nature communications

Article

https://doi.org/10.1038/s41467-026-71035-5

Leveraging weighted embedding and Transformer architecture to improve phenotype prediction of complex traits for crops

Received: 7 August 2025

Accepted: 11 March 2026

Published online: 26 March 2026

Check for updates

Jing Li[1,5] ✉, Linfeng Yu[1,5], Mengfan Li[2,5], Rui Han[2], Yecheng Li[1], Abdulwahab Saliu Shaibu[3], Kwadwo Gyapong Agyenim-Boateng[4], Zhaoyi Hao[1], Yitian Liu[1], Bin Li[1], Shengrui Zhang[1], Liang Li[1], Lijuan Qiu[1] ✉ & Junming Sun[1] ✉

Understanding the relationship between genomic variation and phenotype is fundamental to deciphering the genetic architecture underlying complex traits. Yet, existing statistical models struggle to balance massive genomic datasets with biological interpretability. Here, we introduce GP-WAITER, a deep learning framework integrating GWAS-derived SNP weights into a hybrid convolutional neural network and Transformer architecture. By utilizing a weighted embedding mechanism and multi-head self-attention, GP-WAITER effectively captures long-range dependencies across ultra-long genomic sequences. The model consistently outperforms seven state-of-the-art genomic prediction models across six datasets, achieving up to a 77.5% improvement in prediction accuracy, a 78% reduction in mean squared error, and a 1.8-2.4fold increase in computational efficiency. Furthermore, GP-WAITER offers biological transparency by pinpointing key genetic variants driving specific traits. This scalable, interpretable framework provides a powerful tool for precision breeding and the functional interpretation of trait-associated variants.

The genetic architecture of complex traits in crops poses a major challenge for breeding programs, as numerous loci with small effect sizes typically control such traits. Genomic prediction (GP) has emerged as a transformative approach for accelerating breeding by enabling the estimation of phenotypic values from genome-wide high-density molecular markers. Traditional breeding methods, which rely on phenotypic selection through iterative crossing, are increasingly being supplanted by data-driven approaches that reduce time and resource costs. By shifting selection decisions from empirical observations to computational inference, GP has the potential to enhance selection efficiency across diverse agricultural systems. Since its inception, GP has been applied in a range of species, including livestock, crops, and aquaculture[1–3]. Among existing genomic selection (GS) tools, ridge regression best linear unbiased prediction (rrBLUP)[4] is a widely recognized state-of-the-art method. It employs a linear mixed-effect model to infer the genomic kinship among breeding materials,

[1]The State Key Laboratory of Crop Gene Resources and Breeding, National Engineering Laboratory for Crop Molecular Breeding, Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing, China. [2]Institute of Environment and Sustainable Development in Agriculture, Chinese Academy of Agricultural Sciences, Beijing, China. [3]Department of Agronomy, Bayero University Kano, Kano, Nigeria. [4]DynaMo Center, Department of Plant and Environmental Sciences, University of Copenhagen, Frederiksberg, Denmark. [5]These authors contributed equally: Jing Li, Linfeng Yu, Mengfan Li.
✉e-mail: lijing02@caas.cn; qiulijuan@caas.cn; sunjunming@caas.cn

applies maximum-likelihood algorithm to estimate fixed effects and SNP effects, and ultimately predicts phenotypes traits based on genome-wide SNP data.

The application of machine learning (ML) and deep learning (DL) has further enhanced genomic prediction by capturing non-linear and high-dimensional genotype-phenotype relationships[5,6]. ML models including Support Vector Regression (SVR)[7], eXtreme Gradient Boosting (XGBoost)[8], Light Gradient Boosting Machine (LightGBM)[9], DL models including convolutional neural networks (CNNs)[10], deep neural network genomic prediction (DNNGP)[11], and transformer-based model like Cropformer[12], GPformer[13], and DPCformer[14], have demonstrated promising results without prior assumptions about genetic models[15,16]. However, these models often rely on convolutional operations, which inherently limit information propagation and restrict the capture of long-range dependencies among distant single-nucleotide polymorphisms (SNPs). Additionally, previous approaches often rely on a small subset of variants that pass stringent significance thresholds[17] or public GWAS result[18], which may overlook informative signals and increase the risk of false-positive associations. Empirical studies across various species further indicate that models using only pre-selected SNPs generally do not outperform[19], and can even underperform[20,21], they discard substantial polygenic information by neglecting small-effect variants, which restricts their ability to capture the full genetic architecture of complex traits, exhibit poor generalizability across populations and environments due to the limited portability of significance-based SNP sets, underscoring the need for integrated prediction frameworks that more comprehensively leverage prior biological knowledge.

The Transformer architecture, initially developed for machine translation[22], has revolutionized natural language processing[23,24] and shown strong applicability in domains such as remote sensing[25], environmental time-series modeling[26,27], and human genomics[28–30]. These advances underscore the architecture's capacity to model long-range dependencies and assign dynamic attention weights across high-dimensional input spaces, avoiding local inductive biases in genetic feature extraction[31]. However, its application to crop genomics remains limited, the high-dimensional nature of genomic data likely to cause overfitting remains a challenge. Most transformer-based genomic prediction approaches limit the applicability of the method due to depend heavily on dimensionality reduction or feature selection like maximum information coefficient (MIC) or principal component analysis (PCA) to avoid overfitting[11,14], which are susceptible to false positive associations that introduce noise and reduce reliability. On the other hands, current Transformer-based models often lack interpretability, which hampers their adoption in biologically meaningful inference[12,13]. This gap highlights the need for genomic prediction frameworks that combine the modeling power of Transformers with efficient high-dimensional nature genomic data representations and interpretable.

In this study, we propose GP-WAITER (Genome-Phenotype prediction using Weighted self-AttentIon TransformER), a deep learning model that integrates GWAS-derived SNP weights into a hybrid CNN-Transformer architecture via an efficient tokenized embedding scheme. The model's design enables dynamic learning of trait-associated feature weights and effective modeling of long-range interactions within ultra-long genomic sequences. GP-WAITER achieves state-of-the-art performance on multiple benchmark datasets across soybean, maize, rice and wheat, and further provides biological insight through attention-weight and SHapley Additive exPlanations (SHAP) interpretability analyses. These contributions position GP-WAITER as a robust, scalable, and interpretable genomic prediction framework with strong potential to accelerate molecular breeding programs and deepen understanding of trait architecture in crops.

Results

Design and development of a deep learning model for genomic prediction

To construct the GP-WAITER multi-input deep learning model, integrating GWAS-derived SNP weights with a hybrid CNN-Transformer architecture, we designed three core components, comprising a weighted embedding block, multiple encoder blocks, and a predictor block. For inputs, the weighted embedding block takes genomic nucleotide sequences and SNP weights derived from GWAS to represent the potential phenotypic contribution of each variant. Extract the weight scores derived from GWAS *p*-values via -log(x) transformation, which correspond to each SNP in the aforementioned SNP genotype matrix, and arrange them into a one-dimensional vector with a length equal to the total number of SNPs in the soybean samples. Genomic sequence inputs are first tokenized and transformed into a three-dimensional tensor, while tokenized SNP weight information is assembled into a two-dimensional matrix. This tokenization process breaks down the sequence into smaller pieces that ensure computational efficiency in processing high-dimensional genomic data. Through element-wise multiplication, the model preserves the proportional influence of each SNP while enabling effective information compression and biologically meaningful representation learning. The output is an optimized four-dimensional tensor, wherein each SNP feature is scaled according to its phenotypic contribution. This tensor then serves as input for subsequent convolutional layers (Fig. 1A).

GP-WAITER employs a hybrid architecture that integrates CNN and Transformer components. The model accepts a three-dimensional tensor as input, which is first processed by a one-dimensional convolutional layer with a kernel size of (90,101) and a stride of (1,4), followed by batch normalization. Subsequently, three layers of two-dimensional convolution and Hyperbolic Tangent (tanh) activation function are applied, with kernel sizes of (60,101), (1,101), and (1,101), and strides of (1,2), (1,1), and (1,1), respectively. The resulting features are then passed to a three-layer Transformer encoder, each layer of which utilizes a multi-head self-attention mechanism. The multiple encoder blocks integrate a multi-head attention mechanism layer, two LayerNorm layers, and two linear transformation layers employing Gaussian Error Linear Unit (GELU) activation functions. The core, multi-head self-attention mechanism of the Transformer computes position-wise correlations across the optimized four-dimensional tensor. This approach ensures that all other sequence positions are given attention when processing specific sites to capture long-range dependencies and dynamically assign differential weights to each position in the genome (Fig. 1B). The LayerNorm layers stabilize output distributions from the self-attention operations, implicitly enhancing capacity for model generalization. The predictor block features a LayerNorm layer, a linear transformation layer with GELU activation, a one-dimensional batch normalization layer, and additional one-dimensional convolution and tanh activation layers. A combination of fully connected neural networks and convolutional layers is used to perform advanced feature extraction and transformation (Supplementary Fig. 1). Based on attention mechanisms and fully connected layers without recurrent structures, the model enables more efficient parallel data processing and improved computational efficiency. We refer to this hybrid attention-based framework as GP-WAITER, which enables interpretable and efficient modeling of complex trait architectures.

At each training step, a batch of tokenized sequences was sampled. The batch size was adapted to available hardware and model size. For model training, tokenized sequences were processed in batches of size 32 over 200 epochs using the backpropagation algorithm and mean squared error loss, optimized with the Adam optimizer (learning rate = 0.001). Dropout with a rate of 0.2 was also applied to enhance training efficiency. Training was conducted on a GeForce RTX 3080

https://doi.org/10.1038/s41467-026-71035-5

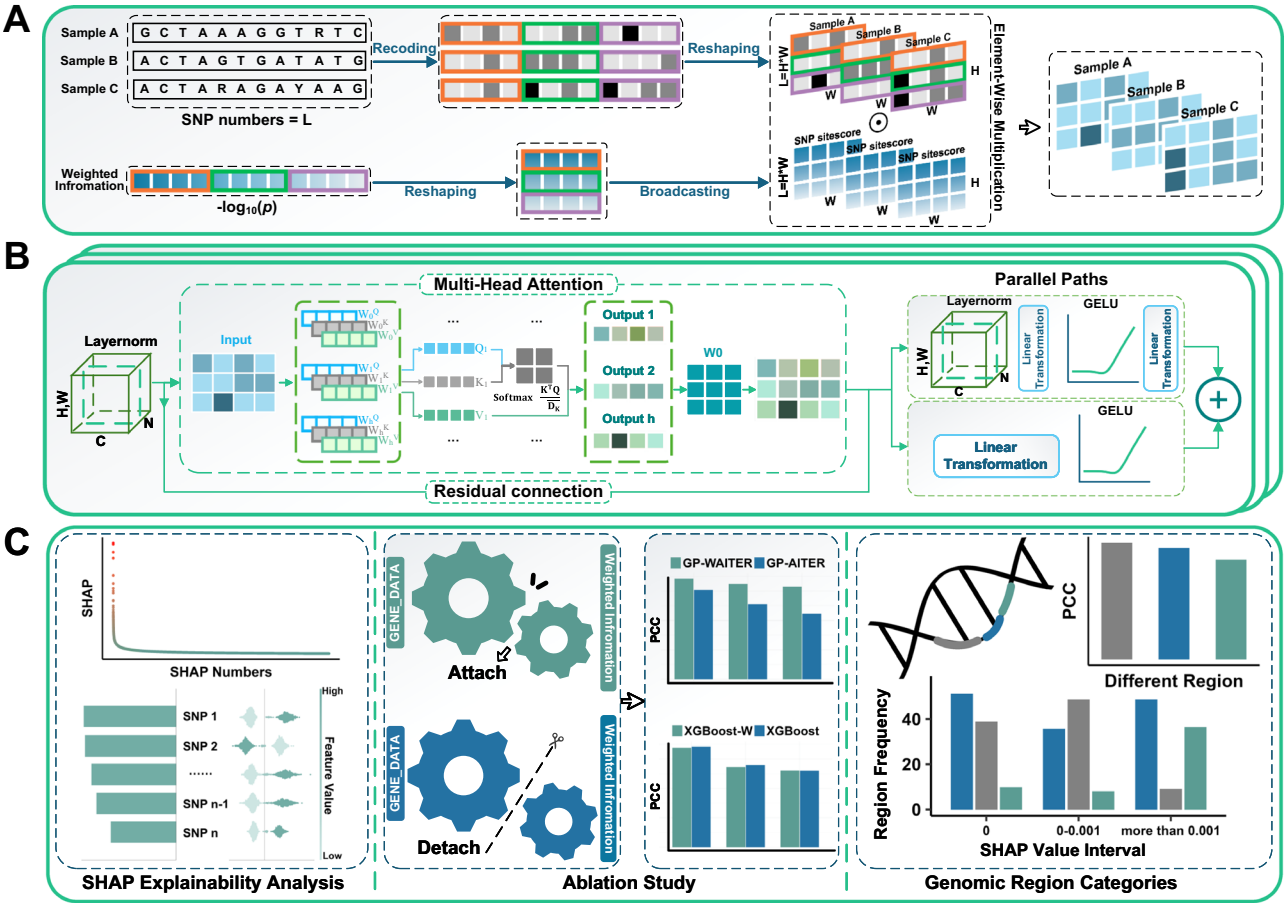

Fig. 1 | Overview of the GP-WAITER model for complex trait analysis in crops. **A** Input preprocessing for genomic data and SNP weights in the weighted embedding block. Genomic data and SNP weight vector derived from genome-wide association analysis are transformed into a 3D tensor and a 2D matrix, respectively. A Hadamard product operation integrates weight information into the genotype data before convolutional feature extraction. **B** Transformer-based encoder structure. The weighted features pass through N stacked Transformer blocks, each comprising multi-head self-attention, layer normalization, residual connections, and feed-forward networks with GELU activation, enabling capture of long-range dependencies and non-linear interactions among SNP loci. **C** Model interpretation through causal variant identification, weighted information ablation analysis, and prediction accuracy comparison across genome regions.

using PyTorch (version 1.13.0), and early stopping was applied when the loss stabilized. We adopt the Pearson correlation coefficient, the mean squared error (MSE), the root mean squared error (RMSE), and the mean absolute error (MAE) between normalized observed and predicted phenotypic values as our evaluation metrics. Further implementation details, including dataset partitioning and cross validation strategies, are provided in the Methods.

Performance evaluation of GP-WAITER

To evaluate the GP-WAITER performance across diverse molecular phenotypes, we systematically compared its prediction accuracy with that of seven state-of-the-art genomic prediction approaches. This benchmark suite spans a wide range of model architectures, including a traditional linear method (rrBLUP[4]), three machine learning methods (SVR[7], XGBoost[8] and LightGBM[9]), and three deep learning models (CNN[10], DNNGP[11], Cropformer[12]). The evaluation of the three MLs and three DLs models was performed under their respective optimal hyperparameter configurations, determined through grid search (Supplementary Table 1). All models were evaluated on six well-established datasets, including (1) previous Soybean1861 dataset, containing eight nutritional traits in a natural population of 1,861 accessions across five environments; (2) The Soybean192 dataset, containing four nutritional traits from a soybean recombinant inbred line (RIL) population; (3) The Soybean14460 dataset, including 11 agronomic and nutritional traits from the GRIN-Global database; (4) The Maize244 dataset, consisting of four agronomic traits in a natural population of 244 inbred lines; (5) The Rice529 dataset, covering 10 agronomic traits of 529 accessions that represent both the usefulness in rice improvement and the genetic diversity in the cultivated species; (6) The Wheat406 dataset, containing three agronomic traits evaluated across three environments in a natural population of 406 bread wheat accessions.

To evaluate model stability across diverse datasets, we first compared prediction accuracy based on Pearson correlation analysis, the mean squared error (MSE), the root mean squared error (RMSE), and the mean absolute error (MAE) across all six datasets. GP-WAITER demonstrated strong generalization performance, consistently outperforming the other seven comparative models with improvements in accuracy ranging from 8.9% to 77.5%, corresponds to an absolute accuracy increase of 4.81% to 19.54%, MSE reductions of 63.9% to 95.9%, RMSE reductions of 25.5% to 57%, MAE reductions of 36.7% to 62.5% (Fig. 2A–F, Supplementary Fig. 2A-F, Supplementary Tables 2–6 and Supplementary Data 1–7). On the soybean192 dataset, GP-WAITER demonstrated superior performance. Its accuracy improvements on four key traits reached 67.53% to 97.78% compared to all seven models tested. When directly compared to the widely used rrBLUP, GP-WAITER outperformed it by 40.89% to 103.09% on these same traits, and attained a mean predictive correlation of 0.45, representing a 56.85% increase compared to the mean correlation of 0.29 achieved by rrBLUP (Fig. 2B, Supplementary Table 2 and Supplementary Data 3).

https://doi.org/10.1038/s41467-026-71035-5

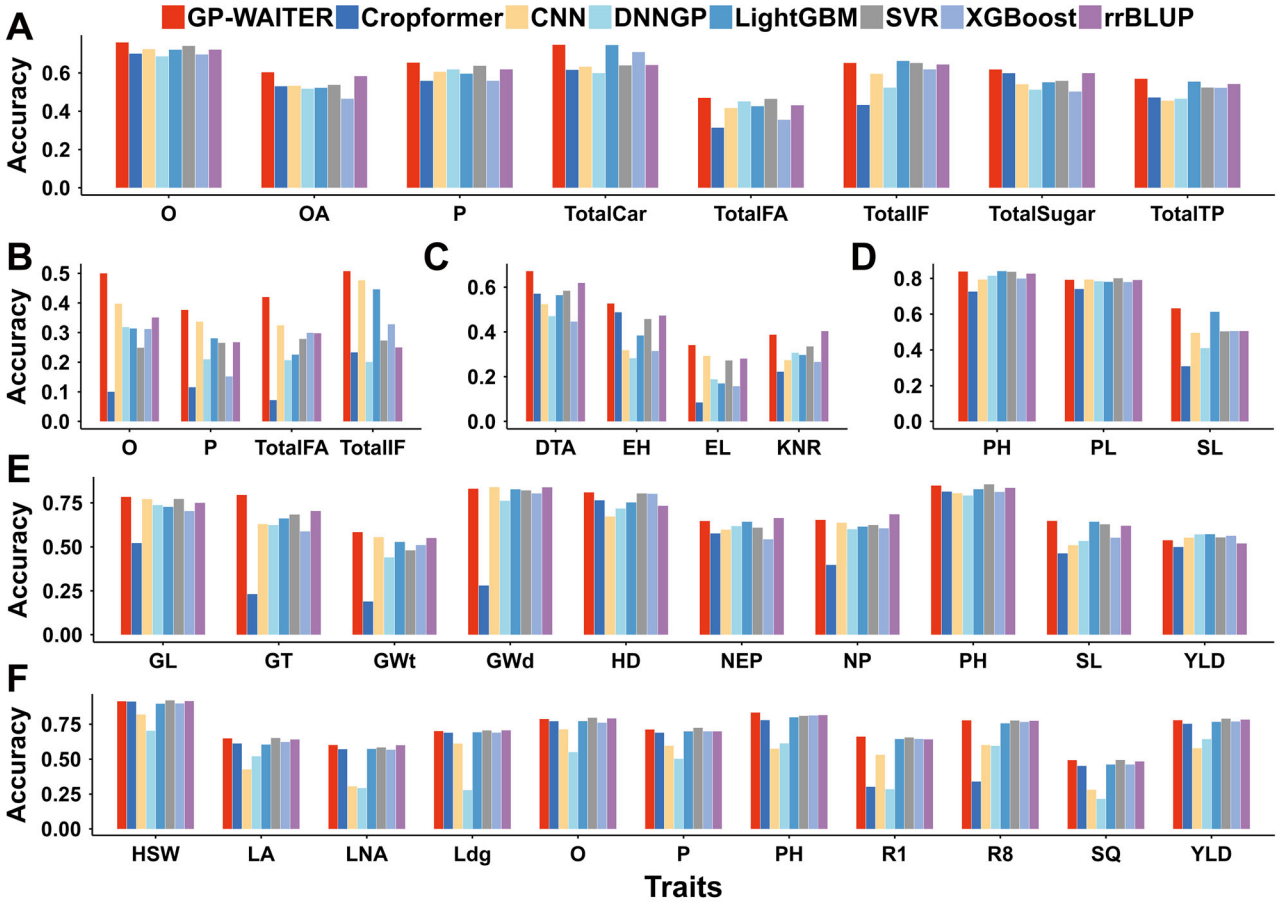

Fig. 2 | Performance evaluation of genomic prediction models across diverse crop datasets. Comparative analysis of the average prediction accuracy (Pearson correlation coefficient, *r*) for eight genomic prediction models evaluated across six distinct datasets: (**A**) soybean1861, (**B**) soybean192, (**C**) maize244, (**D**) wheat406, (**E**) rice529, and (**F**) soybean14460. O oil, P protein, OA oleic acid, TotalIF total isoflavone, TotalFA total folate acid, TotalSugar total sugar, TotalTP total tocopherol, TotalCar total carotenoid, DTA days to anthesis, EH ear height, EL ear length, KNR kernel number per row, R1 Flowering date, R8 Maturity date, PH Plant Height, HSW hundred-seed weight, YLD Yield, Ldg Lodging, SQ Seed quality, LA Linoleic acid, LNA Linolenic acid, HD heading date, NP number of panicles, NEP number of effective panicles, GWt grain weight, SL spikelet length, GL grain length, GWd grain width, GT grain thickness, PL peduncle length. Source data are provided as a Source Data file.

When compared specifically to different modeling approaches, GP-WAITER surpassed deep learning models by 9.8% to 124.6% in accuracy, reduced MSE by 77.3% to 93.9%, reduced RMSE by 34.9% to 75.6%, reduced MAE by 53.8% to 79.7%. Against conventional machine learning models, it improved accuracy by 2% to 40.8%, and compared to the traditional linear model rrBLUP, accuracy gains ranged from 0.6% to 56.9% (Fig. 2A–F, Supplementary Fig. 2A–F, Supplementary Tables 2–6 and Supplementary Data 1–7). It is also noteworthy that the large-sample dataset soybean14460 exhibited consistently strong prediction performance with minimal variability across models, whereas CNN and DNNGP models performed poorly in comparison (Fig. 2F, Supplementary Fig. 2F, Supplementary Table 4 and Supplementary Data 5). These results collectively established the robust accuracy of GP-WAITER in predicting complex traits.

We then systematically evaluated its computational efficiency and memory usage across datasets of varying scales. Conventional machine learning methods, such as rrBLUP, SVR, XGBoost, and LightGBM, typically operate on CPUs. While these models exhibit rapid execution on small to moderate datasets (typically under 400 s). However, as data size increases, the computational time for rrBLUP rises substantially on the Soybean14460 dataset, reaching up to 7766 s (Supplementary Table 7). In contrast, deep learning (DL) architectures (CNN, DNNGP, GP-WAITER, and Cropformer) are characterized by massive parameter spaces and high epoch counts, necessitating GPU-accelerated parallel processing for feasible training. Within this GPU-based category, performance varies significantly. Although Cropformer and DNNGP utilize dimensionality reduction prior to training, Cropformer and DNNGP's efficiency gains are limited to small-scale data, whereas maintains high computational overhead across large dataset. On small to medium-scale datasets such as Maize and Rice, GP-WAITER delivers runtimes comparable to CNN. The advantage of GP-WAITER is most pronounced on the high-dimensional Soybean14460 dataset (~ 574 million data points). GP-WAITER completes training in only 4,216 s, achieving a 1.8-fold speed-up over DNNGP (7,552 s) and a 2.4-fold speed-up over Cropformer (10,049 s) (Supplementary Table 7). Furthermore, GP-WAITER demonstrates exceptional GPU memory optimization, requiring a peak of only 536 MB, whereas Cropformer and DNNGP demand 1134 MB and 1668 MB, respectively (Supplementary Table 8). These results highlight the superior scalability and practical efficiency of GP-WAITER, establishing it as a highly robust solution for large-scale genomic prediction tasks.

Interpretation of GP-WAITER to facilitate genomic breeding in crops

To better understand how input features may influence model predictions and identify the most critical features for phenotype discrimination, we utilized SHAP, an advanced explainable artificial intelligence (XAI) method (Fig. 1C). After obtaining the top 20 genomic

https://doi.org/10.1038/s41467-026-71035-5

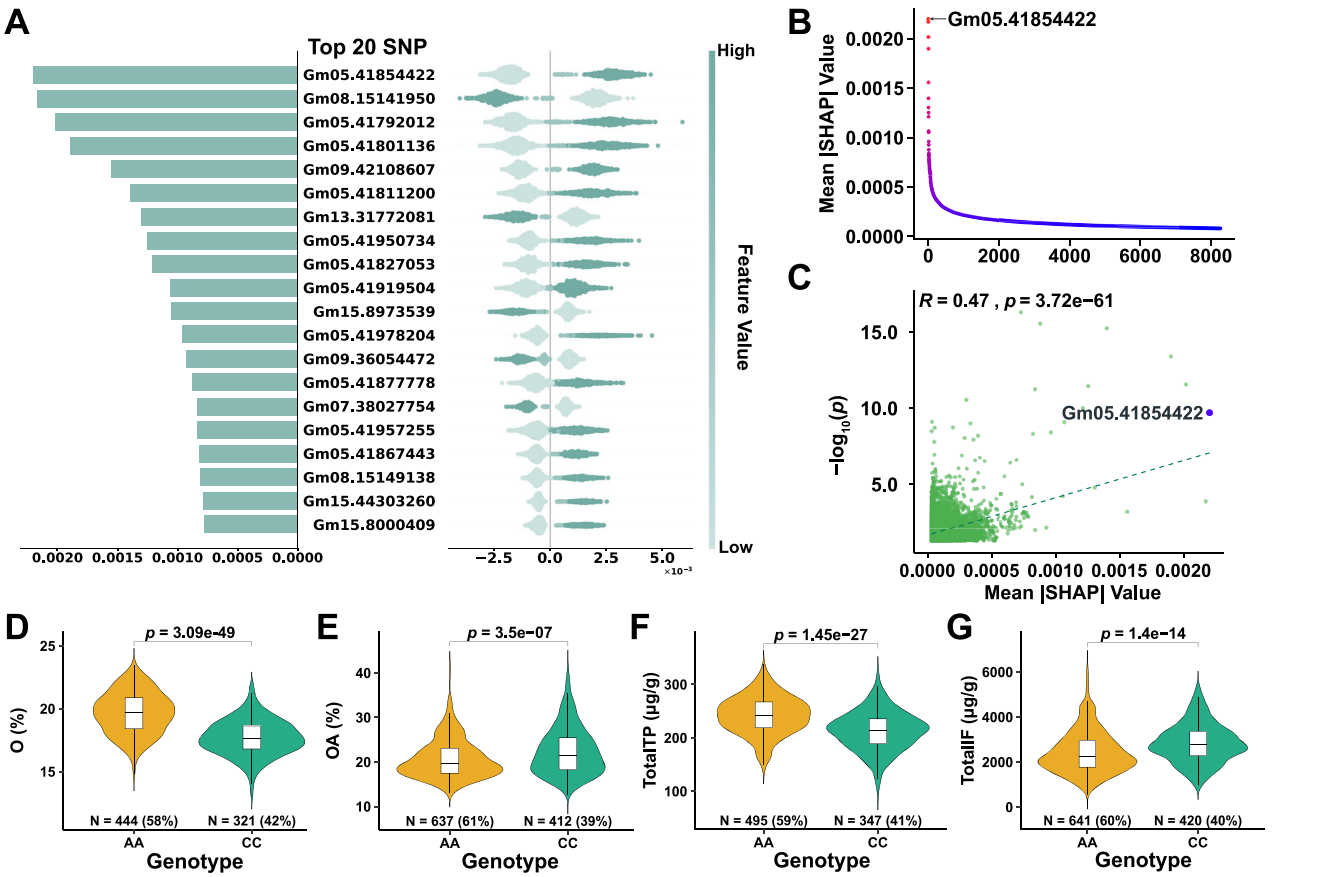

Fig. 3 | Interpretation of the GP-WAITER model and application to total isoflavone prediction. A Bar plot of the top 20 most important genomic features and SHAP beeswarm plot showing the distribution of their influence on predictions. **B** Trait-relevant genes ranked by SHAP values across individuals. **C** Correlation of SHAP-ranked trait-relevant genes with GWAS results for TotalIF; the top-ranked variant of *MFT*, is labeled. Statistical significance was assessed using a Pearson correlation test (two-sided; 95% confidence intervals). Violin plots showing *MFT* variation effects on O (**D**), OA (**E**), TotalTP (**F**), and TotalIF (**G**). The number of accessions (**N**) for each haplotype category and their corresponding proportions are indicated within the plots. In each violin plot, the centre line represents the median, the box limits indicate the interquartile range (IQR), and the whiskers extend to the minima and maxima. The width of the violin represents the data density. Statistical comparisons between haplotype groups were performed using a two-sided *t*-test. O oil, P protein, OA oleic acid, TotalIF total isoflavone, TotalFA total folate acid. Source data are provided as a Source Data file.

features for importance scores (absolute SHAP values) and effect directions (positive/negative) across eight traits in soybean1861, we found that 29 genes harbored variants with potential functional impacts, including splicing alterations, missense mutations, stop-gain, and stop-loss variants (Supplementary Data 8). To further elucidate whether these high-SHAP-value genes are functionally related to the predicted traits, we performed Gene Ontology (GO) enrichment analysis on the genes identified by SHAP analysis for each trait. The results revealed significant functional enrichments that align closely with the biological contexts of the respective traits (Supplementary Data 9 and Supplementary Figs. 3–10), the strongest associations were observed for tocopherols with phenolic compounds and vitamin E metabolism ($p = 6.5e\text{-}07$) (Supplementary Fig. 3), carotenoids with floral development ($p = 0.00044$) (Supplementary Fig. 4), and isoflavones with light response ($p = 8.7e\text{-}05$) (Supplementary Fig. 5). Other traits showed more generalized enrichment patterns, including stress responses and developmental regulation (Supplementary Figs. 6–10), which may reflect pleiotropic effects or shared regulatory networks rather than direct metabolic pathway involvement.

SHAP analysis identified the missense variant, Gm05.41854422, as the top-ranked feature for total isoflavone (Fig. 3A, B). This variant was also significantly associated with total isoflavone in GWAS (Fig. 3C and Supplementary Fig. 11), and was also present among the top 20 features associated with oil (ranked 4th), oleic acid (ranked 4th), and total tocopherol (ranked 7th) (Supplementary Data 8). As this variant resides in the oil and isoflavone content-related gene, *MFT* (Glyma.05G244100), encoding a *Mother of FT* (*Flowering Locus T*)-like protein[32–34], we further investigated the functional link between this variant and phenotypic variation. To this end, we stratified 1,861 accessions into two haplotype groups, MFT-AA (495 accessions) and MFT-CC (347 accessions), based on Gm05.41854422 alleles. We found that all four traits significantly differed between haplotype groups (Fig. 3D–G), aligning well with previous reports of these alleles contributing to oil and total isoflavone phenotype[32–34], and further supporting the pleiotropy of *MFT* in co-modulating these nutritional traits of oleic acid and total tocopherol.

Subsequent examination of SHAP values and GWAS signals for total carotenoid revealed that the top-ranked SNP, Gm01.53229579 (stop-loss) (Supplementary Fig. 12A–C), was located within a known quantitative trait locus (QTL) associated with the *stay-green G* gene (Glyma.01G198500), encoding a CAAX amino-terminal protease identified to regulate seed dormancy, and which reportedly exhibits evidence of parallel selection during soybean domestication[35]. Haplotype analysis of Gm01.53229579 in the soybean1861 population also identified two distinct groups, consisting of 726 (G-AA) and 246 (G-GG) lines, respectively. We observed that total carotenoid levels

significantly differed (p = 5.25e-44, t-test) between the two haplotype groups, with the GG allele associated with higher total carotenoid (Supplementary Fig. 12D), which was consistent with previous variant analysis and further reinforcing the functional relevance of this locus[35]. The variant's significant association with total carotenoid levels was also confirmed in our GWAS (Supplementary Fig. 13). In parallel, the SHAP interpretation highlighted a candidate locus, Gm08.8472159 (an upstream gene variant), which ranked fifth in feature importance of total carotenoid but was not detected in the conventional GWAS due to its modest effect size or epistatic interactions (Supplementary Data 8). This SNP is located near Glyma.08G110300, encoding chalcone synthase, a key enzyme in plant secondary metabolism, particularly in the flavonoid and anthocyanin biosynthetic pathways, the total carotenoid content of materials with TT is significant higher than CC (p = 1.68e-46, t-test) (Supplementary Fig. 14). Taken together, these results indicated that SHAP interpretation of GP-WAITER features could not only effectively capture trait-associated SNPs significantly contributing to phenotypic variation in soybean, but also identifies candidate loci beyond the detection limit of conventional GWAS, thereby providing a more comprehensive genetic dissection of complex traits.

Evaluation of factors affecting GP-WAITER performance

To investigate factors potentially affecting the prediction accuracy of our Transformer architecture, we first examined SNP-based heritability and functional regions. As expected, higher heritability scores for oil and total carotenoid (0.88 and 0.85) generally benefited prediction accuracy (r = 0.76 and 0.75, respectively) provided by leveraging the GP-WAITER accuracy, while prediction accuracy was the lowest for total folate acid (0.47), which also had the lowest heritability (0.62). Further analysis for eight traits indicated that heritability was positively correlated with prediction accuracy (r^2 = 0.58) (Fig. 4A). To next determine how specific genomic regions might affect prediction accuracy, we classified the genome-wide soybean variants into two main categories, regulatory region and genic region, in the Soybean1861 population based on their functional and structural characteristics. Applying GP-WAITER to predict eight traits across five environments based on SNPs from regulatory regions, genic regions, or the whole genome, we found that prediction accuracy showed a relatively minor, gradual decrease along with the number of SNPs selected by GP-WAITER. That is, accuracy was the highest in predictions based on the genome-wide SNP set (826,020 SNPs; r = 0.47–0.76), followed by regulatory region SNPs (295,075 SNPs; r = 0.36–0.71), and the lowest using only genic region SNPs (66,864 SNPs; r = 0.31–0.70) across the eight nutritional traits. In particular, prediction accuracy for total folate acid and total carotenoid improved by 32% and 10% using whole-genome SNPs compared to regulatory SNPs (Fig. 4B and Supplementary Data 10).

To identify whether specific regions or types of SNPs provide greater influence on model accuracy, we categorized the genome-wide SNPs as either regulatory, intergenic, genic, intron, synonymous, nonsynonymous, STOP, or splice, then predicted oil contents in the Soybean1861 population, spanning five environments, based on these SNP subsets as features (Fig. 1C). The prediction accuracy varied across different genomic regions, with regulatory regions showing relatively higher accuracy (average 0.71) compared to intergenic regions (average 0.69), while the difference was more pronounced in genic regions, nonsynonymous ans synonymous exhibited the highest prediction accuracy (both at 0.69), splice sites showed the lowest (0.44) (Supplementary Fig. 15 and Supplementary Data 11). Random sampling of SNP subsets across a gradient of sizes from the whole genome revealed the prediction accuracy for intergenic regions reached 0.71, exceeding that of the real regions. Conversely, SNP subsets from nonsynonymous and synonymous regions yielded accuracies roughly 2% below those of the actual genomic sites. Most strikingly, for splice regions, the randomly sampled subsets outperformed the actual regions by a margin of 3.3% in prediction accuracy (Supplementary Data 11). We then calculated the proportion of specific regions to assess whether region size influenced prediction accuracy. The result revealed that categories with more SNPs generally resulted in higher prediction accuracy, gradually declining along with SNP number in each category (Supplementary Fig. 16 and Supplementary Data 11). To further investigate whether variants with higher SHAP values were clustered in specific genomic regions, we analyzed their distribution across different SHAP value intervals. More than 90% of variants with high SHAP values were enriched in regulatory and genic regions (Fig. 4C and Supplementary Table 9). This finding strongly suggests that functional elements in these regions (particularly regulatory and coding sequences) make substantial contributions to the predictive performance for soybean complex traits. Together, these results highlight the importance of functional genomic regions, particularly regulatory and genic elements, in driving model accuracy and interpretability.

Contributions of weighted information to prediction accuracy

To investigate the effects of SNP weighting on prediction accuracy, we performed an ablation study comparing weighted (GP-WAITER) and unweighted (GP-AITER) models to define the impacts on phenotype prediction (Fig. 1C). The prediction accuracy was improved to 0.64 for GP-WAITER, representing a 7.9% increase compared to the 0.59 achieved by GP-AITER. This improvement was particularly noteworthy in protein, oleic acid, total isoflavone, total folate acid, total sugar, total tocopherol, and total carotenoid predictions, increasing by 5.6%, 8.4%, 4.3%, 14.9%, 15.1%, 7.5% and 12.9%, respectively (Fig. 4D and Supplementary Data 12). Furthermore, the integration of environmental markers associated with genomic variants was explored as a potential strategy to further improve prediction accuracy[36]. Environmental covariates were extracted, and environmental genome-wide association studies (envGWAS) were performed using the soybean1861 dataset to identify loci adaptive to environmental conditions. For the traits of oil and protein content, the conventional weighting scheme within the GP-WAITER model was substituted with loci prioritized by envGWAS. This approach yielded prediction accuracy (average r = 0.69), which was lower than that of the original GP-WAITER model (average r = 0.71) and did not significantly surpass the baseline rrBLUP model (average r = 0.67) (Supplementary Table 10), which suggest that integration of envGWAS-identified variants as environmental covariates offers limited additional predictive benefit for these traits. Overall, these findings demonstrate that weighted information can be informative in predicting complex traits with our framework, as it effectively leverages these weights to integrate and prioritize informative features.

To access whether weighted information could improve prediction accuracy in other models, we incorporated SNP locus weights into the XGBoost framework to create XGBoost_W, and compared its performance with the standard XGBoost model across multiple traits (Fig. 1C). Comparative analysis revealed nearly identical prediction accuracy between XGBoost_W and standard XGBoost (average accuracy: 0.57 vs 0.56, respectively), with minimal differences across traits (maximum difference of 0.0037 observed for total sugar content) (Supplementary Fig. 17). In addition, GP-WAITER demonstrated stronger performance (average r = 0.64) than other machine learning models, even in the absence of weight information, as GP-AITER (0.59) outperformed XGBoost (0.55), CNN (0.56), DNNGP (0.55), and Cropformer (0.54), respectively (Fig. 4D and Supplementary Data 1). Although GP-AITER achieved comparable average prediction accuracy to that of rrBLUP, LightGBM, and SVR, our unweighted model exhibited superior performance in specific scenarios, such as oil content in Beijing 2017 and Hainan 2018, protein content in Beijing 2017, oleic acid in Beijing 2017, total tocopherol in Hainan 2017 (Supplementary Data 12). To further evaluate GP-WAITER performance in different

https://doi.org/10.1038/s41467-026-71035-5

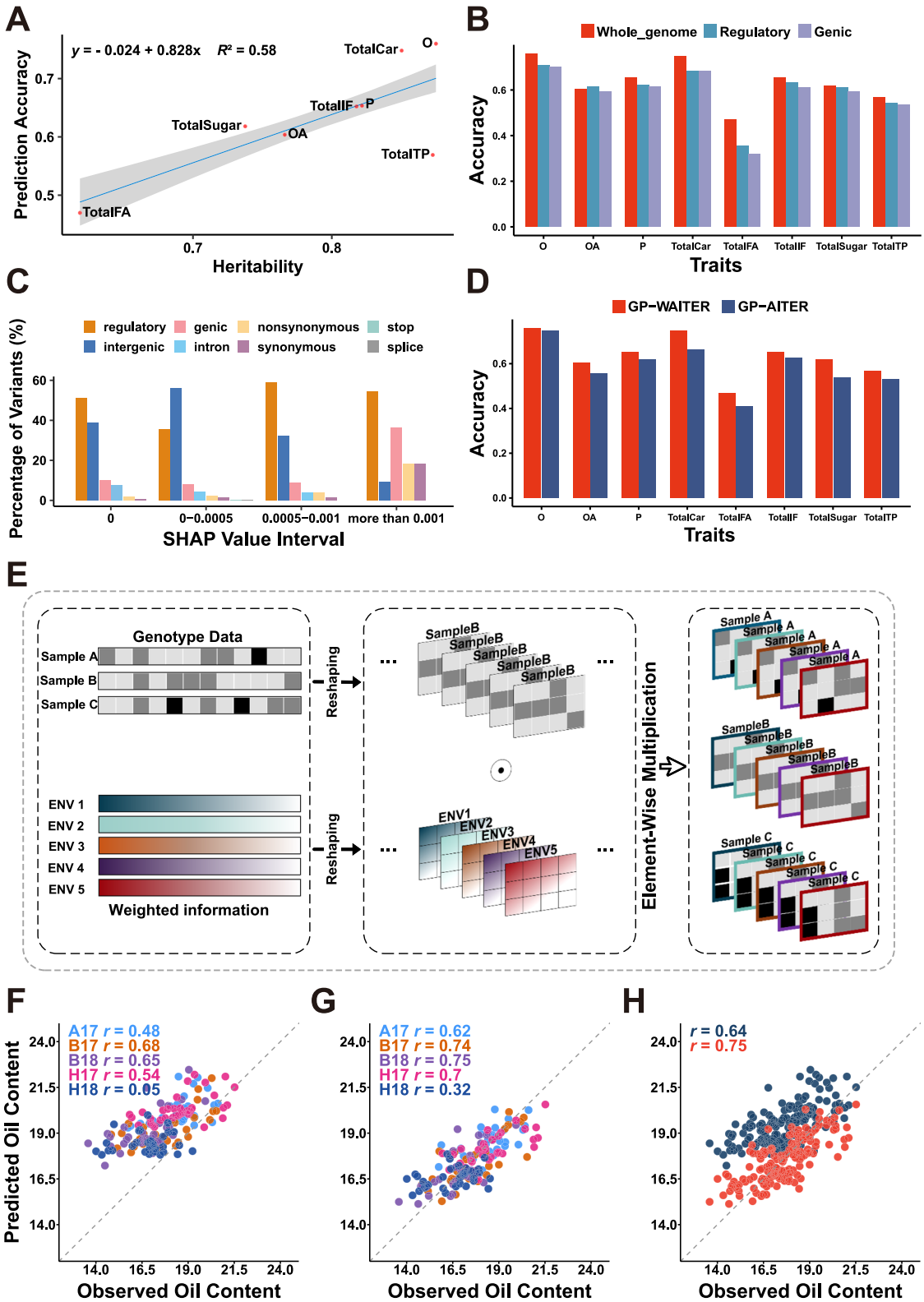

Fig. 4 | Contributions of heritability, functional categories, and weighted information to model performance. A Correlation between heritability and prediction accuracy in soybean traits. The blue solid line represents the linear regression line (fitted values), with the shaded area indicating the 95% confidence interval for the mean prediction. **B** Prediction accuracy for eight traits using different genomic regions. **C** Distribution of genomic regions across SHAP value intervals for TotalIF. **D** Model performance comparison between GP_WAITER and GP_AITER. **E** Architecture of the input preprocessing for the weighted multi environment model. **F** Baseline model (GP_AITER) performance across environments. **G** Weighted model (GP_WAITER) performance across environments. **H** Integrated comparison of baseline GP_AITER (blue) and weighted GP_WAITER (red) multi-environment models. O oil, P protein, OA oleic acid, TotalIF total isoflavone, TotalFA total folate acid, TotalSugar total sugar, TotalTP total tocopherol, TotalCar total carotenoid. Source data are provided as a Source Data file.

environments, we expanded the model for multi-environment prediction by incorporating association results from five distinct environments (Anhui 2017, Beijing 2017, Beijing 2018, Hainan 2017, and Hainan 2018) as weighted information (Fig. 4E). We evaluate the prediction accuracy in the scenario of untested genotypes in tested environments. The enhanced GP-WAITER (multi-environment weighted) model achieved a prediction accuracy of 0.75, outperforming the baseline multi-environment model lacking association-based weighting ($r = 0.64$) (Fig. 4F–H). These results suggest that SNP weighting alone does not universally improve performance, and the enhanced accuracy observed in our primary model likely results from the synergistic integration of weighted information with the attention mechanism inherent in the Transformer architecture.

Discussion

Here, we present GP-WAITER, a Transformer-based method for predicting crop phenotype across environments that leverages GWAS summary statistics of complete SNP sets for scalable and accurate prediction tasks with large-scale genomic data. Our approach introduces three key innovations: (1) A weighted embedding layer that tokenizes input data, then transforms discrete data into continuous vectors using locus weight information to effectively capture long-range dependencies in genomic sequences while maintaining computational efficiency. Moreover, our analyses demonstrate that weighted information embedded in sequencing data can be learned and utilized by the model to better estimate phenotypes for traits, with GP-WAITER outperforming its unweighted counterpart (GP-AITER) (Fig. 4D and Supplementary Data 12), and providing a 0.11 gain in accuracy in multi-environment modeling (Fig. 4H). (2) Built with a multi-head attention framework, GP-WAITER preserves shallow feature significance while modeling complex interactions through parallel processing of diverse relationship patterns[37,38]. This architecture incorporates residual connections to address vanishing gradients, employs regularization techniques to prevent overfitting, and uses batch normalization to ensure stable convergence, leverages the complementary strengths of tanh and GELU across the hybrid architecture to effectively balances training stability and expressive power (Supplementary Fig. 1), collectively enhancing the model's generalization capacity, as demonstrated by its average accuracy by 27.2% (+0.1 correction), with maximum gains reaching 77.5% (an absolute increase of 0.195), and slashing MSE, RMSE, and MAE by 77.5%, 45.5%, and 50.6%, respectively (Fig. 2, Supplementary Fig. 2, Supplementary Tables 2–6 and Supplementary Data 1–7). (3) Combining CNNs with Transformer mechanisms allows simultaneous modeling of both local and global genomic information. CNN demonstrate particular efficacy in breeding populations characterized by minimal population stratification (e.g., RIL population, Soybean192; Fig. 2B and Supplementary Fig. 2B), where convolutional layers efficiently capture localized features such as epistatic interactions and haplotype blocks. In contrast, models like CNN, DNNGP, and Cropformer exhibit suboptimal performance on the Soybean14460 dataset (Fig. 2F and Supplementary Fig. 2F). This is primarily attributed to their inability to fully encompass the complex spectrum of genetic relationships inherent in diverse populations; furthermore, the restricted SNP selection in these models may fail to capture the full breadth of genetic variation contributing to complex traits. Conversely, GP-WAITER has more stable and encouraging performance on all datasets, which advances prediction accuracy across both designed breeding populations and diverse natural populations, offering a versatile and robust framework for genomic prediction.

GP-WAITER effectively balances computational speed and resource utilization, offering a robust and scalable solution for high-dimensional genomic selection across diverse data scales. Benchmarking reveals that while the efficiency of rrBLUP is primarily constrained by sample size, it remains relatively insensitive to SNP number. Conversely, although LightGBM generally exhibits higher computational overhead than other conventional methods on smaller sets, it demonstrates robustness to expanding sample sizes, completing the analysis of the Soybean14460 dataset with high efficiency (Supplementary Table 7). Deep learning (DL) architectures typically require iterative training leading to longer runtimes than linear models, while GP-WAITER achieves superior execution speeds and significantly lower GPU memory footprints compared to its DL counterparts. This is particularly evident in large-scale datasets (Supplementary Tables 7 and 8), where many existing models rely heavily on aggressive dimensionality reduction to remain computationally feasible or to mitigate overfitting. GP-WAITER's architectural efficiency allows for the seamless training of massive datasets on a single NVIDIA GeForce RTX 3080 GPU, democratizing high-performance GS. The core of this efficiency lies in its structural innovation: rather than processing one-dimensional sequential data, GP-WAITER reshapes genomic inputs into a two-dimensional format (Fig. 1A). This transformation significantly reduces algorithmic complexity and bypasses the scalability bottlenecks typical of ultra-long sequence processing on standard hardware. The resulting 3D tensor representation aligns natively with the parallel processing architecture of GPUs, maximizing hardware utilization. Furthermore, the strategic integration of convolutional layers with BatchNorm (BN) (Supplementary Fig. 1) stabilizes the internal distribution of data, mitigates covariate shift, and alleviates gradient vanishing. This synergy not only accelerates convergence but also reduces sensitivity to weight initialization and hyperparameter selection, substantially enhancing the model's practical deployability and robustness in real-world breeding programs.

Our model's strong performance and robust generalization properties suggest it may serve as a particularly valuable tool for molecular breeding programs requiring high precision in data-driven decision-making, where accurate phenotypic prediction and reliable genomic analyses are essential for success and efficiency. Deployment of GP-WAITER can enable breeders to systematically identify optimal parental lines, design superior hybrid combinations, and accelerate cultivar development through enhanced genetic gains and shorter breeding cycles. The incorporation of biological prior knowledge into the design of these models could help guide the feature extraction process and lead to more interpretable models. SHAP analysis of the Transformer-based model revealed biologically significant patterns that enabled identification of key predictive features (Supplementary Data 9), while SHAP analysis uncovered the disproportionate influence of functional regions, especially regulatory and promoter region SNPs, on model accuracy. This pattern of selective attention of multiple attention heads towards specific promoter variants (Fig. 4C and Supplementary Data 11) aligns well with established principles of explainability in Transformer-based models[39] and is strongly supported by recent functional genomics studies[40,41]. Our approach thus transforms conventional black-box predictions into biologically grounded insights, consequently enabling a clearer understanding of causal relationships between genotype and phenotype, facilitating identification of biologically meaningful predictive features, and allowing effective translation of computational findings into candidate gene discovery for precision breeding applications.

In view of these capabilities for handling ever-growing complexity and high-dimensional data and addressing non-linear problems, future genomic prediction studies should prioritize three critical advances that are necessary to revolutionize precision plant breeding. First, real-time adaptive models incorporating sparse Transformer architectures should be developed to enable continuous learning from emerging data. Second, fusion frameworks that systematically integrate functional genomic annotations and other biological priors with deep learning should be established to enhance both predictive accuracy and breeding applicability. Third, robust discovery-validation pipelines combining computational predictions with experimental

verification should be implemented to refine GWAS mappings while providing a biological basis for model interpretation. Such advances will pave new avenues for transformative innovation in precision crop breeding.

Methods

Datasets used for genomic prediction

The soybean1861 dataset consisted of 1,861 soybean accessions selected from a global germplasm collection, encompassing broad geographic and genetic diversity[42]. This panel included 218 wild soybean accessions, 1,323 landraces, and 320 cultivated varieties. Field trials were conducted in 2017 and 2018 across three locations: Changping, Beijing (40°13′N, 116°12′E); Sanya, Hainan (18°24′N, 109°5′E); and Hefei, Anhui (33°61′N, 117°E). A randomized block design with three replications was employed. Each row measured 3.0 meters in length with a spacing of 0.5 meters between rows and 0.1 meters between plants[43]. Soybean seeds were harvested at maturity, immediately threshed, and stored at −4 °C before trait analysis. Whole-genome resequencing was performed on all 1,861 accessions using the Illumina HiSeq X platform, yielding a total of 16.41 Tb of sequencing data at an average depth of 5.6x per accession. After filtering, 826,020 high-confidence variant sites were retained for downstream analysis. Nutritional quality traits were quantified using multiple high-throughput analytical platforms. Protein (P) and oil (O) contents were determined using a Fourier Transform Near-Infrared Spectrometer (MATRIX-1, Bruker, Germany). Oleic acid contents (OA) were measured via gas chromatography (GC)[44], and isoflavone content (TotalIF) via high-performance liquid chromatography (HPLC)[45], tocopherol levels (TotalTP) were assessed using HPLC[46], soluble sugar (TotalSugar) via ultra-high-performance liquid chromatography with refractive index detection (UPLC-RID)[47], folate acid content (TotalFA) via HPLC-MS/MS[48], and carotenoid and chlorophyll (TotalCar) via HPLC-UV-Vis[49].

The soybean192 dataset, comprisint 192 $F_{7:8}$ recombinant inbred line (RILs) derived from a cross between Zhonghuang 35 and Zhonghuang 13, was used to assess the performance of genomic prediction on biparental populations. Field experiments were conducted in 2020 at Changping (40°13′N, 116°12′E) and Shunyi (40°13′N, 116°34′E), and in 2021 at Nankou (40°13′N, 116°06′E) and Shunyi (40°13′N, 116°34′E)[50]. Four nutritional quality traits were evaluated: oil content (O), protein content (P), folate acid content (TotalFA), and isoflavone content (TotalIF). Analytical methods for these traits were consistent with those described above for the soybean1861 panel. Genotypic yielded 1,355,959 high-quality SNPs.

The soybean14460 dataset developed by an SoySNP50K Illumina Infinium II BeadChip was acquired from a publicly available repository (https://github.com/IndigoFloyd/SoybeanWebsite/tree/main). It includes 14,460 soybean accessions genotyped with 39,707 SNP markers. The dataset provides phenotypic measurements for 11 key traits: Flowering date (R1), Maturity date (R8), Plant Height (PH), hundred-seed weight (HSW), Yield (YLD), Lodging (Ldg), Seed quality (SQ), Oil (O), Protein (P), Linoleic acid (LA), Linolenic acid (LNA).

The maize244 dataset consisted of 244 inbred lines drawn from a natural population representing nine major subgroups to ensure broad genetic coverage[51]. Field trials were conducted in 2022 and 2023 at the Institute of Crop Science, Chinese Academy of Agricultural Sciences (Shunyi, Beijing; 40°N, 116°28′E), and at the Mishan Experimental Station, Bayi Agricultural University (Mishan, Heilongjiang; 131°87′N, 45°54′E). Four agronomic traits were measured: days to anthesis (DTA), ear length (EL), ear length (EL), and kernel number per row (KNR). A total of 308,136 high-quality SNPs were retained.

The rice529 dataset was sourced from the RiceVarmap database (https://ricevarmap.ncpgr.cn/download/). It consists of 529 rice accessions genotyped with 669,384 high-quality single nucleotide polymorphism (SNP) markers using the Illumina HiSeq 2000. All sequencing data were mapped to rice reference genome (Nipponbare, MSU version 6.1). The dataset includes phenotypic records for the following 10 agronomic traits: heading date (HD), plant height (PH), number of panicles (NP), number of effective panicles (NEP), grain yield (YLD), grain weight (GWt), spikelet length (SL), grain length (GL), grain width (GWd), and grain thickness (GT).

The wheat406 dataset was obtained from the Plant Wheat Genotype and Phenotype Database (PWGBD; http://resource.iwheat.net/PWGBD/). It comprises 406 bread wheat accessions genotyped with 234,219 SNP markers. All sequencing data were mapped to IWGSC RefSeq v1.0. Phenotypic data were collected for three traits: peduncle length (PL), plant height (PH), and spike length (SL). Each traits was measured across three independent environments (E1, E2, and E3), E1, normal sowing at Sichuan province (30.63′N, 103.67′E) in 2018; E2, normal sowing at Shaanxi Province (34.28′N, 108.07′E) in 2018; E3, late sowing at Shaanxi Province (34.28′N, 108.07′E) in 2018.

Genotypic data

Genotypic data across all datasets comprised biallelic single-nucleotide polymorphisms (SNPs). Marker data underwent standardized quality control and imputation prior to downstream analyses. Quality control was performed using PLINK v1.9[52]. SNPs with a minor allele frequency (MAF) < 0.01, a missing rate > 0.2, and those that were multi-allelic were filtered out. Remaining missing genotypes were imputed using Beagle 5.4 program (version 22Jul22.46e)[53]. To mitigate the effects of linkage disequilibrium (LD) and reduce redundancy among correlated markers for soybean1861 dataset, LD-based pruning was conducted in PLINK v1.9 with the parameters -indep-pairwise 50 5 0.4 (a window size of 50 SNPs, a step of 5 SNPs, and an r^2 threshold of 0.4).

SNP genotypes were encoded numerically to represent allele configurations: homozygous reference alleles were assigned a value of 1, homozygous alternate alleles as −1, and heterozygous genotypes as 0. The reference allele ('A') was defined as either the designated reference base from the source database or, when unspecified, the most frequent allele within the respective population. Genotypic data corresponding to the phenotypic samples were extracted and formatted as a two-dimensional matrix, with rows denoting individual samples and columns representing SNP loci.

Weighted information

GWAS analyses were conducted using the EMMAX linear mixed model framework for all six datasets. Population structure was accounted for by including the first three principal components, derived from principal component analysis (PCA), as covariates. SNP weights were extracted based on GWAS p-values and aligned to match the genotypic data matrix. A log(x) transformation was applied to stabilize variance and enhance interpretability. The weight Wi is formalized as follows:

$$Wi = -\log_{10}(2 \cdot (1 - \Phi|\hat{\beta}i/\sqrt{\mathrm{Var}(\hat{\beta}i)}|)) \quad (1)$$

where Wi is the weighted information for the i-th SNP; $\hat{\beta}i$ is the effect size for the i-th SNP, Φ is the cumulative distribution function of the standard normal distribution.

The resulting weights were stored as a one-dimensional vector with a length equal to the number of SNPs. For weight alignment, we matched SNPs between the GWAS summary statistics and the target genotypic dataset based on their genomic positions. SNPs present in the GWAS results but missing in the target genotype matrix were excluded. Similarly, SNPs with unavailable p-values in the summary statistics were removed. This process ensured that the final weight vector corresponded exactly, in both SNP identity and order, to the columns of the aligned genotype matrix. During cross-validation iterations, weighted data were recalculated separately for each training subset.

Phenotypic data

Phenotypic trait values were obtained from multi-environment field trials and structured into a two-dimensional matrix, with each row corresponding to a genotype and each column representing a specific trait under a given environment.

Environmental data

For each accession in Soybean1861 dataset, environmental data were extracted from extrapolated climate grids at 30 s (~1 km) resolution. Raw data files for monthly long-term average minimum (tmin°C), maximum (tmax), averaged temperature (tavg), solar radiation (srad), wind speed (wind), vapour pressure (vapr), and rainfall (prec) were downloaded from WorldClim (http://worldclim.org/version2) and converted to ESRI grid format for storage and extraction. Curated georeferenced collection site locations were used to extract climate for accessions using the spatial analyst toolbox in ESRI ArcMap 10.6. To evaluate the variants identified by envGWAS, we tested for associations between genetic variation and environmental gradients of GIS data parameters (tmax, tmin, tavg, srad, wind, vapr, prec, lat, lng, height) using latent factor mixed models (LFMM). To select the optimal number of latent factors for this analysis, the mixing coefficients were estimated using Sparse Nonnegative Matrix Factorization (sNMF), choosing the cluster number (K) with the lowest cross-entropy. Parameters for LFMM were set for 10 repetitions, 100,000 iterations, and a burn-in period of 50,000 steps. Merge all the *z*-scores from the runs and recalculate the *p*-values using the genomic inflation factor (λ). The corrected *p*-values were incorporated as environmental variables in the model input.

Overview of GP-WAITER

The GP-WAITER model consists of three primary components: a weighted embedding block, multiple encoder blocks, and a predictor block. The encoder integrates a three-layer Transformer architecture, each containing 27 multi-head attention heads. Layer normalization and residual connections are applied before and after each encoder layer to stabilize training and preserve representational integrity. The predictor block is implemented as a two-layer feed-forward neural network.

In the weighted embedding block, genotype data and SNP weight information are transformed and combined through element-wise multiplication, resulting in a weighted tensor. This tensor is processed by a sequence of convolutional operations, including four two-dimensional convolutional layers, two BatchNorm2d layers (batch normalization), three additional two-dimensional convolution layers, and three tanh activation layers. The input matrix is defined as $I_{N\times M}$, where N is the number of individuals and M is the number of markers. The convolutional operation is defined as follows:

$$O(i,j)=\sum_{m=0}^{M-1}\sum_{n=0}^{N-1} I(i+m,j+n)\cdot K(m,n) \tag{2}$$

Here, $O(i,j)$ represents the output feature map at position (i,j), $I(m,n)$ represents the element of input matrix at position (m,n), and $K(m,n)$ is the element of convolution kernel at position (m,n). M and N are the height and width of the kernel, and i,j and m,n are the spatial indices of the input and kernel, respectively.

Each of the three encoder layers contains a multi-head attention mechanism, two LayerNorm operations, three linear transformation layers, and two GELU activation functions. Two residual connections are incorporated. In the first step, the input data undergoes LayerNorm and multi-head attention, and the attention output is then added back to the input. The second residual connection further processes the result of the first connection using linear transformations and GELU activation. Specifically, one path involves a single linear transformation followed by GELU activation, while the other applies LayerNorm, two linear transformations, and a GELU activation. The outputs of both branches are summed to produce the final result of the encoder layer.

The multi-head self-attention mechanism, central to the Transformer block, dynamically assigns attention weights to each token by computing relevance scores across all input positions. Each token is projected into a query vector (Q), key vector (K), and value vector (V) via linear transformations. Attention weights are computed by:

$$\text{Attention}(Q,K,V)=\text{softmax}\left(\frac{Q\cdot K^{T}}{\sqrt{d_k}}\right)\cdot V \tag{3}$$

where Q, K, and V are matrices of query, key, and value vectors, respectively, and d_k is the dimension of the key vectors. The softmax function normalizes the weights such that they sum to 1. The multi-head mechanism repeats this process in parallel, enabling the model to capture diverse patterns and long-range dependencies within the genomic sequence.

The predictor block includes one LayerNorm layer, a linear transformation layer followed by a GELU activation function, a one-dimensional batch normalization layer (BatchNorm1d), two one-dimensional convolutional layers, and three additional tanh activations. The fully connected transformation is defined as:

$$y=f(Wx+b) \tag{4}$$

where $W\in R^{d_{out}\times d_{in}}$ is the weight matrix, $x\in R^{d_{in}}$ is the input vector, $b\in R^{d_{out}}$ is the bias vector, and $f(\cdot)$ is the tanh activation function. The output of the one-dimensional convolutional operation is given by:

$$O(i)=\sum_{m=0}^{M-1} I(i+m)\cdot K(m) \tag{5}$$

where $O(i)$ denotes the convolution output at position i, $I(i)$ is a one-dimensional input vector, $K(m)$ is the convolution kernel, M is the kernel length, and m is the convolution kernel index.

As the data propagate through fully connected and convolutional layers, dimensionality is progressively reduced, and trait-relevant features are effectively mapped to the final output. After each embedding operation compresses feature representations, attention mechanisms are applied to extract high-level dependencies, thereby enhancing the model's generalization ability.

Training was performed for 200 epochs for each trait on an NVIDIA GeForce RTX 3080 GPU. The model checkpoint with the highest test set performance was retained and saved for evaluation. The GP-WAITER architecture was implemented using PyTorch (version 1.1.3).

Interpretation of GP-WAITER

To interpret the model's predictions and assess the contributions of individual genetic variants, we used SHAP, implemented via the GradientExplainer module from the SHAP library. SHAP values quantify the marginal contribution of each feature, specifically individual SNPs, to the model's output. Features with higher absolute SHAP values are considered more influential in the model's decision-making process for trait prediction.

To ensure robustness, SHAP values were calculated across varying baseline and test set sizes. For each test instance, the SHAP value was estimated by approximating expected gradients over a defined background dataset, thereby providing local feature attributions. For trait classification, a high-specificity threshold corresponding to a probability of 0.90 (approximately 95% specificity) was applied to designate positive SNPs, thereby minimizing false-positive classifications. Based on this threshold, the top 20 SNPs with the highest mean absolute SHAP values were selected for interpretability analysis.

Gene Ontology (GO) enrichment analysis was performed using the agriGO v2.0 toolkit (http://systemsbiology.cau.edu.cn/agriGOv2/), a specialized platform for agricultural organism functional analysis. The genes identified through SHAP analysis as harboring high-importance SNPs were used as the query gene set, with the soybean

(*Glycine max*) Wm82.a2.v1 genome as the reference background. Analysis was conducted separately for each of the eight agronomic traits using the Singular Enrichment Analysis (SEA) tool. GO terms were considered significantly enriched at p-value < 0.05, with the Fisher's exact test employed for statistical evaluation. The analysis covered all three GO categories: Biological Process, Molecular Function, and Cellular Component.

A positive SHAP value indicates that a given SNP increases the predicted trait value when included in the model, whereas a negative value suggests a decrease in predicted trait value. The greater the absolute SHAP value, the more significant the SNP's contribution to the prediction for a specific instance. To visualize feature contributions, both SHAP beeswarm and bar plots were generated. The beeswarm plots facilitated a direct comparison of SHAP values, thereby elucidating the influence of genomic features across different samples. Bar plots were used to highlight the global importance of each SNP, aggregated across all test samples.

Cross-validation schemes

Model performance was evaluated using five cross-validation (CV) schemes, each scheme employed twenty repetitions of 5-fold cross-validation. In each repetition, phenotypic data were partitioned into five subsets. Four subsets were used to train the model, while the remaining subset served as the validation set. This process was iterated such that each subset was used once as the validation set across five folds.

Predictive accuracy was quantified by calculating the Pearson correlation coefficient between the predicted and observed phenotypic values in the validation sets. This metric reflects the model's ability to generalize across unseen data and was adopted as the primary indicator of prediction performance. The prediction precision was assessed using three error metrics: the mean squared error (MSE), the root mean squared error (RMSE), and the mean absolute error (MAE). These were computed between the normalized observed and predicted phenotypic values for each trait across different methods. While MSE and RMSE emphasize larger prediction deviations due to their squared terms, MAE provides a more robust measure of average error magnitude that is less sensitive to outliers.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The genotype and phenotype of soybean1861, soybean192, maize244, wheat406, rice529, and soybean14460 datasets, the environmental data of soybean1861 are deposited on Zenodo [https://zenodo.org/records/18779208]. Source data are provided with this paper.

Code availability

The GP-WAITER scripts are available in the release package on Github [https://github.com/snowo-w/GP-WAITER/] under the Apache License. A specific version (v1.0.0) used for this study has been archived via Zenodo [https://doi.org/10.5281/zenodo.18809685][54].

References

1. Goddard, M. E. & Hayes, B. J. Mapping genes for complex traits in domestic animals and their use in breeding programmes. *Nat. Rev. Genet.* **10**, 381–391 (2009).
2. Xu, Y. et al. Smart breeding driven by big data, artificial intelligence, and integrated genomic-enviromic prediction. *Mol. Plant* **15**, 1664–1695 (2022).
3. Alemu, A. et al. Genomic selection in plant breeding: key factors shaping two decades of progress. *Mol. Plant* **17**, 552–578 (2024).
4. Endelman, J. B. Ridge regression and other kernels for genomic selection with R Package rrBLUP. *Plant Genome* **4**, 250–255 (2011).
5. Zhao, T. et al. Integration of eQTL and machine learning to dissect causal genes with pleiotropic effects in genetic regulation networks of seed cotton yield. *Cell Rep.* **42**, 113111 (2023).
6. Wu, Y. et al. Phylogenomic discovery of deleterious mutations facilitates hybrid potato breeding. *Cell* **186**, 2313–2328 (2023).
7. Long, N. et al. Application of support vector regression to genome-assisted prediction of quantitative traits. *Theor. Appl. Genet.* **123**, 1065–1074 (2011).
8. Chen, T. & Guestrin, C. XGBoost: a scalable tree boosting system. *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. (KDD)* 785–794 (2016).
9. Yan, J. et al. LightGBM: accelerated genomically designed crop breeding through ensemble learning. *Genome Biol.* **22**, 271 (2021).
10. Ma, W. et al. A deep convolutional neural network approach for predicting phenotypes from genotypes. *Planta* **248**, 1307–1318 (2018).
11. Wang, K. et al. DNNGP, a deep neural network-based method for genomic prediction using multi-omics data in plants. *Mol. Plant* **16**, 279–293 (2023).
12. Wang, H. et al. Cropformer: an interpretable deep learning framework for crop genomic prediction. *Plant Commun.* **6**, 101223 (2025).
13. Wu, C. et al. A transformer-based genomic prediction method fused with knowledge-guided module. *Brief. Bioinform.* **25**, 1–11 (2023).
14. Deng, P. et al. DPCformer: an interpretable deep learning model for genomic prediction in crops. arXiv preprint arXiv:2510.08662 (2025).
15. Ma, C. et al. Machine learning–based differential network analysis: a study of stress-responsive transcriptomes in *Arabidopsis*. *Plant Cell* **26**, 520–537 (2014).
16. Abdollahi-Arpanahi, R., Gianola, D. & Peñagaricano, F. Deep learning versus parametric and ensemble methods for genomic prediction of complex phenotypes. *Genet. Sel. Evol.* **52**, 12 (2020).
17. Spindel, J. E. et al. Genome-wide prediction models that incorporate de novo GWAS are a powerful new tool for tropical rice improvement. *Heredity* **116**, 395–408 (2016).
18. Jubair, S. et al. GPTransformer: a transformer-based deep learning method for predicting Fusarium-related traits in barley. *Front. Plant Sci.* **12**, 761402 (2021).
19. Li, J. et al. Natural variation of domestication-related genes contributed to latitudinal expansion and adaptation in soybean. *BMC Plant Biol.* **24**, 651 (2024).
20. Zhang, Z. et al. Improving the accuracy of whole genome prediction for complex traits using the results of genome wide association studies. *PLoS ONE* **9**, e93017 (2014).
21. Li, B. et al. Genomic prediction of breeding values using a subset of SNPs identified by three machine learning methods. *Front. Genet.* **9**, 237 (2018).
22. Vaswani, A. et al. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **30** (2017).
23. Choi, S. R. & Lee, M. Transformer architecture and attention mechanisms in genome data analysis: a comprehensive review. *Biology* **12**, 1033 (2023).
24. Benegas, G. et al. A DNA language model based on multispecies alignment predicts the effects of genome-wide variants. *Nat. Biotechnol.* **43**, 1960–1965 (2025).
25. Lin, F. et al. MMST-ViT: climate change-aware crop yield prediction via multi-modal spatial-temporal vision transformer. In *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)* 5774–5784 (2023).
26. Xiong, X. et al. Daily DeepCropNet: a hierarchical deep learning approach with daily time series of vegetation indices and climatic variables for corn yield estimation. *ISPRS J. Photogramm. Remote Sens.* **209**, 249–264 (2024).
27. Xu, Y., Ma, Y. & Zhang, Z. Self-supervised pre-training for large-scale crop mapping using Sentinel-2 time series. *ISPRS J. Photogramm. Remote Sens.* **207**, 312–325 (2024).

28. Avsec, Ž et al. Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods* **18**, 1196–1203 (2021).
29. Dalla-Torre, H. et al. Nucleotide transformer: building and evaluating robust foundation models for human genomics. *Nat. Methods* **22**, 287–297 (2025).
30. Hollmann, N. et al. Accurate predictions on small data with a tabular foundation model. *Nature* **637**, 319–326 (2025).
31. Consens, M. E. et al. Transformers and genome language models. *Nat. Mach. Intell.* **7**, 346–362 (2025).
32. Cai, Z. et al. MOTHER-OF-FT-AND-TFL1 regulates the seed oil and protein content in soybean. *New Phytol.* **239**, 905–919 (2023).
33. Duan, Z. et al. Natural allelic variation of *GmST05* controlling seed size and quality in soybean. *Plant Biotechnol. J.* **20**, 1807–1818 (2022).
34. Zhang, C. et al. High-quality genome of a modern soybean cultivar and resequencing of 547 accessions provide insights into the role of structural variation. *Nat. Genet.* **56**, 2247–2258 (2024).
35. Wang, M. et al. Parallel selection on a dormancy gene during domestication of crops from multiple families. *Nat. Genet.* **50**, 1435–1441 (2018).
36. Crossa, J. et al. Expanding genomic prediction in plant breeding: harnessing big data, machine learning, and advanced software. *Trends Plant Sci.* **30**, 756–774 (2025).
37. He, K. et al. Deep residual learning for image recognition. *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* **2016**, 770–778 (2016).
38. Balduzzi, D. et al. The shattered gradients problem: if ResNets are the answer, then what is the question? *Proc. 34th Int. Conf. Mach. Learn.* **70**, 342–350 (2017).
39. Clauwaert, J., Menschaert, G. & Waegeman, W. Explainability in transformer models for functional genomics. *Brief. Bioinform.* **22**, 1–11 (2021).
40. Feng, Y. et al. Dual-function C2H2-type zinc-finger transcription factor GmZFP7 contributes to isoflavone accumulation in soybean. *New Phytol.* **237**, 1794–1809 (2023).
41. Liu, Y. et al. An R2R3-type MYB transcription factor, GmMYB77, negatively regulates isoflavone accumulation in soybean [*Glycine max* (L.) Merr. *Plant Biotechnol. J.* **23**, 824–838 (2025).
42. Li, Y. et al. Genome-wide signatures of the geographic expansion and breeding of soybean. *Sci. China Life Sci.* **66**, 350–365 (2023).
43. Azam, M. et al. Seed isoflavone profiling of 1168 soybean accessions from major growing ecoregions in China. *Food Res. Int.* **130**, 108957 (2020).
44. Abdelghany, A. M. et al. Profiling of seed fatty acid composition in 1025 Chinese soybean accessions from diverse ecoregions. *Crop J.* **8**, 635–644 (2020).
45. Sun, J. et al. Rapid HPLC method for determination of 12 isoflavone components in soybean seeds. *Agric. Sci. China* **10**, 70–77 (2011).
46. Ghosh, S. et al. Seed tocopherol assessment and geographical distribution of 1151 Chinese soybean accessions from diverse ecoregions. *J. Food Compos. Anal.* **100**, 103932 (2021).
47. Qi, J. et al. Profiling seed soluble sugar compositions in 1164 Chinese soybean accessions from major growing ecoregions. *Crop J.* **10**, 1825–1831 (2022).
48. Agyenim-Boateng, K. G. et al. Profiling of naturally occurring folates in a diverse soybean germplasm by HPLC-MS/MS. *Food Chem.* **384**, 132520 (2022).
49. Gebregziabher, B. S. et al. Identification of genomic regions and candidate genes underlying carotenoid accumulation in soybean using next-generation sequen-cing based bulk segregant analysis. *J. Integr. Agric.* **24**, 2063–2079 (2025).
50. Agyenim-Boateng, K. G. et al. Identification of quantitative trait loci and candidate genes for seed folate content in soybean. *Theor. Appl. Genet.* **136**, 149 (2023).
51. Li, Y. et al. Study on multi-environment genome-wide prediction of inbred agronomic traits in maize natural populations. *Chin. Bull. Bot.* **59**, 1041 (2024).
52. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
53. Browning, B. L. & Browning, S. R. Genotype imputation with millions of reference samples. *Am. J. Hum. Genet.* **98**, 116–126 (2016).
54. Li, J. et al. Leveraging weighted embedding and Transformer architecture to improve phenotype prediction of complex traits for crops. bioRxiv https://doi.org/10.5281/zenodo.18809685 (2026).

Acknowledgements

This work was supported by the Biological Breeding-National Science and Technology Major Project (2023ZD0403301 to J.L.) and National Natural Science Foundation of China (32272178 to J.S, 32472193 to B.L, 32001574 to J.L.).

Author contributions

J.L. and J.S. designed the experiments and managed the project; L.Y. and J.L. analysed the data and wrote the manuscript, M.L., R.H., Yecheng Li, Z.H., and Yitian Liu performed part of the work. B.L., S.Z., and L.L. collected the dataset and performed data preprocessing, J.L., J.S., L.Q., A.S.S., and K.G.A-B. revised and edited the manuscript; and all authors read and approved the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-026-71035-5.

Correspondence and requests for materials should be addressed to Jing Li, Lijuan Qiu or Junming Sun.

Peer review information *Nature Communications* thanks Lanzhi Li and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.