CLCNet: a contrastive learning and chromosome-aware network for genomic prediction in plants

Jiangwei Huang[1,5,#], Zhihan Yang[2,5,#], Mou Yin[3,5], Chao Li[1,5], Jinmin Li[1,5], Yu Wang[1,5], Lu Huang[1,5], Miaomiao Li[2,5], Chengzhi Liang[2,5], Fei He[3,4,5], Rongcheng Han[1,5,*], and Yuqiang Jiang[1,5,*]

[1]Laboratory of Integrative Physiology, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing, China

[2]Key Laboratory of Seed Innovation, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing, China

[3]Laboratory of Advanced Breeding Technologies, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing, China

[4]CAS-JIC Centre of Excellence for Plant and Microbial Science (CEPAMS), Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing, China

[5]University of Chinese Academy of Sciences, Beijing, China

* Correspondence (Tel: +86-10-64806367; Email: hanrch@genetics.ac.cn (R.H.); Tel: +86-10-64806362; Email: yqjiang@genetics.ac.cn (Y.J.))

Keywords: Genomic Prediction, Deep Learning, Feature Selection, Contrastive Learning, Chromosome-aware

ABSTRACT

Genomic selection (GS) uses genome-wide markers and phenotypes to predict complex traits and breeding values. The effectiveness of GS critically depends on the accuracy of genomic prediction (GP) models. However, traditional GP frequently encounter difficulties in accurately capturing inter-individual variability and are often confronted with the challenge posed by the curse of dimensionality, with features like SNPs far exceeding sample sizes, thereby severely restricting their predictive performance. To address these challenges, we present CLCNet (Contrastive Learning and Chromosome-aware Network), a novel deep learning framework that integrates multi-task learning with contrastive learning for GP. CLCNet comprises two key components: (i) a contrastive learning module that enhances the model's ability to capture fine-grained, genotype-dependent phenotypic differences among individuals, and (ii) a chromosome-aware module that performes structured feature selection at both chromosome and genome levels, retaining the most informative SNPs. CLCNet was evaluated across four major crop species, including maize (*Zea mays*), cotton (*Gossypium hirsutum*), rapeseed (*Brassica napus*), and soybean (*Glycine max*), covering ten agronomically important traits, and was compared with a diverse set of classical linear, machine learning, and deep learning models. Across most traits, CLCNet achieved top prediction performance, with statistically significant improvements in Pearson correlation coefficient (PCC), typically ranging from 0.3% to 6.5% over strong baseline models, together with reduced mean squared error (MSE). Notably, the advantages of CLCNet were pronounced for traits in maize, rapeseed, and soybean, while for cotton traits largely governed by additive genetic effects, its performance remained stable and did not show any decline. Overall, these results demonstrated that CLCNet provided a robust and effective framework for improving genomic prediction accuracy and holds substantial potential for accelerating genetic gain in plant breeding.

INTRODUCTION

Genomic selection (GS) has become an increasingly important approach in modern plant breeding by enabling the prediction of genomic estimated breeding values (GEBVs) through the integration of genome-wide molecular markers with phenotypic data (Meuwissen et al., 2001). In contrast to traditional marker-assisted selection (MAS), which primarily focuses on loci with large effects, GS captures the collective contribution of a large number of locis with small to moderate effects distributed across the genome, thereby providing a more comprehensive and accurate framework for the prediction of complex agronomic traits (Crossa et al., 2017; Xu et al., 2020; Budhlakoti et al., 2022). A wide range of genomic prediction (GP) models has been developed for GS, including classical linear models, such as rrBLUP, GBLUP, and Bayesian regression approaches, such as Bayesian A/B/C/Cπ, as well as machine learning methods such as support vector regression and random forest regression (Maenhout et al., 2007; Crossa et al., 2010; Endelman, 2011; Habier et al., 2011; Heslot et al., 2012; Wang et al., 2018). More recently, deep learning (DL) has garnered increasing attention for its strong capacity to nonlinear genotype-phenotype relationships through hierarchical architectures (Liu et al., 2020; Wang et al., 2020; Gao et al., 2025).

DL-based GP models have expanded across diverse learning paradigms, including single-task, multi-modal, and transfer learning (Eraslan et al., 2019; Washburn et al., 2021). Representative single-task learning models include DeepGS, DLGWAS, DNNGP, Cropformer, and SoyDNGP (Ma et al., 2018; Liu et al., 2019b; Wang et al., 2023; Gao et al., 2023; Wang et al., 2025), while DEM, DeepCCR, and GEFormer exemplify multi-modal learning (Ren et al., 2024; Ma et al., 2024; Yao et al., 2025), and TrG2P represents transfer learning model-based GP (Li et al., 2024). Nevertheless, multi-task learning (MTL) that offers a principled framework for jointly modeling multiple traits and promoting information sharing across related tasks-has been only sporadically explored in GP, and its systematic integration with DL architectures remains limited. Moreover, most DL-based GP models fail to capture fine-grained inter-individual differences in genotype–phenotype relationships.

Despite these methodological advances, the practical performance of GP remains fundamentally constrained by the “curse of dimensionality”. High-throughput genotyping platforms generate hundreds of thousands to millions of single nucleotide polymorphisms (SNPs), while the number of available phenotypic samples is often

limited. This considerable imbalance (e.g., >100,000 features vs. <10,000 samples) frequently leads to overfitting, instability, and reduced generalizability of predictive models, particularly in the presence of genomic noise or sparse phenotype data (Habier et al., 2007; Wang et al., 2018; Tong & Nikoloski, 2021). Consequently, effective feature selection has been recognized as a critical step for improving signal-to-noise ratio and model stability. However, many existing approaches generally treat SNPs as independent or unordered features and fail to explicitly exploit the inherent chromosomal organization of the genome, thereby overlooking epistasis, which plays a crucial role across the genome (Fu et al., 2025; Yin et al., 2025; MacNish et al., 2026). Feature selection strategies that explicitly preserve chromosomal structure, capturing both local (intra-chromosomal) and global (inter-chromosomal) genetic interactions thus remain insufficiently developed.

To address these challenges, we propose CLCNet (Contrastive Learning and Chromosome-aware Network), a novel multi-task deep learning framework. CLCNet integrates two core modules: (i) a contrastive learning module, designed to capture and emphasize inter-sample differences in genotype-phenotype relationships; and (ii) a chromosome-aware module, which implements a hierarchical SNP selection strategy at both inter- and intra-chromosomal levels. The performance of CLCNet was assessed across four major crop species-maize (*Zea mays*), cotton (*Gossypium hirsutum*), rapeseed (*Brassica napus*), and soybean (*Glycine max*), across ten distinct traits. The evaluation process involved a thorough comparison with three classical linear models (rrBLUP, Bayesian Ridge, Bayesian Lasso), two machine learning models (LightGBM, SVR), and two deep learning models (DNNGP, DeepGS). Our results collectively demonstrate that CLCNet provides a robust and biologically informed framework for improving GP and holds substantial potential for accelerating genetic gain in plant breeding programs.

RESULTS

1. The architectural design of CLCNet

CLCNet is a multi-task learning framework designed to predict phenotypic traits from high-dimensional SNP data as its primary objective, while simultaneously learning low-dimensional SNP feature embeddings as an auxiliary task. The core architecture adopted a dual-pathway backbone, comprising a non-linear pathway with three fully

connected layers and a parallel linear pathway. These two pathways were integrated through a residual connection to preserve linear additive effects while enabling non-linear effects (Figure 1 A). The shared backbone outputs were then fed into two task-specific branches: a regression head for phenotypic prediction, optimized using mean squared error (MSE), and an auxiliary head that generates SNP embeddings for contrastive learning (Figure 1C).

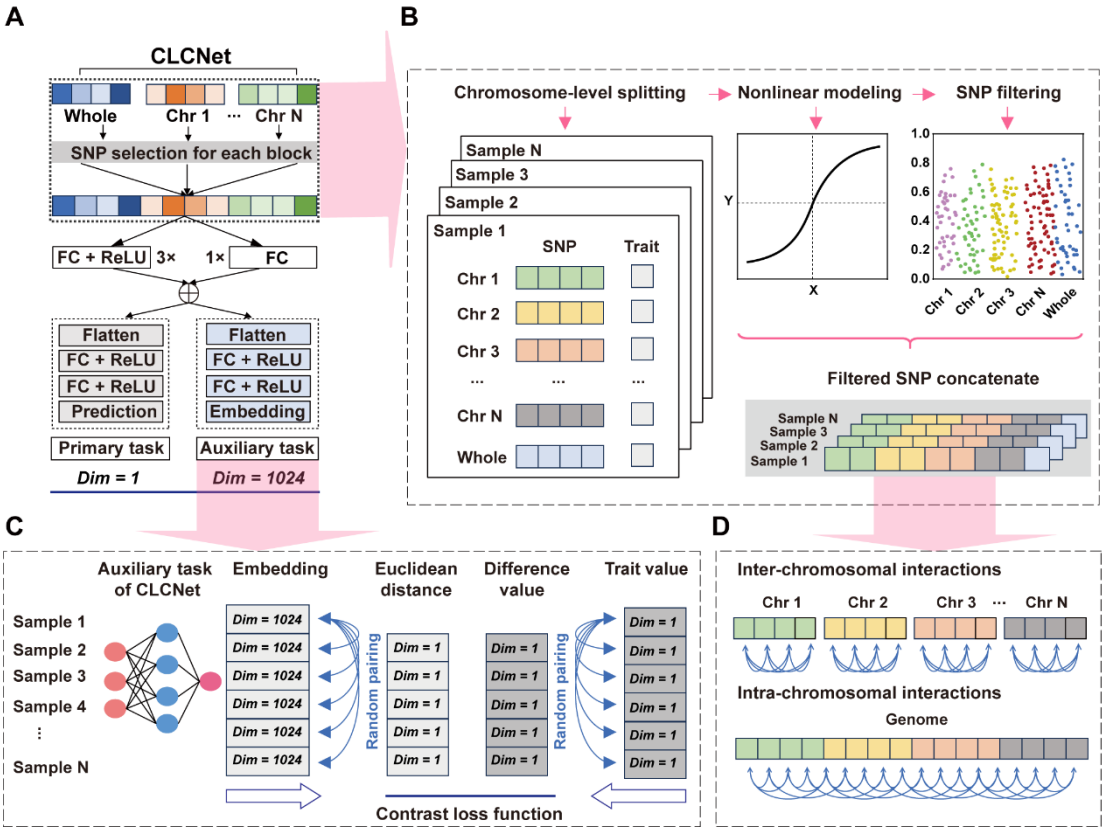

Figure 1. The Framework of CLCNet Model.

(A) CLCNet architecture. FC denotes fully connected layers, and ReLU denotes the activation function. "3×" and "1×" indicate the number of repeated blocks. "Dim" refers to feature dimensionality; the output dimension of the primary task (trait prediction) is 1, while the auxiliary task (embedding) is set to 1024. **(B) Chromosome-aware module.** First, genome-wide SNPs were partitioned into chromosome SNP groups while retaining the original genome SNPs. Second, nonlinear feature selection methods, such as LightGBM, were applied to identify SNPs with positive contribution values. Third, the selected SNP subsets were concatenated to form the final feature set. **(C) Contrastive learning module.** For each sample, the 1024-dimensional embedding generated by the auxiliary task was used for contrastive learning. Randomly paired samples were used to compute the Euclidean distance between embeddings, yielding a scalar value, while the absolute difference between their phenotypic values served as a one-dimensional difference value. These two quantities were incorporated into a customized contrastive loss function (see Equation 2 in Materials and Methods section). **(D) Design concept of the chromosome-aware module.** Local selection

retained within-chromosome dependencies, such as LD blocks and short-range epistasis, while global selection captured inter-chromosomal and long-range effects. Combining both levels preserves complementary genetic dependencies.

The framework incorporates two key methodological innovations. First, a chromosome-aware SNP preprocessing module was introduced to exploit the inherent structural organization of the genome. In this module, SNPs were grouped by chromosome, and a feature selection algorithm was applied independently at both chromosome SNP (local) and genome SNP (global) levels (Figure 1B). Specifically, LightGBM was used to identify informative variants based on information gain (IG > 0). Local selection preserved within-chromosome dependency structures, such as linkage disequilibrium blocks and short-range epistatic interactions, whereas global selection considered all chromosomes simultaneously and therefore retained signals arising from inter-chromosomal relationships and long-range regulatory effects. By jointly retaining variants identified from both levels, the resulting input representation explicitly reflects complementary short- and long-range genetic dependencies (Figure 1D).

Secondly, a contrastive learning module was integrated into the multi-task framework to further stabilize feature representation learning. During training, samples were randomly paired at each epoch, and a composite loss function was optimized that combined the phenotypic prediction loss, measured by the MSE between predicted and observed phenotypes, with a customized contrastive loss (see Equation 2 in Materials and Methods section). The contrastive loss was defined as the difference between the Euclidean distance between paired samples in the SNP embedding space and the absolute difference in their observed phenotypic values. In this way, the model was encouraged to learn embeddings in which phenotypically similar samples were mapped closer together and phenotypically dissimilar samples were pushed farther apart. This design is particularly important in GP, where densely distributed markers can induce substantial random effects at the sample level, potentially masking true genetic signal structures. By enforcing consistency between embedding distance and phenotypic difference, contrastive learning helps constrain such stochastic sample-level variation

CLCNet jointly learns accurate trait predictions and informative sample-level representations by integrating chromosome-aware feature selection with contrastive learning. It is particularly well suited for genomic selection scenarios with extremely high-dimensional marker data (often >100,000 SNPs) and relatively limited sample

sizes (often <10,000), conditions that are common in modern sequencing-based breeding programs.

2. Superior performance of CLCNet over other models

Publicly available data from CropGS-Hub, a large-scale genomic selection resource that integrates genotypic and phenotypic information from 14 populations across 7 major crop species (Chen et al., 2024), were utilized in our study. From this resource, four representative datasets covering maize (*Zea mays*), cotton (*Gossypium hirsutum*), rapeseed (*Brassica napus*), and soybean (*Glycine max*) were curated. The final datasets span ten agronomically important traits: days to tasseling (DTT), ear weight (EW), plant height (PH), flowering time (FT), oil content (OC), protein content (PC), hundred seed weight (HSW), seed length (SL), fiber strength (FStr), and fiber micronaire (FMic). These datasets exhibit substantial heterogeneity in both genomic and phenotypic scales, with sample sizes ranging from fewer than 1,000 to over 8,600 individuals, and raw SNP counts spanning from approximately one million to nearly 30 million markers (Supplementary Table 1). After linkage disequilibrium (LD) filtering, the final number of SNPs used for model training remained high, ranging from ~125,000 to over 660,000 markers, reflecting a typical high-dimensional, low-sample-size setting in genomic selection.

We evaluated CLCNet against a diverse set of baseline models, including three classical statistical methods, rrBLUP (Endelman, 2011), Bayesian Ridge (Pérez and de Los Campos, 2010), and Bayesian Lasso (Park and Casella, 2008); two machine learning algorithms, LightGBM (Ke et al., 2017) and SVR (Basak et al., 2007); and two deep learning frameworks, DNNGP (Wang et al., 2023) and DeepGS (Ma et al., 2018), using the Pearson correlation coefficient (PCC) and mean squared error (MSE) across all ten traits. Statistical significance was evaluated using one-tailed paired t-tests based on three independent repetitions of randomly stratified ten-fold cross-validation.

For the maize dataset, CLCNet achieved the highest PCC across all three traits (DTT, EW, and PH), reaching 95.34%, 80.62%, and 95.14%, respectively (Figure 2). The performance gains were statistically significant in nearly all pairwise comparisons, with only two exceptions: the comparison with DNNGP on EW (Figure 2B) and with SVR on PH (Figure 2C). Notably, for DTT, CLCNet significantly outperformed other models (Figure 2 A). The best-performing baseline models within the statistical, machine learning, and deep learning model were consistently rrBLUP, SVR, and DNNGP,

respectively. Compared with these, CLCNet improved PCC by 0.34%, 0.46%, and 0.77% on DTT; by 1.07%, 0.58%, and 0.23% on EW; and by 0.36%, 1.03%, and 0.69% on PH. In addition, LightGBM, DeepGS, and DNNGP showed the weakest performance on DTT, EW, and PH, respectively, where CLCNet achieved larger improvements of 1.31%, 2.76%, and 1.15%. In terms of MSE, CLCNet yielded the lowest MSE values, with all improvements being statistically significant (Supplementary Figure 1 A-C).

A similar performance pattern was observed for rapeseed. CLCNet ranked first in PCC for all three traits, with values of 82.96% for FT, 41.54% for OC, and 30.03% for PC (Figure 2D-F). Compared with rrBLUP, which ranked second across rapeseed traits, CLCNet improved PCC by 0.62% for FT, 4.35% for OC, and 6.53% for PC. Relative to other representative models like DeepGS, DNNGP, LightGBM, and SVR, CLCNet achieved even larger gains: for FT, PCC increased by 1.16%, 9.09%, 4.65%, and 11.44%, respectively; for OC, gains were 10.02%, 8.98%, 7.59%, and 13.31%; and for PC, improvements reached 8.37%, 10.13%, 12.19%, and 10.12%. The MSE results were highly consistent with the PCC-based evaluations (Supplementary Figure 1D-F).

For soybean, CLCNet demonstrated clear and robust advantages for HSW and SL, achieving significantly higher PCC values at 51.71% and 57.17%, respectively, together with lower MSE values than all competing models (Figure 2 G-H and Supplementary Figure 1 G-H). Specifically, for HSW, CLCNet's PCC exceeded the second-best model, DeepGS, by 1.35% and surpassed the worst-performing DNNGP by 8.41%. For SL, CLCNet's PCC was 0.80% higher than the second-best rrBLUP and 5.90% higher than the lowest-performing LightGBM.

In two traits from cotton, the predictive performance of different models was highly similar across the two traits, and no significant differences were detected between CLCNet and six strong baseline models, including rrBLUP, Bayesian Ridge, Bayesian Lasso, SVR, LightGBM, and DeepGS (Figure 2 I and Supplementary Figure 1 I). For FMic, CLCNet showed comparable PCC performance to rrBLUP, Bayesian Ridge, Bayesian Lasso, and LightGBM, achieving PCC values of 67.70%, 67.36%, 67.17%, 66.80%, and 66.51%, respectively, with no statistically significant differences among these models (Figure 2I). In contrast, CLCNet significantly outperformed SVR, DeepGS, and DNNGP, with PCC improvements of 2.33%, 1.86%, and 3.56%, respectively. In terms of MSE, CLCNet exhibited significantly lower prediction errors than SVR and DNNGP, while no significant differences were observed relative to the remaining five models.

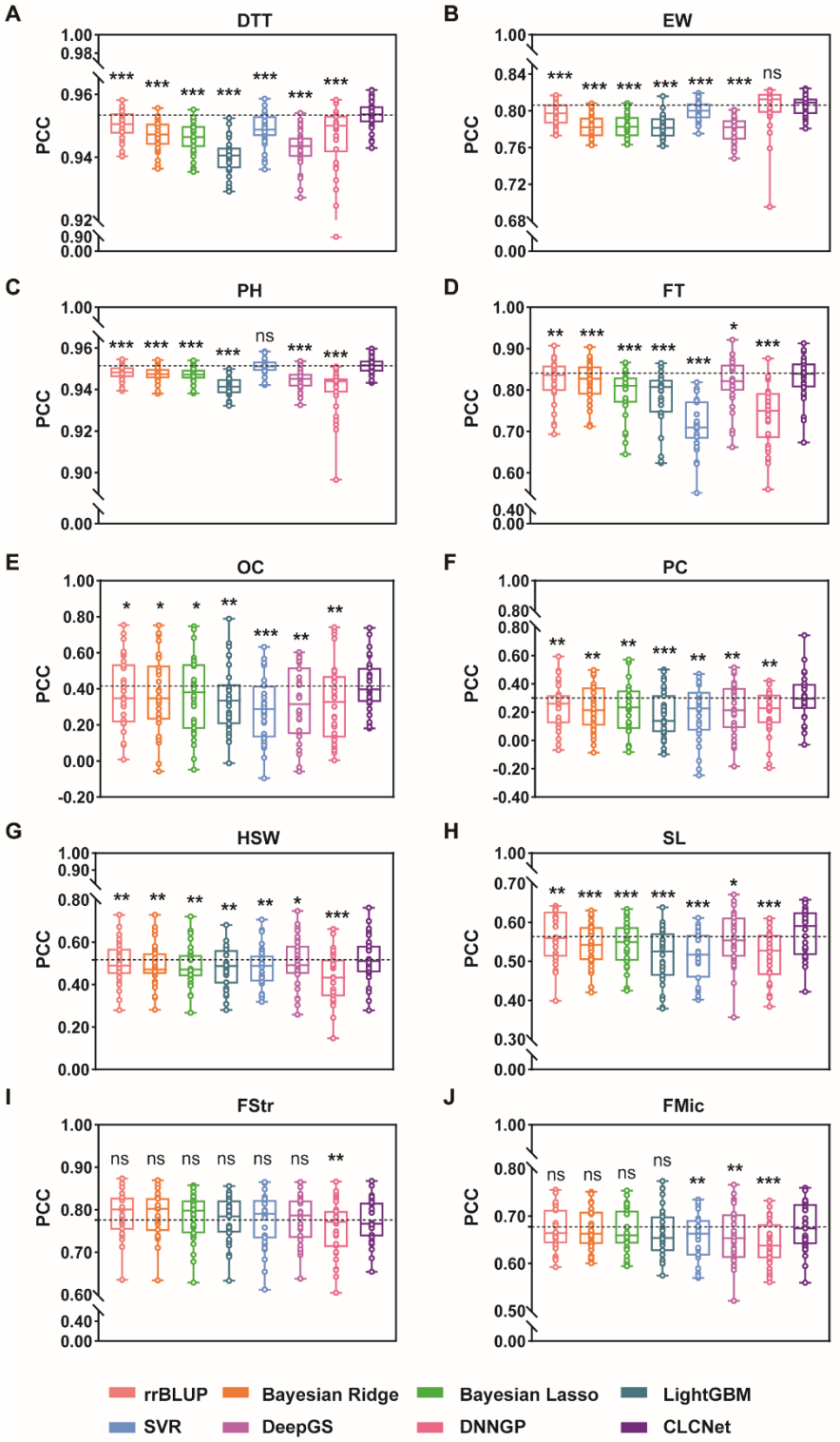

Figure 2. Prediction performance of CLCNet on ten traits across multiple crop datasets: (A) maize DTT, (B) maize EW, (C) maize PH, (D) rapeseed FT, (E) rapeseed OC, (F) rapeseed PC, (G) soybean HSW, (H) soybean SL, (I) cotton FStr and (J) cotton FMic. Trait abbreviations correspond to: DTT-days to tasseling, EW-ear weight, PH- plant height, FT-flowering time, OC-oil content, PC-protein content, HSW-hundred seed weight, SL-seed length, FStr-fiber strength, and FMic-fiber micronaire. PCC, Pearson correlation coefficient. The dashed line represented the

average PCC value of CLCNet. A one-tailed paired t-test was used to compare CLCNet with other models. *** denoted $p < 0.001$, ** denoted $p < 0.01$, * denoted $p < 0.05$, and ns indicated no significant difference.

Overall, CLCNet achieved top predictive performance, with statistically significant improvements in PCC of approximately 0.3%–6.5% over strong baseline models, accompanied by reduced MSE. For maize, rapeseed, and soybean, CLCNet showed clear and often statistically significant gains, reflecting its ability to model complex and non-linear genotype–phenotype relationships. In contrast, for cotton traits such as FStr and FMic, performance differences among models were much smaller, suggesting a ceiling effect in which largely additive architectures are already well captured by classical linear models. Importantly, under these conditions CLCNet did not suffer performance degradation. Taken together, these results support CLCNet as a promising and robust tool for accelerating genetic gain through more accurate genomic prediction in plant breeding.

3. Strong feature extraction capability of Chromosome-aware module

Feature extraction is critical in genomic prediction to identify informative SNPs, reduce redundancy, and improve model accuracy and generalizability, while mitigating computational cost (Pudjihartono et al., 2022). We compared chromosome-aware (CA) module with two widely used feature selection methods: genome-wide association study (GWAS) and principal component analysis (PCA). For GWAS, SNPs with $p < 0.01$ were selected, whereas for PCA, principal components explaining more than 98% of the variance were retained as input. All eight models were trained using each of the three feature selection strategies, and performance was compared in terms of PCC and MSE.

Across all ten traits, CLCNet with the CA module consistently achieved higher PCC and lower MSE compared with PCA, with statistically significant improvements observed for all traits. Specifically, PCC gains of CA over PCA ranged from 1.40% (PH) to 18.76% (PC), while gains over GWAS ranged from 0.20% (PH) to 4.06% (HSW), with FStr, OC, and PC showing no significant difference versus GWAS. MSE comparisons followed a similar trend, with CA showing lower prediction errors across most traits and significant reductions relative to PCA and GWAS for the most of traits (Figure 3B).

Consistent trends were observed across other models, including rrBLUP, Bayesian Ridge, Bayesian Lasso, and DeepGS. In these models, the CA module generally yielded higher PCC and lower MSE than both GWAS and PCA, with statistically significant improvements detected in the majority of traits (typically 6-9 out of 10 traits per model) (Supplementary Figures 2-5). Specifically, CA achieved PCC gains ranging from approximately 0.1% to over 5%, with larger improvements consistently observed for genetically complex traits such as FMic, PC, and HSW. When compared with models trained without feature selection (green bars), CA resulted in significant PCC improvements in 7-10 traits, with increases up to 5.50% (PC) in Bayesian Ridge and 5.15% (PC) in Bayesian Lasso, accompanied by corresponding reductions in MSE. These results indicated that CA can lead to robust and consistent performance gains across both linear and nonlinear genomic prediction models.

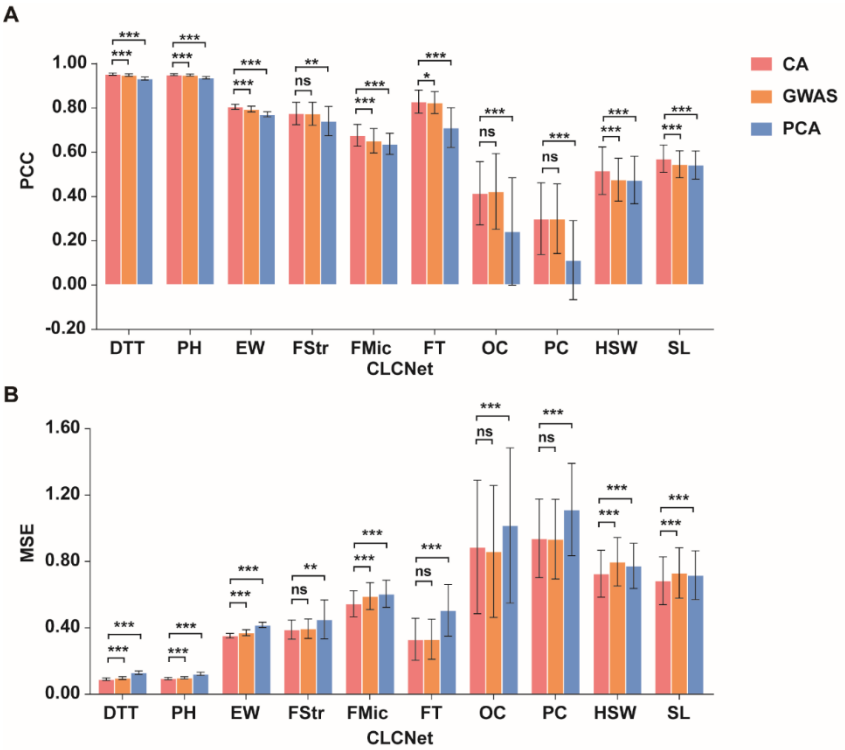

Figure 3. Comparison the PCC and MSE of CA (chromosome-aware module), GWAS (genome wide association study) and PCA (principal component analysis) feature selection methods in CLCNet.(A) Comparison the PCC of CA, GWAS and PCA feature selection methods. (B) Comparison the MSE of CA, GWAS and PCA feature selection methods. A one-tailed paired t-test was used to compare CA with other feature selection methods. *** denoted $p < 0.001$, ** denoted $p < 0.01$, * denoted $p < 0.05$, and ns indicated no significant difference.

Model-dependent exceptions were also observed, where the magnitude and significance of CA's advantage varied with model-feature compatibility. In LightGBM,

no significant differences in PCC were detected between CA and the no-selection setting across all 10 traits (Supplementary Figure 6), and no trait showed a significant advantage of CA over GWAS. This outcome was expected, as the CA module was derived from LightGBM feature importance, resulting in feature subsets highly consistent with the model's internal split and gain criteria. Nevertheless, when compared with PCA, CA still achieved significant PCC improvements in 6 traits (PH, FStr, FMic, OC, HSW, and SL), with gains ranging from 0.35% to 10.27%, highlighting the limitations of linear projection-based dimensionality reduction in tree-based models.

In SVR, GWAS exhibited strong compatibility with the kernel-based framework. Compared with GWAS, CA showed significant PCC advantages in only 2 traits (FMic and FT), with modest gains of 1.12% and 1.27%, respectively (Supplementary Figure 7). However, relative to the no-selection setting, CA significantly improved PCC in 9 out of 10 traits, with gains ranging from 0.13% (PH) to 8.60% (OC), and also outperformed PCA in 5 traits (DTT, PH, EW, FStr, and FMic), where improvements reached up to 20.14% (PH).

DNNGP was not evaluated under the no-selection setting because it inherently requires PCA-based dimensionality reduction. Consistent with this architectural dependence, CA exhibited significant advantages over PCA in only 3 traits (FStr, FMic, and FT), with PCC increases of 1.56%, 3.61%, and 3.54%, respectively (Supplementary Figure 8), while no significant differences were detected for the remaining traits. This further reflectd the strong coupling between DNNGP and PCA-derived features, and explained the relatively limited performance margin of CA in this model.

Overall, the chromosome-aware (CA) module demonstrated a clear and consistent advantage in feature extraction across models and traits. Across the eight evaluated models, CA achieved statistically significant PCC improvements in the majority of traits (typically 6-10 out of 10 per model), with performance gains ranging from ~0.1% to 18.76%, and the largest improvements repeatedly observed for genetically complex traits such as FMic, PC, and HSW. Compared with PCA, CA showed broad and often substantial advantages, whereas relative to GWAS, CA delivered more moderate but still consistent improvements (up to ~4-8% in several models), indicating its ability to retain informative SNPs while reducing redundancy. Model-dependent effects further revealed that CA's benefit is maximized when feature representations are not intrinsically constrained by the model architecture (e.g., PCA-dependent DNNGP or LightGBM-derived importance). Collectively, these results confirm that CA provided

robust, scalable, and model-agnostic feature extraction, leading to improved prediction accuracy and error reduction across diverse genomic prediction frameworks.

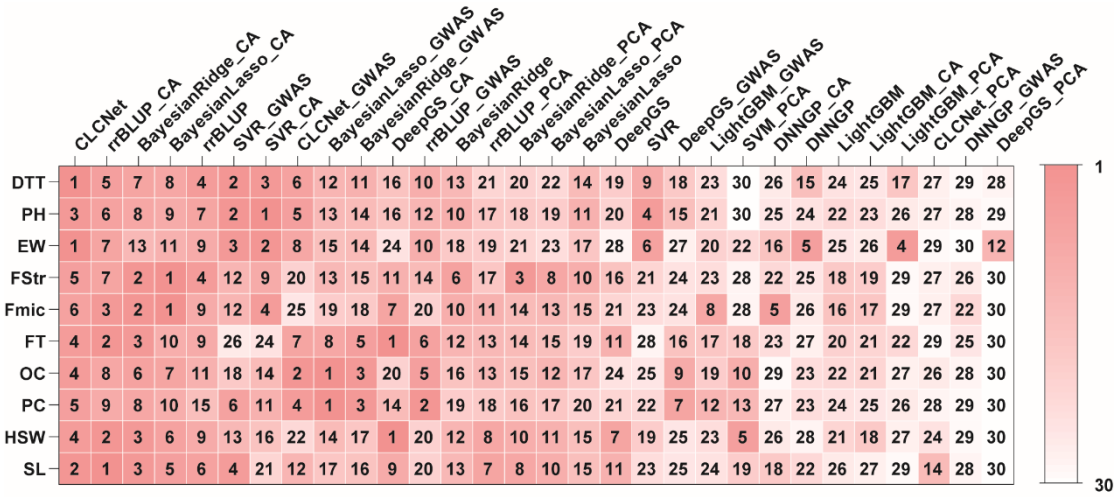

Figure 4. PCC rank distribution of feature selection strategies in eight models across 10 traits. Models were evaluated using different selection strategies: no-selection, CA selection, GWAS selection, and PCA selection. The ranks (value: 1-30) for each trait were based on mean PCC values averaged over three independent runs of randomly generated ten-fold cross-validation.

We further evaluated the overall performance of the chromosome-aware (CA) module across different models and traits. For each of the ten traits, all prediction models under three feature selection strategies (CA, GWAS, and PCA) were ranked based on the mean PCC, averaged over three independent runs of ten-fold cross-validation (Figure 4). Among the top-performing models, eight traits achieved their highest PCC with CA-based feature selection: DTT and EW were topped by CLCNet_CA; PH by SVR_CA; FStr and FMic by Bayesian Lasso_CA; FT and HSW by DeepGS_CA; and SL by rrBLUP_CA. The remaining two traits, OC and PC, reached optimal performance with GWAS combined with Bayesian Lasso. CLCNet consistently ranked among the top-performing models, achieving first place for two traits (DTT and EW), and ranking no lower than sixth (FMic), indicating superior stability compared with other methods. Notably, the top four positions (Figure 4, first four columns) from all models ranked were associated with either CLCNet or CA-based feature selection, highlighting the central contribution of the chromosome-aware design to robust genomic prediction.

354 **4. Demonstrating the effectiveness of independently designed modules in CLCNet**

355 We conducted comprehensive ablation studies across all phenotypic prediction tasks.
356 Four ablated variants, CLCNet_1 to CLCNet_4 were constructed to isolate the effects
357 of the chromosome-aware feature selection and contrastive learning modules.
358 Specifically, CLCNet_1 removed the global level (genome SNP) SNP selection,
359 retaining chromosome-level (chromosome SNP) feature selection, to evaluate the
360 added value of integrating genome-wide information. CLCNet_2 further excluded the
361 contrastive learning module from CLCNet_1, reducing the model to a single-task
362 setting and enabling assessment of contrastive learning when only local features were
363 used. In contrast, CLCNet_3 preserved only the global feature selection module by
364 removing local selection, testing whether global signals alone were sufficient for
365 accurate prediction. CLCNet_4 further removed the contrastive learning module from
366 CLCNet_3, yielding a minimal single-task baseline that captures the cumulative effect
367 of excluding both local feature and contrastive learning. The architectural differences
368 among these models were illustrated in Supplementary Figure 9.

369 Across all traits, no statistically significant differences were observed between
370 CLCNet and CLCNet_1, indicating that removing the global feature selection alone had
371 a limited impact on predictive performance (Figure 5). This suggested that genome SNP
372 selection contributed marginally when chromosome SNP feature selection was retained.

373 Comparisons between CLCNet and CLCNet_2 revealed statistically significant
374 differences only for DTT and FT, while comparisons between CLCNet_1 and
375 CLCNet_2 showed significant differences only for DTT and HSW (Figure 5 A, F, I),.
376 For the remaining traits, no significant differences were detected. These results
377 indicated that the contrastive learning module does not serve as a primary determinant
378 of prediction accuracy, but instead acted as an auxiliary enhancement mechanism that
379 benefits certain traits or task conditions.

380 In contrast, statistically significant differences were consistently observed between
381 CLCNet and CLCNet_3 across nearly all traits (except FStr) (Figure 5 and
382 Supplementary Figure 10). Specifically, PCC decreased by 0.43% for DTT, 0.23% for
383 PH, 0.58% for EW, 2.43% for FMic, 4.63% for FT, 13.17% for OC, 14.22% for PC,
384 5.69% for HSW, and 5.68% for SL. This result demonstrated that local feature selection
385 played a critical role in maintaining predictive accuracy. When local selection was
386 retained, model performance remained comparable to the full CLCNet, whereas its

removal led to substantial declines in performance. Similarly, CLCNet significantly outperformed CLCNet_4, the most simplified variant lacking both local selection and contrastive learning, with PCC improved 0.67% for DTT, 0.36% for PH, 1.08% for EW, 1.29% for FMic, 4.62% for FT, 16.43% for OC, 13.88% for PC, 6.22% for HSW, and 5.69% for SL.

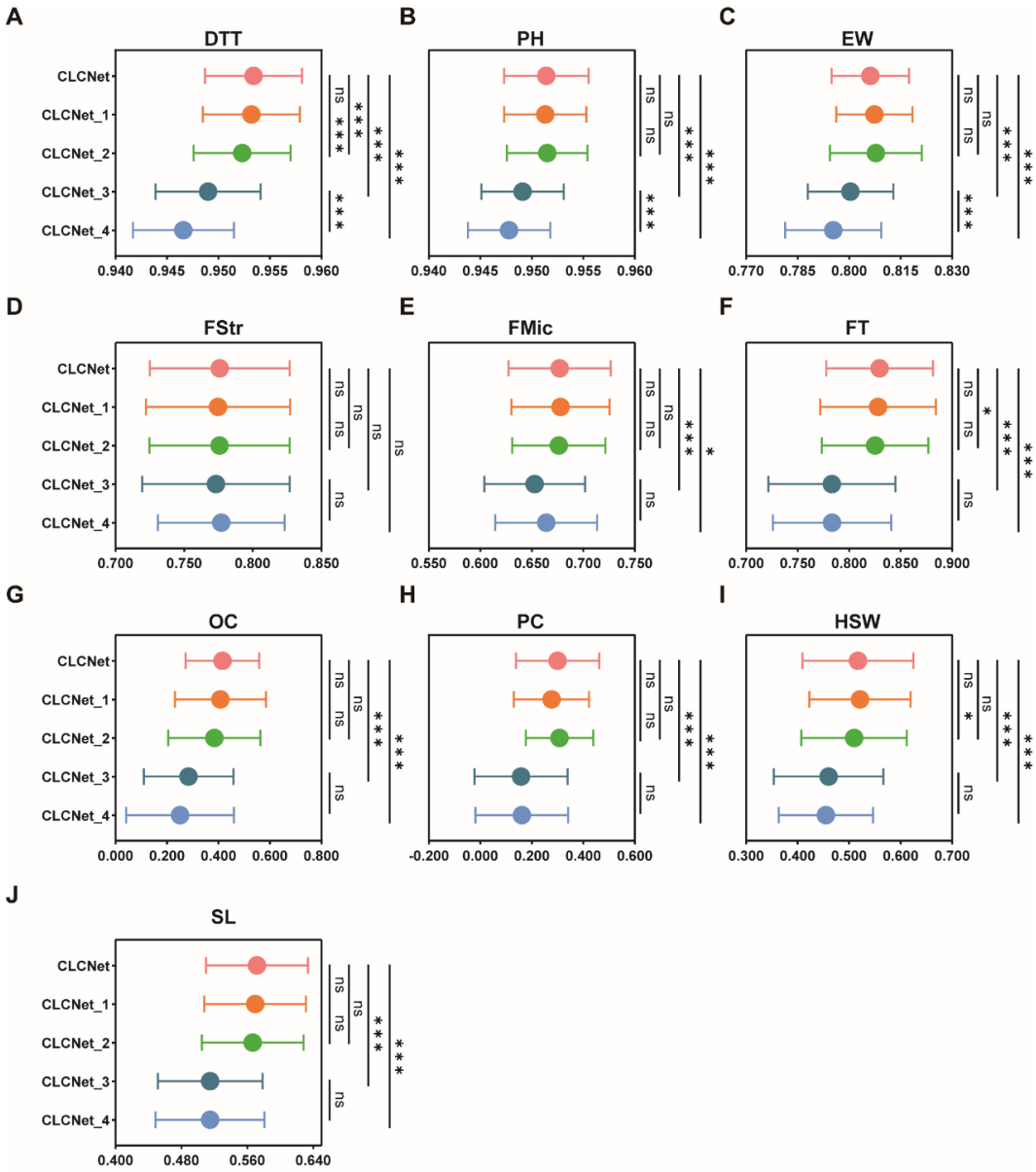

Figure 5. Analysis of CLCNet's core modules on PCC. (A) maize DTT, (B) maize PH, (C) maize EW, (D) cotton FStr, (E) cotton FMic, (F) rapeseed FT, (G) rapeseed OC, (H) rapeseed PC, (I) soybean HSW and (J) soybean SL. A one-tailed paired t-test was used. *** denoted $p < 0.001$, ** denoted $p < 0.01$, * denoted $p < 0.05$, and ns indicated no significant difference.

Comparisons between CLCNet_3 and CLCNet_4 showed statistically significant differences only for DTT, PH, and EW, whereas the differences for the remaining traits

were not significant. Furthermore, for FMic, FT, and PC, removing contrastive learning (i.e., comparing CLCNet_3 with CLCNet_4) did not lead to an additional performance decline; the performance gaps between CLCNet and CLCNet_3 were almost identical to those between CLCNet and CLCNet_4 (FMic:2.43% vs 1.29%; FT: 4.63% vs 4.62%, and PC :14.22% vs 13.88%). These results indicate that, when local feature selection is absent, contrastive learning provides only limited compensatory benefit and cannot recover the performance loss caused by the removal of chromosomal structural modeling.

Overall, the ablation results indicated a clear hierarchy in module importance: local feature selection provided the largest contribution, followed by contrastive learning, while global feature selection had a relatively minor effect.

5. Chromosomal structure guided SNP selection from Chromosome-aware

Our ablation demonstrated that the local feature (chromosome SNP) selection of chromosome-aware (CA) module contributed most substantially compared with contrastive learning module and global feature selection (genome SNP). We hypothesized that its advantage arised from explicitly exploiting chromosomal structural information, including linkage disequilibrium (LD) patterns and epistatic interactions among SNPs. We designed a perturbation experiment aimed at disrupting chromosomal structure while preserving SNP composition (Supplementary Figure 9 F). Specifically, SNP positions within each chromosome were randomly permuted, thereby breaking the original positional order and destroying native LD structure and local epistatic dependencies, while keeping the same number of SNPs for each chromosome. This manipulation removed biologically meaningful positional context without altering allele frequencies or overall feature dimensionality. By comparing model performance before and after chromosomal shuffling using CLCNet, we were able to directly assess whether CA module relied on chromosome structural information rather than merely on the presence of informative SNPs.

Consistent with our expectation, the CLCNet_Shuffle variant exhibited a systematic decline in prediction accuracy across nearly all traits when compared with the original CLCNet. Except for FStr, which did not show a statistically significant difference, all remaining traits experienced significant reductions in PCC. Specifically, the decreases were 0.15% for DTT, 0.12 % for PH, 0.29 % for EW, 2.01 % for FMic, 1.12 % for FT, 9.89 % for OC, 9.23 % for PC, 2.25% for HSW, and 1.65% for SL (Figure 6). Notably,

the largest performance declines were observed for genetically complex traits such as OC and PC, further underscoring the importance of preserving chromosomal structure for accurate prediction.

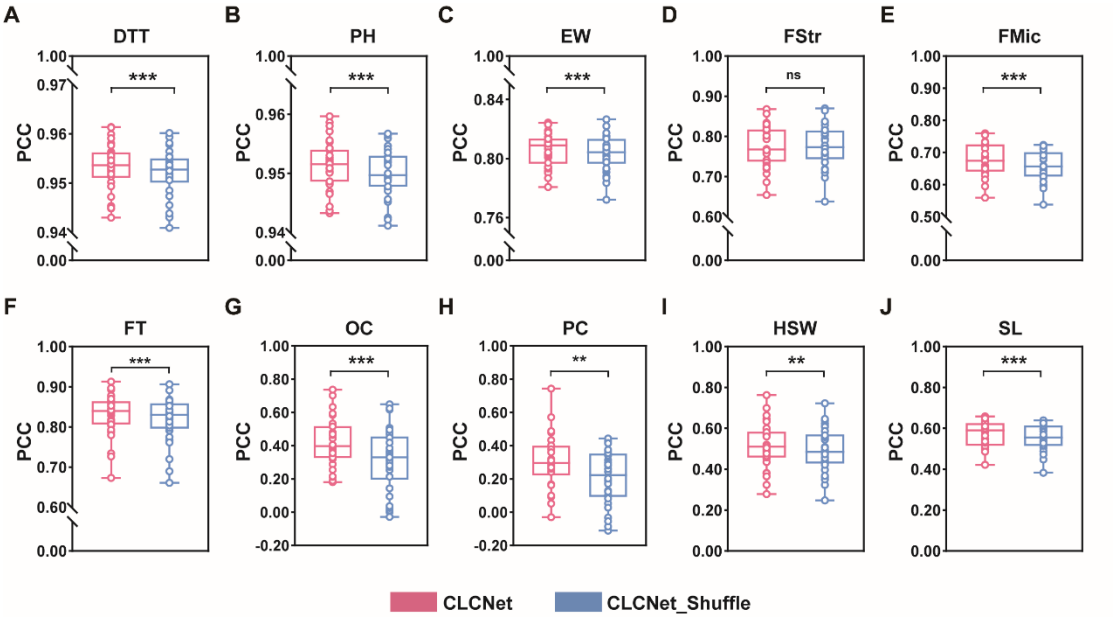

Figure 6. Model PCC performance under SNP position perturbation. (A) maize DTT, (B) maize PH, (C) maize EW, (D) cotton FStr, (E) cotton FMic, (F) rapeseed FT, (G) rapeseed OC, (H) rapeseed PC, (I) soybean HSW and (J) soybean SL. A one-tailed paired t-test was used to CLCNet with CLCNet_Shuffle version. *** denoted $p < 0.001$, ** denoted $p < 0.01$, * denoted $p < 0.05$, and ns indicated no significant difference.

Same trends were observed for the MSE, where chromosomal shuffling led to increased prediction errors across the same set of traits (Supplementary Figure 11). This concordance between PCC and MSE further confirmed that the performance degradation arised from the disruption of chromosome-aware structural information rather than random variation.

Together, these results highlighted the critical role of SNP positional information in chromosome-aware modeling. The pronounced performance degradation under random chromosomal perturbation demonstrated that CLCNet successfully captured structured genomic signals like epistasis.

6. Evaluating SHAP-Based SNP importance against GWAS results

GWAS have been widely used to identify genetic variants associated with complex traits (Liu et al., 2019a). GWAS typically analyzes each SNP individually in relation to a phenotype, thereby ignoring potential interactions among SNPs as well as nonlinear relationships between genotype and phenotype. Deep learning models, such as CLCNet,

can complement these limitations by capturing complex and nonlinear patterns. We used SHAP (SHapley Additive exPlanations) (Lundberg and Lee, 2017) to interpret the contribution of individual SNPs in CLCNet, and the resulting importance scores were compared with GWAS p-values and estimated effect sizes. This comparison can evaluate the interpretability of CLCNet, explore its potential for gene discovery, and highlight the biological significance of the identified SNPs.

GWAS were performed for three traits in the maize dataset using a total of 125,820 SNPs (Supplementary Table 1). The chromosome-aware (CA) module of CLCNet selected a subset of these SNPs (approximately 20,000) for phenotype prediction in the DTT, EW, and PH traits. It was observed that most SNPs involved by CLCNet were associated with relatively lower -log10 (p) values in the GWAS Manhattan plots (Figure 7A; Supplementary Figures 12 A; Supplementary Figures 13 A), indicating that the same SNPs contributed differently in the two approaches. Notably, the SNPs involved by CLCNet were evenly distributed across chromosomes and showed substantial overlap with GWAS peaks. In addition, examination of the SHAP values of SNPs involved by CLCNet revealed peak patterns similar to those observed in GWAS (Figure 7C; Supplementary Figures 12 C; Supplementary Figures 13 C), although they were not entirely identical, suggesting that the two methods capture overlapping yet distinct genetic signals.

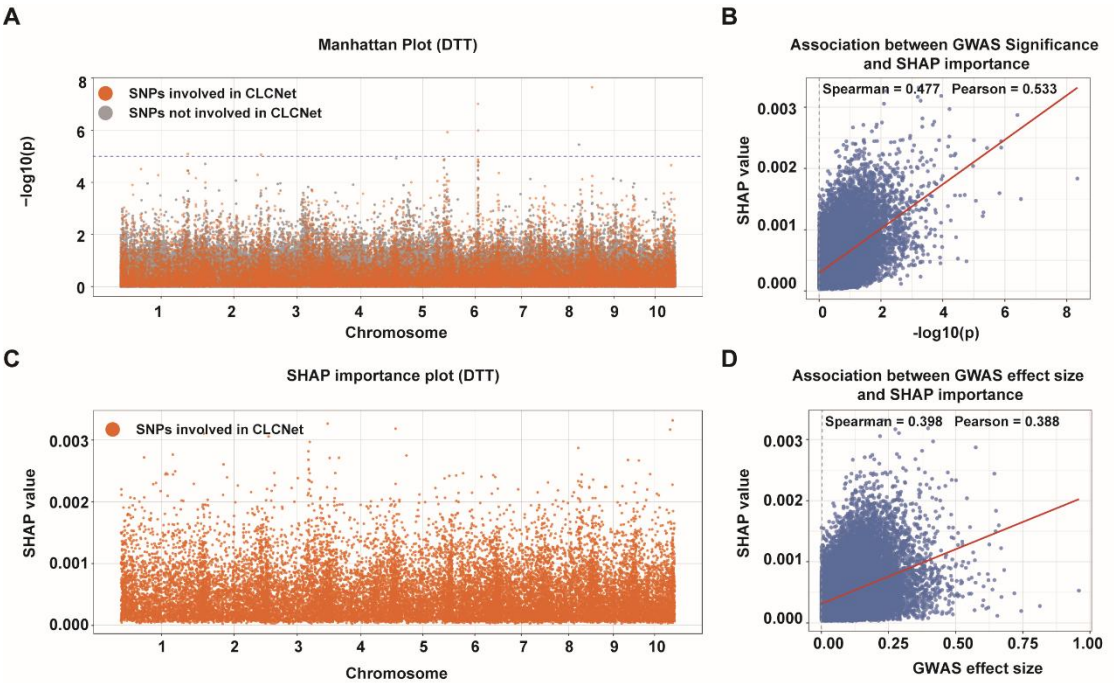

Figure 7. Relationship between SHAP-based SNP importance derived from CLCNet and GWAS for DTT. (A) Manhattan plot of GWAS results for DTT. (B) Relationship between

GWAS association significance and SNP SHAP importance learned by CLCNet for DTT. (C) SHAP-based SNP importance distribution derived from the CLCNet for DTT. (D) Relationship between GWAS estimated effect size and SNP SHAP importance learned by CLCNet for DTT.

Furthermore, the relationships between SHAP values of SNPs involved by CLCNet and their corresponding GWAS p-values and estimated effect sizes were examined (Figure 7B and 7D; Supplementary Figures 12 B and D; Supplementary Figures 13 B and D). Spearman and Pearson correlation coefficients across the three traits ranged from 0.398 to 0.509 and from 0.388 to 0.573, respectively, indicating a moderate but consistent correlation. Such moderate correlations were expected, given the fundamental methodological differences between the two approaches. While GWAS evaluated the marginal effect of each SNP independently using linear models, CLCNet captured complex, nonlinear relationships, including epistatic interactions. As a result, SNPs with weak marginal effects in GWAS may still receive high SHAP importance due to their contributions through these nonlinear interactions. These findings suggested that CLCNet not only recover biologically meaningful GWAS signals but also provided complementary insights for gene discovery beyond what traditional linear association analyses can reveal.

DISCUSSION

CLCNet adopts a multi-task learning (MTL) framework, which fundamentally distinguishes it from most existing GP models that rely on single-task learning, multi-modal learning, or transfer learning paradigms. Previous studies have shown that MTL can improve performance across related tasks by exploiting shared representations and information, highlighting its effectiveness as a powerful deep learning paradigm (Zhang et al., 2018). In this study, contrastive learning is further integrated into the MTL framework to reinforce representation learning by explicitly modeling inter-sample differences in genotype-phenotype relationships. Our results indicate that contrastive learning serves as a complementary enhancement. In practical GP settings, sample sizes are frequently limited while feature dimensionality remains much higher. Under these situations, a small number of samples with pronounced phenotypic or genotypic

differences can cause large fluctuations in association estimates, resulting in spurious associations and elevated false-positive rates. Consequently, effective control of this random effects is essential for enhancing model robustness. In CLCNet, contrastive learning explicitly captures phenotypic and genotypic characteristics at the individual-sample level and alleviates this problem by clustering similar samples while separating dissimilar ones. This process promotes the learning of shared, stable patterns across samples, stabilizes genotype-phenotype associations, and ultimately improves predictive accuracy. Through the integration of contrastive learning into the MTL framework, CLCNet effectively constrains the influence of random effects.

The chromosome-aware module proposed in this study serves as an effective and biologically informed feature selection strategy for GP. In high-dimensional genomic datasets, where the number of SNP features vastly exceeds the number of phenotypic samples, dimensionality reduction is essential for improving computational efficiency and predictive robustness-a challenge commonly known as the "curse of dimensionality" (Pudjihartono et al., 2022). Feature selection addresses this issue by reducing noise and redundancy while retaining the most informative genetic signals (Ho et al., 2019; Urbanowicz et al., 2019). However, conventional GWAS-based feature selection methods primarily rely on linear models to evaluate associations between individual SNPs and phenotypes, thereby largely ignoring epistasis and other nonlinear genotype-phenotype interactions (Liu et al., 2019a). Likewise, commonly used GP feature selection methods such as principal component analysis (PCA), maximal information coefficient (MIC), and Pearson-collinearity selection (PCS) typically do not explicitly account for epistatic interactions (Wang et al., 2023; Wang et al., 2025; Zhao et al., 2025).

To overcome these limitations, CLCNet implements an explicit chromosome-based

532 grouping strategy that directly leverages biological chromosome structure. By first
533 modeling SNPs within individual chromosomes, the framework captures local genetic
534 dependencies, while genome-wide SNP integration captures global genomic effects.
535 These locally and globally selected SNP subsets are then integrated into a unified and
536 complementary feature representation. Our results demonstrate that local, within-
537 chromosome SNP selection plays a decisive role in improving prediction accuracy.
538 Moreover, the pronounced performance degradation observed after random
539 permutation of SNP positions provides strong evidence that the chromosome-aware
540 module exploits biologically meaningful positional and structural information, such as
541 LD pattern and epistasis, rather than relying solely on individual SNPs. These findings
542 are consistent with recent large-scale studies showing that epistasis is a pervasive
543 component of the genetic architecture of complex traits, contributing substantially to
544 phenotypic variation and yield-related traits in plants (Fu et al., 2025; Yin et al., 2025).
545 In crop species, epistatic network modules have been shown to reorganize under
546 selection and to play a critical role in enhancing yield potential, underscoring the
547 biological importance of explicitly modeling genetic interactions rather than treating
548 variants as independent predictors (Yin et al., 2025). Importantly, the ability of CLCNet
549 to capture such epistatic effects is central to both its predictive performance and its
550 biological interpretability.

551 The biological relevance and interpretability of CLCNet were further supported by
552 comparisons between SHAP-based SNP importance and GWAS results. The observed
553 moderate but consistent correlations indicate that CLCNet successfully recovers
554 established GWAS signals while simultaneously identifying additional, complementary
555 genetic information. The incomplete concordance between the two approaches reflects
556 their fundamental methodological differences: GWAS evaluates marginal linear effects

independently, whereas CLCNet captures nonlinear genotype–phenotype relationships and epistatic interactions among SNPs. As a result, variants with weak marginal effects in GWAS may still receive high importance in CLCNet due to their involvement in complex interaction patterns. Notably, SNPs highlighted by CLCNet were evenly distributed across chromosomes and overlapped with GWAS peaks, while SHAP importance also emphasized broader genomic regions, suggesting that the model captures polygenic architectures beyond isolated loci.

Despite these strengths, several limitations of this study should be acknowledged. First, the proposed dual-pathway architecture of CLCNet required relatively long training time. For example, on maize with 8,652 samples and 125,820 SNPs, CLCNet ranged from 134 s to 795 s per trait, while deep learning baselines required 74-373 s (DNNGP) or 174-216 s (DeepGS), machine learning models 0.08-106 s (SVR, LightGBM), and statistical models 108-271 s (rrBLUP, Bayesian Ridge, Bayesian Lasso) per trait (Supplementary Table 7). Second, CLCNet should be compared with a broader range of baseline models, especially newly developed deep-learning-based genomic prediction approaches such as VMGP (Zhao et al., 2025) and MeNet (Li et al., 2026). Third, CLCNet currently lacks multi-modal prediction capability. Although molecular marker information is relatively cost-effective, previous studies have demonstrated that integrating multi-omics or environmental variables can further improve prediction performance, and future work will extend CLCNet in this direction.

In conclusion, we developed CLCNet, a novel deep learning model for plant GP, integrating contrastive learning modules and chromosome-aware for the first time (Huang et al., 2024). By explicitly exploiting chromosomal structure to capture both local and global epistatic interactions and by optimizing marker selection, CLCNet consistently outperforms classical linear models, conventional machine learning

approaches and deep learning methods. The incorporation of a multi-task learning strategy, in which phenotypic differences are introduced as an auxiliary objective, primarily serves to reduce random effects by encouraging shared information across samples. Collectively, these results demonstrate that CLCNet provides a robust and scalable solution for high-dimensional GP, with strong potential to accelerate genetic gain in plant breeding programs.

MATERIALS AND METHODS

Data description and preprocessing

We utilized the CropGS-Hub database (Chen et al., 2024), from which four representative datasets were selected for analysis. Data dimensionality was reduced using PLINK (v1.9) (Purcell et al., 2007) for linkage disequilibrium (LD) pruning (*--indep-pairphase 100 10 0.2*). The final datasets comprised maize (*Zea mays*) from GSTP001, cotton (*Gossypium hirsutum*) from GSTP010, rapeseed (*Brassica napus*) from GSTP013, and soybean (*Glycine max*) from GSTP014, encompassing a total of 10 traits. A summary of the preprocessed datasets was provided in Supplementary Table 1.

The architecture of CLCNet

We CLCNet was implemented in PyTorch (v2.1.0; https://pytorch.org/). It is a multi-task deep learning model designed to take genomic single-nucleotide polymorphism (SNP) data as input and to output phenotypic predictions as the primary task, together with low-dimensional feature representations as an auxiliary task. Genotypes were encoded using a standard 0/1/2 coding. The backbone network of CLCNet adopted a dual-pathway architecture. One pathway consisted of three nonlinear fully connected layers with ReLU activation functions, with layer dimensions of {4096,2048, 1024}. The other pathway comprised a single linear fully connected layer with an output dimension of 1024. The outputs of the two pathways were summed at the 1024-dimensional level and subsequently fed into both the primary and auxiliary task branches. Each branch was composed of two nonlinear fully connected layers. For the primary task, the layer dimensions were {4096, 2048,1}, producing the final phenotypic

prediction, whereas for the auxiliary task, the dimensions were {4096,2048,1024}, generating low-dimensional feature embeddings.

A chromosome-aware module was incorporated as a preprocessing step for genomic data. SNPs were first grouped into chromosomal level according to their genomic positions. Local feature selection was then performed within each chromosome to identify informative SNPs (referred to as local SNPs). In parallel, global feature selection was applied across the entire genome to capture genome-wide informative SNPs (referred to as global SNPs). The selected local and global SNPs were concatenated to form the final input feature set for the model. Feature selection was conducted using the LightGBM algorithm, a tree-based method capable of modeling nonlinear relationships among SNPs. SNPs with positive information gain (IG > 0) were retained, while those with zero information gain (IG = 0) were discarded, thereby identifying key loci relevant to the prediction task (Yan et al., 2021).

A contrastive learning module was introduced through a customized training strategy. During each training epoch, a random pairing mechanism was used when loading samples, such that each sample was paired with another randomly selected sample. The paired samples were fed into the model simultaneously, and both their phenotypic values and low-dimensional feature embeddings were predicted. Model optimization was guided by three loss components: the mean squared error (MSE) between the predicted and true phenotypic values for each of the two input samples (Equation 1), and a contrastive loss based on the Euclidean distance between the embeddings of the paired samples and the difference in their phenotypic values (Equation 2). The total training loss was defined as the sum of the two MSE losses and the contrastive loss.

equation 1 $$Loss_{mse} = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2$$

Here, N is the number of samples. y_i is the true value for the i-th observation. $\hat{y}_i$ is the predicted value for the i-th observation.

equation 2 $$Loss_{contrastive} = \frac{1}{N}\sum_{i=1}^{N}\left(\left\|\text{embedding}_1^i - \text{embedding}_2^i\right\| - \text{diff}^i\right)^2$$

Here, N is the number of samples. $\left\|\text{embedding}_1^i - \text{embedding}_2^i\right\|$ denotes the Euclidean distance between the i-th pair of embeddings. diff^i is the phenotypic difference for the i-th pair of samples.

All experiments were conducted on a server equipped with two Intel Xeon® Gold 6330 CPUs, 512 GB of system memory, and an NVIDIA GeForce RTX 3090 GPU.

Model comparison and cross-validation

rrBLUP (Endelman, 2011), Bayesian Ridge (Pérez and de Los Campos, 2010), and Bayesian Lasso (Park and Casella, 2008); two machine learning algorithms, LightGBM (Ke et al., 2017) and SVR (Basak et al., 2007); and two deep learning frameworks, DNNGP (Wang et al., 2023) and DeepGS (Ma et al., 2018),

The predictive performance of CLCNet was evaluated against seven baseline models on benchmark datasets, including three classical statistical methods, rrBLUP (Endelman, 2011), Bayesian Ridge (Pérez and de Los Campos, 2010), and Bayesian Lasso (Park and Casella, 2008); two machine learning algorithms, LightGBM (Ke et al., 2017) and SVR (Basak et al., 2007); and two deep learning frameworks, DNNGP (Wang et al., 2023) and DeepGS (Ma et al., 2018). Classical methods were implemented using the R packages rrBLUP (v4.6.2) and BGLR (v1.1.3). For machine learning, LightGBM was implemented with Python's lightgbm package (v3.3.5), while SVR utilized Python's scikit-learn library (v0.24.2). DeepGS utilized a convolutional neural network (CNN) architecture comprising input, convolutional, max-pooling, dropout, fully connected, and output layers, with dropout rates of 0.2, 0.1, and 0.05 applied at various stages to mitigate overfitting. The ReLU activation function was employed throughout the architecture. Similarly, DNNGP adopted a featuring three convolutional layers, batch normalization, dropout, a flattening layer, a dense layer, and an output layer. Principal components accounting for over 98% of the variance were extracted from the genomic data using PLINK (v1.9) as input features.

A 3 × 10-fold cross-validation (CV) strategy was adopted to ensure robust and unbiased performance evaluation. Specifically, ten-fold CV was repeated three times using different random data partitions (random seed 42,84 and 162), resulting in a total of 30 independent training and testing processes. In each fold, the training data were further randomly split into an internal training set (80%) and an internal validation set (20%), while the test set was strictly held out for final evaluation.

Phenotype values were standardized using z-score normalization based solely on the mean and variance computed from the internal training set to prevent data leakage. The same normalization parameters from the internal training set were then applied to the internal validation and test sets. All models were trained exclusively on the internal training set. Feature selection strategies, including chromosome-aware selection (CA), GWAS-based selection, and principal component analysis (PCA), were conducted only

on the internal training set, and the selected features were subsequently applied to the internal validation and test sets. The test sets were used exclusively for final performance assessment and were not involved in any stage of model training, feature selection, or hyperparameter optimization

Predictive performance was evaluated using the Pearson correlation coefficient (PCC) and mean squared error (MSE). Based on the 30 test results obtained from the 3 × 10-fold CV, one-tailed paired t-tests were performed to assess whether CLCNet achieved statistically significant improvements over the competing baseline models, thereby reducing the likelihood that observed performance gains arose from random variation.

Hyperparameter optimization

Hyperparameter tuning was performed using grid search for both machine learning and deep learning models. For machine learning models, including LightGBM and support vector regression (SVR), the hyperparameters considered were as follows. For LightGBM, the number of leaves {num_leaves: 31, 63}, the maximum tree depth (max_depth: 5, 10, -1), the learning rate {learning_rate: 0.01, 0.05, 0.1}, and the number of boosting rounds (n_estimators: 100, 200, 500) were systematically explored. For SVR, the regularization parameter {C: 0.1, 1, 10, 100}, the epsilon-insensitive loss parameter {epsilon: 0.01, 0.1, 0.5}, the kernel coefficient {gamma: 'scale', 'auto'}, were evaluated, with the RBF kernel chosen for its robustness in genomic data.

For deep learning models, including CLCNet, DeepGS, and DNNGP, the key hyperparameters considered included the learning rate, batch size, and number of training epochs, with candidate values of {0.01, 0.001, 0.0001}, {16, 32, 64}, and {20, 50, 100}, respectively.

The optimal hyperparameter combination for each model was automatically selected based on the highest Pearson correlation coefficient (PCC) on the internal validation set across all ten traits. To reduce the computational cost of hyperparameter tuning, the grid search was performed using only the internal training and validation sets of the first fold from the first of three independent 10-fold cross-validation runs, primarily based on chromosome-aware selected data, except for DNNGP, for which PCA-selected data were used. Detailed hyperparameter optimization results were provided in Supplementary Table 2-6.

Feature selection comparison

To benchmark the effectiveness of our chromosome-aware (CA) module, we compared it with two widely used feature selection strategies: genome-wide association study (GWAS) and principal component analysis (PCA). For GWAS filtering, SNPs were scored using a mixed linear model implemented via GCTA (Yang et al., 2011), and markers with significant associations ($p < 0.01$) were retained for model input. For PCA, dimensionality reduction was performed using the *sklearn.decomposition.PCA* function in Python. Consistent with DNNGP, principal components explaining more than 98% of the total variance in the genotype data were retained as model input. Each of the baseline models—rrBLUP, Bayesian Ridge, Bayesian Lasso, LightGBM, SVR, DNNGP and DeepGS, was evaluated using features selected by all three strategies: GWAS, PCA, and CA method. Model performance was quantified by the Pearson correlation coefficient (PCC) and Mean Squared Error (MSE). Based on the results from 30 independent test sets (3 × 10-folds), one-tailed paired t-tests were conducted to assess whether the CA-based feature selection strategy achieved statistically significant performance improvements over GWAS and PCA across statistical, machine learning, and deep learning paradigms.

SNP position perturbation

The input genotype data were represented as an $n \times m$ matrix, where n denotes the number of samples and m denotes the number of single-nucleotide polymorphisms (SNPs), ordered according to their physical positions along the genome. To perform SNP position perturbation, a single global random permutation was applied to the m SNP columns, while keeping the permutation identical across all samples. This operation preserved the original genotype values for each sample but completely disrupted the positional ordering of SNPs. After perturbation, the input matrix retained the same $n \times m$ dimensionality as the original data.

Model performance before and after perturbation was evaluated using the Pearson correlation coefficient (PCC). Based on the results from 30 independent test evaluations (3 × 10-folds), one-tailed paired t-tests were conducted to determine whether performance degradation after perturbation was statistically significant.

SHAP-Based SNP importance analyzes

CLCNet predictions were interpreted using SHAP (SHapley Additive exPlanations)

values (Lundberg and Lee, 2017). For each trait in maize dataset, the mean absolute SHAP value of each SNP was computed across all samples to obtain a global importance score. GWAS were conducted for the maize traits (DTT, EW, and PH) using all 125,820 SNPs (Supplementary Table 1). A linear mixed model was applied to evaluate the association between individual SNPs and the phenotype, producing p-values and estimated effect sizes for each SNP. Manhattan plots were generated to visualize the genome-wide association signals. SHAP importance scores were compared with GWAS p-values (–log10 transformed) and estimated effect sizes. SNPs involved by the CLCNet were mapped onto the genome, and their SHAP distribution across chromosomes was analyzed relative to GWAS peaks. Spearman and Pearson correlation coefficients were calculated between SHAP values and GWAS statistics for each trait.

Acknowledgements

We would like to express our gratitude to Prof. Fei Lu, Prof. Ni Jiang, Postdoc Dr. Zeyu Zhang rand Postdoc Dr. Jiachao Xu, from the Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, for their professional and insightful discussions. We also thank Dr. Shihao Shao from Peking University Health Science Center for his assistance with the deep learning model. In addition, we are grateful to Prof. Jianxiao Liu from Huazhong Agricultural University for their guidance. We are especially grateful to Prof. Shichao Jin from Nanjing Agricultural University for his thoughtful suggestions and helpful discussions regarding the manuscript.

This research was supported by the National Key R&D Program of China (Grant Numbers: 2024YFA1209900, 2024YFA1209901, 2024YFA1209902).

Data and code availability

All datasets we used is accessible at https://iagr.genomics.cn/CropGS/#/Datasets. The scripts are available on GitHub (https://github.com/SuppurNewer/CLCNet).

Conflict of interest

The authors declare no conflicts of interest.

Authors' contributions

Y.J., and R.H. conceived the study and supervised the research; J.H. developed the

model and drafted the manuscript; J.H., Z. Y., M.Y., C.L., and J.L analyzed data. R.H. Y.W., L.W., F.H., C.L., and M.L. discussed the results and revised the manuscript. All authors read and approved the final manuscript.

Supplemental information

Supplementary Table 1. Benchmark dataset in this study.

Supplementary Table 2. SVR hyperparam results.

Supplementary Table 3. LightGBM hyperparam results.

Supplementary Table 4. DeepGS hyperparam results.

Supplementary Table 5. DNNGP hyperparam results.

Supplementary Table 6. CLCNet hyperparam results.

Supplementary Table 7. Training and inference times of different models.

Supplementary Figure 1. Prediction performance of CLCNet on ten traits across multiple crop datasets.

Supplementary Figure 2. Comparison the PCC and MSE of CA (Chromosome-Aware), GWAS (Genome Wide Association Study) and PCA (Principal Component Analysis) feature selection methods in rrBLUP.

Supplementary Figure 3. Comparison the PCC and MSE of CA (Chromosome-Aware), GWAS (Genome Wide Association Study) and PCA (Principal Component Analysis) feature selection methods in Bayesian Lasso.

Supplementary Figure 4. Comparison the PCC and MSE of CA (Chromosome-Aware), GWAS (Genome Wide Association Study) and PCA (Principal Component Analysis) feature selection methods in Bayesian Ridge.

Supplementary Figure 5. Comparison the PCC and MSE of CA (Chromosome-Aware), GWAS (Genome Wide Association Study) and PCA (Principal Component Analysis) feature selection methods in DeepGS.

Supplementary Figure 6. Comparison the PCC and MSE of CA (Chromosome-Aware), GWAS (Genome Wide Association Study) and PCA (Principal Component Analysis) feature selection methods in LightGBM.

Supplementary Figure 7. Comparison the PCC and MSE of CA (Chromosome-Aware), GWAS (Genome Wide Association Study) and PCA (Principal Component Analysis) feature selection methods in SVR.

Supplementary Figure 8. Comparison the PCC and MSE of CA (Chromosome-Aware), GWAS (Genome Wide Association Study) and PCA (Principal Component Analysis)

feature selection methods in DNNGP.

Supplementary Figure 9. Model architecture diagrams.

Supplementary Figure 10. Analysis of CLCNet's core modules on MSE.

Supplementary Figure 11. Model MSE performance under SNP position perturbation.

Supplementary Figure 12. Relationship between SHAP-based SNP importance derived from CLCNet and GWAS for EW.

Supplementary Figure 13. Relationship between SHAP-based SNP importance derived from CLCNet and GWAS for PH.

REFERENCES

Basak, D., Pal, S., and Patranabis, D.C. (2007). Support vector regression. Neural Inf. Process. Lett. Rev. 11: 203–224.

Budhlakoti, N., Kushwaha, A.K., Rai, A., Chaturvedi, K., Kumar, A., Pradhan,A.K., Kumar, U. et al. (2022). Genomic selection: a tool for accelerating the efficiency of molecular breeding for development of climate-resilient crops. Front. Genet. 13, 832153.

Crossa, J., de los Campos, G., Pérez, P., Gianola, D., Burgueño, J., Araus, J. L., Makumbi, D., Singh, R. P., Dreisigacker, S., Yan, J., et al. (2010). Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics, 186:713-724.

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., de los Campos, G., Burgueño, J., González-Camacho, J. M., Pérez-Elizalde, S., Beyene, Y., Dreisigacker, S., Singh, R., Zhang, X., Gowda, M., Roorkiwal, M., Rutkoski, J., and Varshney, R. K. (2017). Genomic selection in plant breeding: Methods, models, and perspectives. Trends Plant Sci. 22: 961-975.

Endelman, J.B. (2011). Ridge regression and other kernels for genomic selection with R package rrBLUP. Plant Genome 4: 250-255.

Eraslan, G., Avsec, Ž., Gagneur, J., and Theis, F.J. (2019). Deep learning: new computational modelling techniques for genomics. Nat. Rev. Genet. 20, 389-403.

Fu, B., Pazokitoroudi, A., Shi, Z., Kar, A., Xue, A., Anand, A., Anand, P., Liu, Z., Border, R., Pajukanta, P., Zaitlen, N., and Sankararaman, S. (2025). A biobank-scale test of marginal epistasis reveals genome-wide signals of polygenic interaction effects. Nature Genetics 57, 3175–3184.

Gao, P., Zhao, H., Luo, Z., Lin, Y., Feng, W., Li, Y., Kong, F., Li, X., Fang, C., and Wang, X. (2023). SoyDNGP: A web-accessible deep learning framework for genomic prediction in soybean breeding. Brief. Bioinform. 24(6): 1-12.

Gao, S., Yu, T., Rasheed, A., Wang, J., Crossa, J., Hearne, S., and Li, H. (2025). Fast-forwarding plant breeding with deep learning-based genomic prediction. J. Integr. Plant Biol. 00: 1-6.

González-Camacho, J.M., Crossa, J., Pérez-Rodríguez, P., Ornella, L., and Gianola, D. (2016). Genome-enabled prediction using probabilistic neural network classifiers. BMC Genomics 17: 1-16.

Habier, D., Fernando, R. L., and Dekkers, J. C. M. (2007). The impact of genetic relationship information on genome-assisted breeding values. Genetics, 177(4), 2389-2397.

Habier, D., Fernando, R.L., Kizilkaya, K., and Garrick, D.J. (2011). Extension of the Bayesian alphabet for genomic selection. BMC Bioinformatics 12: 186.

Heslot, N., Yang, H.-P., Sorrells, M. E., and Jannink, J.-L. (2012). Genomic selection in plant breeding: A comparison of models. Crop Science, 52(1), 146-160.

Ho, D. S. W., Schierding, W., Wake, M., Saffery, R., and O'Sullivan, J. (2019). Machine learning SNP based prediction for precision medicine. Front. Genet.10, 267.

Huang, J., Yang, Z., Yin, M., Li, C., Li, J., Wang, Y., Huang, L., Liang, C., He, F., Liang, C., Li, M., Han, R., Jiang, Y. (2024). CLCNet: a contrastive learning and chromosome-aware network for genomic prediction in plants. Preprint at bioRxiv.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In Advances in Neural Information Processing Systems, Vol. 30.

Li Y., Ren S., Li J., Lee J., Wan J., and Gan X. (2026). MeNet: A mixed-effect deep neural network for multi-environment genomic prediction of agronomic traits. Plant Comm. 7, 101620.

Li, J., Zhang, D., Yang, F., Yang, J., Wang, K., and Zhao, C. (2024). TrG2P: A transfer-learning-based tool integrating multi-trait data for accurate prediction of crop yield. Plant Commun. 5: 100975.

Liu, H. J., and Yan, J. (2019a). Crop genome-wide association study: A harvest of biological relevance. Plant J. 97: 8-18.

Liu, J., Li, J., Wang, H., and Yan, J. (2020). Application of deep learning in genomics. Science China Life Sciences, 63, 1860-1878.

Liu, Y., Wang, D., He, F., Wang, J., Joshi, T., and Xu, D. (2019b). Phenotype prediction and genome-wide association study using deep convolutional neural network of soybean. Front. Genet. 10: 1091.

Lundberg SM and Lee S I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems 30: 4765–4774.

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

Ma, X., Wang, H., Wu, S., Han, B., Cui, D., Liu, J., Zhang, Q., Xia, X., Song, P., Tang,

C., Geng, L., Yang, Y., Yan, S., Zhou, K., and Han, L. (2024). DeepCCR: Large-scale genomics-based deep learning method for improving rice breeding. Plant Biotechnol. J. 1-3.

MacNish, TR., Bergmann, T., Edwards, D. (2026). Haplotype applications in genomic selection. Genome Biology.

Maenhout, S., De Baets, B., Haesaert, G., and Van Bockstaele, E. (2007). Support vector machine regression for the prediction of maize hybrid performance. Theoretical and Applied Genetics, 115(7), 1003-1013.

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

Park, T., and Casella, G. (2008). The Bayesian Lasso. J. Am. Stat. Assoc. 103: 681–686.

Pérez, P., de Los Campos, G., Crossa, J., and Gianola, D. (2010). Genomic-enabled prediction based on molecular markers and pedigree using the Bayesian linear regression package in R. Plant Genome 3.

Pudjihartono, N., Fadason, T., Kempa-Liehr, A.W., and O'Sullivan, J.M. (2022). A review of feature selection methods for machine learning-based disease risk prediction. Front. Bioinform. 2: 927312.

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

Ren, Y., Wu, C., Zhou, H., Hu, X., and Miao, Z. (2024). Dual-extraction modeling: A multi-modal deep-learning architecture for phenotypic prediction and functional gene mining of complex traits. Plant Commun. 5: 100102.

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

Urbanowicz, R. J., Meeker, M., La Cava, W., Olson, R. S., and Moore, J. H. (2018). Relief-based Feature Selection: Introduction and Review. J. Biomed. Inf. 85,189-203.

Wang, H., Cimen, E., Singh, N., and Buckler, E. (2020). Deep learning for plant genomics and crop improvement. Curr. Opin. Plant Biol. 54: 34-41.

Wang, H., Yan, S., Wang, W., Chen, Y., Hong, J., He, Q., Diao, X., Lin, Y., Chen, Y., Cao, Y., Guo, W., and Fang, W. (2025). Cropformer: An interpretable deep learning framework for crop genomic prediction. Plant Commun. 6, 101223.

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

Wang, X., Xu, Y., Hu, Z., and Xu, C. (2018). Genomic selection methods for crop improvement: Current status and prospects. Crop J. 6: 330-340.

Washburn, J.D., Cimen, E., Ramstein, G., Reeves, T., O'Briant, P.,McLean, G., Cooper, M., Hammer, G., and Buckler, E.S. (2021). Predicting phenotypes from genetic, environment, management, and historical data using CNNs. Theor. Appl. Genet. 134: 3997-4011.

Wu, C., Zhang, Y., Ying, Z., Li, L., Wang, J., Yu, H., Zhang, M., Feng, X., Wei, X., and Xu, X. (2023). A transformer-based genomic prediction method fused with knowledge-guided module. Brief. Bioinform. 25: 1-11.

Xu Y., Liu X., Fu J., Wang H., Wang J., Huang C., Prasanna B.M., Olsen M.S., Wang G., Zhang A. (2020). Enhancing genetic gain through genomic selection: From livestock to plants. Plant Communications 1: 100005.

Yan, J., Xu, Y., Cheng, Q., Jiang, S., Wang, Q., Xiao, Y., Ma, C., Yan, J., and Wang, X. (2021). LightGBM: Accelerated genomically designed crop breeding through ensemble learning. Genome Biology, 22, 271.

Yang, J., Lee, S.H., Goddard, M.E., and Visscher, P.M. (2011). GCTA: a tool for genome-wide complex trait analysis. Am. J. Hum. Genet. 88, 76–82.

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

Yin, M., Han, H., Liu, W., Kong, L., Li, X., Ji, W., Wu, J., Deng, G., Dong, H., Zhang, J., Zhou, S., Zhang, Z., Hou, Y., Wang, H., Zhao, C., Li, X., Luo, Q., Guo, X., Yang, X., Ling, H.-Q., He, F., and Li, L. (2025). Reshaping epistatic network modules enhances wheat yield potential. Preprint at bioRxiv.

Zhang, Y., and Yang, Q. (2018). An overview of multi-task learning. Natl. Sci. Rev. 5: 30-43.

Zhao, L., Tang, P., Luo, J., Liu, J., Peng, X., Shen, M., Wang, C., Zhao, J., Zhou, D., Fan, Z., Chen, Y., Wang, R., Tang, X., Xu, Z., and Liu, Q. (2025). Genomic prediction with NetGP based on gene network and multi-omics data in plants. Plant

Biotechnol. J. 23: 1190-1201.

Zhao, X., Sun, F., Li, J., Zhang, D., Zhang, Q., Liu, Z., Tan, C., Ma, H., and Wang, K. (2025). VMGP: A unified variational auto-encoder based multitask model for multi-phenotype, multi-environment, and crosspopulation genomic selection in plants. Artif. Intell. Agric. 15:829–842.