A Foundational Large Language Model for Edible Plant Genomes

Javier Mendoza-Revilla[*1], Evan Trop[*1], Liam Gonzalez[1], Maša Roller[1], Hugo Dalla-Torre[1], Bernardo P. de Almeida[1], Guillaume Richard[1], Jonathan Caton[2], Nicolas Lopez Carranza[1], Marcin Skwark[1], Alex Laterre[1], Karim Beguir[1], Thomas Pierrot[†‡1], Marie Lopez[†‡1]

[1]**InstaDeep** [2]**Google DeepMind**

Abstract

In recent years, significant progress has been made in the field of plant genomics, demonstrated by the increased use of high-throughput methodologies that allow for the characterization of multiple genome-wide molecular phenotypes. These results have provided valuable insights into plant traits and their underlying genetic mechanisms, especially in well-researched model plant species. Nonetheless, although acquiring and characterizing these molecular phenotypes can offer valuable insights into plant traits, effectively leveraging them to make accurate predictions represents a critical step in crop genomic improvement. We present AgroNT, a novel foundational large language model trained on reference genomes from 48 plant species with a predominant focus on crop species. We show that AgroNT can obtain state-of-the-art predictions for many genomic elements, including polyadenylation sites, splice sites, open chromatin and enhancer regions. Furthermore, AgroNT can be used to predict the strength of promoter sequences and tissue-specific gene expression levels or prioritize functional variants. Using the cassava genome as an example of an understudied species, we perform a large-scale *in silico* saturation mutagenesis analysis to assess the impact of >10 million mutations on gene expression levels and enhancer elements in the cassava genome, and provide the results as a valuable resource for regulatory causal variant characterization. Furthermore, owing to the lack of comprehensive benchmarks in the context of deep learning-based methods in plant genomic research, we propose the use of the multiple datasets encompassing seven distinct genomic prediction tasks, which have been compiled here, as the Plants Genomic Benchmark (PGB). The pre-trained AgroNT model is now publicly available on Hugging Face at `https://huggingface.co/InstaDeepAI/agro-nt` for future research purposes.

Introduction

The advent of high-throughput next-generation sequencing has led to a vast increase in available genomic data in the field of plant sciences. Since the completion of the genome sequence of the model plant *Arabidopsis thaliana* over 20 years ago [1], more than 200 plant species' genome sequences have been published to date [2]. However, the generation of a species' assembly is merely the initial step in the process of understanding the genome. Extensive additional experiments and computational processing are necessary for the crucial task of structural and functional annotation of important genomic regions, such as genes or regulatory elements. Furthermore, many of the plant species whose genomes have been sequenced still lack sufficient experimental resources from these complementary analyses. These include the so called 'orphan crops', i.e. plants important for regional food and economic security that are not traded world-wide. Due to the lack of comprehensive transcriptomic, regulatory, or proteomic experiments we are limited in understanding their growth, senescence, yield, and responses to biotic and abiotic stresses [3, 4, 5, 6], hampering the utilization of modern improvement tools, such as high-throughput (molecular) phenotyping, genomic selection, and genome editing. In light of

[*]**These authors contributed equally**
[†]**These authors jointly supervised this work**
[‡]**Corresponding authors**: t.pierrot@instadeep.com & m.lopez@instadeep.com

this, novel approaches that can accurately predict gene annotations and regulatory genomic features directly from DNA sequences have the potential to provide valuable biological insights and assist in genomic editing applications. This is mainly because once an accurate predictive model for a particular biological process of interest has been developed, researchers can enjoy complete freedom to assess the effect of any variant. For instance, in the work of Rodriguez-Leal et al., the authors edited the tomato *CLAVATA3* gene (SlCLV3) promoter to increase fruit size and optimize inflorescence branching [7]. Due to the lack of functional annotations in the SlCLV3 promoter, they had to employ saturated promoter mutagenesis using the CRISPR/Cas9 system, followed by the selection of mutants with desirable fruit and inflorescence traits. An accurate predictive model could have potentially facilitated the identification of key regulatory elements in the SlCLV3 promoter enabling the implementation of a model-guided approach for promoter editing based purely on *in silico*-based analysis. Indeed, such approaches have demonstrated their feasibility and power in human genetics, successfully revealing variants, including rare alleles, underlying specific genetic diseases [8, 9].

The complexity associated with sequence determinants of gene structure and regulatory features makes end-to-end deep learning-based approaches highly suitable for direct learning from DNA sequences to achieve accurate predictions of specific outcomes. The majority of deep learning approaches have heavily relied on supervised learning to train neural networks in a task specific manner [10, 11, 12, 13, 14]. While successful across various applications, this approach depends on abundant labeled data, which can lead to sub-optimal performance and limited usability, especially in scenarios with data scarcity. Obtaining high-quality labeled data is often a time-consuming and costly process. Additionally, most of these models have used convolutional neural networks (CNNs), recurrent neural networks (RNN's), or combinations of the two as their model architectures. Although effective these architectures have been outperformed by transformer architectures as evidenced by their adoption in both the computer vision and natural language processing fields [15, 16, 17]. To address these limitations and exploit the transformer architecture, large language models (LLMs) have gained considerable popularity in the field of biology [18, 19, 20, 21, 22, 23, 24, 25, 26, 27], offering the ability to be trained on unlabeled data and generate general-purpose representations capable of solving specific tasks. Furthermore, LLMs overcome a current limitation of other deep learning-based models, as they are not reliant on single reference genomes, which often provide an incomplete and biased genomic diversity depiction from a limited number of individuals. LLMs can leverage multiple reference genomes, including those from genetically distant species, thereby increasing overall diversity, which has been shown to significantly enhance prediction performance [24]. This diversity is particularly relevant in plant species due to the structural complexities of their genomes, which hinder accurate mapping of polymorphisms across whole-genome alignments. Moreover, depending on the plant species under study, there may be a lack of large and representative collections of individual genomes, necessitating the use of genomic diversity from other species to improve prediction performance. Finally, LLMs are well-suited for zero-shot learning, a transfer learning approach that enables the model to recognize and classify samples from new classes not encountered during training. This ability stems from their generalizability and comprehensive language understanding [28]. Zero-shot predictions represent an alternative approach to the traditional method of training supervised models on large amounts of functional genomic data. Since LLMs are trained solely on DNA sequences, zero-shot predictions can be readily obtained, even for understudied plant species like orphan crops, in the absence of such functional genomic data. In the context of plant genomic research, only two LLMs have been developed so far. Nonetheless, one of these was exclusively evaluated for its capability to predict the functional impact of genetic variants in *Arabidopsis thaliana* [29], while the other focused solely on predicting gene expression levels in various tissues of maize (*Zea mays*) [30].

In this study, we introduce a novel DNA large language model called the Agronomic Nucleotide Transformer (AgroNT), which is based on the transformer architecture. AgroNT was trained using reference genomes from 48 plant species, with a primary focus on edible plant species. We assessed the performance of AgroNT across several prediction tasks ranging from regulatory features, RNA processing, and gene expression, and show that AgroNT can obtain state-of-the-art performance. We also illustrate the capability of AgroNT to improve prioritization of functional variants through zero-shot prediction. By exploiting the scalability of AgroNT, we further characterized the mutation space of promoter se-

quences and enhancer regulatory regions in the agriculturally significant 'orphan crop', *Manihot esculenta*, commonly known as cassava. Through *in silico* saturation mutagenesis we profiled over 10 million single point mutations. As a major contribution of this study, we provide the predicted impact of these variants for future research. Furthermore, we have made the pre-trained AgroNT model available for future research. Given the lack of comprehensive benchmarks in the context of deep learning-based methods in plant genomic research, we propose the use of the datasets covering several distinct genomic prediction tasks compiled here as the Plant Genomic Benchmark (PGB). Our aim is to enhance and provide a comprehensive assessment of the performance of deep learning methods in plant genomic studies. Overall, the results of our study emphasize the significant capabilities of AgroNT in the field of plant genomics and highlight its particular relevance for under-researched plant species.

Results

AgroNT: a novel large language model that integrates genomes across plants species

We developed a transformer-based DNA language model named the Agronomic Nucleotide Transformer (AgroNT), which learned general nucleotide sequence representations from genomic DNA sequences of 48 different plant species (Fig. 1a; Methods). Building upon our previous work [24], our pre-training strategy involves performing masked language modeling (MLM) on a DNA sequence consisting of 6,000 base pairs (bp). Our tokenization algorithm splits the DNA sequence into 6-mers, treating each 6-mer as a token, and masks 15% of the tokens for prediction (Fig. 1b; Methods). For our fine-tuning strategy, we implemented parameter-efficient fine-tuning using the IA3 technique [31]. In this approach, we replaced the language model head with a prediction head, using either a classification or regression head based on the task. We kept the weights of the transformer layers and embedding layers frozen, or alternatively, unfroze a small number of the final layers to reduce training time for specific downstream tasks (Fig. 1c; Methods). We introduced new learnable weights for the fine-tuning process. Throughout the study, we evaluated the performance of AgroNT across 7 genomic prediction tasks (Table 1) using exclusively our fine-tuning method and compared the obtained performance to other state-of-the-art methods, where available.

Polyadenilation site, splice site, and long non-coding RNA prediction

We first evaluated the predictive capability of AgroNT in predicting alternative polyadenylation (APA), a process characterized by the selection of different polyadenilation sites within the same gene. APA has been recognized as a key regulator of gene expression in various eukaryotes, including plants [32]. For this analysis, we leveraged PlantAPAdb, a manually curated resource that offers an extensive catalog of APA sites in plants encompassing diverse biological samples [33]. The compiled dataset included more than 240,000 APA sites distributed across different genomic features from reference annotations in five plant species (Fig. 2a). As expected, the majority of polyadenilation sites are located at annotated 3' UTR regions across plant species. AgroNT demonstrated high accuracy in predicting APA sites across various plant species. It achieved area under the receiver operating characteristic (ROC) curve (AUC) values ranging from 0.89 to 0.96 and area under the precision-recall curve (AUPRC) values ranging from 0.82 to 0.93 (Fig. 2b). To the best of our knowledge, no previous study has used PlantAPAdb to develop a method for predicting APA sites. Hence, for a comparative analysis of AgroNT's APA prediction performance, we trained the model on an APA dataset specifically composed of 3'UTR sequences from *Arabidopsis thaliana*, which was used to train the CNN-based model DeepPolyA [34]. Remarkably, AgroNT achieved an almost perfect performance with an AUC of 0.99, surpassing DeepPolyA's reported AUC of 0.97, as well as outperforming other methods such as DanQ and DeepSEA, which obtained AUC values below 0.97 on the same dataset [34]. We hypothesize that the presence of exclusively introns, CDS, and 5' UTR sequences in the negative training dataset of DeepPolyA, without any nearby or non-APA 3' UTR signals, contributed to the reduced complexity of the classification task. This could explain the close-to-perfect performance obtained by AgroNT.

Next, we evaluated the performance of AgroNT in predicting splice sites, which are crucial contributors to transcriptome diversification in both animals and plants [35, 36]. For this assessment, we used a dataset encompassing acceptor and donor sites, as well as non-acceptor and non-donor sequences across the entire genome of

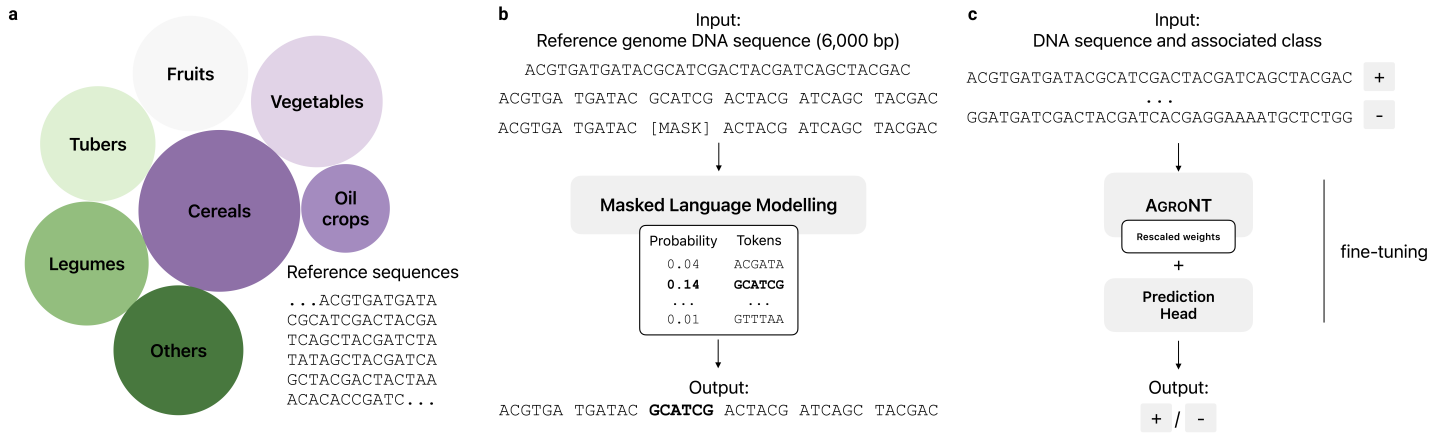


Figure 1: **Overview of Agronomic Nucleotide Transformer. a)** Data collection of plant species' reference genomes used for pre-training. The size of the circles is proportional to the number of species in each category. **b)** Pre-training strategy. The input is a 6,000 base pair (bp) DNA sequence randomly selected from 48 plant species' reference genomes. Each sequence is tokenized into a 6-mer (i.e., a string of 6 nucleotides), and 15% of these tokens are randomly masked. The objective of masked language modeling (MLM) is to obtain the model parameters that best predict the masked tokens. **c)** Fine-tuning strategy. After pre-training, we fine-tuned AgroNT using a supervised learning approach. We employed the parameter-efficient fine-tuning method, IA3, which introduces re-scaling weight vectors at the model's activations, enabling significant performance improvements while reducing computational resources compared to full model fine-tuning.

Arabidopsis thaliana (Fig. 2c). Notably, AgroNT demonstrated high performance in predicting both acceptor and donor sites, achieving AUC values of 0.97 and 0.95, respectively. The AUPRC for acceptor and donor site prediction was 0.97 and 0.98, respectively (Fig. 2d).

We also evaluated the classification performance of AgroNT in distinguishing long non-coding RNAs (lncRNAs) from messenger RNAs (mRNAs). LncRNAs in plant species have been recognized for their significant roles in growth, development, responses to biotic and abiotic stresses, and the regulation of cell differentiation [37]. To assess the classification capability of AgroNT, we used a comprehensive repository of annotated lncRNAs in plants and applied it to classify lncRNAs and annotated mRNAs across six plant species. The length of the two sequence types differed significantly, with lncRNAs having a median length of 457 bp, while mRNAs had a median sequence length of 2604 bp (Wilcoxon rank sum test, P-value $< 2.2 \times 10^{-16}$) (Fig. 2e). To mitigate the effect of the model relying solely on sequence length and encourage it to learn more meaningful features, we opted to retain only mRNA sequences that matched the length and GC content of lncRNAs, resulting in a more challenging and matched dataset (see Methods). AgroNT demonstrated high classification performance across the six plant species tested, achieving AUC values ranging from 0.79 to 0.89 and AUPRC values ranging from 0.80 to 0.87 (Fig. 2f). We found only one study that used the GreeNC database to develop a LSTM-based model for predicting plant lncRNAs, namely lncRNA-LSTM [38]. Therefore, we trained AgroNT on the dataset compiled in that study, which differed from our dataset by including only mRNAs from

Task	No. datasets	No. classes/Regression	Sequence length (bp)	Source
Polyadenilation	6	2	400	This study
Splicing site	2	2	398	Baten et al.
LncRNA	6	2	101-6000	This study
Promoter strength	2	Regression	170	Jores et al.
Chromatin accessibility	7	9-19	1000	Zhao et al.
Gene expression	6	Regression	6000	This study
Enhancer region	1	2	1000	This study

Table 1: **Summary of the Plants Genomic Benchmark (PGB).** The table details the number of datasets, supervised learning task, sequence length, and source.

Zea mays and not matching by length or sequence content. AgroNT exhibited almost perfect performance with an AUC of 0.991, which was comparable to the reported performance of lncRNA-LSTM (0.9934). We hypothesize that the much higher performance obtained in this classification setting by AgroNT is primarily due to the less challenging dataset, which was not matched by length nor GC content.

Promoter activity and genome accessibility prediction

We also assessed the predictive performance of AgroNT in determining the regulatory activity of core promoters, which play a crucial role in governing transcription initiation [39]. We used a recently developed dataset that extensively examined the core promoters of three plant species. These promoters were analyzed through STARR-seq assays, which were individually conducted in transiently transformed tobacco leaves and maize protoplasts [40]. Consistent with this study, we observed higher activity in the tobacco system compared to the maize protoplasts (Fig.3a). AgroNT demonstrated the ability to predict the strength of promoters in held-out sequences with high performance, achieving an R^2 value ranging from 0.62 to 0.71 in the maize system and an R^2 value ranging from 0.62 to 0.75 in the tobacco model (Fig.3b). Notably, AgroNT outperformed a CNN-based model [40] trained on this dataset across all plant species and systems, on average, by 3 R^2 points, showcasing its capability to learn this type of regulatory activity.

As an additional example of regulatory sequence prediction, we also evaluated the performance of AgroNT in predicting genome-wide chromatin profiles. By analyzing chromatin profiles, it is possible to identify key regulatory elements, such as enhancers and promoters, where many putative causal variants in humans and plants are often found [41, 42, 43]. We leveraged a recent collection of ATAC-seq data from multiple tissues of six model and important crop species [11]. The dataset included an average of 81,000 open chromatin profiles across all the tissues for the six plant species, with the profiles distributed among various tissues showcasing the sparsity of these annotations across the genome. (Fig. 3c). AgroNT demonstrated strong performance in predicting chromatin profile peaks across different plant species and tissues, achieving AUC values ranging from 0.94 to 0.98 (Fig. 3d) and AUPRC values ranging from 0.51 to 0.67 (Supplementary Fig. 3). We note that these performance metrics highlight the predictive capability of AgroNT in identifying these regulatory sequences, considering the low proportion of positive sequences in the datasets across plant species (average 0.12). Further, AgroNT either outperformed or achieved comparable performance to a CNN-based based model, plantDeepSEA [11], in five out of the six tested plant species when comparing the performance across all tissues within each plant. On average AgroNT exhibited a marginally higher performance, i.e. less than 1 AUC or AUPRC point variation. It is worth noting that these predictions were made using a limited sequence length of 1000 bp for predicting chromatin profiles, ensuring a fairer comparison against the CNN-based model. We anticipate that by extending the sequence context length to 6000 bp length, the capability of AgroNT, the model's performance is likely to further improve [44, 45].

Tissue-specific gene expression prediction

To assess the performance of AgroNT in predicting the levels of gene expression across tissues, we collected and processed data from gene expression atlases of five plant species. Our compiled dataset included tissue-specific gene expression values that ranged from 7 tissues in rice and up to 56 tissues in *Arabidopsis thaliana* (see Methods). To predict gene expression, we trained AgroNT on promoter-proximal sequences based on transcription starting sites (TSS) annotations and ensured that the sequence overlapped slightly with the coding sequence, as this has been shown to aid in gene expression prediction accuracy across different species, including plants [46]. AgroNT obtains moderate prediction performance of gene expression levels for genes across the five plants species and across all tissues ($R^2 = 0.419 - 0.621$, P-value $< 2.2 \times 10^{-16}$) (Fig. 4a). The prediction performance was relatively consistent across individual tissues with the exception of pollen which had consistently lower prediction performance across plant species (Supplementary Tables 4-8, Supplementary Fig. 4-8). This might be related to the inaccessible nature of pollen developing in the anther and the resistant pollen wall that might affect proper pollen isolation, as recently suggested [38]. Considering the absence of models capable of predicting quantitative gene expression levels, and to conduct a comparative analysis for this task, we also trained AgroNT to predict whether genes are expressed or not. To accomplish this, we employed a dataset derived from a collection

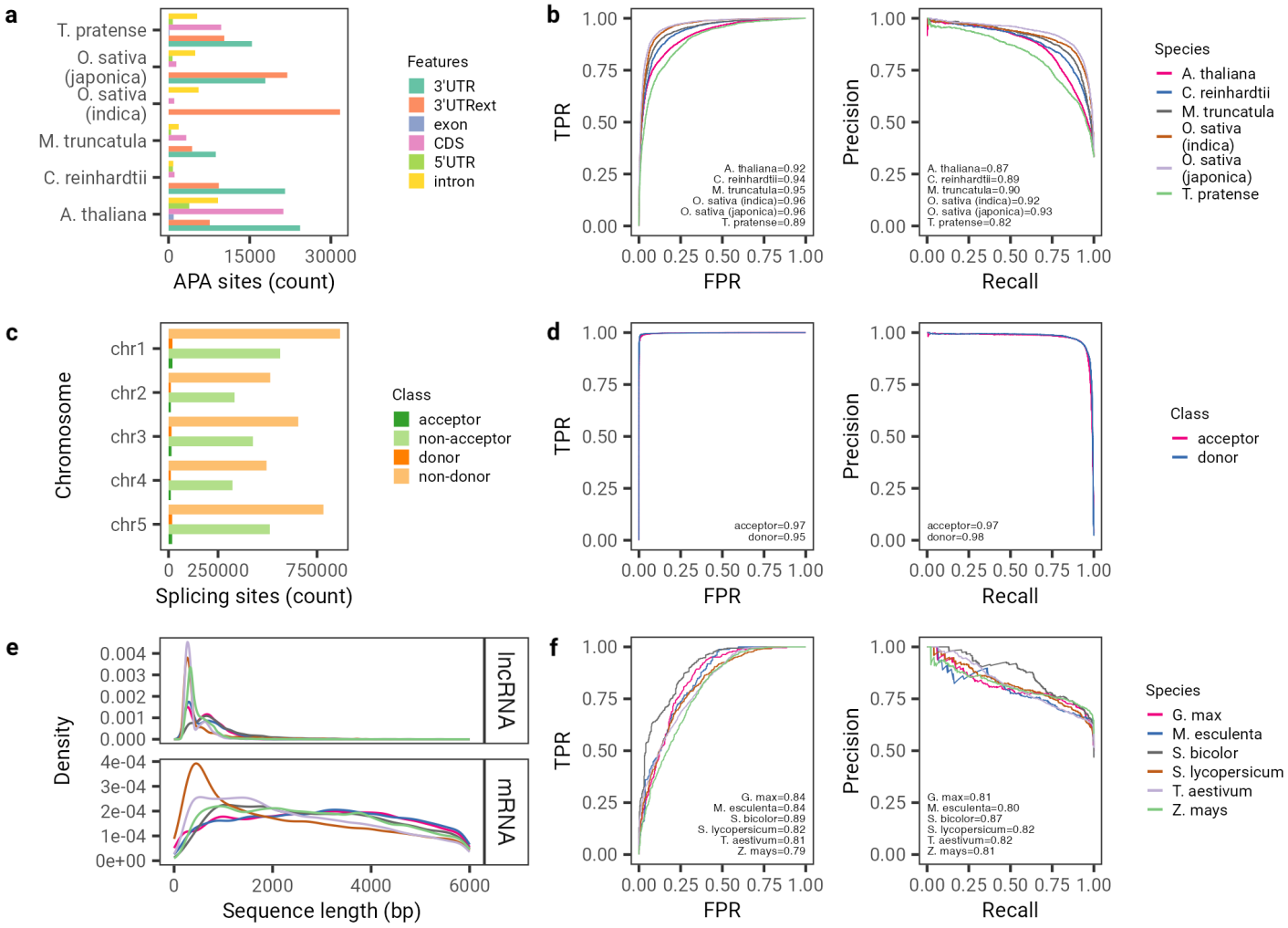


Figure 2: **AgroNT predicts polyadenilation, splicing, and long non-coding RNAs. a)** Distribution of alternative polyadenilation (APA) sites across different genomic features in 5 species as annoated in the PlantAPAdb. 3′UTRext refers to the 3′UTR extended annotation. See [33] for further details on the annotation of APA sites. **b)** Receiver operating characteristic curve (left) and precision-recall curve (right) for APA prediction. **c)** Distribution of splicing sites (acceptor or donors) and non-splicing sites across the genome in *Arabidopsis thaliana*. **d)** Same as in **b)**, but for splicing site prediction. **e)** Distribution of the lenght of long non-coding RNAs and mRNA sequences across 6 species from the GreeNC database. **f)** Same as in **b)**, but for long non-coding RNA prediction.

of *Zea mays* tissues, which was previously utilized by a recent study to train a CNN-based model [47]. Notably, AgroNT achieved superior performance compared to the model trained on promoter sequences (AUC 0.89 versus 0.81) as well as terminator sequences (0.89 versus 0.82).

To further explore the capabilities of AgroNT for gene expression prediction, we also evaluated the performance of AgroNT in reproducing tissue-specific expression patterns. To do this, we conducted hierarchical clustering using the distances between observed gene expression levels across tissues. We then assessed how well the predicted distances replicated these patterns (Fig. 4b,c). See (Supplementary Fig. 4-8) for gene expression heat maps with tissue labels. Predictions were made solely on promoter-proximal regions as carried out in the fine-tuning. Encouragingly, AgroNT successfully reproduced the general patterns of similarity across tissues. Nonetheless, we observed that the predicted distances were significantly shorter than the observed distances, indicating that, on average, the predicted gene expression levels were more similar across tissues than the observed. To further evaluate the tissue-specific performance, we also computed log-fold changes (LFC) for all pairs of tissues and assessed the correlation between the observed and predicted LFC values for each species separately (Fig. 4d). The correlation between the observed

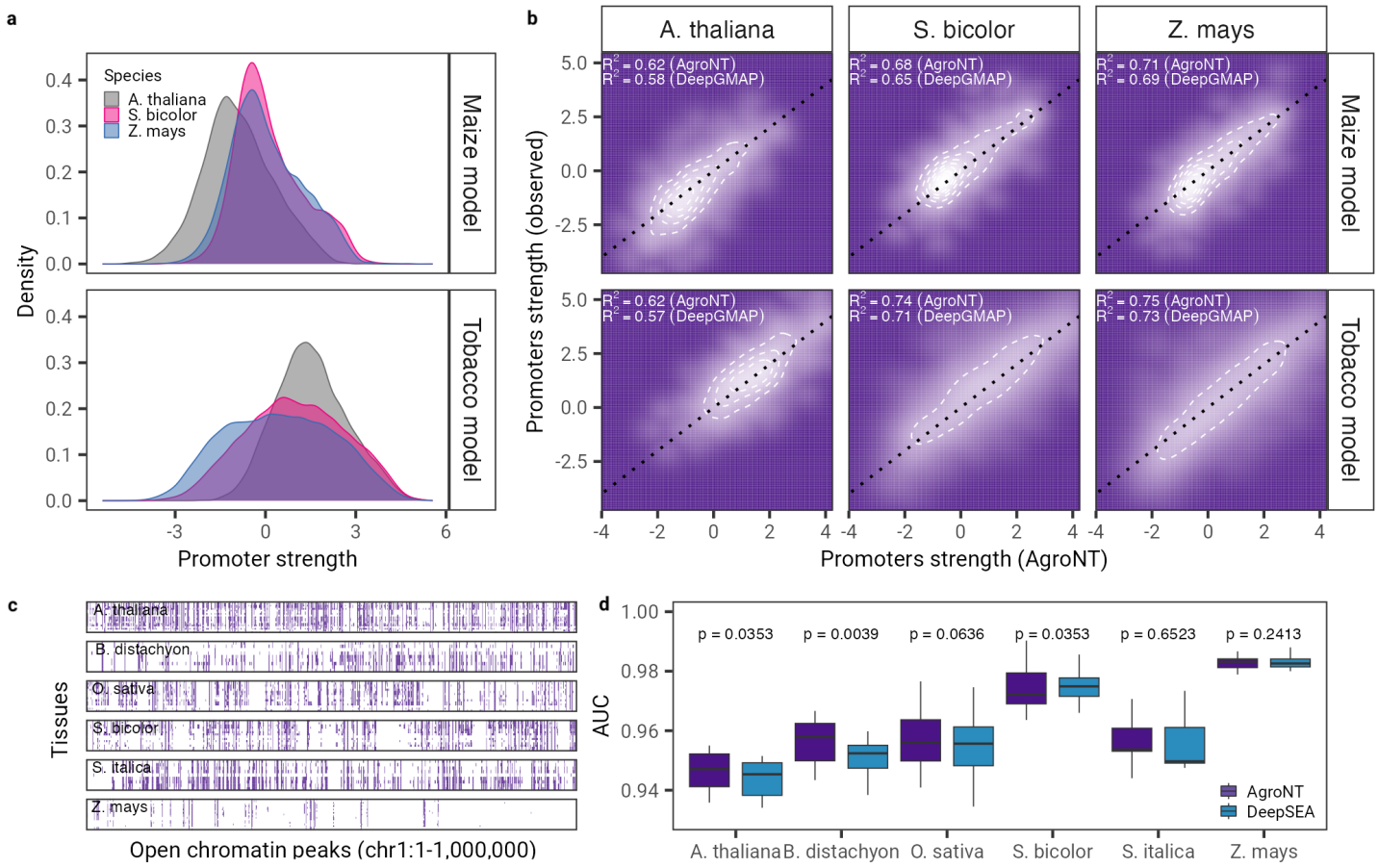


Figure 3: **AgroNT predicts promoter activity and genome accessibility across tissues.** **a)** Promoter strength distribution based on STARR-seq assays conducted in transiently transformed maize protoplasts and tobacco leaves. **b)** AgroNT prediction of promoter strength. The coefficient of determination (R^2) from AgroNT and from a CNN model based on the DeepGMAP architecture are shown. **c)** Open chromatin profiles across tissues. The heatmap illustrates the genomic position of chromatin profiles from six plant species across tissues in a 1-megabase window of chromosome 1. Each cell in the heatmap represents a 1000-base pair (bp) region. The training strategy of AgroNT involved detecting chromatin profiles in the center of a 1000 bp region. **d)** The performance of AgroNT in predicting chromatin profiles across species and tissues is measured using the area under the curve (AUC) of the receiver operating characteristic curve. P-values shown are based on a two-sided Wilcoxon signed test comparing the performance of AgroNT with a model based on the DeepSEA architecture.

and predicted LFC values was relatively low, with similar correlation values observed across species (ranging from 0.18 to 0.24). Similarly, when examining the predicted LFC directions, which indicate the accuracy of the model in predicting whether genes are upregulated or downregulated between tissues, we found that slightly over half of these predictions were correct (with a median ranging from 0.54 to 0.58).

It is likely that the model has primarily learned to recognize essential features associated with basal transcriptional activity. Indeed, when evaluating the importance of specific regions within the prediction sequences, we found that sequences positioned approximately 100 base pairs from the annotated TSS had the most significant influence on the predictions (Fig. 4e). This observation is consistent with the understanding that these sequences serve primarily as binding sites for transcription factors that are crucial for basal transcription levels, independent of tissue or cell types. This result also illustrates the challenge of accurately predicting tissue specificity based on promoter-proximal sequences only. The model's inability to capture tissue-specific regulatory regions, such as enhancers that can be located far away from the promoter region and exhibit more variable and cell type-specific activity, contributes to this difficulty.

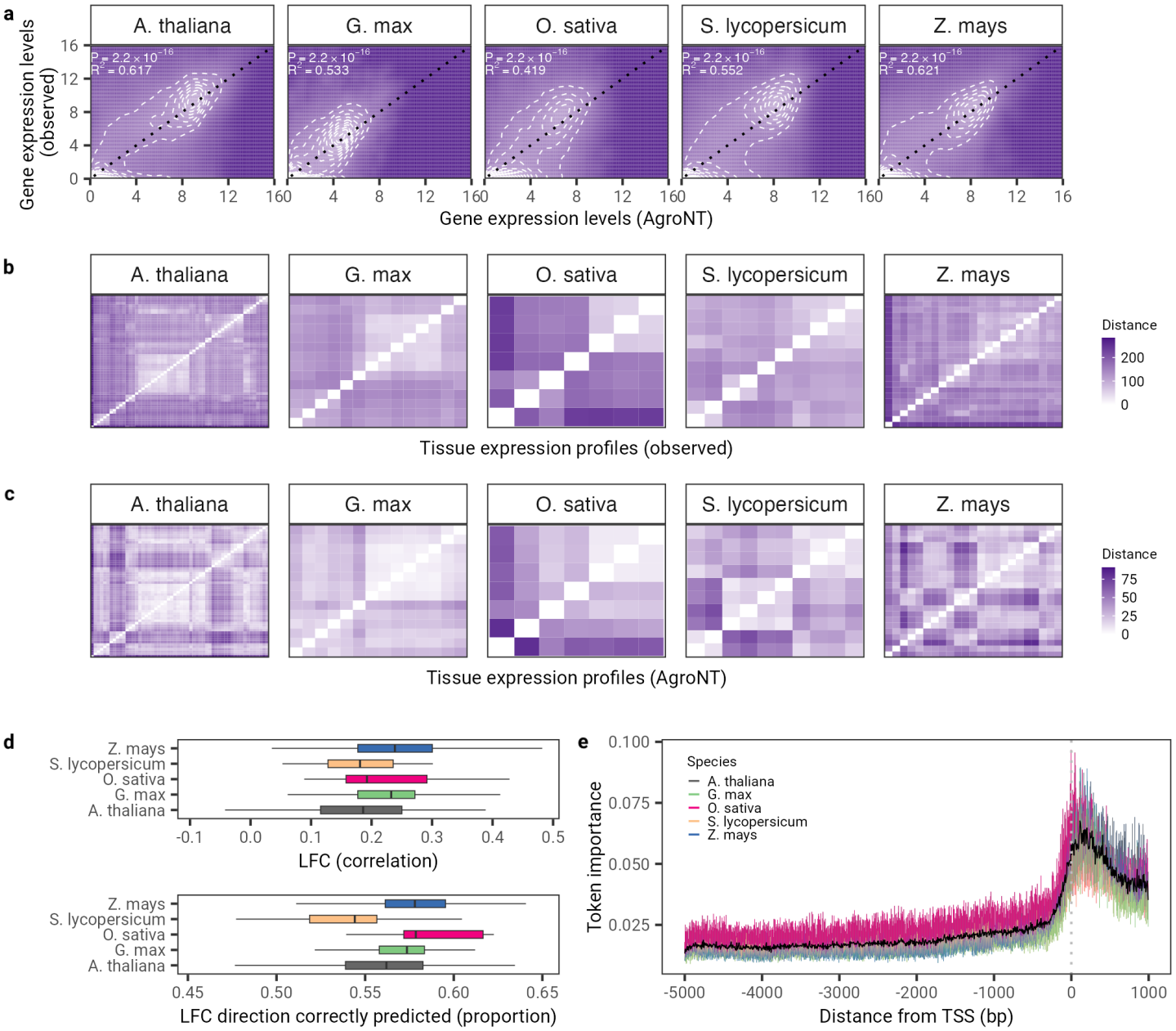


Figure 4: **AgroNT provides gene expression prediction across different plant species. a)** Gene expression prediction on holdout genes across all tissues are correlated with observed gene expression levels. The coefficient of determination (R^2) from a linear model and associated P-values between predicted and observed values are shown. **b/c)** Tissue expression profiles between observed (upper panels) and predicted gene expression levels (lower panels) on holdout genes. The order of the cells in the heatmap is based on a hierarchical clustering using the euclidean distance of the observed gene expression levels across tissues. **d)** Correlation between observed and predicted log fold changes (LFC) for all pairwise tissue comparisons (upper plot) and LFC directions, i.e. accuracy of the model in predicting upregulated or downregulated genes between tissues. **e)** Token importance analysis across promoter-proximal sequences. Token importance is computed as the absolute difference between the original model prediction and the predicted gene expression level after inserting a random token across the promoter-proximal sequence. Token importance is averaged across tissue for each species. Black line denotes the mean value across species. Vertical dotted line indicated the transcription starting site (TSS) position.

Zero-shot prediction of functional variants

AgroNT can also be leveraged for zero-shot learning, a transfer learning approach that enables the model to recognize and classify samples from new classes that were not encountered during training. Hence, we evaluated the ability of AgroNT to assess the functional consequences of

different types of genetic variants through various zero-shot scores. Specifically, we computed zero-shot scores by considering different aspects of vector distances in the embedding space, along with a score derived from the loss function. To do this, we calculated the log-likelihood ratio (LLR) between the probabilities of alternate and reference alleles (see Methods). As a baseline for comparison, we also included a LLR score called the Genomic Pre-trained Network (GPN) score [29]. Like our approach, this score is based on a multi-species DNA language model pre-trained on reference genomes of *Arabidopsis thaliana* and seven related species within the Brassicales order. Considering that the functional variants analysed here are determined exclusively from annotations obtained from *Arabidopsis thaliana*, we regard this comparison as challenging, given that the GPN model was solely trained on species related to Brassicales, where its performance is anticipated to be high. In addition to these, we also included two conservation scores, phastCons [48] and phyloP [49], in the comparison. These scores assess the evolutionary constraint on specific genomic positions and provide valuable information regarding potential functional implications.

We first investigated the utility of zero-shot scores in evaluating deleteriousness. The presence of aggregated genomes from multiple *Arabidopsis thaliana* samples as part of the 1001 Arabidopsis Genome Project provided an opportunity to assess the impact of natural selection on missense and synonymous mutations across a range of allele frequencies. When comparing the ratio of missense to synonymous variants, we observed that very rare variants (frequency lower than 0.1%) and rare variants (frequency between 0.1% and 1%) exhibited a significantly higher missense/synonymous ratio of 1.8 (95% CI [1.78-1.82]) and 1.64 (95% CI [1.63-1.66]), respectively, compared to the expected ratio of 1 for non-deleterious variants (Fig. 5a). As allele frequency increased, the missense to synonymous ratio decreased consistently, suggesting that a considerable proportion of missense variants in *Arabidopsis thaliana* with frequencies lower than 1% may have deleterious effects. Based on this observation, we examined the enrichment of missense variants in these two frequency categories compared to common variants (frequency >5%) using odds ratios (OR) across different quantiles of the score distributions (Fig. 5b). Across quantiles, GPN and the two conservation scores showed the highest ORs for both frequency classes and across quantiles. LLR and the dot product of embeddings (based on layers 11 and 40) from AgroNT also exhibited significant ORs, particularly at the lowest quantiles, although they were approximately two times lower than those observed for GPN and conservation scores. Other zero-shot scores based on the embedding space and other layers did not exhibit superior performance either (Supplementary Fig. 1). Considering that GPN is based on pre-trained sequences from Brassicales-related species, it is likely that the model leverages this variation to better capture the frequency of variants in *Arabidopsis thaliana*, thereby aiding in the detection of these rare missense variants. Indeed, across all missense variants, GPN scores displayed a stronger correlation with allele frequency ($r = 0.1$, P-value $< 2.2 \times 10^{-16}$) compared to LLR from AgroNT ($r = 0.05$, P-value $< 2.2 \times 10^{-16}$), which was trained on multiple plant species.

We next evaluated the performance of these scores to detect genome-wide associated variants (GWAS) based on the AraGWAS Catalog [50], a manually curated database for standardised GWAS results for *Arabidopsis thaliana*. As in our previous analysis we examined the enrichment of coding and non-coding GWAS variants relative to non-GWAS variants. Across quantiles, GPN and LLR exhibited comparable performance for both coding and non-coding GWAS variants (Fig. 5b). The two conservation scores performed similarly to GPN and LLR for coding variants, as expected, given the higher selective pressure typically exerted within these regions, but showed lower performance for non-coding variants. Interestingly, the dot product score demonstrated the highest OR for both coding and non-coding GWAS variants, and this score exhibited consistent performance when calculated from different layers (Supplementary Fig. 2). This observation suggests that a measure of embedding similarity might be more effective in capturing the effects of GWAS variants, which are typically common in the population. Unlike loss-based scores, the dot product score is not primarily influenced by the model's training objective, which focuses on the disparity between predicted and true labels. Instead, we hypothesize that it can better capture the relevance of the variants themselves, potentially taking into account other intrinsic characteristics within the genomic context. Interestingly, we also observed that the zero-shot scores derived from AgroNT exhibited a very low correlation with the conservation scores, indicating that they capture genomic information beyond just the degree of conservation (Fig. 5c). This observation suggests the possibility of integrating these scores to create a more robust annota-

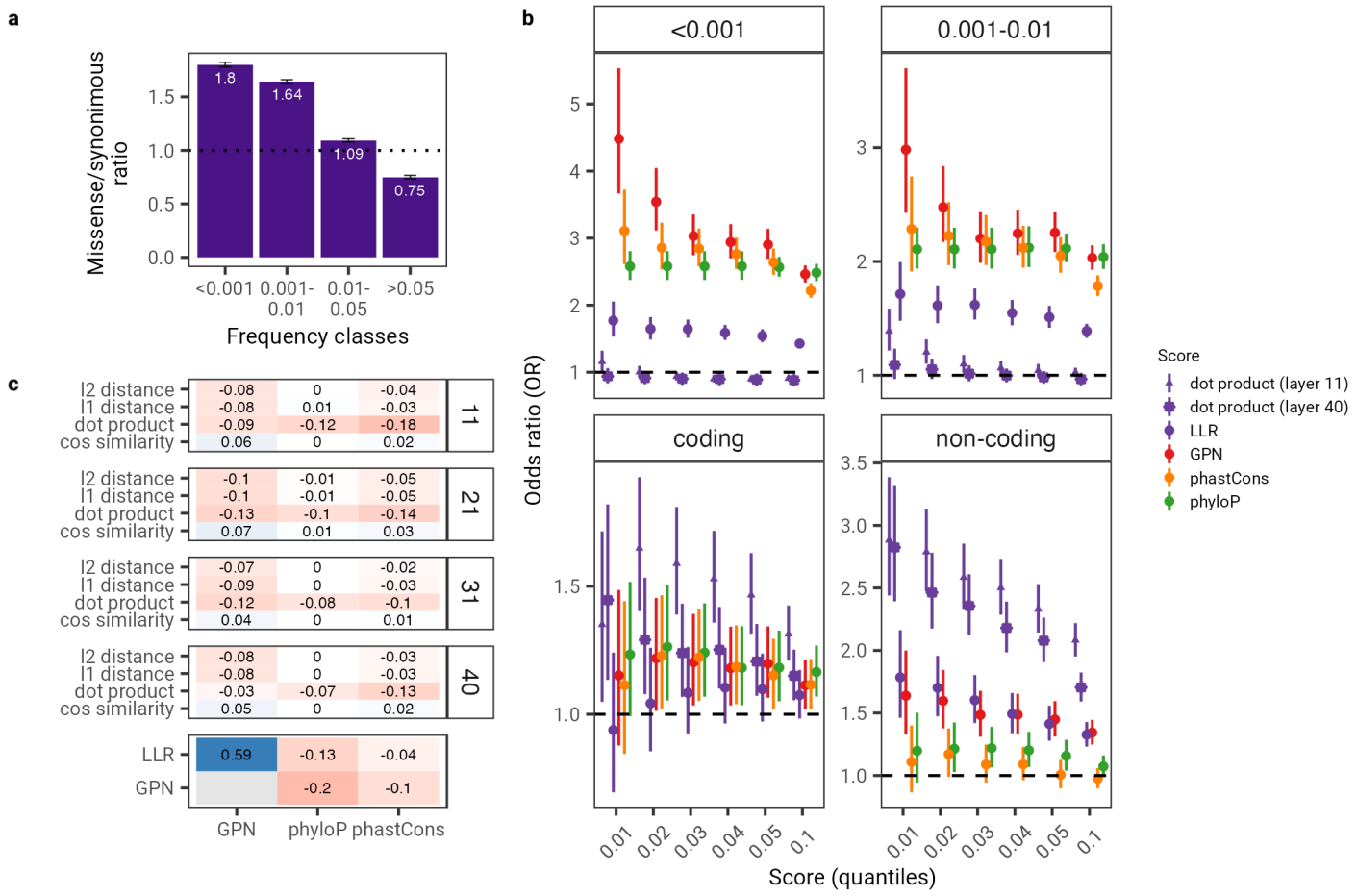


Figure 5: **Zero-shot based prediction for variants of different functional classes in *Arabidopsis thaliana*. a)** Ratio of missense (i.e. non-synonymous) to synonymous variants in the 10001 Arabidopsis thalaiana Genome Project across frequency classes. Error bars denote 95% confidence intervals based on 100 chromosome-based samplings with replacement. **b)** Odd ratios (OR) across different score thresholds used to assess enrichment of different functional variants. The ORs of the upper panel is based on assessing the enrichment of two frequency classes of missense variants. The bottom panel is based on coding and non-coding GWAS variants. Error bars denote 95% confidence intervals. Dashed black line indicates an OR of 1. **c)** Correlations between AgroNT zero-shot based scores and conservation (phyloP and phastCons) and GPN scores. Facets indicate the layer used to compute the zero-shot scores. LLR refers to the log-likelihood ratio between the alternate and reference allele probabilities from AgroNT.

tor capable of predicting the potential impact of genetic variations in plant genomes.

Profiling the regulatory potential via *in silico* mutagenesis

Lastly, we leveraged the capability of AgroNT to obtain accurate predictions across a variety of regulatory features by further training AgroNT to predict intergenic enhancer elements and gene expression levels across 11 tissue/organ types in the *Manihot esculenta* (cassava) genome, and performed *in silico* saturation mutagensisis analyses (see Methods). AgroNT demonstrated strong performance in detecting intergenic enhancer sequences, achieving an AUC value of 0.88 (Fig. 6a). Through a large-scale *in silico* mutagenesis analysis, we systematically mutated each site to all three possible distinct nucleotides across the central 400 bp region of the top 1000 most confidently predicted enhancer regions by AgroNT (probabilities ranging 0.975 to 0.999). As a result, mutating the nucleotides within these regions is more likely to produce an effect that AgroNT can accurately identify. By estimating the log-fold change (LFC) between the probability of the mutated sequence and that of the original sequence, we note, as expected, that the majority of mutations do not strongly influence the predicted probability of an enhancer sequence (median absolute LFC probability $= -1.10 \times 10^{-3}$) (Fig. 6b), since the majority of the evaluated

mutations are likely not impacting relevant enhancer functionality. To further assess the impact of the evaluated mutations we identified potentially strongly-affecting mutations as those with a LFC probability less than −0.042 (0.01% quantile), implying a disruption of enhancer elements. This resulted in 120 mutations which impacted 15 distinct enhancer sequences. Notably, the most potent mutation overlapped the TCGAAT and DOF motifs (Fig. 6c), while the second most influential mutation targeted the DOF motif (Fig. 6d). This observation suggests that the AgroNT-trained model has likely acquired an understanding of the significance of these motifs in relation to enhacer regions. To assess the impact of mutations on promoter-proximal regions, we also trained AgroNT to predict gene expression levels across 11 tissue/organ types in cassava, achieving moderate predictive performance ($R^2 = 0.51$, P-value $< 2.2 \times 10^{-16}$ across all tissues) (Fig.6e and Supplementary Fig.9). Similar to the enhancer-based mutagenesis analysis, we then calculated the LFC between the predicted gene expression levels of the mutated sequence and the original sequence for the 1000 promoter sequences derived from genes with the highest average gene expression levels across all tissues (Fig. 6f). Focusing on mutations with the most substantial impact, specifically those with LFC less than −0.28 (0.01% quantile), we identified 136 mutations across 20 distinct promoter sequences. The most significant mutation affected the Trihelix motif (Fig. 6g), while the second most significant mutation impacted the DOF motif. These findings suggest that the AgroNT-trained model has also developed an understanding of the importance of these motifs in relation to gene expression levels. We have provided the predicted impact of all possible single-nucleotide substitutions across all enhancer and promoter sequences for future research purposes.

Discussion

In this paper, we showcase the capabilities of AgroNT, a novel DNA-based LLM for accurately making predictions on a wide range of genomic tasks across various plant species. Due to its pre-training on a large collection of plant genomes, AgroNT inherently possesses transfer learning capabilities, allowing it to aggregate knowledge from all assembled plant genomes, and enhancing its predictive power. In the context of regulatory-based prediction tasks, it is reasonable to hypothesize that AgroNT has learned to capture genomic regulatory features that likely hold functional importance across plant species, leading to improved performance. Recent studies have indicated that many key promoter elements identified in animals are also enriched in plant promoters [51, 52, 53, 54], suggesting that leveraging more phylogenetically-distinct species could potentially offer further performance improvements. AgroNT differs from the majority of previously developed models that rely on training with specific genomic sequences from individual plant species and use distinct deep learning architectures for each prediction task. In contrast, we exclusively employed a parameter-efficient fine-tuning approach for all predictive tasks. This flexible and effective strategy either matched or outperformed current state-of-the-art methods in the assessed tasks. It simplifies model selection, ensuring reliable high performance without the necessity for extensive architecture analysis for each prediction task.

Several paths to further improve model performance appear promising, particularly in two areas with a direct impact on breeding engineering. The first of these relates to the prediction of functional and deleterious genetic variants. Several studies on domesticated plant species have found that domestication is often associated with an increase in harmful deleterious variants [55, 56, 57, 58]. With low rates of sexual reproduction, these deleterious variants are maintained in the population and can hinder breeding efforts [59]. As such, confidently identifying these mutations could facilitate targeted decisions to purge genetic load from crop species and advance genetic gains [60]. Our analysis shows that AgroNT could be used to prioritize functional variants based solely on DNA sequences, which may be particularly relevant for understudied plant species, such as orphan crops, that currently lack functional genomic annotations. Interestingly, our analysis comparing zero-shot scores from AgroNT and the GPN model [29] in prioritizing deleterious missense variants in *Arabidopsis thaliana* suggests that leveraging the genetic variability from species of relatively short phylogenetic distance during the pre-training leads to better performance in detecting these variants. As a direct application, the recent study of Long et al. [61], which assembled 52 *Euphorbiaceae* genomes to compute evolutionary conservation scores and assess its use in detecting deleterious variants in cassava, could be compared against zero-shot scores derived from AgroNT after additional pre-training with these genomes. In that study, over half of all base pairs across orthologous genes had an align-

ment depth of >31 species, meaning that >50% of bases could not be assessed for their deleteriousness. In contrast, AgroNT can leverage multiple genomes without having to rely on whether they have been aligned or not, and as such, can compute, for example, a zero-shot score for any variant. In addition to zero-shot scores, we envision using supervised models as another potential approach to enhance the detection of functional and deleterious variants. Current studies predominantly relying on sequence conservation metrics to predict deleterious mutations in plant species have shown varying performance (AUC $\sim$ 0.5 to AUC > 0.9) [62, 63]. Nonetheless, these methods still trail behind those developed for humans, suggesting an opportunity for improvement. Given the low correlation that we observe between the zero-shot based scores from AgroNT and conservation metrics, it is feasible to assume that their addition as features into the currently developed models may lead to an improvement in performance. Lastly, another potential avenue for improvement involves harnessing AgroNT's transfer learning capabilities when dealing with a lack of putative deleterious variants in a specific plant species. For instance, the recently introduced model PrimateAI-3D [64] has demonstrated that systematically cataloging common variants in non-human primates can accurately infer the pathogenicity of missense variants in humans. Similarly, utilizing plant catalogs like the 1001 Arabidopsis Genome Project, one could develop a classifier to predict rare (or absent) variants, assuming their high enrichment in deleterious variants. The second area with direct impact on breeding engineering relates to regulatory genomics. Cis-regulatory elements do not work in isolation. In plant systems, it has been demonstrated that interactions between promoters and terminators affect gene expression, and the relative strength of a terminator depends on which promoter it is paired with [65]. While our gene expression prediction focused on promoter-proximal sequences, incorporating terminators could improve results, as observed in binary expression prediction in *Zea mays* [47]. An additional approach could also leverage distinct omics assays. Leveraging AgroNT, one could combine candidate regulatory elements based on chromatin accessibility assays with promoter sequences to enhance gene expression prediction. As demonstrated in our evaluated gene expression models, accurately predicting tissue-specific gene expression remains a challenge. Data based on gene regulation at single-cell resolution have the potential to unravel complex regulatory interactions and identify targets for genetic manipulation. In this vein, single-cell-based assays have recently been employed to study cell-type-dependent gene regulation in plants [66, 67, 68, 69]. These studies identified cell-type-specific changes in gene expression and chromatin accessibility, illustrating the potential use of single-cell assays in prediction models. Encouragingly, at least based on single-cell assays from human cells, LLMs trained on single-cell data have shown high performance across numerous downstream tasks [70, 71]. Nonetheless, given the current data sparsity of single-cell experiments and the difficulty in capturing sufficient cells of rarer types, at least in plants, leveraging this type of data in deep learning approaches will likely depend on novel methods that can significantly reduce the cost per cell. In addition to leveraging distinct assays, another avenue relates to expanding the sequence context to capture additional regulatory elements such as enhancers, silencers, and insulators [72, 73]. Models like Enformer [44] and the recent HyenaDNA [74], which can handle larger regions and longer context lengths, represents yet another promising approach to better predict tissue-specific gene expression. Despite the potential approaches for improvement, as of now, AgroNT stands as the most accurate plant gene expression predictor. Comparative studies achieved R^2 values of 0.27 [46] and 0.09-0.19 [30] using promoter-proximal sequences to predict gene expression levels in *Arabidopsis thaliana* and *Zea mays*, respectively. In comparison to these two models and based on our accuracy on both *Arabidopsis thaliana* and *Zea mays*, AgroNT demonstrates up to a three-fold improvement in performance.

To aid the field of plant genomics, we have made the pre-trained AgroNT model available for future research, as well as the datasets compiled here, which we term the Plants Genomic Benchmark (PGB). Furthermore, we have released the effect predictions for promoter sequences and enhancer regulatory regions in the still understudied crop cassava (*Manihot esculenta*). Altogether, continued progress in deep learning-based models in plant genomics will open exciting opportunities to explore the expanding catalogs of genetic variants linked to critical plant traits, ultimately enabling the successful engineering of high-yielding and stress-resistant crops, ensuring food security and agricultural sustainability in the face of changing climatic conditions.

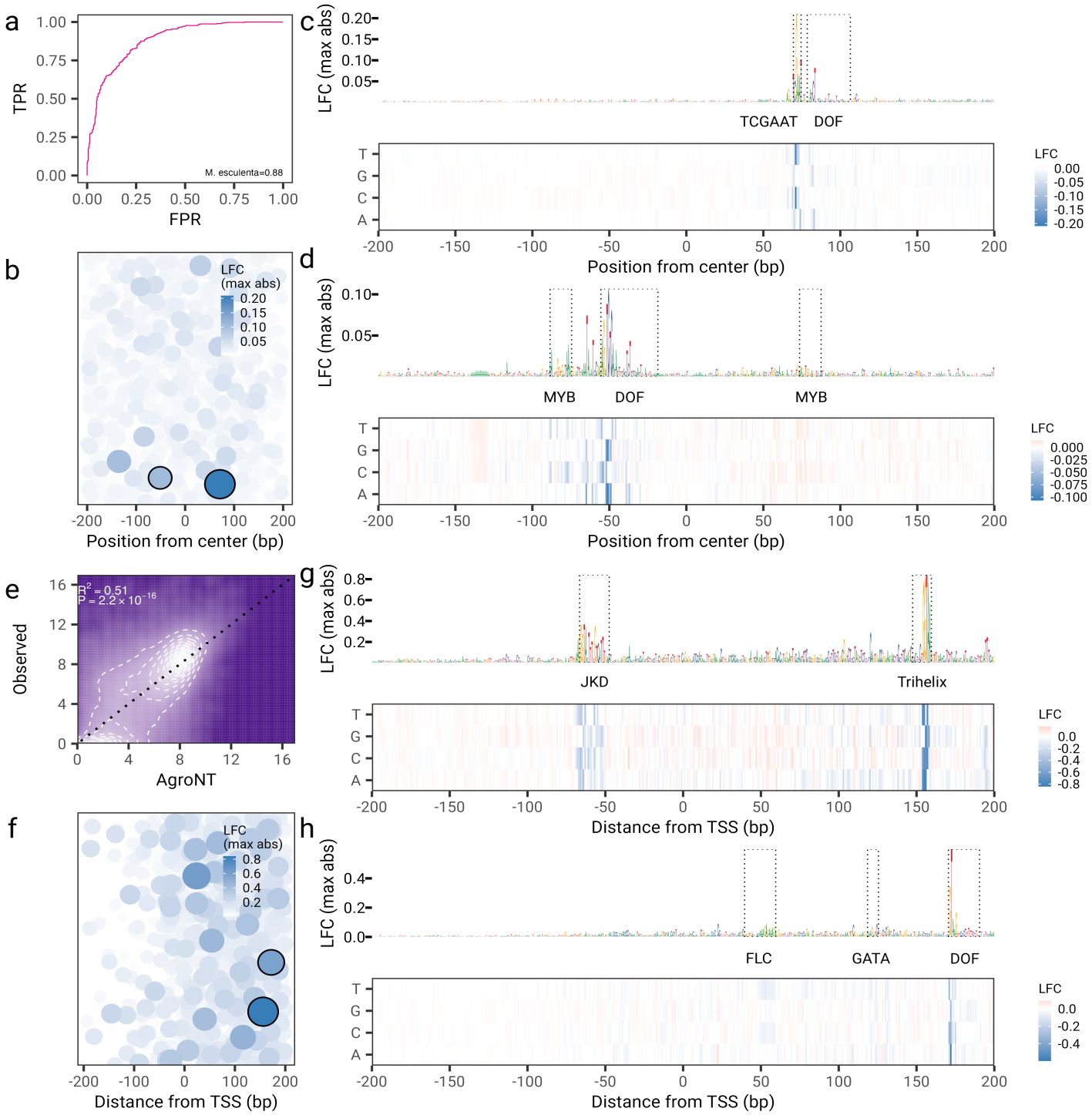


Figure 6: ***In silico* mutagenesis in the Cassava genome.** **a)** Receiver operating characteristic curve for the prediction of intergenic enhancer regions. **b)** The mutations with the strongest log-fold change (LFC) probabilities for each of the top 1000 enhancer regions are shown in relation to their distance from the central 400 bp region. The size of the points is proportional to the LFC value. Circled points represent the mutations with the strongest LFC values shown in **c)** and **d)**. **c)** The *in silico* mutagenesis map for the enhancer region with coordinates chr1:6334152-6335151. The bottom panel displays the LFC value for mutating each site to all three possible distinct nucleotides across the central 400 bp region. The upper panel features the DNA sequence logo, with the nucleotide heights representing the maximum absolute LFC value. Motif positions are indicated within a dotted rectangle, with the motif names displayed below. **d)** Same as in **c)**, but for the enhancer region with coordinates chr2:15863252-15864251. **e)** Gene expression prediction on holdout genes across all tissues. The coefficient of determination (R^2) from a linear model and associated P-values between predicted and observed values are shown. **f)** Same as in **b)**, but for LFC gene expression predictions across promoter sequences. **g)** Same as in **c)**, but for promoter sequence with coordinates chr12:27521008-27527007. **h)** Same as in **c)**, but for promoter sequence with coordinates chr15:862119-868118.

References

[1] A. G. I. genomeanalysis@ tgr. org genomeanalysis@ gsf. de, "Analysis of the genome sequence of the flowering plant arabidopsis thaliana," *nature*, vol. 408, no. 6814, pp. 796–815, 2000.

[2] Y. Sun, L. Shang, Q.-H. Zhu, L. Fan, and L. Guo, "Twenty years of plant genome sequencing: achievements and challenges," *Trends in Plant Science*, 2022.

[3] H. Yaqoob, A. Tariq, B. A. Bhat, K. A. Bhat, I. B. Nehvi, A. Raza, I. Djalovic, P. V. Prasad, and R. A. Mir, "Integrating genomics and genome editing for orphan crop improvement: a bridge between orphan crops and modern agriculture system," *GM Crops & Food*, vol. 14, no. 1, pp. 1–20, 2023.

[4] R. Jamnadass, R. H. Mumm, I. Hale, P. Hendre, A. Muchugi, I. K. Dawson, W. Powell, L. Graudal, H. Yana-Shapiro, A. J. Simons, *et al.*, "Enhancing african orphan crops with genomics," *Nature Genetics*, vol. 52, no. 4, pp. 356–360, 2020.

[5] A. Bohra, U. C. Jha, P. K. Kishor, S. Pandey, and N. P. Singh, "Genomics and molecular breeding in lesser explored pulse crops: current trends and future opportunities," *Biotechnology advances*, vol. 32, no. 8, pp. 1410–1428, 2014.

[6] R. K. Varshney, T. J. Close, N. K. Singh, D. A. Hoisington, and D. R. Cook, "Orphan legume crops enter the genomics era!," *Current opinion in plant biology*, vol. 12, no. 2, pp. 202–210, 2009.

[7] D. Rodríguez-Leal, Z. H. Lemmon, J. Man, M. E. Bartlett, and Z. B. Lippman, "Engineering quantitative trait variation for crop improvement by genome editing," *Cell*, vol. 171, no. 2, pp. 470–480, 2017.

[8] J. Zhou, C. L. Theesfeld, K. Yao, K. M. Chen, A. K. Wong, and O. G. Troyanskaya, "Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk," *Nature genetics*, vol. 50, no. 8, pp. 1171–1179, 2018.

[9] J. Zhou, C. Y. Park, C. L. Theesfeld, A. K. Wong, Y. Yuan, C. Scheckel, J. J. Fak, J. Funk, K. Yao, Y. Tajima, *et al.*, "Whole-genome deep-learning analysis identifies contribution of noncoding mutations to autism risk," *Nature genetics*, vol. 51, no. 6, pp. 973–980, 2019.

[10] D. Quang and X. Xie, "Danq: a hybrid convolutional and recurrent deep neural network for quantifying the function of dna sequences," *Nucleic acids research*, vol. 44, no. 11, pp. e107–e107, 2016.

[11] H. Zhao, Z. Tu, Y. Liu, Z. Zong, J. Li, H. Liu, F. Xiong, J. Zhan, X. Hu, and W. Xie, "Plantdeepsea, a deep learning-based web service to predict the regulatory effects of genomic variants in plants," *Nucleic Acids Research*, vol. 49, no. W1, pp. W523–W529, 2021.

[12] N. Scalzitti, A. Kress, R. Orhand, T. Weber, L. Moulinier, A. Jeannin-Girardon, P. Collet, O. Poch, and J. D. Thompson, "Spliceator: Multi-species splice site prediction using convolutional neural networks," *BMC bioinformatics*, vol. 22, no. 1, pp. 1–26, 2021.

[13] B. P. de Almeida, F. Reiter, M. Pagani, and A. Stark, "Deepstarr predicts enhancer activity from dna sequence and enables the de novo design of synthetic enhancers," *Nature Genetics*, vol. 54, no. 5, pp. 613–624, 2022.

[14] D. R. Kelley, Y. A. Reshef, M. Bileschi, D. Belanger, C. Y. McLean, and J. Snoek, "Sequential regulatory activity prediction across chromosomes with convolutional neural networks," *Genome research*, vol. 28, no. 5, pp. 739–750, 2018.

[15] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[16] S. Karita, N. Chen, T. Hayashi, T. Hori, H. Inaguma, Z. Jiang, M. Someki, N. E. Y. Soplin, R. Yamamoto, X. Wang, *et al.*, "A comparative study on transformer vs rnn in speech applications," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 449–456, IEEE, 2019.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[18] M. Zaheer, G. Guruganesh, K. A. Dubey, J. Ainslie, C. Alberti, S. Ontanon, P. Pham,

A. Ravula, Q. Wang, L. Yang, *et al.*, "Big bird: Transformers for longer sequences," *Advances in neural information processing systems*, vol. 33, pp. 17283–17297, 2020.

[19] S. Mo, X. Fu, C. Hong, Y. Chen, Y. Zheng, X. Tang, Y. Lan, Z. Shen, and E. Xing, "Multi-modal self-supervised pre-training for large-scale genome data," in *NeurIPS 2021 AI for Science Workshop*, 2021.

[20] M. Yang, L. Huang, H. Huang, H. Tang, N. Zhang, H. Yang, J. Wu, and F. Mu, "Integrating convolution and self-attention improves language model of human genome for interpreting non-coding regions at base-resolution," *Nucleic acids research*, vol. 50, no. 14, pp. e81–e81, 2022.

[21] A. Hoarfrost, A. Aptekmann, G. Farfañuk, and Y. Bromberg, "Deep learning of a bacterial and archaeal universal language of life enables transfer learning and illuminates microbial dark matter," *Nature communications*, vol. 13, no. 1, p. 2606, 2022.

[22] H.-J. Gwak and M. Rho, "Vibe: a hierarchical bert model to identify eukaryotic viruses using metagenome sequencing data," *Briefings in Bioinformatics*, vol. 23, no. 4, p. bbac204, 2022.

[23] Z. Bai, Y.-z. Zhang, S. Miyano, R. Yamaguchi, K. Fujimoto, S. Uematsu, and S. Imoto, "Identification of bacteriophage genome sequences with representation learning," *Bioinformatics*, vol. 38, no. 18, pp. 4264–4270, 2022.

[24] H. Dalla-Torre, L. Gonzalez, J. Mendoza-Revilla, N. L. Carranza, A. H. Grzywaczewski, F. Oteri, C. Dallago, E. Trop, H. Sirelkhatim, G. Richard, *et al.*, "The nucleotide transformer: Building and evaluating robust foundation models for human genomics," *bioRxiv*, pp. 2023–01, 2023.

[25] E. C. Alley, G. Khimulya, S. Biswas, M. AlQuraishi, and G. M. Church, "Unified rational protein engineering with sequence-based deep representation learning," *Nature methods*, vol. 16, no. 12, pp. 1315–1322, 2019.

[26] R. Rao, N. Bhattacharya, N. Thomas, Y. Duan, P. Chen, J. Canny, P. Abbeel, and Y. Song, "Evaluating protein transfer learning with tape," *Advances in neural information processing systems*, vol. 32, 2019.

[27] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, *et al.*, "Highly accurate protein structure prediction with alphafold," *Nature*, vol. 596, no. 7873, pp. 583–589, 2021.

[28] J. Meier, R. Rao, R. Verkuil, J. Liu, T. Sercu, and A. Rives, "Language models enable zero-shot prediction of the effects of mutations on protein function," *Advances in Neural Information Processing Systems*, vol. 34, pp. 29287–29303, 2021.

[29] G. Benegas, S. S. Batra, and Y. S. Song, "Dna language models are powerful zero-shot predictors of non-coding variant effects," *bioRxiv*, pp. 2022–08, 2022.

[30] B. Levy, Z. Xu, L. Zhao, K. Kremling, R. Altman, P. Wong, and C. Tanner, "Florabert: cross-species transfer learning withattention-based neural networks for geneexpression prediction," 2022.

[31] H. Liu, D. Tam, M. Muqeeth, J. Mohta, T. Huang, M. Bansal, and C. A. Raffel, "Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 1950–1965, 2022.

[32] B. Tian and J. L. Manley, "Alternative polyadenylation of mrna precursors," *Nature reviews Molecular cell biology*, vol. 18, no. 1, pp. 18–30, 2017.

[33] S. Zhu, W. Ye, L. Ye, H. Fu, C. Ye, X. Xiao, Y. Ji, W. Lin, G. Ji, and X. Wu, "Plantapadb: a comprehensive database for alternative polyadenylation sites in plants," *Plant physiology*, vol. 182, no. 1, pp. 228–242, 2020.

[34] X. Gao, J. Zhang, Z. Wei, and H. Hakonarson, "Deeppolya: a convolutional neural network approach for polyadenylation site prediction," *IEEE Access*, vol. 6, pp. 24340–24349, 2018.

[35] D. Staiger and J. W. Brown, "Alternative splicing at the intersection of biological timing, development, and stress responses," *The Plant Cell*, vol. 25, no. 10, pp. 3640–3656, 2013.

[36] A. S. Reddy, Y. Marquez, M. Kalyna, and A. Barta, "Complexity of the alternative splicing landscape in plants," *The Plant Cell*, vol. 25, no. 10, pp. 3657–3683, 2013.

[37] X. Zhang, W. Wang, W. Zhu, J. Dong, Y. Cheng, Z. Yin, and F. Shen, "Mechanisms and functions of long non-coding rnas at multiple regulatory levels," *International journal of molecular sciences*, vol. 20, no. 22, p. 5573, 2019.

[38] J. Meng, Q. Kang, Z. Chang, and Y. Luan, "Plncrna-hdeep: plant long noncoding rna prediction using hybrid deep learning based on two encoding styles," *BMC bioinformatics*, vol. 22, no. 3, pp. 1–16, 2021.

[39] R. Andersson and A. Sandelin, "Determinants of enhancer and promoter activities of regulatory elements," *Nature Reviews Genetics*, vol. 21, no. 2, pp. 71–87, 2020.

[40] T. Jores, J. Tonnies, T. Wrightsman, E. S. Buckler, J. T. Cuperus, S. Fields, and C. Queitsch, "Synthetic promoter designs enabled by a comprehensive analysis of plant core promoters," *Nature Plants*, vol. 7, no. 6, pp. 842–855, 2021.

[41] M. E. Hauberg, J. Creus-Muncunill, J. Bendl, A. Kozlenkov, B. Zeng, C. Corwin, S. Chowdhury, H. Kranz, Y. L. Hurd, M. Wegner, *et al.*, "Common schizophrenia risk variants are enriched in open chromatin regions of human glutamatergic neurons," *Nature communications*, vol. 11, no. 1, p. 5581, 2020.

[42] E. Rodgers-Melnick, D. L. Vera, H. W. Bass, and E. S. Buckler, "Open chromatin reveals the functional maize genome," *Proceedings of the National Academy of Sciences*, vol. 113, no. 22, pp. E3177–E3184, 2016.

[43] K. A. Maher, M. Bajic, K. Kajala, M. Reynoso, G. Pauluzzi, D. A. West, K. Zumstein, M. Woodhouse, K. Bubb, M. W. Dorrity, *et al.*, "Profiling of accessible chromatin regions across multiple plant species and cell types reveals common gene regulatory principles and new control modules," *The Plant Cell*, vol. 30, no. 1, pp. 15–36, 2018.

[44] Ž. Avsec, V. Agarwal, D. Visentin, J. R. Ledsam, A. Grabska-Barwinska, K. R. Taylor, Y. Assael, J. Jumper, P. Kohli, and D. R. Kelley, "Effective gene expression prediction from sequence by integrating long-range interactions," *Nature methods*, vol. 18, no. 10, pp. 1196–1203, 2021.

[45] D. R. Kelley, "Cross-species regulatory sequence activity prediction," *PLoS computational biology*, vol. 16, no. 7, p. e1008050, 2020.

[46] J. Zrimec, C. S. Börlin, F. Buric, A. S. Muhammad, R. Chen, V. Siewers, V. Verendel, J. Nielsen, M. Töpel, and A. Zelezniak, "Deep learning suggests that gene expression is encoded in all parts of a co-evolving interacting gene regulatory structure," *Nature Communications*, vol. 11, no. 1, p. 6141, 2020.

[47] J. D. Washburn, M. K. Mejia-Guerra, G. Ramstein, K. A. Kremling, R. Valluru, E. S. Buckler, and H. Wang, "Evolutionarily informed deep learning methods for predicting relative transcript abundance from dna sequence," *Proceedings of the National Academy of Sciences*, vol. 116, no. 12, pp. 5542–5549, 2019.

[48] A. Siepel, G. Bejerano, J. S. Pedersen, A. S. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L. W. Hillier, S. Richards, *et al.*, "Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes," *Genome research*, vol. 15, no. 8, pp. 1034–1050, 2005.

[49] K. S. Pollard, M. J. Hubisz, K. R. Rosenbloom, and A. Siepel, "Detection of nonneutral substitution rates on mammalian phylogenies," *Genome research*, vol. 20, no. 1, pp. 110–121, 2010.

[50] M. Togninalli, Ü. Seren, D. Meng, J. Fitz, M. Nordborg, D. Weigel, K. Borgwardt, A. Korte, and D. G. Grimm, "The aragwas catalog: a curated and standardized arabidopsis thaliana gwas catalog," *Nucleic acids research*, vol. 46, no. D1, pp. D1150–D1156, 2018.

[51] C. Molina and E. Grotewold, "Genome wide analysis of arabidopsis core promoters," *BMC genomics*, vol. 6, no. 1, pp. 1–12, 2005.

[52] Y. Y. Yamamoto, H. Ichida, T. Abe, Y. Suzuki, S. Sugano, and J. Obokata, "Differentiation of core promoter architecture between plants and mammals revealed by ldss analysis," *Nucleic acids research*, vol. 35, no. 18, pp. 6219–6226, 2007.

[53] S. Kumari and D. Ware, "Genome-wide computational prediction and analysis of core promoter elements across plant monocots and dicots," *PloS one*, vol. 8, no. 10, p. e79011, 2013.

[54] T. Morton, J. Petricka, D. L. Corcoran, S. Li, C. M. Winter, A. Carda, P. N. Benfey, U. Ohler, and M. Megraw, "Paired-end analysis of transcription start sites in arabidopsis reveals plant-specific promoter signatures," *The Plant Cell*, vol. 26, no. 7, pp. 2746–2760, 2014.

[55] J. Lu, T. Tang, H. Tang, J. Huang, S. Shi, and C.-I. Wu, "The accumulation of deleterious mutations in rice genomes: a hypothesis on the cost of domestication," *Trends in Genetics*, vol. 22, no. 3, pp. 126–131, 2006.

[56] B. Nabholz, G. Sarah, F. Sabot, M. Ruiz, H. Adam, S. Nidelet, A. Ghesquière, S. Santoni, J. David, and S. Glémin, "Transcriptome population genomics reveals severe bottleneck and domestication cost in the a frican rice (o ryza glaberrima)," *Molecular ecology*, vol. 23, no. 9, pp. 2210–2227, 2014.

[57] S. Renaut and L. H. Rieseberg, "The accumulation of deleterious mutations as a consequence of domestication and improvement in sunflowers and other compositae crops," *Molecular biology and evolution*, vol. 32, no. 9, pp. 2273–2283, 2015.

[58] D. Koenig, J. M. Jiménez-Gómez, S. Kimura, D. Fulop, D. H. Chitwood, L. R. Headland, R. Kumar, M. F. Covington, U. K. Devisetty, A. V. Tat, *et al.*, "Comparative transcriptomics reveals patterns of selection in domesticated and wild tomato," *Proceedings of the National Academy of Sciences*, vol. 110, no. 28, pp. E2655–E2662, 2013.

[59] D. McKey, M. Elias, B. Pujol, and A. Duputié, "The evolutionary ecology of clonally propagated domesticated plants," *New Phytologist*, vol. 186, no. 2, pp. 318–332, 2010.

[60] J. G. Wallace, E. Rodgers-Melnick, and E. S. Buckler, "On the road to breeding 4.0: unraveling the good, the bad, and the boring of crop quantitative genomics," *Annual review of genetics*, vol. 52, pp. 421–444, 2018.

[61] E. M. Long, M. C. Romay, G. Ramstein, E. S. Buckler, and K. R. Robbins, "Utilizing evolutionary conservation to detect deleterious mutations and improve genomic prediction in cassava," *Frontiers in Plant Science*, vol. 13, p. 1041925, 2023.

[62] R. Lozano, E. Gazave, J. P. Dos Santos, M. G. Stetter, R. Valluru, N. Bandillo, S. B. Fernandes, P. J. Brown, N. Shakoor, T. C. Mockler, *et al.*, "Comparative evolutionary genetics of deleterious load in sorghum and maize," *Nature Plants*, vol. 7, no. 1, pp. 17–24, 2021.

[63] M. S. Kovalev, A. A. Igolkina, M. G. Samsonova, and S. V. Nuzhdin, "A pipeline for classifying deleterious coding mutations in agricultural plants," *Frontiers in plant science*, vol. 9, p. 1734, 2018.

[64] H. Gao, T. Hamp, J. Ede, J. G. Schraiber, J. McRae, M. Singer-Berk, Y. Yang, A. S. Dietrich, P. P. Fiziev, L. F. Kuderna, *et al.*, "The landscape of tolerated genetic variation in humans and primates," *Science*, vol. 380, no. 6648, p. eabn8153, 2023.

[65] A. I. Andreou, J. Nirkko, M. Ochoa-Villarreal, and N. Nakayama, "Mobius assembly for plant systems highlights promoter-terminator interaction in gene regulation," *bioRxiv*, pp. 2021–03, 2021.

[66] J. T. Cuperus, "Single-cell genomics in plants: current state, future directions, and hurdles to overcome," *Plant Physiology*, vol. 188, no. 2, pp. 749–755, 2022.

[67] A. P. Marand and R. J. Schmitz, "Single-cell analysis of cis-regulatory elements," *Current Opinion in Plant Biology*, vol. 65, p. 102094, 2022.

[68] M. W. Dorrity, C. M. Alexandre, M. O. Hamm, A.-L. Vigil, S. Fields, C. Queitsch, and J. T. Cuperus, "The regulatory landscape of arabidopsis thaliana roots at single-cell resolution," *Nature communications*, vol. 12, no. 1, p. 3334, 2021.

[69] A. Farmer, S. Thibivilliers, K. H. Ryu, J. Schiefelbein, and M. Libault, "Single-nucleus rna and atac sequencing reveals the impact of chromatin accessibility on gene expression in arabidopsis roots at the single-cell level," *Molecular Plant*, vol. 14, no. 3, pp. 372–383, 2021.

[70] H. Cui, C. Wang, H. Maan, K. Pang, F. Luo, and B. Wang, "scgpt: Towards building a foundation model for single-cell multi-omics using generative ai," *bioRxiv*, pp. 2023–04, 2023.

[71] F. Yang, W. Wang, F. Wang, Y. Fang, D. Tang, J. Huang, H. Lu, and J. Yao, "scbert as a large-scale pretrained deep language model for cell type annotation of single-cell rna-seq data," *Nature Machine Intelligence*, vol. 4, no. 10, pp. 852–866, 2022.

[72] R. J. Schmitz, E. Grotewold, and M. Stam, "Cis-regulatory sequences in plants: Their importance, discovery, and future challenges," *The plant cell*, vol. 34, no. 2, pp. 718–741, 2022.

[73] A. P. Marand, A. L. Eveland, K. Kaufmann, and N. M. Springer, "cis-regulatory elements in plant development, adaptation, and evolution," *Annual review of plant biology*, vol. 74, pp. 111–137, 2023.

[74] E. Nguyen, M. Poli, M. Faizi, A. Thomas, C. Birch-Sykes, M. Wornow, A. Patel, C. Rabideau, S. Massaroli, Y. Bengio, *et al.*, "Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution," *arXiv preprint arXiv:2306.15794*, 2023.

[75] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)* (J. Burstein, C. Doran, and T. Solorio, eds.), pp. 4171–4186, Association for Computational Linguistics, 2019.

[76] A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C. L. Zitnick, J. Ma, *et al.*, "Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences," *Proceedings of the National Academy of Sciences*, vol. 118, no. 15, p. e2016239118, 2021.

[77] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[78] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized BERT pretraining approach," *CoRR*, vol. abs/1907.11692, 2019.

[79] A. D. Yates, J. Allen, R. M. Amode, A. G. Azov, M. Barba, A. Becerra, J. Bhai, L. I. Campbell, M. Carbajo Martinez, M. Chakiachvili, *et al.*, "Ensembl genomes 2022: an expanding genome resource for non-vertebrates," *Nucleic acids research*, vol. 50, no. D1, pp. D996–D1003, 2022.

[80] J. C. Loke, E. A. Stahlberg, D. G. Strenski, B. J. Haas, P. C. Wood, and Q. Q. Li, "Compilation of mrna polyadenylation signals in arabidopsis revealed a new signal element and potential secondary structures," *Plant physiology*, vol. 138, no. 3, pp. 1457–1468, 2005.

[81] A. K. Baten, B. C. Chang, S. K. Halgamuge, and J. Li, "Splice site identification using probabilistic parameters and svm classification," in *BMC bioinformatics*, vol. 7, pp. 1–15, BioMed Central, 2006.

[82] J. Mergner, M. Frejno, M. List, M. Papacek, X. Chen, A. Chaudhary, P. Samaras, S. Richter, H. Shikata, M. Messerer, *et al.*, "Mass-spectrometry-based draft of the arabidopsis proteome," *Nature*, vol. 579, no. 7799, pp. 409–414, 2020.

[83] H. Parkinson, M. Kapushesky, M. Shojatalab, N. Abeygunawardena, R. Coulson, A. Farne, E. Holloway, N. Kolesnykov, P. Lilja, M. Lukk, *et al.*, "Arrayexpress—a public database of microarray experiments and gene expression profiles," *Nucleic acids research*, vol. 35, no. suppl_1, pp. D747–D750, 2007.

[84] J. W. Walley, R. C. Sartor, Z. Shen, R. J. Schmitz, K. J. Wu, M. A. Urich, J. R. Nery, L. G. Smith, J. C. Schnable, J. R. Ecker, *et al.*, "Integration of omic networks in a developmental atlas of maize," *Science*, vol. 353, no. 6301, pp. 814–818, 2016.

[85] T. Barrett, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, M. Holko, *et al.*, "Ncbi geo: archive for functional genomics data sets—update," *Nucleic acids research*, vol. 41, no. D1, pp. D991–D995, 2012.

[86] T. G. C. K. D. R. I. S. S. . T. S. . H. H. . A. E. . S. K. . I. S. . K. T. . N. Y. . S. D. . A. K. 1, a. R. c. E. M. . K. J. . 454 Life Sciences, A. E. I. B. R. 3, B. A. of Agriculture, F. S. L. C. . 5, B.-S. S. Y. . X. X. . P. S. . C. S. . L. X. . R. Y. . W. J. 6, B.-G. S. A. A. . D. P. F. . T. S. 7, C. for BioSystems Genomics Stiekema Willem J. 9, C. N. de Análisis Genómico (CNAG) Ribeca Paolo 10 Alioto Tyler 10, C. A. U. Y. W. 11, *et al.*, "The tomato genome sequence provides insights into fleshy fruit evolution," *Nature*, vol. 485, no. 7400, pp. 635–641, 2012.

[87] I. Papatheodorou, P. Moreno, J. Manning, A. M.-P. Fuentes, N. George, S. Fexova, N. A.

Fonseca, A. Füllgrabe, M. Green, N. Huang, *et al.*, "Expression atlas update: from tissues to single cells," *Nucleic acids research*, vol. 48, no. D1, pp. D77–D83, 2020.

[88] A. J. Severin, J. L. Woody, Y.-T. Bolon, B. Joseph, B. W. Diers, A. D. Farmer, G. J. Muehlbauer, R. T. Nelson, D. Grant, J. E. Specht, *et al.*, "Rna-seq atlas of glycine max: a guide to the soybean transcriptome," *BMC plant biology*, vol. 10, pp. 1–16, 2010.

[89] A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras, "Star: ultrafast universal rna-seq aligner," *Bioinformatics*, vol. 29, no. 1, pp. 15–21, 2013.

[90] M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for rna-seq data with deseq2," *Genome biology*, vol. 15, no. 12, pp. 1–21, 2014.

[91] R. Lozano, G. T. Booth, B. Y. Omar, B. Li, E. S. Buckler, J. T. Lis, D. P. Del Carpio, and J.-L. Jannink, "Rna polymerase mapping in plants identifies intergenic regulatory elements enriched in causal variants," *G3*, vol. 11, no. 11, p. jkab273, 2021.

[92] M. C. Wilson, A. M. Mutka, A. W. Hummel, J. Berry, R. D. Chauhan, A. Vijayaraghavan, N. J. Taylor, D. F. Voytas, D. H. Chitwood, and R. S. Bart, "Gene expression atlas for the food security crop cassava," *New Phytologist*, vol. 213, no. 4, pp. 1632–1641, 2017.

A Methods

A.1 Model

In our work we employ Language Models (LMs) as the approach to learning the underlying representations of plant genomes. LMs have been proven to exhibit strong performance and generalization in NLP, proteomics, and recently in our work in genomics [75, 76, 24]. Our previous work showed the ability of our pre-trained LM's to match or out perform task specific models on variety of human genomics tasks. LM's are statistical systems used to estimate the probability of a sequence of tokens occurring given some vocabulary of tokens, where tokens could be words, amino acids, or nucleotide k-mers. Two different methods for Language Modeling are Masked Language Modelling (MLM), where a proportion of the input sequence is masked and/or randomly replaced with another word [75], and Causal Language Modelling (CLM), where a sequence is used to predict the next word [77]. More specifically, we use a deep learning model called transformer whose primary component,the self-attention mechanism, is capable of modelling global dependencies within sequences[17]. Our training technique closely follows BERT, the Bidirectional Encoder Representation from Transformer, which employs MLM, thus allowing it to benefit from the self-attention bidirectional property[75].

A.1.1 Architecture

Our model architecture can be described broadly as an encoder only transformer that contains one billion trainable parameters. Our tokenizer consists of a vocabulary of all possible 6-mers using the four nucleotides A, T, C and G, additional special tokens, and N to represent all unknown nucleotides in a sequence. The tokenization algorithm works by first splitting genomic DNA sequences on N's and subsequently identifying 6-mers in the chunked sequences. Our tokenization strategy relies exclusively on non-overlapping 6-mers. The model takes as input a sequence of 1024 tokens which are passed into an embedding layer that projects each token into a 1500 dimensional space and are subsequently added to a learned positional embedding. This combined embedding vector is then fed through 40 attention blocks, with each attention block consisting of a layer normalization layer, multi-headed attention layer, and multi-layer perceptron (MLP). The final layer,a Roberta LM head [78], transforms the final attention layer output into a distribution over the tokenizer's vocabulary at each position of the input sequence.

A.2 Dataset

Our pre-training dataset was built from (mostly) edible plants reference genomes contained in the Ensembl Plants database [79]. The dataset consists of approximately 10.5 million genomic sequences across 48 different species. See Supplementary Table 1 for the dataset details, including genome annotation version and source. For data preparation, we followed the strategy employed in our previous work [24]. Once we obtained the reference genome FASTA for each species, we assembled them into a single FASTA file. In this FASTA file, all nucleotides other than A, T, C, G were replaced by N. We used a tokenizer to convert strings of letters into sequences of tokens. The tokenizer's alphabet consisted of the $4^6 = 4096$ possible 6-mer combinations obtained by combining A, T, C, G, as well as five additional tokens representing standalone A, T, C, G, and N. It also included three special tokens: the padding [PAD], masking [MASK], and the beginning of sequence (also called class; [CLS]) token. This resulted in a vocabulary of 4104 tokens. To tokenize an input sequence, the tokenizer started with a class token and then converted the sequence from left to right, matching 6-mer tokens when possible, or using the standalone tokens when necessary (for instance, when the letter N was present or if the sequence length was not a multiple of 6). Similar to our previous work, the first step of the pre-training involved splitting each reference genome into chunks of 6100 nucleotides, where each chunk shared the first and last 50 nucleotides with the previous and last chunk, respectively. As a data augmentation exercise, for each epoch and chunk, a starting nucleotide index was randomly sampled between 0 and 100, and the sequence was then tokenized from this nucleotide until 1000 tokens were reached. At each step, a batch of sequences randomly sampled from the epoch set was fed to the model.

A.3 Pre-training

The MLM objective was used to pre-train our model in a self-supervised manner. In a self-supervised learning setting annotations (supervision) for each sequence are not needed as we can mask some proportion of the sequence and use the information contained in the unmasked portion of the sequence to predict the masked locations. This allows us to leverage the vast amount of unlabeled genomic sequencing data available. Specifically, 15% of the tokens in the input sequence are selected to be augmented with 80% being replaced with a mask token, 10% randomly replaced by another token from the vocabulary, and the final 10% maintaining the same token. The tokenized sequence is passed through the model and a cross entropy loss is computed for the masked tokens. Updates to the model parameters are computed through gradient descent with Adam, the adaptive learning rate optimizer. Pre-training was carried out with a sequence length of 1024 tokens and an effective batch size of 1.5M tokens for 315k update steps, resulting in the model training on a total of 472.5B tokens. The Adam optimizer was used with a modified square decay learning rate (lr) schedule where a warm up period with linear increase was used for 64K steps with initial lr of 0.00005 and end lr of 0.0001. A square decay function was then applied to the lr for all subsequent steps. The exponential decay rates $\beta_1 = 0.9$ and $\beta_2 = 0.999$ were used in the optimizer.

A.4 Fine-tuning

After pre-training our model with the MLM objective we further fine-tune the model in a supervised learning setting using several annotated, plant genomics datasets. We fine-tune our pre-trained model with IA^3 [31], a method which injects re-scaling weight vectors at the model's activation's, allowing us to achieve high levels of performance while only computing gradient updates for approximately 1% of the model parameters. The tasks upon which we fine-tune include single output regression, multi output regression, binary classification, multi-label classification.

A.5 Hardware

Model pre-training was carried out using Google TPU-V4 accelerators, specifically a TPU v4-1024 machine containing 512 devices. Since the model was able to fit on a single TPU device, pre-training was carried out in a data parallel fashion where model parameters were replicated and the effective batch sharded across the 512 devices with subsequent gradient accumulation. We trained for a total of approx. four days.

A.6 Downstream tasks

A.6.1 Polyadenilation site prediction

We obtained data for alternative polyadenylation (APA) sites from PlantAPAdb, which currently represents the most comprehensive manually curated catalog of APA sites in several plants. The dataset is available at http://www.bmibig.cn/plantAPAdb. For our APA site prediction, we focused on five specific species: *Oryza sativa L.* (japonica and indica), *Arabidopsis thaliana*, *Chlamydomonas reinhardtii*, *Medicago truncatula*, and *Trifolium pratense*. We downloaded all possible sequences around polyadenylation (poly[A]) sites for each of these species. These sequences encompassed different genomic regions such as 3' UTR, 5' UTR, CDS, intron, and exon. Each sequence had a length of 400 bp, with the upstream 300 bp and downstream 100 bp sequence relative to the respective poly(A) site cluster (PAC). The poly(A) site itself was located at the 301st position. To obtain a negative set, we randomly shifted the position of the poly(A) site by both negative and positive values between 1 and 50. We then extracted a 400 bp sequence based on these shuffled positions. The number of negative sequences was twice that of the positive sequences. It is important to note that this strategy of generating negative sequences is more challenging compared to approaches used in other studies. This is because the negative sequences still contain similar gene-related sequences, including those from 3' UTR, 5' UTR, CDS, intron, and exon regions. Furthermore, we emphasize that the model was trained to predict whether a given sequence had a poly(A) site at the 301st position or not. To prevent data leakage, we limited the test set, for each plant species, to sequences that were exclusively derived from a single sampled chromosome. Due to the imbalance in our dataset resulting from the strategy of recreating negative sequences, we evaluated the performance of the models using both

the area under the receiver operating characteristic curve (AUROC) and the precision-recall curve (AUPRC). Additionally, we recreated the dataset utilized in Gao et al. [34] for APA site prediction. The positive sequences were sampled from the 16K dataset originally introduced in Loke et al. [80], which consists of over 16,000 downstream sequences from the 3' UTR of *Arabidopsis thaliana*. Following the approach in Gao et al., we initially selected sequences longer than 162 base pairs and then trimmed them to a sequence length of 162 base pairs. These trimmed sequences included 131 base pairs upstream and 31 base pairs downstream of the annotated poly(A) site. To create negative samples, we randomly sampled 9,704 coding sequences, 3,222 intronic sequences, and 501 5'-UTR sequences from the Arabidopsis Information Resources (TAIR) database. This sampling ensured a balanced dataset, as described in Gao et al. Similarly to the positive sequences, we trimmed the negative sequences to a length of 162 base pairs. Since the Gao et al. study did not specify the training and test splitting ratio, we used a standard 0.9/0.1 split ratio, maintaining the same proportion of negative coding, intronic, and 5'-UTR sequences in both the training and test datasets.

A.6.2 Splicing site prediction

We used the dataset compiled in Baten et al. [81] for splice site prediction, which is available at https://public.bmi.inf.ethz.ch/user/behr/splicing/. The dataset consists of sequences of 398 bp each, containing both acceptor and donor sites, as well as non-acceptor and non-donor sequences across the genome in *Arabidopsis thaliana*. To prevent data leakage, the test set was limited to chromosome 5. Given the highly imbalanced nature of the dataset, with 76,871 acceptor sequences and 2,157,898 non-acceptor sequences, as well as 76,659 donor sequences and 3,311,934 non-donor sequences, we evaluated the performance of the models using both the AUROC and AUPRC.

A.6.3 Long non-coding RNA prediction

We obtained data for long non-coding RNAs (lncRNA) from the Green Non-Coding (GreeNC) database version 2.0, which is available at http://greenc.sequentiabiotech.com/. Our lncRNA prediction focused on six specific species: *Glycine max*, *Manihot esculenta*, *Solanum lycopersicum*, *Sorghum bicolor*, *Triticum aestivum*, and *Zea mays*. For each of these species, we selected lncRNA sequences with a length smaller than 6,000 bp. Additionally, we only included lncRNA sequences that were reported on scaffolds that were present in the version of the associated annotation file for the given reference assembly mentioned in the GreeNC database. To construct the set of negative sequences, we extracted all mRNA sequences from the reference genome, using the assembly mentioned in the GreeNC database. We removed mRNA sequences longer than 6,000 bp to maintain consistency with the lncRNA dataset. To prevent the model from relying solely on sequence length for classification and to encourage the learning of more meaningful features, we performed a matching process between lncRNA and mRNA sequences based on both sequence length and GC content. Specifically, during the matching process, mRNA sequences were selected if they had a length equal to or less than 100 bp compared to the corresponding lncRNA sequence. Similarly, the GC content of the mRNA sequences was within a range of 1% less or 1% more than the GC content of the corresponding lncRNA sequence. To prevent data leakage, we created a test set for each plant species that exclusively consisted of sequences derived from a single sampled chromosome. In the case of *Triticum aestivum*, the test set was limited to chromosomes 1A, 1B, and 1D. We evaluated the performance of the models using both the AUROC and AUPRC. In addition, we replicated the dataset compiled in Meng et al. [38] to compare our lncRNA prediction performance with that of PlncRNA-HDeep. This dataset was also based on the GreeNC database, but included sequences exclusively from *Zea mays*. This dataset consisted of 18,110 lncRNA sequences from the GreeNC database that were downsampled to generate a positive set of 18,000 samples. The negative set was based on 57,776 mRNA sequences obtained from the RefSeq database, and also downsampled to create a balanced dataset. To ensure a fair comparison, we used the same training and test datasets available at https://github.com/kangzhai/PlncRNA-HDeep/. The only difference was that we trimmed down sequences longer than 6,000 bp since our model does not handle longer sequences. This trimming affected 43 sequences in the training dataset and 9 sequences in the test dataset, which are unlikely to greatly affect the performance comparison between models.

A.6.4 Chromatin profiles prediction

We used the dataset compiled in Zhao et al [11] for chromatin profiles prediction. The compiled dataset is based on transposase-accessible chromatin with sequencing (ATAC-seq) data from multiple tissues of six plant species including *Arabidopsis thaliana*, *Oryza sativa*, *Zea mays*, *Setaria italica*, *Sorghum bicolor*, and *Brachypodium distachyon*. NCBI and SRA accession number of the resources and publications used to compile the dataset can be found in the original publication. For data generation, we followed the same strategy as outlined in the publication. We generated the same number of training sequences samples, each of size 1000 bp, retrieved from a reference genome. For each ATAC-seq sample, we labeled each training sequence as 1 (i.e. positive sample) if the middle 200 bp region overlaps with an open chromatin region (OCR) by more than 50% of the sequence length or as 0 (i.e. negative sample) otherwise. To ensure a fair comparison with the Zhao et al predictions we selected the same chromosomes used as validation and test sets for each plant species. As in the Zhao et al study, we evaluated the performance of the models using the area under the AUROC and AUPRC metrics.

A.6.5 Promoter strength prediction

We used the publicly available self-transcribing active regulatory region sequencing (STARR-seq) assays for *Arabidopsis thaliana*, *Zea mays* and *Sorghum bicolor*[40]. The STARR-seq assays included all known promoter regions for the species, which were defined as -165 to +5 bp relative to annotated transcription start sites (TSSs), and their strength was measured in two systems: tobacco leaves and maize protoplasts. To enable direct comparison with the CNN model trained in the original publication, we downloaded their training and test datasets available at https://github.com/tobjores/Synthetic-Promoter-Designs-Enabled-by-a-Comprehensive-Analysis-of-Plant-Core-Promoters/, which included the 170 bp promoter sequences and their measured STARR-seq strength in the tobacco leaf and maize protoplast systems.

A.6.6 Tissue-specific quantitative gene expression level

To test our performance on predicting tissue-specific gene expression levels, we collected data from large-scale studies of expression in five plants (Supplementary Table 2). All the data were downloaded as raw sequencing files and reprocessed with a uniform pipeline. We downloaded all publicly available RNA-seq experiments from the *Arabidopsis thaliana* tissue atlas [82] stored in ArrayExpress [83] (https://www.ebi.ac.uk/biostudies/arrayexpress/studies/E-MTAB-7978), which included 56 experiments across 27 tissues. For *Zea mays*, we downloaded the RNA-seq data published by Walley et al. [84] stored in the GEO Database [85] (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE50191). The *Zea mays* dataset consisted of 68 RNA-seq experiments across 23 tissues. For *Oryza sativa*, we collected samples from two NCBI BioProjects (https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJEB47919 and https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJEB32629), totalling 20 experiments across seven tissues. For *Solanum lycopersicum*, we downloaded the RNA-seq experiments associated to the whole-genome sequencing project [86] and stored in the Gene Expression Atlas [87] (https://www.ebi.ac.uk/gxa/experiments/E-MTAB-4812) which included ten tissues with one replicate each. Finally, we also retrieved the *Glycine max* tissue atlas [88] from the European Nucleotide Archive (https://www.ebi.ac.uk/ena/browser/view/SRA012188) which included 14 samples across nine tissues. The genome and annotation version used for gene expression analyses were downloaded from Ensembl Plants version 56 [79] and are listed in Supplementary Table 3. RNA-seq read mapping was performed with the STAR aligner [89] version 2.7.10b. The genome indices for each species were prepared using STAR's *–runMode genomeGenerate* mode and the options *–sjdbOverhang 99 –genomeSAindexNbases 12* using the genome files and gtf genome annotation files listed in Supplementary Table 3. The raw RNA-seq reads were mapped using STAR's *–quantMode GeneCounts* mode, and with the options *–outFilterType BySJout –outFilterMultimapNmax 40 –alignSJoverhangMin 8 –alignSJDBoverhangMin 1 –outFilterMismatchNmax 999 –outFilterMismatchNoverReadLmax 0.04 –alignIntronMin 20 –alignIntronMax 1000000*. Within each species, the read counts across all tissues and replicates were imported into R and normalized with DESeq2 [90] version 1.38.3. Size factors were first estimated using the *estimateSizeFactors()* method and then used to normalize the reads using the *counts()* method with the *normalized=TRUE*

option. For species that had multiple replicates per tissue, we averaged the resulting counts in each tissue to get one value per tissue. Finally, a pseudocount of one was added to each count and they were log2 transformed. For training the model, we followed a similar strategy as outlined in Washburn et al. [47]. We used promoter-proximal sequences based on the TSS annotations and ensured that the sequence overlapped with the coding sequence by extracting 5000 bp upstream and 1000 bp downstream of the TSS, resulting in a sequence of 6000 bp. To prevent potential data leakage, we divided the promoter sequences into gene families using the Washburn et al. pipeline available at https://bitbucket.org/bucklerlab/p_strength_prediction/. Specifically, we conducted an all-by-all BLAST on a given plant's proteome sequence to evaluate pairwise similarity between proteins. As a given gene can encode multiple protein isoforms, the results of the BLAST search were collapsed to gene-level similarity. Then, a graph was constructed with nodes representing genes and edges connecting paralogous genes. The graph was further divided into clusters (i.e., gene families) using a clustering algorithm, and genes not assigned to a family were considered as a family of their own. Using this set of partly independent gene sequences based on their family assignments, we allocated promoter sequences into training and test datasets, thus ensuring that similar promoter sequences were not leaked. Since the number of promoter sequences allocated to families might differ across species, we randomly chose a number of families that would result in an 80%/20% split for each plant species separately. Lastly, we used the coefficient of determination (R^2) to evaluate the performance of gene expression prediction.

To further evaluate the performance of AgroNT in predicting gene expression and compare it to a previously developed method, we trained a model specifically for predicting expressed and non-expressed genes. We used the dataset compiled by Washburn et al. [47], which is based on 422 individual samples from seven studies, providing a comprehensive collection of *Zea mays* tissues at various developmental stages. The NCBI and SRA accession numbers can be found in the original publication. The dataset included 39,470 sequences with gene expression labels mentioned in the paper, but sequences were available for only 35,402 of those sequences. As we did not retrieve the missing sequences for training, the comparison to their reported performance becomes somewhat more challenging, as the training dataset was reduced. Following the methodology described by Washburn et al., we trained two separate models, each trained on either promoter or terminator sequences with a fixed size of 1500 bp. We classified each sequence as associated with an expressed gene or a non-expressed gene. As done in the Washburn et al. study, we evaluated the performance of the models using the area under the receiver operating characteristic curve (AUC).

A.6.7 Token importance for gene expression level prediction

We also evaluated the significance of a genomic region within the 6000 bp promoter-proximal sequence by performing an importance analysis similar to the one described in [30]. To do this, we randomly replaced each of the 1000 possible tokens within the promoter-proximal sequence with a random token, ensuring that the newly generated token was not the same as the original one. Then, we compared the predicted expression levels of the sequence with the randomly altered token to the predicted expression levels based on the original sequence. In other words, for a given sequence, we generated 1000 new sequences where only a single token differed from the original sequence. We carried out this importance analysis using 100 randomly selected promoter-proximal sequences, drawn only from the test sets, for each of the four plant species that were employed to train the gene expression prediction models. Given that the four plant species gene expression models included various tissues, we averaged the differences across tissues and sequences to derive our final importance scores.

A.6.8 Zero-shot scores for variant consequence prediction

To calculate zero-shot scores for a specific site of interest, we followed these steps: First, for each single nucleotide polymorphism (SNP), we retrieved a 6,000 bp sequence centered on the SNP of interest based on the reference genome of *Arabidopsis thaliana*. Next, we created two sequences: one carrying the reference allele and the other carrying the alternative allele at the SNP position. Then, we computed several zero-shot scores that capture different aspects of the vector distances in the embedding space between these two sequences. These scores include the L1 distance (Manhat-

tan), L2 distance (Euclidean), cosine similarity, and dot-product (not normalized cosine similarity). Additionally, we computed a loss-based score, namely the log-likelihood ratio (LLR), which compares the probabilities of alternate and reference alleles. For the evaluation of zero-shot scores on potential deleterious variants, we utilized the genomes from multiple *Arabidopsis thaliana* samples available through the 1001 Arabidopsis Genome Project (accessible at https://1001genomes.org/). In this analysis, we only considered bi-allelic SNPs and removed any SNP with a genotyping rate lower than 95% to ensure reliable allele frequencies. To annotate missense and non-synonymous SNPs, we used the General Feature Format (GFF) files from the TAIR10 reference genome obtained from Ensembl. For the assessment of zero-shot scores on phenotype-associated variants, we referred to the AraGWAS catalog found at https://aragwas.1001genomes.org. Specifically, we considered all variants annotated as "top association". Similar to the previous analysis, we focused on biallelic SNPs. To distinguish between coding and non-coding associated variants, we used the gene coordinates from the GFF of the TAIR10 reference genome. Variants that overlapped with gene coordinates were labeled as coding, while those that did not were labeled as non-coding. The LLR score, referred to as the Genomic Pre-trained Network (GPN) score in [29], as well as the phastCons and phyloP conservation scores, were downloaded from https://huggingface.co/datasets/gonzalobenegas/processed-data-arabidopsis/. As the two conservation scores relied on multiple alignments across species to compute a score, resulting in some SNPs lacking annotations, we ensured fair comparison across methods by considering only variants with associated scores across all methods.

A.6.9 *In silico* mutagenesis in the cassava genome

To leverage the capability of AgroNT for obtaining accurate predictions across a variety of regulatory features, we further trained AgroNT to predict enhancer elements and gene expression across multiple tissues in the understudied crop, *Manihot esculenta* (cassava). For enhancer elements prediction, we used the PRO-seq data generated in *Manihot esculenta* seedlings from the study conducted by Lozano et al. [91]. Specifically, we used a set of 9665 intergenic regulatory elements (IRE) regions, which the authors defined as those located at least 1000 bp away from any gene. To maintain consistency and avoid reprocessing the PRO-seq data, we used the *Manihot esculenta* reference genome v6.1 from Phytozome, which corresponds to the reference genome employed in the study. The training dataset was constructed as follows: For positive sequences, which represent positive enhancer samples, we extracted a 1000bp sequence centered in the middle of the PRO-seq peak. To create matched negative sequences for each positive sequence, we randomly selected a 1000bp intergenic sequence from the same chromosome, ensuring it neither overlapped with any PRO-seq peak nor differed significantly in GC% from the corresponding positive sequence. Specifically, the GC% of the negative sequence was within 5% points of the GC% of the positive sequence. This approach resulted in a balanced and GC-matched dataset. We then split the training and test sets by chromosomes, ensuring they were strictly non-overlapping. Chromosomes 17 was designated as the test set, and chromosome 9 as the validation set. The remaining chromosomes were used as the training set. The model's performance was evaluated using the AUROC metric.

For gene expression prediction, we used the gene expression atlas from the study conducted by Wilson et al. [92], which measured gene expression levels across 11 tissue/organ types. We employed the reference genome and annotation from Ensembl Plants version 56 and followed the same pipeline to process the RNA sequencing data described above. As for our previous gene expression models, we used promoter-proximal sequences based on TSS annotations to obtain sequences of 6000 bp and followed the same clustering approach to assign each promoter sequence to a gene family. We then used this information to produce a training and test set, using an 80%/20% split. Based on these trained models, we performed a comprehensive profiling of regulatory potential. Specifically, for a given sequence obtained from the reference genome, we systematically mutated each site to all three possible distinct nucleotides. In our initial analyses to detect motifs, we predicted the impact of all possible single-nucleotide substitutions across 200 bp upstream and 200 bp downstream of the midpoint of the top 1000 best predicted enhancer regions. Similarly, for the gene expression analyses, we conducted all possible single-nucleotide substitutions across promoter sequences located within 200 bp upstream and 200 bp downstream of the TSS. We selected 1000 promoter sequences of genes that had the highest average gene expression levels across all tissues. To measure the impact of a

mutation, we computed the log-fold change as $LFC = log_2(P_1/P_0)$. Here, P_0 represents the probability (or gene expression level) predicted for the original sequence, and P_1 represents the probability (or gene expression level) predicted for the mutated sequence. Given the observation that AgroNT has gained an understanding of the importance of disrupting critical motifs in enhancer and promoter-proximal regions concerning gene expression values, we also assessed all possible single-nucleotide substitutions within 50 bp upstream and 50 bp downstream of the midpoint for all enhancer regions. Similarly, we performed all possible single-nucleotide substitutions within 50 bp upstream and 50 bp downstream of the TSS for all promoter-proximal regions. This resulted in a total of ∼2.9 and ∼7.3 million mutations, respectively, which we have made available for future research purposes.

We visualized the mutation scores of a given sequence using the ggseqlogo function from the R package ggseqlogo. To identify transcription factor motifs corresponding to patterns predicted to be important for enhancer or promoter activity, we mapped the positions of 656 plant TF motifs from the non-redundant Jaspar CORE 2022 collection (`https://jaspar.genereg.net/download/data/2022/CORE`) in these sequences. We accomplished this using the matchMotifs function from the R package motifmatchr, with the following parameters: p.cutoff = 1e-04 and bg = "even". The figure panel highlights TF motif types that match positions with high prediction mutation scores.

B Code availability

The pre-trained AgroNT model has been made publicly available via HuggingFace at https://huggingface.co/InstaDeepAI/agro-nt

C Data availability

Pre-training reference sequences were acquired from publicly available resources and are listed in Supplementary Table 1. Gene annotations were obtained from Ensembl Plants (`https://plants.ensembl.org`) and Phytozome (`https://phytozome-next.jgi.doe.gov/`). A comprehensive description of the datasets, including links for all the datasets presented in the study, can be found in the Methods section. The Plant Genomic Benchmark has been made publicly available via HuggingFace at https://huggingface.co/datasets/InstaDeepAI/plant-genomic-benchmark.

D Competing interests

This work was done in the course of employment at InstaDeep and Google DeepMind, with no other competing financial interests.

E Acknowledgements

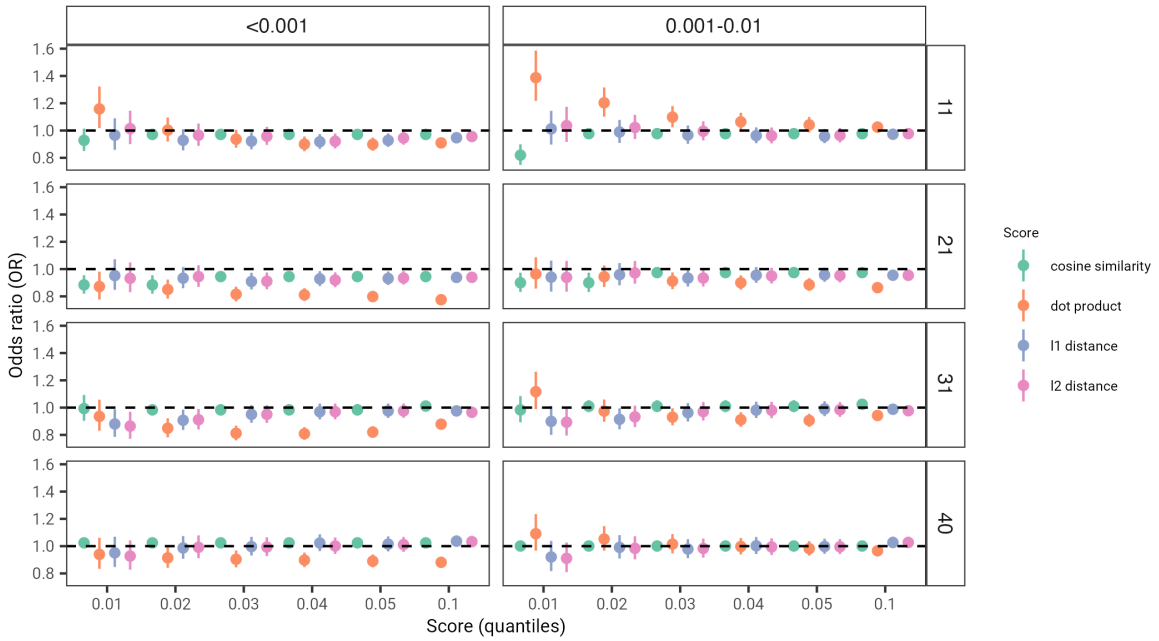
This research was supported by Cloud TPUs from Google's TPU Research Cloud (TRC). We also extend our gratitude to the researchers who deposited experimental data in public databases, those who maintain these databases, and those who make analytical and predictive methods accessible to the scientific community.

F Supplementary Tables

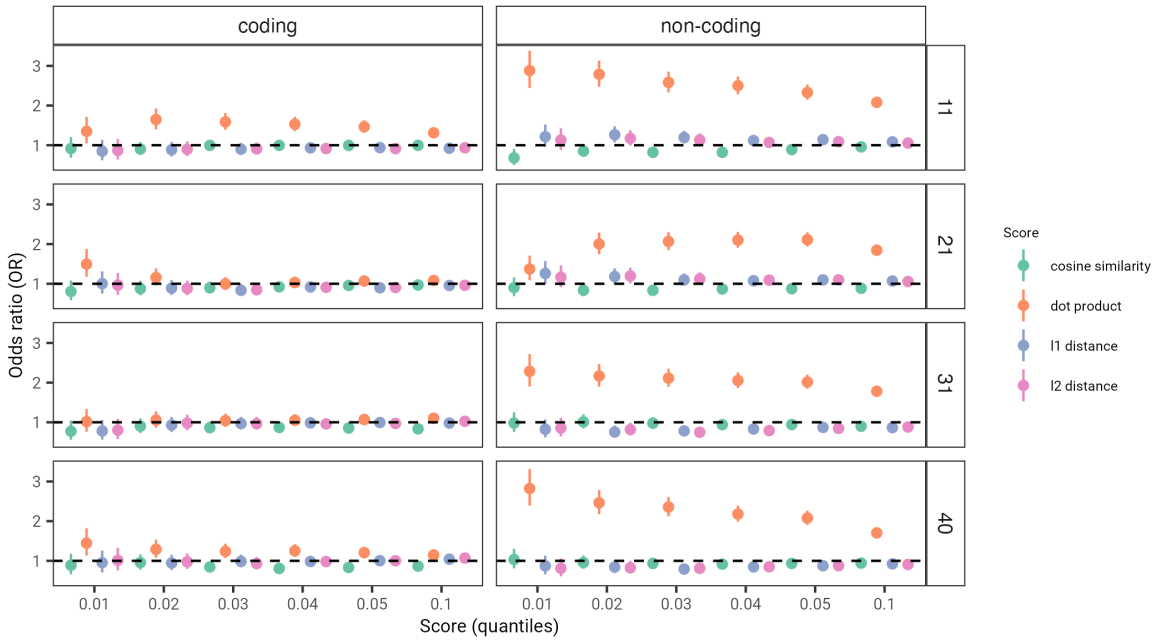
Supplementary tables 1-8 (separate file).

Supplementary table 1. Genome, annotation version, and source of reference genomes used for pre-training.
Supplementary table 2. RNA sequencing datasets used for gene expression prediction.
Supplementary table 3. Genome, annotation version, and source of reference genomes used for gene expression prediction.
Supplementary table 4. R^2 values between observed and predicted expression across tissues on holdout genes for *Arabidopsis thaliana*.
Supplementary table 5. R^2 values between observed and predicted expression across tissues on holdout genes for *Glycine max*.
Supplementary table 6. R^2 values between observed and predicted expression across tissues on holdout genes for *Oryza sativa*.
Supplementary table 7. R^2 values between observed and predicted expression across tissues on holdout genes for *Solanum lycopersicum*.
Supplementary table 8. R^2 values between observed and predicted expression across tissues on holdout genes for *Zea mays*.

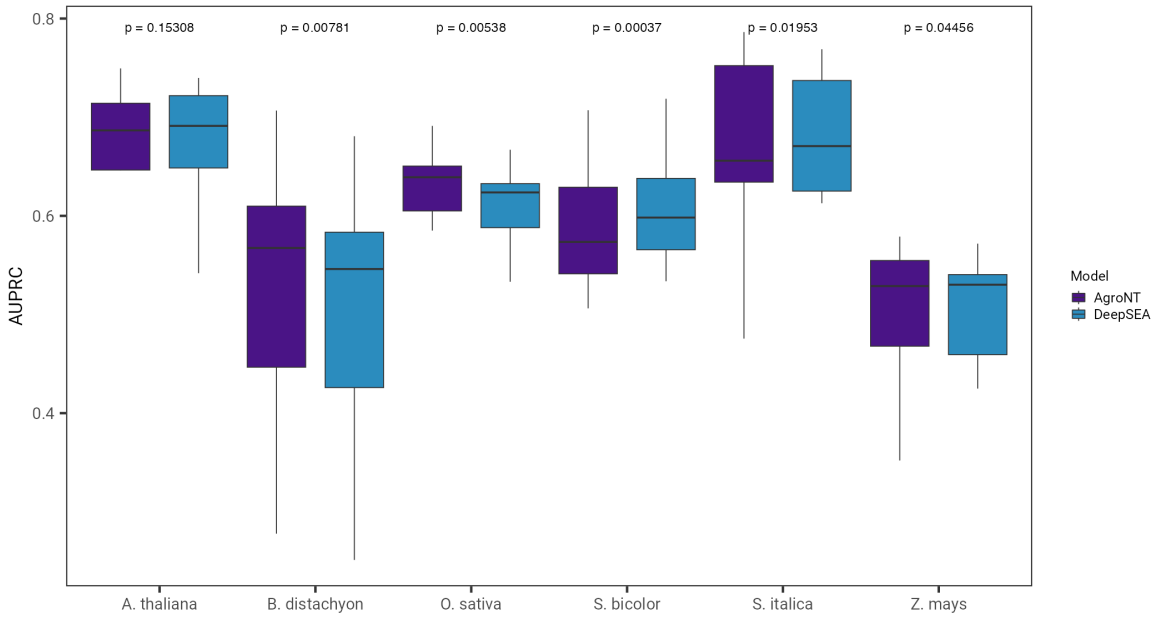
G Supplementary Figures



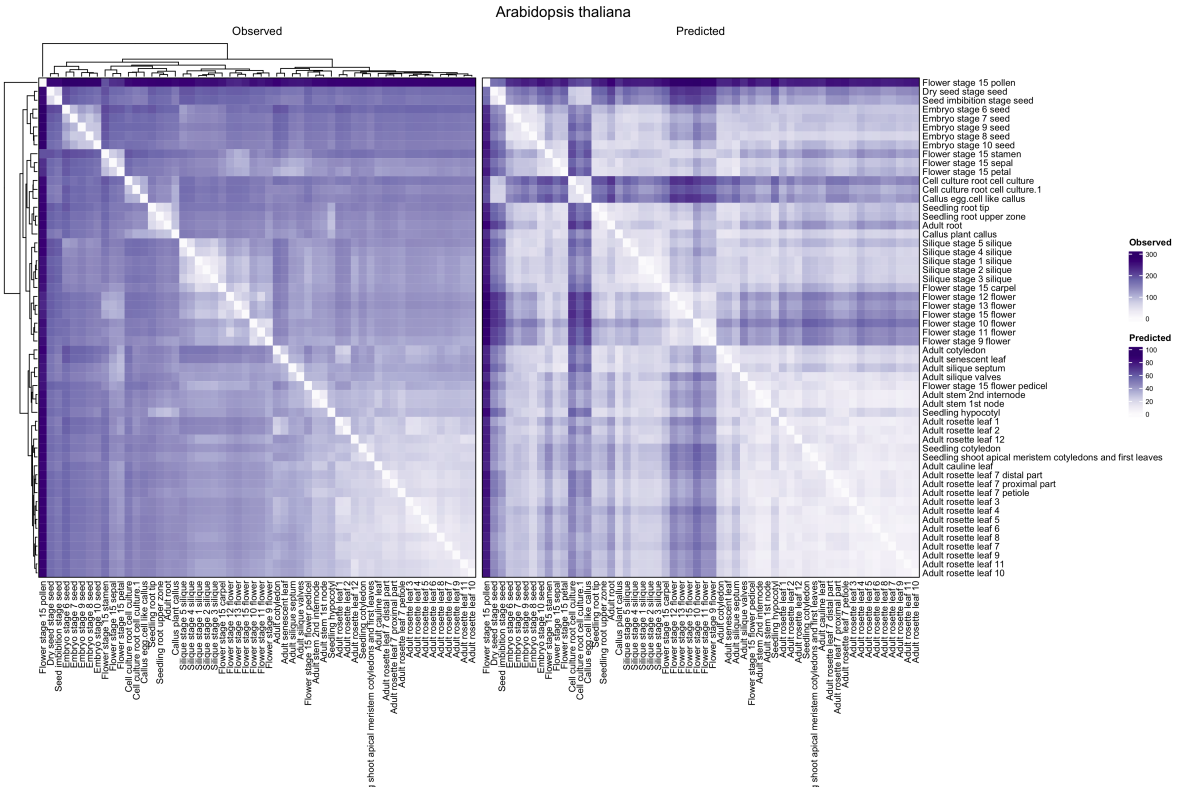
Supplementary Figure 1: Odd ratios (OR) across different score thresholds used to assess enrichment of two frequency classes of missense variants. The performance of four zero-shot scores are shown. Error bars denote 95% confidence intervals. Dashed black line indicates an OR of 1.



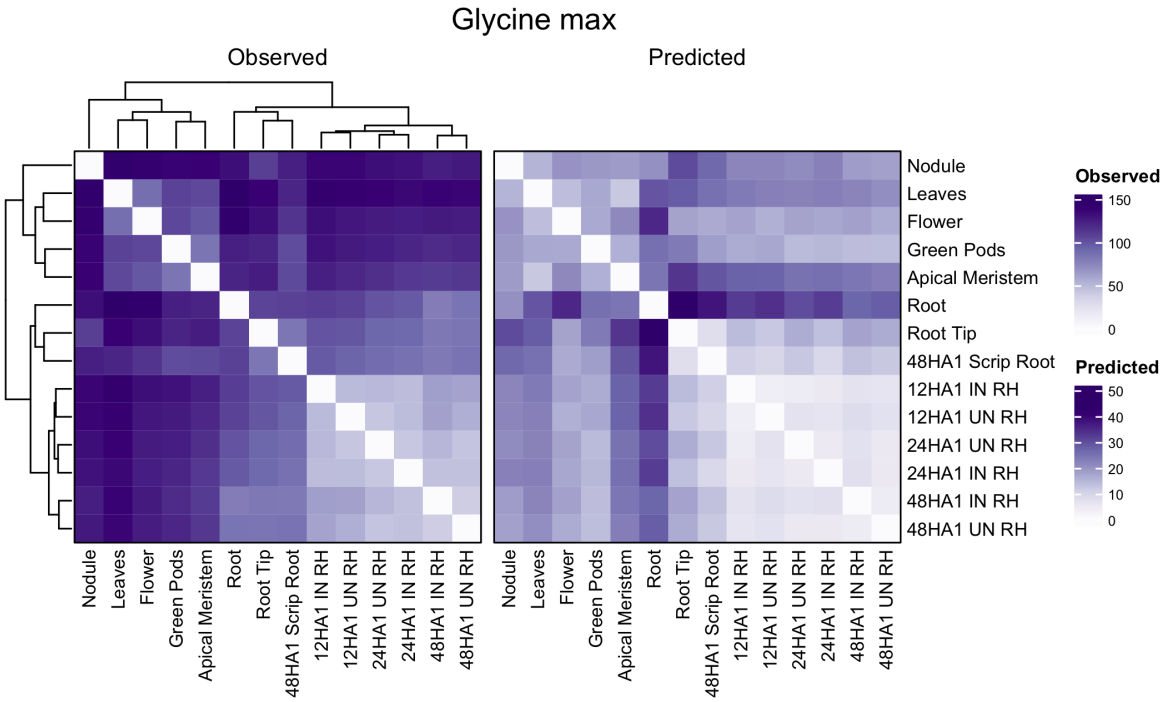
Supplementary Figure 2: Odd ratios (OR) across different score thresholds used to assess enrichment of coding and non-coding GWAS variants. The performance of four zero-shot scores are shown. Error bars denote 95% confidence intervals. Dashed black line indicates an OR of 1.



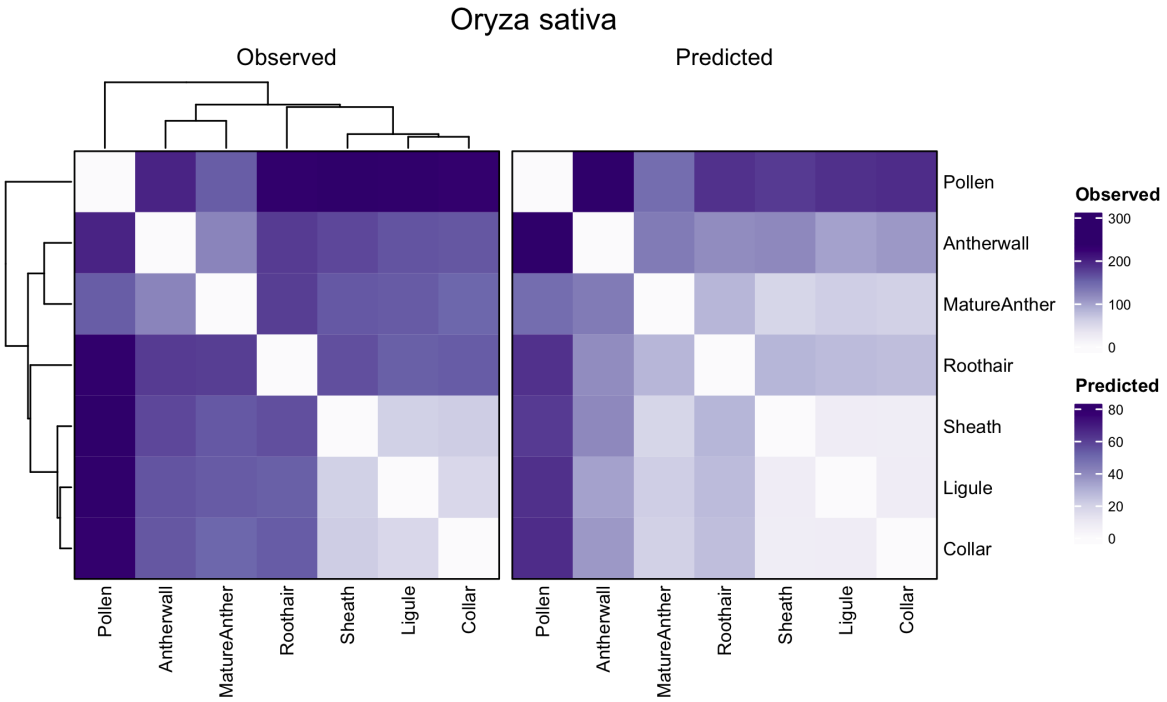
Supplementary Figure 3: The performance of AgroNT in predicting chromatin profiles across species and tissues is measured using the area under the curve of the precision-recall curve (AUPRC). P-values shown are based on a two-sided Wilcoxon signed test comparing the performance of AgroNT with a model based on the DeepSEA architecture.



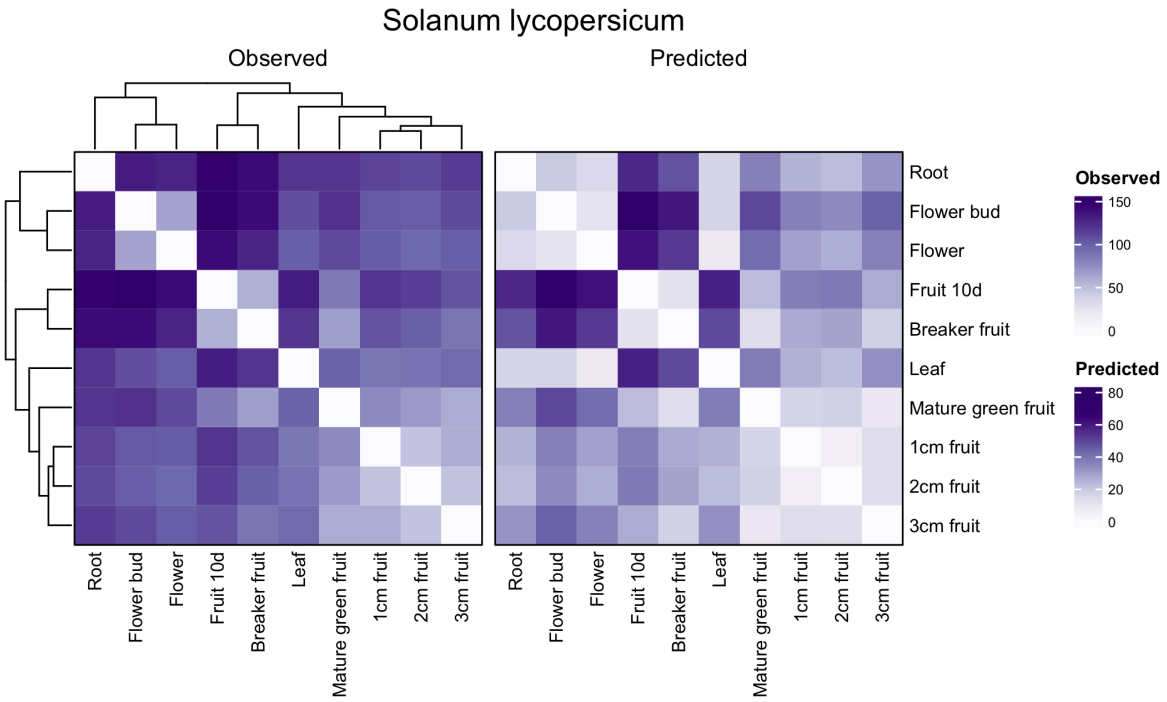
Supplementary Figure 4: Tissue expression profiles between observed and predicted gene expression levels on holdout genes for *Arabidopsis thaliana*. The order of the cells in the heatmap is based on a hierarchical clustering using the euclidean distance of the observed gene expression levels across tissues.



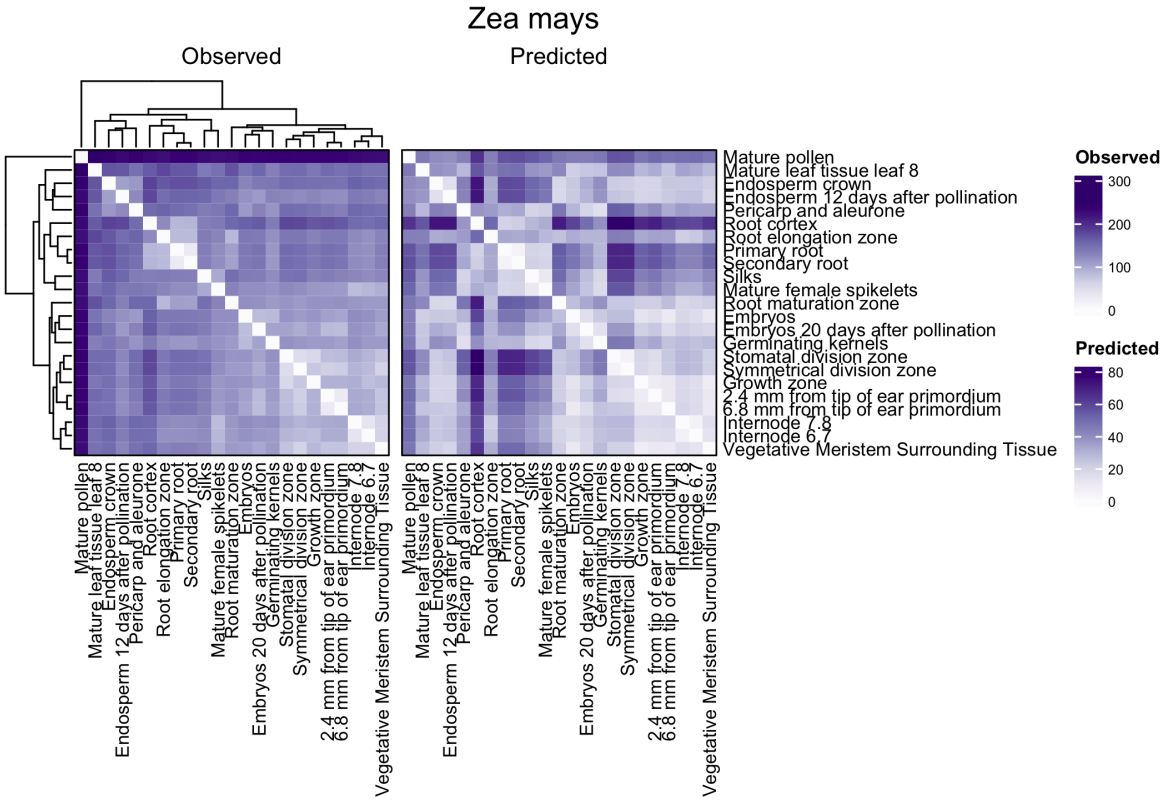
Supplementary Figure 5: Tissue expression profiles between observed and predicted gene expression levels on holdout genes for *Glycine max*. The order of the cells in the heatmap is based on a hierarchical clustering using the euclidean distance of the observed gene expression levels across tissues.



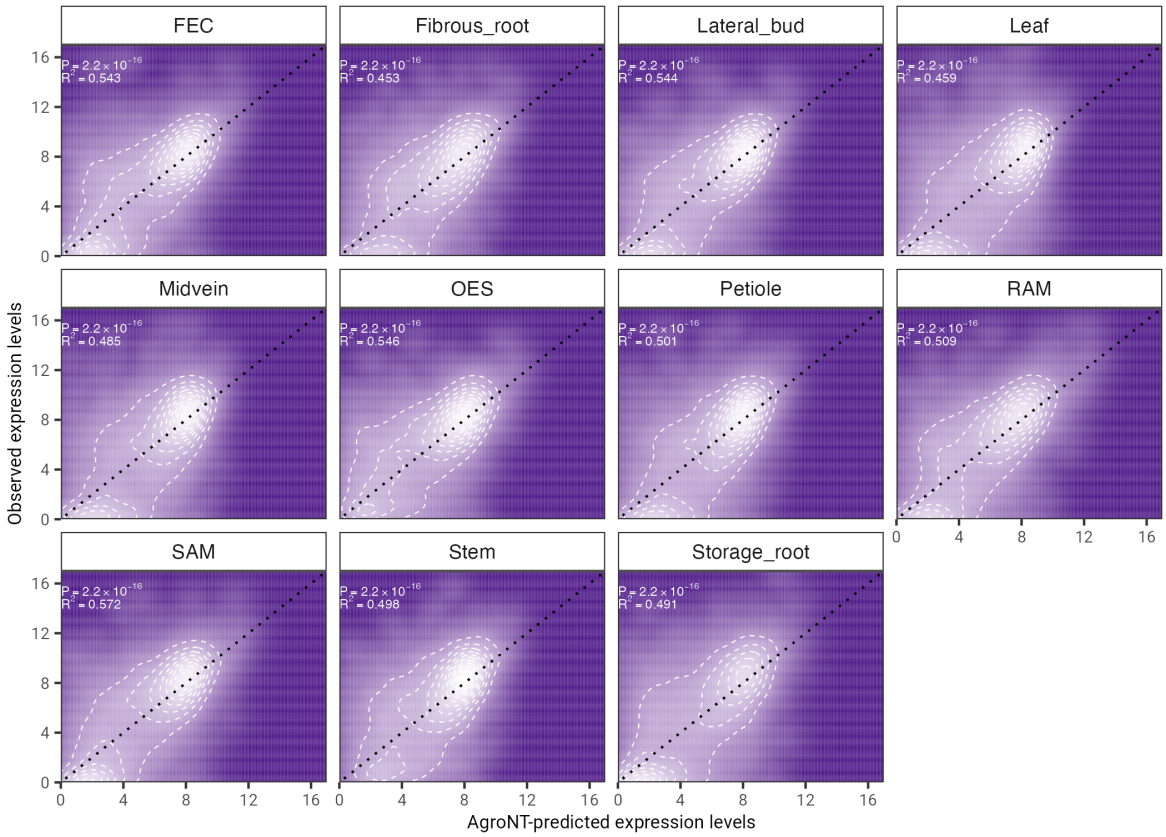
Supplementary Figure 6: Tissue expression profiles between observed and predicted gene expression levels on holdout genes for *Oryza sativa*. The order of the cells in the heatmap is based on a hierarchical clustering using the euclidean distance of the observed gene expression levels across tissues.



Supplementary Figure 7: Tissue expression profiles between observed and predicted gene expression levels on holdout genes for *Solanum lycopersicum*. The order of the cells in the heatmap is based on a hierarchical clustering using the euclidean distance of the observed gene expression levels across tissues.



Supplementary Figure 8: Tissue expression profiles between observed and predicted gene expression levels on holdout genes for *Zea mays*. The order of the cells in the heatmap is based on a hierarchical clustering using the euclidean distance of the observed gene expression levels across tissues.



Supplementary Figure 9: Gene expression prediction on holdout genes across cassava tissues. The coefficient of determination (R^2) from a linear model and associated P-values between predicted and observed values are shown.