Genome Biology

RESEARCH

Open Access

Cross-species prediction of histone modifications in plants via deep learning

Tongxuan Lv[1], Quan Han[1], Yilin Li[1], Chen Liang[1], Zhonghao Ruan[1], Haoyu Chao[2], Ming Chen[2] and Dijun Chen[1*]

*Correspondence: dijunchen@nju.edu.cn

[1] Key Laboratory of Pharmaceutical Biotechnology, School of Life Sciences, Nanjing University, Nanjing 210023, China
[2] Department of Bioinformatics, College of Life Sciences, Zhejiang University, Hangzhou 310058, China

Abstract

Background: The regulation of gene expression in plants is governed by complex interactions between cis-regulatory elements and epigenetic modifications such as histone marks. While deep learning models have achieved success in predicting regulatory features from DNA sequence, their cross-species generalizability in plants remains largely unexplored.

Results: We systematically evaluate the ability of deep learning models to predict histone modifications across plant species using a multi-stage framework based on the Sei architecture. We train species-specific models for Arabidopsis (*Arabidopsis thaliana*), rice (*Oryza sativa*), and maize (*Zea mays*), achieving high within-species predictive performance and strong agreement between predictions and experimental ChIP-seq profiles. However, cross-species predictions show reduced performance with increasing phylogenetic distance, highlighting limited model transferability between monocots and dicots. To improve generalization, we construct a Poaceae family-level model by jointly training on rice and maize, and an Arabidopsis-trained model based solely on Arabidopsis. These models demonstrate robust predictive power in completely unprofiled species that are not used in training set, highlighting the model's adaptability to novel plant genomes based solely on conserved regulatory syntax. In contrast, cross-family models produce less consistent results, with reliable performance only in species sharing conserved regulatory features. We also develop an easy-to-use pipeline that predicts genome-wide chromatin signals directly from DNA sequences.

Conclusions: Our findings demonstrate that phylogenetically informed model training significantly improves cross-species epigenomic prediction, offering a scalable computational strategy for functional annotation in non-model and agriculturally important plants.

Keywords: Cross-species prediction, Plant epigenomics, Deep learning, Regulatory sequence modeling, Histone modification

Background

The precise regulation of gene expression in plants is jointly controlled by complex interactions between DNA sequence elements (such as transcription factor binding sites (TFBS), promoters, enhancers, and inducible elements) and epigenetic modifications (particularly histone modifications and DNA methylation) [1–3]. These elements and modifications are collectively referred to as cis-regulatory elements. Among these, histone modifications — including methylation (e.g., H3K4me3) and acetylation (e.g., H3K27ac) — play a key role in regulating chromatin accessibility and transcriptional activity, thereby significantly influencing plant development, environmental adaptability, and phenotypic plasticity [1, 4, 5]. While DNA sequence–encoded *cis*-elements provide a foundational scaffold for the targeting of histone modifications, their deposition is also shaped by trans-acting factors, including transcription factors, chromatin remodelers, and chromatin-associated complexes. These trans regulators interact with *cis*-elements or act through chromatin loops and protein–protein interactions, enabling dynamic, context-specific regulation of the epigenomic landscape [6, 7].

In recent years, machine learning strategies — especially deep learning methods — have emerged as the state-of-the-art approach, achieving significant success in predicting various genomic features [8], including cis-regulatory elements [9–12], chromatin states [13, 14], and gene expression levels [15, 16]. These advancements have significantly enhanced the ability to infer regulatory activity directly from DNA sequences, thereby opening new avenues for understanding genomic function [17, 18].

Despite these advances, the field still faces two major challenges. First, there is a lack of systematic evaluation of the cross-species generalization ability of deep learning models [19–22]. Most existing models are trained and validated in a single-species context and have not been evaluated for their transferability to phylogenetically distant plant lineages. Second, the vast majority of deep learning-based regulatory models are developed and evaluated using data from animal systems, humans, or a few model plant species [21, 23]. Therefore, their applicability to diverse and agriculturally important plant species remains largely unexplored [24, 25]. The complexity of plant gene regulation and evolutionary diversity further exacerbate these challenges [26]. Due to adaptive evolution and the diversification of regulatory mechanisms, regulatory elements and histone modification patterns may exhibit significant differences across plant species [27]. Consequently, deep learning models trained on a single species or closely related species may demonstrate limited generalization ability when predicting genome-wide regulation in distantly related species.

To address the limited cross-species transferability of deep learning models in plant regulatory genomics, we adopted and tailored the Sei deep learning framework — a multi-task, sequence-based architecture originally designed to predict thousands of regulatory features directly from DNA sequences [28]. Previous studies have demonstrated the applicability of the Osei model — a Sei variant — in plant systems, validating its potential for plant-specific regulatory prediction [29]. Leveraging its capacity to integrate diverse regulatory signals, we retrained the model using plant-specific epigenomic datasets to evaluate its predictive performance across multiple species. We hypothesised that training on phylogenetically related or taxonomically diverse species would enhance generalization to untrained lineages. To test this, we developed a

four-stage experimental framework comprising: (1) species-specific modeling, (2) cross-species prediction, (3) within-family generalization, and (4) cross-family evaluation. This systematic design enabled a comprehensive assessment of how sequence-based models capture regulatory logic and transfer across evolutionary scales in plants.

Results

As illustrated in Fig. 1a, we implemented this framework using chromatin profiling datasets of histone modifications, collected from multiple plant species in the ChIP-Hub database [30]. For each species, signal tracks corresponding to the same chromatin feature were filtered and unified to identify high-confidence regulatory intervals. A 1,024-bp genomic window centered on each region was extracted and one-hot encoded for model input. We trained the Sei model using a hierarchical stack of convolutional layers followed by multi-task classification to learn multi-scale sequence features associated with chromatin activity. The resulting species-specific models were first evaluated within their respective species and subsequently applied to cross-species settings, including both within-family (e.g., Poaceae, Brassicaceae) and cross-family generalization. The following sections detail the performance outcomes and biological insights derived from each stage.

Species-specific modeling of chromatin features in Arabidopsis, Rice, and Maize

We first constructed species-specific models using chromatin feature datasets derived from three representative plant species: rice (*Oryza sativa*), maize (*Zea mays*), and Arabidopsis (*Arabidopsis thaliana*). These species represent major lineages of monocotyledons and dicotyledons and possess rich ChIP-seq datasets, thereby allowing us to evaluate model performance across evolutionarily divergent taxa.

As shown in Fig. 1b the trained models exhibited high accuracy in within-species prediction of chromatin features. Specifically, the rice model achieved an area under the receiver operating characteristic curve (AUROC, reflecting overall classification performance across thresholds) of 0.970 and an area under the precision-recall curve (AUPRC, more informative for imbalanced data) of 0.754; the Arabidopsis model achieved an AUROC of 0.941 and an AUPRC of 0.714; and the maize model achieved an AUROC of 0.965 and an AUPRC of 0.718. All performance metrics were calculated on held-out test chromosomes — chromosome 3 for rice and Arabidopsis, and chromosome 6 for maize — ensuring rigorous separation between training and evaluation sets.

To determine the optimal input length, we trained SeiPlant using window sizes of 512 bp, 1 kb, 2 kb, and 4 kb across all three species (Additional file 1: Fig. S1). A 1 kb window achieved the best trade-off between predictive accuracy and computational efficiency. Smaller windows underperformed due to limited sequence context, while larger windows increased memory usage and training time without improving performance. Moreover, larger windows inflated the number of positive labels — since labels are assigned per window — leading to higher false positive rates and reduced specificity. This issue was particularly pronounced in large, repeat-rich genomes like maize, where longer windows introduced considerable label noise. These findings support the use of 1 kb as the default input window for all subsequent analyses, balancing biological resolution, labeling specificity, and computational cost.

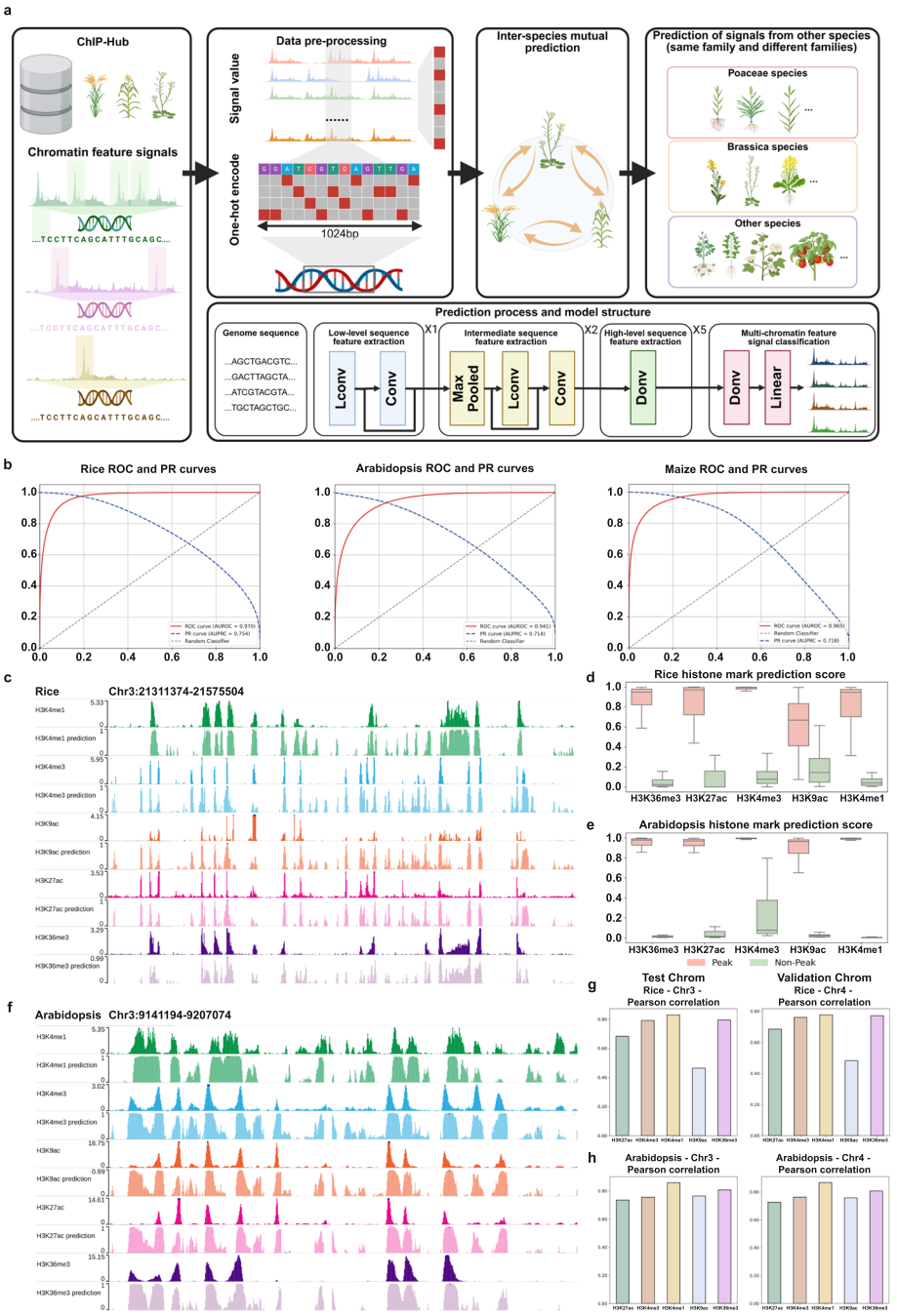

Fig. 1 Framework and performance of species-specific chromatin feature prediction in plants. **a** Overview of the cross-species chromatin feature prediction framework. **b** Receiver operating characteristic (ROC) and precision-recall (PR) curves for species-specific models in Rice, Arabidopsis, and Maize. ROC curves are shown in red; PR curves are shown in blue. **c** Predicted and experimental signal tracks for representative histone marks in Rice (Chr3:21,311,374–21,575,504). **d** The distribution Boxplot of prediction scores for peak and non-peak regions across five histone marks in Rice. **e** The distribution Boxplot of prediction scores for peak and non-peak regions across five histone marks in Arabidopsis. **f** Predicted and experimental signal tracks for representative histone marks in Arabidopsis (Chr3:9,141,194–9,207,074). **g** Pearson correlation coefficients between predicted and experimental signals across test and validation chromosomes for each histone mark in Rice. **h** Pearson correlation coefficients between predicted and experimental signals across test and validation chromosomes for each histone mark in Arabidopsis

Meanwhile, to evaluate the competitiveness of the SeiPlant architecture, we compared its performance with several representative models originally developed for regulatory genomics [9, 10, 12, 31]. Across all three species, SeiPlant consistently outperformed (Additional file 1: Fig. S2), While Enformer excels in human chromatin prediction with long input sequences (196 kb), its performance declined significantly on plant datasets with shorter input windows (1 kb), likely due to domain shift and insufficient inductive bias.

To assess the contribution of core architectural components in SeiPlant, we conducted ablation experiments targeting the three signature modules of the Sei architecture: spline transformation layers, residual connections, and dilated convolutions. We defined three model variants: (i) a No Spline version (with spline layers removed), (ii) a Spline Only version (retaining only spline layers, with residual and dilated convolutions removed), and (iii) the Full Model, which includes all three components. All variants were trained and evaluated under identical settings using 1 kb input windows. The Full Model consistently outperformed both ablated variants across all three plant species, evaluated on held-out chromosomes (Additional file 1: Fig. S3). These results demonstrate that all three modules—spline transformation, residual connections, and dilated convolutions—contribute synergistically to the model's overall accuracy.

To further assess the resolution and biological fidelity of the model predictions, we compared the output signal profiles to ground-truth ChIP-seq tracks from the ChIP-Hub database. Predicted signals in representative genomic intervals closely matched experimentally measured chromatin features in Rice (Chr3:21,311,374–21,575,504; Fig. 1c), Arabidopsis (Chr3:9,141,194–9,207,074; Fig. 1f), and Maize (Chr6:75,059,101–75,148,699; Additional file 1: Fig. S4a). To quantitatively evaluate signal specificity, we further conducted peak-versus-non-peak classification by selecting top-ranked peak regions and pairing them with flanking regions of equal length as negative controls. The models consistently assigned significantly higher scores to peak regions, demonstrating strong discriminatory power across multiple histone modifications, including H3K36me3, H3K27ac, H3K4me3, H3K4me1, and H3K9ac (Fig. 1d, e, and Additional file 1: Fig. S4b).

To explore the latent regulatory information captured by the model, we performed unsupervised clustering on the predicted chromatin profiles of 1,024-bp genomic windows (Additional file 1: Figs. S5b, S6b, S7b). To assess the model's ability to distinguish fine-grained regulatory features beyond histone marks, we also incorporated transcription factor binding site (TFBS) data into this analysis. Using t-SNE for dimensionality reduction, we observed distinct clusters in two-dimensional space, indicating that the model learns biologically meaningful differences in regulatory sequence composition. Enrichment analysis based on ChIP-seq peak annotations further revealed that these clusters corresponded to specific chromatin states: promoter-like clusters were enriched for H3K4me3 and H3K27ac, enhancer-like clusters for H3K4me1 and transcription factor binding, and heterochromatin clusters for CENH3 and H3K9me2 (Additional file 1: Figs. S5a, S6a).

To quantitatively validate these assignments, we used BEDTools to calculate intersection lengths between each cluster and regulatory marker regions, followed by $\log_2$-transformed ratio-of-ratios enrichment scoring with non-positive intersections set to NA. This produced a matrix-form enrichment score table (Additional file 1: Figs. S5c,

S6c, S7c) revealing clear signatures: heterochromatin types were enriched for repressive marks (H3K9me2/3, H3K27me3) and depleted for activating marks, while promoters and enhancers showed the opposite trend. CENH3 was markedly enriched in centromeric heterochromatin, validating cluster accuracy. In rice, H3K27me3 and H3K9me3 patterns distinguished facultative from constitutive heterochromatin; in Arabidopsis, H3K4me1 and H3K4me3 differentiated enhancers from promoters. RNA polymerase binding enrichment also matched expected functional roles. Based on dominant chromatin features and genomic context, we annotated clusters into promoters, enhancers, centromeric heterochromatin, and general heterochromatin. In maize, certain clusters showed marked enrichment for individual transcription factors, enabling finer annotation (Additional file 1: Fig. S7a).

Although promoter–enhancer assignments followed the general rule of higher promotive histone modifications and RNAP binding in promoters, motif analysis showed no strong separation, likely due to their frequent genomic proximity (< 1 kb) and the ~ 200 bp span of promoter motifs, which limits resolution at the 1,024 bp input scale. Smaller windows or alternative modeling may improve separation.

Lastly, we quantified the Pearson correlation between predicted and experimentally measured signal intensities on both test and validation chromosomes (Fig. 1g, h, Additional file 1: Fig. S4c). Most histone marks showed correlation coefficients above 0.7, indicating that the model reliably captures both the presence and quantitative variation of regulatory features across chromatin marks and genomic contexts. Together, these results demonstrate that the clustering and enrichment analyses—supported by strong correlation with experimental data—robustly capture biologically meaningful regulatory categories across species.

Cross-species prediction of histone modifications using species-specific models

In the second phase of our study, we evaluated the cross-species generalization ability of the species-specific models. Specifically, each model trained on one of the three species (Arabidopsis, rice, or maize) was used to predict histone modification signals in the other two species. This setup allowed us to assess how well regulatory features learned from one species can be transferred to phylogenetically distinct plant genomes. As illustrated in Fig. 2a, c, we compared the predicted histone modification signals for rice using models trained on maize and Arabidopsis, and conversely, predicted signals for maize using models trained on rice and Arabidopsis. Visual inspection of selected genomic regions revealed substantial differences in signal recovery, with models trained on the more closely related species generally producing signal profiles more consistent with the ground truth.

To quantitatively evaluate cross-species prediction performance, we assessed multiple metrics, including Pearson correlation between predicted and observed values across test chromosomes, BigWig-based genome-wide signal correlation, peak-region enrichment, and ROC-AUC scores (Fig. 2b, d). For rice targets, the model trained on maize consistently produced more accurate signal profiles than the one trained on Arabidopsis, as evidenced by higher correlation coefficients and improved discrimination between peak and non-peak regions. A similar trend was observed when predicting maize signals:

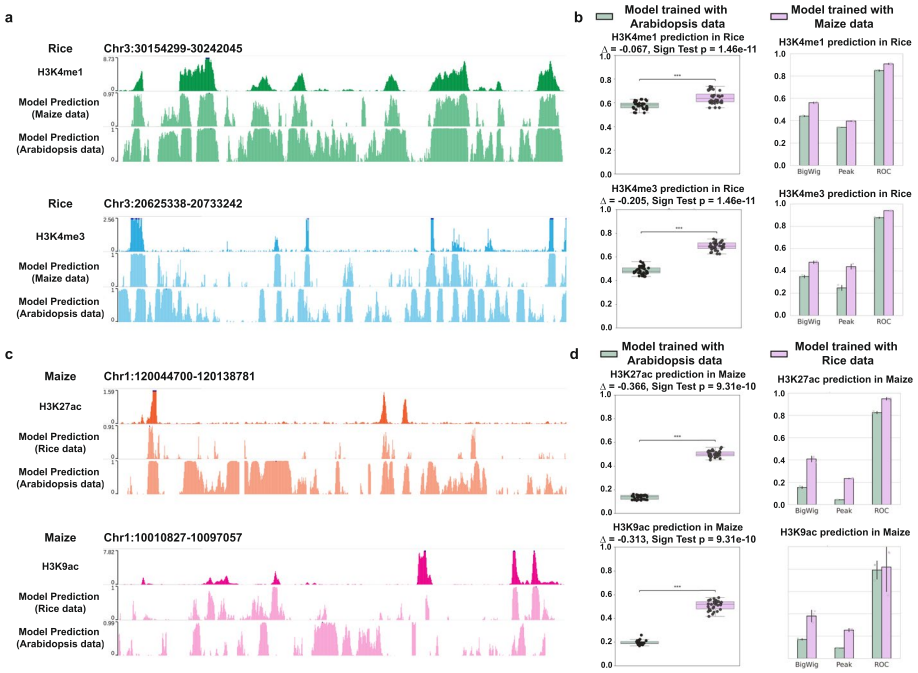

Fig. 2 Cross-species prediction performance of chromatin feature models across plant genomes. **a** Representative genome browser views showing predicted versus observed signal tracks for histone modifications in Rice using cross-species models. **b** Boxplots and barplots showing cross-species prediction scores and evaluation metrics for H3K4me1 and H3K4me3 in Rice using Maize- and Arabidopsis-trained models. **c** Representative genome browser views showing predicted versus observed signal tracks for histone modifications in Maize, using cross-species models. **d** Boxplots and barplots showing cross-species prediction scores and evaluation metrics for H3K27ac and H3K9ac in Maize using Rice- and Arabidopsis-trained models

the rice-trained model more closely reproduced ChIP-seq patterns than the Arabidopsis-trained model.

As a representative case, we assessed the prediction of H3K27ac histone modification signals in maize using models trained on rice and Arabidopsis. We selected independent genomic regions with high-confidence H3K27ac ChIP-seq signals as ground truth. The rice-trained model achieved a maximum Pearson correlation of 0.52 and a BigWig correlation of 0.42 with the observed signals, while the Arabidopsis-trained model showed markedly lower values of 0.15 and 0.16, respectively. Similarly, we evaluated the prediction of H3K4me3 signals in rice using models trained on maize and Arabidopsis. Across three representative genomic regions, the maize-trained model achieved a maximum Pearson correlation of 0.72 and a BigWig correlation of 0.49, whereas the Arabidopsis-trained model achieved lower correlations of 0.52 and 0.36, respectively.

To further validate the generality of this trend across additional histone marks, we extended our cross-species prediction experiments to include three more modifications. For each histone mark, we trained models on Arabidopsis, rice, and maize data separately and evaluated their prediction performance on the remaining species. The full results are shown in Additional file 1: Fig. S8. Consistent with the findings for H3K4me3 and H3K27ac, we observed that for all five marks, the species-specific models trained on closely related species (i.e., within the same family) consistently outperformed those trained on more distant species. For instance, in the prediction of rice histone

marks, maize-trained models yielded significantly higher signal correlation and ROC-AUC scores compared to Arabidopsis-trained models across all five marks. Likewise, for maize predictions, rice-trained models performed better than Arabidopsis-trained models.

These results consistently demonstrate the substantial advantage of training on phylogenetically closer species, reinforcing the notion that epigenomic signatures learned from one plant species are more readily transferable to another species within the same family, while highlighting the limited transferability of regulatory models across more distantly related plant taxa.

Family-level cross-species prediction with group-trained models

Building on the observation that models trained on phylogenetically closer species exhibit superior predictive performance, we further explored the potential of family-level model generalization. To this end, we constructed a Poaceae family-level model by jointly training on filtered data from rice and maize. In contrast, due to data availability, we trained a model solely on Arabidopsis and referred to it as the Arabidopsis-trained model throughout this study. While not a true family-level model, it was nonetheless evaluated on related species within the Brassicaceae family to investigate the extent to which regulatory syntax learned from A. thaliana can generalize across closely related dicots.

As shown in Fig. 3a, the Poaceae-trained model was used to predict chromatin features in other Poaceae family species, including *Setaria italica* (*S. italica*), *Sorghum bicolor* (*S. bicolor*), *Brachypodium distachyon* (*B. distachyon*), and *Hordeum vulgare* (*H. vulgare*). Similarly, the Arabidopsis-trained model was evaluated on related species such as *Brassica napus* (*B. napus*), *Arabidopsis alpina* (*A. alpina*), *Brassica rapa* (*B. rapa*), and *Arabidopsis lyrata* (*A. lyrata*). Figure 3b presents the comparison of predicted signal values with true ChIP-seq signals (such as H3K4me1 and H3K4me3) within fixed chromosomal regions across these species. The visualization demonstrates high consistency between the predicted and true signals, particularly in specific chromosomal regions, confirming the model's robustness in capturing chromatin feature patterns.

To further assess the sensitivity of the family-level models to peak and non-peak regions, we quantified the signal values for both types of regions. Figure 3c shows the prediction results for peak and non-peak regions, with the models consistently assigning higher prediction scores to peak regions, reflecting their ability to accurately capture regulatory signals. In Fig. 3d, in silico mutagenesis of important sequence motif (ISM) sites revealed that clipping mutations within peak regions had a significantly higher impact on the predicted signal values compared to mutations in non-peak regions. This further highlights the model's sensitivity to regulatory features in peak regions. Figure 3e quantifies the correlation between predicted and true signal values, including Pearson correlation, BigWig correlation, and ROC-AUC values, showing that family-level models effectively capture chromatin modification signals for species within the same family.

To further verify the robustness of our conclusions across multiple histone modifications, in addition to H3K4me3, we evaluated other histone modifications (H3K4me1, H3K9ac, H3K36me3, and H3K27ac) across representative species in each family (Additional file 1: Figs. S9, S10). To strengthen the evaluation, we further

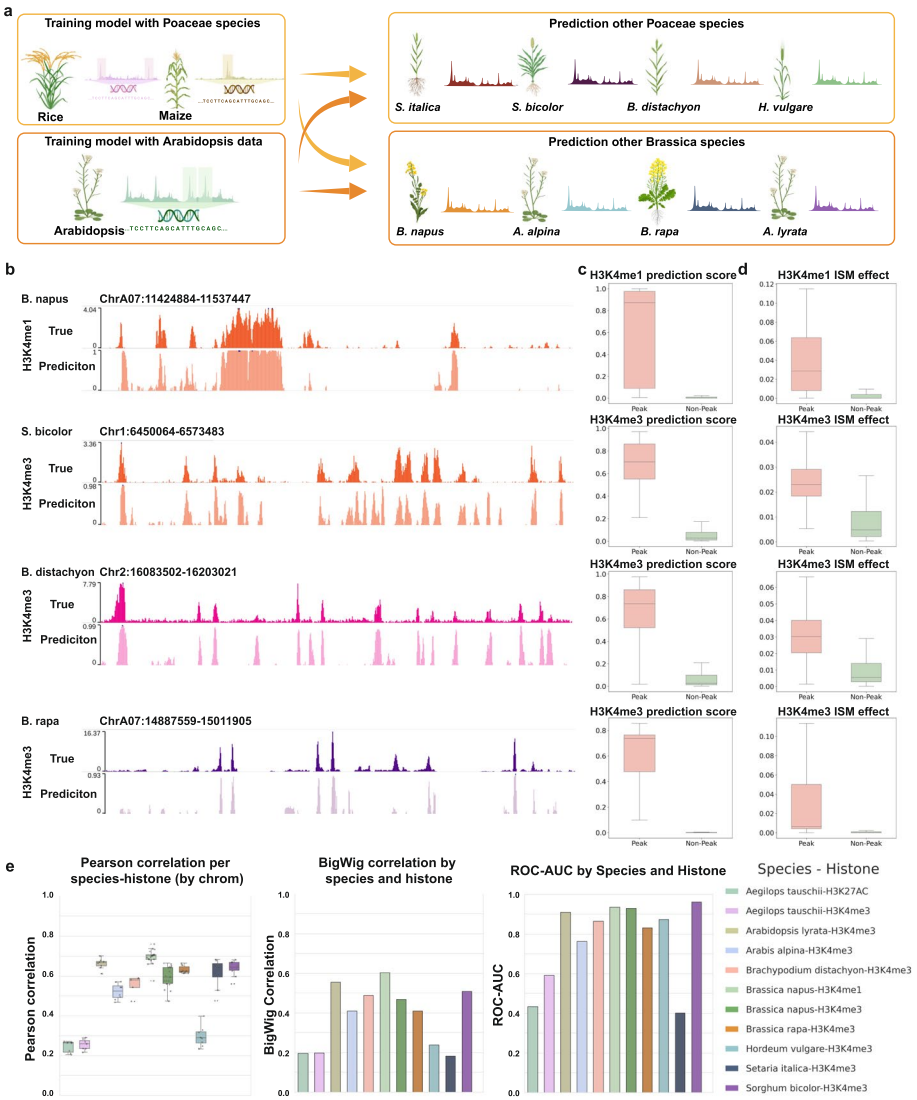

Fig. 3 Family-level models improve cross-species prediction of chromatin features. **a** Schematic illustration of family-level training and prediction: the Poaceae model trained on Rice and Maize was used to predict signals in other Poaceae species; the Arabidopsis-trained model was used to predict signals in other Brassicaceae species. **b** Comparison of predicted and observed histone modification signals across representative genomic intervals in multiple Poaceae and Brassicaceae species. **c** Boxplots showing prediction scores for peak and non-peak regions across three histone modifications. **d** Boxplots showing ISM-based Δ scores for peak and non-peak regions across three histone modifications. **e** Quantitative evaluation of family-level prediction performance using Pearson correlation, BigWig correlation, and ROC-AUC for each species-histone pair

included *Aegilops tauschii* (*A. tauschii*), a species with a large and complex genome that poses higher prediction challenges. The Poaceae-trained model achieved high concordance between predicted and true signals for different chromatin modifications, supporting the strong transferability of family-level models. We performed quantitative metric evaluations for these histone modifications, as summarized in Additional file 1: Fig. S11, the models trained within the same family exhibited markedly higher predictive accuracy than those trained on distantly related species.

Additionally, we explored the cross-prediction ability of the Poaceae model and Arabidopsis-trained model by using them to predict each other's species signals (Fig. 4). Compared to within-family models, cross-family models led to false positive peak regions and noise in non-peak regions (Fig. 4a). Moreover, the performance of within-family models significantly outperformed cross-family models across different chromosomes (Fig. 4b, c). To further verify the uncertainty of cross-family prediction compared to within-family prediction, we supplemented prediction experiments for the same histone modifications used above (Additional file 1: Figs. S12, S13). The results consistently showed that cross-family models produced lower correlation values and more pronounced noise in predicted signal profiles, reinforcing the importance of phylogenetic proximity in model transferability.

Finally, to evaluate the impact of including multiple species from the same family, we compared the performance of models trained on rice + maize (Poaceae model) with those trained on data from only rice or maize. As shown in Fig. 4d, the inclusion of both species significantly improved the model's performance across various Poaceae species, such as *S. italica*, *S. bicolor*, *B. distachyon*, and *H. vulgare*.

In summary, our results demonstrate that family-level models, trained on phylogenetically related species, significantly enhance the accuracy and generalizability of cross-species chromatin feature prediction, particularly for histone modifications.

Assessment of cross-family models for chromatin feature prediction across diverse plant families

To assess whether cross-family models could further enhance predictive performance, we designed the experiment shown in Fig. 5a, where we trained a model using filtered signal data from rice, maize, and Arabidopsis to predict chromatin feature signals in both Poaceae and Brassicaceae species, as well as species from other plant families. This multi-family model aimed to determine if incorporating species from multiple families would improve the model's cross-species prediction capability.

The cross-family model was trained using a combined dataset of chromatin signals from rice, maize, and Arabidopsis, and was then applied to predict histone modification signals in species from both the Poaceae and Brassicaceae families, including *S. italica*, *S. bicolor*, *B. distachyon*, *H. vulgare*, *B. napus*, *A. alpina*, and *A. lyrata*.

We first evaluated the performance of the cross-family model relative to family-specific models. As shown in Fig. 5b, the cross-family model did not consistently outperform the family-specific models. Notably, it showed improved performance in predicting H3K4me3 signals in *B. napus* and *A. lyrata*, but not in other histone marks or species. In contrast, within the Poaceae family, the cross-family model exhibited no performance advantage over models trained exclusively on related species. These results suggest that while cross-family training may offer limited improvements in select cases, especially for specific marks or species, family-specific models generally provide more reliable predictions across most plant taxa evaluated.

We then evaluated the performance of the cross-family model on species from other plant families (Fig. 6a, b), including *Glycine max* (*G. max*, Fabaceae), *Gossypium arboreum* (*G. arboreum*, Malvaceae), *Solanum lycopersicum* (*S. lycopersicum*, Solanaceae), and *Kalanchoe pinnata* (*K. pinnata*, Crassulaceae). For histone modification

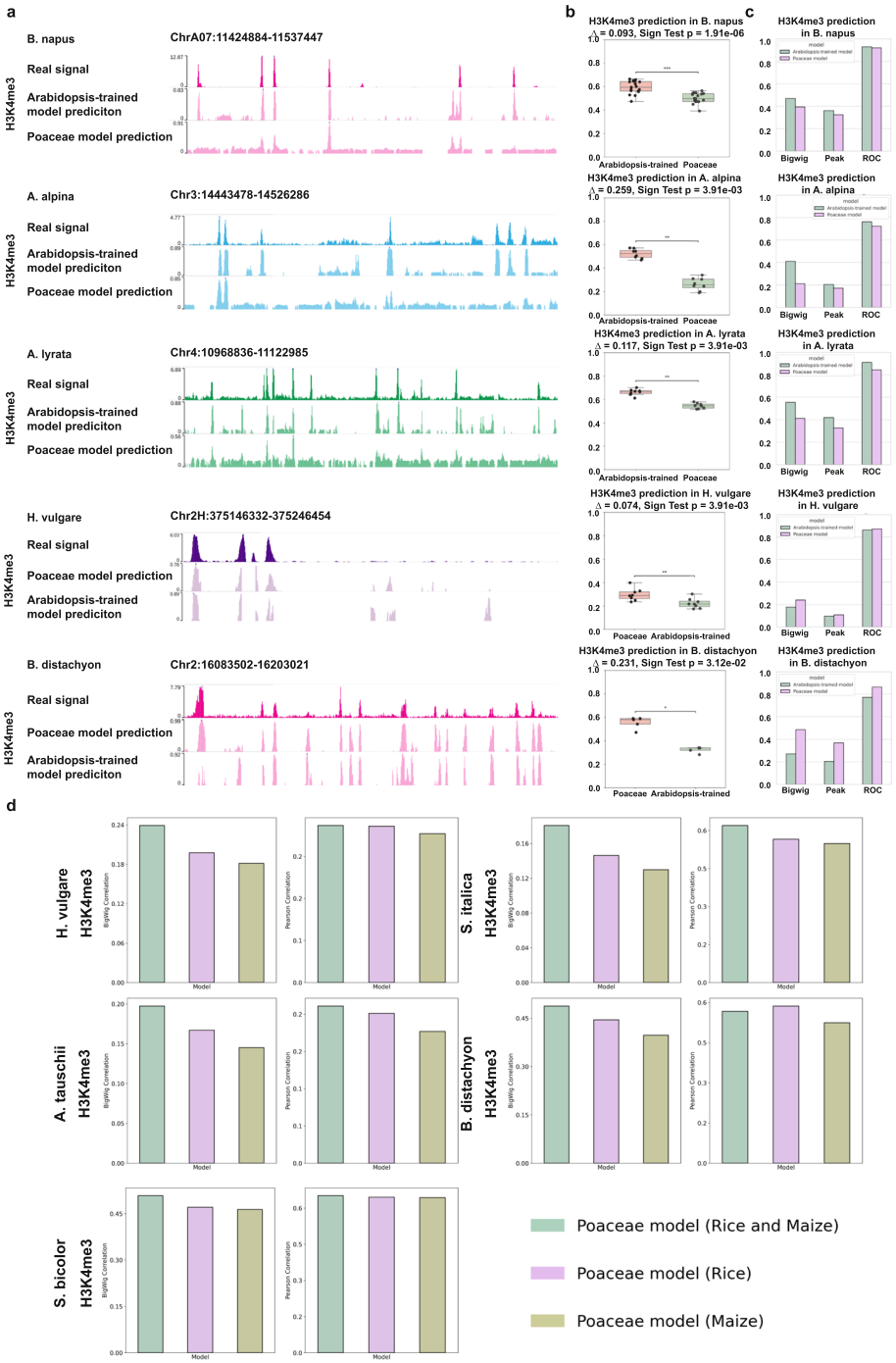

Fig. 4 Family-level training improves accuracy and mitigates noise in cross-species prediction. **a** Comparison of predicted H3K4me3 signal profiles in five species using different models (Poaceae and Arabidopsis-trained). Each row shows the real ChIP-seq signal, predictions from the Arabidopsis-trained model, and predictions from the Poaceae model across fixed chromosomal intervals. **b** Boxplots of prediction scores for H3K4me3 across species, comparing Arabidopsis-trained and Poaceae models. Significance was assessed using the Sign test. **c** Bar charts showing BigWig correlation, peak enrichment score, and ROC-AUC for each species, comparing predictions from Arabidopsis-trained and Poaceae models. **d** Bar charts comparing model performance across Poaceae species using models trained on Rice only, Maize only, or both combined. Metrics include peak score and average coverage for H3K4me3

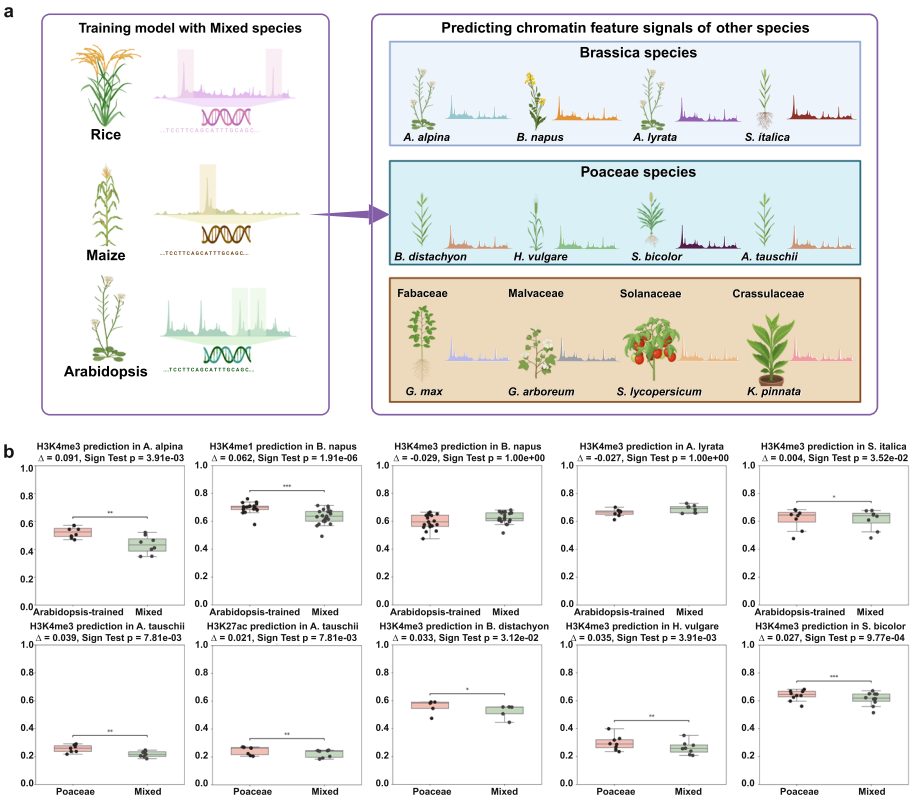

Fig. 5 Evaluating cross-family models for chromatin feature prediction across diverse plant lineages. **a** Schematic of the cross-family prediction experiment. A multi-species model was trained on Rice, Maize, and Arabidopsis, and used to predict chromatin features in species from Brassicaceae (e.g., *B. napus*, *A. alpina*), Poaceae (e.g., *B. distachyon*, *S. bicolor*), and other families (e.g., *G. max*, *S. lycopersicum*). **b** Boxplots comparing family-specific models and the mixed cross-family model across multiple species. Each panel shows prediction scores for a histone mark (mainly H3K4me3 or H3K27ac) in a target species, evaluated using the Sign test. Predictions from the mixed model are shown alongside those from Poaceae or Arabidopsis-trained models

signals such as H3K4me3, H3K4me1, H3K9ac, and H3K27ac, the predicted signals from the cross-family model showed strong correlation with the true signal regions in these species. These results indicate that the cross-family model retains a certain degree of predictive ability for chromatin features even in species from distantly related families. Although the predictive performance of the cross-family model is weaker compared to family-specific models, it still demonstrates a certain ability to accurately capture and localize key signal regions.

Model consistency in duplicated genomic regions suggests cis-regulatory dominance

To further investigate whether cis-regulatory sequence elements alone are sufficient to drive histone modification patterns, we examined model prediction consistency across duplicated genomic regions within individual plant genomes. Specifically, we identified pairs of homologous regions with high sequence similarity through BLASTN-based self-alignment within the genomes of rice and Arabidopsis. For each duplicated region pair, we used the corresponding within-species model to generate histone modification predictions, and assessed the similarity between the predicted

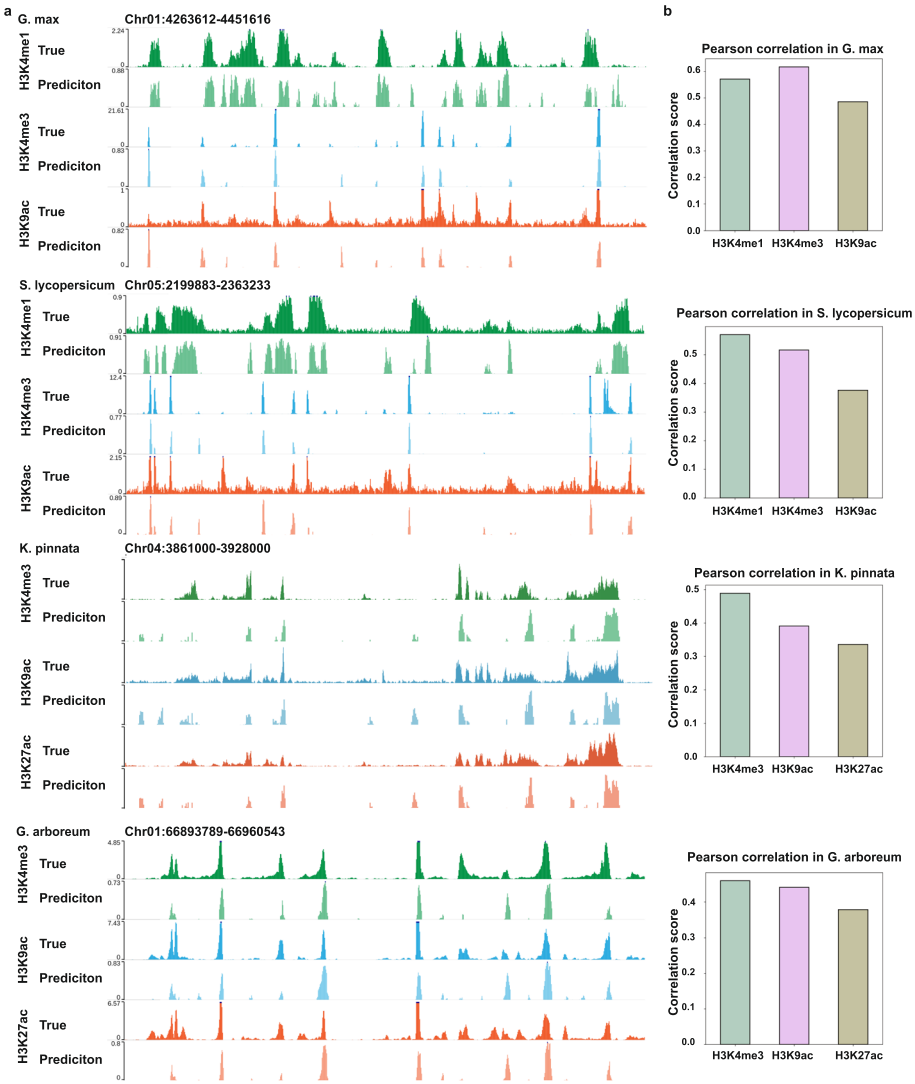

Fig. 6 Prediction of chromatin features in non-training lineages using the cross-family model. **a** Visualization of predicted versus true ChIP-seq signal tracks in non-training species using the cross-family model. **b** Bar plots showing total Pearson correlation between predicted and observed signals for each histone modification

signals using three metrics: average cosine similarity, 1 minus mean squared error (MSE), and Pearson correlation coefficient. To assess whether this pattern extends beyond individual species, we applied the same analysis to several other species in the Brassicaceae family and the Poaceae family. The results consistently showed high prediction similarity across duplicated pairs, indicating that the model makes similar predictions for homologous sequences regardless of their chromosomal location or surrounding chromatin context. Across all tested species, predictions on duplicated regions remained highly consistent (Additional file 1: Fig. S14), with average cosine similarity exceeding 0.95 in most cases. These findings strongly suggest that the model relies primarily on cis-encoded sequence features to determine histone modification profiles, and is minimally influenced by positional context or broader

epigenomic background. This supports the central hypothesis that local sequence features are the primary determinants of histone mark recruitment.

Model usability and efficient data output generation

To improve model usability, we developed an efficient and user-friendly workflow. As illustrated in Fig. 7, users can simply provide the genome sequence file as input, and the model processes it to generate histone modification signal outputs in both BedGraph and BigWig formats. This streamlined, one-click pipeline facilitates the rapid generation of chromatin signal tracks from raw sequence data, allowing for easy integration into downstream analyses. The intuitive interface and accessible output formats significantly reduce the computational burden, enabling researchers to focus on biological interpretation. Importantly, this approach is particularly valuable for species in which chromatin features are difficult or costly to measure experimentally, providing a practical solution for generating epigenomic signal files for rare or under-studied plant species.

Discussion

In this study, we developed a deep learning framework for predicting histone modifications across diverse plant species using the Sei architecture—a multi-scale convolutional network originally designed for regulatory sequence modeling—and implemented a streamlined user pipeline. This pipeline enables direct generation of genome browser–compatible visualization files (BigWig and BedGraph) from raw FASTA sequences, facilitating efficient and interpretable chromatin feature track generation, especially for species lacking extensive experimental epigenomic data.

Species-specific models trained on Arabidopsis, rice, and maize achieved strong within-species performance (AUROC > 0.94, AUPRC > 0.71), with predicted profiles closely matching ChIP-seq tracks and showing strong enrichment in peak regions,

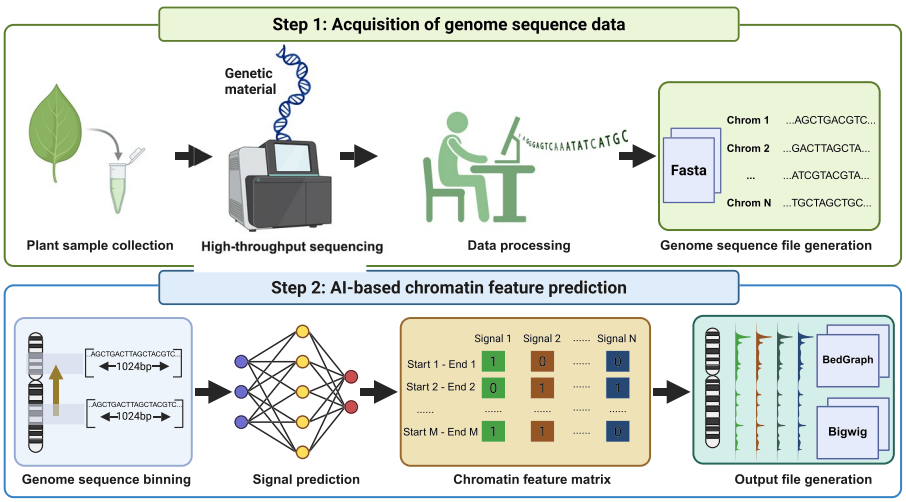

Fig. 7 User-friendly pipeline for chromatin signal prediction from genomic sequences. In Step 1, plant samples are collected and sequenced using next-generation sequencing (NGS) to obtain genomic data, which is processed into a standardized FASTA file containing chromosome-level sequences. In Step 2, the genome is binned into fixed-length intervals, and the AI-based prediction model is applied to infer the presence and intensity of histone modification signals

consistent with prior findings in human and animal systems [12, 21, 24, 32]. However, model generalization declined with increasing phylogenetic distance. Models trained on monocots (e.g., maize) performed poorly when applied to dicots (e.g., Arabidopsis), and vice versa, highlighting the evolutionary divergence in regulatory syntax between lineages. This phylogeny-dependent loss of accuracy mirrors observations from cross-species chromatin accessibility studies [33], underscoring the challenge of transferring regulatory models across major taxonomic boundaries.

To mitigate this limitation, we trained a family-level model for Poaceae by integrating data from rice and maize, and a model solely on Arabidopsis, which was subsequently evaluated across related species within the family. These models significantly improved prediction accuracy in related species, showing higher correlation with experimental data, better peak resolution, and increased robustness. ISM experiments further validated the functional importance of sequence features learned by the model, as perturbations within peak regions caused notable drops in predicted signal strength.

In contrast, cross-family models trained on mixed-lineage datasets (e.g., Poaceae + Brassicaceae) yielded inconsistent results. While some conservation was observed — for instance, H3K4me3 predictions in B. napus and A. lyrate — overall performance lagged behind family-specific models. Interestingly, this improved performance may reflect the relatively high conservation of promoter-associated chromatin features within the Brassicaceae family. This likely reflects the higher conservation of promoter-associated chromatin features within closely related dicots [34, 35], highlighting the importance of phylogenetically informed training for high-fidelity cross-species prediction.

While our model is designed to capture sequence-encoded cis-regulatory patterns, histone modification recruitment is also shaped by trans-acting influences such as transcription factors, chromatin remodelers, non-coding RNAs, and histone-modifying complexes. These components direct histone mark deposition through protein–DNA interactions, chromatin context, and higher-order nuclear organization. Consequently, our predictions reflect the intrinsic cis-regulatory potential of genomic sequences but does not account for the full spectrum of dynamic, context-dependent regulation mediated by trans-acting components.

In addition, the ChIP-seq datasets used for training span multiple developmental stages and tissue types. To maximize robustness, we intentionally aggregated high-quality datasets across diverse developmental contexts. This strategy avoids overfitting to stage-specific artifacts and better reflects the biological variation encountered in cross-species prediction, albeit at the cost of removing stage- or tissue-specific resolution. While this aggregation inevitably masks developmental heterogeneity, our binary presence/absence modeling of histone modifications mitigates this limitation by emphasizing stable and evolutionarily conserved regulatory features. Genomic positions of histone marks such as promoter H3K4me3 or Polycomb-targeted H3K27me3 domains tend to remain conserved even when their quantitative levels fluctuate across developmental stages. Consequently, our models reflect a developmentally averaged regulatory landscape and may not capture dynamic chromatin changes during key developmental transitions such as seed germination or floral induction. Similarly, the Brassicaceae model was trained solely on *A. thaliana*, limiting its capacity to learn family-level regulatory features. Incorporating multiple Brassicaceae

species—particularly those occupying different phylogenetic positions—would expand the diversity of cis-regulatory sequences and epigenomic patterns, helping disentangle conserved family-wide signals from species-specific noise and potentially narrowing the performance gap with the Poaceae-trained model.

Beyond these biological and dataset constraints, our approach does not integrate additional regulatory layers such as chromatin accessibility (ATAC-seq) [36, 37], transcriptomic output (e.g., RNA-seq), or three-dimensional chromatin architecture (e.g., Hi-C) [38]. Furthermore, the reliance on ChIP-Hub datasets—though high in quality—restricts training to a limited set of histone marks and species with available ChIP-seq data, potentially biasing generalization performance [27]. The absence of trans-acting regulatory information further limits the biological completeness of our predictions. Integrating data such as transcription factor binding, chromatin loops, or chromatin remodeler occupancy may provide a more mechanistically grounded understanding of histone modification patterns, particularly in developmentally or environmentally responsive contexts.

To improve upon these limitations, future work should explore (i) transformer-based architectures capable of modeling long-range dependencies and integrating prior biological knowledge [23, 39, 40], (ii) domain adaptation techniques that explicitly account for species differences, and (iii) multimodal models incorporating transcriptomic and chromatin accessibility data. In particular, future models that explicitly represent both cis-encoded sequences and trans-acting factors could yield more comprehensive and biologically interpretable predictions of epigenomic landscapes across plant species. Beyond architectural innovations, expanding the taxonomic breadth and depth of training datasets—especially for under-represented and economically important crops—will be critical to building more inclusive and generalizable models. In addition, stage-aware or multi-condition modeling strategies could be developed to explicitly disentangle stage-specific chromatin dynamics, particularly for highly dynamic marks such as H3K27ac. Such approaches would be especially informative when comparing closely related species, where quantitative differences across developmental stages may reflect meaningful regulatory divergence.

Conclusions

This study demonstrates that phylogenetically informed training substantially enhances the cross-species prediction of histone modifications in plants. By jointly modeling multi-species epigenomic profiles within an explicit evolutionary framework, our approach improves generalization to divergent and data-poor species relative to conventional single-species models. These results indicate that phylogeny can serve as an effective inductive bias for regulatory genomics in plants and provide a scalable computational strategy for functional annotation in non-model and agriculturally important species.

Methods

Training label curation and genome-wide windowing

To systematically generate high-confidence and reproducible training labels across diverse plant genomes, we developed a multi-step strategy that integrates multiple ChIP-seq datasets per histone modification, aggregates regulatory signals, and assigns window-level scores based on overlap and signal consistency.

Integration of ChIP-seq datasets

We collected histone modification and transcription factor binding site (TFBS) data for rice, maize, and Arabidopsis from the ChIP-Hub platform, which hosts large-scale, curated chromatin profiling datasets across multiple plant species. For each chromatin feature (e.g., H3K4me3, TFBS), we integrated peak regions from multiple independent ChIP-seq experiments corresponding to the same feature. For each chromatin feature, we gathered all available narrowPeak BED files corresponding to independent ChIP-seq experiments from distinct SRA accessions. To ensure robustness and mitigate dataset-specific artifacts, only BED files with metadata-supported high-confidence peak calls were retained. For each histone mark within a species, we merged the retained narrowPeak files using bedtools multiinter to identify overlapping peak regions. Genomic intervals observed in at least two independent experiments were retained as consensus peaks, representing biologically consistent and reproducible chromatin features.

To further quantify the reliability of each consensus peak, we computed an overlap frequency score:

> If all retained intervals had the same overlap count (i.e., all with exactly two datasets), a uniform confidence score of 1 was assigned.
> Otherwise, each interval's overlap count was normalized by the total number of overlapping intervals across all peak files and linearly rescaled to a confidence score in the range [0.1,1.0], reflecting reproducibility strength.

For cross-species model evaluation, we focused on five histone modifications with relatively large sample sizes and narrow peak profiles in the ChIP-Hub dataset: H3K4me1, H3K4me3, H3K9ac, H3K27ac, and H3K36me3. To further refine signal confidence, we scored genomic regions based on their overlap frequency across experiments. If all retained regions had exactly two overlaps, they were uniformly assigned a score of 1. Otherwise, we first normalized each region's overlap count by the total overlap count across all regions, then ranked and linearly mapped the normalized values to a predefined confidence score range (e.g., 0.1 to 1). This final score reflects the consistency of a given region across multiple experimental replicates, with higher scores indicating greater reproducibility and confidence.

Genome windowing and label generation

After obtaining high-confidence chromatin feature intervals, we loaded the corresponding species' reference genome sequences. To assign training labels from chromatin features across the genome, we first defined a sliding window approach over the reference genome sequence. To scan the genome in a uniform and non-overlapping fashion, we first defined a sliding window approach over the input genome G. Each window w_j spans a 1024-bp interval and slides with a step size of 512 bp. The complete set of genomic windows is defined by:

$$W = \{w_j = [a_j,\ a_j + L) \mid a_j = j \cdot S,\ j = 0,\ 1,\ \ldots,\ N\} \tag{1}$$

where $L = 1024$ bp and $S = 512$ bp represent the window size and step size, respectively, and a_j is the starting position of window w_j.

To assign a chromatin activity score to each window, we quantified the total contribution of all overlapping chromatin features derived from BED annotations. Each region $(s_i, e_i, c_i) \in R$ includes start and end positions s_i, e_i and a confidence score $c_i \in [0, 1]$. The contribution of region i to window w_j is proportional to the fraction of overlap and its confidence score. The aggregated raw score for window w_j is calculated as:

$$\text{Score}(w_j) = \sum_{(s_i,\ e_i,\ c_i)\in\mathcal{R}} \frac{\left|w_j \cap [s_i,\ e_i)\right|}{L} \cdot c_i \tag{2}$$

This formula (2) reflects the average per-base contribution of overlapping chromatin signals to the window and effectively integrates signal intensity and genomic span.

Binary label generation and thresholding

For species-specific classification tasks, we generated binary labels for each chromatin feature k. A window was labeled as positive if it overlapped any annotated region associated with that feature. Formally, the binary label $y_j^{(k)}$ for window w_j is defined as:

$$y_j^{(k)} = \begin{cases} 1, \text{ if } \exists (s_i, e_i) \in R_k \text{ such that } w_j \cap [s_i, e_i) \neq \emptyset \\ 0, \text{ otherwise} \end{cases} \tag{3}$$

where R_k denotes the set of genomic intervals associated with histone mark k. In cross-species prediction tasks where regression on signal strength is preferred, we normalized the aggregated score for each window to the [0,1] range. To avoid over-smoothing and to retain interpretability, we rounded each window score to one decimal place and clipped values above 1.0:

$$\hat{y}_j = \min\left(\text{Round}(\text{Score}(w_j),\ 1),\ 1.0\right) \tag{4}$$

To further suppress noisy or marginal signals that may arise due to weak overlap or low-confidence annotations, we applied a score threshold. Windows with normalized scores below 0.1 were set to zero:

$$\hat{y}_j = \begin{cases} \hat{y}_j, \text{ if } \hat{y}_j \geq 0.1 \\ 0, \text{ otherwise} \end{cases} \tag{5}$$

Only windows with non-zero final scores were retained as training candidates in cross-species prediction settings, ensuring that the model focused on informative regions with meaningful regulatory signals.

Model architecture

The SeiPlant model is adapted from the original Sei architecture for multi-label chromatin feature prediction. The input is a one-hot encoded 1,024-bp DNA sequence. The model begins with three hierarchical convolutional stages:

The first stage uses two 1D convolutional layers with 480 channels (kernel size 9, padding 4), followed by a residual block with two additional convolutions (480 channels) and ReLU activations. The second stage applies max pooling (kernel size 4), dropout (rate 0.2), and two convolutional layers with 640 channels. A residual block with the same structure follows. The third stage repeats this pattern with 960 channels. Skip

connections add the output of the local convolutions to the corresponding residual blocks in each stage.

After feature extraction, the model applies five dilated convolutional layers with dilation rates of 2, 4, 8, 16, and 25. Each uses 960 channels, kernel size 5, dropout rate 0.1, and ReLU activation. Residual summation is performed at each step.

The output is passed through a learnable B-spline transformation layer that compresses the temporal dimension into 16 spline basis functions. The flattened result is fed into a two-layer fully connected head: a linear layer with ReLU activation followed by another linear layer and sigmoid activation. The output dimension matches the number of target chromatin features.

Training strategy

SeiPlant was trained using mini-batch gradient descent with the Adam optimizer. The learning rate was set to 1e-5. For classification tasks, the loss function was binary cross-entropy (BCELoss); for regression tasks, mean squared error (MSELoss) was used. Chromosome data for each species were divided into non-overlapping training, validation, and test sets to ensure the independence and generalisation ability of model evaluation. For rice and Arabidopsis, chromosome 3 was used as the test set, chromosome 4 as the validation set, and the remaining chromosomes were used for model training. For maize, chromosome 6 is used as the test set, chromosome 8 as the validation set, and the remaining chromosomes for training.

Genome-wide signal prediction and file generation

To enable genome-wide regulatory signal prediction in non-training species, we implemented a three-step pipeline to generate standardized output files in BedGraph and BigWig formats, facilitating downstream visualization and analysis.

Step 1: genome sequence fragmentation

Following model training, we performed genome-wide inference by first generating candidate sequence fragments from the reference genome of the target species. A sliding window approach was applied across each chromosome using a fixed window size (e.g., 1,024 bp) and step size (e.g., 128 bp). Based on the chromosome size file, genomic intervals were systematically extracted and filtered to retain only those containing standard nucleotides (A, T, C, G). The corresponding genomic coordinates were saved in BED format, and the nucleotide sequences were exported in FASTA format to serve as input for downstream prediction.

Step 2: model-based signal inference

Using the pre-trained Sei model, we conducted inference on the unlabeled genomic fragments generated in Step 1. For each input FASTA sequence, the model performs forward propagation to produce multi-label probability scores corresponding to predefined histone modifications. These predicted signal values were saved in NumPy (.npy) format, with each score array aligned to the input window and histone mark.

Step 3: BedGraph and BigWig file assembly

To minimize edge effects caused by incomplete receptive field coverage near sequence boundaries, only the central high-confidence region of each window was retained. Specifically, for a 1,024 bp window, the midpoint lies at 512 bp, and the central region was defined as ± 64 bp around this midpoint, corresponding to positions 448–576 bp within the window. This range equals the sliding step size (128 bp), ensuring that consecutive retained segments from adjacent windows align seamlessly without gaps.

After extracting central-region predictions for all windows, we paired these scores with their corresponding genomic coordinates to generate per-label BedGraph files. Weak or background signals (< 0.01) were set to zero, and the remaining values were min–max normalized to the range 0.1–1.0 while preserving relative signal magnitude. Finally, BedGraph files were converted to BigWig format using the UCSC bedGraphToBigWig utility for efficient storage and visualization.

Clustering and quantitative enrichment analysis of predicted chromatin features

For each target genome, we first segmented the genome into fixed-length intervals and obtained the predicted chromatin feature values for each segment from the trained model. If the genome contained M segments and each segment had N chromatin features, the resulting dataset was structured as an $M \times N$ feature matrix, where rows correspond to genomic segments and columns correspond to individual chromatin features.

This high-dimensional feature matrix was reduced to two dimensions using t-distributed Stochastic Neighbor Embedding (t-SNE) to preserve local similarities between genomic segments while enabling visualization. The Louvain community detection algorithm was then applied to the t-SNE projection to partition the segments into discrete clusters, which served as the basis for downstream functional enrichment analysis.

To quantitatively validate the functional annotations assigned to the t-SNE–derived clusters, we computed enrichment scores for each chromatin mark or transcription factor binding site within each cluster. Genomic intervals corresponding to each cluster were intersected with the genomic regions of interest using BEDTools multiinter, and the overlap lengths were recorded. Enrichment was then quantified using a log-transformed ratio-to-ratio metric:

$$E = \log_2 \left(\frac{\frac{O_c}{L_c}}{\frac{O_g}{L_g}} \right) \tag{6}$$

where E means Enrichment score, O_c means the overlap length in cluster (bp), L_c means the total length of the cluster (bp), O_g means the genome-wide overlap length (bp), L_g means the total genome length (bp). Positive enrichment scores indicate that the feature is overrepresented in the cluster relative to genome-wide background levels, whereas negative scores indicate depletion. Clusters with zero or non-positive overlap lengths were assigned NA to avoid spurious enrichment calls. The resulting enrichment score matrix across all clusters and features was visualized as heatmaps in the supplementary figures, providing quantitative support for cluster annotation and functional interpretation.

Identification of duplicated genomic regions via BLASTN

To identify duplicated genomic regions within individual species for prediction consistency analysis, we first constructed a nucleotide BLAST database from each species' genome using makeblastdb. Self-alignment was then performed using blastn, with parameters optimized for high-confidence local alignments (-evalue 1e-10, -perc_identity 90, -outfmt 6). BLASTN results were filtered to retain alignments $\geq$ 1024 bp in length and involving non-identical coordinates on the same chromosome. From each retained pair, 1024 bp segments centered within the aligned regions were extracted and converted into model-compatible input formats using custom preprocessing scripts. Histone modification predictions were then generated using the trained models, and prediction similarity between duplicated regions was evaluated using average cosine similarity, mean squared error (MSE), and Pearson correlation coefficient.

Implementation of baseline models for comparison

To benchmark the performance of SeiPlant, we implemented and trained several widely used regulatory sequence prediction models under consistent settings. All baseline models were trained independently for each species using the same datasets and window sizes (1,024 bp), and their outputs were compared to SeiPlant predictions in terms of AUROC and AUPRC.

SVM-based

To implement a classical machine learning baseline, we trained a support vector machine (SVM) classifier using k-mer-based sequence features, following the general design of QHistone. Each 1,024 bp one-hot encoded input sequence was flattened into a binary feature vector, effectively capturing the presence or absence of nucleotide patterns. We employed a one-vs-rest strategy using LogisticRegression with the liblinear solver to handle multilabel classification. Model training was conducted using scikit-learn with 8 CPU threads and a maximum of 500 iterations. Evaluation was performed on a held-out test chromosome using macro- and micro-averaged AUROC and AUPRC metrics, and only labels with non-trivial binary distributions were retained for scoring.

Basenji

This is a deep convolutional neural network designed to predict quantitative epigenomic profiles across long genomic regions using DNA sequence as input. In this study, we re-implemented a modified version of Basenji in PyTorch to suit plant genome datasets with shorter sequence lengths. Our implementation accepts 1,024 bp one-hot encoded DNA sequences and outputs multi-label chromatin feature predictions. The architecture consists of three convolutional blocks with GELU activations and max-pooling, followed by six layers of dilated convolutions to capture long-range dependencies. The final outputs are obtained via a fully connected layer with sigmoid activation for each target label. Training was performed using the Adam optimizer (learning rate 1e-5) with a batch size of 64 for up to 100 epochs, and binary cross-entropy loss was used as the training objective. We applied early stopping with a patience of 20 epochs, based on validation loss. During inference, per-label scores were obtained by applying max-pooling across the sequence dimension. Evaluation was conducted on a held-out chromosome

using macro- and micro-averaged AUROC and AUPRC metrics. All training, validation, and testing were performed on the same data splits as used for SeiPlant. Implementation details were adapted from the original Basenji publication [10] and its official GitHub repository (https://github.com/calico/basenji). For reproducibility, the modified PyTorch implementation used in this study is available (https://github.com/goodarzilab/Basenji-Pytorch).

DanQ

DanQ is a hybrid neural network architecture that combines convolutional neural networks (CNNs) for motif detection and bidirectional long short-term memory networks (BiLSTMs) for capturing long-range dependencies between sequence motifs. In our benchmark, we implemented DanQ in PyTorch using 1,024 bp one-hot encoded DNA sequences as input. The architecture consisted of one 1D convolutional layer (320 filters, kernel size = 19, padding = 9) followed by a max pooling layer (kernel size = 13, stride = 13) and dropout ($p = 0.2$). The pooled sequence output was then processed by a two-layer bidirectional LSTM (each direction: 320 hidden units), whose outputs were flattened and passed through a fully connected layer (925 hidden units, ReLU activation, dropout $p = 0.5$) and a final sigmoid output layer for multi-label classification. We trained the model for up to 50 epochs using the Adam optimizer (learning rate = 1e-5) and binary cross-entropy loss. Early stopping was employed with a patience of 10 epochs, based on validation loss. Performance was evaluated using AUROC and AUPRC, both macro- and micro-averaged across all labels with non-trivial distributions. Evaluation was conducted on a held-out test chromosome, and all data splits were consistent with those used for SeiPlant. Our implementation was based on the original DanQ architecture [9] and customized to support plant chromatin prediction. Our research used a customized PyTorch implementation of DanQ (https://github.com/HelloWorldLTY/DanQ_pytorch).

Enformer

This is a transformer-based deep learning model developed to predict a wide range of genomic and epigenomic signals from long DNA sequences by explicitly modeling long-range regulatory interactions [12]. In this study, we adapted Enformer for plant chromatin feature prediction using a PyTorch implementation. Due to the relatively compact regulatory architecture of plant genomes and to reduce GPU memory requirements, we downscaled the original input sequence length (196,608 bp) to 1,024 bp. Our implementation retains the key architectural components of Enformer, including initial convolutional and pooling blocks for local feature extraction, followed by a stack of transformer layers to model distal dependencies across sequence bins. Specifically, our PyTorch model takes one-hot encoded 1,024 bp sequences as input and outputs a tensor of shape [batch_size, 8, num_targets], corresponding to eight 128-bp bins. The model was trained using binary cross-entropy loss and optimized with the Adam optimizer (learning rate = 1e-4) for up to 100 epochs. We used a batch size of 64 and applied early stopping based on validation loss with a patience of 20 epochs. During evaluation, we aggregated predictions across bins using max pooling, and assessed model performance using macro- and micro-averaged AUROC and AUPRC metrics on a held-out

test chromosome. All training, validation, and test splits were identical to those used for SeiPlant. Our implementation was adapted from the official DeepMind TensorFlow codebase (https://github.com/google-deepmind/deepmind-research/tree/master/enformer) and the PyTorch version provided by the community (https://github.com/boxiangliu/enformer-pytorch).

Ablation design for key components

To systematically evaluate the functional contributions of key architectural components in the Sei model, we performed a series of ablation experiments targeting three critical modules: residual connections, dilated convolutional layers, and B-spline transformation layers. Each component can be selectively disabled via configuration flags, and the network internally substitutes them with neutral alternatives to preserve structural connectivity and enable end-to-end gradient propagation.

Residual blocks disabled

In the full Sei architecture, residual (skip) connections are applied within convolutional blocks to promote gradient flow and facilitate feature reuse. When residual connections are disabled, these additive operations are omitted. The model simply passes the output of each convolutional block directly to the next, without merging with earlier layer outputs.

Dilated blocks disabled

The standard model incorporates a series of five dilated convolutional layers to enlarge the receptive field without increasing parameter count. When this module is removed, the entire dilated block is skipped. The output from the final standard convolutional layer is routed directly to the downstream transformation layer (either spline or flatten), with no substitute layers inserted.

Spline transformation disabled

The B-spline transformation acts as a learnable spatial projection layer prior to classification. When this component is removed, it is replaced by a flattening operation (nn.Identity()), which directly reshapes the preceding convolutional output into a 2D tensor compatible with the classifier head. This replacement ensures dimensional alignment without introducing additional transformation.

Evaluation of predicted signal accuracy

To quantitatively evaluate the accuracy of predicted histone modification signals, we implemented a multi-metric comparison framework based on both global signal similarity and peak-level agreement between predicted and experimentally measured BigWig files. This evaluation was conducted in three main steps:

1. Genome binning and signal extraction

   The reference and predicted BigWig files were processed using pyBigWig to extract average signal values across non-overlapping genomic bins of fixed size (default:

1,024 bp). For each chromosome present in both tracks, signal values were extracted for each bin, and bins containing NaN values were replaced with zero. The extracted values were used to compute global correlation metrics and define peak regions.

2. Global correlation metrics
   To assess the genome-wide similarity between predicted and experimentally measured histone modification signals, we calculated global correlation metrics using both Pearson methods. The genome was first divided into non-overlapping bins of fixed size (1,024 bp), and the average signal value for each bin was extracted from the predicted and reference BigWig files using the stats() function from the pyBigWig Python package. The resulting signal vectors were then compared using the pearsonr() functions from the scipy.stats module to compute Pearson correlation coefficients. Pearson correlation measures the linear relationship between the predicted and true signal profiles. In addition, for an efficient whole-genome correlation estimate, we employed UCSC's command-line tool bigWigCorrelate, which directly calculates the Pearson correlation between two BigWig tracks. The output of this tool was recorded as the fast global correlation metric and used as a reference to validate the signal consistency at scale.
3. Peak region overlap analysis
   To assess the agreement between predicted and reference histone modification signals in high-signal (peak) regions, we computed the peak overlap ratio based on bin-level peak detection and set intersection. Peak regions were then determined by applying a percentile-based threshold. Specifically, bins whose signal values exceeded the 90th percentile of all bin values in the track were considered peak bins. This process was applied independently to both the predicted signal track and the reference signal track.

$$\text{Peak Overlap Ratio} = \frac{|P_{\text{pred}} \cap P_{\text{ref}}|}{|P_{\text{pred}} \cup P_{\text{ref}}|} \tag{7}$$

   where P_{pred} is the set of peak bins from the predicted signal, P_{ref} is the set of peak bins from the reference (ground truth) signal, $|P_{\text{pred}} \cap P_{\text{ref}}|$ is the number of overlapping peak bins (shared peaks), $|P_{\text{pred}} \cup P_{\text{ref}}|$ is the total number of unique peak bins across both tracks.
4. ROC-AUC assessment
   To further evaluate the predictive power, we treated the predicted signal values as continuous scores and generated binary labels from the normalized reference signal. Specifically, reference signal values were log-transformed using log1p, followed by min–max normalization. A threshold of 0.5 on the normalized values was used to define binary labels. The ROC-AUC (receiver operating characteristic area under the curve) was computed using sklearn.metrics.roc_auc_score to measure how well the predicted signal distinguished high and low signal regions.

In silico mutagenesis (ISM) and impact score calculation

To assess the functional relevance of individual nucleotide positions within regulatory sequences, we performed in silico saturation mutagenesis (ISM) analysis. Each input

sequence (length = 1,024 bp) was first encoded using one-hot encoding (A, T, C, G → 4 channels). For each sequence, we generated all possible single-nucleotide mutations at every position, resulting in $3 \times 1{,}024 = 3{,}072$ mutated variants per sequence. This mutation set included both the reference sequence and all its single-base substitutions. All encoded sequences were passed through a pretrained Sei model to obtain predicted chromatin feature scores. The model outputs, for each sequence or mutation variant, a vector of predicted probabilities across all trained histone modification labels. Then the mutation effect (Δ-score) is defined as:

$$\Delta^{(i)} = \mathbf{x}_{\text{mut}}^{(i)} - \mathbf{x}_{\text{ref}} \in \mathbb{R}^F \tag{8}$$

where $\mathbf{x}_{\text{ref}} \in \mathbb{R}^F$ is the prediction vector for the reference sequence, $\mathbf{x}_{\text{mut}}^{(i)} \in \mathbb{R}^F$ is the prediction vector for the i-th mutation (where F is the number of genomic features, e.g., histone marks). For downstream comparison of peak and non-peak regions, we focused on the most impactful mutations by computing the top-5% average absolute Δ-score within each sequence. Specifically, for a target histone mark (indexed by tag), we computed:

$$\text{Impact Score} = \frac{1}{k} \sum_{j=1}^{k} \left| \Delta_{\text{tag}}^{(j)} \right|, \quad \text{where } k = \lceil 0.05 \times 3072 \rceil \tag{9}$$

This score reflects the sensitivity of a region's chromatin feature prediction to small sequence perturbations. Separate impact scores were computed for sequences in peak and non-peak regions.

Identification of important genomic regions for ISM analysis

To evaluate model sensitivity to specific loci, we extracted high-confidence peak regions and matched flanking non-peak controls from genome-wide predicted signal tracks. For each histone modification, the reference genome (FASTA) and predicted BigWig files were processed by first converting BigWig to BedGraph, suppressing background signals (< 0.1 set to zero), and removing non-standard chromosomes. Intervals were sorted by signal intensity, the top 1% were discarded to avoid extreme outliers, and up to top_n (default: 50) non-overlapping peaks were selected by centering a fixed-length window (1024 bp, flank = 512 bp) on the highest remaining scores. Only sequences with standard nucleotides (A, T, C, G) were retained. For each peak, flanking non-peak regions were identified by sliding a 1024 bp window (step = 128 bp) upstream and downstream until a region with all values < 0.5 and no ambiguous bases was found. Final peak and non-peak sets were saved as CSV files containing species name, chromosome, and coordinates. This ensured ISM was applied to biologically relevant loci with well-matched negative controls, enabling interpretable comparisons of sequence perturbation effects across different chromatin contexts.

Clustering and functional annotation of predicted chromatin profiles

Unsupervised clustering was performed on the predicted signal vectors. First, we used t-Distributed Stochastic Neighbor Embedding (t-SNE) for dimensionality reduction to two dimensions. Clustering was then applied to the t-SNE-embedded coordinates to identify distinct groups of regulatory sequences.

To assign functional annotations to these clusters, we performed enrichment analysis utilizing experimentally validated ChIP-seq peak annotations of specific histone modifications and transcription factor (TF) binding sites, including H3K4me3, H3K27ac, H3K4me1, H3K9me2, CENH3, and RNA polymerase (RNAP). Clusters enriched in promoter-associated marks—primarily indicated by RNAP binding sites and H3K4 trimethylation (H3K4me3)—were annotated as promoter regions. Although acetylation marks (such as H3K27ac) are commonly associated with promoters and enhancers, we prioritized RNAP occupancy and H3K4 methylation states for robust promoter annotation due to their specificity and consistent enrichment patterns. Clusters significantly enriched for enhancer-associated marks (H3K4me1 and specific TF binding sites) were annotated as enhancers. Heterochromatic regions were annotated based on characteristic histone modifications: clusters enriched for centromeric-specific CENH3 were labeled as centromeric heterochromatin, clusters enriched for H3K27me3—associated with facultative heterochromatin (regions of reversible chromatin repression)—were annotated accordingly, and clusters enriched for H3K9me2 were categorized as constitutive heterochromatin. For euchromatic regions, the methylation state of H3K4 distinguished enhancer-like regions (monomethylation) from promoter-like regions (trimethylation). In maize, clusters demonstrating strong enrichment of specific transcription factors were further annotated according to the dominant TF identity.

Final functional annotations were systematically assigned based on predominant chromatin features and genomic context, resulting in biologically meaningful categories, including "promoter," "enhancer," "centromeric heterochromatin," "facultative heterochromatin," and "constitutive heterochromatin."

Evaluation metrics

AUROC and AUPRC were used as evaluation metrics to assess model classification performance. AUROC reflects the model's ability to distinguish positive from negative samples across thresholds, while AUPRC summarizes precision–recall trade-offs and is especially informative under class imbalance. Both metrics were computed from predicted probabilities and true labels using scikit-learn on held-out test sets.

Macro-averaged AUROC/AUPRC computes the metric independently for each label (i.e., histone mark or target) and then takes the unweighted average across all labels. This reflects how the model performs on each class regardless of its prevalence.

$$AUROC_{macro} = \frac{1}{L}\sum_{l=1}^{L} AUROC^{(l)} \tag{10}$$

$$AUPRC_{macro} = \frac{1}{L}\sum_{l=1}^{L} AUPRC^{(l)} \tag{11}$$

where L is total number of labels, $AUROC^{(l)}$ and $AUPRC^{(l)}$ are the score calculated for the l-th label (class).

$$AUROC_{micro} = AUROC\left(\bigcup_{l=1}^{L}\hat{y}^{(l)}, \bigcup_{l=1}^{L}y^{(l)}\right) \tag{12}$$

$$AUPRC_{micro} = AUPRC\left(\bigcup_{l=1}^{L}\hat{y}^{(l)}, \bigcup_{l=1}^{L}y^{(l)}\right) \tag{13}$$

$\hat{y}^{(l)}$ predicted probability vector for the l-th label. $y^{(l)}$ is the ground truth binary label vector for the l-th label. $\bigcup_{l=1}^{L}\hat{y}^{(l)}$ Denotes concatenation (flattening) of prediction or label vectors across all classes.

We employed three metrics to quantify the similarity of model predictions across duplicated genomic regions within species and between orthologous regions across related species:

Average Cosine Similarity (pairwise): For each homologous region pair, we computed the cosine similarity between their predicted signal vectors. The final score was obtained by averaging the pairwise cosine similarities across all duplicated pairs. Cosine similarity is defined as

$$\text{Cosine Similarity}(x, y) = \frac{x \cdot y}{||x|| \cdot ||y||} \tag{14}$$

where x and y are the predicted signal vectors of two homologous regions.

Mean Squared Error (MSE): The mean squared error between the predicted signal vectors of each duplicated pair, then averaged across all pairs:

$$\text{MSE}(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2 \tag{15}$$

Pearson Correlation Coefficient (flattened): To assess linear correlation between predicted signal values, we concatenated predictions from all duplicated region pairs into two flattened vectors and computed the Pearson correlation coefficient between them:

$$r = \frac{\text{Cov}(x, y)}{\sigma_x \cdot \sigma_y} \tag{16}$$

where $\text{Cov}(x, y)$ is the covariance, and σ_x, σ_y are standard deviations of the predicted values from region sets x and y.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-025-03929-4.

Additional file 1: Fig. S1. The effect of different input window lengths on the prediction performance of three species. Fig. S2. Comparison with representative models. Fig. S3. SeiPlant architecture ablation experiment. Fig. S4. Quantitative evaluation of chromatin signal predictions in Maize. Fig. S5. Functional annotation of chromatin states based on predicted regulatory profiles in Rice. Fig. S6. Functional annotation of chromatin states based on predicted regulatory profiles in Arabidopsis. Fig. S7. Functional annotation of chromatin states based on predicted regulatory profiles in Maize. Fig. S8. Cross-species prediction performance comparison between models trained on different plant species across five histone marks. Fig. S9. Representative predictions and quantitative summaries for histone modifications in Brassicaceae species. Fig. S10. Representative predictions and quantitative summaries for histone modifications in Poaceae species. Fig. S11. Per-chromosome prediction accuracy for additional histone modifications in representative Brassicaceae and Poaceae species. Fig. S12. Cross-family versus within-family prediction on

Brassicaceae targets. Fig. S13. Cross-family versus within-family prediction on Poaceae targets. Fig. S14. Prediction Consistency on Duplicated Genomic Regions Across Plant Species. Fig. S15. Motif occurrence between promoter-from enhancer-labeled clusters across species

Additional file 2: Table S1. Details of species sequence information. Table S2. Chromatin feature data used for rice, maize, and Arabidopsis. Table S3. Verified species data sets for Poaceae and Brassicaceae. Table S4. Verified specises data sets for other family

Acknowledgements
We thank the Center for Information Technology and the High Performance Computing Center of Nanjing University for providing computational resources used in this study. We also acknowledge the support from the Yachen Foundation of Nanjing University.

Peer review information
Pil Joon Seo and Wenjing She were the primary editors of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team. The peer-review history is available in the online version of this article.

Data and code availability
All datasets used for model training and evaluation in this study were obatined from the ChIP-Hub database (https://biobigdata.nju.edu.cn/ChIPHub/) [30]. The genome assemblies included in our analyses are summarized in Additional file 2: Table S1; Chromatin feature datasets for rice, maize, and Arabidopsis are listed in Additional file 2: Table S2; datasets for other Poaceae and Brassicaceae species are listed in Additional file 2: Table S3; and datasets for all remaining species are listed in Additional file 2: Table S4. The genomic sequence data and chromatin feature information related to *Kalanchoe pinnata* were independently generated and assembled by our laboratory, and deposited at the National Genomics Data Center and are publicly accessible (https://ngdc.cncb.ac.cn/gsa/browse/CRA026198) [41].
All source code used in this study is available at our GitHub repository (https://github.com/compbioNJU/SeiPlant), which under the MIT License [42]. A versioned archive of the Poaceae, Brassicaceae, and mixed-species model training parameters and configuration files has been deposited in Zenodo (https://doi.org/10.5281/zenodo.15421964) [43].

Authors' contributions
D.C. designed the research. T.L. completed the experiments and data preprocessing involved in this paper, completed the first draft of this research paper. Q.H. and Z.R. assisted in data collection and processing. Y.L. and C.L. assisted in the clustering and annotation of chromatin feature profiles. H.C. and M.C. assisted in the experimental design of this paper. All the authors reviewed and approved the paper.

Funding
This work was supported by the National Natural Science Foundation of China (Grant No. T2541063 and 32270709); the Yachen foundation of Nanjing University; 151 Talent Project, and Science and Technology Innovation Leader of Zhejiang Province (2022R52035); and Jiangsu Collaborative Innovation Center for Modern Crop Production co-sponsored by province and ministry.

Data availability
No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 26 May 2025 Accepted: 31 December 2025
Published online: 09 January 2026

References

1. Liu C, Lu F, Cui X, Cao X. Histone methylation in higher plants. Annu Rev Plant Biol. 2010;61:395–420.
2. Yocca AE, Edger PP. Current status and future perspectives on the evolution of *cis*-regulatory elements in plants. Curr Opin Plant Biol. 2022;65:102139.
3. Schmitz RJ, Grotewold E, Stam M. Cis-regulatory sequences in plants: their importance, discovery, and future challenges. Plant Cell. 2022;34:718–41.

4. Swinnen G, Goossens A, Pauwels L. Lessons from domestication: targeting *cis* -regulatory elements for crop improvement. Trends Plant Sci. 2016;21:506–15.
5. Hu X, Fernie AR, Yan J. Deep learning in regulatory genomics: from identification to design. Curr Opin Biotechnol. 2023;79:102887.
6. Allis CD, Jenuwein T. The molecular hallmarks of epigenetic control. Nat Rev Genet. 2016;17:487–500.
7. Kouzarides T. Chromatin modifications and their function. Cell. 2007;128:693–705.
8. Novakovsky G, Dexter N, Libbrecht MW, Wasserman WW, Mostafavi S. Obtaining genetics insights from deep learning via explainable artificial intelligence. Nat Rev Genet. 2023;24:125–37.
9. Quang D, Xie X. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. Nucleic Acids Res. 2016;44:e107–e107.
10. Kelley DR, Reshef YA, Bileschi M, Belanger D, McLean CY, Snoek J. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. Genome Res. 2018;28:739–50.
11. Pei G, Hu R, Dai Y, Manuel AM, Zhao Z, Jia P. Predicting regulatory variants using a dense epigenomic mapped CNN model elucidated the molecular basis of trait-tissue associations. Nucleic Acids Res. 2021;49:53–66.
12. Avsec Ž, Agarwal V, Visentin D, Ledsam JR, Grabska-Barwinska A, Taylor KR, et al. Effective gene expression prediction from sequence by integrating long-range interactions. Nat Methods. 2021;18:1196–203.
13. Hoffman GE, Bendl J, Girdhar K, Schadt EE, Roussos P. Functional interpretation of genetic variants using deep learning predicts impact on chromatin accessibility and histone modification. Nucleic Acids Res. 2019;47:10597–611.
14. Cofer EM, Raimundo J, Tadych A, Yamazaki Y, Wong AK, Theesfeld CL, et al. Modeling transcriptional regulation of model species with deep learning. Genome Res. 2021;31:1097–105.
15. Zhou J, Theesfeld CL, Yao K, Chen KM, Wong AK, Troyanskaya OG. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. Nat Genet. 2018;50:1171–9.
16. Agarwal V, Shendure J. Predicting mRNA abundance directly from genomic sequence using deep convolutional neural networks. Cell Rep. 2020;31(7):107663.
17. Talukder A, Barham C, Li X, Hu H. Interpretation of deep learning in genomics and epigenomics. Brief Bioinform. 2021;22:bbaa177.
18. Toneyan S, Tang Z, Koo PK. Evaluating deep learning for predicting epigenomic profiles. Nat Mach Intell. 2022;4:1088–100.
19. Zhang Q, Wang S, Li Z, Pan Y, Huang D. Cross-species prediction of transcription factor binding by adversarial training of a novel nucleotide-level deep neural network. Adv Sci. 2024. https://doi.org/10.1002/advs.202405685.
20. Minnoye L, Taskiran II, Mauduit D, Fazio M, Van Aerschot L, Hulselmans G, et al. Cross-species analysis of enhancer logic using deep learning. Genome Res. 2020;30:1815–34.
21. Kelley DR. Cross-species regulatory sequence activity prediction. Ma J, editor. PLoS Comput Biol. 2020;16:e1008050.
22. Chen L, Fish AE, Capra JA. Prediction of gene regulatory enhancers across species reveals evolutionarily conserved sequence properties. Ma J, editor. PLoS Comput Biol. 2018;14:e1006484.
23. Mendoza-Revilla J, Trop E, Gonzalez L, Roller M, Dalla-Torre H, De Almeida BP, et al. A foundational large language model for edible plant genomes. Commun Biol. 2024;7:835.
24. Peleke FF, Zumkeller SM, Gültas M, Schmitt A, Szymański J. Deep learning the *cis*-regulatory code for gene expression in selected model plants. Nat Commun. 2024;15:3488.
25. Zhao T, Zhan Z, Jiang D. Histone modifications and their regulatory roles in plant development and environmental memory. J Genet Genomics. 2019;46:467–76.
26. Jiang D, Borg M, Lorković ZJ, Montgomery SA, Osakabe A, Yelagandula R, et al. The evolution and functional divergence of the histone H2B family in plants. Malik HS, editor. PLoS Genet. 2020;16:e1008964.
27. Lu Z, Marand AP, Ricci WA, Ethridge CL, Zhang X, Schmitz RJ. The prevalence, evolution and chromatin signatures of plant regulatory elements. Nat Plants. 2019;5:1250–9.
28. Chen KM, Wong AK, Troyanskaya OG, Zhou J. A sequence-based global map of regulatory activity for deciphering human genetics. Nat Genet. 2022;54:940–9.
29. Zhou X, Ruan Z, Zhang C, Kaufmann K, Chen D. Deep learning on chromatin profiles reveals the cis-regulatory sequence code of the rice genome. Journal of Genetics and Genomics. 2024;S1673852724003564.
30. Fu L-Y, Zhu T, Zhou X, Yu R, He Z, Zhang P, et al. ChIP-hub provides an integrative platform for exploring plant regulome. Nat Commun. 2022;13(1):3413.
31. Hsieh C-H, Chang Y-TS, Yen M-R, Hsieh J-WA, Chen P-Y. Predicting protein synergistic effect in *Arabidopsis* using epigenome profiling. Nat Commun. 2024;15:9160.
32. Ricci WA, Lu Z, Ji L, Marand AP, Ethridge CL, Murphy NG, et al. Widespread long-range *cis*-regulatory elements in the maize genome. Nat Plants. 2019;5:1237–49.
33. Maher KA, Bajic M, Kajala K, Reynoso M, Pauluzzi G, West DA, et al. Profiling of Accessible Chromatin Regions across Multiple Plant Species and Cell Types Reveals Common Gene Regulatory Principles and New Control Modules. Plant Cell. 2018;30:15–36.
34. Schranz ME, Lysak MA, Mitchell-Olds T. The ABC's of comparative genomics in the Brassicaceae: building blocks of crucifer genomes. Trends Plant Sci. 2006;11:535–42.
35. Ma M, Zhong W, Zhang Q, Deng L, Wen J, Yi B, et al. Genome-wide analysis of transcriptome and histone modifications in *Brassica napus* hybrid. Front Plant Sci. 2023;14:1123729.
36. TA Cazares FW Rizvi B Iyer X Chen M Kotliar AT Bejjani et al Przytycka TM, editor 2023 MaxATAC: genome-scale transcription-factor binding prediction from ATAC-seq with deep neural networks PLoS Comput Biol 19 e1010863
37. Thibodeau A, Khetan S, Eroglu A, Tewhey R, Stitzel ML, Ucar D. CoRE-ATAC: A deep learning model for the functional classification of regulatory elements from single cell and bulk ATAC-seq data. Roy S, editor. PLoS Comput Biol. 2021;17:e1009670.
38. Fudenberg G, Kelley DR, Pollard KS. Predicting 3D genome folding from DNA sequence with Akita. Nat Methods. 2020;17:1111–7.
39. Gao Z, Liu Q, Zeng W, Jiang R, Wong WH. EpiGePT: a pretrained transformer-based language model for context-specific human epigenomics. Genome Biol. 2024;25:310.

40. Liu G, Chen L, Wu Y, Han Y, Bao Y, Zhang T. PDLLMs: A group of tailored DNA large language models for analyzing plant genomes. Molecular Plant. 2024;S1674205224003903.
41. Ruan Z. ATAC-seq and CHIP-seq of kalanchoe daigremontiana. Datasets. National Genomics Data Center. 2025. https://ngdc.cncb.ac.cn/gsa/browse/CRA026198.
42. Lv T, Han Q, Li Y, L C, Ruan Z, Chao H, et al. Cross-Species Prediction of Histone Modifications in Plants via Deep Learning. Github. 2025. https://github.com/compbioNJU/SeiPlant.
43. Lv T, Han Q, Li Y, Liang C, Ruan Z, Chen D. Cross-Species Prediction of Histone Modifications in Plants via Deep Learning. 2025. Zenodo. https://doi.org/10.5281/zenodo.15421964.

Publisher's Note