive and non-core genes as negative, and AUROC was calculated based on the masked motif accuracy scores for each gene.

Accessible chromatin region prediction

To evaluate the capability of PlantCAD2 to capture regulatory sequence features, fine-tuning experiments were performed using Low-Rank Adaptation (LoRA) [60] on two accessible chromatin prediction tasks: (1) cross-species accessible regions (ACRs) prediction, and (2) cell-type-specific ACR prediction.

For the cross-species task, ATAC-seq peak regions from 12 plant species [59] were downloaded from NCBI (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE128434). We followed the data processing pipeline as described by Wrightsman et al [80]. For each species, peak regions were processed by extracting the midpoint of each peak and symmetrically extending it by half the target input window size (300bp, 600bp and 1,000bp) in both directions to generate positive observations. To reflect the real-world scenario in which most of the genome is inaccessible, the rest of the genome was used as negative examples, ensuring no overlap with known peaks.

For the cell-type-specific task, we used the single-cell ATAC-Seq [61] and used a similar preprocessing pipeline, but each genomic region could be associated with accessibility across 92 cell types. As such, the task was framed as a multi-label classification problem, where each region was assigned a binary accessibility label for each of the 92 cell types based on its overlap with experimentally identified peaks.

Gene expression prediction in leaf

To evaluate the models' ability to predict gene expression, we designed two tasks: (1) leaf absolute expression and (2) leaf on/off expression classification. The training dataset was derived from 15 Andropogoneae species [62]. For validation, we held out two species closest to Zea mays—

Tripsacum zopilotense and Zea diploperennis—both members of the Tripsacinae subtribe, which diverged from maize approximately 0.6 to 4 million years ago. This setup enabled evaluation of the models' cross-species generalization to closely related taxa. For the leaf absolute expression task, the $\log_{10}$TPM values were used as regression targets during fine-tuning. For the on/off expression task, genes with TPM > 1 were labeled as expressed (positive), and those with TPM ≤ 1 were considered non-expressed (negative). This setup enabled evaluation of the models' ability to generalize expression predictions across closely related species within the clade.

Leaf protein abundance prediction task

To evaluate the models' ability to predict protein abundance, we designed two tasks analogous to the gene expression analysis: absolute abundance and on/off classification. Ribo-Seq data were obtained from Arabidopsis [81] and Zea mays [82]. Raw reads were downloaded from NCBI, and Trimmomatic [83] was used to trim adapters and filter low-quality reads. Cleaned reads were first aligned to rRNA reference sequences using Bowtie [84] to remove contaminating rRNA. The remaining reads were then mapped to the reference genomes of Arabidopsis and maize using STAR [85]. Gene-level translation abundance was quantified using StringTie [86] based on uniquely mapped reads. We then designed two tasks analogous to the expression prediction setup: (1) absolute protein abundance, where the log10-transformed Ribo-Seq expression values were used as regression targets, and (2) on/off classification, where genes with TPM > 1 were considered expressed (positive) and those with TPM ≤ 1 were labeled as non-expressed (negative).

Fine-tuning PlantCAD2

To adapt the pre-trained PlantCAD2 model to downstream tasks, we employed Low-Rank Adaptation (LoRA) [60], a parameter-efficient fine-tuning strategy that inserts trainable low-rank matrices into the attention layers of the transformer. This approach enables effective adaptation while keeping the vast majority of model parameters frozen. Fine-tuning was performed using the PEFT library [87] with LoRA rank = 8, α = 32, and dropout = 0.1, targeting the "x_proj", "in_proj", and "out_proj" modules. Models were trained using the Hugging Face Trainer with a learning rate of 1e−4, a global batch size of 128, and one training epoch. BF16 precision and linear learning rate scheduling with 50 warm-up steps were used. Over 98% of the model parameters remained frozen, enabling efficient and scalable fine-tuning across tasks. All tasks were fine-tuned for a single epoch without hyperparameter tuning to ensure stability and consistency across experiments. Fine-tuning objectives for all models were a binary cross entropy loss for classification tasks and a mean squared error loss for regression tasks.

Fine-tuning AgroNT

To directly compare the performance of fine-tuned PlantCAD2 with AgroNT [22], we applied the same parameter-efficient fine-tuning strategy using LoRA. All LoRA hyperparameters were kept consistent with those used for PlantCAD2, including rank = 8, α = 32, and dropout = 0.1. For AgroNT, LoRA adapters were inserted into the "query" and "key" projection layers of the transformer [46], reflecting its architecture. This setup ensured a fair comparison between models under matched fine-tuning conditions.

Supervised CNN + LSTM baseline

To benchmark against traditional supervised models, we implemented a CNN+LSTM architecture based on DanQ [34], a widely used hybrid model for DNA sequence classification. For each task, the model was trained from scratch using one-hot encoded sequences. We used the Adam optimizer with a learning rate of 0.01, a batch size of 2,048, and trained for up to 200 epochs, with early stopping after 20 epochs without validation improvement.

Supervised PlantCAD2 baseline

To assess the contribution of pretraining, we trained a small PlantCAD2 model from scratch for each downstream task. This supervised baseline used the same architecture and hyperparameters as the fine-tuned version but was initialized without pretrained weights—by loading only the Hugging Face model configuration.

Zero-shot evaluation of PlantCAD2, PlantCAD and GPN models

All three models were pre-trained with masked language modeling. For PlantCAD, we used the largest available model "kuleshov-group/PlantCaduceus_l32" for evaluation. For GPN, we used "songlab/gpn-brassicales". Due to the 512 bp context window limitation of both PlantCAD and GPN, we cropped input sequences to 512 bp centered on the target position. All other evaluation configurations remained identical to those used for PlantCAD2.

Zero-shot evaluation of Evo2 model

All zero-shot tasks were also benchmarked using the Evo2 [19] model ("evo2_7b") for comparison. Since Evo2 is autoregressive (predicting the next token rather than masked tokens), masked token accuracy could not be directly computed. Therefore, for the evolutionary constraint task, we fed the full input sequence into the model and extracted the likelihood of the reference allele at the target site as the conservation score. To ensure a fair comparison, we used an 8,192 bp context window for Evo2, matching the input length used for PlantCAD2 evaluations. The same approach was applied for benchmarking structural variants.

For the masked motif accuracy task, we evaluated Evo2 using two configurations to compensate for its unidirectional architecture: (1) forward sequences (Evo2-fwd), where the model uses upstream context to predict the junction—for example, for TIS prediction, we used the 4,094 bp upstream of the TIS as a prompt for Evo2 to generate the next three tokens; and (2) reverse-complement sequences (Evo2-rc), where the model uses downstream context (reverse complemented) to predict the junction in the opposite direction.

Code availability

All pre-trained models, datasets, and benchmark tasks are available at https://huggingface.co/collections/kuleshov-group/plantcad2-67e437e241a382671371a572. Fine-tuning pipelines and code are available at https://github.com/kuleshov-group/PlantCaduceus.

Acknowledgements

This work is funded by a cooperative agreement from US Department of Agriculture Agricultural Research Service to Cornell University (MCR), NSF grant (#2240888), NSF CAREER grant (#2145577), NIH Maximizing Investigators' Research Award grant (#1R35GM151243-01) and NIH Grant (5R35GM151348). We thank Travis Wrightsman (Cornell University) for sharing gene expression tasks, Michelle C. Stitzer (Cornell University) for plotting the tree, Hai Wang (China Agricultural University) for sharing max expression data of 17 species, Sara Miller (Cornell University) for helpful comments, and all members of Cornell's Institute of

Genomic Diversity (MCR & ESB) for helpful discussions. We would also like to thank the SCINet project, Texas Advanced Computing Center at The University of Texas at Austin, and MosaicML for providing compute resources for pretraining and fine-tuning experiments.

Author contributions

J.Z., A.G., V.K., and E.S.B. designed research; J.Z., A.G., S.-K.H., Z.-Y.L., S.-P.C., E.M., E.C., B.C., A.B., M.C.R., M.P., V.K., and E.S.B. performed research and analyses; J.Z., E.M., S.-K.H., M.P., and E.S.B. wrote the manuscript with all other authors' suggestions and comments.

Competing interests

The authors declare no competing interests.

Supplemental Information

Supplemental Table 1. Pretraining species and masked language modeling performance across 65 angiosperm genomes

Supplemental Table 2. Cross-species evolutionary conservation prediction performance

Supplemental Table 3. Masked motif prediction accuracy for transcriptional and translational junction sites

Supplemental Table 4. Zero-shot structural variant impact prediction performance

Supplemental Table 5. Cross-species chromatin accessibility prediction trained on Arabidopsis

Supplemental Table 6. Multi-species chromatin accessibility prediction performance

Supplemental Table 7. Cell-type-specific chromatin accessibility prediction in maize

Supplemental Table 8. Gene expression prediction performance across species

Supplemental Table 9. Translation prediction performance across species

References

1. Vancaester, E., and Blaxter, M. (2023). Phylogenomic analysis of Wolbachia genomes from the Darwin Tree of Life biodiversity genomics project. PLoS Biol. *21*, e3001972.

2. Lewin, H.A., Robinson, G.E., Kress, W.J., Baker, W.J., Coddington, J., Crandall, K.A., Durbin, R., Edwards, S.V., Forest, F., Gilbert, M.T.P., et al. (2018). Earth BioGenome Project: Sequencing life for the future of life. Proc. Natl. Acad. Sci. U. S. A. *115*, 4325–4333.

3. Rhie, A., McCarthy, S.A., Fedrigo, O., Damas, J., Formenti, G., Koren, S., Uliano-Silva, M., Chow, W., Fungtammasan, A., Kim, J., et al. (2021). Towards complete and error-free genome assemblies of all vertebrate species. Nature *592*, 737–746.

4. Cheng, S., Melkonian, M., Smith, S.A., Brockington, S., Archibald, J.M., Delaux, P.-M., Li, F.-W., Melkonian, B., Mavrodiev, E.V., Sun, W., et al. (2018). 10KP: A phylodiverse genome sequencing plan. Gigascience *7*, 1–9.

5. Sun, Y., Shang, L., Zhu, Q.-H., Fan, L., and Guo, L. (2022). Twenty years of plant genome sequencing: achievements and challenges. Trends Plant Sci. *27*, 391–401.

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

7. Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., et al. (2021). On the opportunities and risks of foundation models. arXiv [cs.LG].

8. Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C.L., Ma, J., et al. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proc. Natl. Acad. Sci. U. S. A. *118*, e2016239118.

9. Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., et al. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. Science *379*, 1123–1130.

10. Hayes, T., Rao, R., Akin, H., Sofroniew, N.J., Oktay, D., Lin, Z., Verkuil, R., Tran, V.Q., Deaton, J., Wiggert, M., et al. (2025). Simulating 500 million years of evolution with a language model. Science *387*, eads0018.

11. Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., et al. (2022). ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. IEEE Trans. Pattern Anal. Mach. Intell. *44*, 7112–7127.

12. Brandes, N., Ofer, D., Peleg, Y., Rappoport, N., and Linial, M. (2022). ProteinBERT: a universal deep-learning model of protein sequence and function. Bioinformatics *38*, 2102–2110.

13. Kulmanov, M., Guzmán-Vega, F.J., Duek Roggli, P., Lane, L., Arold, S.T., and Hoehndorf, R. (2024). Protein function prediction as approximate semantic entailment. Nat. Mach. Intell.

6, 220–228.

14. Chowdhury, R., Bouatta, N., Biswas, S., Floristean, C., Kharkar, A., Roy, K., Rochereau, C., Ahdritz, G., Zhang, J., Church, G.M., et al. (2022). Single-sequence protein structure prediction using a language model and deep learning. Nat. Biotechnol. *40*, 1617–1623.

15. Brandes, N., Goldman, G., Wang, C.H., Ye, C.J., and Ntranos, V. (2023). Genome-wide prediction of disease variant effects with a deep protein language model. Nat. Genet. *55*, 1512–1522.

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

17. Zhou, Z., Ji, Y., Li, W., Dutta, P., Davuluri, R.V., and Liu, H. (2023). DNABERT-2: Efficient Foundation Model and Benchmark For Multi-Species Genomes.

18. Nguyen, E., Poli, M., Durrant, M.G., Kang, B., Katrekar, D., Li, D.B., Bartie, L.J., Thomas, A.W., King, S.H., Brixi, G., et al. (2024). Sequence modeling and design from molecular to genome scale with Evo. Science *386*, eado9336.

19. Brixi, G., Durrant, M.G., Ku, J., Poli, M., Brockman, G., Chang, D., Gonzalez, G.A., King, S.H., Li, D.B., Merchant, A.T., et al. (2025). Genome modeling and design across all domains of life with Evo 2. bioRxiv. https://doi.org/10.1101/2025.02.18.638918.

20. Benegas, G., Batra, S.S., and Song, Y.S. (2023). DNA language models are powerful predictors of genome-wide variant effects. Proc. Natl. Acad. Sci. U. S. A. *120*, e2311219120.

21. Benegas, G., Albors, C., Aw, A.J., Ye, C., and Song, Y.S. (2025). A DNA language model based on multispecies alignment predicts the effects of genome-wide variants. Nat. Biotechnol., 1–6.

22. Mendoza-Revilla, J., Trop, E., Gonzalez, L., Roller, M., Dalla-Torre, H., de Almeida, B.P., Richard, G., Caton, J., Lopez Carranza, N., Skwark, M., et al. (2024). A foundational large language model for edible plant genomes. Commun. Biol. *7*, 835.

23. Zhai, J., Gokaslan, A., Schiff, Y., Berthel, A., Liu, Z.-Y., Lai, W.-Y., Miller, Z.R., Scheben, A., Stitzer, M.C., Romay, M.C., et al. (2025). Cross-species modeling of plant genomes at single-nucleotide resolution using a pretrained DNA language model. Proc. Natl. Acad. Sci. U. S. A. *122*, e2421738122.

24. Schiff, Y., Kao, C.-H., Gokaslan, A., Dao, T., Gu, A., and Kuleshov, V. (2024). Caduceus: Bi-Directional Equivariant Long-Range DNA Sequence Modeling. arXiv [q-bio.GN].

25. Dalla-Torre, H., Gonzalez, L., Mendoza-Revilla, J., Lopez Carranza, N., Grzywaczewski, A.H., Oteri, F., Dallago, C., Trop, E., de Almeida, B.P., Sirelkhatim, H., et al. (2025). Nucleotide Transformer: building and evaluating robust foundation models for human genomics. Nat. Methods *22*, 287–297.

26. Wang, N., Bian, J., Li, Y., Li, X., Mumtaz, S., Kong, L., and Xiong, H. (2024). Multi-purpose RNA language modelling with motif-aware pretraining and type-guided fine-tuning. Nat. Mach. Intell. *6*, 548–557.

27. Shen, T., Hu, Z., Sun, S., Liu, D., Wong, F., Wang, J., Chen, J., Wang, Y., Hong, L., Xiao, J., et al. (2024). Accurate RNA 3D structure prediction using a language model-based deep learning approach. Nat. Methods *21*, 2287–2298.

28. Yu, H., Yang, H., Sun, W., Yan, Z., Yang, X., Zhang, H., Ding, Y., and Li, K. (2024). An interpretable RNA foundation model for exploring functional RNA motifs in plants. Nat. Mach. Intell. *6*, 1616–1625.

29. Hao, M., Gong, J., Zeng, X., Liu, C., Guo, Y., Cheng, X., Wang, T., Ma, J., Zhang, X., and Song, L. (2024). Large-scale foundation model on single-cell transcriptomics. Nat. Methods *21*, 1481–1491.

30. Zeng, Y., Xie, J., Shangguan, N., Wei, Z., Li, W., Su, Y., Yang, S., Zhang, C., Zhang, J., Fang, N., et al. (2025). CellFM: a large-scale foundation model pre-trained on transcriptomics of 100 million human cells. Nat. Commun. *16*, 4679.

31. Cui, H., Wang, C., Maan, H., Pang, K., Luo, F., Duan, N., and Wang, B. (2024). scGPT: toward building a foundation model for single-cell multi-omics using generative AI. Nat. Methods *21*, 1470–1480.

32. Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv [cs.CL].

33. Zhou, J., and Troyanskaya, O.G. (2015). Predicting effects of noncoding variants with deep learning-based sequence model. Nat. Methods *12*, 931–934.

34. Quang, D., and Xie, X. (2016). DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. Nucleic Acids Res. *44*, e107.

35. Kelley, D.R., Snoek, J., and Rinn, J.L. (2016). Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. Genome Res. *26*, 990–999.

36. Benegas, G., Ye, C., Albors, C., Li, J.C., and Song, Y.S. (2025). Genomic language models: opportunities and challenges. Trends Genet. *41*, 286–302.

37. Gu, A., and Dao, T. (2023). Mamba: Linear-Time Sequence Modeling with Selective State Spaces. arXiv [cs.LG].

38. Schmitz, R.J., Grotewold, E., and Stam, M. (2022). Cis-regulatory sequences in plants: Their importance, discovery, and future challenges. Plant Cell *34*, 718–741.

39. Marand, A.P., Eveland, A.L., Kaufmann, K., and Springer, N.M. (2023). Cis-regulatory elements in plant development, adaptation, and evolution. Annu. Rev. Plant Biol. *74*, 111–137.

40. Engelhorn, J., Snodgrass, S.J., Kok, A., Seetharam, A.S., Schneider, M., Kiwit, T., Singh, A., Banf, M., Khaipho-Burch, M., Runcie, D.E., et al. (2024). Genetic variation at transcription factor binding sites largely explains phenotypic heritability in maize. bioRxiv, 2023.08.08.551183. https://doi.org/10.1101/2023.08.08.551183.

41. Marand, A.P., Jiang, L., Gomez-Cano, F., Minow, M.A.A., Zhang, X., Mendieta, J.P., Luo, Z., Bang, S., Yan, H., Meyer, C., et al. (2025). The genetic architecture of cell type-specific cis regulation in maize. Science *388*, eads6601.

42. Magallón, S., and Castillo, A. (2009). Angiosperm diversification through time. Am. J. Bot. *96*, 349–365.

43. Bar-On, Y.M., Phillips, R., and Milo, R. (2018). The biomass distribution on Earth. Proc. Natl. Acad. Sci. U. S. A. *115*, 6506–6511.

44. The Angiosperm Phylogeny Group (2016). An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. Bot. J. Linn. Soc. *181*, 1–20.

45. Dao, T., and Gu, A. (2024). Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. arXiv [cs.LG].

46. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., and Polosukhin, I. (2017). Attention Is All You Need. arXiv [cs.CL].

47. Warner, B., Chaffin, A., Clavié, B., Weller, O., Hallström, O., Taghadouini, S., Gallagher, A., Biswas, R., Ladhak, F., Aarsen, T., et al. (2024). Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. arXiv [cs.CL].

48. Gu, A., Goel, K., and Ré, C. (2021). Efficiently Modeling Long Sequences with Structured State Spaces. arXiv [cs.LG].

49. Novák, P., Guignard, M.S., Neumann, P., Kelly, L.J., Mlinarec, J., Koblížková, A., Dodsworth, S., Kovařík, A., Pellicer, J., Wang, W., et al. (2020). Repeat-sequence turnover shifts fundamentally in species with large genomes. Nat Plants *6*, 1325–1329.

50. Sun, S., Wang, B., Li, C., Xu, G., Yang, J., Hufford, M.B., Ross-Ibarra, J., Wang, H., and Wang, L. (2023). Unraveling Prevalence and Effects of Deleterious Mutations in Maize Elite Lines across Decades of Modern Breeding. Mol. Biol. Evol. *40*. https://doi.org/10.1093/molbev/msad170.

51. Lozano, R., Gazave, E., dos Santos, J.P.R., Stetter, M.G., Valluru, R., Bandillo, N., Fernandes, S.B., Brown, P.J., Shakoor, N., Mockler, T.C., et al. (2021). Comparative evolutionary genetics of deleterious load in sorghum and maize. Nature Plants *7*, 17–24.

52. Lye, Z., Choi, J.Y., and Purugganan, M.D. (2022). Deleterious Mutations and the Rare Allele Burden on Rice Gene Expression. Mol. Biol. Evol. *39*. https://doi.org/10.1093/molbev/msac193.

53. Mezmouk, S., and Ross-Ibarra, J. (2014). The pattern and distribution of deleterious mutations in maize. G3 *4*, 163–171.

54. Song, B., Buckler, E.S., and Stitzer, M.C. (2024). New whole-genome alignment tools are needed for tapping into plant diversity. Trends Plant Sci. *29*, 355–369.

55. Huber, C.D., Kim, B.Y., and Lohmueller, K.E. (2020). Population genetic models of GERP scores suggest pervasive turnover of constrained sites across mammalian evolution. PLoS Genet. *16*, e1008827.

56. Bennetzen, J.L., and Wang, H. (2014). The contributions of transposable elements to the structure, function, and evolution of plant genomes. Annu. Rev. Plant Biol. *65*, 505–530.

57. Kozak, M. (1986). Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. Cell *44*, 283–292.

58. Tian, F., Yang, D.-C., Meng, Y.-Q., Jin, J., and Gao, G. (2020). PlantRegMap: charting functional regulatory maps in plants. Nucleic Acids Res. *48*, D1104–D1113.

59. Lu, Z., Marand, A.P., Ricci, W.A., Ethridge, C.L., Zhang, X., and Schmitz, R.J. (2019). The prevalence, evolution and chromatin signatures of plant regulatory elements. Nat. Plants *5*, 1250–1259.

60. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2021). LoRA: Low-Rank Adaptation of large language models. arXiv [cs.CL].

61. Marand, A.P., Chen, Z., Gallavotti, A., and Schmitz, R.J. (2021). A cis-regulatory atlas in maize at single-cell resolution. Cell *184*, 3041–3055.e21.

62. Wrightsman, T., Ferebee, T.H., Romay, M.C., Seetharam, A.S., AuBuchon-Elder, T., Phillips, A.R., Syring, M., Hufford, M.B., Kellogg, E.A., and Buckler, E.S. (2024). Current genomic deep learning architectures generalize across grass species but not alleles. bioRxiv, 2024.04.11.589024. https://doi.org/10.1101/2024.04.11.589024.

63. Huang, C., Shuai, R.W., Baokar, P., Chung, R., Rastogi, R., Kathail, P., and Ioannidis, N.M. (2023). Personal transcriptome variation is poorly explained by current genomic deep learning models. Nat. Genet. *55*, 2056–2059.

64. Tang, Z., Toneyan, S., and Koo, P.K. (2023). Current approaches to genomic deep learning struggle to fully capture human genetic variation. Nat. Genet. *55*, 2021–2022.

65. Liu, T., Zhang, X., Ying, R., and Zhao, H. (2025). Pre-training genomic language model with variants for better modeling functional genomics. bioRxiv, 2025.02.26.640468. https://doi.org/10.1101/2025.02.26.640468.

66. Jaganathan, K., Ersaro, N., Novakovsky, G., Wang, Y., James, T., Schwartzentruber, J., Fiziev, P., Kassam, I., Cao, F., Hawe, J., et al. (2025). Predicting expression-altering promoter mutations with deep learning. Science *389*, eads7373.

67. Li, T., Xu, H., Teng, S., Suo, M., Bahitwa, R., Xu, M., Qian, Y., Ramstein, G.P., Song, B., Buckler, E.S., et al. (2024). Modeling 0.6 million genes for the rational design of functional cis-regulatory variants and de novo design of cis-regulatory sequences. Proc. Natl. Acad. Sci. U. S. A. *121*, e2319811121.

68. Clark, R.M., Wagler, T.N., Quijada, P., and Doebley, J. (2006). A distant upstream enhancer at the maize domestication gene tb1 has pleiotropic effects on plant and inflorescent architecture. Nat. Genet. *38*, 594–597.

69. Studer, A., Zhao, Q., Ross-Ibarra, J., and Doebley, J. (2011). Identification of a functional transposon insertion in the maize domestication gene tb1. Nat. Genet. *43*, 1160–1163.

70. Tu, X., Mejía-Guerra, M.K., Valdes Franco, J.A., Tzeng, D., Chu, P.-Y., Shen, W., Wei, Y., Dai, X., Li, P., Buckler, E.S., et al. (2020). Reconstructing the maize leaf regulatory network using ChIP-seq data of 104 transcription factors. Nat. Commun. *11*, 5089.

71. Smith, S.A., and Brown, J.W. (2018). Constructing a broadly inclusive seed plant phylogeny.

Am. J. Bot. *105*, 302–314.

72. Song, B., Buckler, E.S., Wang, H., Wu, Y., Rees, E., Kellogg, E.A., Gates, D.J., Khaipho-Burch, M., Bradbury, P.J., Ross-Ibarra, J., et al. (2021). Conserved noncoding sequences provide insights into regulatory sequence and loss of gene expression in maize. Genome Res. *31*, 1245–1257.

73. Loshchilov, I., and Hutter, F. (2017). Decoupled Weight Decay Regularization. arXiv [cs.LG].

74. Welker, C.A.D., McKain, M.R., Estep, M.C., Pasquet, R.S., Chipabika, G., Pallangyo, B., and Kellogg, E.A. (2020). Phylogenomics enables biogeographic analysis and a new subtribal classification of Andropogoneae (Poaceae—Panicoideae). J. Syst. Evol. *58*, 1003–1030.

75. Stitzer, M.C., Seetharam, A.S., Scheben, A., Hsu, S.-K., Schulz, A.J., AuBuchon-Elder, T.M., El-Walid, M., Ferebee, T.H., Hale, C.O., La, T., et al. (2025). Extensive genome evolution distinguishes maize within a stable tribe of grasses. bioRxivorg, 2025.01.22.633974. https://doi.org/10.1101/2025.01.22.633974.

76. Song, B., Marco-Sola, S., Moreto, M., Johnson, L., Buckler, E.S., and Stitzer, M.C. (2022). AnchorWave: Sensitive alignment of genomes with high sequence diversity, extensive structural polymorphism, and whole-genome duplication. Proc. Natl. Acad. Sci. U. S. A. *119*. https://doi.org/10.1073/pnas.2113075119.

77. Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R., and Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. *20*, 110–121.

78. Hufford, M.B., Seetharam, A.S., Woodhouse, M.R., Chougule, K.M., Ou, S., Liu, J., Ricci, W.A., Guo, T., Olson, A., Qiu, Y., et al. (2021). De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. Science *373*, 655–662.

79. Gao, L., Gonda, I., Sun, H., Ma, Q., Bao, K., Tieman, D.M., Burzynski-Chang, E.A., Fish, T.L., Stromberg, K.A., Sacks, G.L., et al. (2019). The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. Nat. Genet. *51*, 1044–1051.

80. Wrightsman, T., Marand, A.P., Crisp, P.A., Springer, N.M., and Buckler, E.S. (2022). Modeling chromatin state from sequence across angiosperms using recurrent convolutional neural networks. Plant Genome *15*, e20249.

81. Wu, H.-Y.L., Ai, Q., Teixeira, R.T., Nguyen, P.H.T., Song, G., Montes, C., Elmore, J.M., Walley, J.W., and Hsu, P.Y. (2024). Improved super-resolution ribosome profiling reveals prevalent translation of upstream ORFs and small ORFs in Arabidopsis. Plant Cell *36*, 510–539.

82. Zhu, W., Miao, X., Qian, J., Chen, S., Jin, Q., Li, M., Han, L., Zhong, W., Xie, D., Shang, X., et al. (2023). A translatome-transcriptome multi-omics gene regulatory network reveals the complicated functional landscape of maize. Genome Biol. *24*, 60.

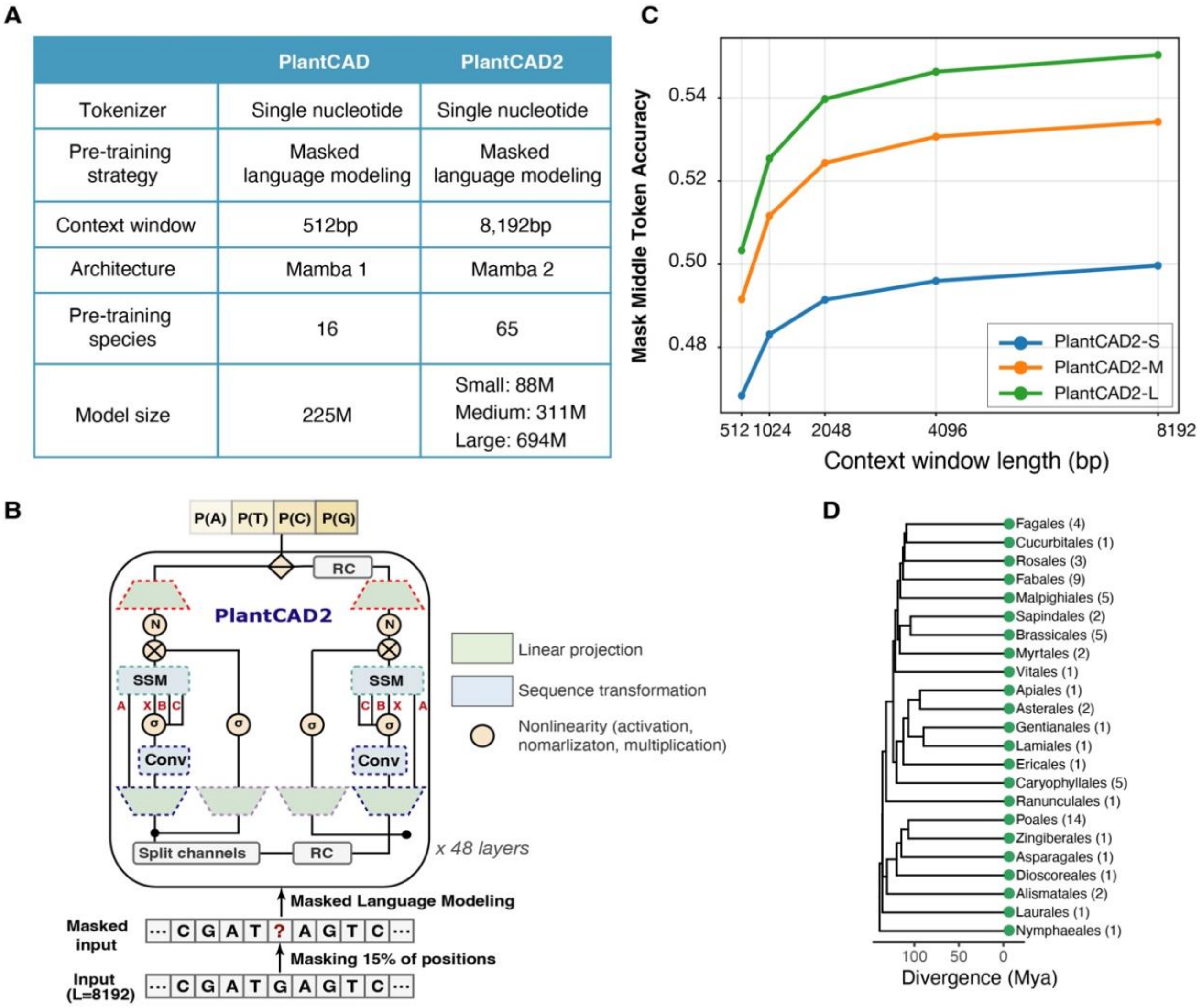
83. Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120.

84. Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. *10*, R25.

85. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

86. Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.-C., Mendell, J.T., and Salzberg, S.L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat. Biotechnol. *33*, 290–295.

87. Mangrulkar, S., Gugger, S., Debut, L., Belkada, Y., Paul, S., and Bossan, B. (2022). Peft: State-of-the-art parameter-efficient fine-tuning methods. In Peft: State-of-the-art parameter-efficient fine-tuning methods.



	PlantCAD	PlantCAD2
Tokenizer	Single nucleotide	Single nucleotide
Pre-training strategy	Masked language modeling	Masked language modeling
Context window	512bp	8,192bp
Architecture	Mamba 1	Mamba 2
Pre-training species	16	65
Model size	225M	Small: 88M Medium: 311M Large: 694M

Figure 1. Overview of the PlantCAD2 model. (A) Comparison of PlantCAD1 and PlantCAD2 model configurations. PlantCAD2 introduces a longer context window, upgraded architecture (Mamba2), expanded pre-training species set, and scaled model sizes (small: 88M, medium: 311M, large: 694M parameters), while maintaining single-nucleotide tokenization. **(B)** Schematic of the PlantCAD2 architecture based on Mamba2 with reverse-complement (RC) equivariance, convolutional and state space modules (SSM), and a masked language modeling objective applied to 8,192 bp input sequences. **(C)** Effect of context window length on model performance. The y-axis shows the prediction accuracy of three models when masking the single central token in the held-out test set. **(D)** Phylogenetic distribution of the 65 angiosperm genomes across flowering plant orders. Numbers in parentheses indicate the number of species included from each order.

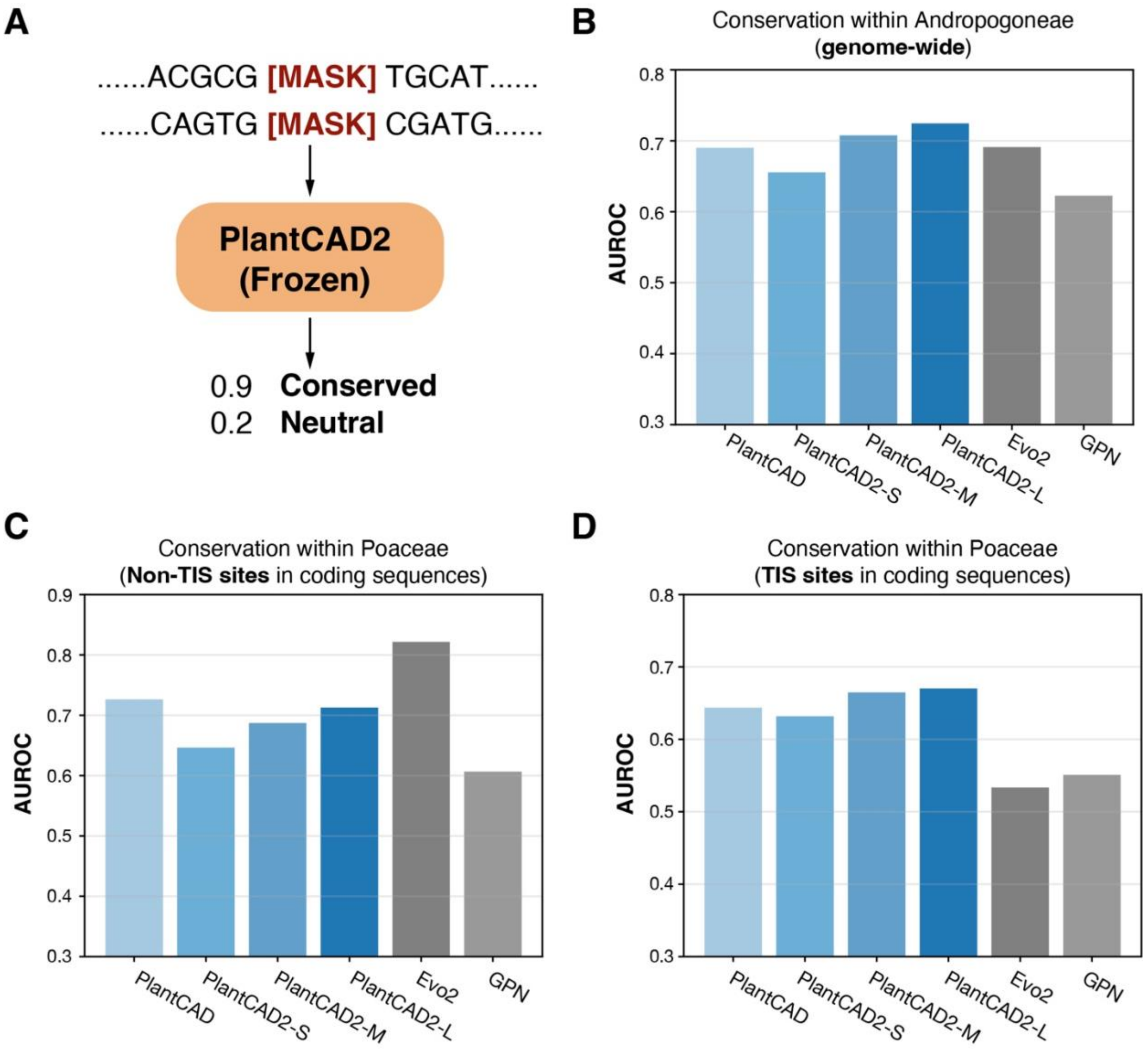


Figure 2. PlantCAD2 accurately predicts evolutionary conservation using zero-shot strategy. (A) Zero-shot conservation prediction approach using masked token probabilities. **(B)** AUROC of conservation of the Sorghum genome within the Andropogoneae tribe. **(C)** AUROC of conservation within Poaceae for non-TIS sites in coding sequences. **(D)** AUROC of conservation within Poaceae for TIS sites in coding sequences.

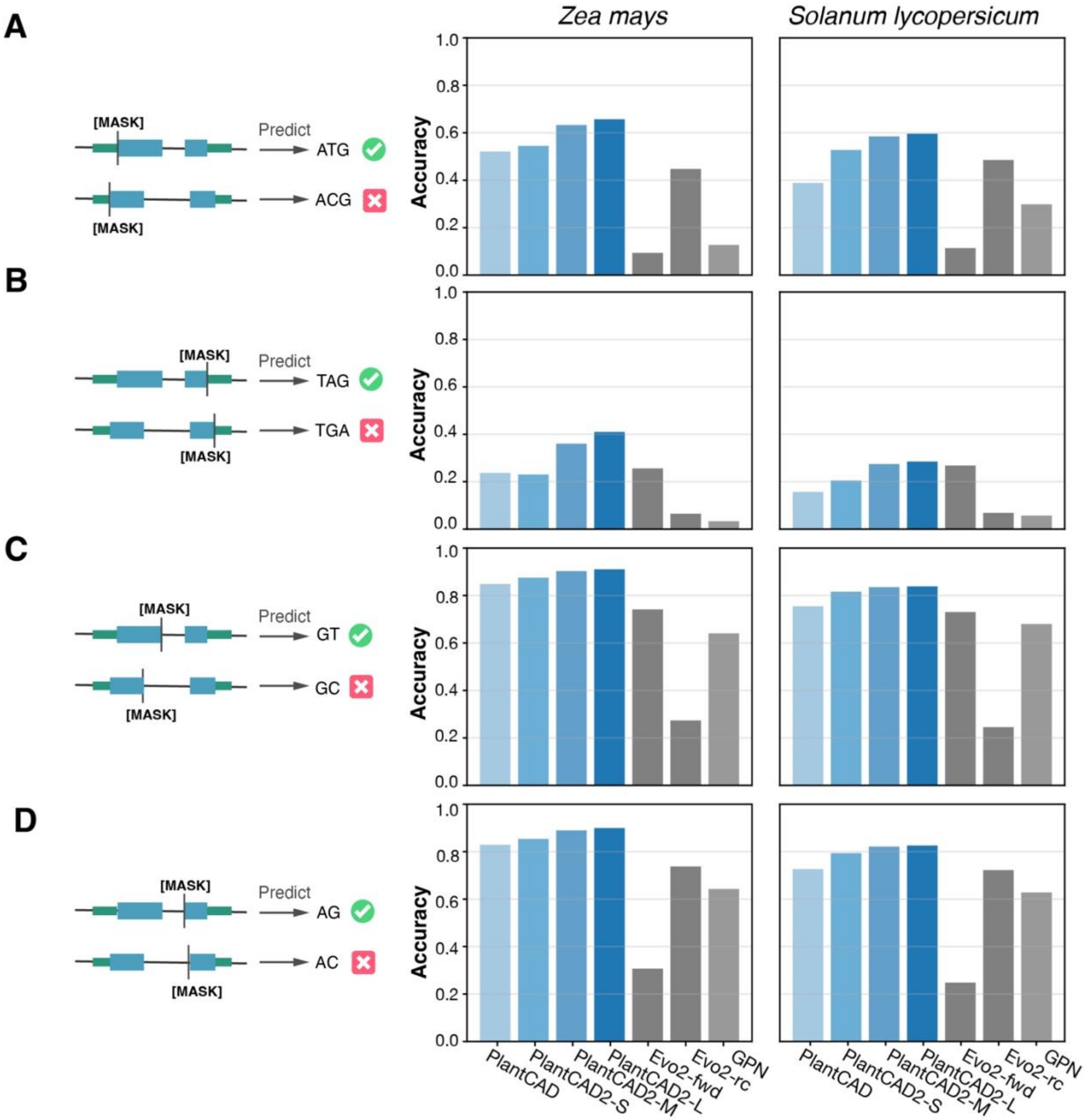


Figure 3. PlantCAD2 accurately predicts transcriptional and translational junction sites using zero-shot masked motif prediction. Left panels show the masking strategy where canonical motifs are replaced with [MASK] tokens and models predict the correct sequence. Right panels show prediction accuracy for each model on maize (left, included in training) and tomato (right, excluded from training). **(A)** Translation initiation sites (ATG masking). **(B)** Translation termination sites (TAG/TGA/TAA masking). **(C)** Splice donor sites (GT masking). **(D)** Splice acceptor sites (AG masking).

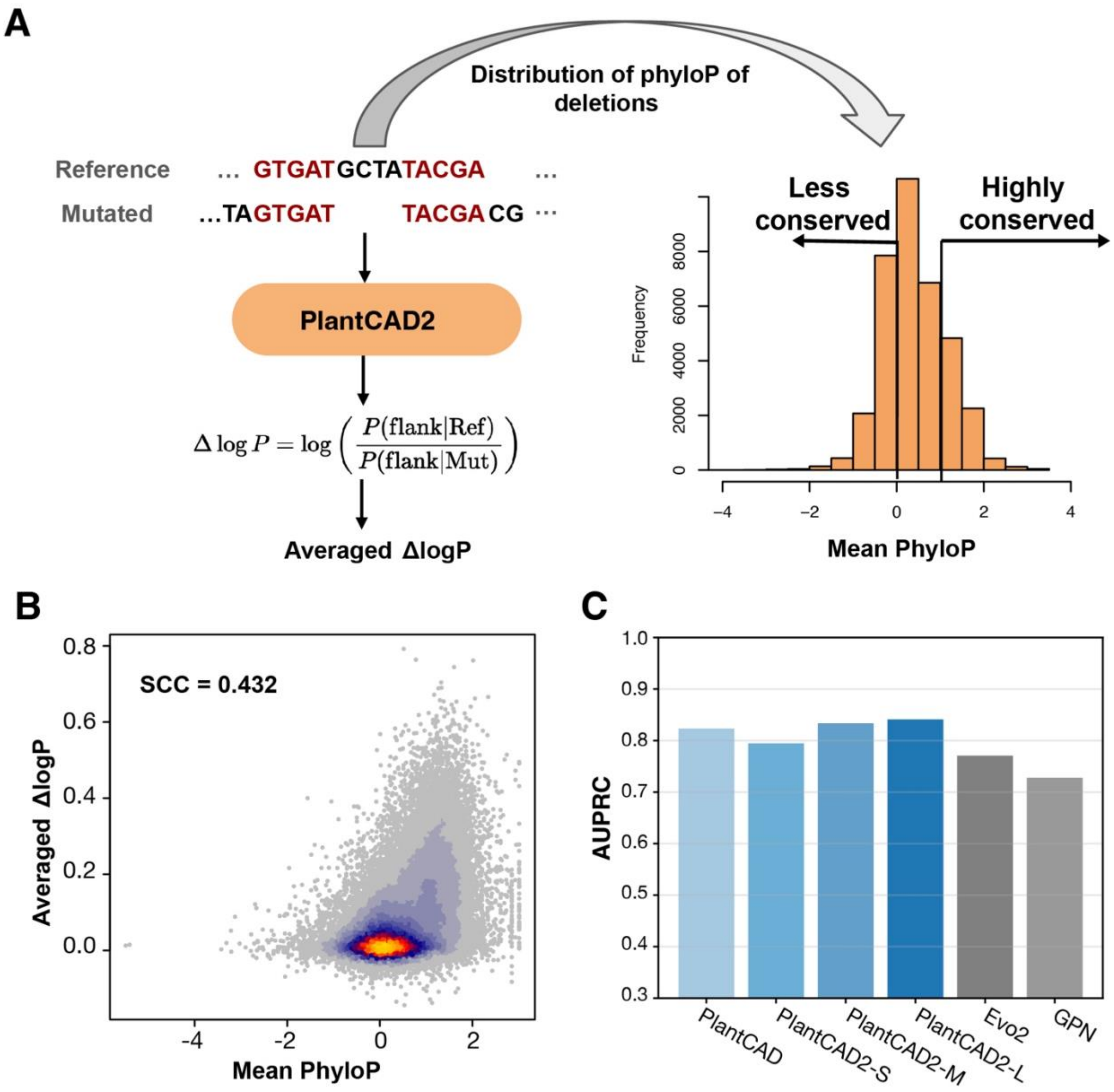


Figure 4. PlantCAD2 predicts functional impact of structural variants using zero-shot strategy. (A) ΔlogP calculation approach for deletion variants and phyloP score distribution for classification. **(B)** Scatter plot showing the positive correlation between PlantCAD2's ΔlogP scores and phyloP-based conservation scores. **(C)** AUROC performance distinguishes highly conserved from less conserved deletions.

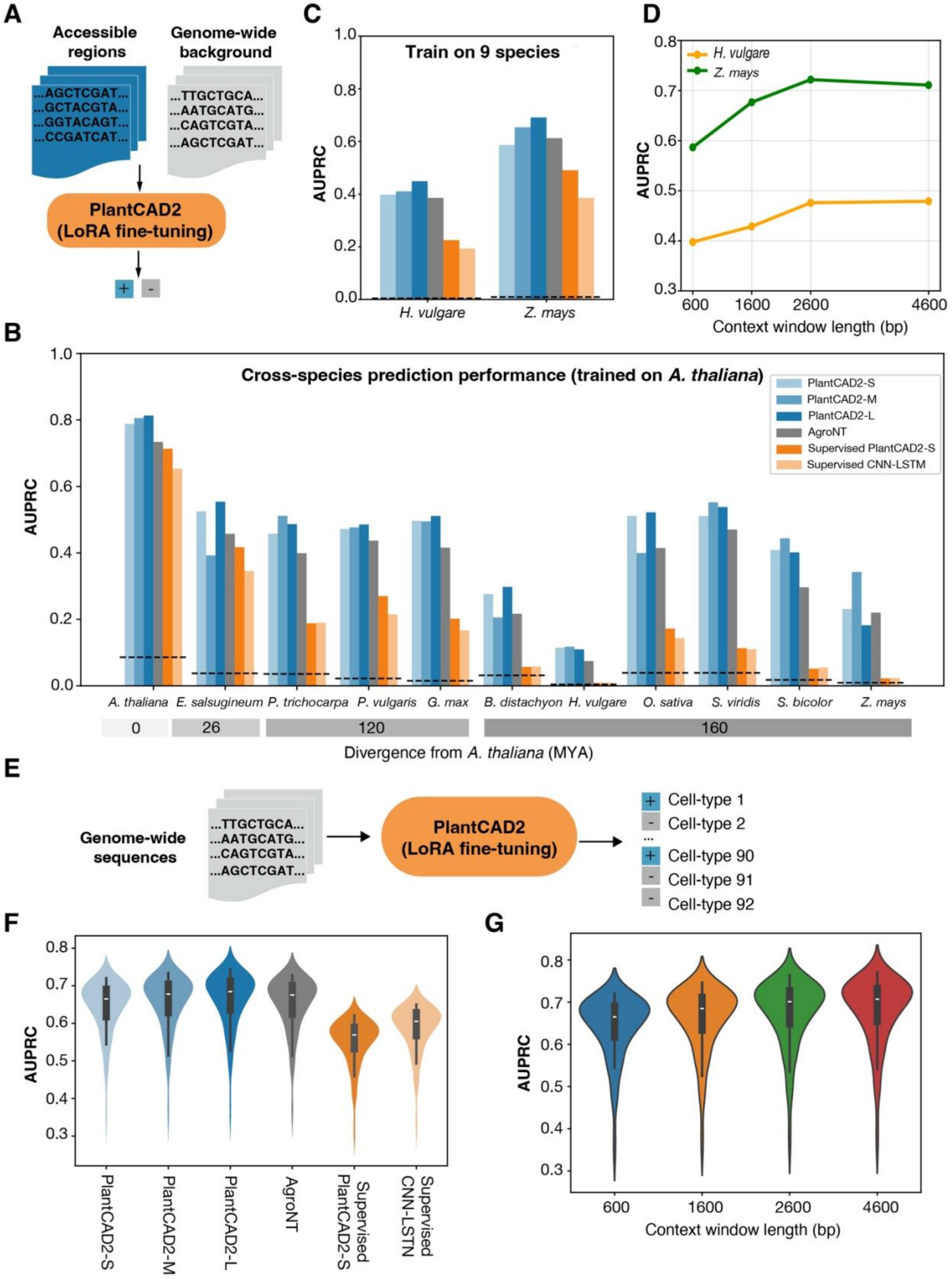

A
Accessible regions
Genome-wide background
...AGCTCGAT...
...GCTACGTA...
...GGTACAGT...
...CCGATCAT...
...TTGCTGCA...
...AATGCATG...
...CAGTCGTA...
...AGCTCGAT...
PlantCAD2 (LoRA fine-tuning)
+
-
C
Train on 9 species
AUPRC
H. vulgare
Z. mays
D
H. vulgare
Z. mays
AUPRC
Context window length (bp)
600
1600
2600
4600
B
Cross-species prediction performance (trained on A. thaliana)
AUPRC
PlantCAD2-S
PlantCAD2-M
PlantCAD2-L
AgroNT
Supervised PlantCAD2-S
Supervised CNN-LSTM
A. thaliana
E. salsugineum
P. trichocarpa
P. vulgaris
G. max
B. distachyon
H. vulgare
O. sativa
S. viridis
S. bicolor
Z. mays
0
26
120
160
Divergence from A. thaliana (MYA)
E
Genome-wide sequences
PlantCAD2 (LoRA fine-tuning)
Cell-type 1
Cell-type 2
...
Cell-type 90
Cell-type 91
Cell-type 92
F
AUPRC
PlantCAD2-S
PlantCAD2-M
PlantCAD2-L
AgroNT
Supervised PlantCAD2-S
Supervised CNN-LSTN
G
AUPRC
Context window length (bp)
600
1600
2600
4600

Figure 5. PlantCAD2 predicts chromatin accessibility across species and cell types. **(A)** LoRA fine-tuning approach for binary accessibility prediction using ATAC-seq peaks versus genomic background. **(B)** Cross-species AUPRC performance when trained on Arabidopsis, showing superior generalization of PlantCAD2 models compared to supervised baselines across evolutionary distances. **(C)** Multi-species training performance on held-out barley and maize. **(D)** Effect of context window length on accessibility prediction accuracy for PlantCAD2-S. **(E)** Multi-label classification approach for cell-type-specific accessibility prediction. **(F)** Performance comparison across models for 92 cell types . **(G)** Context window effects on cell-type-specific prediction accuracy for PlantCAD2-S.

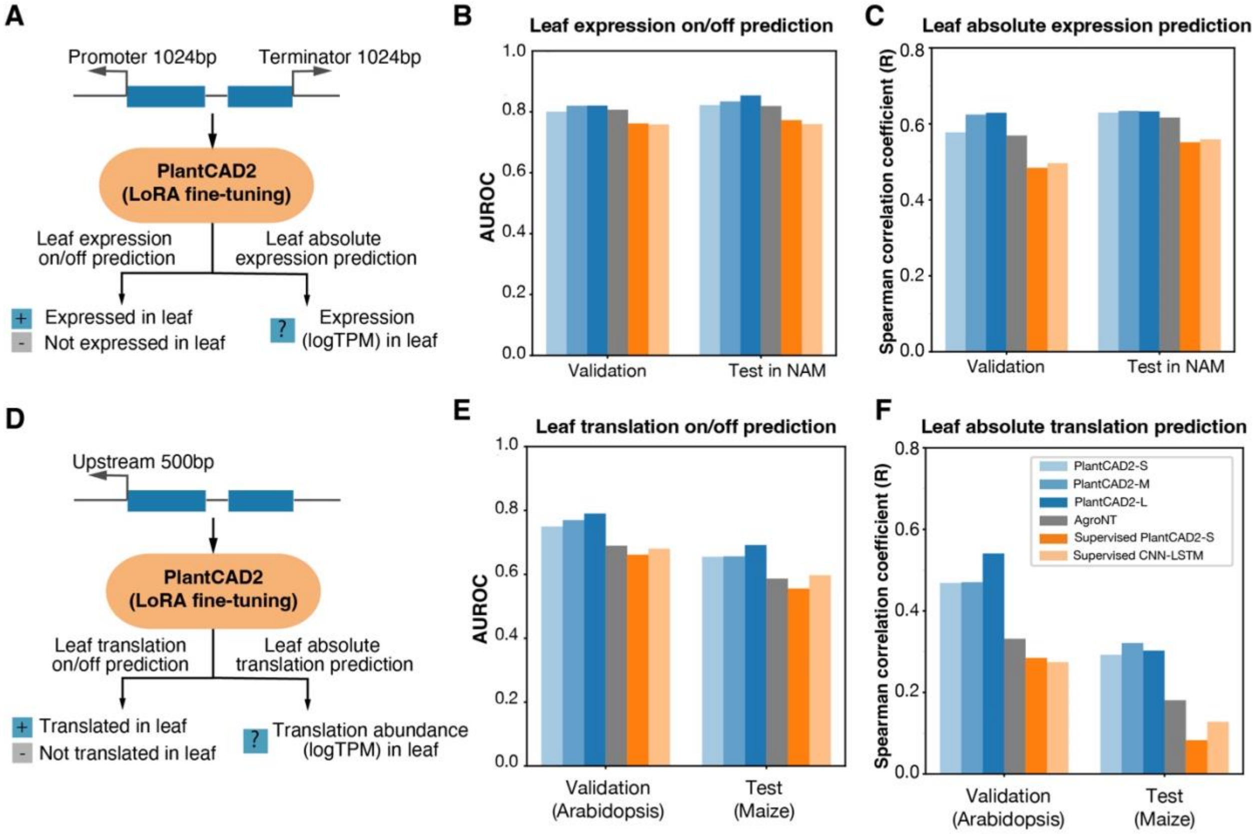


Figure 6. PlantCAD2 predicts gene expression and translation across species. (A) Gene expression prediction pipeline using promoter and terminator sequences (1024 bp each) for binary classification and regression tasks. **(B-C)** Cross-species gene expression performance on maize NAM population for binary on/off prediction (B) and absolute expression levels (C). **(D)** Translation prediction pipeline using 500 bp upstream sequences. **(E-F)** Translation prediction performance trained on Arabidopsis and tested cross-species on maize for binary on/off prediction (E) and absolute translation levels (F).