Theoretical and Applied Genetics (2025) 138:103
https://doi.org/10.1007/s00122-025-04894-z

ORIGINAL ARTICLE

EBMGP: a deep learning model for genomic prediction based on Elastic Net feature selection and bidirectional encoder representations from transformer's embedding and multi-head attention pooling

Lu Ji[1,2] · Wei Hou[3] · Heng Zhou[1] · Liwen Xiong[4] · Chunhai Liu[1] · Zheming Yuan[1] · Lanzhi Li[1]

Received: 13 October 2024 / Accepted: 27 March 2025 / Published online: 19 April 2025

Abstract

Enhancing early selection through genomic estimated breeding values is pivotal for reducing generation intervals and accelerating breeding programs. Recently, deep learning (DL) approaches have gained prominence in genomic prediction (GP). Here, we introduce a novel DL framework for GP based on Elastic Net feature selection and bidirectional encoder representations from transformer's embedding and multi-head attention pooling (EBMGP). EBMGP applies Elastic Net for the selection of features, thereby diminishing the computational burden and bolstering the predictive accuracy. In EBMGP, SNPs are treated as "words," and groups of adjacent SNPs with similar LD levels are considered "sentences." By applying bidirectional encoder representations from transformers embeddings, this method models SNPs in a manner analogous to human language, capturing complex genetic interactions at both the "word" and "sentence" scales. This flexible representation seamlessly integrates into any DL network and demonstrates a marked improvement in predictive performance for EBMGP and SoyDNGP compared to the widely used one-hot representation. We propose multi-head attention pooling, which can adaptively assign weights to features while learning features from multiple subspaces through multi-heads for a high level of semantic understanding. In a comprehensive comparative analysis across four diverse plant and animal datasets, EBMGP outperformed competing models in 13 out of 16 tasks, achieving accuracy gains ranging from 0.74 to 9.55% over the second-best model. These results underscore EBMGP's robustness in genomic prediction and highlight its potential for deep learning applications in life sciences.

Introduction

Genomic prediction (GP), initially proposed by Meuwissen et al., utilizes genome-wide genotype markers to predict the breeding values of unobserved populations, thereby facilitating the rapid identification of superior genotypes and accelerating the breeding process (Li et al. 2023; Meuwissen et al. 2001). The significant reduction in genotyping costs has propelled the widespread adoption of GP over the past decade, yielding substantial genetic gains across various plant and animal breeding programs. Extensive GP research has focused on optimizing marker density, training population size, family relationships, and the selection of GP models. Genomic best linear unbiased prediction (GBLUP) is a prevalent GP model that relies on a marker-based relationship matrix for predictions (Clark and van der Werf 2013). Conversely, Bayesian models incorporate prior distributions, necessitating distinct models for different traits (Pérez

Communicated by Mikko J. Sillanpää.

✉ Zheming Yuan
zhmyuan@hunau.edu.cn

✉ Lanzhi Li
lancy0829@163.com

Lu Ji
kobejilu@163.com

[1] Hunan Engineering and Technology Research Center for Agricultural Big Data Analysis and Decision-Making, Hunan Agricultural University, Changsha 410128, China

[2] Basic Biology Laboratory, Hunan First Normal University, Changsha 410205, China

[3] College of Bioscience and Biotechnology, Hunan Agricultural University, Changsha 410128, China

[4] College of Life Sciences, University of Chinese Academy of Sciences, Beijing, Beijing 100049, China

and de los Campos 2014). Bayes B, for instance, utilizes a Gaussian mixture, assuming that not all markers contribute to the genetic variance (Pérez and de los Campos 2014). Bayesian lasso (BL) applies a double exponential prior for continuous shrinkage and variable selection, employing a long-tailed Student-t distribution for marker effects (Li et al. 2010). However, the precise influence of individual SNPs remains elusive and may not strictly confirm to any specific distribution. Furthermore, these parametric models often fail to capture the sophisticated interplay between SNPs, particularly pertinent in complex traits arising from epistasis (Johnson et al. 2023; Webber 2017).

Deep learning, a subfield of machine learning, harnesses complex neural networks with multiple nonlinear transformations across layers, making it well-suited to address the challenges of GP. Deep learning genome-wide association study (DLGWAS) is a dual-stream deep learning model (Liu et al. 2019). It has demonstrated superior predict performance over traditional statistical methods on both simulated and soybean datasets. The DNNGP model integrates three convolutional neural networks (CNN) layers, a batch normalization (BN) layer to prevent overfitting, and two dropout layers (Wang et al. 2023). It efficiently processes complex omics data, surpassing commonly used GP methods such as GBLUP, LightGBM, SVR, DeepGS, and DLGWAS (Wang et al. 2023). SoyDNGP is a deep network comprising 12 convolutional blocks and a fully connected layer (Gao et al. 2023). It incorporates a coordinate attention (CA) mechanism after the first and final convolutional layers to enhance spatial information extraction (Gao et al. 2023). SoyDNGP outperforms methods like AdaBoost, Decision Tree, Naive Bayes, and Random Forest in classification tasks and exceeds both DeepGS and DNNGP in regression tasks, highlighting its versatility and strength in GP (Gao et al. 2023).

Despite deep learning's notable advancements in GP, there is ample scope for further exploration. Firstly, the $p >> n$ problem, where the number of features (p) vastly exceeds the number of individuals (n), has become one of the limiting factors in the development of GP deep learning models. Feature selection provides an efficient solution to this issue. Jubair et al. employed mutual information feature selection, a filter-based method, where genetic markers and phenotypes are used as inputs to generate a mutual information score as the output (Jubair et al. 2021). Secondly, most existing GP deep learning models represent SNPs using one-hot, which treats each SNP in isolation and overlooks their interrelationships. Furthermore, this poses significant challenges for the model to discern functional (semantic) differences among SNPs that share the same genotype. Recent studies have concentrated on developing more efficient biological sequence representations, such as word embedding. Le et al. applied word embedding to represent DNA sequences, incorporating sub-word information (Le et al. 2019). This approach led to remarkable performance in enhancer identification (Le et al. 2019). Bidirectional encoder representations from transformers (BERT) embedding has been utilized to convert RNA sequences into feature descriptors, enabling the model to capture hidden information more effectively and enhance m7G site prediction (Zhang et al. 2021a). This indicates the promising potential of using contextual word embedding to represent biological sequences in conjunction with deep learning networks. In these studies, when BERT embeddings are used to represent biological sequences, semantic segmentation within the sequence is not conducted, with the entire sequence being treated as a singular "sentence," an approach that lacks optimality. Previous studies suggest that LD blocks, which are likely inherited from shared ancestors, provide a richer source of information compared to individual SNPs in genomic selection (Karimi et al. 2018). If LD is used for semantic segmentation of SNP sequences, it may help the model better understand the genetic structure of the data and make more accurate predictions. Thirdly, some GP models employing traditional max pooling and average pooling methods can result in information loss and fail to dynamically optimize features (Abdollahi-Arpanahi et al. 2020; Azodi et al. 2019; Zingaretti et al. 2020). Various pooling strategies have been introduced to preserve vital information within activation maps. Soft pooling uses a softmax-weighted sum of activations, and local importance-based pooling (LIP) dynamically enhances key features during down sampling by adapting importance weights based on input characteristics (Gao et al. 2019; Stergiou et al. 2021). Inspired by the self-attention mechanism in Transformers, multi-head self-attention pooling was introduced to comprehensively account for the contribution of each feature to the final outcome (Chen et al. 2024; Yan et al. 2022).

To overcome the challenges mentioned above, we propose a novel deep learning model (EBMGP) for GP. Our approach incorporates the following innovations: (1) Utilizing elastic net (EN) to select key SNPs and comprehensively analyzed how varying feature subset sizes influence model accuracy. (2) Through BERT embeddings, we conceptualize SNPs as analogous to human natural language. This allows for the dynamic detection of interactions at both the SNP and linkage disequilibrium (LD) block levels. (3) Inspired by Transformer's multi-head self-attention mechanism, we introduce multi-head attention pooling (MAP), which assigns adaptive weights to features and employs multiple heads to capture diverse subspace features. Comparison with seven widely adopted GP models on four plant and animal datasets confirms EBMGP's robustness in genomic prediction and underscores its promise for deep learning applications in life sciences.

Materials and methods

Dataset

This study utilized four datasets with varying numbers of SNPs, as well as different genomic architectures, to assess the accuracy of the model's predictions and its ability to generalize. The rice datasets included 413 diverse inbred rice accessions (Oryza sativa) from 82 countries (Zhao et al. 2011). Each plant was genotyped using a 44-K chip (44,100 SNPs). The SNPs with minor allele frequency (MAF) lower than 5% were deleted. After quality control, 36,901 SNPs were retained. Five traits were evaluated: seed width (SW), flag leaf width (FLW), plant height (PH), amylose content (AC), and seed number per panicle (SNPP). The genotype of soybean was procured from SoyBase, including 13,974 accessions and 42,509 SNPs (Grant et al. 2009). The Beagle 5.4 program (version 22Jul22.46e) was used to phase the SNPs and fill in the missing data (Ayres et al. 2011). The sorghum dataset was derived from sorghum lines provided by the US National Plant Germplasm System and grown in Urbana, IL. It consists of 451 lines and 58,961 SNPs. Three traits were analyzed in this study: plant height (HT), grain moisture (MO), and yield (YLD) (Azodi et al. 2019). The SNPs with a MAF below 5% were excluded, leaving a total of 36,468 SNPs. The phenotypic data for soybean were obtained from the GRIN-Global database (https://npGPweb.ars-grin.gov/gringlobal/search). We focused on 5 quantitative traits: protein content (protein), stearic acid content (Stearic), the maturity date (R8), hundred-seed weight (SdWgt), and yield. The Holstein bull dataset contains 1,508 samples. According to the paper, genotype imputation was performed using BEAGLE 3.3.1, resulting in 52,886 SNPs on 29 autosomes. SNPs with a genotype call rate below 90%, a minor allele frequency under 0.01, or a Hardy–Weinberg equilibrium p-value less than 1e-6 were excluded (Yin et al. 2019). Ultimately, 44,074 SNPs were retained for the study. Three traits were used in this study: motility of sperm (MS), the number of motile sperm (NMSP), and volume (VE). All phenotypes were standardized, and missing values were imputed using the mean.

Overview of EBMGP

A schematic diagram of EBMGP is shown in Fig. 1a. The EBMGP's architecture comprises four key components: (1) feature selection, (2) BERT embedding layer, (3) convolution layer, and (4) linear layer.

To decrease noise and cut down on computation costs, Elastic Net (EN) is applied to preselect the top n features before training. The EN combines L1 and L2 penalties, allowing for adjustment of their proportions through parameter tuning (Zou and Hastie 2005). The parameters were set as follows: alpha at 0.5 and max_iter at 1000. To meet the desired number of nonzero coefficients (1,000, 3,000, 5,000, 7,000, and 9,000), the l1_ratio was incrementally adjusted in steps of 0.0001. To avoid artificially inflating prediction accuracies, feature selection was conducted exclusively on the training set.

After feature selection, we adopt BERT embedding to represent SNPs. Each SNP is represented by two letters. The first letter corresponds to the genotype of the SNP, where the major allele is designated by "H," the heterozygous state is denoted by "M," and the minor allele is signified by "L." The second letter is dedicated to representing the linkage disequilibrium (LD) coefficient R^2 between adjacent SNPs. The calculation of LD is as follows:

$$LD = \frac{(\mathrm{Conv}(i,j))^2}{\mathrm{Var}(i) \cdot \mathrm{Var}(j)}$$

where i represents the ith SNP, j represents the jth SNP, $\mathrm{Conv}(i,j)$ is the covariance between the two loci, $\mathrm{Var}(i)$ and $\mathrm{Var}(j)$ are the variances of ith and jth SNP, respectively. Based on the literature, a stringent LD threshold of 0.8 was applied to aid the model in identifying tightly associated SNPs (Carlson et al. 2004; Kemppainen et al. 2015). When the LD value is 0.8 or higher, the second letter is labeled as "Y"; conversely, it is labeled as "J" when the LD value is below 0.8. Therefore, the model can distinguish those adjacent SNPs that are closely related, that is, LD blocks. For the final SNP on each chromosome, the second letter is consistently represented by "N." The LD calculations were performed using PLINK version 1.9 (Chang et al. 2015). BERT embedding comprises three components: token embedding, segment embedding, and position embedding (Fig. 1b) (Devlin et al. 2019). Token embeddings is the basic embedding layer in BERT. Its main role is to convert the two letters representing SNP into vector representations with fixed dimensions. In natural language processing, segment embedding is added to the input representation to distinguish two sentences. This study uses the letter represents LD between adjacent SNPs as input for segment embedding. If the LD of adjacent SNPs is entirely "Y," they form part of a high-LD "sentence." Conversely, if all are labeled "J," they belong to a low-LD "sentence". Position embedding is used for the model to learn the relative positional information between SNPs. To explore the impact of different SNP representation methods, we replaced the BERT embedding layer in EBMGP with a batch normalization layer, a 1D convolutional layer, and a dropout layer when encoding SNPs using the one-hot.

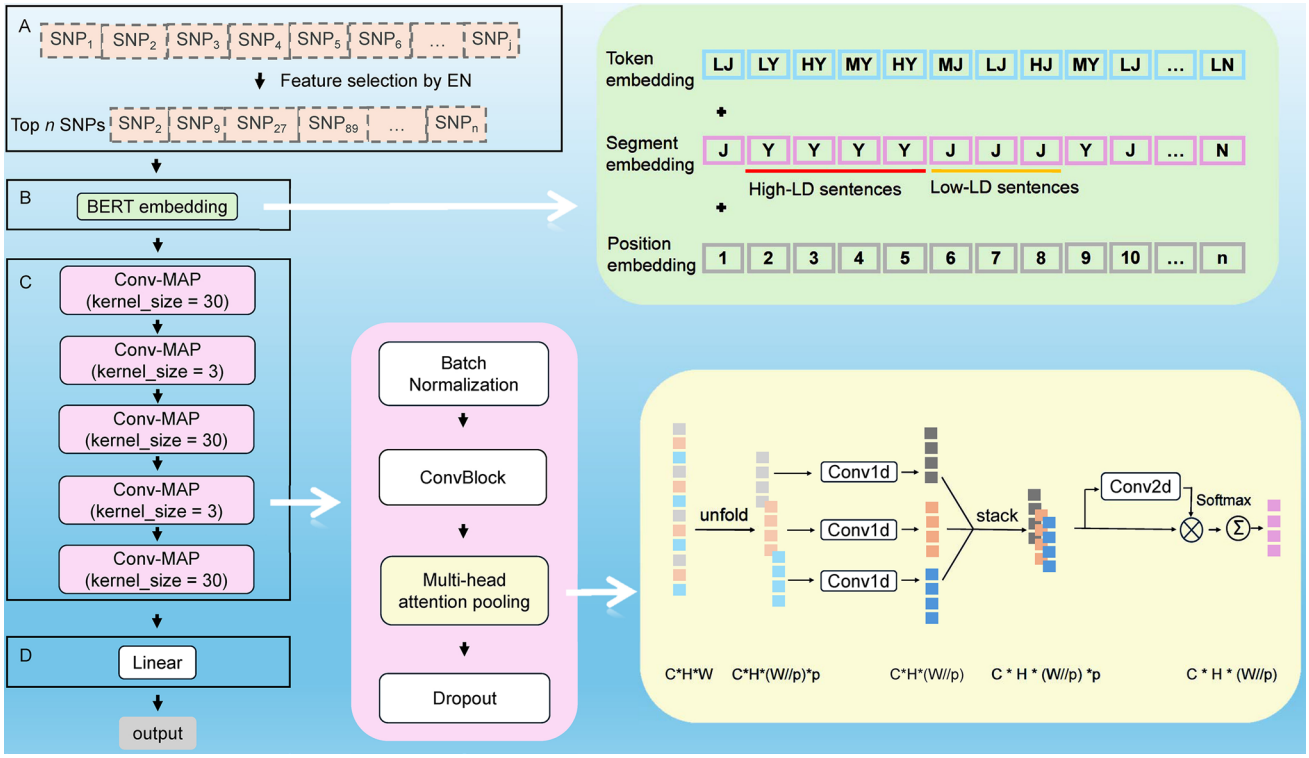


Fig. 1 Overview of EBMGP. **A** Feature selection, j: the number of raw SNP, n: the number of selected SNPs. **B** BERT embedding. **C** convolution layer. **D** linear layer. The green square, the three components of BERT embedding: (1) Token embedding: converts each SNP or token into a numerical vector; (2) Segment embeddings: distinguishes between different sentences; (3) Position embedding: adds positional information, ensuring the model understands the order of SNPs or tokens. H: major alleles, M: heterozygous alleles, L: minor alleles, J: SNP's LD with next SNP lower than 0.8, Y: SNP's LD with next SNP equal or greater than 0.8, N: the last SNP's LD of each chromosome, n: the number of SNPs. The pink square, details of Conv-MAP module. The yellow square, details of MAP, C: batch size of features, H: input channel of features, W: length of features, p: pool_size

The convolution layer contains 5 Conv-MAP modules. Conv-MAP module incorporates a batch normalization, a ConvBlock, a MAP, and a Dropout. The Three of the five Conv-MAP use ConvBlock with larger kernels (30) and two use smaller kernels (3), strategically cross-stacked to effectively capture both fine-grained local variations and broader high-level conceptual patterns (Mishkin et al. 2017). ConvBlock incorporate a batch normalization, an activation (GELU) and a 1D convolution. We propose a novel pooling strategy: MAP, designed to address the need for a high level of semantic understanding (Fig. 1). We first use the unfold function to split the features into multiple subfeatures. Then each subfeatures pass through 1D convolutions to obtain multi-spatial features. All multi-spatial features were stacked to capture a broader range of potential semantic associations. The last step is performing a weighted sum on these subfeatures by softmax.

The linear layer contains a linear module.

The detailed parameter settings for the model can be found in the accompanying GitHub repository (https://github.com/luxixi2021/EBMGP).

Comparative analysis of different SNP representation strategies

We compared the effects of two SNP representations, BERT embeddings and one-hot SNP, on BMADNPG and SoyDNGP by fivefold cross-validation. The BERT embedding representation method is described in Sect. "Overview of EBMGP" part. When using one-hot, the BERT embedding layer in EBMGP was substituted with a ConvBlock with a kernel size of 1 and a stride of 1. When employing BERT embeddings, we systematically evaluate the influence of varying LD threshold settings (0.2, 0.4, 0.6, and 0.8) on EBMGP's performance. In addition, we compared these results with omitting LD information and retaining only the genotype information. To substantiate the efficacy of our SNP representation method, we retrofitted the SoyDNGP model by substituting its initial convolutional layer with BERT embedding (LD threshold = 0.8). In one-hot representation, the major allele was encoded as (0, 0, 0), the heterozygous state as (0, 1, 0), and the minor allele as (0, 0, 1).

Comparison of pooling strategies

We substitute multi-head attention pooling with four alternative pooling modules, including max pooling, average pooling, soft pooling, and local importance pooling (LIP), to assess the performance of different strategies by fivefold cross-validation. Max pooling and average pooling are staples in CNNs, valued for their simplicity and lack of required parameter tuning. Max pooling identifies the maximum value within a pooling region, whereas average pooling segments the input feature map into rectangles for mean calculation (Nirthika et al. 2022). Soft pooling, a kernel-based approach, uses a softmax-weighted sum of activations, with activation gradients contingent on their weights (https://github.com/alexandrosstergiou/SoftPool) (Stergiou et al. 2021). This technique avoids additional parameters and computational complexity. Local importance pooling (LIP) develops adaptive, discriminative importance maps for efficient feature aggregation during downsampling (https://github.com/sebgao/LIP) (Gao et al. 2019).

Training and evaluation

The EBMGP model was developed using PyTorch version 11.6. Training was conducted with a batch size of 32, a learning rate of 0.0005, and a dropout rate of 0.3. For the rice, sorghum, and Holstein bull datasets, the model underwent 100 training epochs, while the soybean dataset was trained for 30 epochs. We utilized the AdamW optimizer with a weight decay of 0.00001 and the CosineAnnealingLR scheduler with T_max set to the number of training steps to update the model's weights and modulate the learning rate, respectively. To evaluate the model's predictive performance, we first fine-tuned the hyperparameters using rice SW, bulls MS, and soybean protein as validation datasets. The remaining 10 prediction tasks in rice, soybean, and bulls were used as independent test datasets. Additionally, we introduced a sorghum dataset as a fully independent test dataset to further evaluate the model's generalizability. In this study, fivefold cross-validation was used to evaluate the prediction performance. The Pearson correlation coefficient (R) and mean squared error (MSE) were applied as the evaluation criterion.

Comparison with other models

To establish the efficacy of EBMGP, it was compared against seven cutting-edge models, including GBLUP, RKHS, Bayes B, Bayesian LASSO, DLGWAS, SoyDNGP, and DNNGP. GBLUP was implemented using the "rrBLUP" package in R (Endelman 2011). The reproducing kernel Hilbert space (RKHS) was executed through the "BGLR" R package, a semi-parametric method employing the Gaussian kernel function (Friedman et al. 2010). The Bayes B model was also implemented using the "BGLR" R package, relying on a Monte Carlo–Markov chain (MCMC) strategy with 12,000 iterations and a burn-in period of 2,000 (Endelman 2011). Additionally, the Bayesian Lasso (BL) was implemented using the "glment" R package with the same MCMC specifications (Endelman 2011). These models offer precise predictions of breeding values, which are a result of additive effects. DLGWAS and SoyDNGP were also implemented in PyTorch 11.6 according to the code provided in the paper (Gao et al. 2023; Liu et al. 2019). While the DNNGP model was downloaded from https://github.com/AIBreeding/DNNGP/releases/download/v1.0.0/DNNGP-v1.0.0.zip) (Wang et al. 2023). All models were evaluated using fivefold cross-validation.

Results

Feature selection's influence on EBMGP

In this study, we employed the embedded feature selection method, Elastic Net (EN), to select the most critical features. As demonstrated in Table 1, we compared the average prediction accuracy of EBMGP from fivefold cross-validation using various SNP subsets. For the rice dataset, EBMGP demonstrated optimal predictive performance for SW, FLW, and SNPP when employing a preselected subset of 5,000 SNPs, whereas the highest accuracy for AC was achieved with a subset of 1,000 SNPs. This approach yielded accuracy improvements of 2.28%, 0.01%, 10.52%, and 6.5%, respectively, in comparison to using the complete SNP set. However, prediction accuracy for PH declined regardless of the number of preselected SNPs utilized. In the sorghum dataset, EBMGP attained the highest predictive accuracy for MO and HT using a subset of 5,000 SNPs, while the optimal prediction for YLD was achieved with 9,000 SNPs. This subset-based strategy enhanced prediction accuracy by 1.89%, 5.16%, and 1.82%, respectively, relative to employing the full SNP dataset. For the soybean dataset, the most accurate predictions were obtained with a subset of 3,000 SNPs, resulting in accuracy improvements of 8.10%, 10.19%, 4.14%, 1.95%, and 5.32% for protein, stearic, R8, SdWgt, and yield, respectively. In the bull dataset, EBMGP achieved superior predictive accuracy for NMSP and VE using a subset of 5,000 SNPs, with accuracy enhancements of 37.82%, 18.09%, and 12.56%, respectively, over the predictions made with the complete SNP set.

These findings suggest that datasets with a smaller sample-to-feature ratio require a greater number of features to retain sufficient information for accurate predictions. In contrast, for datasets with a larger sample-to-feature ratio, selecting fewer features is advantageous; as an excessive

Table 1 Average predictive accuracy of EBMGP model derived from fivefold cross-validation utilizing subset features selected via elastic net

Datasets	Traits	All	Top1000	Top3000	Top5000	Top7000	Top9000
Rice	SW	0.8072	0.8095	0.8088	0.8256	0.8220	0.8204
	FLW	0.7299	0.7131	0.7261	0.7347	0.7323	0.7324
	AC	0.7564	0.8360	0.8329	0.8311	0.8199	0.8193
	PH	0.7141	0.6889	0.6945	0.7039	0.7015	0.6958
	SNPP	0.5569	0.5664	0.5636	0.5931	0.5793	0.5631
Sorghum	MO	0.5717	0.5541	0.5519	0.5825	0.5721	0.5623
	YLD	0.3738	0.2990	0.3326	0.3774	0.3745	0.3806
	HT	0.5945	0.5818	0.5995	0.6252	0.5975	0.5984
Soybean	Protein	0.7051	0.7219	0.7622	0.7517	0.7544	0.7410
	Steartic	0.6456	0.6617	0.7114	0.6984	0.6953	0.6851
	R8	0.8280	0.8349	0.8623	0.8531	0.8525	0.8486
	SdWgt	0.9147	0.9121	0.9325	0.9287	0.9280	0.9278
	Yield	0.7851	0.7956	0.8269	0.8159	0.8176	0.8107
Bulls	MS	0.2771	0.3456	0.3819	0.3697	0.3695	0.3742
	NMSP	0.3139	0.3488	0.3537	0.3707	0.3662	0.3493
	VE	0.3510	0.3857	0.3881	0.3951	0.3902	0.3736

number of features may introduce noise, hinder model interpretability, and compromise efficiency. This highlights the importance of balancing feature selection with dataset characteristics to optimize predictive performance.

Comparison of SNP representations

Based on the findings in Sect. "Feature Selection's Influence on EBMGP," we selected the top 5,000 features for the rice, sorghum, and bull datasets, and the top 3,000 features for the soybean dataset to evaluate the impact of various SNP representations on model performance. As presented in Table 2, beyond the MS prediction task in bulls, EBMGP achieves higher accuracy with BERT embedding compared to one-hot encoding, with the most significant improvement of 5.52% in the rice AC prediction task. The average prediction accuracy of EBMGP across 16 tasks with BERT embedding (LD threshold = 0.8) reached 0.6565, representing a

Table 2 Comparison the effects of different SNP representations on EBMGP

Datasets	Traits	one-hot	BERT-1	BERT-2	BERT-3	BERT-4	BERT-5
Rice	SW	0.8123	0.8149	0.8212	0.8200	0.8262	0.8256
	FLW	0.7063	0.7191	0.7257	0.7274	0.7218	0.7347
	AC	0.7876	0.8196	0.8325	0.8283	0.8320	0.8311
	PH	0.6867	0.7011	0.6936	0.7001	0.7077	0.7039
	SNPP	0.5812	0.5757	0.5743	0.5901	0.5837	0.5931
Sorghum	MO	0.5559	0.5543	0.5595	0.5629	0.5663	0.5825
	YLD	0.3725	0.3725	0.3746	0.3722	0.3846	0.3774
	HT	0.6233	0.5968	0.6279	0.6171	0.6126	0.6252
Soybea	Protein	0.7563	0.7617	0.7641	0.7622	0.7644	0.7622
	Steartic	0.7070	0.7116	0.7137	0.7113	0.7110	0.7114
	R8	0.8592	0.8630	0.8626	0.8627	0.8626	0.8623
	SdWgt	0.9310	0.9319	0.9317	0.9326	0.9317	0.9325
	Yield	0.8245	0.8253	0.8264	0.8271	0.8268	0.8269
Bulls	MS	0.3770	0.3740	0.3729	0.3662	0.3704	0.3697
	NMSP	0.3542	0.3517	0.3604	0.3787	0.3616	0.3707
	VE	0.3839	0.3735	0.3719	0.3844	0.3879	0.3951
Mean		0.6449	0.6467	0.6508	0.6527	0.6532	0.6565

one-hot: EBMGP use one-hot to represent SNP genotype; BERT-1: EBMGP use BERT embedding to represent SNP genotype, without LD information; BERT-2, BERT-3, BERT-4, and BERT-5: EBMGP use BERT embedding to represent both SNP genotype and LD with adjacent SNPs, the criterion for dividing LD is 0.2, 0.4, 0.6, and 0.8, respectively

1.80% improvement compared to one-hot encoding (0.6449). Moreover, as the LD threshold gradually decreases (from 0.8 to 0.2), the model's average prediction accuracy of 16 task declines from 0.6565 to 0.6508, with the lowest accuracy observed when no LD is used for semantic segmentation (0.6467). These results suggest that BERT embeddings capture the complex associations between SNPs more effectively than one-hot encoding. Furthermore, applying a high-LD threshold for semantic segmentation enhances the model's understanding of the genetic structure, leading to improved prediction accuracy.

To substantiate the efficacy of the SNP representation methodology, we retrofitted the SoyDNGP model by substituting its initial convolutional layer with BERT embedding (LD threshold = 0.8). The SoyDNGP model, when augmented with BERT embedding, demonstrated a spectrum of enhancements in predictive accuracy across 12 tasks, with improvements spanning from 0.08 to 10.60% (Fig. 2). The mean predictive accuracy for 16 traits using SoyDNGP with BERT embedding reached 0.6258, which exceeds the accuracy achieved through one-hot encoding by 1.51%. These experimental results underscore the superiority of our proposed SNP representation method in enabling deep learning models to more effectively decipher the underlying patterns within SNPs, as compared to the conventional one-hot encoding approach.

Impact of pooling strategies on EBMGP performance

In response to the stringent requirements for enhanced model semantic comprehension and to mitigate the loss of critical features, we introduced an innovative multi-head attention pooling methodology. We conducted a fivefold cross-validation to assess the performance of MAP in comparison to four widely used pooling strategies. We trained the rice, sorghum, and bull datasets using the top 5,000 SNPs and the soybean dataset using the top 3,000 SNPs, with SNP representation specified in the methods section(LD threshold = 0.8). The average predictive accuracies across 16 tasks for average pooling(AP), max pooling(MP), soft pooling(SP), local importance-based pooling (LIP), and MAP were 0.6441, 0.6379, 0.6447, 0.6449, and 0.6565, respectively (Fig. 3A). When equipped with MAP, EBMGP achieved an average predictive accuracy of 0.7377 on the rice dataset, consisting of 413 lines, outperforming the second-best method, SP (0.7248), by 1.77%. Similarly, on the sorghum dataset with 451 lines, EBMGP with MAP reached an average accuracy of 0.5284, exceeding the next best method, AP (0.5097), by 3.67%. For the soybean dataset, which encompasses a substantial 13,974 lines, the predictive accuracy across the five pooling methods showed no significant differences. Meanwhile, for the bull dataset, comprising 1,508 lines, EBMGP with MAP attained an average accuracy of 0.3785, marginally outperforming the second-ranked method, LIP (0.3749), by 0.95%. The MSE of phenotype prediction for rice, sorghum, soybean, and bulls is shown in Fig. 4. With MAP, EBMGP achieves the lowest MSE in 15 out of 16 prediction tasks (Fig. 3B). The average MSE across 16 tasks for AP, MP, SP, LIP, and MAP were 0.7833, 0.7530, 0.7333, 0.7725, and 0.5704, respectively. The paired t-test results for different pooling strategies are presented in Supplemental Table 1. The results indicated that the prediction performance of MAP differs significantly from that of the other four pooling strategies, reaching a highly significant level ($p\text{-value} < 0.01$). The computation time required by different pooling strategies is presented in Supplemental Table 2.

These findings indicate that MAP is particularly effective in enhancing model prediction accuracy and reducing MSE when the sample size is small, outperforming other pooling techniques. In scenarios with a sufficiently large sample size, while differences in prediction accuracy across pooling

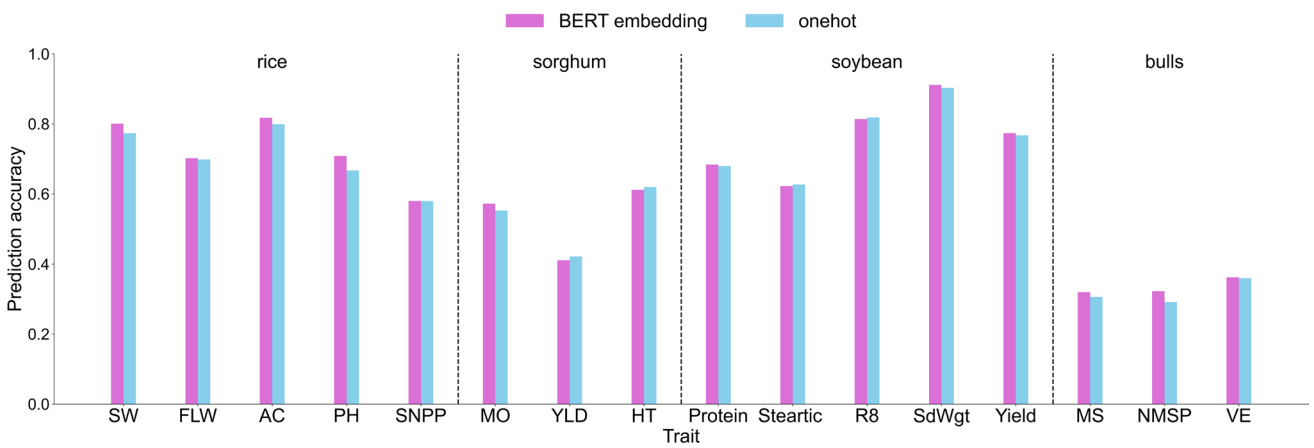


Fig. 2 Comparative analysis of prediction accuracies achieved by SoyDNGP utilizing one-hot encoding and BERT embedding

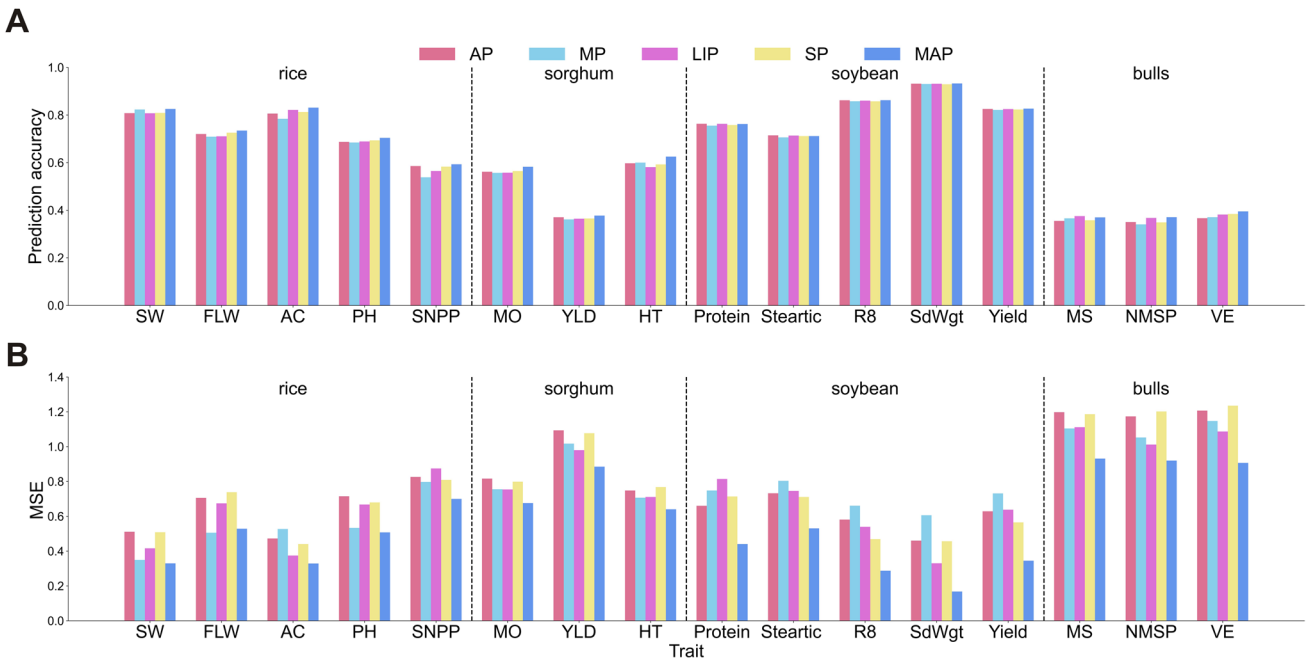


Fig. 3 Comparison between different pooling strategies. **A** The average prediction accuracy of EBMGP from fivefold cross-validation using different pooling. **B** The average MSE of EBMGP from fivefold cross-validation using different pooling. AP: avg pooling, MP: max pooling, LIP: local importance pooling, SP: soft pooling, MAP: multi-head attention pooling

methods are not significant, MAP consistently achieves the lowest MSE. Overall, MAP emerges as a sophisticated pooling technique that effectively minimizes information loss while simultaneously reducing dimensionality, enhances prediction accuracy, and ensures stability in predictions.

Assessment of nonlinear relationships

The nonlinear relationship between observer values and predicted values of EBMGP was calculated with kernel density plots as the criterion of model evaluation. EBMGP was trained on top 5,000 SNPs from the rice, sorghum, and bull datasets and top 3,000 from the soybean dataset, with SNP representation outlined in the methods section (LD threshold = 0.8). Density plots were employed to illustrate the relationships between observed values and various features, providing a clear visualization of their distributions and potential associations. As illustrated in Fig. 4 A-P, all density plots exhibit distinct clustering patterns rather than a completely random distribution, indicating that the model effectively captures the underlying patterns in the observations. For rice, sorghum, and bulls (Fig. 4 A-H, N-P), the data points display irregular distributions, reflecting nonlinear trends between predicted and observed values, which the model is able to capture effectively. In contrast, for soybean (Fig. 4 I-M), the data points are densely concentrated along the diagonal (red line), suggesting that the relationship between predicted and observed values is predominantly linear. Additionally, orthogonal distance regression (ODR) was performed to fit the observed and predicted values (blue line). The angle between the ODR regression line and the diagonal represents the degree of deviation from linearity; a larger angle indicates a weaker linear relationship and greater dispersion of data points relative to the ideal linear relationship (diagonal), while a smaller angle signifies a stronger linear relationship. The results obtained from orthogonal distance regression closely correspond to those of the two-dimensional density plots. Given the pivotal role of activation functions in enabling deep learning models to capture nonlinear relationships, we further investigated EBMGP's ability to model such relationships through ablation experiments that excluded the GELU activation function (Fig. 4 Q, R). The average prediction accuracy across 16 tasks of EBMGP is 0.6846 with GELU and 0.6724 without GELU, while the corresponding average MSE values are 0.5356 and 0.5373. Across the rice, sorghum, and bull datasets, the absence of the GELU activation function resulted in decreased prediction accuracy and increased MSE in most cases. Conversely, for the soybean dataset, all five traits experienced slight gains in accuracy and significant reductions in MSE. This evidence indicates that the soybean data may primarily reflect linear relationships, in contrast to the other datasets, where nonlinear relationships are more influential.

Over all, EBMGP effectively captures nonlinear relationships across multiple environments, enabling highly accurate

predictions, particularly in densely populated regions of the phenotypic distribution.

Comparative analysis of EBMGP with other GP models

We assessed the performance of EBMGP against seven commonly used GP models (Fig. 5). EBMGP was trained on top 5,000 SNPs from the rice, sorghum, and bull datasets and top 3,000 from the soybean dataset, with SNP representation outlined in the methods section (LD threshold = 0.8). Each reported accuracy is an average derived from fivefold cross-validation. The overall average prediction accuracies (R) across 16 tasks for GBLUP, RKHS, Bayes B, BL, DLGWAS, SoyDNGP, DNNGP, and EBMGP were 0.6371, 0.6370, 0.6427, 0.6359, 0.5870, 0.6165, 0.5927, and 0.6565, respectively, while the corresponding average MSE values were 0.5684, 0.5672, 0.5574, 0.5688, 0.6692, 0.6086, 0.6398, and 0.5704. Remarkably, EBMGP demonstrated superior performance, attaining the highest prediction accuracy in 13 tasks and the lowest MSE in 5 tasks, reflecting its adaptability and precision in addressing a wide range of prediction challenges.

In the analysis of the rice dataset (Fig. 5 A, E), EBMGP demonstrated exceptional predictive accuracy for SW, FLW, AC, and SNPP, surpassing the next best-performing model by margins of 1.39%, 2.52%, 0.74%, and 2.51%, respectively. Conversely, Bayes B outperformed EBMGP in predicting PH, achieving an accuracy advantage of 2.31%. Furthermore, EBMGP achieved the lowest MSE for SW, while Bayes B recorded the lowest MSE across the other four traits. As shown in the sorghum dataset (Fig. 5 B, F), EBMGP outshone the next best model by 0.90% for MO and 0.94% for HT, while SoyDNGP excelled in YLD accuracy. Bayes B resulted in the lowest MSE across all three traits. In the soybean dataset (Fig. 5 C, G), EBMGP achieved the highest prediction accuracy for five traits, with improvements ranging from 1.74 to 9.55% over the second-best model. Additionally, it recorded the lowest MSE for four traits, reducing by 8.44% to 11.3%. Within the bull dataset (Fig. 5 D, H), EBMGP attained the highest prediction accuracy for MS and NMSP, outperforming BayesB—the second highest model—by 2.90% and 1.09%. For VE, GBLUP respectively achieved the highest prediction accuracies, exceeding EBMGP by 6.52%. BL, Bayes B, and GBLUP recorded the lowest MSE for MS, NMSP, and VE, respectively.

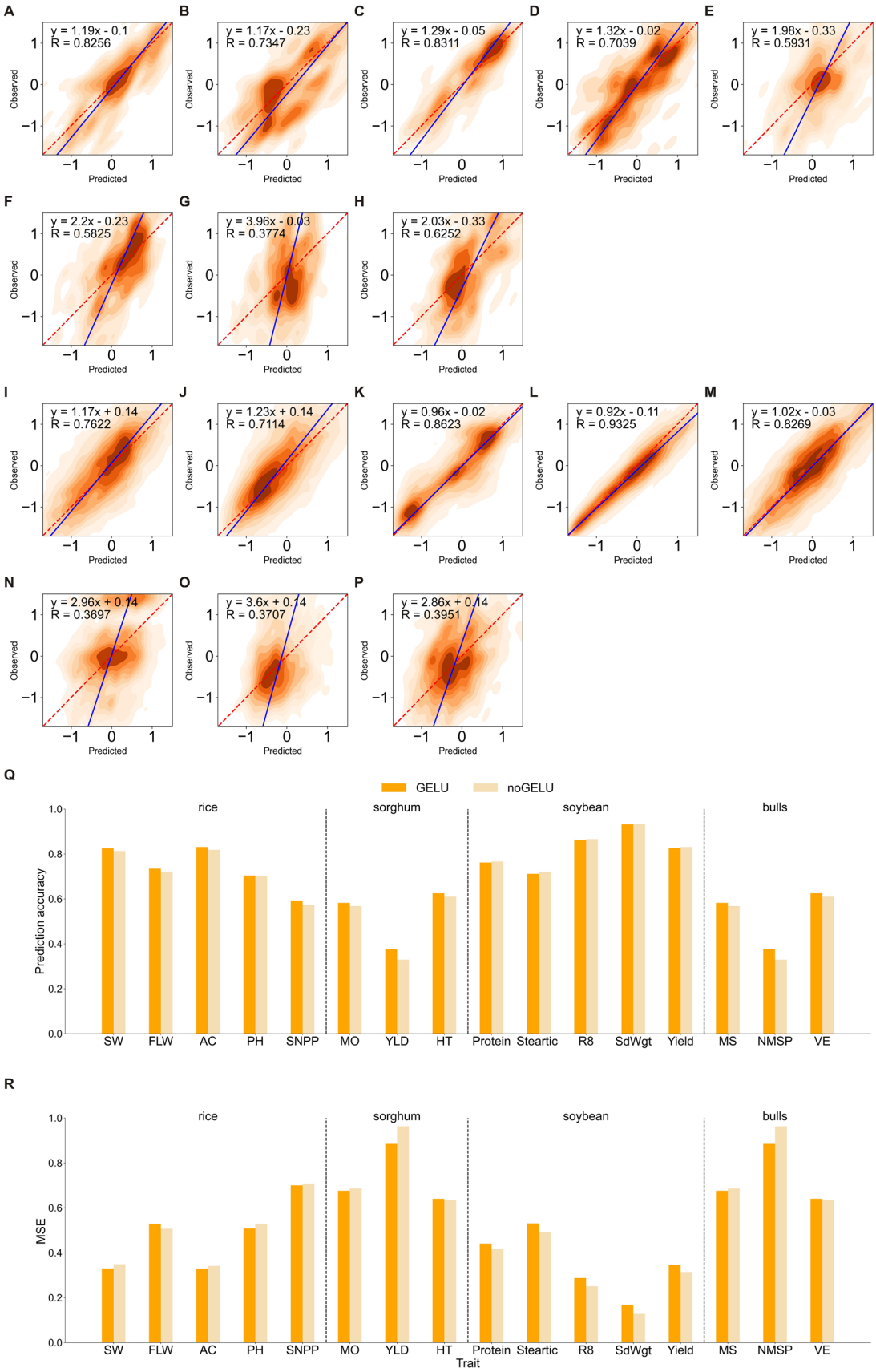
To further assess the robustness of these models, we calculated the standard error (SE) of prediction accuracy through fivefold cross-validation (Supplementary Table 3). Among the models, RKHS demonstrated the lowest SE of 0.0153, indicating its high consistency and stability in predictions. In contrast, DLGWAS exhibited the highest SE of 0.0195, suggesting greater variability in its predictions. EBMGP ranked fourth with an SE of 0.0175, reflecting moderate variability in its performance across the tasks.

In general, EBMGP excels with high prediction accuracy and robust generalization ability, especially in larger datasets. Its performance, however, may be less consistent in smaller datasets, as the feature selection process benefits from a larger volume of data. Traditional models such as Bayes B maintain stability with smaller datasets but are unable to capture more intricate relationships as the dataset grows.

Discussion

Despite significant long-term genetic gains achieved through modern breeding methods and technologies, the current rate of genetic improvement must be accelerated to meet growing agricultural demands. Genomic selection (GS) has been validated as a powerful tool for boosting genetic gain in both plant and animal breeding. The precision of prediction model holds an important position in GS. Evidence indicates that deep learning model captures nonlinear patterns more efficiently than conventional models and integrates data from various sources without the need for feature engineering (Montesinos-López et al. 2021). However, ensuring their robustness and practical applicability in breeding programs remains a critical area of research. Our study highlights that EBMGP effectively models genomic relationships, particularly in datasets with a high sample-to-feature ratio, making it a promising tool for GS applications. By integrating linkage disequilibrium (LD) information into SNP representations, EBMGP improves the accuracy of genomic predictions, particularly for traits governed by intricate genetic architectures. This capability is particularly relevant for breeding programs aiming to enhance complex traits controlled by multiple loci. Additionally, our findings demonstrate that EBMGP maintains stable performance across different datasets, suggesting its potential adaptability across diverse breeding populations and environments. Moreover, the practical implications of our results extend beyond theoretical advancements. The improved predictive accuracy of EBMGP could aid breeders in making more informed selection decisions, ultimately accelerating genetic gains. Unlike conventional deep learning models, which often require extensive computational resources, EBMGP achieves a balance between accuracy and efficiency, making it a feasible option for large-scale breeding programs. Future research should focus on optimizing its computational efficiency further and expanding its applicability to multi-environment trials and large-scale breeding populations.

The so-called $p >> n$ challenge, where the number of features (p) greatly exceeds the number of samples (n), has emerged as a significant barrier in GP. Feature selection


A
y = 1.19x - 0.1
R = 0.8256
B
y = 1.17x - 0.23
R = 0.7347
C
y = 1.29x - 0.05
R = 0.8311
D
y = 1.32x - 0.02
R = 0.7039
E
y = 1.98x - 0.33
R = 0.5931
F
y = 2.2x - 0.23
R = 0.5825
G
y = 3.96x - 0.03
R = 0.3774
H
y = 2.03x - 0.33
R = 0.6252
I
y = 1.17x + 0.14
R = 0.7622
J
y = 1.23x + 0.14
R = 0.7114
K
y = 0.96x - 0.02
R = 0.8623
L
y = 0.92x - 0.11
R = 0.9325
M
y = 1.02x - 0.03
R = 0.8269
N
y = 2.96x + 0.14
R = 0.3697
O
y = 3.6x + 0.14
R = 0.3707
P
y = 2.86x + 0.14
R = 0.3951
Observed
Predicted
Q
GELU
noGELU
rice
sorghum
soybean
bulls
Prediction accuracy
SW FLW AC PH SNPP MO YLD HT Protein Steartic R8 SdWgt Yield MS NMSP VE
R
MSE
Trait

◂**Fig. 4** Density plots of phenotypes predicted by EBMGP. Panels **A**–**E** represent rice phenotypes: SW, FLW, AC, PH, and SNPP. Panels **F**, **G**, and **H** represent sorghum panels phenotypes: MO, YLD, HT. **I**—**M** represent soybean phenotypes: protein, Stearic, R8, SdWgt, and yield, respectively. Panels **N**, **O**, and **P** represent bulls phenotypes: MS, NMSP, and VE. R, Pearson's correlation coefficient, which is the mean of the fivefold cross-validation. **Q** The prediction accuracy of EBMGP with and without GELU. **P** The MSE of EBMGP with and without GELU

plays a pivotal role in overcoming this issue in GP through three primary methodologies: principal component analysis (PCA), SNP ranking via GWAS results, and machine learning-based SNP prioritization. Both DNNGP and PNNGS have adopted PCA to mitigate the high dimensionality of data (Wang et al. 2023; Xie et al. 2024). Heinrich developed a feature selection framework that ranks SNPs based on GWAS results, leading to a marked improvement in prediction accuracy (Heinrich et al. 2023; Li et al. 2018). Random Forests (RFs) and gradient boosting machines (GBMs) have been effective in identifying SNPs associated with growth traits in Brahman cattle, with the top 3,000 SNPs yielding GEBV prediction accuracies comparable to using all SNPs (Li et al. 2018). Azodi employed three feature selection methods—Random Forest (RF), EN, and Bayes A—to rank SNPs (Azodi et al. 2019). In species with a low *p*:*n* ratio, model performance tends to peak more rapidly with an increasing number of features (*p*) compared to species with a higher *p*:*n* ratio (Azodi et al. 2019). This trend was also observed in our study. For instance, in soybeans, which have a low *p*:*n* ratio, EBMGP's performance peaked at $p = 3000$. In contrast, for rice, sorghum, and bull, with higher *p*:*n* ratios, the performance plateaued at $p = 5000$. Therefore, with an adequate sample size, the number of features selected for subsequent training can be minimized.

Previous research has established that biological sequences share significant parallels with human language, spanning from lexical to grammatical structures [33, 34]. The application of BERT embedding to representing biological sequences has, in recent years, achieved remarkable results in various tasks, including the prediction of promoters, splice sites, identification of antibacterial peptides, and recognition of RNA N7-methylguanosine sites (Ji et al. 2021; Yang et al. 2022; Zhang et al. 2021a, 2021b). However, these studies have typically employed BERT embeddings to represent nucleic acid or protein genotypes alone, without considering the semantic segmentation of biological sequences, thus limiting the model's capacity to grasp more profound semantic insights. It has been observed that SNPs with high LD can exhibit identical gene models, such as those related to salt tolerance in spring wheat (Hasseb et al. 2022), suggesting a potential shared semantic context among high-LD SNPs. Moreover, LD blocks, which typically represent inherited haplotype structures, offer a more comprehensive genetic context than analyzing SNPs in isolation (Karimi et al. 2018). In this study, we utilized BERT embeddings not only to represent SNPs as "words," but also to model adjacent SNPs with similar LD levels, termed LD blocks, as "sentences." This representation method has proven to exhibit robust generalization capabilities in both EBMGP and SoyDNGP models. This effectiveness may stem from the method's ability to concurrently extract complex associations between SNPs and LD blocks, while reducing redundancy by treating high-LD SNPs as a single entity. The inclusion of LD blocks has also been noted to enhance statistical power in genome-wide association studies (GWAS) (Wu et al. 2020). Future advancements in modeling biological sequences could potentially benefit from clustering sites with analogous semantic traits.

Pooling strategies are predominantly employed to downsample feature maps and extract larger-scale features that are invariant under minor local transformations. In this paper, we introduce multi-head attention pooling, inspired by the Transformer's attention mechanism, to adaptively extract crucial features from various subspaces that influence prediction outcomes (Xiong et al. 2020). Unlike traditional pooling methods, MAP allows for flexible, context-dependent feature aggregation, which is crucial for handling the heterogeneity and high dimensionality of genomic data. Furthermore, its capability to preserve important genetic signals while filtering noise enhances its effectiveness in genomic prediction tasks. There were also a variety of pooling strategies have been proposed, each tailored to specific objectives. For instance, soft pooling retains more information in the reduced activation maps (Stergiou et al. 2021), while spatial pyramid pooling captures spatial or structural features (He et al. 2014).We compare MAP with two novel pooling strategies, soft pooling and LIP, both of which perform downsampling by weighting. However, their outputs are heavily influenced by high-valued input features, making them more prone to noise. MAP addresses this issue by integrating weighting for feature preservation with the incorporation of diverse spatial features across multiple heads, resulting in more robust and comprehensive feature extraction. The results presented in Fig. 3 highlight the effectiveness of our pooling method. While sufficient sample sizes enable any pooling strategy to extract adequate information for accurate predictions, our method's versatility and broad applicability are particularly beneficial given the often limited sample sizes in GP data due to cost constraints.

A plenty of previous studies has delved into comparisons between statistical models, machine learning (ML) models, and DL models (Abdollahi-Arpanahi et al. 2020; Azodi et al. 2019; Gill et al. 2022). Gill et al.'s comparison of DL and ML models across 14 prediction tasks revealed that DL only surpassed ML models on one occasion (Gill et al. 2022). XGBoost and RF have been found to better

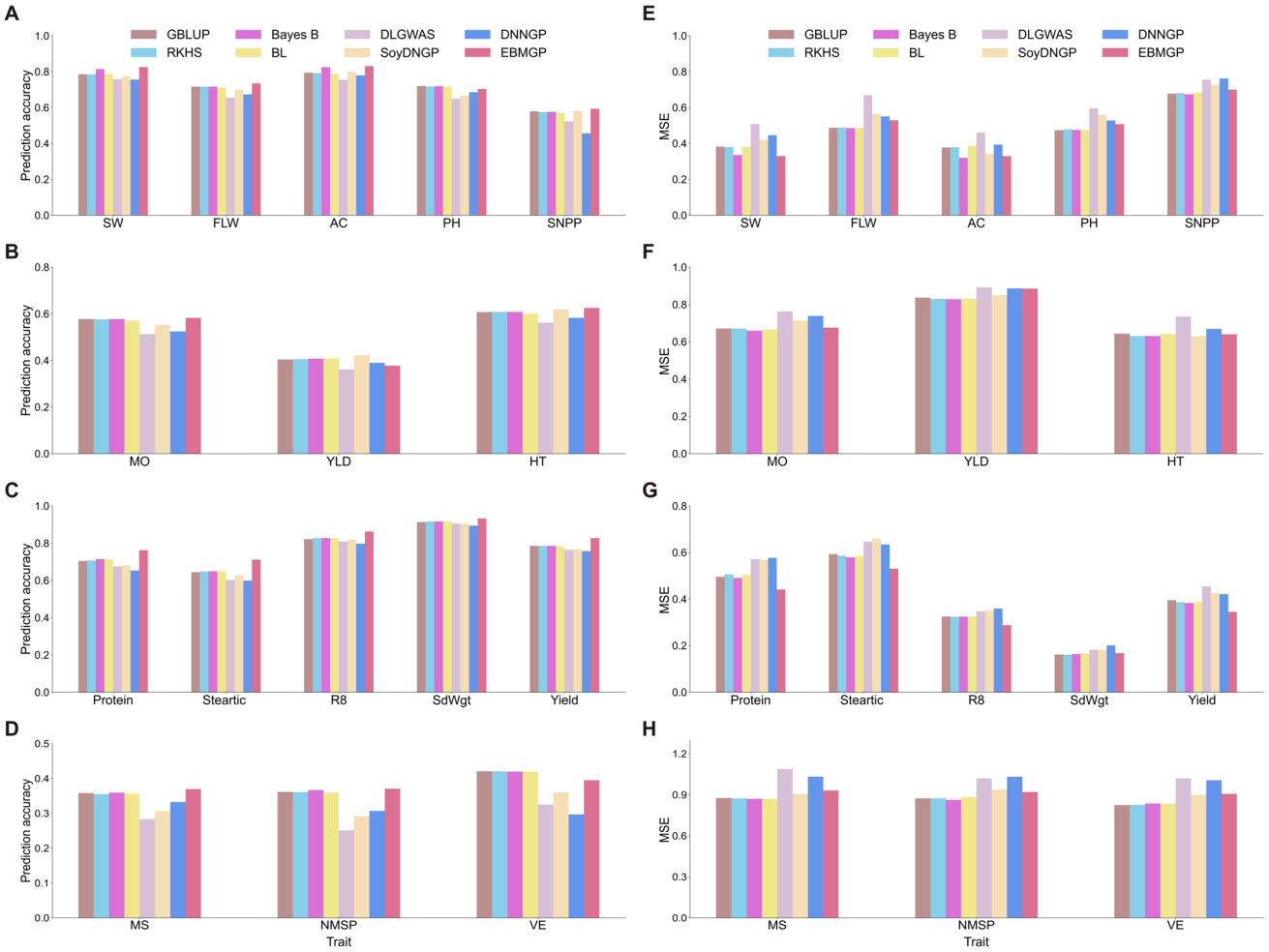


Fig. 5 Comparative analysis of predictive performance between EBMGP and other models. **A** Prediction accuracy of 8 GP models for rice dataset. **B** Prediction accuracy of 8 GP models for sorghum dataset. **C** Prediction accuracy of 8 GP models for soybean dataset. **D** Prediction accuracy of 8 GP models for bull dataset. **E** to **H** separately represent the MSE of 8 GP models for rice dataset, sorghum dataset, soybean dataset, and bull dataset

capture nonadditive effects in GP than DL architectures (Gill et al. 2022). A comparison utilizing the bull dataset demonstrated that gradient boosting is a reliable method for predicting traits influenced by nonadditive gene action, while DL approaches only showed an advantage in GP when nonadditive variance was substantial (Abdollahi-Arpanahi et al. 2020). However, studies in this domain often compare statistical and ML models with shallow and basic DL architectures, such as artificial neural networks (ANNs), CNN, and multilayer perceptrons (MLP) (Abdollahi-Arpanahi et al. 2020; Azodi et al. 2019; Gill et al. 2022). DNNGP has performed as well as or even surpassed commonly used linear, ML, and DL models across a wide range of tasks (Wang et al. 2023). SoyDNGP, a cutting-edge DL model for GP, has shown balanced performance across all classification traits, even when compared to several high-accuracy ML models (Gao et al. 2023). Moreover, it has outperformed both deepGS and DNNGP in regression tasks (Gao et al. 2023). In this study, we developed EBMGP and compared it with seven leading GP models: a linear model (GBLUP), three ML models (RKHS, Bayes B, and Bayesian LASSO), and three DL models (DLGWAS, SoyDNGP, and DNNGP) across 13 prediction tasks. Our model demonstrated robust effectiveness and generalization. However, it is noteworthy that in both the current study and the SoyDNGP paper, DNNGP did not exhibit the advantages claimed in its original work (Gao et al. 2023). This may be attributed to DNNGP's requirement for adjusting a substantial number of parameters, resulting in a vast array of potential combinations and complicating the identification of the optimal configuration. In contrast, both EBMGP and SoyDNGP necessitate fewer parameter adjustments, rendering them more user-friendly. Nevertheless, DNNGP possesses a unique

strength in its ability to process diverse and complex inputs, including genomes, transcriptomes, and proteomes.

Previous studies have consistently shown that plants generally exhibit lower population genetic diversity than animals, a discrepancy that profoundly impacts genetic parameters such as allele frequency distributions and genetic relationships (De Kort et al. 2021). These differences can introduce biases in model predictions. To explore this, we analyzed genetic diversity across our four datasets using observed heterozygosity (Ho) and expected heterozygosity (He) (Supplementary Table 4). The plant datasets revealed a notable reduction in Ho relative to He (Ho < < He), a pattern indicative of lower genetic diversity likely driven by strong population structure, inbreeding, or selective breeding practices. In contrast, the bull dataset exhibited a more balanced Ho-He relationship, reflecting stable genetic diversity. Interestingly, EBMGP and baseline models consistently achieved lower prediction accuracy in the bull dataset compared to the plant datasets, suggesting that genetic diversity patterns significantly influence model performance. One plausible explanation is that higher genetic diversity, as seen in the bull dataset, may result in weaker SNP-trait associations, complicating prediction tasks. To address these challenges, we employed Elastic Net feature selection to reduce dataset-specific noise by prioritizing informative SNPs. Furthermore, the integration of BERT embeddings in EBMGP enables the capture of context-dependent SNP relationships, moving beyond reliance on raw allele frequencies and enhancing model robustness across diverse genetic architectures.

EBMGP shows promising performance, but several aspects require further refinement. The model's performance on smaller datasets is less stable, suggesting the need for improved feature selection methods tailored to limited data to reduce overfitting. Additionally, the approach to semantic segmentation of SNP sequences could be enhanced to better capture complex genetic interactions, particularly for traits influenced by multiple loci. A more sophisticated approach to semantic segmentation could provide deeper insights into the underlying genetic structures and lead to improved model performance. While the model benefits from BERT embeddings and multi-head attention, their computational cost can be a bottleneck when scaling to larger datasets. Future work should focus on improving computational efficiency to better handle large-scale genomic data. This can be achieved through alternative embedding strategies, such as lightweight transformer architectures (e.g., FastEmbed), to reduce processing time and memory requirements (Fang et al. 2020). Parallel processing using distributed computing frameworks like TensorFlow Distributed and PyTorch Distributed can accelerate model training across multiple GPUs or high-performance computing clusters. Dimensionality reduction techniques can streamline computation by identifying the most informative SNPs. Model compression methods, such as pruning, quantization, and knowledge distillation, can further enhance efficiency by optimizing model size and inference speed. By integrating these optimizations, future versions of EBMGP can be more scalable and resource-efficient, facilitating its application in real-world breeding programs and large-scale genomic studies. Collectively, no single model dominates across all tasks, yet DL holds greater potential in GP, driven by its ongoing technical innovations, such as the cosine annealing learning rate adjustment and multi-head attention mechanism highlighted in this study, coupled with its flexible and adaptable structure.

Conclusion

We introduce EBMGP, a deep learning model for genomic prediction that leverages EN feature selection, a novel SNP representation, and multi-head attention pooling. Employing EN for feature selection streamlines EBMGP by minimizing the feature space. By utilizing BERT embeddings to treat SNPs as natural language, EBMGP captures complex associations at both the SNP and LD block levels. This SNP representation method has shown its effectiveness across applications in both EBMGP and SoyDNGP. Our multi-head attention pooling proves highly effective in small datasets, yet remains competitive with top-tier pooling methods in larger datasets. EBMGP's efficacy is validated in four animal and plant datasets, and we believe it holds great potential for advancing data-driven decisions in animal and plant breeding programs.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00122-025-04894-z.

Acknowledgements This study was supported by the Special Funds for Construction of Innovative Provinces in Hunan Province (2021NK1011), the Science and Technology Innovation Program of Hunan Province (2023NK2001, 2024RC9014), and Changsha Natural Science Foundation Project (kq2402109).

Author contributions LJ finished the study and wrote this manuscript. WH reviewed edited this paper. HZ contributed to formal analysis. LLX involved in data curation and visualization. CHL carried out formal analysis. ZMY developed conceptualization and funding acquisition. LZL performed conceptualization, review and editing, supervision, and funding acquisition.

Funding This study was funded by the Hunan Engineering Laboratory for Analyse and Drugs Development of Ethnomedicine in Wuling Mountain, 2021NK1011, Zheming Yuan, Science and Technology Innovation Program of Hunan Province, 2023NK2001, Lanzhi Li, Science and Technology Innovative Research Team in Higher Educational Institutions of Hunan Province, 2024RC9014, Wei Hou, and Changsha Science and Technology Project, kq2402109, Lanzhi Li.

Data and code availability The authors declare that all data supporting the findings of this study come from public datasets. The source code of EBMGP can be freely downloaded from https://github.com/luxixi2021/EBMGP.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Abdollahi-Arpanahi R, Gianola D, Peñagaricano F (2020) Deep learning versus parametric and ensemble methods for genomic prediction of complex phenotypes. Genet Sel Evol 52:12

Ayres DL, Darling A, Zwickl DJ, Beerli P, Holder MT, Lewis PO, Huelsenbeck JP, Ronquist F, Swofford DL, Cummings MP, Rambaut A, Suchard MA (2011) BEAGLE: an application programming interface and high-performance computing library for statistical Phylogenetics. Syst Biol 61:170–173

Azodi CB, Bolger E, McCarren A, Roantree M, de Los Campos G, Shiu SH (2019) Benchmarking parametric and machine learning models for genomic prediction of complex traits. G3 Genes Genomes Genet 9(11):3691–3702

Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L, Nickerson DA (2004) Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. Am J Human Genet 74:106–120

Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaSci 4:7

Chen R, Wu J, Luo Y, Xu G (2024) PointMM: point cloud semantic segmentation CNN under multi-spatial feature encoding and multi-head attention pooling. Remote Sens 16:1246

Clark SA, van der Werf J (2013) Genomic best linear unbiased prediction (gBLUP) for the estimation of genomic breeding values. In: Gondro C, van der Werf J, Hayes B (eds) Genome-Wide Association Studies and Genomic Prediction. Humana Press, Totowa, NJ, pp 321–330

De Kort H, Prunier JG, Ducatez S, Honnay O, Baguette M, Stevens VM, Blanchet S (2021) Life history, climate and biogeography interactively affect worldwide genetic diversity of plant and animal populations. Nat Commun 12:516

Devlin J, Chang M-W, Lee K, Toutanova K (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. Association for computational linguistics, Minneapolis, Minnesota, pp 4171–4186

Endelman JB (2011) Ridge regression and other kernels for genomic selection with R Package rrBLUP. Plant Genome 4

Fang Y, Liu Y, Huang C, Liu L (2020) FastEmbed: predicting vulnerability exploitation possibility based on ensemble machine learning algorithm. PLoS ONE 15:e0228439

Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. J Stat Softw 33:1–22

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. Brief Bioinform. https://doi.org/10.1093/bib/bbad349

Gao Z, Wang L, Wu G (2019) LIP: Local Importance-Based Pooling. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp 3354–3363

Gill M, Anderson R, Hu H, Bennamoun M, Petereit J, Valliyodan B, Nguyen HT, Batley J, Bayer PE, Edwards D (2022) Machine learning models outperform deep learning models, provide interpretation and facilitate feature selection for soybean trait prediction. BMC Plant Biol 22:180

Grant D, Nelson RT, Cannon SB, Shoemaker RC (2009) SoyBase, the USDA-ARS soybean genetics and genomics database. Nucleic Acids Res 38:D843–D846

Hasseb NM, Sallam A, Karam MA, Gao L, Wang RRC, Moursi YS (2022) High-LD SNP markers exhibiting pleiotropic effects on salt tolerance at germination and seedlings stages in spring wheat. Plant Mol Biol 108:585–603

He K, Zhang X, Ren S, Sun J (2014) Spatial pyramid pooling in deep convolutional networks for visual recognition. Springer International Publishing, Cham, pp 346–361

Heinrich F, Lange TM, Kircher M, Ramzan F, Schmitt AO, Gültas M (2023) Exploring the potential of incremental feature selection to improve genomic prediction accuracy. Genet Sel Evol 55:78

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

Johnson MS, Reddy G, Desai MM (2023) Epistasis and evolution: recent advances and an outlook for prediction. BMC Biol 21:120

Jubair S, Tucker JR, Henderson N, Hiebert CW, Badea A, Domaratzki M, Fernando WGD (2021) GPTransformer: a transformer-based deep learning method for predicting fusarium related traits in barley. Front Plant Sci 12:761402

Karimi Z, Sargolzaei M, Robinson JAB, Schenkel FS (2018) Assessing haplotype-based models for genomic evaluation in Holstein cattle. Can J Anim Sci 98:750–759

Kemppainen P, Knight CG, Sarma DK, Hlaing T, Prakash A, Maung Maung YN, Somboon P, Mahanta J, Walton C (2015) Linkage disequilibrium network analysis (LDna) gives a global view of chromosomal inversions, local adaptation and geographic structure. Mol Ecol Resour 15:1031–1045

Le NQK, Yapp EKY, Ho Q-T, Nagasundaram N, Ou Y-Y, Yeh H-Y (2019) iEnhancer-5Step: Identifying enhancers using hidden information of DNA sequences via Chou's 5-step rule and word embedding. Anal Biochem 571:53–61

Li J, Das K, Fu G, Li R, Wu R (2010) The Bayesian lasso for genome-wide association studies. Bioinformatics 27:516–523

Li B, Zhang N, Wang Y-G, George AW, Reverter A, Li Y (2018) Genomic prediction of breeding values using a subset of SNPs identified by three machine learning methods. Front Genet. https://doi.org/10.3389/fgene.2018.00237

Li L, Zheng X, Wang J, Zhang X, He X, Xiong L, Song S, Su J, Diao Y, Yuan Z, Zhang Z, Hu Z (2023) Joint analysis of

phenotype-effect-generation identifies loci associated with grain quality traits in rice hybrids. Nat Commun 14:3930
Liu Y, Wang D, He F, Wang J, Joshi T, Xu D (2019) Phenotype prediction and genome-wide association study using deep convolutional neural network of soybean. Front Genet 10:1091
Meuwissen TH, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829
Mishkin D, Sergievskiy N, Matas J (2017) Systematic evaluation of convolution neural network advances on the Imagenet. Comput vis Image und 161:11–19
Montesinos-López OA, Montesinos-López A, Pérez-Rodríguez P, Barrón-López JA, Martini JWR, Fajardo-Flores SB, Gaytan-Lugo LS, Santana-Mancilla PC, Crossa J (2021) A review of deep learning applications for genomic selection. BMC Genomics 22:19
Nirthika R, Manivannan S, Ramanan A, Wang R (2022) Pooling in convolutional neural networks for medical image analysis: a survey and an empirical study. Neural Comput Appl 34:5321–5347
Pérez P, de Los Campos G (2014) Genome-wide regression and prediction with the BGLR statistical package. Genetics 198(2):483–495
Stergiou A, Poppe R, Kalliatakis G (2021) Refining activation downsampling with SoftPool. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp 10337–10346
Wang K, Abid MA, Rasheed A, Crossa J, Hearne S, Li H (2023) DNNGP, a deep neural network-based method for genomic prediction using multi-omics data in plants. Mol Plant 16:279–293
Webber C (2017) Epistasis in neuropsychiatric disorders. Trends Genet 33:256–265
Wu G, Guo X, Xu B (2020) BAM: a block-based Bayesian method for detecting genome-wide associations with multiple diseases. Tsinghua Sci Technol 25:678–689
Xie Z, Weng L, He J, Feng X, Xu X, Ma Y, Bai P, Kong Q (2024) PNNGS, a multi-convolutional parallel neural network for genomic selection. Front Plant Sci. https://doi.org/10.3389/fpls.2024.1410596
Xiong R, Yang Y, He D, Zheng K, Zheng S, Xing C, Zhang H, Lan Y, Wang L, Liu T-Y (2020) On layer normalization in the transformer architecture. In: Proceedings of the 37th International Conference on Machine Learning. JMLR.org, p Article 975
Yan S, Wang J, Song Z (2022) Microblog sentiment analysis based on dynamic character-level and word-level features and multi-head self-attention pooling. Future Int 14:234
Yang M, Huang L, Huang H, Tang H, Zhang N, Yang H, Wu J, Mu F (2022) Integrating convolution and self-attention improves language model of human genome for interpreting non-coding regions at base-resolution. Nucleic Acids Res 50:e81
Yin H, Zhou C, Shi S, Fang L, Liu J, Sun D, Jiang L, Zhang S (2019) Weighted single-step genome-wide association study of semen traits in holstein bulls of China. Front Genet. https://doi.org/10.3389/fgene.2019.01053
Zhang L, Qin X, Liu M, Liu G, Ren Y (2021a) BERT-m7G: a transformer architecture based on BERT and stacking ensemble to identify RNA N7-methylguanosine sites from sequence information. Comput Math Methods Med 2021:7764764
Zhang Y, Lin J, Zhao L, Zeng X, Liu X (2021b) A novel antibacterial peptide recognition algorithm based on BERT. Brief Bioinform. https://doi.org/10.1093/bib/bbab200
Zhao K, Tung C-W, Eizenga GC, Wright MH, Ali ML, Price AH, Norton GJ, Islam MR, Reynolds A, Mezey J, McClung AM, Bustamante CD, McCouch SR (2011) Genome-wide association mapping reveals a rich genetic architecture of complex traits in Oryza sativa. Nat Commun 2:467
Zingaretti LM, Gezan SA, Ferrão LFV, Osorio LF, Monfort A, Muñoz PR, Whitaker VM, Pérez-Enciso M (2020) Exploring deep learning for complex trait genomic prediction in Polyploid outcrossing species. Front Plant Sci. https://doi.org/10.3389/fpls.2020.00025
Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. J R Stat Soc Ser B Stat Methodol 67:301–320