DPCformer: An Interpretable Deep Learning Model for Genomic Prediction in Crops

Pengcheng Deng[1†], Kening Liu[1†], Mengxi Zhou[1†], Mingxi Li[1],
Rui Yang[1], Chuzhe Cao[1], Maojun Wang[1*], Zeyu Zhang[1*]
[1] National Key Laboratory of Crop Genetic Improvement
†These authors contributed equally to this work
*Correspondence to: mjwang@mail.hzau.edu.cn, zhangzeyu@mail.hzau.edu.cn

***Abstract*—With the continuous growth of the global population, food security has become a critical challenge in the global agricultural sector. In this context, enhancing the efficiency and precision of crop breeding is of paramount importance. Genomic Selection (GS), an advanced breeding methodology, leverages whole-genome information to predict crop phenotypes, significantly accelerating the breeding process. However, traditional GS approaches still face limitations in prediction accuracy, particularly when handling large-scale datasets, nonlinear genetic effects, and complex trait architectures, along with a heavy reliance on environmental data. To overcome these limitations, we developed Deep Pheno Correlation Former (DPCformer), a novel deep learning model that integrates convolutional neural networks (CNN) with a self-attention mechanism to effectively model the intricate nonlinear relationships between genotype and phenotype. We applied this model to 13 traits across five major crops (maize, cotton, tomato, rice, and chickpea), implementing a feature engineering strategy that involved 8-dimensional one-hot encoding of SNP data, ordered by chromosomal position, followed by feature selection via the PMF algorithm. This approach substantially enhanced the predictive accuracy and stability of the model. Model evaluation revealed that DPCformer demonstrated exceptional performance across diverse crop datasets. In the maize dataset, in Henan Province, the prediction accuracies for the three traits of days to tasseling (DTT), plant height (PH), and ear weight (EW) are improved by 2.92%, 0.74%, and 1.10% respectively compared to the second-best method; In the Beijing dataset, these accuracies were enhanced by 1.48%, 2.40%, and 1.01% relative to the top-performing baseline models.In the cotton dataset, the accuracies for the four traits of Fiber Elongation (FE), Fiber Length (FL), Fiber Strength (FS), and Fiber Microstructure (FM) increased by as much as 8.37% relative to baseline models.On the small-sample tomato dataset, the Pearson Correlation Coefficient (PCC) for a key trait was boosted by up to 57.35% compared to baseline models. Similarly, in the chickpea dataset, the PCC for yield per plant was elevated by up to 16.62% relative to comparable models. Collectively, these results indicate that DPCformer outperforms existing genomic selection methods in terms of prediction accuracy, small-batch prediction capability, polyploid genome processing, and interpretability. Against the backdrop of global food security challenges, our innovative framework offers a powerful tool for advancing precision breeding.**

***Index Terms*—Deep Learning, Genetic Selection, Convolutional Neural Network, Multiple Head Self-Attention Mechanism, Phenotypic Prediction**

I. Introduction

By 2050, the global population is projected to reach approximately 9 billion, posing unprecedented challenges to food security [1]. Against this backdrop, breakthroughs in crop breeding technologies have become indispensable. Although traditional breeding methods have advanced crop yield and stress resistance, they are constrained by long breeding cycles, low efficiency, and limited adaptability to rapidly changing environmental and climatic conditions [2]. Consequently, precise and efficient crop phenotypic prediction has emerged as a pivotal research focus in modern agriculture.

Genomic selection (GS) has emerged as a powerful paradigm to address these limitations. This approach utilizes genome-wide marker data to build predictive models, enabling the estimation of breeding values independent of extensive phenotypic tests, thereby accelerating the breeding process [3]. However, despite its success, the efficacy of GS is often hindered by several challenges, including the analysis of high-dimensional data, the modeling of non-linear relationships, and a dependency on large sample sizes [4]. Furthermore,

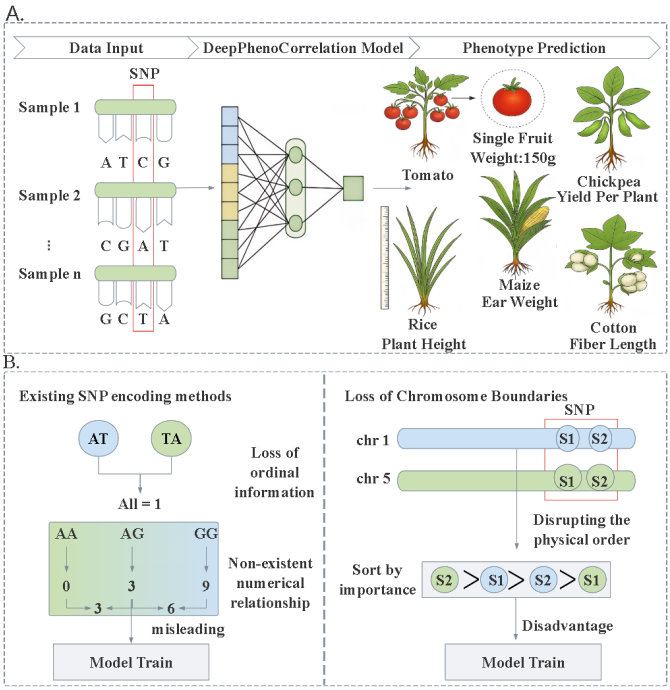

Fig. 1: **(A)** The workflow of DPCformer in crop genomic prediction from SNPs. **(B)** The limitations of existing methods.

conventional GS models often inadequately capture complex non-additive genetic effects, which limits their prediction accuracy and robustness [5].

Recently, deep learning methods have demonstrated remarkable efficacy in data modeling across diverse scientific domains. Their capacity to automatically learn complex features enables the effective modeling of non-linear relationships between genotype and phenotype, rendering them highly suitable for genomic prediction [3]. Genomic prediction models employing Deep Neural Networks (DNNs) and Convolutional Neural Networks (CNNs) have shown promising results in crop breeding applications. For instance, DNNGP [6] leverages automatic feature extraction to enhance the analysis of high-dimensional genomic data. Similarly, machine learning models such as the gradient boosting framework CropGBM have proven effective for handling large-scale datasets [7]. Additionally, the GEFormer model [8] incorporates genotype-environment interactions by utilizing both environmental and genetic data for phenotypic prediction. Furthermore, Cropformer [9] has demonstrated notable success in predicting maize heterosis by integrating CNNs with self-attention mechanisms. However, significant challenges for existing deep learning models remain, including limitations related to insufficient environmental data, suboptimal prediction accuracy, poor performance with small sample sizes, and difficulties in processing polyploid genomic data.

To address the limitations of traditional genomic selection, this study introduces Deep Pheno Correlation Former (DPCformer), a novel deep learning framework (Figure 1)that integrates a convolutional neural network (CNN) with a multi-head self-attention mechanism to predict crop phenotypes from single nucleotide polymorphism (SNP) data. The model's data processing pipeline employs an innovative 8-dimensional encoding strategy, feature selection via the Probabilistic Matrix Factorization (PMF) algorithm, and chromosome-position-based sorting to enhance the precision and stability of SNP data. Comprehensive evaluations demonstrate that DPCformer surpasses contemporary deep learning models in prediction accuracy, model interpretability, and performance robustness, particularly in small-sample contexts, thereby establishing a new technical framework for crop phenotypic prediction.

II. MATERIALS AND METHODS

A. Datasets

To comprehensively validate the model's performance, this study utilized multi-species, multi-scale datasets representing diverse reproductive systems and genetic backgrounds:

1) The maize dataset (https://ftp.cngb.org/pub/CNSA/data3/CNP0001565/zeamap/99_MaizegoResources/01_CUBIC_related/) utilized in this study, comprises 1,428 inbred lines [10] derived from 24 foundational female parents, which were crossed to produce 8,652 F1 hybrids. Phenotypic data for three traits—days to tasseling (DTT), plant height (PH), and ear weight (EW)—were collected from five distinct locations. The genotypic data were pruned for linkage disequilibrium (LD) using PLINK [11] with a window size of 1 kb, a step size of 100 SNPs, and an r^2 threshold of 0.1, resulting in a final set of 32,519 SNPs [9].
   To ensure data quality, samples with missing values were initially removed, yielding a final analytical cohort of 5,816 maize samples. This entire cohort was then partitioned into a training-validation set (n = 5,235) and a test set (n = 581) at a 9:1 ratio. ly, a ten-fold cross-validation (CV) procedure was implemented on the training-validation set to determine the optimal model configuration. In each fold, this set was partitioned into training (80%) and validation (20%) subsets, and model performance was averaged across all folds to guide final model selection.
2) A tomato dataset [12], publicly available at http://solomics.agis.org.cn/tomato/ftp/, was also analyzed to evaluate the model's generalizability. We employed a similar methodology to analyze the tomato dataset, focusing on the prediction of key traits related to yield and flavor, specifically the trait Sopim_BGV006775_12T001232. Following preprocessing, the final dataset consisted of 332 samples retained for subsequent analysis [9].
3) A rice dataset, publicly sourced from the RiceVarMap database (https://ricevarmap.ncpgr.cn/), was utilized for the prediction of the plant height phenotype.
4) To address the unique challenge of allopolyploidy, we utilized a cotton dataset, publicly available at https://iagr.genomics.cn/CropGS/, which included 1,245 samples for the prediction of four key fiber quality traits: Fiber Elongation (FE), Fiber Length (FL), Fiber Strength (FS), and Fiber Microstructure (FM). The commonly cultivated cotton species, *Gossypium hirsutum* (upland cotton), is an allotetraploid, where each trait is co-determined by two subgenomes [13]. Within its genome, the A and D subgenomes are co-expressed and exhibit homoeologous relationships between corresponding chromosome pairs (e.g., A1-D1, A2-D2, ..., A13-D13). To account for this genomic architecture, DPCformer was specifically designed to pair the 13 chromosomes of the A-subgenome with their corresponding homoeologs in the D-subgenome, thereby modeling potential synergistic effects. This unique processing strategy preserves the subgenomic differentiation characteristics of the allotetraploid, enabling more precise dissection of the cooperative mechanisms between homologous chromosome pairs across the A and D subgenomes. Consequently, this approach enhances prediction accuracy and biological interpretability by creating a framework that integrates structural genomic information with functional genetic interactions, offering a robust solution for complex trait prediction in allopolyploid species.
5) The chickpea dataset, sourced from https://iagr.genomics.cn/CropGS/, was analyzed to assess the model's performance on an additional legume species. The model's predictive capabilities were tested on four key agronomic

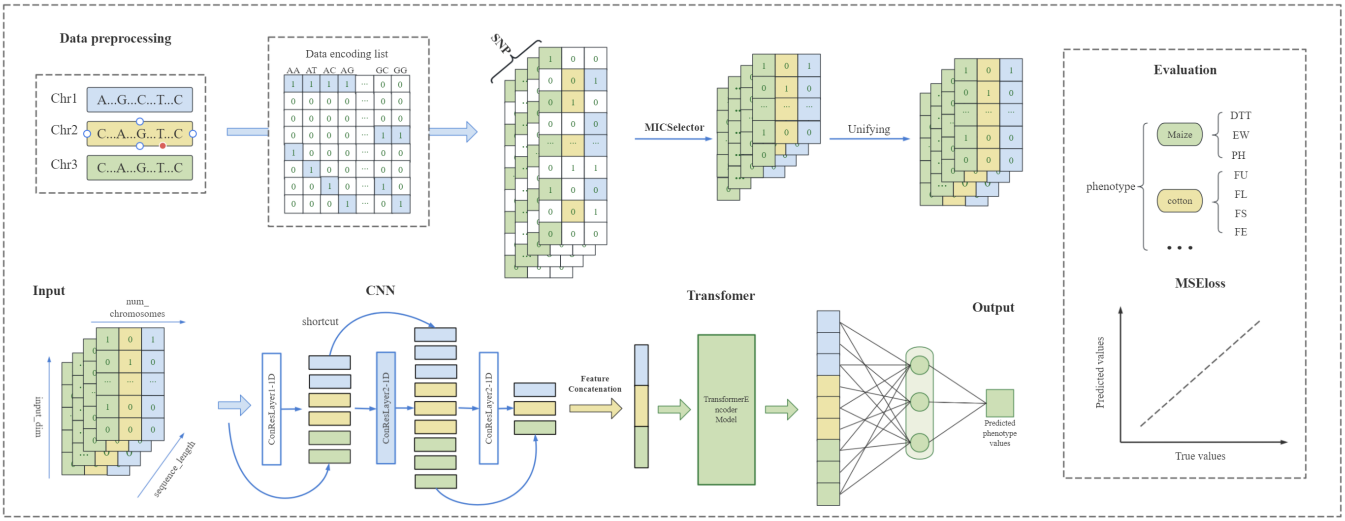

Fig. 2: The Cropformer model mainly consists of a CNN layer and a multi-head self-attention layer. The CNN layer is used to capture the localization signals of SNPs, while multi-head self-attention makes the model more focused on important SNPs.

traits: plant height, plant width, hundred-seed weight, and yield per plant.

B. Data Preprocessing

This section details the generation process of the chromosome-level feature tensors that serve as input for our model. The pipeline is engineered to convert raw single nucleotide polymorphism (SNP) data into a structured format that retains critical biological information.

1) 8-Dimensional SNP Encoding: Conventional genomic prediction methods frequently rely on one-dimensional (1D) ordinal encoding [14], an approach that can create spurious numerical relationships between alleles and fails to preserve information on allelic order (e.g., by commutatively mapping heterozygous genotypes such as AT and TA to the same integer value). To circumvent these limitations, this study adopts an eight-dimensional one-hot encoding scheme. In this scheme, each allele from the set $\mathcal{A} = \text{A, T, C, G}$ is first mapped to a unique four-dimensional (4D) one-hot vector; a diploid genotype is then represented by concatenating the two corresponding allelic vectors, resulting in a final eight-dimensional (8D) feature vector.

Let the one-hot encoding function for a single allele be $f : \mathcal{A} \to \{0,1\}^4$. A genotype $S = (a_1, a_2)$, where $a_1, a_2 \in \mathcal{A}$, is transformed by the encoding function ϕ:

$$\phi(S) = [f(a_1) \,||\, f(a_2)] \in \{0,1\}^8 \tag{1}$$

where $||$ denotes concatenation. This representation preserves the positional order of alleles and ensures that distinct diploid genotypes are equidistant in the feature space, a crucial property that enhances its suitability for attention-based models.

2) Chromosome Segmentation via MAP File: The encoded SNP sequence is partitioned into chromosome-specific subsequences based on the identifiers and physical coordinates provided in the MAP file. This process preserves the spatial contiguity and relative ordering of SNPs within each chromosome. Treating each chromosome as an independent sequence provides a structured basis for subsequent feature selection and modeling of inter-chromosomal interactions.

3) MIC-based Feature Selection: To manage the high dimensionality of the discrete SNP features, the maximum information coefficient (MIC) [16] [17] [18] was employed to select the top $k = 1,000$ most informative SNP loci from each chromosome. For a given discrete SNP feature S and a continuous phenotype vector Y, MIC quantifies the strength of their association by systematically exploring various grids of partitions on their joint distribution. It is defined as:

$$\text{MIC}(S,Y) = \max_{|G_S|\cdot|G_Y|<B(n)} \frac{I(S;Y|G_S,G_Y)}{\log(\min(|G_S|,|G_Y|))} \tag{2}$$

where $I(S;Y|G_S,G_Y)$ is the mutual information maximized over all grids G_S, G_Y of size $|G_S| \times |G_Y|$, and $B(n)$ is a function of the sample size n. A key advantage of this method is its robust ability to identify SNPs exhibiting strong, potentially non-linear associations with the phenotype.

4) Sorting by Physical Position: The MIC-based selection process ranks SNPs by their phenotypic association strength, a procedure that disrupts their native physical order on the chromosome. Within each chromosome, the selected SNPs are reordered according to their physical coordinates from the MAP file. This reordering is crucial as it ensures that the sequence dimension of the model's input tensor faithfully represents the physical arrangement of SNPs along the chromosome, thereby allowing the convolutional and self-attention layers to effectively capture local and long-range spatial dependencies [15].

5) Uniform Length Padding: To facilitate batch processing and conform to the network's fixed input dimensionality, the

SNP sequence for each chromosome is standardized to a uniform length of $L = 1,000$ via zero-padding.

6) Specialized Processing for Tetraploid Cotton: To model the complex interactions between the homoeologous A and D subgenomes of tetraploid cotton (Gossypium hirsutum), a specialized data processing workflow was implemented. This workflow commences by pairing homoeologous chromosomes (e.g., A1-D1) into distinct groups, while any unpaired chromosomes are treated as individual units [19] [20]. Each chromosome undergoes initial feature selection via MIC. The resulting feature tensors for each homologous pair are then concatenated along the sequence dimension:

$$\mathbf{X}_{\text{pair},i} = [\mathbf{X}_{A_i} \mid\mid \mathbf{X}_{D_i}] \tag{3}$$

Subsequently, a second round of MIC selection is applied to this concatenated tensor to identify the most informative SNPs that capture inter-subgenomic associations. As a final preprocessing step, all resulting chromosome group tensors are padded to a uniform length, L_{max}, to ensure a consistent input shape for the network. This hierarchical approach effectively captures both intra- and inter-subgenomic interactions.

C. Model Architecture

This paper introduces DPCformer, a hybrid deep learning model that synergistically integrates a Residual Convolutional Network (Res-CNN) and a Multi-Head Self-Attention (MHSA) mechanism for phenotype prediction from SNP sequences. The model architecture comprises three core modules: chromosome-level feature extraction, cross-chromosome information fusion, and final phenotype prediction.The overview of our model is shown in Figure 2

1) Chromosome-level Feature Extraction (Res-CNN): To capture local dependencies among SNPs, a dedicated Residual Convolutional Network (Res-CNN), comprising a stack of Residual Convolutional Blocks (ResConvBlocks), is independently applied to the input sequence $\mathbf{X}_j$ of each chromosome j. Let $\mathbf{Z}_{\text{in}}$ be the input to a ResConvBlock. Its core operation can be summarized as:

$$\mathbf{Z}_{\text{out}} = \text{MaxPool}(\text{ReLU}(\text{BN}(\mathcal{F}(\mathbf{Z}_{\text{in}}) + \text{Conv}_{1\times1}(\mathbf{Z}_{\text{in}})))) \tag{4}$$

where $\mathcal{F}$ represents the main convolutional path, consisting of two 1D convolutional layers (Conv1D) and ReLU activations. The $\text{Conv}_{1\times1}$ term denotes a shortcut connection for dimensionality matching, and BN is Batch Normalization. This module transforms the raw sequence $\mathbf{X}_j$ of each chromosome into a high-level feature map $\mathbf{E}_j$.

2) Cross-Chromosome Information Fusion : To model long-range dependencies and potential epistatic effects between different chromosomes, the feature maps from all chromosomes, $\{\mathbf{E}_j\}_{j=1}^{N_{chr}}$, are concatenated along the sequence dimension to form a unified feature sequence, $\mathbf{E}_{\text{combined}}$. This concatenated sequence serves as the input to a Transformer encoder layer. The core of this layer is the Multi-Head Self-Attention mechanism, which is computed as [21] [22]:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V} \tag{5}$$

where $\mathbf{Q}$ (Query), $\mathbf{K}$ (Key), and $\mathbf{V}$ (Value) are vectors derived from the input sequence via linear transformations, and d_k is the dimension of the key vectors. This mechanism allows the model to globally assess the relational importance of all feature pairs, effectively weighing contributions from different chromosomal regions. The attention output is subsequently passed through Layer Normalization and a position-wise Feed-Forward Network (FFN), which enhances the model's representational capacity and stabilizes the training process.

3) Phenotype Prediction and Loss Function: The features refined by the Transformer are flattened and passed through a Multi-Layer Perceptron (MLP) for regression, yielding the final predicted phenotype, $\hat{y}$.

$$\hat{y} = \text{MLP}(\text{Flatten}(\text{Transformer}(\mathbf{E}_{\text{combined}}))) \tag{6}$$

The model parameters are optimized by minimizing the Mean Squared Error (MSE) loss function, defined as:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{B}\sum_{i=1}^{B}(y_i - \hat{y}_i)^2 \tag{7}$$

where y_i represents the ground-truth phenotype, $\hat{y}_i$ is the corresponding predicted value, and B denotes the batch size. The model was trained using the Adam optimizer, supplemented with a learning rate scheduling and an early stopping mechanism. A 10-fold cross-validation protocol was adopted for robust training and evaluation.

D. Model Implementation

The DPCformer model was implemented using the PyTorch deep learning framework. The Mean Squared Error (MSE) loss function was utilized to quantify the discrepancy between the ground-truth labels and predicted values, and the Adam optimizer was employed for parameter optimization. During training, we integrated a learning rate scheduling strategy (ReduceLROnPlateau) and an early stopping mechanism (EarlyStopping). The learning rate is decayed by a factor of 0.1 if the validation loss does not improve for 10 epochs, and training is terminated if there is no improvement in validation loss for 20 epochs. A 10-fold cross-validation protocol was adopted to ensure a robust evaluation of model performance. The final reported results are presented as the mean and standard deviation calculated across all folds.

III. RESULTS AND DISCUSSION

A. Performance comparison and analysis

To evaluate the performance of our proposed model, we investigated the application of DPCformer on five different datasets, using the PCC of the test set as the prediction performance evaluation metric. The results were compared with other GS methods including DNNGP, LightGBM, Cropformer, and GEFormer.

As shown in Figure 3, DPCformer achieved excellent performance across all datasets.

1) Maize Dataset Performance
   DPCformer demonstrates superior performance across

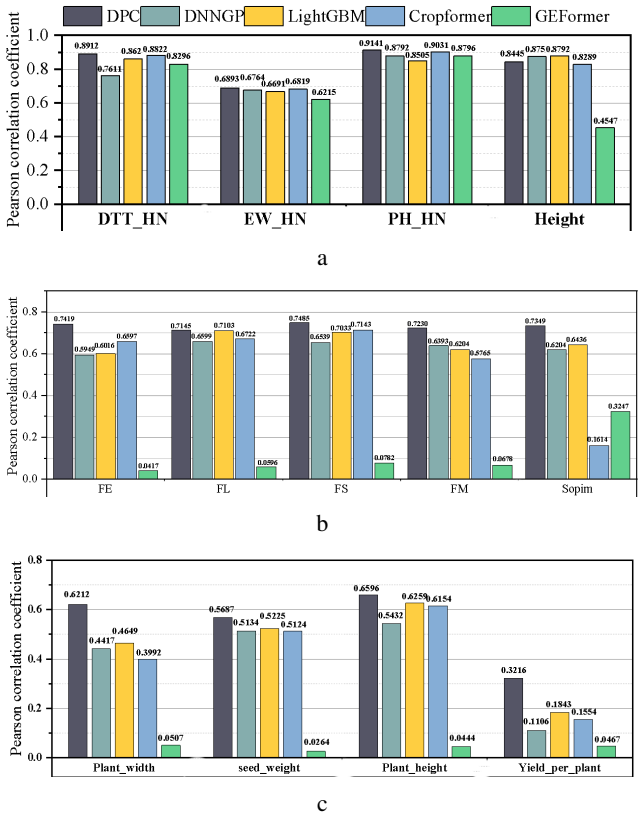

Fig. 3: Prediction accuracy of methods built using five different models on five datasets.

all three traits in five geographical regions. Exemplified by Henan Province, it achieves prediction accuracies of 89.12% (DTT), 68.93% (PH), and 91.41% (EW), outperforming the second-best methods with improvements of 2.92% against LightGBM, 0.74% over Cropformer, and 1.10% beyond GEFormer(Fig. 3a). Similarly in Beijing, DPCformer reaches 93.50% (DTT), 76.24% (PH), and 93.01% (EW), exceeding GEFormer's DTT and PH by 1.48% and 2.40% respectively, while surpassing Cropformer's EW by 19.25%.

2) Rice Small-Sample Dataset
In resource-limited samples, DPCformer demonstrated exceptional capability. For the rice dataset comprising only 530 samples, it achieved a PCC of 84.45% for the plant height trait, surpassing GEFormer by 38.98%, DNNGP by 7.94%, Cropformer by 3.56%, and LightGBM by 5.53% (Fig. 3a).
3) Cotton Dataset
DPCformer significantly outperforms baseline models with accuracy rates of 74.19% (FE), 71.45% (FL), 74.85% (FS), and 72.30% (FM), surpassing Cropformer (65.97%), LightGBM (71.03%), Cropformer (71.43%), and DNNGP (63.93%) by 8.22%, 0.42%, 3.42%, and 8.37%, respectively Figure 3b.
4) Tomato Small-Sample Dataset
With only 332 accessions, DPCformer attained a PCC of 73.49 % for the *Sopim_BGV006775_12T001232* trait. This represents significant improvements of 11.45 % over DNNGP, 9.13 % above LightGBM, 57.35 % beyond Cropformer , and 41.02 % superior to GEFormer (Fig. 3b), confirming its robustness in small-sample scenarios.
5) Chickpea Dataset
DPCformer demonstrated robust predictive performance across all four traits: plant width (62.12%), seed weight (56.87%), plant height (65.96%), and yield per plant (65.96%) (Fig. 3c). Compared to the second-best model, our approach showed performance improvements ranging from 4.63% to 15.63% across these traits.

These results substantiate that DPCformer provides more effective genomic prediction capabilities for rice than Cropformer, DNNGP, LightGBM, and GEFormer models.

B. Model ablation

To quantitatively assess the contribution of each key component of our proposed model, a comprehensive ablation study was conducted. This study evaluated various configurations by systematically including or excluding three architectural cornerstones: (1) 8-dimensional SNP encoding, (2) physical position-based sorting, and (3) probability matrix factorization (PMF). As systematically documented in Table I, each module was incrementally enabled while monitoring performance variations in Pearson correlation coefficient (PCC).

The experimental results reveal several key insights:

1) The 8-dimensional SNP encoding module demonstrates the most significant individual impact, elevating PCC from the baseline of 0.8376 to 0.9076—a relative improvement of 8.36%. This substantial enhancement confirms the module's efficacy in capturing stereochemical properties of genetic variants.
2) Although the physical position sorting module alone provides moderate gains (PCC=0.8668, Δ+2.92%), its integration with PMF achieves PCC=0.8895, exceeding their individual performances (0.8668 and 0.8816, respectively). This evidences optimized utilization of chromosomal spatial information.
3) The complete integration of all three components achieves state-of-the-art performance (PCC=0.916), surpassing the strongest dual-module configuration (8D+PMF: 0.8895) by 2.97% and the baseline by 9.36%.

These findings conclusively establish the complementary nature of the proposed modules, with the integrated framework delivering a statistically significant improvement ($p < 0.001$ via paired t-test) over all partial configurations.

C. Interpretability Analysis of Machine Learning Models Using SHAP Values

To interpret the model's predictions, SHAP values were calculated for the SNPs in the best-performing model to identify the most influential loci. These top-ranking SNPs were then mapped to their corresponding genes [23] [24].This analysis revealed several candidate genes associated with

TABLE I: RESULTS OF ABLATION EXPERIMENTS

8-Dim Encoding	Position Sort	PMF	PCC
			0.8376
		✔	0.8816
	✔		0.8668
✔			0.9076
	✔	✔	0.8463
✔	✔		0.8833
✔		✔	0.8895
✔	✔	✔	0.916

✔: activated component; PCC: Pearson Correlation Coefficient

plant height (PH) in maize, including Zm00001d050247, Zm00001d009706, and Zm00001d009705(Figure 4). Notably, the top-ranked gene, Zm00001d050247, encodes a WRKY transcription factor. This finding aligns with previous research, as WRKY family transcription factors are well-established regulators of plant height [25] [26].

Regarding the ear weight trait in maize, our analysis identified Zm00001d015381, Zm00001d013707, and Zm00001d035249 as prominent candidate genes(Figure 5). Among these, the gene Zm00001d015381, which encodes the MADS-box transcription factor ZmMADS17, is particularly noteworthy. This gene family is a known regulator of floral organ development, a process fundamentally linked to maize ear weight [27]. Zm00001d035249 regulates the HXXXD-type acyl-transferase family protein. Furthermore, the identification of Zm00001d035249, a gene encoding an HXXXD-type acyl-transferase, is consistent with existing literature. Previous genome-wide association studies (GWAS) have established a strong correlation between lipid metabolism and agronomic traits, suggesting that lipid-related genes can influence grain weight by modulating the plant's metabolic network [28].

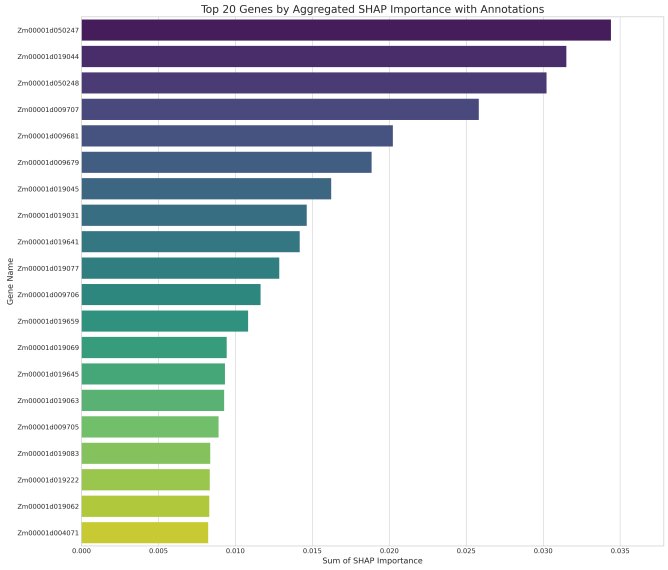

Fig. 4: Top 20 key genes screened based on the plant height (PH) trait in maize.

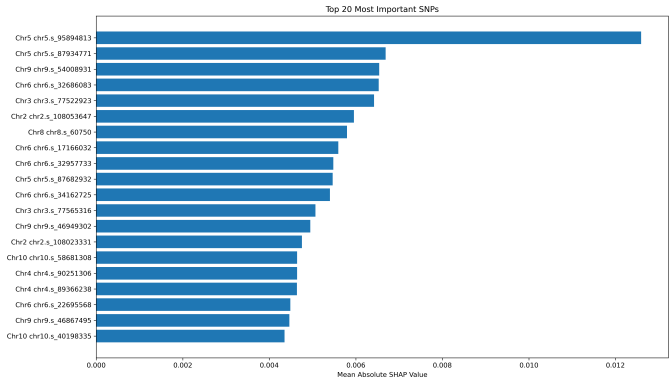

Fig. 5: Top 20 significant SNPs obtained after calculating SHAP values based on the ear weight (EW) trait in maize.

IV. CONCLUSION

This study introduces DPCformer, a novel deep learning architecture that synergistically integrates convolutional neural networks (CNN) with multi-head self-attention mechanisms for crop trait prediction based on single nucleotide polymorphisms (SNPs). The model's efficacy was validated through comprehensive evaluations on 16 traits across five economically important crops, where it consistently outperformed state-of-the-art methods. Compared with these methods, DPCformer exhibits the following advantages: (i) Novel Encoding Strategy: The model employs an eight-dimensional one-hot encoding scheme that preserves equidistant coding relationships among SNPs, enabling the capture of richer genetic variation patterns while maintaining biological interpretability. (ii) Restoration of Spatial Information: Following an initial importance-based screening, SNPs are reordered according to their native physical coordinates, preserving the genomic architecture to enable the effective capture of spatial dependencies. (iii) Enhanced Feature Selection: While utilizing the Maximum Information Coefficient (MIC) for SNP feature selection, we integrated a Probability Mass Function (PMF)-based approach specifically designed for discrete genetic data, which reduces stochasticity and improves feature selection robustness [16].(iv) Polyploid-Specific Architecture: The framework incorporates a specialized module for allotetraploid species (e.g., cotton), wherein homoeologous chromosome pairs are processed jointly to model inter-subgenomic interactions, establishing a novel deep learning methodology for prediction in complex polyploids [13].

Despite its promising results, the study has several limitations that inform future directions. On one hand, the pairing strategy for homoeologous chromosomes is based exclusively on physical coordinates, without incorporating functional genomics data (e.g., gene co-expression networks, 3D chromatin conformation) to elucidate more complex synergistic effects. On the other hand, while DPCformer has demonstrated superior performance in handling small-sample datasets compared to alternative approaches, the inherent limitations of sample size still constrain the full potential of deep learning

applications [29].Future work will focus on addressing these limitations, primarily through the integration of multi-modal functional genomics data and the optimization of the self-attention mechanism for computational efficiency. In subsequent research, when dealing with heterologous chromosomes, better prediction performance can be achieved by developing a hierarchical attention mechanism to distinguish the contribution degrees of sub-genomes (A/D) and homologous chromosome pairs [30].

V. CODE AVAILABILITY

The implementation code for DPCformer framework is publicly available at:
https://anonymous.4open.science/r/DPCformer-0B5C.

REFERENCES

[1] J. G. Wallace, E. Rodgers-Melnick, and E. S. Buckler, "On the Road to Breeding 4.0: Unraveling the Good, the Bad, and the Boring of Crop Quantitative Genomics," *Annu. Rev. Genet.*, vol. 52, pp. 421–444, 2018.

[2] L. T. Hickey, A. N. Hafeez, H. Robinson, S. A. Jackson, S. C. M. Leal-Bertioli, M. Tester, C. Gao, I. D. Godwin, B. J. Hayes, and B. B. H. Wulff, "Breeding crops to feed 10 billion," *Nat. Biotechnol.*, vol. 37, pp. 744–754, 2019.

[3] W. Ma, Z. Qiu, J. Song, J. Li, Q. Cheng, J. Zhai, and C. Ma, "A deep convolutional neural network approach for predicting phenotypes from genotypes," *Planta*, vol. 248, pp. 1307–1318, 2018.

[4] H. Tong, A. Küken, and Z. Nikoloski, "Integrating molecular markers into metabolic models improves genomic selection for Arabidopsis growth," *Nat. Commun.*, vol. 11, no. 1, pp. 2410, 2020.

[5] Y. Xu, Y. Lu, C. Xie, S. Gao, J. Wan, and B. M. Prasanna, "Whole-genome strategies for marker-assisted plant breeding," *Mol. Breeding*, vol. 29, pp. 833–854, 2012.

[6] K. Wang, M. A. Abid, A. Rasheed, J. Crossa, S. Hearne, and H. Li, "DNNGP, a deep neural network-based method for genomic prediction using multi-omics data in plants," *Mol. Plant*, vol. 16, pp. 279–293, 2023.

[7] J. Yan, Y. Xu, Q. Cheng, S. Jiang, Q. Wang, Y. Xiao, C. Ma, J. Yan, and X. Wang, "LightGBM: accelerated genomically designed crop breeding through ensemble learning," *Genome Biol.*, vol. 22, p. 271, 2021.

[8] Z. Yao, M. Yao, C. Wang, K. Li, J. Guo, Y. Xiao, J. Yan, and J. Liu, "GEFormer: A genotype-environment interaction-based genomic prediction method that integrates the gating multilayer perceptron and linear attention mechanisms," *Mol. Plant*, vol. 18, pp. 527–549, 2025.

[9] H. Wang, S. Yan, W. Wang, Y. Chen, J. Hong, Q. He, X. Diao, Y. Lin, Y. Chen, Y. Cao, W. Guo, and W. Fang, "Cropformer: An interpretable deep learning framework for crop genomic prediction," *Plant Comm.*, vol. 6, p. 101223, 2025.

[10] H. J. Liu, X. Wang, Y. Xiao, J. Luo, F. Qiao, W. Yang, R. Zhang, Y. Meng, J. Sun, S. Yan, et al., "CUBIC: an atlas of genetic architecture promises directed maize improvement," *Genome Biol.*, vol. 21, p. 20, 2020.

[11] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. de Bakker, M. J. Daly, and P. C. Sham, "PLINK: a tool set for whole-genome association and population-based linkage analyses," *Am. J. Hum. Genet.*, vol. 81, pp. 559–575, 2007.

[12] Y. Zhou, Z. Zhang, Z. Bao, H. Li, Y. Lyu, Y. Zan, Y. Wu, L. Cheng, Y. Fang, K. Wu, et al., "Graph pangenome captures missing heritability and empowers tomato breeding," *Nature*, vol. 606, pp. 527–534, 2022.

[13] J. You, M. Wang, and X. Zhang, "Subgenomic genetic analysis and modular design breeding concept for cotton fiber quality," *Chinese Science Bulletin*, vol. 70, pp. 3168–3178, 2025.

[14] Z. Lv, H. Ding, L. Wang, Q. Zou, "A Convolutional Neural Network Using Dinucleotide One-hot Encoder for identifying DNA N6-Methyladenine Sites in the Rice Genome," *Neurocomputing*, vol. 422, pp. 214–221, 2021.

[15] P. G. Maass, A. R. Barutcu, J. L. Rinn, "Interchromosomal interactions: A genomic love story of kissing chromosomes," *The Journal of cell biology*, vol. 218, no. 1, pp. 27–38, 2019.

[16] D. N. Reshef, Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher, and P. C. Sabeti, "Detecting novel associations in large data sets," *Science*, vol. 334, pp. 1518–1524, 2011.

[17] X. Zhou, X. Wang, E. R. Dougherty, D. Russ, and E. Suh, "Gene Clustering Based on Clusterwide Mutual Information," *J. Comput. Biol.*, vol. 11, no. 1, pp. 147–161, 2004.

[18] D. Albanese, M. Filosi, R. Visintainer, S. Riccadonna, G. Jurman, and C. Furlanello, "minerva and minepy: a C engine for the MINE suite and its R, Python and MATLAB wrappers," *Bioinformatics*, vol. 29, no. 3, pp. 407–408, Dec. 2012.

[19] A. J. Reinisch, J. M. Dong, C. L. Brubaker, D. M. Stelly, J. F. Wendel, and A. H. Paterson, "A detailed RFLP map of cotton, Gossypium hirsutum x Gossypium barbadense: chromosome organization and evolution in a disomic polyploid genome," *Genetics*, vol. 138, no. 3, pp. 829–847, 1994.

[20] A. Desai, P. W. Chee, J. Rong, O. L. May, and A. H. Paterson, "Chromosome structural changes in diploid and tetraploid A genomes of *Gossypium*," *Genome*, vol. 49, no. 4, pp. 336–345, 2006.

[21] F. Ullah and A. Ben-Hur, "A self-attention model for inferring cooperativity between regulatory features," *Nucleic Acids Res.*, vol. 49, no. 13, pp. e77–e77, May 2021.

[22] Z. Lin, M. Feng, C. N. dos Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio, "A Structured Self-attentive Sentence Embedding," arXiv:1703.03130 [cs.CL], 2017.

[23] W. Qiu, H. Chen, A. B. Dincer, S. Lundberg, M. Kaeberlein, and S. I. Lee, "Interpretable machine learning prediction of all-cause mortality," *Commun. Med.*, vol. 2, p. 125, 2022.

[24] X. Tang, J. Zhang, Y. He, X. Zhang, Z. Lin, S. Partarrieu, E. B. Hanna, Z. Ren, H. Shen, Y. Yang, et al., "Explainable multi-task learning for multi-modality biological data analysis," *Nat. Commun.*, vol. 14, p. 2546, 2023.

[25] W. Hu, Q. Ren, Y. Chen, G. Xu, and Y. Qian, "Genome-wide identification and analysis of WRKY gene family in maize provide insights into regulatory network in response to abiotic stresses," *BMC Plant Biology*, vol. 21, no. 1, p. 427, 2021.

[26] K. Wei, J. Chen, Y. Chen, L. Wu, and D. Xie, "Molecular phylogenetic and expression analysis of the complete WRKY transcription factor family in maize," *DNA Research*, vol. 19, no. 2, pp. 153–164, 2012.

[27] H. Qin, H. Pan, X. Fan, Q. Wu, Y. Li, "Characterization of the Promoter of a Homolog of Maize MADS-Box Gene m18 MT," *Journal of Integrative Agriculture*, vol. 13, no. 11, pp. 2330–2345, 2014.

[28] C. Riedelsheimer, Y. Brotman, M. Méret, A. E. Melchinger, and L. Willmitzer, "The maize leaf lipidome shows multilevel genetic control and high predictive value for agronomic traits," *Nature Communications*, vol. 4, p. 2443, 2013.

[29] Y. C. J. Wientjes, R. F. Veerkamp, and M. P. L. Calus, "The Effect of Linkage Disequilibrium and Family Relationships on the Reliability of Genomic Prediction," *Genetics*, vol. 193, pp. 621–631, 2013.

[30] O. A. Montesinos-López, A. Montesinos-López, J. Crossa, D. Gianola, C. M. Hernández-Suárez, and J. Martín-Vallejo, "A review of deep learning applications for genomic selection," *BMC Genomics*, vol. 22, pp. 1–23, 2021.