Research on Crop Phenotype Prediction Methods Based on SNP-context and Whole-genome Features Embedding

Huan Li[1], Yunpeng Cui[1*], Zhen Chen[2], Ting Wang[1], Chao Wang[3], Wenbo Bian[3],Juan Liu[1], Mo Wang[1], Li Chen[1], Jinming Wu[1], Jie Huang[1]

[1]Institute of Agricultural Information, Chinese Academy of Agricultural Sciences(CAAS), Beijing 100081, China

[2]Collaborative Innovation Center of Henan Grain Crops, Henan Agricultural University, Zhengzhou 450046, China

[3]Zibo Digital Agriculture and Rural Research Institute, Zibo 255000, China

***** Corresponding author: Institute of Agricultural Information, Chinese Academy of Agricultural Sciences(CAAS), Beijing 100081, China. E-mail: cuiyunpeng@caas.cn

Abstract

Modern agriculture demands precise genomic prediction to accelerate elite crop breeding, yet current methods predominantly analyze single nucleotide polymorphisms (SNPs) while neglecting the regulatory context of surrounding sequences. To overcome these limitations, two novel feature embedding modes(SNP-context and whole-genome) to capture both local and global genetic information. These were implemented through a unified computational framework combining the DNABERT-2 cross-species genomic foundation model with optimized dimensionality reduction techniques, including Partial Least Squares and Principal Component Analysis, to mitigate noise in high-dimensional feature spaces. The performance of classical machine learning algorithms, such as Support Vector Regression, ridge regression Best Linear Unbiased Prediction, Random Forest, and Gradient Boosting Regression, was evaluated across different dimensions(4-768), dimensionality reduction methods and SNP-context lengths(500-3000bp) in three crop datasets—rice413, rice395, and maize301. Results demonstrate that while both embedding modes have limitations in predicting traits with ultra-low heritability in rice395, the SNP-context feature embedding significantly outperforms SNPs marker-based predictions for traits with low-to-medium heritability in rice413 and maize301, achieving higher accuracy and lower mean absolute error (MAE). For some traits, the accuracy of SNP-context embedding mode even surpasses that of whole-genome feature embedding. When predicting traits with high heritability, the whole-genome feature embedding mode, combined with conventional machine learning algorithms, significantly improves prediction accuracy and reduces MAE compared to state-of-the-art deep learning models like DNNGP and ResGS. These findings show the capacity to capture non-additive genetic effects often neglected by conventional SNP-based models and provides new insights into the genetic architecture of complex traits.

Keywords: genomic prediction, DNABert-2, Partial Least Squares, Principal Component Analysis, SNP-context feature embedding, whole-genome feature embedding.

Introduction

In recent years, the exponential growth of the global population and significant climate changes have imposed substantial pressure on food production systems worldwide. On one hand, the increasing population demands greater food supplies to combat persistent hunger; on the other hand, challenges such as global warming, water and soil resource scarcity, and frequent pest and disease outbreaks have severely disrupted crop growth and metabolism, ultimately reducing yield and quality (Saraiva et al. 2023; Yuan et al. 2024; Bita and Gerats 2013). According to the Food and Agriculture Organization of the United Nations (FAO), global food production must increase by over 70% by 2050 to meet the needs of a population exceeding 9 billion(He and Li 2020; FAO 2009). Accelerating the selection of crop varieties with high-yield potential is pivotal to addressing this challenge. Traditional breeding methods, however, are constrained by cumbersome experimental cycles, and it often requires several years or even decades to develop crop varieties with desired agronomic traits(Bonea 2022; Martínez-Gómez 2019). The advent of high-throughput sequencing technology has significantly reduced genotyping costs and generated extensive high-density genetic markers, thereby accelerating the shift in crop breeding methodologies from empirical tradition to data-driven paradigms(Singh et al. 2023; Pujolar et al. 2013; Sun et al. 2013). Genomic selection (GS), a method that employs whole-genome markers to predict genomic estimated breeding values (GEBVs), has revolutionized animal and crop breeding by enabling early-stage selection (Meuwissen et al. 2001; Li et al. 2023; Ibtisham et al. 2017; Grenier et al. 2015; Budhlakoti et al. 2022; Sandhu et al. 2021). Nevertheless, the genetic gain achievable through genomic selection (GS) is intrinsically limited by the accuracy of genomic prediction (GP) models. Improving GP accuracy can effectively shorten breeding cycles and reduce associated cost(Crossa et al. 2017; Zhao et al. 2023; Bohra et al. 2020). A fundamental challenge in GP arises from the high-dimensionality of genomic data, where the number of single nucleotide polymorphism (SNP) markers (P) vastly exceeds the available sample size (N) - the so-called 'big P, small N' problem. The dimensionality disparity (P >> N) elevates overfitting risks and complicates predictive model development (Crossa et al. 2017; Schmidt et al. 2016). Furthermore, linkage disequilibrium (LD) among SNPs introduces multicollinearity, amplifying model complexity(Johnston and Carvalho 2014). To mitigate these issues, a subset of significant SNP markers was proposed to be used instead of the entire genome (Jeong et al. 2020). Although whole-genome association studies (GWAS) effectively identify genetic loci associated with complex traits, pinpointing the most predictive markers for specific phenotypes remains challenging due to the polygenic architecture of agronomic traits, confounding factors such as population stratification, and limited generalizability of marker-trait correlations across genetically distinct cohorts.

Current genomic prediction methodologies in plant sciences predominantly focus on statistical modeling and machine learning approaches. Statistical models include the classical Best Linear Unbiased Prediction (BLUP) and its variants, such as Genomic BLUP (GBLUP)(VanRaden 2008), Ridge Regression BLUP (RR-BLUP)(Resende et al. 2012; Endelman 2011), and Single-Step GBLUP(Sánchez-Mayor et al. 2022). These methods differ in their underlying assumptions; for instance, GBLUP assumes that all SNPs have equal effect variances, while RR-BLUP assumes that marker effects follow a normal distribution with equal magnitudes. Bayesian methods, such as BayesA, BayesB, BayesCπ, and Bayes Lasso (Habier et al. 2011; Zhu et al. 2016; Cleveland et al. 2010), also rely on distinct prior assumptions. Machine learning methods, including Support Vector Regression (SVR), Random Forest (RF), and Gradient Boosting Regression (GBR)(Grinberg et al. 2020; Cosgun et al. 2011; Xie et al. 2024), offer flexibility by avoiding strict assumptions about data distribution, making them suitable for small-scale genomic datasets. However, they often struggle to capture the complex nonlinear relationships between genomes and phenotypes. In contrast, deep learning (DL), a machine learning paradigm characterized by multi-layered neural networks, autonomously derives high-order genomic features, overcoming limitations of manual feature selection(Shen et al. 2022). The DeepGS model, developed using a deep convolutional neural network (CNN) with an 8-31-1 layer architecture, demonstrated promising results(Ma et al. 2018). Subsequent work introduced DNNGP, a

three-CNN-layer model that outperformed both traditional machine learning methods and DeepGS across four crop species(Wang et al. 2022). Further progress was achieved with SoyDNGP, a CNN-based architecture validated across multiple crops including cotton, corn, rice, and tomato(Gao et al. 2023). Most recently, the incorporation of residual structures and hierarchical convolutions led to the development of the 35-layer ResGS model, which achieved reliable predictions on four public datasets(Xie et al. 2024).

However, most methods predominantly depend on whole-genome SNP markers. Despite their strong associations with phenotypic traits, these markers are discrete and often overlook critical biological context, such as marker-related genomic backgrounds and linear ordering(Grinberg et al. 2020). Moreover, SNPs represent only a fraction of genomic variations, failing to account for other types of genetic variations, such as insertions, deletions, and structural variations (e.g., copy number variations, inversions, translocations), which also significantly influence phenotypes(Mérot et al. 2020). Additionally, factors like sample quality, sequencing depth, and variation detection algorithms can lead to errors in SNP detection, further complicating phenotype prediction(Kotlarz et al. 2024). To solve these problems, we hypothesize that integrating whole-genome features or sequences adjacent to SNPs could provide more comprehensive genetic variation information, potentially enhancing the interpretability of genomic mechanisms and improving phenotype prediction accuracy.

DNA and RNA sequences are composed of four nucleotide bases (A, G, C, T, or U), analogous to the 26 letters that form words in the English language. The specific arrangement of these nucleotides encodes genetic information, much like how the sequential ordering of letters forms meaningful words and sentences. Within these sequences, distinct patterns such as promoters, terminators, and coding regions play crucial roles in regulating gene expression. This regulatory mechanism parallels the way specific grammatical rules govern the construction of semantically correct sentences in natural language. Furthermore, the three-dimensional folding of biological sequences enables interactions between distant sequence elements, mirroring the long-range syntactic dependencies observed between words in different positions within a sentence. This similarity not only provides us with a novel perspective for understanding biological sequences, but also suggests a certain scientific rationale for applying natural language processing methods to the analysis of biological sequences. The advent of pre-trained models like BERT has revolutionized natural language processing(Devlin et al. 2018), offering a novel perspective and framework for biological sequence analysis. This has significant implications for downstream tasks, such as gene expression regulation and genomic prediction. In bioinformatics, foundational models like DNABERT, pre-trained on the human reference genome GRCh38.p13, have been fine-tuned for tasks such as promoter prediction, transcription factor binding site (TFBS) identification, and splice site prediction(Ji et al. 2021). DNABERT-2, an enhanced multi-species genomic model, improves long-sequence processing efficiency and achieves state-of-the-art performance on the GUE dataset(Zhou et al. 2023). Similarly, HyenaDNA, pre-trained on a single human reference genome, leverages implicit convolutions and instruction tuning to excel in tasks like chromatin spectrum prediction and species classification(Nguyen et al. 2023). However, the computational complexity and substantial time requirements associated with whole-genome sequences, compared to SNP markers, have significantly constrained their application in crop phenotype prediction. Consequently, systematic investigations into whole-genome sequence-based approaches for crop phenotype prediction remain notably lacking in the literature.

We developed two distinct genomic feature representation approaches: (1) whole-genome sequence embedding that captures global genomic patterns, and (2) SNP-context embedding that extracts local sequence features within defined flanking regions (500-3000 bp) surrounding each SNP marker. Then, we generated feature embeddings by the cross-species genomic foundational model DNABERT-2 for three crop datasets (rice413, rice395, and maize301) and employed Principal Component Analysis (PCA) and Partial Least Squares (PLS) for dimensionality reduction. We compared phenotypic prediction accuracy (measured by Pearson correlation coefficient) across

different dimensions, dimensionality reduction methods (PLS, PCA, and no reduction), and SNPs context lengths. Our results demonstrate that while both embedding modes have limitations in predicting traits with low heritability in rice395, the SNPs context feature embedding mode significantly outperforms SNPs marker-based predictions for traits with low-to-medium heritability, achieving higher accuracy and lower mean absolute error (MAE). For some traits, its accuracy even surpasses that of whole-genome feature embedding. When predicting traits with high heritability, the whole-genome feature embedding mode, combined with conventional machine learning algorithms, significantly improves prediction accuracy and reduces MAE compared to state-of-the-art deep learning models like DNNGP and ResGS.

Materials and methods

Plant materials

We have collected publicly available datasets of whole-genome SNP markers and phenotypic traits for three crop datasets: rice413, rice395, and maize301, from various sources. The rice413 and rice395 rice datasets were obtained from the literature(Zhao et al. 2011). The rice413 dataset comprises 413 inbred lines of Asian rice from 82 countries, containing 44,100 SNP markers and 34 traits. Among these, the alkali spreading value (ASV), amylose content (AC), panicle number per plant (PNPP), protein content (PC), seed length (SL), and seed number per panicle (SNPP) were selected as target traits, with non-missing phenotype values for 403, 401, 372, 393, 377, and 376 samples, respectively. The rice395 dataset consists of 395 diverse Asian rice samples spanning five subpopulations: indica, japonica, tropical japonica, temperate japonica, and group v, with a total of 1,536 SNP markers. The target traits for rice395 are amylose content (AC) and seed length (SL), with complete phenotypic records available for 356 and 372 accessions, respectively. The maize301 SNPs data, used as the source for whole-genome data conversion in maize, was selected from the built-in dataset in Tassel software (Bradbury et al. 2007). It includes 3,093 SNP markers, with target traits being days to pollination (Dpoll ; n=276), ear diameter (EarDia; n =249), and ear height (EarHT; n=279).

Reconstructing Whole-genome sequences from SNPs

To enable whole-genome feature embedding, we developed a pipeline for reconstructing complete genome sequences from variant call format (VCF) files containing single nucleotide polymorphisms (SNPs). The conversion of crop whole-genome SNP markers into FASTA-formatted sequences involves several key steps to ensure accuracy and efficiency.

First, preprocessing of the reference genome FASTA file and VCF file is performed to optimize downstream analysis. The reference genome is indexed using samtools, significantly improving the speed of sequence retrieval and alignment. Simultaneously, the VCF file is indexed using bcftools, allowing efficient access to SNP information during processing.

Next, sample names are extracted from the VCF header, a crucial step for generating individual-specific genome sequences. These names are systematically recorded in a file (SampleName.list) to enable batch processing in subsequent steps.

For each sample, a custom Python script integrates the reference genome with sample-specific variants using pysam and pyfaidx. The script sequentially examines each genomic locus in the reference, checks for corresponding entries in the VCF file, and modifies the reference sequence at variant positions. This process reconstructs a personalized genome sequence that incorporates all observed SNPs for the given sample.

Finally, the resulting FASTA files undergo stringent quality control to confirm correct formatting and variant incorporation. This includes validation of header information, sequence integrity, and

consistency between expected and observed variants. These measures ensure the reliability of the reconstructed genomes for subsequent analyses.

Relative concepts

DNABERT-2 is a genome-based model built on the Transformer architecture, distinguishing itself from DNABERT through several key innovations. Unlike DNABERT, which relies on k-mer tokenization, DNABERT-2 employs Byte-Pair Embedding (BPE), a statistical data compression algorithm(Sennrich et al. 2015), to tokenize DNA sequences. Additionally, it replaces learned positional embeddings with Attention with Linear Biases (ALiBi), eliminating input length limitations(Press et al. 2021), and integrates Flash Attention to enhance computational efficiency (Dao et al. 2022). Pre-trained on datasets spanning 135 species across 7 categories, DNABERT-2 demonstrates robust cross-species genomic generalization capabilities, making it a versatile tool for genomic analysis.

To minimize the noise unrelated to traits in high-dimensional feature vectors, we perform dimensionality reduction on the encoded feature vectors. Principal Component Analysis (PCA) is a widely used unsupervised dimensionality reduction technique. Its core principle involves transforming multiple variables in the original dataset into a smaller set of uncorrelated principal components through linear transformation. This process reduces data dimensionality while retaining the most significant information from the original dataset. The resulting principal components are linear combinations of the original features and are mutually orthogonal. By eliminating redundant information, PCA reduces noise in high-dimensional data, thereby improving the accuracy of subsequent analyses(Galinha and Crespo 2022). This method is particularly valuable in metabolomics for analyzing high-dimensional chromatography and mass spectrometry data(Nguyen and Rocke 2002; Broadhurst et al. 1997).

Partial Least Squares (PLS) is another method for dimensionality reduction, first introduced by Herman Wold in the 1960s(Wold et al. 1983), integrates the strengths of Principal Component Analysis, Canonical Correlation Analysis, and Multiple Regression Analysis. It is particularly effective in scenarios where multicollinearity exists between dependent and independent variables. As a supervised feature extraction method, PLS is widely applied in gene expression microarray data analysis. Its core objective is to identify orthogonal projection directions (principal components) that maximize the covariance between projected dependent and independent variables, thereby facilitating the construction of predictive models.

Method implement

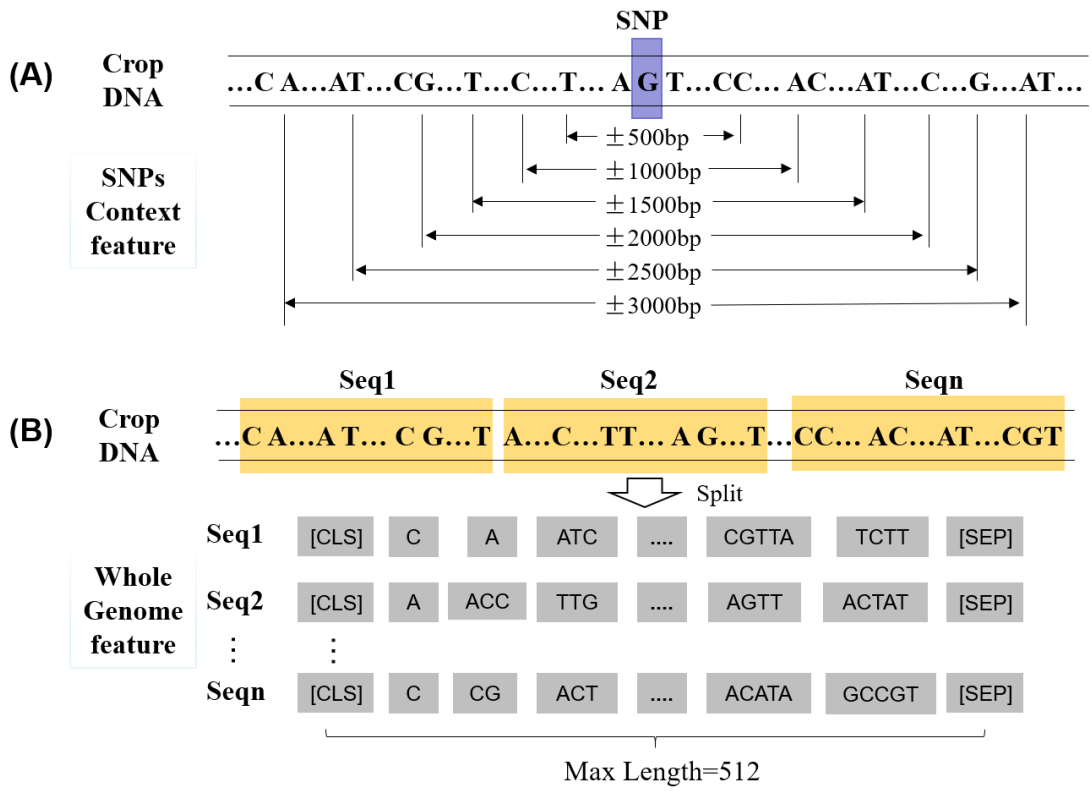
We present two distinct genomic embedding approaches for crop phenotypic prediction: SNP-context feature embedding (Figure 1A) and whole-genome feature embedding (Figure 1B). By integrating DNABERT-2 with dimensionality reduction techniques, we developed a unified framework capable of operating under both embedding paradigms (Figure 1C). To enable rigorous comparison with traditional SNP marker-based prediction methods, we implemented a three-stage analytical pipeline.

The first stage involves sequence preprocessing and whole-genome feature embedding. For the rice413, rice395, and maize301 datasets, we processed reconstructed whole-genome sequences by removing ambiguous bases ("N") arising from sequencing artifacts. Each sample's filtered sequences were partitioned into chromosomal segments, which were further divided into model-compatible subsequences. These subsequences were formatted with [CLS] and [SEP] tokens while maintaining a maximum length of 512 nucleotides to comply with DNABERT-2's input constraints. The model then processed these subsequences in batch mode to generate comprehensive whole-genome feature vectors.

In the second stage, we performed mutation site analysis and SNP-context feature embedding. From the VCF files, we systematically identified all heterozygous (0/1) and homozygous (1/1) SNP loci for each sample. For each mutation site, we extracted flanking sequences of varying context lengths (500-3000 bp), ensuring the fragments remained within chromosomal boundaries and avoided overlapping regions containing adjacent SNPs. DNABERT-2 processed these context-aware sequences to produce SNP-specific feature vectors, with a collision avoidance protocol that skipped secondary SNPs falling within previously captured genomic regions.

Due to the extensive length of whole-genome sequences, they may carry noise that is detrimental to phenotypic prediction. Therefore, in the third stage, dimensionality reduction is applied to the whole-genome feature embedding. Considering the variable sample sizes across different crops after phenotype null value removal, we designed an exponentially increasing target dimensionality search range to mitigate the impact of high-dimensional data on model performance. The dimensionality parameter d was defined as $\{d|d=2^n, n \in Z+, 2 \leq n \leq \lfloor \log_2 N \rfloor\}$, where N represents the sample count post-phenotype filtering. We employed both Principal Component Analysis (PCA) and Partial Least Squares (PLS) dimensionality reduction techniques to process the genomic vectors while preserving the original dimensional integrity of prediction indicators.

The dataset was partitioned into training (80%), testing (10%), and validation (10%) subsets to ensure robust model evaluation. We conducted exhaustive hyperparameter optimization through grid search across multiple machine learning architectures. For the purpose of benchmarking, we implemented several well-established genomic prediction methods, including Support Vector Regression (SVR), Ridge Regression Best Linear Unbiased Prediction (rrBLUP), Random Forest (RF), and Gradient Boosting Regression (GBR), by leveraging the automated machine learning framework provided by PyCaret. This platform streamlined comparative analysis by handling data preprocessing, feature engineering, and model selection through its integrated pipeline.



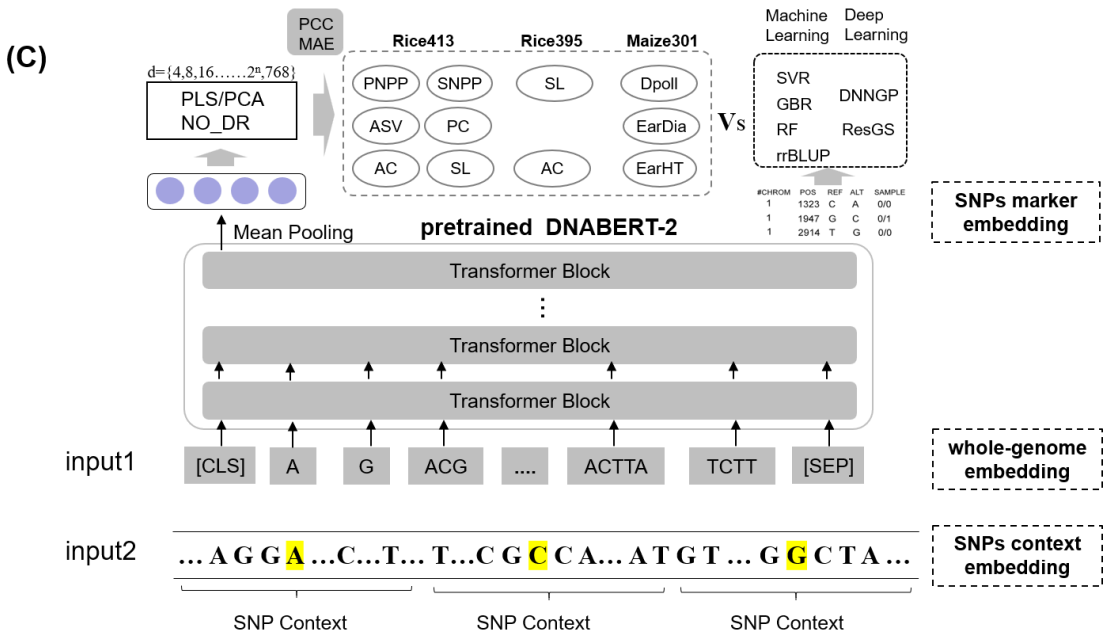


Figure 1 Framework for predicting crop phenotypes based on two distinct embedding modes. (A) Example for SNP context feature embedding mode. The symbol "±" denotes the extraction of a specific number of base pairs upstream and downstream from the current SNP locus. (B) Example for whole-genome features embedding mode. The genome is divided into consecutive sections (Sec1, Sec2, ..., Secn), each further partitioned into sub-sections with a maximum input length of 512 base pairs. Each sub-section, analogous to a sentence in natural language processing, begins with a [CLS] token and ends with a [SEP] token. Sub-sections shorter than 512 base pairs are padded using the [PAD] token from the vocabulary. (C) Generation of crop genomic feature vectors under two embedding modes using the DNABERT-2 pre-trained model.

Indicator evaluation and calculation

Model evaluation was conducted through calculation of the Pearson Correlation Coefficient (PCC) between vectors of true and predicted phenotypic values. We implemented ten-fold cross-validation, with mean PCC values serving as the primary metric for assessing prediction accuracy. Additionally, Mean Absolute Error (MAE) was employed as a supplementary metric for evaluating quantitative trait prediction accuracy in crops. The computational methodologies for these metrics are detailed in the corresponding equations.

We evaluated model performance using rigorous statistical metrics to assess prediction accuracy. The primary evaluation metric was the Pearson Correlation Coefficient (PCC), calculated between vectors of observed and predicted phenotypic values. This measure quantifies the linear relationship between predicted and actual trait values, with higher coefficients indicating better predictive performance.

To ensure robust performance estimation, we implemented ten-fold cross-validation across all datasets. This validation approach partitions the data into ten subsets, using nine for training and one for testing in each iteration. We report the mean PCC values across all folds as our primary accuracy metric, providing a comprehensive assessment of model generalizability.

As a complementary measure of prediction accuracy, we calculated the Mean Absolute Error (MAE) for quantitative trait predictions. This metric provides an intuitive measure of average prediction error magnitude in the original units of measurement, offering additional insights into model performance beyond correlation-based metrics.

The mathematical formulations for these evaluation metrics are provided in Equations (1)-(2), with detailed derivations available in the Supplementary Materials. This dual-metric approach ensures comprehensive assessment of both the strength of association (PCC) and absolute prediction accuracy (MAE) across all tested phenotypes.

$$PCC = \frac{\mathrm{cov}(Y_t, Y_p)}{\sigma_{Y_t}\sigma_{Y_p}} \tag{1}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | Y_t - Y_p | \tag{2}$$

Where, cov represents the covariance between the true phenotypic values Y_t and the predicted phenotypic values Y_p, they represent the standard deviations of the true phenotypic values Y_t and the predicted phenotypic values Y_p, respectively.

We assessed trait heritability using broad-sense heritability (h^2), following established genomic prediction protocols(Wang et al. 2022). This comprehensive metric quantifies the proportion of total phenotypic variance attributable to all genetic components, including additive effects, dominance variance, and epistatic interactions across the rice413, rice395, and maize301 populations.

The broad-sense heritability was calculated as:

$$h^2 = \frac{V_g}{V_p} = \frac{V_a + V_d}{V_a + V_d + V_e} \tag{3}$$

This formulation enables a robust assessment of the genetic contribution to phenotypic variation, accounting for both genetic and environmental influences. In the formula, V_a denotes the additive genetic variance, V_d denotes the dominant genetic variance, and V_e denotes the environmental variance.

Results

Phenotypic prediction results under SNP-context feature embedding mode

We investigated how genetic information in SNP-flanking regions influences phenotypic prediction accuracy by adapting contextual analysis approaches from natural language processing. For each SNP variant site, we extracted genomic segments of increasing lengths (500-3000 bp) centered on the polymorphic position, encompassing both upstream and downstream sequences. These context windows, which included the variant site itself, were processed through our feature embedding pipeline to capture local genomic information.

For the assessment of prediction performance, four representative models were implemented, namely SVR), rrBLUP, RF, and GBR. We evaluated the capability of each model based on rice413 dataset in predicting phenotypic traits across varying window sizes of SNP contexts.

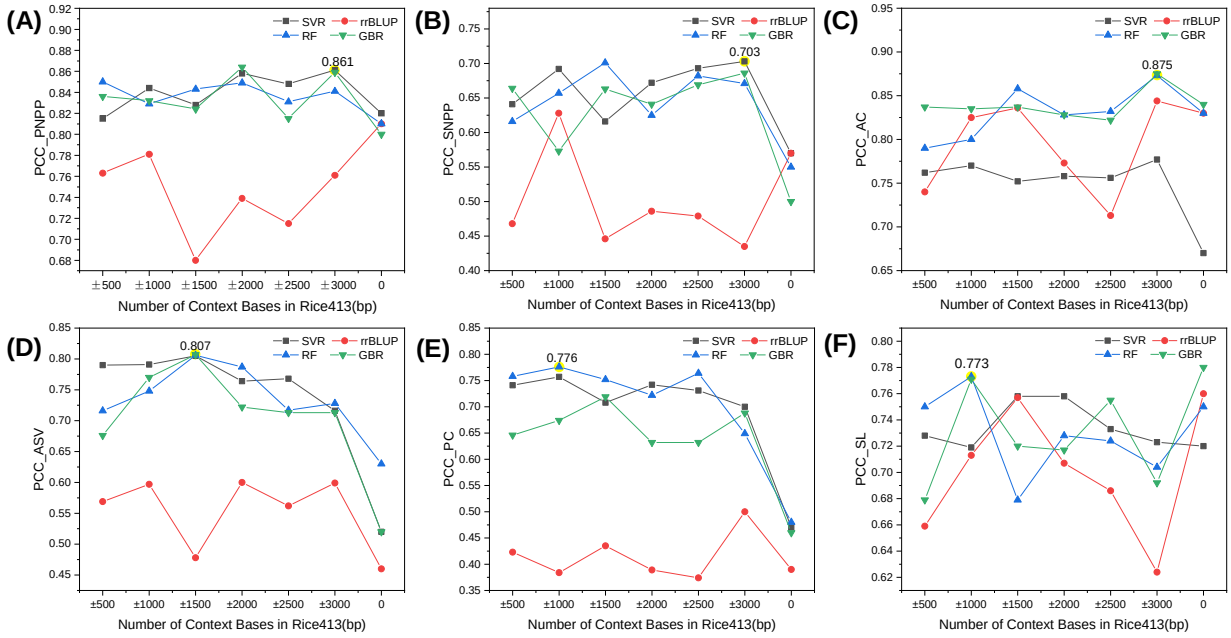


Figure 2 Phenotypic prediction accuracy for six agronomic traits in rice413 using SNP-context features (500bp-3000bp) across four machine learning algorithms. (A) PNPP; (B) SNPP; (C) AC; (D) ASV; (E) PC; (F) SL. Prediction performance was evaluated by PCC for SVR, rrBLUP, RF, and GBR models.

The horizontal axis labels ±500, ±1000, ±1500, ±2000, ±2500, and ±3000 denote the genomic sequences extracted from 500bp, 1000bp, 1500bp, 2000bp, 2500bp, and 3000bp upstream and downstream of the SNP variant sites, respectively(Figure 2). The horizontal coordinate 0 represents the traditional trait prediction method that relies solely on SNP marker features, with the corresponding PCC values derived from the trait prediction results reported in the literature(Xie et al. 2024). Yellow circles highlight the highest PCC values achieved within the context length range of 500bp to 3000bp under the context feature embedding mode.

For traits such as PNPP, SNPP, and PC, the prediction accuracy of rrBLUP is comparatively lower when using the same length of contextual genomic sequences. The highest prediction accuracies for traits PNPP, SNPP, ASV, AC, and PC are achieved at context lengths of 3000bp, 3000bp, 1500bp, 3000bp, and 1000bp, respectively. Notably, the PCC between predicted and true phenotypic values are generally higher than those obtained using only SNP loci for prediction. For SL, certain algorithms demonstrate prediction accuracies that are either superior to or comparable with those achieved using only SNP features when the context length ranges between 1000bp and 2000bp. These findings underscore the importance of incorporating flanking genomic sequences in enhancing phenotypic prediction accuracy.

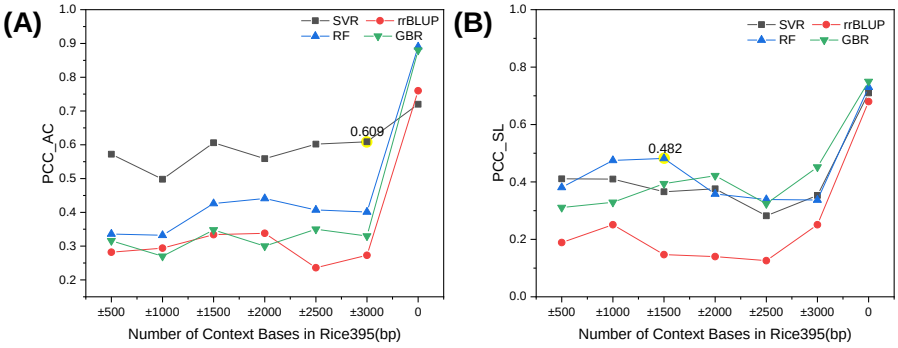


Figure 3 Prediction accuracy for AC and SL traits in rice395 using four machine learning algorithms with SNP-context features (±500bp to ±3000 bp). (A) AC; (B) SL. Performance was evaluated by PCC for SVR, rrBLUP, RF, and GBR models.

As illustrated in Figure 3, for the AC trait in the rice395 dataset, the prediction accuracy of SVR demonstrated statistically significant superiority over other algorithms across SNP context lengths ranging from 500 bp to 3000 bp, achieving its maximum predictive performance of 0.609 at the 3000 bp context length. Regarding the SL trait under SNP-context feature encoding, SVR, RF and GBR all exhibited superior prediction accuracy compared to rrBLUP. Notably, RF attained the highest accuracy of 0.482 at the 1500 bp SNP context length. It is evident that the prediction accuracy for the AC and SL in rice395, based on SNPs context feature encoding, is lower than that of traditional SNPs marker features.

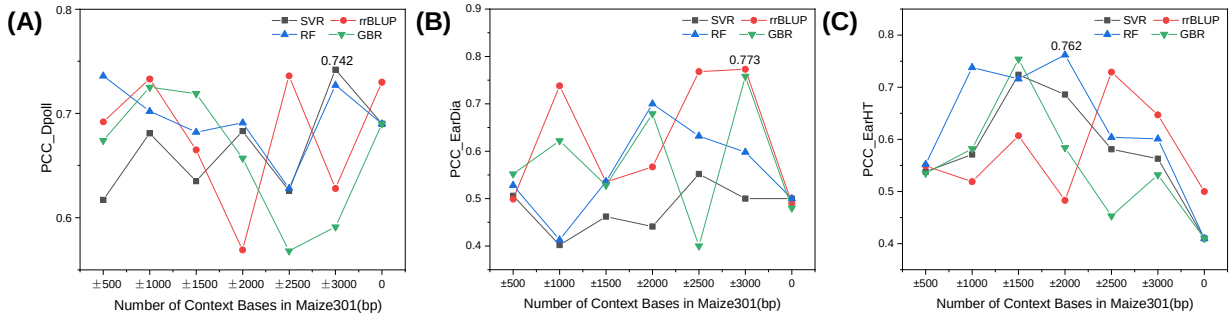


Figure 4 Prediction accuracy in maize301 using SNP-context features (500-3000 bp) across four machine-learning algorithms. (A) Days to pollination (Dpoll); (B) Ear diameter (EarDia); (C) Ear height (EarHT). Performance was evaluated using Pearson correlation coefficients.

The analysis reveals significant variation in prediction accuracy for Dpoll across different algorithms using SNP-context features. Notably, several configurations outperform traditional SNP marker-based approaches. SVR demonstrates the strongest performance for Dpoll prediction, achieving a peak accuracy of 0.742 at a context length of 3000bp.

For EarDia, rrBLUP shows optimal performance with an accuracy of 0.773 at 3000 bp. RF performs best for EarHT prediction, reaching 0.762 accuracy at 2000 bp context length. These results highlight the context length-dependent nature of prediction performance across different traits.

Comparative evaluation demonstrates consistent improvements when using SNP-context features compared to traditional SNP markers. The average accuracy gains are most substantial for EarHT (35.2%), followed by EarDia (28.3%) and Dpoll (5.2%). This pattern suggests that the relative benefit of contextual sequence information varies considerably among traits, potentially reflecting differences in their underlying genetic architectures.

Phenotypic prediction results based on Whole-genome feature embedding mode

The genetic regulation of crop traits involves complex interactions among multiple major genes, minor genes, and quantitative trait loci (QTLs) distributed throughout the genome. These genetic elements often span multiple chromosomes and may exhibit non-linear epistatic interactions, presenting significant challenges for prediction methods relying solely on SNPs or localized sequence contexts.

To overcome these limitations, we implemented a whole-genome feature embedding strategy that captures nucleotide sequence information across complete chromosomal sequences. This comprehensive approach was complemented by dimensionality reduction through both PCA and PLS. By systematically evaluating prediction performance across different reduced dimensions, our

framework provides enhanced capacity to model the distributed genetic architecture underlying complex crop traits.

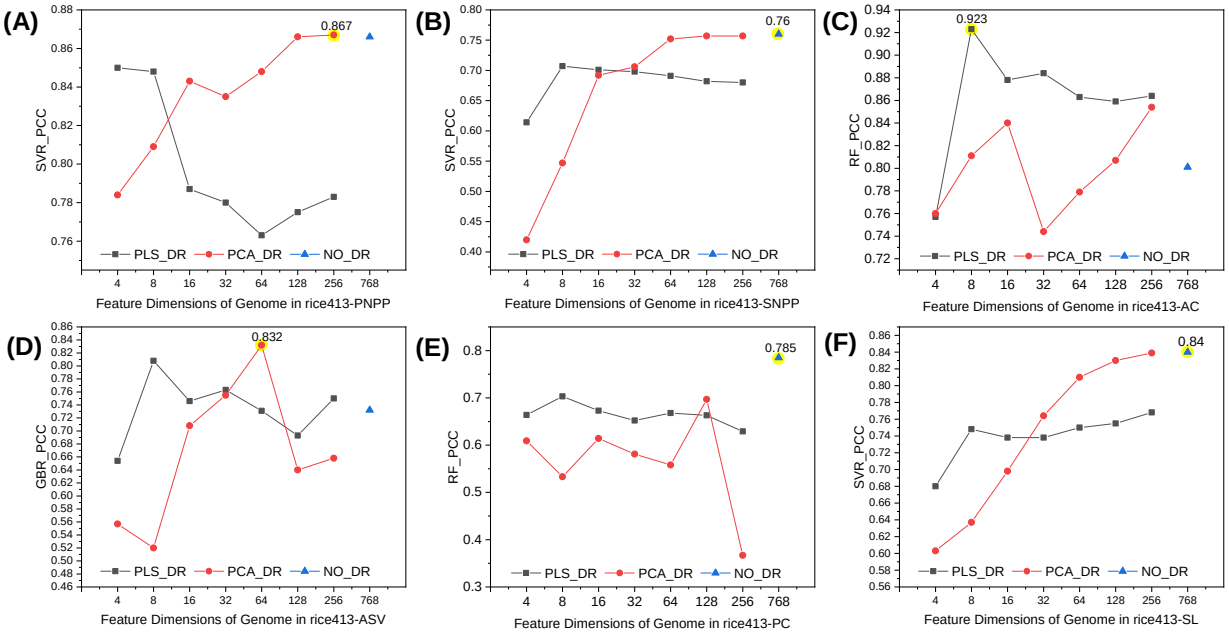


Figure 5 Comparative performance of whole-genome feature embedding with dimensionality reduction (PLS/PCA) versus full-dimensional (768D) approaches for six agronomic traits in rice413. (A) PNPP; (B) SNPP;(C) AC; (D) ASV; (E) PC; (F) SL. Yellow highlights denote optimal performance among four machine learning algorithms (SVR, rrBLUP, RF, GBR). Complete algorithm comparisons available in Figure S1.

Our comprehensive evaluation reveals significant variation in prediction accuracy across different machine learning algorithms and feature dimensionality approaches (Figures 5 and S1). The results demonstrate clear trait-specific patterns in optimal prediction strategies.

For PNPP, PCA demonstrates superior performance to PLS at higher dimensionalities (>16 components). SVR achieves maximal prediction accuracy (PCC = 0.867) at 256 dimensions, matching the performance of full-dimensional embedding. Similar optimal performance is observed for SNPP, where SVR with complete feature sets yields the highest accuracy (PCC = 0.76).

Notably distinct patterns emerge for quality-related traits. AC shows markedly improved prediction using PLS dimensionality reduction, particularly when combined with RF, reaching peak accuracy (PCC = 0.923) at just 8 dimensions. For ASV, GBR with PCA at intermediate dimensionality (64 components) provides optimal results (PCC = 0.832). PC and SL both achieve maximum accuracy without dimensionality reduction, using RF (PCC = 0.785) and SVR (PCC = 0.84) respectively.

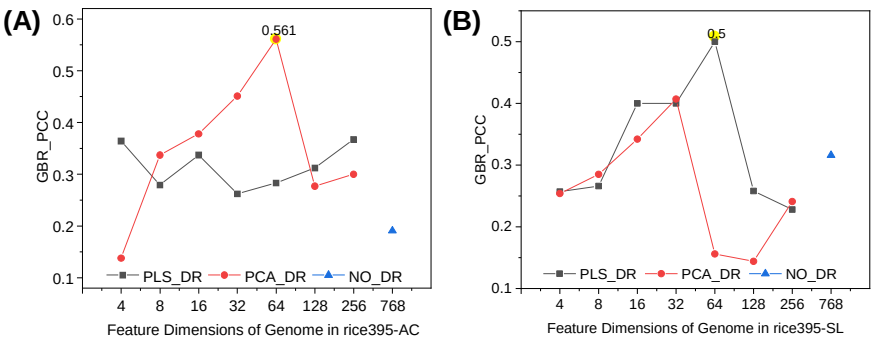


Figure 6 Prediction accuracy for AC and SL traits in rice395 using whole-genome feature embedding with dimensionality reduction (PLS, PCA) or full dimensions (768). (A) AC; (B) SL. Yellow highlights indicate highest accuracy among four algorithms (SVR, rrBLUP, RF, GBR). Complete algorithm comparisons shown in Figure S2.

As shown in Figure 6 and Figure S2(a)~(f), the prediction accuracy for the AC trait in the rice395 dataset remains relatively stable with increasing dimensionality when using SVR, outperforming other algorithms in consistency. The highest prediction accuracy for AC (0.561) is achieved when GBR is combined with PCA at 64 dimensions. For the SL trait, the highest prediction accuracy (0.5) is attained using GBR with PLS at 64 dimensions. These results suggest that dimensionality reduction techniques, particularly at intermediate dimensions (e.g., 64), can enhance prediction accuracy for specific traits. However, the performance varies significantly across traits and algorithms, highlighting the importance of tailoring dimensionality reduction and algorithm selection to the specific trait of interest. Additionally, the relatively lower prediction accuracy for SL compared to AC may reflect the greater genetic complexity or environmental influence on this trait, warranting further investigation into alternative modeling approaches or feature representations.

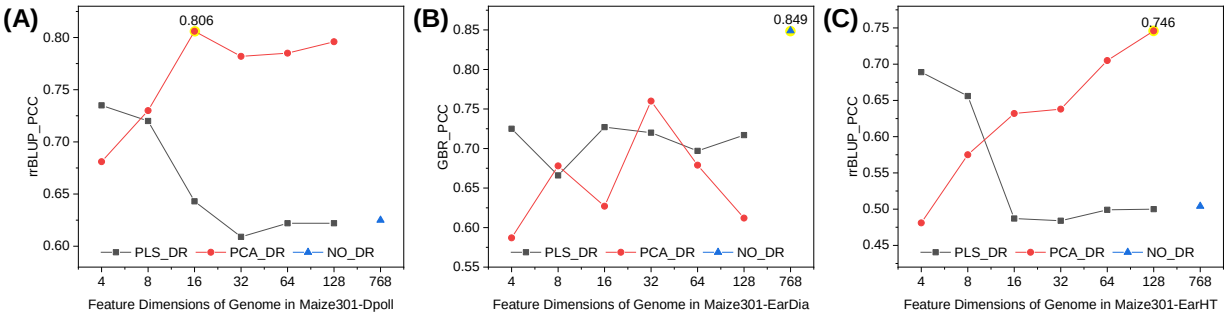


Figure 7 Prediction accuracy for maize301 traits using whole-genome feature embedding with dimensionality reduction (PLS, PCA) or full dimensions (768D). (A) Dpoll; (B) EarDia; (C) EarHT. Yellow highlights indicate optimal performance among four algorithms (SVR, rrBLUP, RF, GBR). Complete results are shown in Figure S3.

We can find PLS dimensionality reduction outperforms PCA when SVR is used to predict Dpoll, EarDia, and EarHT(Figure 7). For Dpoll, the highest prediction accuracy (0.806) is achieved using rrBLUP with PCA at 16 dimensions, representing a 7.6% improvement over traditional SNP feature embedding. For EarDia, GBR without dimensionality reduction yields the highest accuracy (0.849), a 35.9% improvement over traditional methods. For EarHT, rrBLUP combined with PCA at 128 dimensions achieves the maximum accuracy (0.746), a 24.6% increase over traditional SNP feature embedding.

The highest average prediction accuracies for each trait across the rice413, rice395, and maize301 datasets using whole-genome feature embedding, SNP context feature embedding, and traditional SNP features ("SNP*" column) are summarized in Tables 1. The "SNP*" column represents the optimal traditional machine learning algorithm based on SNP marker features, with its values derived from the literature(Xie et al. 2024). When the best-performing algorithms differ between SNP context embedding and whole-genome embedding, the SNP* column reflects the highest precision value from either algorithm. Notably, no single algorithm consistently achieves the best prediction performance across all traits, consistent with findings in prior studies(Azodi et al. 2019; Montesinos-López et al. 2018). Furthermore, certain traits exhibit higher prediction accuracy when whole-genome feature embedding is combined with PCA or PLS dimensionality reduction, compared to no dimensionality reduction. This underscores the importance of dimensionality reduction in retaining core genetic features, minimizing noise, and enhancing model generalization.

Table1 Highest average prediction accuracy for various traits in rice413 under different feature modes

Traits for rice413	SNP*	SNP-context Feature embedding			Whole-genome feature embedding		
	PCC	Algorithm Type	Length of Context	PCC	Algorithm Type	Dimensionality Reduction Method	PCC
PNPP	0.82	GBR	3000bp	0.861	SVR	PCA(d=256)	0867
SNPP	0.57	SVR	3000bp	0.703	SVR	No_DR(d=768)	0.76
AC	0.84	GBR	3000bp	0.875	RF	PLS(d=8)	0.923
ASV	0.63	GBR	1500bp	0.807	GBR	PCA(d=64)	0.832
PC	0.48	RF	1000bp	0.776	RF	NO_DR(d=768)	0.785
SL	0.75	RF	1000bp	0.773	SVR	NO_DR(d=768)	0.84

Table2 Highest average prediction accuracy for various traits in rice395 under different feature modes

Traits for rice395	SNP*	SNP-context Feature embedding			Whole-genome feature embedding		
	PCC	Algorithm Type	Length of Context	PCC	Algorithm Type	Dimensionality Reduction Method	PCC
AC	0.88	SVR	3000bp	0.609	GBR	PCA(d=64)	0.561
SL	0.75	RF	1500bp	0.482	GBR	PLS(d=64)	0.5

Table3 Highest average prediction accuracy for various traits in maize301 under different feature modes

Traits for maize301	SNP*	SNP-context Feature embedding			Whole-genome feature embedding		
	PCC	Algorithm Type	Length of Context	PCC	Algorithm Type	Dimensionality Reduction Method	PCC
Dpoll	0.73	SVR	3000bp	0.742	rrBLUP	PCA(d=16)	0.806
EarDia	0.49	rrBLUP	3000bp	0.773	GBR	NO_DR(d=768)	0.849
EarHT	0.5	RF	2000bp	0.762	rrBLUP	PCA(d=128)	0.746

Comparative analysis of prediction accuracy across different models and feature embedding methods, as illustrated in Figure 8, reveals significant performance variations among the evaluated

traits. For the PNPP trait, whole-genome feature embedding demonstrates nearly equivalent predictive performance to SNP-context feature embedding, with marginal accuracy improvements of 2.7% and 0.7% over the DNNGP and ResGS models, respectively. Notably, the SNPP trait exhibits superior prediction accuracy when utilizing whole-genome feature embedding, outperforming SNP-context feature embedding by 5.7% and substantially exceeding DNNGP and ResGS models by 10% and 9%, respectively.

Regarding the AC trait, whole-genome feature embedding achieves a 4.8% higher accuracy compared to SNP-context feature embedding, while surpassing DNNGP and ResGS models by 10.3% and 5.3%, respectively. The most remarkable performance enhancement is observed in the ASV trait, where whole-genome feature embedding demonstrates a 4.7% improvement over SNP-context feature embedding and substantial advantages of 28.4% and 15.4% over DNNGP and ResGS models, respectively.

For the PC trait, whole-genome feature embedding shows slightly better performance than SNP-context feature embedding, while achieving significant accuracy improvements of 30.5%, 29.5%, and 19.5% over traditional SNPs feature, DNNGP, and ResGS models, respectively. Finally, in the case of the SL trait, whole-genome feature embedding outperforms the traditional SNPs feature embedding method by 9%, and exceeds DNNGP and ResGS models by 6% and 2%, respectively.

These comparative results consistently demonstrate the superior predictive capability of whole-genome feature embedding across multiple traits, particularly in its substantial performance advantages over both traditional SNPs feature methods and advanced deep learning-based models (DNNGP and ResGS).

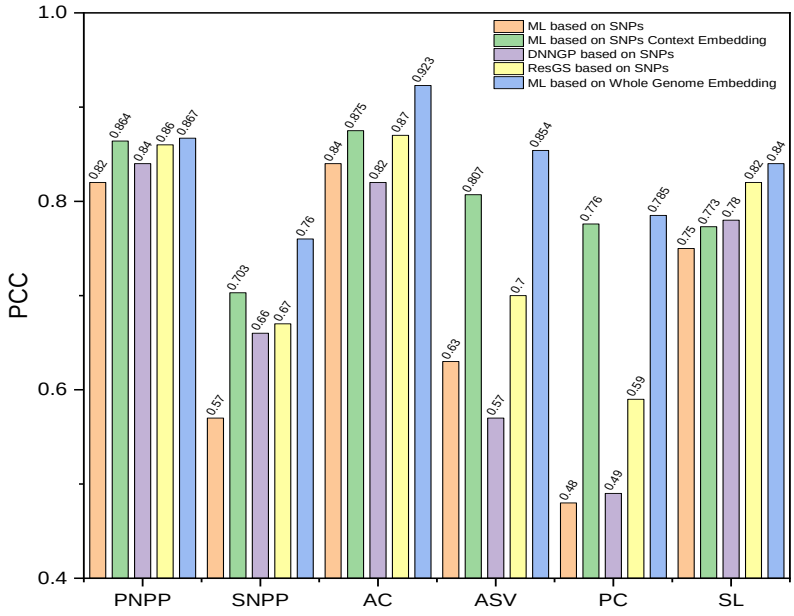


Figure 8 Comparison of prediction accuracy among traditional machine learning algorithms (abbreviated as ML), DNNGP, ResGS (Based on SNP markers) and ML based on SNP context feature embedding and whole-genome feature embedding for PNPP, SNPP, AC, ASV, PC, and SL in Rice413.

Consistent with previous studies(Norman et al. 2018; Spindel et al. 2015; Riedelsheimer et al. 2012; Ogutu et al. 2011), our analysis demonstrates that trait heritability (Table 4) significantly influences the performance of genomic prediction algorithms. The results reveal distinct patterns in prediction accuracy across different traits and methods. For SNPP, ASV, and PC traits, SNP-context feature embedding at specific context lengths (3000bp, 1500bp, and 1000bp, respectively) substantially outperformed traditional SNPs features. The implementation of whole-genome feature embedding

yielded additional accuracy improvements of 5.7%, 4.7%, and 0.9% for these traits, respectively. This enhancement can be attributed to the incorporation of linkage disequilibrium (LD) and genetic associations between SNPs, which are typically neglected in conventional SNPs feature modeling. The superior performance of whole-genome feature embedding may reflect DNABERT-2's capacity to identify major genes or key QTL characteristics within extended genomic contexts. While our investigation was limited to six context lengths within a 3000bp range, the potential for improved accuracy through longer context lengths remains unexplored.

Notably, PNPP trait prediction exhibited comparable accuracy across SNP-context feature embedding, whole-genome feature embedding, and the ResGS model. This observation suggests that the selected SNP markers may encompass core quantitative trait loci crucial for PNPP expression, rendering additional genomic context information redundant. Phenotypes with moderate heritability are particularly susceptible to intricate gene regulatory networks and environmental perturbations, which complicates the disentanglement of causal genetic variants from stochastic noise.

For AC and SL traits, substantial accuracy improvements were observed with advanced embedding methods. SNP-context feature embedding enhanced prediction accuracy by 3.5% and 2.3%, respectively, while whole-genome feature embedding achieved more significant improvements of 8.3% and 9%. The superior performance of whole-genome feature embedding for these traits aligns with their higher heritability values, suggesting greater genetic determination. This phenomenon likely reflects the complex regulation of AC and SL traits by multiple genes or QTLs, where comprehensive genomic information captures additional non-additive and recessive effects. These findings corroborate previous observations in animal genomic prediction studies(Bolormaa et al. 2013; Daetwyler et al. 2010) regarding the positive correlation between heritability and prediction accuracy.

The collective results underscore the importance of considering both methodological approaches and trait-specific genetic architectures in genomic prediction. Whole-genome feature embedding emerges as a particularly powerful tool for traits with complex genetic regulation and higher heritability, while context-specific approaches may suffice for traits dominated by major QTLs.

Table4 Heritability of Traits in rice413 Based on Genetic Analysis

Trait for rice413	Heritability
AC	0.775324
ASV	0.412983
PC	0.417095
PNPP	0.485917
SL	0.982803
SNPP	0.505230

PCC primarily assesses the linear association strength between predicted and actual trait values, whereas the MAE quantifies the magnitude of deviation between predicted and actual values. In this study, MAE was concurrently computed for various prediction methods applied to the rice413 dataset, with results illustrated in Figure 9. The MAE values corresponding to single-nucleotide polymorphisms (SNPs) were sourced from the literature (Xie et al., 2024). In instances where the algorithms achieving the highest average prediction accuracy differed between SNPs context encoding and genome-wide encoding, the minimum MAE based on SNPs marker features from both algorithms was adopted.

Overall, the traits of panicle number per plant (PNPP), spikelet number per panicle (SNPP), amylose content (AC), protein content (PC), and spikelet length (SL) in the rice413 dataset demonstrated the lowest MAE when employing genome-wide feature encoding methods. Specifically, for trait AC, the MAE derived from SNPs context feature encoding was comparable to that of

traditional SNPs features and the ResGS model; however, a substantial decrease in MAE was observed when predictions were based on genome-wide feature encoding. Similarly, trait SL exhibited a further reduction in MAE with genome-wide feature encoding compared to alternative approaches.

For traits such as amylose-starch viscosity (ASV) and PC, the utilization of SNPs context features led to a significant reduction in MAE compared to traditional SNPs features, the deep neural network for genomic prediction (DNNGP), and the ResGS model. This improvement may be attributed to the fact that traditional SNPs features typically concentrate solely on the effects of individual SNPs, often overlooking the genetic context of neighboring regions, which can result in the loss or distortion of trait-genetic association signals. In contrast, encoding SNPs context features using DNABERT-2 or genome-wide sequences across a broader genomic range offers the advantage of capturing distant trait-genetic associations through long-range dependency modeling. This approach effectively addresses signal loss issues arising from linkage disequilibrium in conventional methods, thereby mitigating trait prediction errors.

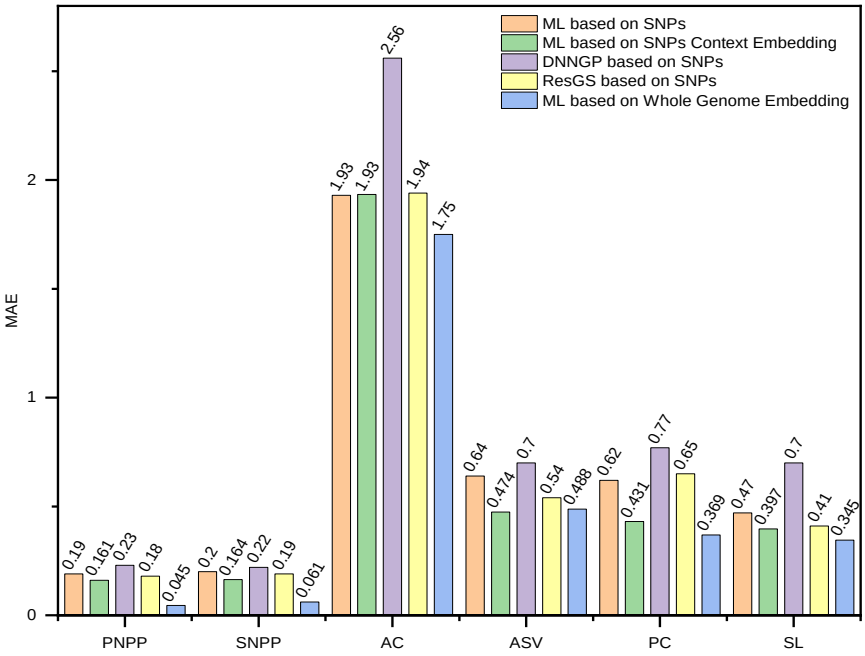


Figure 9 Comparison of Mean Absolute Error (MAE) among traditional machine learning algorithms (abbreviated as ML), DNNGP, ResGS (Based on SNP markers) and ML based on SNP context feature embedding and whole-genome feature embedding for PNPP, SNPP, AC, ASV, PC, and SL in Rice413.

The MAE metric complements PCC by quantifying the absolute deviation between predicted and true trait values. Figure 9 presents the MAE values for predicting six traits (PNPP, SNPP, AC, ASV, PC, and SL) in the rice413 dataset across five methods: SNP markers features, SNP-context feature embedding, DNNGP, ResGS, and whole-genome feature embedding. The MAE values for SNP markers features were derived from Xie et al. (2024). When discrepancies arose in the highest prediction accuracy between SNP-context embedding and whole-genome embedding, the minimum MAE from both algorithms was selected for SNP markers features.

Overall, whole-genome feature embedding achieved the lowest MAE for PNPP, SNPP, AC, PC, and SL traits. For the AC trait, SNP-context feature embedding exhibited comparable MAE to traditional SNPs features and the ResGS model, but whole-genome feature embedding significantly reduced the MAE. Similarly, the SL trait showed further MAE reduction with whole-genome feature embedding compared to other methods. For ASV and PC traits, SNP-context features substantially lowered the MAE relative to traditional SNPs features, DNNGP, and ResGS. This improvement likely stems from the ability of DNABERT-2 to capture long-range dependencies and genetic associations within extended genomic contexts, addressing the limitations of traditional methods that focus solely

on individual SNPs and often overlook neighboring genetic regions. By modeling broader genomic sequences, DNABERT-2 mitigates signal loss due to linkage disequilibrium, thereby reducing prediction errors.

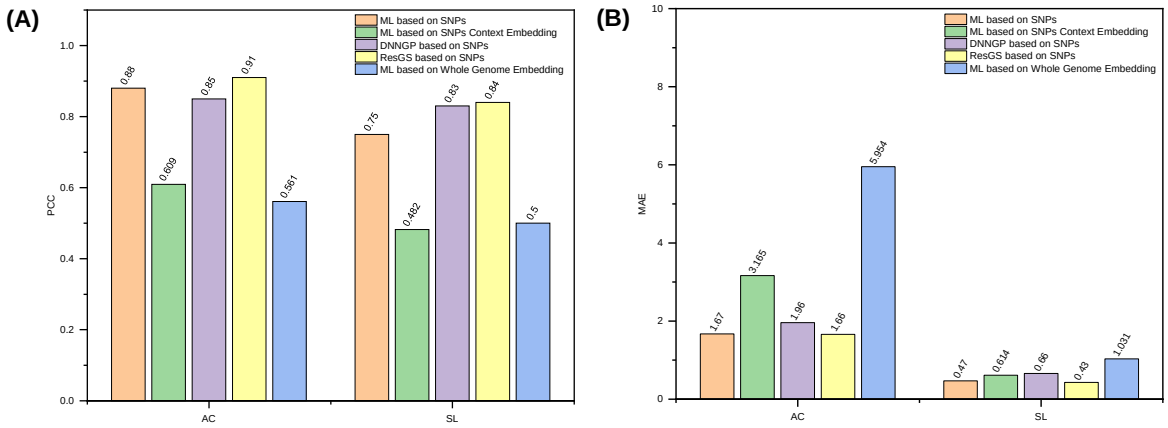


Figure 10 Performance comparison of traditional machine learning algorithms (ML), DNNGP, ResGS (Based on SNP markers), and ML based on SNP context and whole-genome feature embedding for traits in rice395. (A) Pearson Correlation Coefficient (PCC) values for predicting AC and SL traits. (B) Mean Absolute Error (MAE) values for predicting AC and SL traits.

Figure 10 compares the performance of SNP-context feature embedding and whole-genome feature embedding with traditional SNP markers features, DNNGP, and ResGS for predicting AC and SL traits in the rice395 dataset. Contrary to the results in rice413, both SNP-context feature embedding and whole-genome feature embedding underperformed relative to traditional SNPs features, DNNGP, and ResGS in rice395, exhibiting lower PCC values and higher MAE values. Specifically, SNP-context feature embedding achieved 4.8% higher prediction accuracy for the AC trait compared to whole-genome feature embedding, while whole-genome feature embedding showed marginally better performance for the SL trait.

Table5 Heritability of Traits in rice395 Based on Genetic Analysis

Traits for rice395	Heritability
AC	0.040703
SL	0.096478

Further analysis incorporating heritability estimates revealed that the heritability of AC and SL traits in rice395 was significantly lower than in rice413. This discrepancy may arise from differences in the number and effect sizes of genes controlling AC and SL, as well as variations in gene-environment interactions across rice varieties. The ultra-low heritability in rice395 suggests reduced genetic stability across generations, making these traits more susceptible to environmental influences and other confounding factors. Consequently, the predictive power of SNP-context feature embedding and whole-genome feature embedding was limited, resulting in lower accuracy and higher MAE values. This observation aligns with findings from Habier et al. (2011), which indicate that traits with lower heritability exhibit weaker gene-trait correlations, leading to reduced accuracy in estimating marker effects across the genome.

Figure 11 compares the prediction performance of SNP-context feature embedding and whole-genome feature embedding with traditional SNP markers features, DNNGP, and ResGS for three traits (Dpoll, EarDia, and EarHT) in the maize301 dataset. Whole-genome feature embedding achieved the highest prediction accuracy (PCC = 0.806) for the Dpoll trait, accompanied by the lowest mean

absolute error (MAE). This represents improvements of 7.6%, 2.6%, and 2.6% over traditional SNPs features, DNNGP, and ResGS, respectively. For the EarDia trait, whole-genome feature embedding yielded a prediction accuracy of 0.849, surpassing traditional SNPs features, DNNGP, and ResGS by 35.9%, 22.9%, and 19.9%, respectively. Similarly, SNP-context feature embedding outperformed traditional SNPs features, DNNGP, and ResGS for the EarHT trait, with accuracy improvements of 26.2%, 19.2%, and 20.2%, respectively.

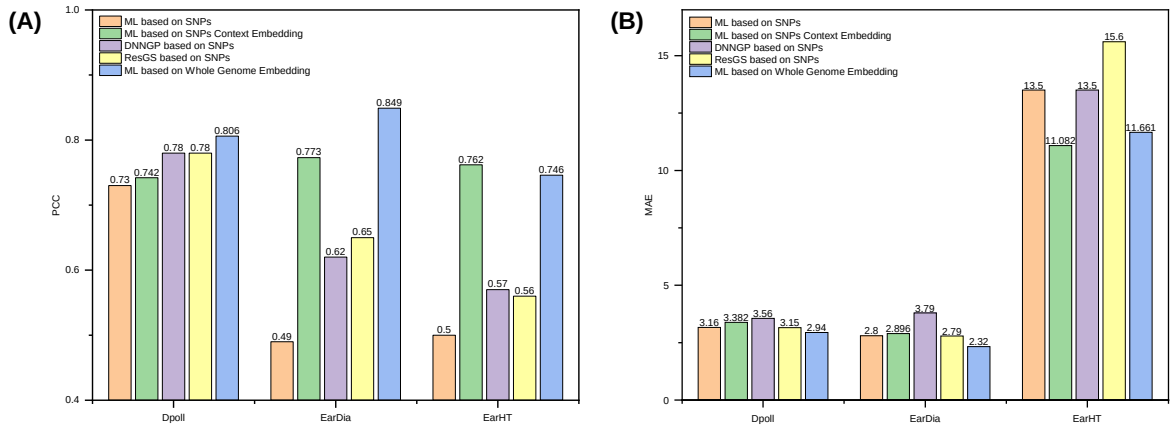


Figure 11 Performance comparison of traditional machine learning algorithms (ML), DNNGP, ResGS (Based on SNP markers), and ML based on SNP context and whole-genome feature embedding for traits in maize301. (A) Pearson Correlation Coefficient (PCC) values for predicting Dpoll, EarDia, and EarHT traits. (B) Mean Absolute Error (MAE) values for predicting Dpoll, EarDia, and EarHT traits.

The significant performance enhancement for the EarDia trait using SNP-context feature embedding (3000bp context length) and further improvement with whole-genome feature embedding suggests the presence of distal regulatory elements (>3000bp from SNPs sites) influencing trait expression. The high heritability of Dpoll and EarDia traits in maize301, combined with the superior performance of whole-genome feature embedding, underscores its effectiveness in capturing genetic features associated with high-heritability traits. This aligns with observations in rice413, where whole-genome feature embedding excelled in predicting high-heritability traits such as SL and AC. Conversely, for medium-heritability traits like EarHT in maize301 and ASV/PC in rice413, SNP-context feature embedding demonstrated comparable or superior performance to whole-genome embedding, highlighting its ability to uncover genetic interactions within localized genomic contexts.

Table6 Heritability of Traits in maize301 Based on Genetic Analysis

Traits for maize301	Heritability
Dpoll	0.805676
EarDia	0.762109
EarHT	0.527156

The fitting effect of typical traits under two embedding modes

To further compare the correlation and distribution ranges between predicted and observed phenotypic values under whole-genome feature embedding and SNP-context feature embedding, we randomly selected one representative trait from each of three crops: EarDia (maize301), AC (rice395), and SL (rice413). The statistical distribution patterns of predicted versus observed values were visualized using boxplots and normal distribution curves. Additionally, regression analysis was performed to derive the

best-fit equations between predicted and observed values, followed by visualization of the linear fitting results.

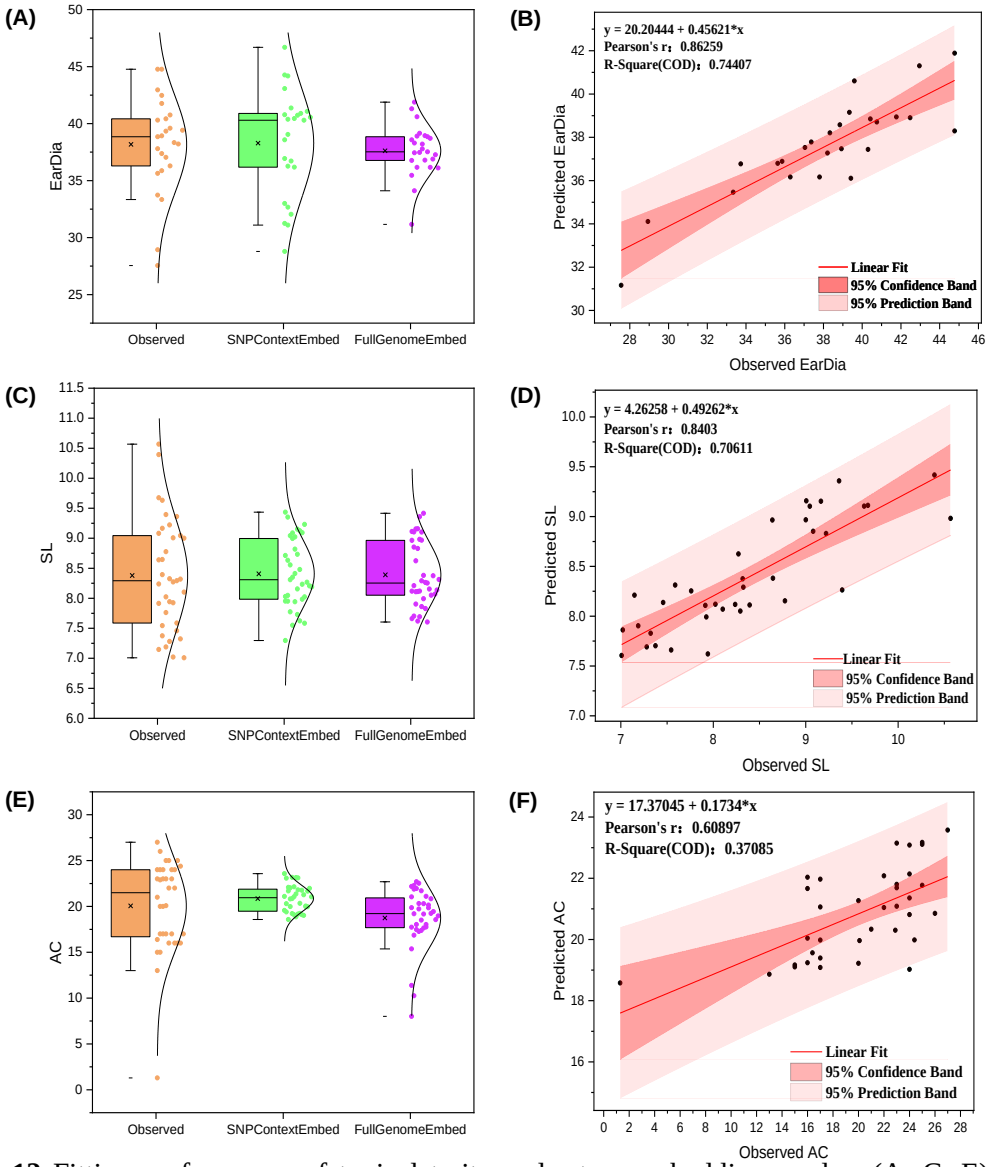


Figure 12 Fitting performance of typical traits under two embedding modes. (A, C, E) Boxplots and statistical distribution curves showing the predicted and observed values of EarDia (maize301), SL (rice413), and AC (rice395) traits under SNP-context feature embedding and whole-genome feature embedding. (B, D, F) Fitting results between predicted and true values for EarDia, SL, and AC traits under optimal embedding conditions(As shown in Table1-3).

The result presents the fitting performance of typical traits (EarDia, SL, and AC) under SNP-context feature embedding and whole-genome feature embedding(Figure 12). For the EarDia trait in maize301, whole-genome feature embedding produced predicted values with a median closer to the true value distribution, achieving the highest average prediction accuracy. The fitting equation exhibited narrow 95% confidence and prediction bands, with most data points closely aligned with the optimal fitting line. The Pearson correlation coefficient (PCC $\approx$ 0.86) and goodness-of-fit ($R^2 \approx$ 0.744) indicate robust model performance, likely due to the method's ability to capture multidimensional genomic features, including structural variations and epigenetic modifications.

For the SL trait in rice413, whole-genome feature embedding also yielded predicted values with a median closer to the true value distribution. The fitting results demonstrated narrow confidence and prediction bands, with data points concentrated around the optimal fitting line (PCC $\approx$ 0.84, $R^2 \approx$ 0.71), reflecting stable prediction performance and strong phenotypic prediction capability.

In contrast, for the AC trait in rice395, SNP-context feature embedding achieved slightly higher accuracy than whole-genome embedding, but the overall prediction accuracy remained low. The wide 95% confidence and prediction bands, along with a low goodness-of-fit ($R^2 \approx 0.37$), indicate high prediction uncertainty. This suggests that low-heritability traits like AC are more susceptible to environmental influences, making it challenging to capture key genetic determinants of trait expression.

Discussion

This study investigated the efficacy of feature embedding, utilizing whole-genome sequence features and SNP-context features, in predicting crop phenotypes with DNABERT-2, comparing its performance to state-of-the-art deep learning models like DNNGP and ResGS which are based on SNPs marker. By employing DNABERT-2, a pre-trained genomic foundation model, we encoded whole-genome sequences and SNP-context into feature vectors and systematically evaluated the impact of key factors, including SNP-context length, dimensionality reduction techniques, and feature dimensions, on phenotypic prediction accuracy. Our findings reveal several critical insights into the genetic architecture of complex traits and the potential of advanced genomic feature embedding methods for improving prediction performance.

Impact of SNP-context and Whole-genome feature embedding

The incorporation of gene sequences surrounding SNPs significantly enhanced phenotype prediction for traits such as PNPP, SNPP, PC, ASV, and AC in the rice413 dataset. The optimal context length varied across traits, with PNPP, SNPP, and AC achieving peak accuracy at 3000bp, while ASV, PC, and SL performed best at 1500bp, 1000bp, and 1000bp, respectively. Similarly, in the maize301 dataset, Dpoll, EarDia, and EarHT exhibited the highest prediction accuracy at context lengths of 3000bp, 3000bp, and 2000bp, respectively. These results underscore the importance of considering the local genomic environment around SNPs, rather than relying solely on SNPs marker, for accurate trait prediction. This aligns with the understanding that genetic regulation often involves interactions between multiple loci and distal regulatory elements, which are better captured by extended genomic contexts.

Whole-genome feature embedding, combined with dimensionality reduction techniques such as PCA and PLS in some case, outperformed traditional SNPs-based methods and SNP-context feature embedding mode in the rice413 dataset, it achieved the highest prediction accuracy across all traits, likely due to its ability to capture non-additive and recessive genetic effects that are often overlooked in traditional methods. This suggests that comprehensive genomic information, beyond individual SNPs, is crucial for unraveling the genetic basis of complex traits. The effectiveness of dimensionality reduction varied by trait and dataset. For instance, PCA with 256 dimensions yielded the highest accuracy for PNPP in rice413, while PLS significantly improved prediction accuracy for AC. In the maize301 dataset, PCA enhanced prediction accuracy for Dpoll and EarHT at dimensions of 16 and 128, respectively. These findings highlight the importance of selecting appropriate dimensionality reduction strategies to optimize model performance.

Trait heritability and prediction accuracy

A clear correlation was observed between trait heritability and prediction accuracy. High-heritability traits, such as AC and SL in rice413, demonstrated improved prediction accuracy with whole-genome feature embedding. This improvement can be attributed to the ability of whole-genome feature embedding to capture comprehensive genetic information, including non-additive and recessive effects, which are often overlooked in traditional SNPs-based methods. Conversely, traits with ultra-low heritability, such as AC and SL in the rice395 dataset, exhibited significantly reduced prediction accuracy, even when using whole-genome sequence data. This suggests that environmental factors and complex gene-environment interactions play a more prominent role in shaping traits with lower heritability, thereby limiting the predictive power of genomic information alone. These results are consistent with previous studies indicating that heritability is a key determinant of genomic prediction accuracy(Habier et al. 2011; Daetwyler et al. 2010).

Comparative performance with existing models

Our whole-genome feature embedding method, when applied to common machine learning algorithms, consistently outperformed existing models such as DNNGP and ResGS in terms of prediction accuracy and mean absolute error (MAE) for the majority of traits in both the rice413 and maize301 datasets. This superior performance can be attributed to the ability of DNABERT-2 to capture long-range genomic dependencies and complex genetic interactions, which are often missed by traditional SNPs-based methods. However, for traits with ultra-low heritability, such as those in the rice395 dataset, the improvement in prediction accuracy was limited, further emphasizing the influence of non-genetic factors on these traits.

Limitations and future directions

While our study demonstrated the advantages of SNP-context and whole-genome feature embedding, it was limited to exploring context lengths within a 3000bp range. We speculate that additional genetic features beyond this range may contribute to trait prediction, particularly for traits influenced by distal regulatory elements. Future research should investigate the impact of more extensive genomic contexts, potentially exceeding 3000bp, to further enhance prediction accuracy. Additionally, the integration of multi-omics data, such as transcriptomics and epigenomics, could provide a more comprehensive understanding of the genetic and environmental factors influencing complex traits.

Data availability

The authors affirm that all data necessary for confirming the conclusions of the article are present within the article/supplementary material. Further inquiries can be directed to the corresponding author.

Supplementary information

Supplementary Material is available at Genetics online.

Acknowledgements

We thank Wen Qiang Zhao for assisting in filtering out the noise in both the genomic data and the phenotypic data, and for providing English language polishing.

Author contributions

HL finished the experiment and wrote the manuscript, TW, CW and WB assisted in collecting phenotypic data, genomic data, and restoring whole-genome sequences. JL and LC played a crucial

role in data collection, participating in the preprocessing of phenotypic and genomic sequences. YC and ZC provided expert guidance in data analysis and offered invaluable feedback on the manuscript. MW jointly supervised the entire research process, ensuring ethical compliance and effectively facilitating communication among authors. Lastly, JW and JH were responsible for proofreading the manuscript. All authors participated in the research and approved the final manuscript.

Funding

This work was supported by the Open Fund of the Key Laboratory of Agricultural Big Data, Ministry of Agriculture and Rural Affairs.

Conflict of interest

The author(s) declare no conflict of interest.

Literature cited

Saraiva R, Ferreira Q, Rodrigues GC, Oliveira M. 2023. Nanofertilizer use for adaptation and mitigation of the agriculture/climate change dichotomy effects. 11(6):129.

Yuan X, Li S, Chen J, Yu H, Yang T, Wang C, Huang S, Chen H, Ao X. 2024. Impacts of global climate change on agricultural production: A comprehensive review. 14(7):1360.

Bita CE, Gerats T. 2013. Plant tolerance to high temperature in a changing environment: Scientific fundamentals and production of heat stress-tolerant crops. Front Plant Sci. 4:273.

He T, Li C. 2020. Harness the power of genomic selection and the potential of germplasm in crop breeding for global food security in the era with rapid climate change. The Crop Journal. 8(5):688-700.

FAO. 2009. High level expert forum-how to feed the world in 2050. Economic and Social Development.

Bonea DJAotUoC-AMCS. 2022. Speed breeding and its importance for the improvement of agricultural crops. Annals of the University of Craiova-Agriculture Montanology Cadastre Series. 52(1):59-66.

Martínez-Gómez PJIJoMS. 2019. Editorial for special issue “plant genetics and molecular breeding”. MDPI. p. 2659.

Singh N, Langyan S, Rai V. 2023. Editorial: Advances and applications of cost-effective, high-throughput genotyping technologies for sustainable agriculture. Front Plant Sci. 14:1335417.

Pujolar J, Jacobsen M, Frydenberg J, Als TD, Larsen P, Maes G, Zane L, Jian J, Cheng L, Hansen MMJMer. 2013. A resource of genome ‐ wide single ‐ nucleotide polymorphisms generated by rad tag sequencing in the critically endangered european eel. Molecular ecology resources. 13(4):706-714.

Sun X, Liu D, Zhang X, Li W, Liu H, Hong W, Jiang C, Guan N, Ma C, Zeng HJPo. 2013. Slaf-seq: An efficient method of large-scale de novo snp discovery and genotyping using high-throughput sequencing. PloS one. 8(3):e58700.

Meuwissen THE, Hayes BJ, Goddard MJG. 2001. Prediction of total genetic value using whole-genome dense marker maps. Genetics. 157(4):1819-1829.

Li C, Li J, Wang H, Zhang R, An X, Yuan C, Guo T, Yue YJA. 2023. Genomic selection for live weight in the 14th month in alpine merino sheep combining gwas information. Animals. 13(22):3516.

Ibtisham F, Zhang L, Xiao M, An L, Ramzan MB, Nawab A, Zhao Y, Li G, Xu YJTTJoVM. 2017. Genomic selection and its application in animal breeding. The Thai Journal of Veterinary Medicine. 47(3):301-310.

Grenier C, Cao T-V, Ospina Y, Quintero C, Châtel MH, Tohme J, Courtois B, Ahmadi NJPo. 2015. Accuracy of genomic selection in a rice synthetic population developed for recurrent selection breeding. PloS one. 10(8):e0136594.

Budhlakoti N, Kushwaha AK, Rai A, Chaturvedi K, Kumar A, Pradhan AK, Kumar U, Kumar RR, Juliana P, Mishra DJFiG. 2022. Genomic selection: A tool for accelerating the efficiency of molecular breeding for development of climate-resilient crops. Frontiers in Genetics. 13:832153.

Sandhu KS, Mihalyov PD, Lewien MJ, Pumphrey MO, Carter AHJFips. 2021. Combining genomic and phenomic information for predicting grain protein content and grain yield in spring wheat. Frontiers in plant science. 12:613300.

Crossa J, Pérez-Rodríguez P, Cuevas J, Montesinos-López O, Jarquín D, De Los Campos G, Burgueño J, González-Camacho JM, Pérez-Elizalde S, Beyene YJTips. 2017. Genomic selection in plant breeding: Methods, models, and perspectives. Trends in plant science. 22(11):961-975.

Zhao H, Lin Z, Khansefid M, Tibbits JF, Hayden MJ. 2023. Genomic prediction and selection response for grain yield in safflower. Frontiers in genetics. 14:1129433.

Bohra A, Chand Jha U, Godwin ID, Kumar Varshney R. 2020. Genomic interventions for sustainable agriculture. Plant biotechnology journal. 18(12):2388-2405.

Schmidt M, Kollers S, Maasberg-Prelle A, Großer J, Schinkel B, Tomerius A, Graner A, Korzun V. 2016. Prediction of malting quality traits in barley based on whole-genome marker data to assess the potential of genomic selection. Theoretical and Applied Genetics. 129(2):203-213.

Johnston I, Carvalho LE. 2014. A bayesian hierarchical gene model on latent genotypes for whole-genome association studies. BMC Proceedings. 8(1):S45.

Jeong S, Kim J-Y, Kim N. 2020. Gmstool: Gwas-based marker selection tool for genomic prediction from genomic data. Scientific Reports. 10(1):19653.

VanRaden PMJJods. 2008. Efficient methods to compute genomic predictions. Journal of dairy science. 91(11):4414-4423.

Resende MFR, Muñoz PR, Resende MDVd, Resende MDVd, Garrick DJ, Fernando RL, Davis JM, Jokela EJ, Martin TA, Peter GF et al. 2012. Accuracy of genomic selection methods in a standard data set of loblolly pine (pinus taeda l.). Genetics. 190:1503 - 1510.

Endelman JBJTPG. 2011. Ridge regression and other kernels for genomic selection with r package rrblup. The Plant Genome. 4:250 - 255.

Sánchez-Mayor M, Riggio V, Navarro P, Gutiérrez-Gil B, Haley CS, De la Fuente LF, Arranz J-J, Pong-Wong R. 2022. Effect of genotyping strategies on the sustained benefit of single-step genomic blup over multiple generations. Genetics Selection Evolution. 54(1):23.

Habier D, Fernando RL, Kizilkaya K, Garrick DJ. 2011. Extension of the bayesian alphabet for genomic selection. BMC Bioinformatics. 12:186.

Zhu B, Zhu M, Jiang J, Niu H, Wang Y, Wu Y, Xu L, Chen Y, Zhang L, Gao X et al. 2016. The impact of variable degrees of freedom and scale parameters in bayesian methods for genomic prediction in chinese simmental beef cattle. PLOS ONE. 11(5):e0154118.

Cleveland MA, Forni S, Deeb N, Maltecca C. 2010. Genomic breeding value prediction using three bayesian methods and application to reduced density marker panels. BMC Proc. 4(Suppl 1 Proceedings of the 13th European workshop on QTL map):S6.

Grinberg NF, Orhobor OI, King RD. 2020. An evaluation of machine-learning for predicting phenotype: Studies in yeast, rice, and wheat. Machine Learning. 109(2):251-277.

Cosgun E, Limdi NA, Duarte CW. 2011. High-dimensional pharmacogenetic prediction of a continuous trait using machine learning techniques with application to warfarin dose prediction in african americans. Bioinformatics. 27(10):1384-1389.

Xie Z, Xu X, Li L, Wu C, Ma Y, He J, Wei S, Wang J, Feng X. 2024. Residual networks without pooling layers improve the accuracy of genomic predictions. Theoretical and Applied Genetics. 137(6):138.

. A brief review on deep learning applications in genomic studies. Frontiers in Systems Biology; 2022.

Ma W, Qiu Z, Song J, Li J, Cheng Q, Zhai J, Ma C. 2018. A deep convolutional neural network approach for predicting phenotypes from genotypes. Planta. 248(5):1307-1318.

Wang K, Abid MA, Rasheed A, Crossa J, Hearne SJ, Li H. 2022. Dnngp, a deep neural network-based method for genomic prediction using multi-omics data in plants. Molecular Plant. 16(1):279-293.

Gao P, Zhao H, Luo Z, Lin Y, Feng W, Li Y, Kong F, Li X, Fang C, Wang X. 2023. Soydngp: A web-accessible deep learning framework for genomic prediction in soybean breeding. Briefings in bioinformatics. 24(6).

Mérot C, Oomen RA, Tigano A, Wellenreuther M. 2020. A roadmap for understanding the evolutionary significance of structural genomic variation. Trends in Ecology & Evolution. 35(7):561-572.

Kotlarz K, Mielczarek M, Biecek P, Guldbrandtsen B, Szyda J. 2024. Exploring the impact of sequence context on errors in snp genotype calling with whole genome sequencing data using ai-based autoencoder approach. NAR genomics and bioinformatics. 6(3):lqae131.

Devlin J, Chang M-W, Lee K, Toutanova KJae-p. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. [accessed October 01, 2018]:arXiv:1810.04805. https://ui.adsabs.harvard.edu/abs/2018arXiv181004805D. doi:10.48550/arXiv.1810.04805.

Ji Y, Zhou Z, Liu H, Davuluri RV. 2021. Dnabert: Pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. Bioinformatics. 37(15):2112-2120.

Zhou Z, Ji Y, Li W, Dutta P, Davuluri R, Liu HJae-p. 2023. Dnabert-2: Efficient foundation model and benchmark for multi-species genome. [accessed June 01, 2023]:arXiv:2306.15006. https://ui.adsabs.harvard.edu/abs/2023arXiv230615006Z. doi:10.48550/arXiv.2306.15006.

Nguyen E, Poli M, Faizi M, Thomas A, Birch-Sykes C, Wornow M, Patel A, Rabideau C, Massaroli S, Bengio Y et al. 2023. Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. ArXiv.

Zhao K, Tung C-W, Eizenga GC, Wright MH, Ali ML, Price AH, Norton GJ, Islam MR, Reynolds A, Mezey J et al. 2011. Whole-genome association mapping reveals a rich genetic architecture of complex traits in oryza sativa. Nature Communications. 2(1):467.

Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. 2007. Tassel: Software for association mapping of complex traits in diverse samples. Bioinformatics. 23(19):2633-2635.

Sennrich R, Haddow B, Birch AJA. 2015. Neural machine translation of rare words with subword units. ArXiv. abs/1508.07909.

Press O, Smith NA, Lewis MJae-p. 2021. Train short, test long: Attention with linear biases enables input length extrapolation. [accessed August 01, 2021]:arXiv:2108.12409. https://ui.adsabs.harvard.edu/abs/2021arXiv210812409P. doi:10.48550/arXiv.2108.12409.

Dao T, Fu DY, Ermon S, Rudra A, Ré CJae-p. 2022. Flashattention: Fast and memory-efficient exact attention with io-awareness. [accessed May 01, 2022]:arXiv:2205.14135. https://ui.adsabs.harvard.edu/abs/2022arXiv220514135D. doi:10.48550/arXiv.2205.14135.

Galinha CF, Crespo JGJM. 2022. Development and implementation of mbr monitoring: Use of 2d fluorescence spectroscopy. 12.

Nguyen DV, Rocke DM. 2002. Tumor classification by partial least squares using microarray gene expression data. Bioinformatics (Oxford, England). 18(1):39-50.

Broadhurst DI, Goodacre R, Jones A, Rowland JJ, Kell DBJACA. 1997. Genetic algorithms as a method for variable selection in multiple linear regression and partial least squares regression, with applications to pyrolysis mass spectrometry. 348:71-86.

Wold S, Albano C, Dunn M, Esbensen K, Hellberg S, Johansson E, Sjöström M. 1983. Pattern regression finding and using regularities in multivariate data. Analysis Applied Science Publication.

Azodi CB, Bolger E, McCarren A, Roantree M, de los Campos G, Shiu S-H. 2019. Benchmarking parametric and machine learning models for genomic prediction of complex traits. G3 Genes|Genomes|Genetics. 9(11):3691-3702.

Montesinos-López A, Montesinos-López OA, Gianola D, Crossa J, Hernández-Suárez CM. 2018. Multi-environment genomic prediction of plant traits using deep learners with dense architecture. G3 (Bethesda, Md). 8(12):3813-3828.

Norman A, Taylor J, Edwards J, Kuchel H. 2018. Optimising genomic selection in wheat: Effect of marker density, population size and population structure on prediction accuracy. G3 (Bethesda, Md). 8(9):2889-2899.

Spindel J, Begum H, Akdemir D, Virk P, Collard B, Redoña E, Atlin G, Jannink JL, McCouch SR. 2015. Genomic selection and association mapping in rice (oryza sativa): Effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite, tropical rice breeding lines. PLoS genetics. 11(2):e1004982.

Riedelsheimer C, Technow F, Melchinger AE. 2012. Comparison of whole-genome prediction models for traits with contrasting genetic architecture in a diversity panel of maize inbred lines. BMC Genomics. 13(1):452.

Ogutu JO, Piepho HP, Schulz-Streeck T. 2011. A comparison of random forests, boosting and support vector machines for genomic selection. BMC Proc. 5 Suppl 3(Suppl 3):S11.

Bolormaa S, Pryce JE, Kemper K, Savin K, Hayes BJ, Barendse W, Zhang Y, Reich CM, Mason BA, Bunch RJ et al. 2013. Accuracy of prediction of genomic breeding values for residual feed intake and carcass and meat quality traits in bos taurus, bos indicus, and composite beef cattle1. Journal of Animal Science. 91(7):3088-3104.

Daetwyler HD, Hickey JM, Henshall JM, Dominik S, Gredler B, Werf JHVD, Hayes BJJAPS. 2010. Accuracy of estimated genomic breeding values for wool and meat traits in a multi-breed sheep population. 50:1004-1010.