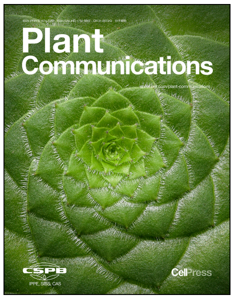


Cropformer: An Interpretable Deep Learning Framework for Crop Genome Prediction

Hao Wang, Shen Yan, Wenxi Wang, Yongming Cheng, Jingpeng Hong, Qiang He, Xianmin Diao, Yunan Lin, Yanqing Chen, Yongsheng Cao, Weilong Guo, Wei Fang

PII: S2590-3462(24)00644-8

DOI: https://doi.org/10.1016/j.xplc.2024.101223

Reference: XPLC 101223

To appear in: *PLANT COMMUNICATIONS*

Received Date: 3 September 2024

Revised Date: 15 October 2024

Accepted Date: 12 December 2024

Cropformer: An Interpretable Deep Learning Framework for Crop Genome Prediction

Hao Wang[1,†], Shen Yan[1,†], Wenxi Wang[2,†], Yongming Cheng[2], Jingpeng Hong[3], Qiang He[1], Xianmin Diao[1], Yunan Lin[4], Yanqing Chen[1], Yongsheng Cao[1,*], Weilong Guo[2,*], Wei Fang[1,*]

[1] Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing 100081, China.

[2] Frontiers Science Center for Molecular Design Breeding, Key Laboratory of Crop Heterosis and Utilization (MOE), and Beijing Key Laboratory of Crop Genetic Improvement, China Agricultural University, Beijing 100193, China

[3] College of Information and Management Science, Henan Agricultural University, Zhengzhou 450002, China

[4] School of Engineering and Design, Technical University Munich, 85521, Munich, Germany

* Corresponding authors: Wei Fang, Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing 100081, China. Email: fangwei@caas.cn; Weilong Guo, Frontiers Science Center for Molecular Design Breeding, Key Laboratory of Crop Heterosis and Utilization (MOE), and Beijing Key Laboratory of Crop Genetic Improvement, China Agricultural University, Beijing 100193, China Email: guoweilong@cau.edu.cn; Yongsheng Cao, Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing 100081, China. Email: caoyongsheng@caas.cn.

Summary

Machine learning and deep learning have become transformative tools in genomic selection (GS) to improve prediction accuracy and accelerate crop breeding. Cropformer, a novel deep learning framework combining convolutional neural networks and self-attention mechanisms, demonstrates superior performance in predicting phenotypic traits across five major crops. By improving prediction robustness and interpretability, Cropformer assists gene mining and supports genomic-assisted breeding strategies.

Abstract

Machine learning and deep learning have been employed in genomic selection (GS) to expedite the identification of superior genotypes and accelerate breeding cycles. However, a significant challenge for current data-driven deep learning models in GS is their low robustness and interpretability. To address this challenge, we developed Cropformer, a deep learning framework for predicting crop phenotypes and exploring downstream tasks. The framework consists of a combination of convolutional neural networks and multiple self-attention mechanisms to improve accuracy. Here, Cropformers ability to predict complex phenotypic traits was extensively evaluated on more than 20 traits across five major crops: maize, rice, wheat, foxtail millet, and tomato. Evaluation results show that Cropformer outperforms other GS methods in precision and robustness. Compared to the runner-up model, Cropformer's prediction accuracy improved by up to 7.5%. Additionally, Cropformer enhances the ability to analyze and assist the mining of genes associated with traits. With Cropformer, we identify dozens of single nucleotide polymorphisms (SNPs) with potential effects on maize phenotypic traits and reveal key genetic variations t underlying these differences. Cropformer makes considerable advances in predictive performance and assisted gene identification, representing a powerful general approach to facilitating the genomic design of crop breeding. Cropformer is freely accessible at https://cgris.net/cropformer.

Keywords: Deep learning; Genomic selection; Multiple self-attention mechanisms; Phenotypic prediction;

Introduction

By 2050, approximately 9 billion people will live on earth, and utilizing limited resources is a serious challenge for ensuring the demand for global food production can be met(Wallace et al., 2018). Furthermore, changing lifestyles, altered population demographics, deterioration of natural resources, climate change, and diminished water supplies are equally challenging problems for crop breeders aiming to achieve precision plant breeding to improve crop performance(Hickey et al., 2017). With the advancement of next-generation sequencing technologies, knowledge acquired from basic plant biology research has dramatically enhanced our understanding of the structure and function of plant genomes and has accelerated crop improvement in recent decades(Varshney et al., 2005). However, the time-consuming nature and even inability to capture "minor" genetic effects in marker–QTL associations remain the major barriers to the selection of suitable breeding materials(Desta and Ortiz, 2014; Xu et al., 2012).

The introduction of genomic selection (GS) has paved the way for overcoming these limitations through the use of whole-genome prediction models(Ma et al., 2018). GS was initially proposed by Meuwissen et al. to improve breeding efficiency by reducing phenotyping costs and shortening the cycle time for early-generation selection(Meuwissen et al., 2001). GS utilizes machine learning to determine the correlation between phenotypic data and high-density molecular markers, such as single nucleotide polymorphisms (SNPs), in the training population(Tong et al., 2020). The model was subsequently used to predict genomic estimated breeding values (GEBV) for genotypes in the test population(Habyarimana et al., 2020; Werner et al., 2020). Most importantly, GS allows for the consideration of minor-effect QTLs that cannot be detected by traditional association methods, thus improving the ability to predict these QTLs and drastically reducing the duration of the breeding process(Tong and Nikoloski, 2021). Such advances in genotyping techniques are allowing samples to be genotyped at a lower cost, and GS in particular is actively being incorporated into plant breeding(Krishnappa et al., 2021).

Over the last few decades, a series of models using statistics and machine learning have been well advanced in genome prediction based on genome sequences(Covarrubias-Pazaran, 2016; Endelman, 2011a; Misztal, 2008). For instance, ridge regression BLUP (rrBLUP), using linear mixed-effects models to infer genomic kinship and marker effects in breeding material for phenotypic prediction(Endelman, 2011b). Expanding on Light Gradient Boosting Machines (LightGBM), Yan et al. developed CropGBM to achieve genotype-to-phenotype prediction. With a large dataset of inbred and hybrid maize lines, CropGBM exhibited superior performance in terms of prediction precision, model stability, and computing efficiency(Yan et al., 2021). In addition to the above methods, there are many other deep learning-based genome prediction methods such as DEM(Ren et al., 2024), DNNGP(Wang et al., 2023b), DeepGS(Ma *et al.*, 2018) and SoyDNGP(Gao et al., 2023). Although deep learning has been successfully applied to whole-genome prediction, current methods still follow a "black-box" model and lack interpretability. This limitation restricts our ability to understand the relationship between features and prediction outcomes. Additionally, the predictive accuracy and training efficiency can be further improved.

Here, we present Cropformer, a GS framework that combines convolutional neural network and self-attention mechanism. Evaluation showed that Cropformer outperformed all other methods in the prediction of both discrete and quantitative traits. Comparing with previous deep learning-based GS algorithms, Cropformer can assess the correlation between variations and crop traits with high resolution, facilitating the understanding of the "black-box" mechanisms of deep learning models. In summary, we present a deep learning-based method in conjunction with genomic big data for genomic prediction in crops with improved accuracy, supporting the interpretation of key variations associated with phenotypes that have not been previously reported and suggesting a promising future for understanding how the genome produces phenotypes.

Results

Design of Cropformer

To predict complex phenotypic traits in crops, we developed and trained a deep neural network model, namely Cropformer (Figures 1, and S1). Cropformer takes the sequences of SNPs from genomic variation data and phenotypic values as input to train and make predictions (Figure 1A). The core components of Cropformer consists a convolutional neural network (CNN) layer and a multiheaded self-attention mechanism (Figure 1B). The convolution layer of CNN can automatically extract features from the raw input data and map them into information representations during the training process without human intervention(Krizhevsky et al., 2017). It transforms the input genomic data into informative representations, optimizing the model's learning during training. The output features of the CNN are fed into the attention module to obtain a decision vector for prediction. To demonstrate the effectiveness of integrating CNN and the multiheaded self-attention mechanism in improving prediction, an ablation study was performed, which resulted in the pearson correlation coefficient (PCC) of Cropformer on the Maize data being 92.21% for days to tasselling (DTT), 91.82% for plant height (PH), and 76.31% for ear weight (EW), which were 10.6%, 3.9%, and 6.9% higher than the attention module alone, and 3.42%, 2.0%, and 10.3% higher than CNN only, respectively (Figure S3). The same performances were also demonstrated in four other datasets (Wheat, Foxtail millet, Rice, and Tomato). Furthermore, Cropformer has comparable training time with CNN and Attention (Figures S4-7).

The weights of the attention mechanism can be extracted to evaluate the impact of each loci on modeling decisions (Figure 1C). Based on these attention weights, loci associated with crop phenotype prediction can be further identified. The entire analytical framework is applicable for supporting various downstream tasks, such as genomic selection and SNPs mining.

Cropformer outperforms existing models for genomic prediction

Five crop species (maize, wheat, foxtail millet, rice, and tomato), each with a dataset of multidimensional genomic variation information, were collected from public studies. We applied Cropformer to these datasets with different population sizes to assess the prediction performance for both discrete (regression) and quantitative (classification) traits. A range of widely used models specifically designed for crop genomic selection prediction were compared, including CropGBM, DNNGP, extreme gradient boosting (XGBoost), Support Vector Regression (SVR), Multilayer Perceptron (MLP), ridge regression Best Linear Unbiased Prediction (rrBLUP), and Dual-extraction modelling (DEM) (Supplementary Tables 2-6). We randomly divided the data of the five datasets into 80% training and 20% testing sets. To avoid overfitting, we employed nested cross-validation to train the model and used callback functions to guide early stopping, ultimately validating the model's robustness on the test set.

We first trained and tested Cropformer using the maize dataset to evaluate the model's performance in predicting phenotypes for days to tasselling (DTT), plant height (PH), and ear weight (EW) (Figure 2A and Supplementary Tables 7-9). According to the final performance evaluation of all the methods on the test dataset, Cropformer exhibited the optimal performance according to PCC (DTT=92.2%%, PH=91.8%, and EW=76.3%), followed by DEM, and CropGBM (DTT=89.5%, PH=88.7%, and EW=70.8%) (Figure 2B).

The performance of the compared methods on the other three datasets was also fully evaluated to assess model generalizability. Cropformer achieved the best performance with all the datasets. Specifically, Cropformer's performance in predicting wheat traits was 63.1% for thousand kernel weight (TKW), 68.7% for grain width (GW), 66.8% for grain height (GH), 49.5% for grain pressure (GP), and 72.4% for grain length (GL). Cropformer outperformed CropGBM by 11.0%, 0.6%, 4.5%, 1.5%, and 1.6%, respectively (Figure 3A and Supplementary Tables 10-14). As indicated in Figure 3B and Supplementary Tables 15–19, analysis of the trait straw weight of the foxtail millet dataset was performed using Cropformer for the regions of Anyang (83.8%), Beijing

(84.1%), Changzhi (86.0%), Dingxi (81.3%), and Urumqi (85.5%). These values were greater than those achieved with the runner-up model (7.5%, 7.6%, 3.2%, 5.3%, and 6.5%, respectively) (Figure 3B and Supplementary Tables 15–19). With the rice dataset, Cropformer had the best prediction performance in predictions of all five traits: 72.1% for Culm_length, 69.5% for days_to_heading_2018HN, 65.5% for grain_length_width_ratio, 72.6% for plant_height_2018HN, and 63.3% for thousand_grain_weight (Figure 3C and Supplementary Tables 20–24). Compared to the other methods, our model improves the prediction performance by 0.3% to 10.0%. Therefore, we conclude that our approach is more effective than CropGBM, DNNGP, XGBoost, SVR, MLP, rrBLUP, and DEM.

Furthermore, incorporating additional molecular features was feasible with the Cropformer model, and here, we assessed the effect of different dimensions of molecular features on the model predictions. On the tomato Sopim_BGV006775_12T001232 (an enzyme-encoding gene affecting flavonoids) trait test set, Cropformer achieved PCC values of 59.3%, 64.7%, 54.7%, and 52.4% on the basis of SNP, insertion and deletion (InDel), gene expression (GE), and structural variation (SV) features, respectively (Figure 3D and Supplementary Tables 25–28). We extracted the top 1500 weighted features from the four types of genomic variants to construct fusion features. Through fusion, Cropformer achieved a prediction PCC of 71.5% for the Sopim_BGV006775_12T001232 trait, which was 12.2%, 6.8%, 16.8%, and 19.1% better than that achieved when using SNP, SV, InDel, and GE features, respectively.

Finally, we benchmarked the runtime of Cropformer with other methods on five datasets. In our study, CropGBM, XGBoost, SVR, MLP, and rrBLUP had the fastest prediction times in small-scale (Tomato, and Foxtail millet) datasets, and Cropformer was able to achieve similar time consumption. DEMs have excellent predictive performance, but require longer computation times and more GPU resources. As the size of the dataset increases (Maize, Rice, and Wheat), Cropformer outperforms the other methods, using only a slight increase in computation time. (Figures S8A-S8E).

Cropformer supports classification prediction

Although the Cropformer model is a regression model for quantitative traits, it also supports performing classification prediction with label-based discrete traits. To test the classification performance of Cropfromer, we divided DTT trait of the maize dataset into three classes, samples with early flowering time (first 25% DTT), moderate flowering time (25 to 75% DTT), and late flowering time (last 25% DTT), referring to the method of Yan *et al* (Yan *et al.*, 2021). Moreover, we also examined performance in the sample balance situation, where DTT traits were split according to early (first 50% DTT) and late flowering time (last 50% DTT). We also employed the maximal information coefficient (MIC) to filter out 10,000 SNPs that had high representativeness. To intuitively assess the importance of the SNPs, we used Uniform Manifold Approximation and Projection (UMAP) for dimension reduction and feature visualization, and the results showed that the samples clustered using filtered SNPs had clearer groupings than those clustered using all SNPs, suggesting that filtering can not only reduce model size but also improve model performance (Figure 4A).

Multiple indicators were calculated for evaluating Cropformer in predicting the DTT (Three-classification): the accuracy was 77.2% (Figure 4B), the precision was 77.6%, the recall was 76.7%, the F1_score was 77.1% (Figure 4C and Supplementary Tables 30–31), and the area under the curve (AUC) was 91.2% (Figures 4D and S9). The accuracy, precision, recall, and F1_score values of Cropformer were 1.7%, 0.6%, 1.7% and 2.4% higher than the runner-up DEM. For the two-classification, Cropformer achieved an accuracy of 83.4%, a precision of 83.8%, an overall accuracy of 83.1%, an F1-score of 83.5% (Supplementary Tables 32–33), and an area under the roc curve (AUC) of 90.5% (Figure 4D), outperforming the other models in prediction.

Next, we evaluated Cropformer's ability to handle different molecular features (SNP, InDel, SV, and GE) in classification tasks. We ranked the Sopim_BGV006775_12T001232 trait values in the tomato dataset and divided them into three classes: class 1 (top 33% of samples), class 2 (middle 33% of samples), and class 3 (bottom 34% of samples). Compared with seven other methods (CropGBM, XGBoost, Support Vector Classifier (SVC), Random Forest Classifier (RFC), MLP, ridge regression best linear unbiased prediction (rrBLUP), and DEM), Cropformer consistently achieved the best phenotypic prediction performance on these test datasets (Supplemental Tables 34-41). Based on the same processing as the regression task, we extracted the 1500 most highly weighted features from the four types of genomic variants to construct fusion features. In classification tasks, the fusion data-trained Cropformer outperforms the single-genomic data-trained Cropformer (Supplemental Tables 42-43). Particularly, our Cropformer exhibited outstanding performance, outperforming the other seven methods using the fusion feature strategy.

Cropformer identifies DTT-related loci by mapping of attention weights

The attention weights underlying the multihead self-attention mechanism can reflect the importance of each locus in phenotype prediction (Figure S10). Here, we visualized the attentional weights of the loci used by the model in the training of the DTT trait data (regression task) in the Manhattan plot (Figure 5A and Extended Data 1). The highly ranked genes included Zm00001d029133, Zm00001d008941, Zm00001d011956, Zm00001d051961 and Zm00001d025617, which are known to be related to flowering time(Berr et al., 2009; Bezerra et al., 2004; Chen et al., 2017; Hong et al., 2009; Kuhn et al., 2007; Liang et al., 2014; Tan et al., 2021; Zhao et al., 2005). The Zm00001d008941 gene, also known as *ATX3*, has been reported to be involved in flowering in maize (Chen *et al.*, 2017). A haplotype analysis of *ATX3* revealed five main haplotypes in the population (Figure 5B), of which samples harbouring haplotype IV exhibited the shortest DTT, which was significantly shorter than that of samples harbouring other haplotypes (Figure 5C). Another gene, Zm00001d011956, is also known as *SDG118* and belongs to the SET domain group (SDG) protein family. The SDG family has been reported to be involved in flowering in multiple species (Berr *et al.*, 2009; Zhao *et al.*, 2005). The haplotype analysis indicated that among the 4 haplotypes observed for *SDG118* (Figures 5B and E), haplotype IV exhibited a significantly shorter DTT. These results indicated that Cropformer can effectively capture quantitative trait loci during training, ensuring powerful predictive performance.

To further expand the ability of the Cropformer framework to highlight genome regions with potential quantitative trait genes, an expansion module based on the XGBoost algorithm (Chen and Guestrin, 2016) was developed, and the SHAP values were calculated to help locate and infer candidate loci. With respect to the module, locus chr8:26,168,415 and chr8:26,166,974 were among the top two according to the SHAP values for the gene *ATX3*, which was consistent with the unique variations in haplotype IV (Figures 5D and S11). Locus chr8:165,145,056, chr8:165,145,371, and chr8:165,146,085 were highlighted on *SDG118* (Figures 5F and S12), including the divergence variations between haplotypes III and IV, as well as those between haplotypes I and IV. The results demonstrate that the Cropformer framework enables haplotype-level analysis and assists identification of trait-related genes.

Identification of loci associated with PH and EW through attention weighting

Attentional weighting was then examined to mine key SNPs associated with EW and PH traits in maize. We present a comprehensive list of attentional weights for SNPs (Extended Data 2-3). Among the SNPs, several have already been reported, suggesting the effectiveness of the list. For the maize PH trait, Zm00001d046014, Zm00001d035104, Zm00001d048865, Zm00001d026791, Zm00001d047614, and Zm00001d002567 were given increased attention from the Cropformer. Research has revealed that Zm00001d046014, a member of the cellulose synthase-like D gene family, is expressed specifically in male plants at the reproductive stage(Proost and Mutwil, 2018). For the EW trait, the Zm00001d038275, Zm00001d039865, Zm00001d050196, Zm00001d002350, Zm00001d011367, and Zm00001d050768 played a role in influencing the model's prediction. Several studies have demonstrated that Zm00001d002350 functions in the synthesis of the phytohormones gibberellin and terpenes(Wang et al., 2019; Wang et al., 2018b; Wang et al., 2023d). The whole list of focused genes can serve as a promising reference locus for future breeding improvements.

Webserver for Cropformer

For the convenience of scientific community, an easy-to-use webserver was established to implement our Cropformer, which could be freely accessed via https://cgris.net/cropformer. A step-by-step guide is given below. Step 1. Access the website at https://cgris.net/cropformer, where users will find a brief overview of Cropformer. Step 2. Click on the “Crop (e.g., maize)” button to access the user-selected prediction module. Then, click the “Example” button to download sample data in CSV format. Users can upload their own files for prediction. Step 3. Finally, click the “Run” button to obtain the predicted result (Figure 6).

Discussion

Predicting crop traits from high-density genomic data facilitates rapid selection of superior genotypes and accelerates the breeding process. As the skills and resources required for genomic selection become broadly applicable, integrating interdisciplinary and collaborative networks brings together different breeding programmes, offering unprecedented opportunities for genomic selection research. In this study, we proposed a convolution combined with a self-attention mechanism-based deep learning architecture, Cropformer, to perform genome prediction utilizing both discrete traits and quantitative traits. We compiled five high-quality crop benchmark datasets to evaluate the predictive performance of different methods. The results demonstrated that the Cropformer method outperforms other methods across the various datasets and evaluation metrics and is applicable to other similar tasks (Supplementary Table 44).

Furthermore, Cropformer demonstrates the ability to assess the contribution of input genotypes to crop phenotype prediction through the multi-head self-attention mechanism at a useful resolution. In previous studies, positional information was often discarded for genotype representations, such as those encoding genotypes as k-mer counts or those generated via PCA for dimensionality reduction. Cropformer offers two primary advantages over these methods. It employs a 0 – 9 encoding scheme for genotype features, preserving all forms of genotypes and enabling the exploration of associations between genotypes and phenotypes. With its multi-head design, the model can simultaneously and independently examine multiple regions, providing a more comprehensive assessment of each genotype's contribution to crop genome prediction. Its attention mechanism can be used to explore the correlation between genotypes and phenotypes.

The following limitations of our study need to be considered. First, the input to the model is genotype data. Crop phenotypes result from genotype–environment interactions (Fu et al., 2022; Xu et al., 2022). However, this study does not include environmental data because of challenges in their collection. Incorporating suitable genotypic and environmental predictors could provide new opportunities for GS. Secondly, while our model helps reveal the importance of SNPs and genotypes in prediction and explores their correlations, several SNPs influencing model performance have been identified. However, the biological impact requires further elucidation. With the advancements in high-throughput molecular biotechnology, integrating multi-omics data, such as metabolomics, offers the potential to further bridge genotypes and phenotypes, uncover downstream interactions, and enhance model predictive performance and interpretability (Xu et al., 2024). Finally, limited data often constrain the application of deep learning, especially when dealing with multimodal data(Qiu et al., 2020). Even though, Cropformer achieved robust and superior performance on all the test datasets.

In summary, Cropformer, as a general framework for crop genomic prediction, provides a new algorithm option for developing superior line selection methods. With Cropformer, researchers can easily perform predictive analyses on crops of interest and assess the correlation of genotypes with model predictions, demonstrating the potential for practical applications. We believe that Cropformer can accelerate the mining of valuable gene resources for crop improvement, enhancing the progress of genomic-design crop breeding and provide a valuable resource for future crop improvement breeding.

Methods

Dataset

We analysed data from five species representing various population sizes and different reproductive systems. The published datasets used in this manuscript are available from websites or the literature: (1) the maize dataset(Liu et al., 2020); (2) the tomato dataset(Zhou et al., 2022); (3) the rice dataset (Oryza sativa L.)(Wang et al., 2018a); (4) the foxtail millet dataset (Setaria italica)(He et al., 2023); and (5) the wheat dataset(Crossa et al., 2016), which can be downloaded from https://hdl.handle.net/11529/10548918.

The maize dataset consisted of 1428 inbred lines derived from 24 founding female crosses(Liu *et al.*, 2020). Three phenotypic traits, days to tasselling (DTT), plant height (PH) and ear weight (EW), were measured in 8652 F1 hybrids at five locations. The procedure for SNP calling and genotype processing of the 8652 samples has been described by Liu et al(Liu *et al.*, 2020)(https://ftp.cngb.org/pub/CNSA/data3/CNP0001565/zeamap/99_MaizegoResources/01_CUBIC_related/). Furthermore, the core SNP set was screened using PLINK(Purcell et al., 2007), where SNPs were removed by linkage disequilibrium pruning with a window size of 1 kb, window step of 100 SNPs, and a r2 threshold of 0.1, resulting in 32,519 SNPs.

We removed the samples containing missing values and finally retained 8439 samples. These maize samples were randomly divided into training set of 6751 samples and test set of 1688 samples in a ratio of 8:2 (Supplementary Table 1, and Figure S2A). To facilitate the calculation, we computed the maximum information coefficient (MIC) (Wang et al., 2023a) of the SNPs in the training dataset and selected the top 10,000 SNPs by ranking them according to the weight of the MIC. Based on the indexing of 10,000 SNPs from the training dataset, the corresponding SNPs are extracted from the test dataset. This ensures that the performance evaluation is objective enough.

The wheat dataset was derived from 2403 Iranian bread wheat (Triticum aestivum) landrace wheat accessions in the CIMMYT wheat gene bank (https://hdl.handle.net/11529/10548918). The dataset was genotyped for these alleles using 33,709 DArT markers, with each allele recorded as 1 (present) or 0 (absent) in each variety(Crossa *et al.*, 2016). For the wheat dataset, the traits measured included thousand-kernel weight (TKW), grain width (GW), grain hardness (GH), grain protein (GP), and grain length (GL). The same strategy was used to select the top 10,000 features based on the MIC, and samples containing missing values were removed, resulting in 2,000 samples. These wheat samples were randomly divided into a 1600-sample training set and a 400-sample testing set at a ratio of 8:2 (Supplementary Table 1, and Figure S2B).

The 3,000 Rice Genomes Project is a gigabyte dataset of genome sequences from 3,000 rice varieties that can represent the genetic and functional diversity of rice on a global scale(Li et al., 2014; Wang *et al.*, 2018a). The rice dataset includes the phenotypes of five measured traits, namely, Culm_length, Days_to_heading_2018H, Grain_length_width_ratio, Plant_height_2018HN, and Thousand_grain_weight (https://snp-seek.irri.org/_download.zul). The 404k core SNP dataset of the rice dataset was downloaded from https://snpseek.irri.org/_download.zul, and the top 10,000 SNPs were selected based on the MIC. The same strategy was applied to remove missing values and the segmented rice dataset, resulting in 2799 samples, a training set containing 2239 samples and a testing set containing 560 samples (Supplementary Table 1, and Figure S2C).

In the foxtail millet dataset, 680 foxtail millet accessions from 13 different geographic locations were sequenced by He et al.(He *et al.*, 2023) (https://www.cgris.net/millet). This dataset includes the phenotypes of five measured traits, namely, straw weight (Anyang), straw weight (Beijing), straw weight (Changzhi), straw weight (Dingxi), and straw weight (Urumqi). We used the high-effect marker

SNPs identified by He et al(He *et al.*, 2023). as feature inputs to the model and obtained 666 samples after removing missing values. These foxtail millet samples were randomly divided into a 566-sample training set and a 100-sample testing set at a ratio of 8:2 (Supplementary Table 1, and Figure S2D).

The tomato dataset was a call set (designated TGG1.1–332) from the tomato graph pangenome consisting of 6,971,059 SNPs, 657,549 InDels, 51,155 GEs, and 54,838 SVs(Zhou, 2022) (http://solomics.agis.org.cn/tomato/ftp/genotypes/). An important traits (Sopim_BGV006775_12T001232) associated with tomato yield and flavor were used for study and analysis. We pruned the SNPs, InDels, and SVs(Zhou, 2022) using PLINK and MIC to obtain the top 10,000 features and removed phenotypes containing missing values, resulting in 332 samples. These tomato samples were randomly divided into a 265-sample training set and a 67-sample testing set at a ratio of 8:2 (Supplementary Table 1, and Figure S2E).

During the data splitting process, we set a random seed. The introduction of a random seed ensures that there are no specific patterns or correlations between different parts of the dataset, thereby making the resulting training and testing sets representative and accurately assessing the model's generalization ability. This approach also ensures that the data splitting procedure remains uniform across different traits, facilitating fair and reliable comparisons of multi-trait predictability. We applied MIC analysis to the maize, wheat, and tomato datasets, selecting the top 10,000 features based on their importance ranking. For the foxtail millet (Setaria italica) and rice datasets, we utilized the core SNPs provided by He et al.(He *et al.*, 2023) and Liu et al.(Wang *et al.*, 2018a), as the feature dimensions did not exceed 10,000, and thus, further MIC processing was not performed.

Feature representations for genotypic data

For ease of inputting the data into the model and interpreting the features, we coded the SNP information using 0–9 as follows: AA (0), AT (1), TA (1), AC (2), CA (2), AG (3), GA (3), TT (4), TC (5), CT (5), TG (6), GT (6), CC (7), CG (8), GC (8), and GG (9).

For the InDel and SV information, we used PLINK to encode them as 0, 1, or 2. For all models, we use the same feature representation scheme to train and test to ensure fairness of comparison. For different gene variants, we extracted the top 1500 MIC weighted features and vertically merged them to train the models.

MIC

The core idea of MIC is: if there is a relationship between two variables, there will be a grid that can split the scatter graph of the two variables to encapsulate this relationship, and then normalize these mutual information values to ensure a fair comparison between grids of different dimensions (Albanese et al., 2013; Reshef et al., 2011; Zhou et al., 2004)

$$\mathrm{I(X;Y)} = \sum_{\mathrm{x,y}} \mathrm{p(x,y)log}\frac{\mathrm{p(x,y)}}{\mathrm{p(x)p(y)}} = \mathrm{H(X)} - \mathrm{H(X|Y)}$$

Where $\|\mathrm{x}-\mathrm{c_i}\|$ represents Euclidean norm; $\mathrm{c_i}$, $\mathrm{R_i}$ and σ_i are the center, the width and the output of the $\mathrm{i\_th}$ hidden unit, respectively.

Cropformer architecture

We introduce Cropformer, a hybrid network based on a convolutional neural network (CNN) combined with a multihead self-attention mechanism that accurately predicts the phenotypic performance of plants from their genome features. The model accepts sequence information of variable lengths. To utilize the mini-batch technique for training and prediction, we fix the length of the input sequence at 10,000 nt. We employed the Maximum Information Coefficient (MIC) method to identify the top 10,000 SNPs with high weights that are closely associated with the phenotype. Specifically, the data pass through a convolutional layer that employs a kernel size of 3×3, with a stride (step size) of one and padding set to one. This configuration is designed to ensure that the dimensionality of the output matches that of the input.

The core component of our network is a multihead self-attention layer. The multi-head self-attention mechanism is used to assess the contribution of sequence regions for localization by multiple heads (head = 8), which has the ability to further detect localization SNPs during the prediction. We borrow the idea proposed by Bengio et al.(Zhouhan Lin et al., 2017) that the overall semantics of a sentence are composed of multiple constituents and that a multihead self-attention mechanism can be used to address different parts of the sentence. Attention can model the dependence of CNN-fed data regardless of their distance, a property we use to capture core SNPs. The attention matrix of self-attention can be obtained by computing the vectors Query (Q), Key (K) and Value (V). The input of the attention layer and its two linear transformations, Q and K, are defined as follows(Ullah and Ben-Hur, 2021) :

$$Q = W_Q^T X$$

$$K = W_K^T X$$

where W_Q and W_K are the corresponding weight matrices for Q and K, respectively. The attention matrix A is then computed using the following expression:

$$A(Q, K) = \mathrm{softmax}(\frac{QK^T}{\sqrt{d_k}})$$

where d_k is the dimension of K. The SoftMax function is applied to each row of the matrix $\frac{QK^T}{\sqrt{d_k}}$, ensuring that the elements of each row sum to 1.

To generate the output of the attention layer, we define the value matrix:

$$V = W_V^T X$$

Finally, we define the output of the attention layer as follows:

$$Z = A \times V$$

Next, we perform a linear transformation of the reshaped data, introduce dropou and normalize the output, which is effective in terms of computational efficiency and leads to better model accuracy. To avoid overfitting, we used early stopping in Cropformer. Finally, for continuous traits, we use the mean square error to define the loss function, and for qualitative traits, we define the loss function using CrossEntropy.

Attention weights

In practical terms, the self-attention mechanism allowed the inputs to interact with themselves and determined which element should receive more attention(Liu et al., 2023b). The attention mechanism was described as mapping a query and a set of key–value pairs to an output, where the query, keys, values, and output were all vectors. We used the excellent data feature extraction potential of CNNs to process encoded genotype data without changing the length of the sequence(Garcia-Gasulla et al., 2018). The output of the CNN was the same length as the input data, so we could calculate the overall attention weights and generate an attention vector for each input(Yan et al., 2022). For each dimension, the attention score indicated the importance of the dimension for model prediction.

In this study, we employed the dynamic weight allocation mechanism to capture attention scores. Specifically, each attention head's output was weighted according to its importance score (Head Importance Score), which was dynamically updated throughout the training. This mechanism ensured that attention heads contributing more significantly to the task received higher weights, thereby preventing the loss of critical information. During training, the importance score of each attention head was learned adaptively, allowing the model to adjust the contribution of each attention head according to its relevance to the task. The final output was a weighted combination of each attention head, where attention heads with higher importance scores contributed more. To ensure the output of each attention head was appropriately scaled before merging, we normalized the importance scores, defined as:

$$\alpha_i = \frac{S_i}{\sum_{j=1}^{N} S_i}$$

where α_i is the weight for the i-th attention head, and S_i is its importance score.

Multimodal data integration

Advances in next-generation sequencing technologies have led to a proliferation of multimodal datasets. The multimodal data, including SNP, InDel, GE, and SV, from 332 tomato samples were used for further analysis. For the SNP data, we employed a 0–9 coding scheme, the details of which are provided in “Feature representations for genotypic". With respect to the InDel and SV information, we utilized PLINK to encode them as 0, 1, or 2. For each modality, we adopt columnwise concatenation to construct fused features for model training.

Clustering

The StandardScaler function of scikit-learn (version: 1.5.1) was used by us to normalize the three-classification data and the two-classification dataset respectively. We use matplotlib for the visualization (version: 3.7.5). The python package umap-learn version 0.5.3 was used for UMAP visualization.

Haplotype analysis

We performed haplotype analysis and generated haplotype networks with Pegas 0.11 (Paradis, 2010) in R. We utilized the ggplot2, gghalves, and ggpubr packages in R to generate boxplots of the DTT trait with at test for different haplotypes. For the gene structure plot, annotation information for the Zm00001d008941_T001 transcript was first extracted from the GFF file (B73, v4.48); subsequently, the three_prime_UTR and five_prime_UTR were plotted as white-filled rectangles, while the CDS features were plotted as red-filled rectangles. The length and relative position of those rectangles follow their physical positions. The physical positions of the various loci were mapped to the gene structure and are marked as red vertical lines. For the genotype heatmap of haplotypes, the consensus genotype for each haplotype at each mutation locus was defined as the genotype with the highest frequency at that locus within the population corresponding to the haplotype, and the consensus genotypes were then plotted in different colours (grey, light blue, and dark blue for the reference genotype, heterozygous mutation, and homozygous mutation, respectively). The gene structure plot and the genotype heatmap of Zm00001d011956 were generated in the same way.

SHapley Additive exPlanations

SHAP (SHapley Additive exPlanations) is a commonly used explanatory machine learning model that shows the magnitude of the overall contribution of features to the prediction of the the whole dataset(Lundberg and Lee, 2017; Qiu et al., 2022; Tang et al., 2023). Based on the highly weighted SNPs extracted by the self-attention mechanism, we annotated them and selected genes Zm00001d011956 and Zm00001d008941 associated with flowering time. For the locus of both genes, we searched for SNPs within an extended region of 400 kbp (half the LD length). We use Explainer, which provides a localized explanation of the impact of input SNPs on the individual predictions of the XGBoost model. Here, a higher SHAP value means more weight.

Evaluation metrics

We use five outer and three inner nested cross-validation to partition the training datasets(Cawley and Talbot, 2010). The inner layer cross-validation is used for hyperparameter optimization and outer layer cross-validation is used to evaluate the generalization performance of the model. Finally, the robustness of the model is evaluated on the test datasets. For qualitative traits, accuracy, recall, precision, and F1_score metrics were used to quantify the performance of the model and are defined as follows(Liu et al., 2023a; Wang et al., 2021; Wang et al., 2023c):

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

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

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

$$\text{F1\_score} = \frac{2 * (\text{precision} * \text{recall})}{\text{precision} + \text{recall}}$$

where TP, TN, FP, and FN represent the numbers of true positives, true negatives, false-positives, and false-negatives, respectively.

The AUC is an indicator of a classification model's performance, representing its ability to classify at varying thresholds. It evaluates the classification effect of the model by calculating the area under the ROC curve, and the closer the AUC value is to 1, the better the classification performance of the model. The Pearson correlation coefficient is used to assess the predictive performance of the model in continuous trait tasks by measuring the linear relationship between true and predicted values. A coefficient closer to 1 indicates a higher predictive accuracy of the model.

Data availability and Code availability

Some of the data that support the findings of this study are publicly available, and some are proprietary datasets provided for this analysis under collaboration agreements. The raw whole genome sequencing of maize is available at NCBI under BioProject Accession No. PRJNA597703. Rice sequencing data are available through NCBI under project accession number PRJEB6180. The tomato dataset can be found in the SolOmics database (http://solomics.agis.org.cn/tomato/ftp).The wheat dataset can be found on the website (https://hdl.handle.net/11529/10548918). The foxtail millet dataset can be found at this link (https://www.cgris.net/millet). The Cropfomer software including documents and tutorial is available on Github (https://github.com/jiekesen/Cropformer).

Acknowledgements

We thank all the investigators and participants in this study. We thank Chunhui Li for his significant contributions in the field of important genes related to Maize DTT phenotype. This Study was supported by National Natural Science Foundation of China (Grant No. 32371996 to S. Y.), and Central Public-interest Scientifc Institution Basal Research Fund of China (S2023QH09 to S. Y.), and the Agricultural Science and Technology Innovation Program (CAAS-ASTIP-2023-ICS01 to the Innovation Team of Crop Germplasm Resources Preservation and Information).

Contributions

H.W., S.Y., W.X.W., W.F., W.L.G., and Y.Q.C. designed this study and wrote the paper. H.W. built the deep learning models. W.X.W., Y.M.C, J.P.H., and Y.Q.C. processed and analyzed the data. H.W., Q.H., and X.M.D. collected the dataset and performed data preprocessing. Y.N.L., Y.S.C., and Y.Q.C., conceived the project and edited the paper. All authors reviewed and approved the final manuscript for submission.

These authors contributed equally: Hao Wang Shen Yan Wenxi Wang.

Corresponding authors

Correspondence to Yongsheng Cao, Weilong Guo, Wei Fang

Competing interests

The authors declare no competing interests.

References

Albanese, D., Filosi, M., Visintainer, R., Riccadonna, S., Jurman, G., and Furlanello, C. (2013). Minerva and minepy: a C engine for the MINE suite and its R, Python and MATLAB wrappers. Bioinformatics **29**:407-408. 10.1093/bioinformatics/bts707.

Berr, A., Xu, L., Gao, J., Cognat, V., Steinmetz, A., Dong, A.W., and Shen, W.H. (2009). Encodes a Histone Methyltransferase and Is Involved in Activation and Repression of Flowering. Plant Physiology **151**:1476-1485. 10.1104/pp.109.143941.

Bezerra, I.C., Michaels, S.D., Schomburg, F.M., and Amasino, R.M. (2004). Lesions in the mRNA cap-binding gene suppress -mediated delayed flowering in. Plant J **40**:112-119. 10.1111/j.1365-313X.2004.02194.x.

Cawley, G.C., and Talbot, N.L. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. The Journal of Machine Learning Research **11**:2079-2107.

Chen, L.-Q., Luo, J.-H., Cui, Z.-H., Xue, M., Wang, L., Zhang, X.-Y., Pawlowski, W.P., and He, Y. (2017). ATX3, ATX4, and ATX5 Encode Putative H3K4 Methyltransferases and Are Critical for Plant Development. Plant Physiology **174**:1795-1806. 10.1104/pp.16.01944.

Chen, T.Q., and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Kdd'16: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining:785-794. 10.1145/2939672.2939785.

Covarrubias-Pazaran, G. (2016). Genome-Assisted Prediction of Quantitative Traits Using the R Package sommer. PLoS One **11**:e0156744. 10.1371/journal.pone.0156744.

Crossa, J., Jarquin, D., Franco, J., Perez-Rodriguez, P., Burgueno, J., Saint-Pierre, C., Vikram, P., Sansaloni, C., Petroli, C., Akdemir, D., et al. (2016). Genomic Prediction of Gene Bank Wheat Landraces. G3 (Bethesda) **6**:1819-1834. 10.1534/g3.116.029637.

Desta, Z.A., and Ortiz, R. (2014). Genomic selection: genome-wide prediction in plant improvement. Trends Plant Sci **19**:592-601. 10.1016/j.tplants.2014.05.006.

Endelman, J.B. (2011a). Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. The Plant Genome **4**:250-255. 10.3835/plantgenome2011.08.0024.

Endelman, J.B. (2011b). Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. The Plant Genome **4**https://doi.org/10.3835/plantgenome2011.08.0024.

Fu, J., Hao, Y., Li, H., Reif, J.C., Chen, S., Huang, C., Wang, G., Li, X., Xu, Y., and Li, L. (2022). Integration of genomic selection with doubled-haploid evaluation in hybrid breeding: From GS 1.0 to GS 4.0 and beyond. Mol Plant **15**:577-580. 10.1016/j.molp.2022.02.005.

Gao, P., Zhao, H., Luo, Z., Lin, Y., Feng, W., Li, Y., Kong, F., Li, X., Fang, C., and Wang, X. (2023). SoyDNGP: a web-accessible deep learning framework for genomic prediction in soybean breeding. Brief Bioinform **24**10.1093/bib/bbad349.

Garcia-Gasulla, D., Pares, F., Vilalta, A., Moreno, J., Ayguade, E., Labarta, J., Cortes, U., and Suzumura, T. (2018). On the Behavior of Convolutional Nets for Feature Extraction. J Artif Intell

Res **61**:563-592. DOI 10.1613/jair.5756.

Habyarimana, E., Lopez-Cruz, M., and Baloch, F.S. (2020). Genomic Selection for Optimum Index with Dry Biomass Yield, Dry Mass Fraction of Fresh Material, and Plant Height in Biomass Sorghum. Genes (Basel) **11**10.3390/genes11010061.

He, Q., Tang, S., Zhi, H., Chen, J., Zhang, J., Liang, H., Alam, O., Li, H., Zhang, H., Xing, L., et al. (2023). A graph-based genome and pan-genome variation of the model plant Setaria. Nat Genet **55**:1232-1242. 10.1038/s41588-023-01423-w.

Hickey, J.M., Chiurugwi, T., Mackay, I., Powell, W., and Implementing Genomic Selection in, C.B.P.W.P. (2017). Genomic prediction unifies animal and plant breeding programs to form platforms for biological discovery. Nat Genet **49**:1297-1303. 10.1038/ng.3920.

Hong, E.H., Jeong, Y.M., Ryu, J.Y., Amasino, R.M., Noh, B., and Noh, Y.S. (2009). Temporal and spatial expression patterns of nine genes encoding Jumonji C-domain proteins. Mol Cells **27**:481-490. 10.1007/s10059-009-0054-7.

Krishnappa, G., Savadi, S., Tyagi, B.S., Singh, S.K., Mamrutha, H.M., Kumar, S., Mishra, C.N., Khan, H., Gangadhara, K., Uday, G., et al. (2021). Integrated genomic selection for rapid improvement of crops. Genomics **113**:1070-1086. 10.1016/j.ygeno.2021.02.007.

Krizhevsky, A., Sutskever, I., and Hinton, G.E. (2017). ImageNet Classification with Deep Convolutional Neural Networks. Communications of the ACM **60**:84-90. 10.1145/3065386.

Kuhn, J.M., Breton, G., and Schroeder, J.I. (2007). mRNA metabolism of flowering-time regulators in wild-type Arabidopsis revealed by a nuclear cap binding protein mutant,. Plant J **50**:1049-1062. 10.1111/j.1365-313X.2007.03110.x.

Li, J.Y., Wang, J., and Zeigler, R.S. (2014). The 3,000 rice genomes project: new opportunities and challenges for future rice research. Gigascience **3**:8. 10.1186/2047-217X-3-8.

Liang, G., He, H., Li, Y., Wang, F., and Yu, D.Q. (2014). Molecular Mechanism of microRNA396 Mediating Pistil Development in Arabidopsis. Plant Physiology **164**:249-258. 10.1104/pp.113.225144.

Liu, H.J., Wang, X., Xiao, Y., Luo, J., Qiao, F., Yang, W., Zhang, R., Meng, Y., Sun, J., Yan, S., et al. (2020). CUBIC: an atlas of genetic architecture promises directed maize improvement. Genome Biol **21**:20. 10.1186/s13059-020-1930-x.

Liu, M., Zhou, J., Xi, Q., Liang, Y., Li, H., Liang, P., Guo, Y., Liu, M., Temuqile, T., Yang, L., et al. (2023a). A computational framework of routine test data for the cost-effective chronic disease prediction. Brief Bioinform **24**10.1093/bib/bbad054.

Liu, T., Zou, B., He, M., Hu, Y., Dou, Y., Cui, T., Tan, P., Li, S., Rao, S., Huang, Y., et al. (2023b). LncReader: identification of dual functional long noncoding RNAs using a multi-head self-attention mechanism. Brief Bioinform **24**10.1093/bib/bbac579.

Lundberg, S.M., and Lee, S.-I. (2017). A unified approach to interpreting model predictions. Proceedings of the 31st International Conference on Neural Information Processing Systems. Curran Associates Inc.

Ma, W., Qiu, Z., Song, J., Li, J., Cheng, Q., Zhai, J., and Ma, C. (2018). A deep convolutional neural network approach for predicting phenotypes from genotypes. Planta **248**:1307-1318.

10.1007/s00425-018-2976-9.

Meuwissen, T.H., Hayes, B.J., and Goddard, M.E. (2001). Prediction of total genetic value using genome-wide dense marker maps. Genetics **157**:1819-1829. 10.1093/genetics/157.4.1819.

Misztal, I. (2008). Reliable computing in estimation of variance components. J Anim Breed Genet **125**:363-370. 10.1111/j.1439-0388.2008.00774.x.

Paradis, E. (2010). pegas: an R package for population genetics with an integrated-modular approach. Bioinformatics **26**:419-420. 10.1093/bioinformatics/btp696.

Proost, S., and Mutwil, M. (2018). CoNekT: an open-source framework for comparative genomic and transcriptomic network analyses. Nucleic Acids Res **46**:W133-W140. 10.1093/nar/gky336.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet **81**:559-575. 10.1086/519795.

Qiu, W., Chen, H., Dincer, A.B., Lundberg, S., Kaeberlein, M., and Lee, S.I. (2022). Interpretable machine learning prediction of all-cause mortality. Commun Med (Lond) **2**:125. 10.1038/s43856-022-00180-x.

Qiu, Y.L., Zheng, H., Devos, A., Selby, H., and Gevaert, O. (2020). A meta-learning approach for genomic survival analysis. Nat Commun **11**:6350. 10.1038/s41467-020-20167-3.

Ren, Y., Wu, C., Zhou, H., Hu, X., and Miao, Z. (2024). Dual-extraction modeling: A multi-modal deep-learning architecture for phenotypic prediction and functional gene mining of complex traits. Plant Commun **5**:101002. 10.1016/j.xplc.2024.101002.

Reshef, D.N., Reshef, Y.A., Finucane, H.K., Grossman, S.R., McVean, G., Turnbaugh, P.J., Lander, E.S., Mitzenmacher, M., and Sabeti, P.C. (2011). Detecting novel associations in large data sets. Science **334**:1518-1524. 10.1126/science.1205438.

Tan, J.R., Yi, X.W., Luo, L., Yu, C., Wang, J., Cheng, T.R., Zhang, Q.X., and Pan, H.T. (2021). RNA-seq and sRNA-seq analysis in lateral buds and leaves of juvenile and adult roses. Sci Hortic-Amsterdam **290** ARTN 110513 10.1016/j.scienta.2021.110513.

Tang, X., Zhang, J., He, Y., Zhang, X., Lin, Z., Partarrieu, S., Hanna, E.B., Ren, Z., Shen, H., Yang, Y., et al. (2023). Explainable multi-task learning for multi-modality biological data analysis. Nat Commun **14**:2546. 10.1038/s41467-023-37477-x.

Tong, H., and Nikoloski, Z. (2021). Machine learning approaches for crop improvement: Leveraging phenotypic and genotypic big data. J Plant Physiol **257**:153354. 10.1016/j.jplph.2020.153354.

Tong, H., Kuken, A., and Nikoloski, Z. (2020). Integrating molecular markers into metabolic models improves genomic selection for Arabidopsis growth. Nat Commun **11**:2410. 10.1038/s41467-020-16279-5.

Ullah, F., and Ben-Hur, A. (2021). A self-attention model for inferring cooperativity between regulatory features. Nucleic Acids Res **49** ARTN e77 10.1093/nar/gkab349.

Varshney, R.K., Graner, A., and Sorrells, M.E. (2005). Genomics-assisted breeding for crop improvement. Trends Plant Sci **10**:621-630. 10.1016/j.tplants.2005.10.004.

Wallace, J.G., Rodgers-Melnick, E., and Buckler, E.S. (2018). On the Road to Breeding 4.0:

Unraveling the Good, the Bad, and the Boring of Crop Quantitative Genomics. Annu Rev Genet **52**:421-444. 10.1146/annurev-genet-120116-024846.

Wang, H., Liang, P., Zheng, L., Long, C., Li, H., and Zuo, Y. (2021). eHSCPr discriminating the cell identity involved in endothelial to hematopoietic transition. Bioinformatics **37**:2157-2164. 10.1093/bioinformatics/btab071.

Wang, H., Zhang, Z., Li, H., Li, J., Li, H., Liu, M., Liang, P., Xi, Q., Xing, Y., Yang, L., et al. (2023a). A cost-effective machine learning-based method for preeclampsia risk assessment and driver genes discovery. Cell Biosci **13**:41. 10.1186/s13578-023-00991-y.

Wang, K., Abid, M.A., Rasheed, A., Crossa, J., Hearne, S., and Li, H. (2023b). DNNGP, a deep neural network-based method for genomic prediction using multi-omics data in plants. Mol Plant **16**:279-293. 10.1016/j.molp.2022.11.004.

Wang, R., Jiang, Y., Jin, J., Yin, C., Yu, H., Wang, F., Feng, J., Su, R., Nakai, K., Zou, Q., et al. (2023c). DeepBIO: an automated and interpretable deep-learning platform for high-throughput biological sequence prediction, functional annotation and visualization analysis. Nucleic Acids Res **51**:3017-3029. 10.1093/nar/gkad055.

Wang, W., Mauleon, R., Hu, Z., Chebotarov, D., Tai, S., Wu, Z., Li, M., Zheng, T., Fuentes, R.R., Zhang, F., et al. (2018a). Genomic variation in 3,010 diverse accessions of Asian cultivated rice. Nature **557**:43-49. 10.1038/s41586-018-0063-9.

Wang, Y., Wang, X., Deng, D., and Wang, Y. (2019). Maize transcriptomic repertoires respond to gibberellin stimulation. Mol Biol Rep **46**:4409-4421. 10.1007/s11033-019-04896-3.

Wang, Y., Wang, Y., Zhao, J., Huang, J., Shi, Y., and Deng, D. (2018b). Unveiling gibberellin-responsive coding and long noncoding RNAs in maize. Plant Mol Biol **98**:427-438. 10.1007/s11103-018-0788-8.

Wang, Y., Zou, J., Li, J., Kong, F., Xu, L., Xu, D., Li, J., Yang, H., Zhang, L., Li, T., et al. (2023d). Identification and functional analysis of ZmDLS associated with the response to biotic stress in maize. Front Plant Sci **14**:1162826. 10.3389/fpls.2023.1162826.

Werner, C.R., Gaynor, R.C., Gorjanc, G., Hickey, J.M., Kox, T., Abbadi, A., Leckband, G., Snowdon, R., and Stahl, A. (2020). How Population Structure Impacts Genomic Selection Accuracy in Cross-Validation: Implications for Practical Breeding. Front Plant Sci **11**ARTN 592977 10.3389/fpls.2020.592977.

Xu, Y., Lu, Y., Xie, C., Gao, S., Wan, J., and Prasanna, B.M. (2012). Whole-genome strategies for marker-assisted plant breeding. Molecular Breeding **29**:833-854. 10.1007/s11032-012-9699-6.

Xu, Y., Zhang, X., Li, H., Zheng, H., Zhang, J., Olsen, M.S., Varshney, R.K., Prasanna, B.M., and Qian, Q. (2022). Smart breeding driven by big data, artificial intelligence, and integrated genomic-enviromic prediction. Mol Plant **15**:1664-1695. 10.1016/j.molp.2022.09.001.

Xu, Y., Yang, W., Qiu, J., Zhou, K., Yu, G., Zhang, Y., Wang, X., Jiao, Y., Wang, X., Hu, S., et al. (2024). Metabolic marker-assisted genomic prediction improves hybrid breeding. Plant Commun:101199. 10.1016/j.xplc.2024.101199.

Yan, J., Xu, Y., Cheng, Q., Jiang, S., Wang, Q., Xiao, Y., Ma, C., Yan, J., and Wang, X. (2021). LightGBM: accelerated genomically designed crop breeding through ensemble learning. Genome Biol **22**:271. 10.1186/s13059-021-02492-y.

Yan, W., Li, Z., Pian, C., and Wu, Y. (2022). PlantBind: an attention-based multi-label neural network for predicting plant transcription factor binding sites. Brief Bioinform **23** 10.1093/bib/bbac425.

Zhao, Z., Yu, Y., Meyer, D., Wu, C., and Shen, W.H. (2005). Prevention of early flowering by expression of FLOWERING LOCUS C requires methylation of histone H3 K36. Nat Cell Biol **7**:1256-1260. 10.1038/ncb1329.

Zhou, H. (2022). On C-E Translation of Chinese Picture Books on COVID-19 for Children from the Perspective of Skopos Theory—Taking Agan Will Win as An Example. Journal of Educational Research and Policies **4**.

Zhou, X., Wang, X., Dougherty, E.R., Russ, D., and Suh, E. (2004). Gene clustering based on clusterwide mutual information. J Comput Biol **11**:147-161. 10.1089/106652704773416939.

Zhou, Y., Zhang, Z., Bao, Z., Li, H., Lyu, Y., Zan, Y., Wu, Y., Cheng, L., Fang, Y., Wu, K., et al. (2022). Graph pangenome captures missing heritability and empowers tomato breeding. Nature **606**:527-534. 10.1038/s41586-022-04808-9.

Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Bengio, Y. (2017). A Structured Self-attentive Sentence Embedding. arXiv 10.48550/arXiv.1703.03130.

Figures

Figure 1. Workflow of the proposed Cropformer framework. **A** We collected genotype information for five crops. Then, we convert the genotype information into a "one-hot code" representation and input it into the neural network for trait prediction. **B** The Cropformer model mainly consists of CNN filters and a multihead self-attention layer. The CNN layer is used to capture the localization signals of SNPs, and multihead self-attention is used to make the model more focused on important SNPs. **C** From left to right, the sequence shows the results of haplotype analysis, attention weight visualization, feature importance assessment (SHapley Additive exPlanations (SHAP) based explanation of machine learning model outputs.), and clustering analysis.

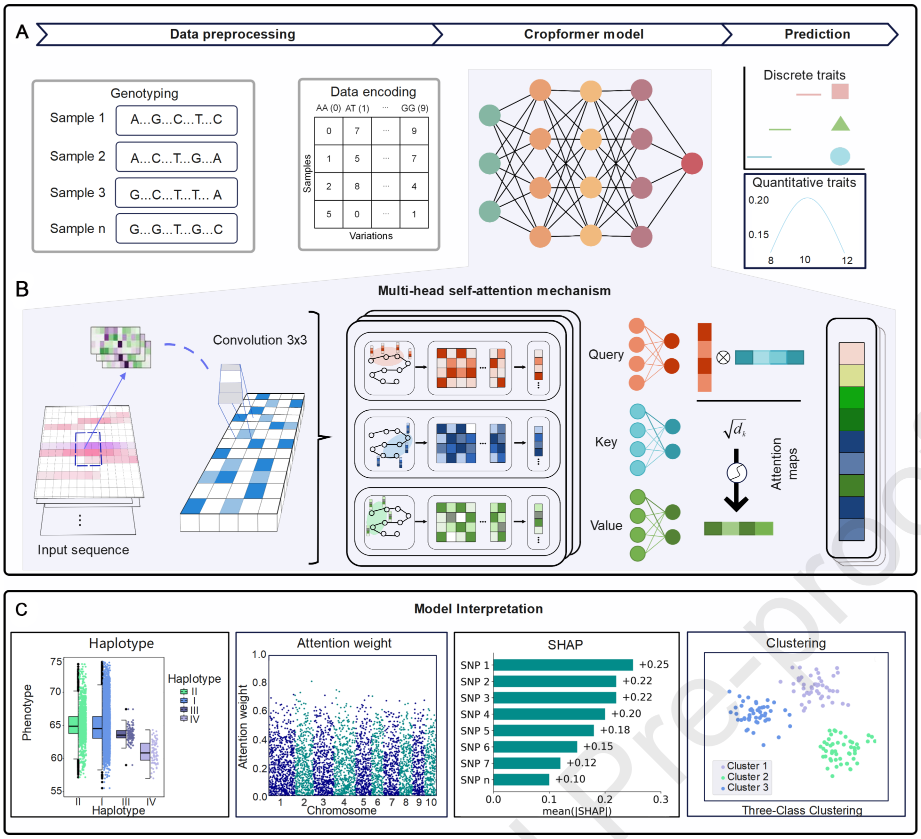
Figure 2. Predictive performance of the Cropformer model on mvvaize data (Train and Test datasets, regression task). **A** The phenotypic distributions of ear weight (EW), plant height (PH), and days to tasselling (DTT) of the maize dataset in the training and test datasets. **B** Comparison of predictive performance of different models on DTT, PH and EW traits in maize for training set (nested cross-validation) and test set. These models include our model, the CropGBM, the DNNGP, XGBoost, SVR, MLP, rrBLUP, and DEM. Model performance was measured using Pearson correlation coefficient.

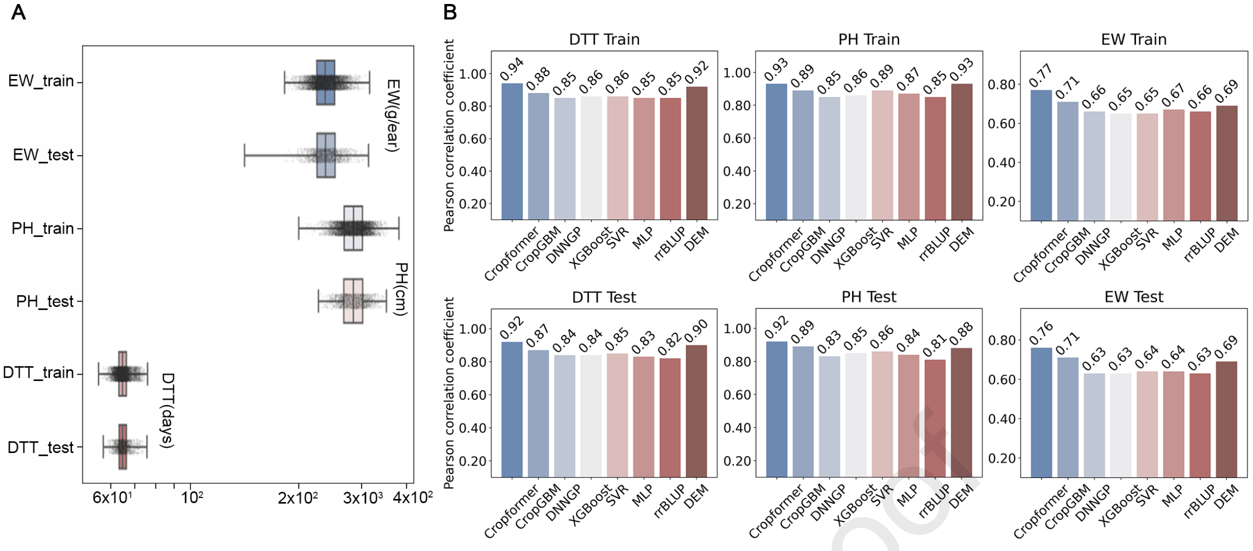
Figure 3. The predictive performance of the Cropformer model on the test datasets of wheat, foxtail millet, rice and tomato (continuous traits, regression task). **A** The predictive performance of different algorithms for five traits, namely, thousand-kernel weight (TKW), grain width (GW), grain hardness (GH), grain protein (GP), and grain length (GL), on the wheat dataset. **B** The prediction performance of different algorithms on the foxtail millet dataset was compared for the straw weight trait from five regions, namely, Anyang, Beijing, Changzhi, Dingxi, and Urumqi. **C** Predictive performance for five traits, Culm_length, Days_to_heading_2018H, Grain_length_width_ratio, Plant_height_2018HN, and Thousand_grain_weight, on the rice dataset according to the different algorithms. **D** Based on the genomic variation information, including single nucleotide polymorphism (SNP), insertion deletion (InDel), gene expression (GE), structural variation (SV), and the fusion of these four types of information, we compared the modelling performance of different algorithms for the Sopim_BGV006775_12T001232 trait in the tomato dataset.

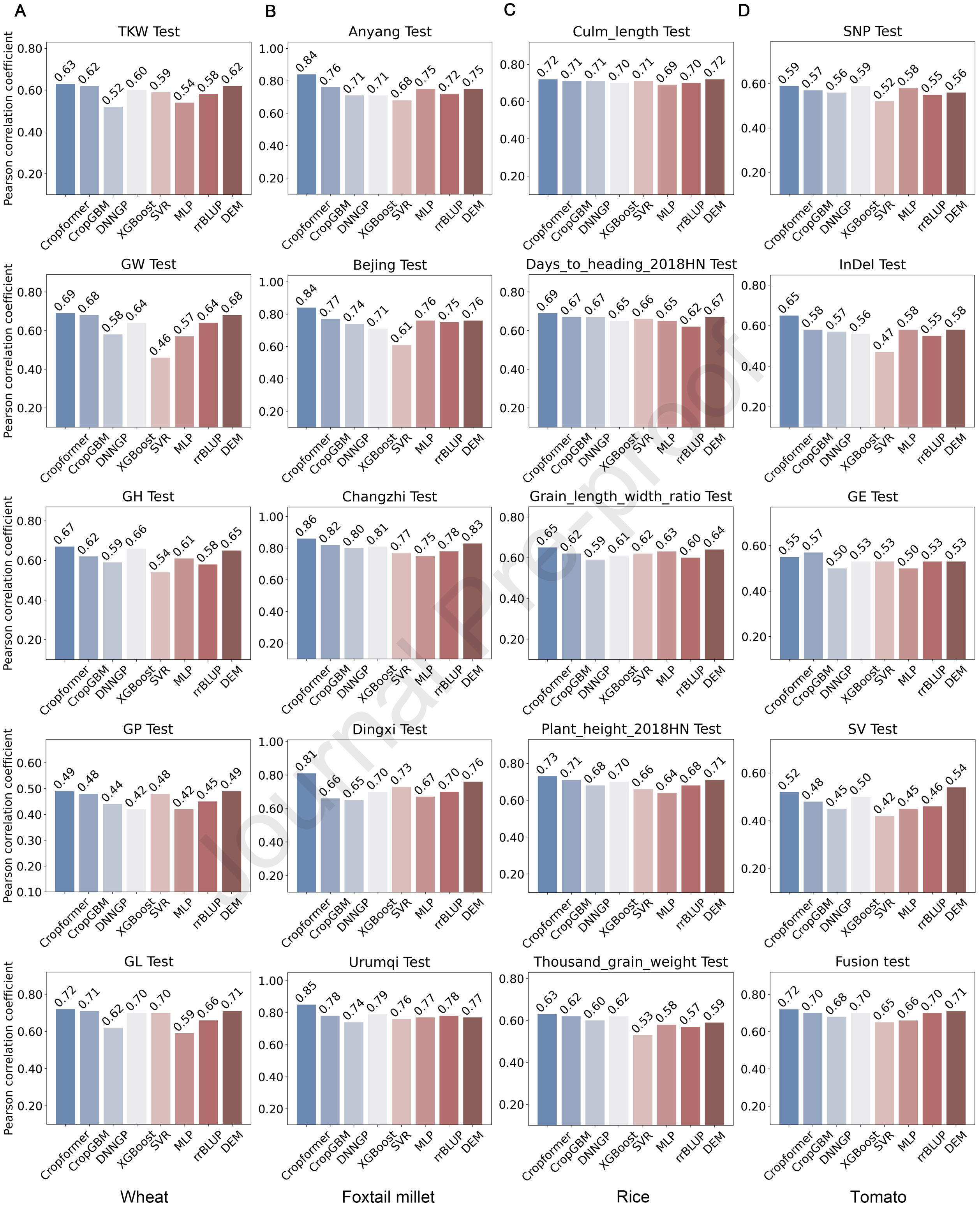
Figure 4. Classification prediction performance of the Cropformer model on the maize dataset (10,000 SNPs, classification task). **A** UMAP visualization of all the SNPs and the 10,000 SNPs extracted from the MIC. From left to right, there are three classifications and two classifications. **B** Comparison of the accuracy of different models on the maize training (nested cross-validation) and test datasets. **C** Comprehensive predictive evaluation of the Cropformer model on a maize test dataset with five metrics: Accuracy, Precision, Recall, F1_score, and Area under the curve (AUC). **D** Comparison of different models for classification of early flowering time (first 25% DTT), moderate flowering time (25 to 75% DTT) and late flowering time (last 25% DTT) DTT based on 10,000 SNPs. The numbers in brackets are AUC values.

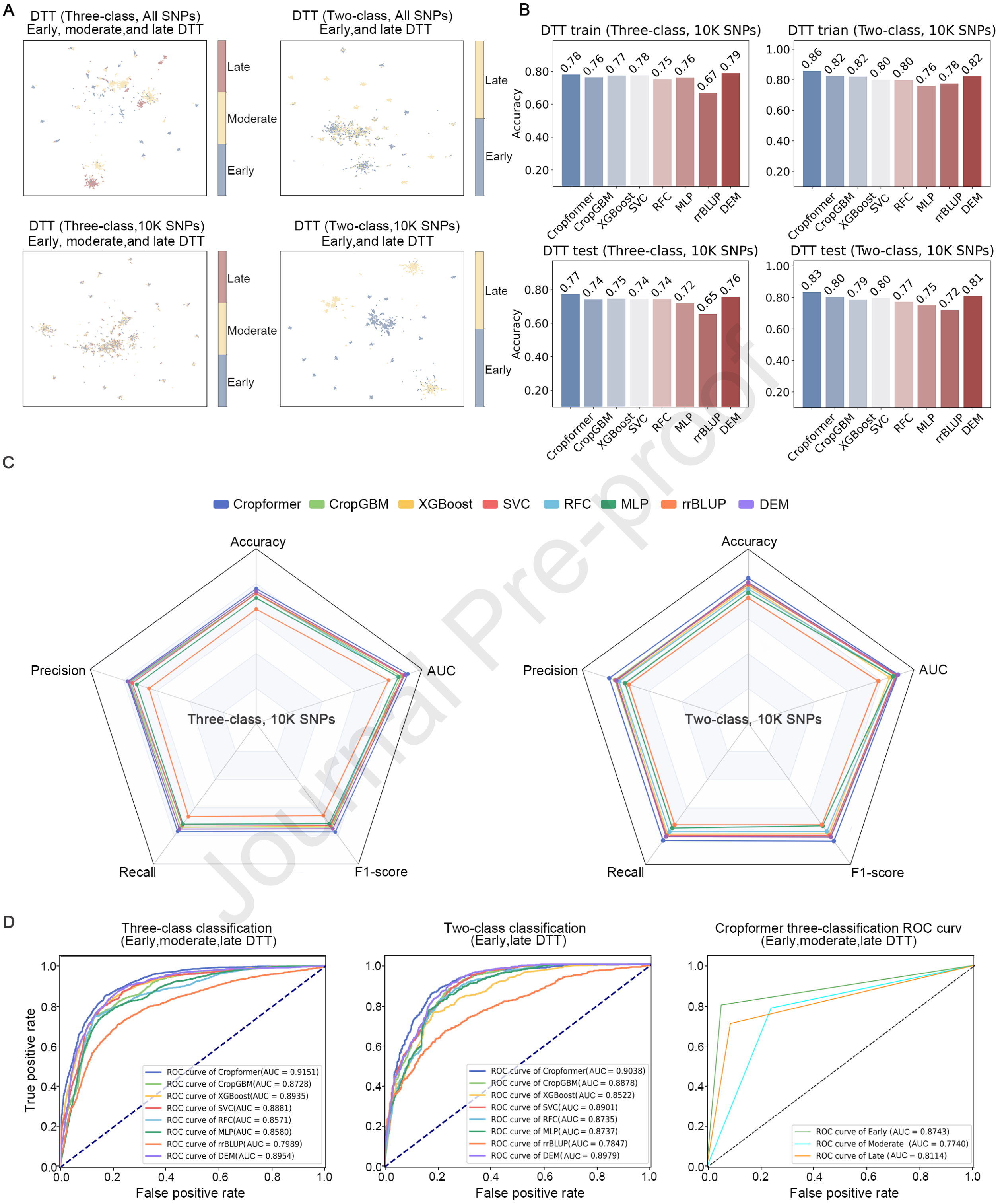
Figure 5. Cropformer can infer the contribution of SNPs to GS (Regression task). **A** Mapping of attentional weights to SNPs for maize DTT traits (Regression). The x-axis represents the SNP index position; the y-axis represents attentional weights (Only SNPs with attention weights greater than 1 are shown). **B** Comparison of traits among haplotypes. DTT comparisons among accessions harbouring different haplotypes of Zm00001d008941 and Zm00001d011956. **C** Haplotype network of Zm00001d008941. Circles represent haplotypes, and haplotypes are linked to their most similar relatives. Short lines indicate the diversity between linked haplotypes. **D** Gene structure and haplotypes of Zm00001d008941 in maize. The consensus genotype of each haplotype is marked in grey, light blue, and dark blue for the reference genotype, heterozygous mutation, and homozygous mutation, respectively. The purple bar graph represents the feature importance analysis based on XGBoost (Regression). **E** Haplotype network of Zm00001d011956. **F** Gene structure and haplotypes of Zm00001d011956 in maize. The purple bar graph represents the feature importance analysis based on XGBoost (Regression

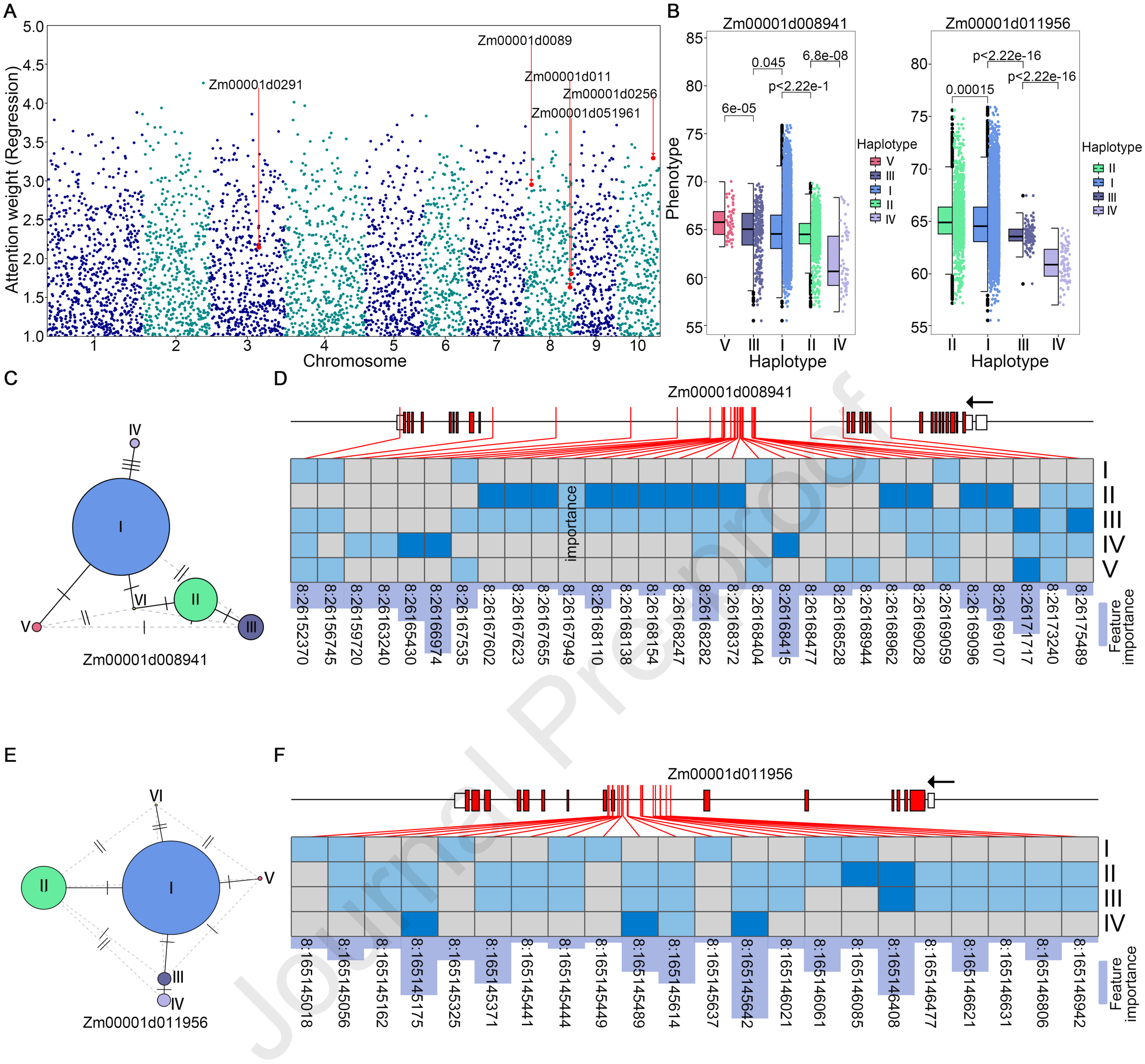
Figure 6. Cropformer web server.

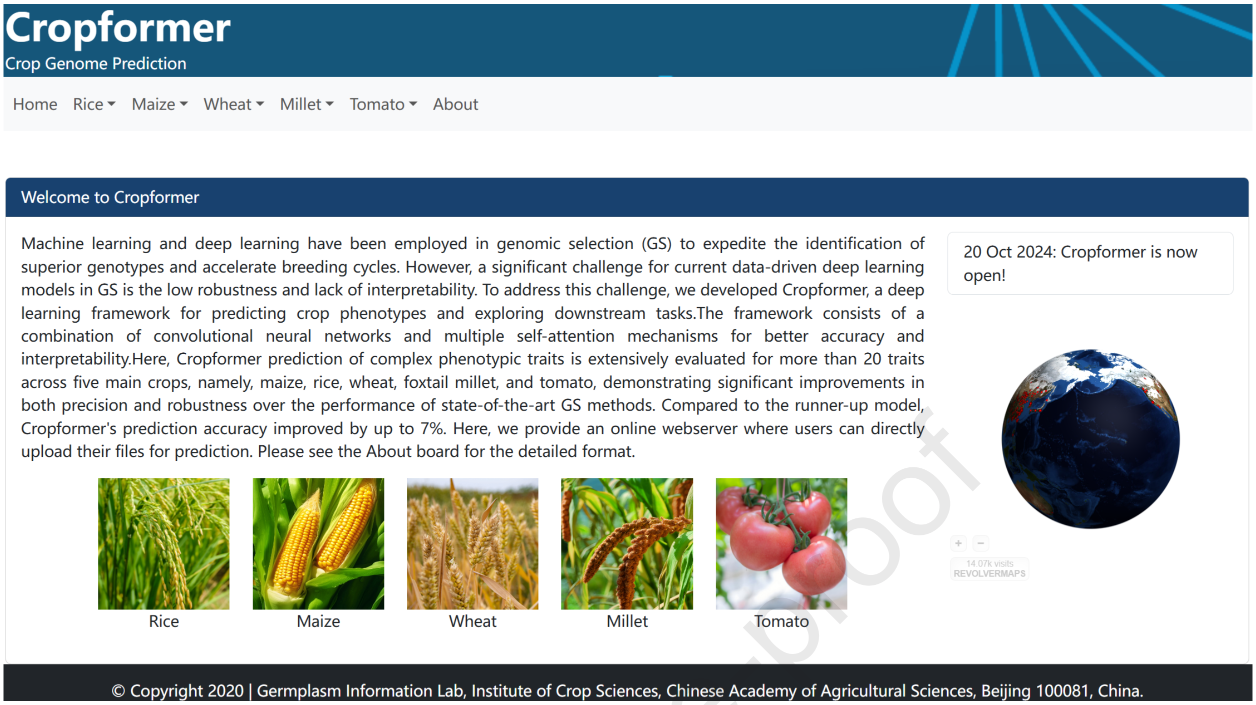

A
Data preprocessing
Cropformer model
Prediction
Genotyping
Sample 1 A...G...C...T...C
Sample 2 A...C...T...G...A
Sample 3 G...C...T...T... A
Sample n G...G...T...G...C
Data encoding
AA (0) AT (1) ··· GG (9)
Samples
Variations
Discrete traits
Quantitative traits
0.20
0.15
8
10
12
B
Multi-head self-attention mechanism
Convolution 3x3
Input sequence
Query
Key
Value
$\sqrt{d_k}$
Attention maps
C
Model Interpretation
Haplotype
Phenotype
Haplotype
II
I
III
IV
Attention weight
Chromosome
1 2 3 4 5 6 7 8 9 10
SHAP
SNP 1 +0.25
SNP 2 +0.22
SNP 3 +0.22
SNP 4 +0.20
SNP 5 +0.18
SNP 6 +0.15
SNP 7 +0.12
SNP n +0.10
mean(|SHAP|)
0.0 0.1 0.2 0.3
Clustering
Cluster 1
Cluster 2
Cluster 3
Three-Class Clustering


A
EW_train
EW_test
PH_train
PH_test
DTT_train
DTT_test
EW(g/ear)
PH(cm)
DTT(days)
6x10¹
10²
2x10²
3x10³
4x10²
B
Pearson correlation coefficient
DTT Train
0.94 0.88 0.85 0.86 0.86 0.85 0.85 0.92
PH Train
0.93 0.89 0.85 0.86 0.89 0.87 0.85 0.93
EW Train
0.77 0.71 0.66 0.65 0.65 0.67 0.66 0.69
DTT Test
0.92 0.87 0.84 0.84 0.85 0.83 0.82 0.90
PH Test
0.92 0.89 0.83 0.85 0.86 0.84 0.81 0.88
EW Test
0.76 0.71 0.63 0.63 0.64 0.64 0.63 0.69
0.20 0.40 0.60 0.80 1.00
Cropformer
CropGBM
DNNGP
XGBoost
SVR
MLP
rrBLUP
DEM


A
B
C
D
Pearson correlation coefficient
TKW Test
0.80 0.60 0.40 0.20
0.63 0.62 0.52 0.60 0.59 0.54 0.58 0.62
GW Test
0.80 0.60 0.40 0.20
0.69 0.68 0.58 0.64 0.46 0.57 0.64 0.68
GH Test
0.80 0.60 0.40 0.20
0.67 0.62 0.59 0.66 0.54 0.61 0.58 0.65
GP Test
0.60 0.50 0.40 0.30 0.20 0.10
0.49 0.48 0.44 0.42 0.48 0.42 0.45 0.49
GL Test
0.80 0.60 0.40 0.20
0.72 0.71 0.62 0.70 0.70 0.59 0.66 0.71
Anyang Test
1.00 0.80 0.60 0.40 0.20
0.84 0.76 0.71 0.71 0.68 0.75 0.72 0.75
Bejing Test
1.00 0.80 0.60 0.40 0.20
0.84 0.77 0.74 0.71 0.61 0.76 0.75 0.76
Changzhi Test
1.00 0.80 0.60 0.40 0.20
0.86 0.82 0.80 0.81 0.77 0.75 0.78 0.83
Dingxi Test
1.00 0.80 0.60 0.40 0.20
0.81 0.66 0.65 0.70 0.73 0.67 0.70 0.76
Urumqi Test
1.00 0.80 0.60 0.40 0.20
0.85 0.78 0.74 0.79 0.76 0.77 0.78 0.77
Culm_length Test
0.80 0.60 0.40 0.20
0.72 0.71 0.71 0.70 0.71 0.69 0.70 0.72
Days_to_heading_2018HN Test
0.80 0.60 0.40 0.20
0.69 0.67 0.67 0.65 0.66 0.65 0.62 0.67
Grain_length_width_ratio Test
0.80 0.60 0.40 0.20
0.65 0.62 0.59 0.61 0.62 0.63 0.60 0.64
Plant_height_2018HN Test
0.80 0.60 0.40 0.20
0.73 0.71 0.68 0.70 0.66 0.64 0.68 0.71
Thousand_grain_weight Test
0.80 0.60 0.40 0.20
0.63 0.62 0.60 0.62 0.53 0.58 0.57 0.59
SNP Test
0.60 0.40 0.20
0.59 0.57 0.56 0.59 0.52 0.58 0.55 0.56
InDel Test
0.80 0.60 0.40 0.20
0.65 0.58 0.57 0.56 0.47 0.58 0.55 0.58
GE Test
0.60 0.40 0.20
0.55 0.57 0.50 0.53 0.53 0.50 0.53 0.53
SV Test
0.60 0.40 0.20
0.52 0.48 0.45 0.50 0.42 0.45 0.46 0.54
Fusion test
0.80 0.60 0.40 0.20
0.72 0.70 0.68 0.70 0.65 0.66 0.70 0.71
Cropformer
CropGBM
DNNGP
XGBoost
SVR
MLP
rrBLUP
DEM
Wheat
Foxtail millet
Rice
Tomato


A
DTT (Three-class, All SNPs)
Early, moderate,and late DTT
Late
Moderate
Early
DTT (Two-class, All SNPs)
Early,and late DTT
Late
Early
DTT (Three-class,10K SNPs)
Early, moderate,and late DTT
Late
Moderate
Early
DTT (Two-class,10K SNPs)
Early,and late DTT
Late
Early
B
DTT train (Three-class, 10K SNPs)
Accuracy
0.78 0.76 0.77 0.78 0.75 0.76 0.67 0.79
DTT trian (Two-class, 10K SNPs)
0.86 0.82 0.82 0.80 0.80 0.76 0.78 0.82
DTT test (Three-class, 10K SNPs)
Accuracy
0.77 0.74 0.75 0.74 0.74 0.72 0.65 0.76
DTT test (Two-class, 10K SNPs)
0.83 0.80 0.79 0.80 0.77 0.75 0.72 0.81
1.00 0.80 0.60 0.40 0.20
Cropformer CropGBM XGBoost SVC RFC MLP rrBLUP DEM
C
Cropformer CropGBM XGBoost SVC RFC MLP rrBLUP DEM
Accuracy
AUC
F1-score
Recall
Precision
Three-class, 10K SNPs
Two-class, 10K SNPs
D
Three-class classification
(Early,moderate,late DTT)
ROC curve of Cropformer(AUC = 0.9151)
ROC curve of CropGBM(AUC = 0.8728)
ROC curve of XGBoost(AUC = 0.8935)
ROC curve of SVC(AUC = 0.8881)
ROC curve of RFC(AUC = 0.8571)
ROC curve of MLP(AUC = 0.8580)
ROC curve of rrBLUP(AUC = 0.7989)
ROC curve of DEM(AUC = 0.8954)
Two-class classification
(Early,late DTT)
ROC curve of Cropformer(AUC = 0.9038)
ROC curve of CropGBM(AUC = 0.8878)
ROC curve of XGBoost(AUC = 0.8522)
ROC curve of SVC(AUC = 0.8901)
ROC curve of RFC(AUC = 0.8735)
ROC curve of MLP(AUC = 0.8737)
ROC curve of rrBLUP(AUC = 0.7847)
ROC curve of DEM(AUC = 0.8979)
Cropformer three-classification ROC curv
(Early,moderate,late DTT)
ROC curve of Early (AUC = 0.8743)
ROC curve of Moderate (AUC = 0.7740)
ROC curve of Late (AUC = 0.8114)
True positive rate
False positive rate
0.0 0.2 0.4 0.6 0.8 1.0


A
Attention weight (Regression)
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
Zm00001d0291
Zm00001d0089
Zm00001d011
Zm00001d0256
Zm00001d051961
1
2
3
4
5
6
7
8
9
10
Chromosome
B
Zm00001d008941
Phenotype
85
80
75
70
65
60
55
6e-05
0.045
6.8e-08
p<2.22e-1
V
III
I
II
IV
Haplotype
Haplotype
V
III
I
II
IV
Zm00001d011956
80
75
70
65
60
55
0.00015
p<2.22e-16
p<2.22e-16
II
I
III
IV
Haplotype
Haplotype
II
I
III
IV
C
IV
I
II
III
V
VI
Zm00001d008941
D
Zm00001d008941
I
II
III
IV
V
importance
8:26152370
8:26156745
8:26159720
8:26163240
8:26165430
8:26166974
8:26167535
8:26167602
8:26167623
8:26167655
8:26167949
8:26168110
8:26168138
8:26168154
8:26168247
8:26168282
8:26168372
8:26168404
8:26168415
8:26168477
8:26168528
8:26168944
8:26168962
8:26169028
8:26169059
8:26169096
8:26169107
8:26171717
8:26173240
8:26175489
Feature importance
E
VI
II
I
V
III
IV
Zm00001d011956
F
Zm00001d011956
I
II
III
IV
8:165145018
8:165145056
8:165145162
8:165145175
8:165145325
8:165145371
8:165145441
8:165145444
8:165145449
8:165145489
8:165145614
8:165145637
8:165145642
8:165146021
8:165146061
8:165146085
8:165146408
8:165146477
8:165146621
8:165146631
8:165146806
8:165146942
Feature importance



Welcome to Cropformer

Machine learning and deep learning have been employed in genomic selection (GS) to expedite the identification of superior genotypes and accelerate breeding cycles. However, a significant challenge for current data-driven deep learning models in GS is the low robustness and lack of interpretability. To address this challenge, we developed Cropformer, a deep learning framework for predicting crop phenotypes and exploring downstream tasks.The framework consists of a combination of convolutional neural networks and multiple self-attention mechanisms for better accuracy and interpretability.Here, Cropformer prediction of complex phenotypic traits is extensively evaluated for more than 20 traits across five main crops, namely, maize, rice, wheat, foxtail millet, and tomato, demonstrating significant improvements in both precision and robustness over the performance of state-of-the-art GS methods. Compared to the runner-up model, Cropformer's prediction accuracy improved by up to 7%. Here, we provide an online webserver where users can directly upload their files for prediction. Please see the About board for the detailed format.

20 Oct 2024: Cropformer is now open!

© Copyright 2020 | Germplasm Information Lab, Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing 100081, China.