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, Complex Traits, Feature Selection, Contrastive Learning

ABSTRACT

Genomic selection (GS), which integrates genomic markers with phenotypic data, has emerged as a powerful breeding strategy for predicting phenotypes and estimating breeding values in candidate populations. The success of GS fundamentally depends on the accuracy of genomic prediction (GP) models. However, conventional GP models often struggle to capture inter-individual variability and are particularly challenged by the curse of dimensionality - where the number of features (*e.g.* single nucleotide polymorphism, SNP) far exceeds the number of samples-ultimately limiting their predictive performance. To overcome these limitations, we proposed Contrastive Learning and Chromosome-aware Network (CLCNet), a novel deep learning framework that, for the first time, integrates multi-task learning with contrastive learning for GP. CLCNet consists of two key components: (i) a contrastive learning module, designed to enhance the model ability to capture fine-grained genotype-dependent phenotypic differences across individuals; and (ii) a chromosome-aware module, which performs feature selection at both the chromosome level and genome-wide level to retain the most informative SNPs. The performance of CLCNet was evaluated across four major crop species-maize (*Zea mays*), cotton (*Gossypium hirsutum*), rapeseed (*Brassica napus*), and soybean (*Glycine max*)-across ten distinct traits. This evaluation involved a comprehensive comparison with three classical linear models (rrBLUP, Bayesian Ridge, Bayesian Lasso), two machine learning models (LightGBM, SVR), and two deep learning models (DNNGP, DeepGS). CLCNet consistently outperformed all comparator models, surpassing the average of classical linear models by ~3.84%, the best-performing machine learning model by 4.84%, and leading deep learning models by over 5%. These results highlight CLCNet as a

promising and robust tool for accelerating genetic gain through more accurate GS in plant breeding.

INTRODUCTION

Genomic selection (GS) has become an increasingly important approach in modern plant and animal 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). Unlike traditional marker-assisted selection (MAS), which typically focuses on loci with large effects, GS captures the aggregate contribution of loci across the whole genome including those with small effects, thus enabling more comprehensive and accurate selection decisions (Crossa et al., 2017; Xu et al., 2020; Budhlakoti et al., 2022). A wide range of genomic prediction (GP) models, including linear models (*e.g.*, rrBLUP, GBLUP, BayesianA/B/C/Cπ) and machine learning methods (*e.g.*, support vector regression, 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), has been developed for GS. More recently, deep learning (DL) has garnered attention for its capacity to model complex, nonlinear genotype-phenotype relationships through hierarchical architectures (Liu et al., 2020; Wang et al., 2020; Gao et al., 2025), and it holds great potential for applications in GS.

Despite significant advances, critical challenges, most notably the "curse of dimensionality," still confront GS. Specifically, high-throughput genotyping platforms generate hundreds of thousands to millions of single nucleotide polymorphisms (SNPs), vastly exceeding available phenotypic samples. This considerable imbalance 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). To address these issues, the process of feature selection has been identified as a critical step in enhancing the signal-to-noise ratio and reducing the complexity of the model. In this study, a chromosome-aware feature selection strategy is developed, which organizes genomic features according to their chromosomal origin. Such an approach has been demonstrated to facilitate the generation of more interpretable and biologically grounded representations, whilst concomitantly preserving structural information both within individual chromosomes and across the entire genome.

In addition to these above developments, DL-based GP models have also gained traction across various learning paradigms, including single-task, multi-modal, and transfer learning (Eraslan et al., 2019; Washburn et al., 2021). Representative examples of 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 illustrate multi-modal learning (Ren et al., 2024; Ma et al., 2024; Yao et al., 2025), and TrG2P exemplifies transfer learning (Li et al., 2024). Despite its potential to enhance model generalization under data-scarce conditions, multi-task learning (MTL)-which provides a natural framework for jointly modeling multiple traits and facilitating information sharing among related tasks-remains underexplored and its systematic application in GP is limited.

To address these gaps, 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). CLCNet demonstrated consistent superiority over all comparator models, surpassing the mean performance of classical linear models by approximately 3.84%, the top-performing machine learning model by 4.84%, and outperforming leading deep learning models by more than 5%. These results underscore CLCNet as a promising and robust tool for accelerating genetic gain through more accurate GP in plant breeding.

RESULTS

1. The architectural design of CLCNet

CLCNet is a multi-task model designed to predict phenotypic traits from SNP data as its primary objective, while concurrently generating low-dimensional SNP feature representations as an auxiliary task. As shown in Figure 1A, the model architecture features a dual-pathway backbone, consisting of one path with three non-linear layers and another with a single linear layer, integrated via a residual connection. The resulting outputs are fed into separate branches corresponding to the primary task for computing the mean squared error (MSE) between predicted and observed phenotypic values, and the auxiliary task for the contrastive loss based on inter-sample differences, respectively.

The framework integrates two key innovations. First, a chromosome-aware module is employed for SNP preprocessing. In this module, SNPs are grouped by chromosome, and a feature selection algorithm is applied to identify informative variants. Specifically, LightGBM is used to identify informative variants based on information gain (IG > 0),

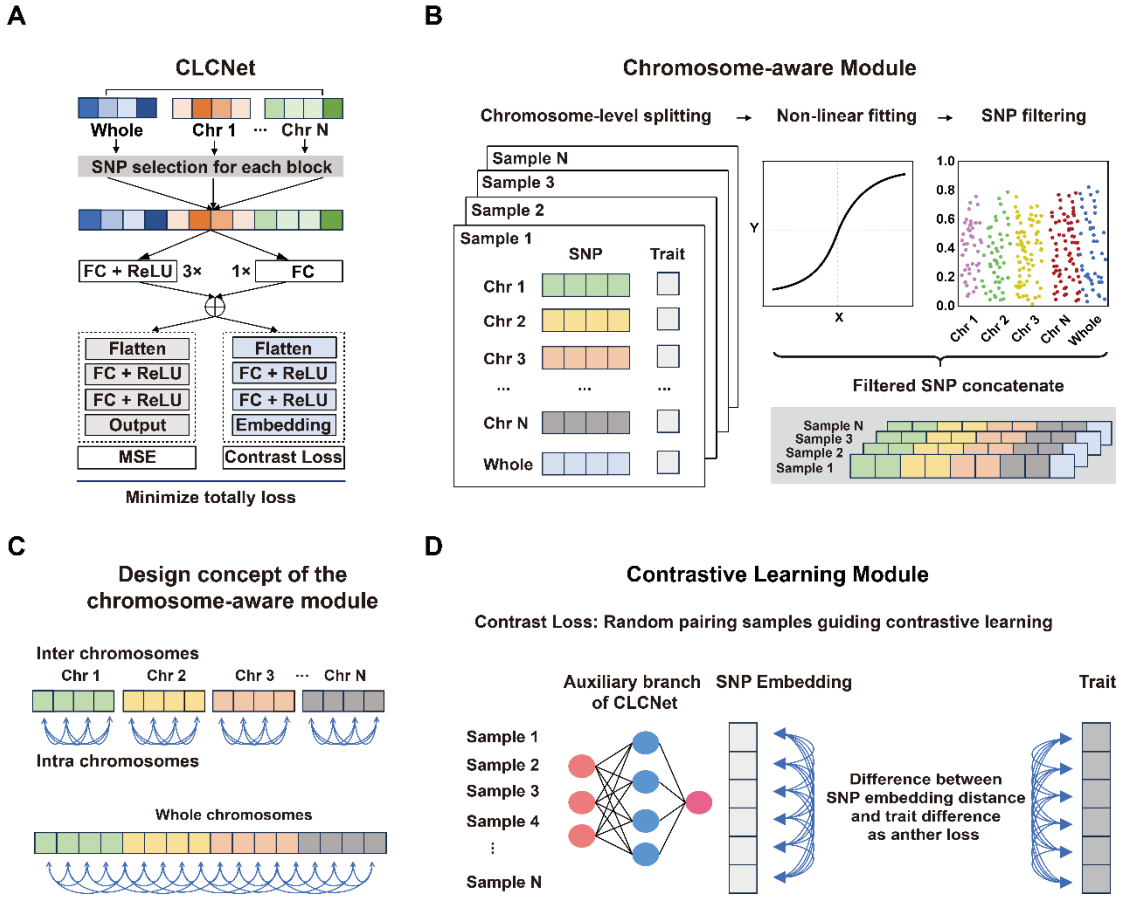

Figure 1. The Framework of CLCNet Model.

(A) CLCNet architecture. CLCNet is a multi-task model that predicts phenotypic traits from SNP data while also extracting compact SNP feature representations. Its architecture features a dual-path backbone that combines nonlinear and linear transformations via a residual connection, allowing it to efficiently handle both the main prediction task and auxiliary objectives. **(B) Chromosome-aware module.** SNPs are first grouped according to their respective chromosomes. Feature selection is then performed using non-linear models capable of capturing complex interactions among features, such as LightGBM, multilayer perceptron (MLP), or convolutional neural networks (CNNs). The module conducts local selection, which identifies important SNPs within each individual chromosome, as well as global selection, which captures influential SNPs. The outputs from both selection strategies are concatenated to form the final input. **(C) Design concept of the chromosome-aware module.** By applying non-linear modeling to both individual chromosome SNP sets and the genome SNP set in relation to phenotypic traits, this module enables the selection of SNPs that capture both intra-chromosomal and inter-chromosomal signals. **(D) Contrastive learning module**. Samples are randomly paired each epoch and processed in parallel. The model learns to predict phenotypic outcomes while simultaneously producing feature vectors. Optimization is driven by two type loss functions: mean squared error (MSE) for phenotype prediction, and contrastive loss, which encourages the Euclidean distance between SNP embedding to reflect phenotypic differences between paired samples.

at both the local (per-chromosome) and global (whole-chromosome) levels (Figure 1 B). The model input is constructed by concatenating SNPs selected at both intra- and inter-chromosomal levels (Figure 1 C). Second, a contrastive learning module is introduced to enhance the training process. This module employs a random pairing strategy during each training epoch and optimizes a composite loss function that combines phenotypic prediction loss with a custom contrastive loss. Specifically, the contrastive loss is defined as the difference between the Euclidean distance of the paired samples' SNP embeddings and the absolute difference in their phenotypic values, encouraging the model to learn meaningful sample-level distinctions (Figure 1D). Our framework integrates contrastive learning within a multi-task setting, allowing the model to simultaneously learn trait predictions and sample-level variation representation.

2. Superior performance of CLCNet over other prediction 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, encompassing over 30,000 individuals and approximately 224 billion genotype records (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 10 agronomically important traits, including both developmental and quality-related phenotypes: 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). Detailed summaries of the processed datasets are provided in Table S1.

As shown in Figure 2, the baselines included three classical statistical models (rrBLUP, Bayesian Ridge, and Bayesian Lasso), two machine learning algorithms (LightGBM and SVR), and two deep learning frameworks (DNNGP and DeepGS). Among these, rrBLUP achieved the highest PCC on majority of traits (7 out of 10). In

contrast, SVM, DNNGP, and Bayesian Ridge showed the best performance on PH, EW, and PC, respectively. However, CLCNet consistently outperformed all baseline models across all ten traits, thereby establishing itself as the most robust and accurate predictor. Compared with the strongest baseline for each trait, CLCNet achieved relative improvements of +0.40% for DTT, +0.21% for EW, +0.02% for PH, +0.43% for FT, +1.07% for OC, +3.54% for PC, +2.29% for HSW, +0.63% for SL, +1.07% for FStr, and +2.10% for FMic. These gains are particularly significant for complex traits such as PC, HSW, and FMic, which are generally characterized by weak genetic signals, polygenic control, or high environmental noise. Notably, even for well-predicted traits such as DTT and PH, where most models already perform strongly, CLCNet still delivered measurable improvements. Overall, CLCNet demonstrated superior robustness and predictive accuracy.

Compared with traditional statistical models, CLCNet exhibited apparent advantages. Although rrBLUP performed competitively and served as the best overall baseline, CLCNet consistently outperformed it, particularly on genetically complex traits such as PC and OC, where absolute gains of +3.55% and +1.07% were observed, respectively (Figure 2 E and F). Bayesian Ridge and Lasso offered only modest improvements on a few traits and were generally less competitive, reflecting the limitations of their linear priors in capturing non-linear interactions and epistasis.

As for the machine learning models, LightGBM and SVR performed well on certain morphological traits, such as PH and EW, but their performance deteriorated significantly for other traits (Figure 2 B and C). For instance, LightGBM achieved the lowest PCC for PC and OC, with values of 21.14% and 25.7%, respectively. In contrast, CLCNet achieved PCC values of 28.80% and 41.53% for these traits, indicating substantial gains of +7.66% and +15.83%, respectively (Figure 2 E and F). Similarly,

SVR underperformed compared to CLCNet, on FMic and PC, with PCC values of 66.34% vs. 70.19% and 22.57% vs. 28.80%, respectively. These results underscore the limitations of tree-based and kernel-based models in handling nonlinear genomic signals (Figure 2 J).

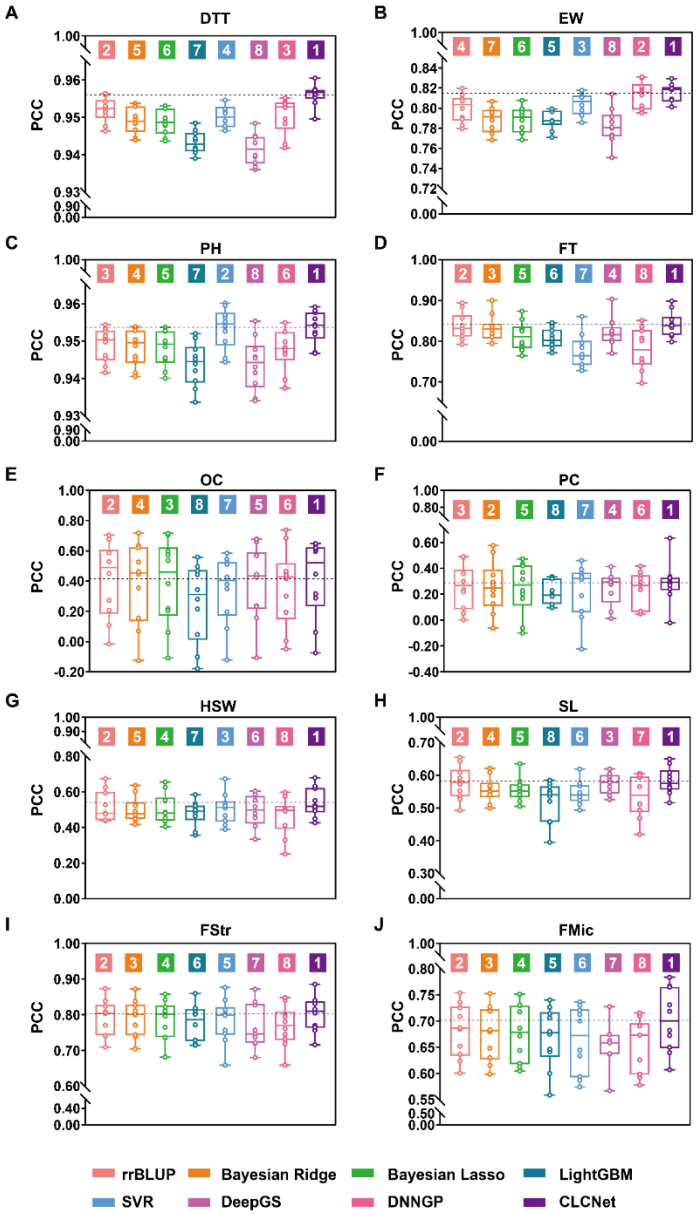

Figure 2. Prediction performance of CLCNet on ten traits across multiple crop datasets: (A) maize DTT, (B) maize EW, (C) maize PH, (D) rapeseed FT, (E) rapeseed OC, (F) rapeseed PC, (G) soybean HSW, (H) soybean SL, (I) cotton FStr and (J) cotton FMic. Trait abbreviations correspond to: DTT-days to tasseling, 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 value of CLCNet. Numbers indicated the ranking of the sort.

Among deep learning models, CLCNet's advantages became more pronounced for traits like HSW and PC, where it delivered absolute gains exceeding +5%, as well as on FMic (+3.02%) and OC (+1.44%). These improvements are particularly meaningful for traits known for their complex inheritance patterns and lower predictive accuracy under conventional models. Furthermore, compared to DeepGS, CLCNet improved performance by +2.12% on OC and +3.97% on PC, confirming its superior ability to model phenotype-genotype relationships with higher biological fidelity.

Similar superiority of CLCNet was also observed when evaluated using MSE (Figure S1). Notably, CLCNet achieved the lowest MSE on eight out of ten traits. On complex biochemical traits, where baseline models tended to exhibit larger variance, CLCNet's lower MSE further confirmed its higher accuracy and stability. For example, compared to rrBLUP, CLCNet reduced MSE by −2.33% on OC, −2.33% on PC, and −2.27% on FMic, showing consistent advantages across both correlation-based and error-based metrics.

3. Chromosome-aware module demonstrate strong feature extraction capability

Feature extraction is crucial in GP due to the high dimensionality and inherent complexity of genomic datasets, which often encompass vast numbers of SNPs. The primary objective is to isolate informative markers that are strongly associated with the target phenotype while eliminating redundant or irrelevant variables. This process not only enhances model accuracy and generalizability but also reduces computational overhead and the likelihood of overfitting.

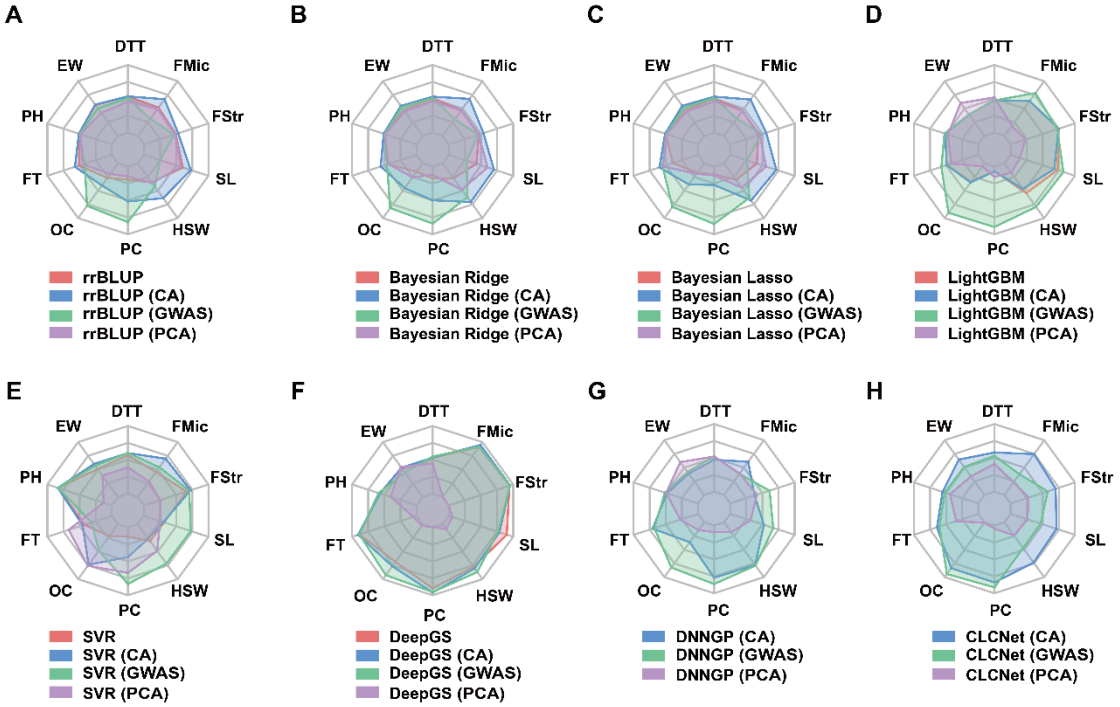

Figure 3. Comparison the PCC of CA (chromosome-aware module), GWAS (genome wide association study) and PCA (principal component analysis) feature selection methods on eight different models cross ten traits. (A) rrBLUP, (B) Bayesian Ridge, (C) Bayesian Lasso, (D) LightGBM, (E) SVR, (F) DeepGS, (G) DNNGP, (H) CLCNet.

To address the challenges of high dimensionality and complexity in genomic datasets, a chromosome-aware (CA) module, a novel feature selection strategy designed to capture both local and global genomic dependencies in a nonlinear manner, was introduced. This stagey is compared with traditional feature selection methods such as those based on genome-wide association studies (GWAS) and principal component analysis (PCA). As shown in Figure 3, the CA module consistently demonstrated superior performance and robustness across various model types. For instance, when integrated with rrBLUP, the CA module outperformed PCA-based and GWAS-based feature selection by 1.2% on FT, 2.11% on OC, and 1.47% on PC. Similarly, in SVR, the CA module improved performance on FMic and EW, with PCC values of 69.12% vs. 66.34% and 81.32% vs. 80.39%, respectively, indicating better stability on noisy traits. In LightGBM, although GWAS showed high scores for OC and PC, the CA module offered better consistency across traits like FMic and hundred-seed weight

(HSW).

Deep learning models benefited significantly from the CA module. For instance, DNNGP (CA) outperformed its PCA-based version on PC, HSW, and FT, with gains of 6–8%. DeepGS (CA) also showed consistent improvement across traits, particularly FMic (+2.43%) and FT (+0.91%). Importantly, CLCNet (CA) achieved the highest performance across all traits, including DTT, EW, PH, HSW, and FMic, confirming that chromosome awareness is complementary.

While GWAS filtering performed well for additive traits like OC and PC in specific models, it showed less generalizability. PCA, although effective at dimensionality reduction, frequently removed biologically relevant features, resulting in substantial performance drops-*e.g.*, DeepGS (PCA) on PC (-6.88%) and OC (-17.21%), demonstrating its limitations in retaining meaningful SNP signals. In contrast, the CA module proved to be the most effective and stable feature selection strategy, particularly in enhancing model accuracy on traits with complex genetic architecture, outperforming generic statistical and univariate filtering methods across both classical and deep learning models.

The model performance was further evaluated using MSE as a complementary metric across the same ten traits and models (Figure S2). Overall, MSE results were highly consistent with PCC-based evaluations and further confirmed the advantages of CA-based filtering. Across nearly all model categories, incorporating CA features led to lower MSEs compared to PCA and GWAS, particularly on complex or low-heritability traits such as OC, PC, FMic, and HSW. For instance, CLCNet with CA selection achieved the lowest MSE for FT, PH, and OC, demonstrating improved

precision in modeling flowering time and biochemical traits. These findings collectively demonstrate that CA feature selection not only enhances PCC but also consistently reduces prediction error (MSE), especially in scenarios where models face challenges of high dimensionality, non-linearity, or phenotype complexity.

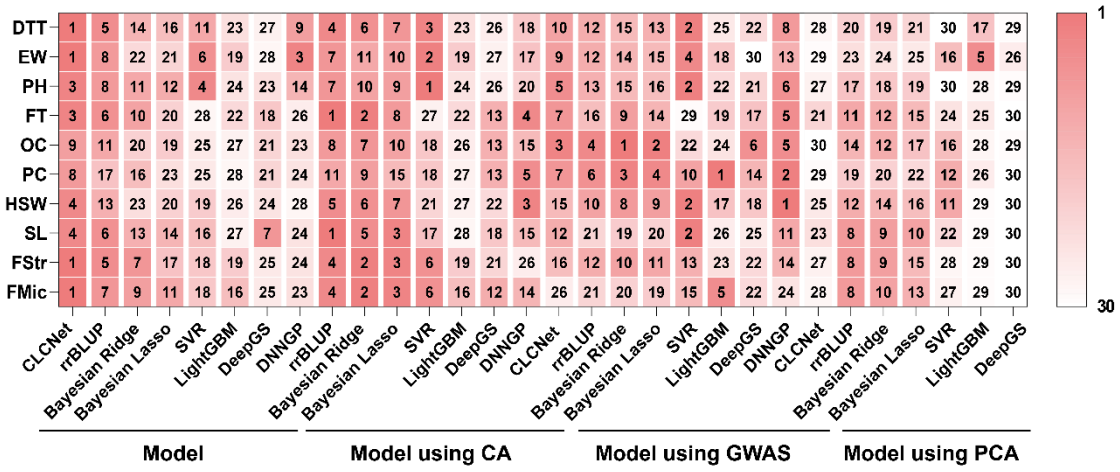

Figure 4. PCC rank distribution of feature selection strategies in eight models across 10 traits. Eight models were evaluated across ten traits using four SNP selection strategies: raw SNPs, CA-based selection, GWAS-based selection, and PCA-based selection. The ranks are based on PCC reflecting predictive performance.

Overall, CA-based methods not only achieve higher predictive accuracy but also maintain more stable rankings across diverse model families and trait types, underscoring the effectiveness of structure-aware feature selection.

Furthermore, the performance of different models was ranked under various feature selection strategies-including the use of raw features-across ten traits based on both Pearson correlation coefficient (PCC) (Figure 4) and mean squared error (MSE) (Figure S3). The heatmap illustrates the rank distribution (1-30) of 30 model-feature selection combinations, where darker colors indicate better rankings and lighter colors represent poorer performance. Models incorporating the CA (Chromosome-Aware) feature selection method clearly outperform others, frequently securing top positions across

288 most traits. Notably, combinations such as CLCNet (CA), DNNGP (CA), and rrBLUP
289 (CA) consistently rank among the top five performers on both morphological traits (*e.g.*,
290 DTT, PH, SL) and complex biochemical traits (*e.g.*, OC, PC, FMic), demonstrating the
291 robustness and generalizability of the CA approach.

292 In contrast, PCA-based models tend to rank lower, particularly for traits with
293 complex genetic architectures. This suggests that unsupervised dimensionality
294 reduction may discard informative genetic signals-as evidenced by the consistently
295 poor rankings of models such as DeepGS (PCA) and LightGBM (PCA). GWAS-based
296 variants perform well on a subset of additive traits, such as OC and PC, but show greater
297 variability across other traits. The MSE-based heatmap (Supplementary Figure 3)
298 reveals trends consistent with those observed in the PCC-based rankings, further
299 validating these findings.

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

301 To evaluate the individual contributions of CLCNet's core components, comprehensive
302 ablation studies were conducted across all phenotypic prediction tasks by assessing four
303 modified models: CLCNet_1, CLCNet_2, CLCNet_3, and CLCNet_4. These ablation
304 variants were specifically designed to isolate and quantify the functional impact of each
305 module-namely, the chromosome-aware feature selection and the contrastive learning
306 modules-on the overall predictive performance.

307 In CLCNet_1, the global (whole-chromosome) feature selection module was
308 removed, retaining only local (per-chromosome) SNP selection. This design aims to
309 assess the added value of integrating genome-wide signals beyond chromosome-level
310 structure. CLCNet_2 builds upon CLCNet_1 by further removing the contrastive
311 learning module, thereby simplifying the architecture from multi-task to single-task
312 learning. This enables evaluation of the unique contribution of contrastive learning

when only local feature information is available.

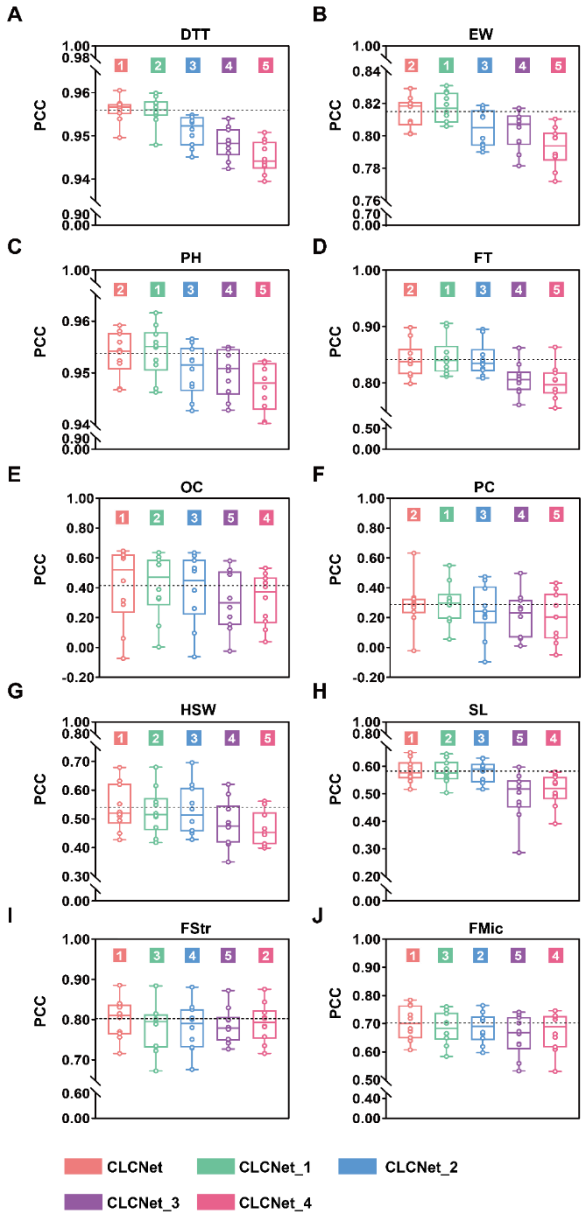

Figure 5. Analysis of CLCNet's core modules on PCC. (A) maize DTT, (B) maize EW, (C) maize PH, (D) rapeseed FT, (E) rapeseed OC, (F) rapeseed PC, (G) soybean HSW, (H) soybean SL, (I) cotton FStr and (J) cotton FMic.

In CLCNet_3, the local feature selection module was excluded, preserving only global feature selection. This tests whether global signals alone are sufficient for accurate prediction, and how their performance compares to the inclusion of fine-grained local structure. CLCNet_4 further removes the contrastive learning module from CLCNet_3, resulting in another single-task variant. This minimal design serves as

a baseline to measure the cumulative benefits of local feature selection and contrastive supervision. The architectural differences among these models are illustrated in Figure S4.

Across all ten traits, the full CLCNet model consistently achieved the best performance, confirming the synergistic effectiveness of integrating global and local feature selection with contrastive learning (Figure 5). In terms of PCC, CLCNet_1, which retains local selection but excludes the global module, showed nearly identical performance to the full model on most traits-*e.g.*, DTT, PH, and FT-demonstrating that local feature selection alone captures highly informative signals. In contrast, CLCNet_3, which removes local selection, resulted in larger performance drops, especially for complex biochemical traits (*e.g.*, PC), highlighting the critical role of local structure-aware features.

These patterns are further corroborated by the MSE results, which follow an inverse but consistent trend with PCC (Supplementary Figure 5). The full CLCNet achieved the lowest error across traits (*e.g.*, PC), while CLCNet_1 maintained similar MSE values, supporting the robustness of local selection. Removing the contrastive module in CLCNet_2 caused moderate error increases, particularly on PC and OC, reaffirming its contribution to generalization. Among all ablation variants, CLCNet_3 and CLCNet_4 showed the largest increases in MSE across multiple traits. For instance, on PC, the MSE increased from 0.7999 (CLCNet) to 1.007 in CLCNet_3 (+25.9%) and 0.9885 in CLCNet_4 (+23.5%), clearly indicating that removing local feature selection severely compromises predictive accuracy.

Overall, both PCC and MSE analyses reveal that local feature selection is the most crucial module, while contrastive learning provides complementary improvements, especially for traits with complex genetic architectures. Global selection also

contributes, though its effects are more modest. The full CLCNet model demonstrates the best balance of accuracy and stability, affirming the effectiveness of its modular design.

5. Chromosome-aware (CA) module leverages chromosomal structure for accurate SNP selection

To determine whether the CA module truly encodes structural dependencies within chromosomes, a perturbation experiment was conducted to simulate chromosome-level disturbances. Specifically, SNP positions within individual chromosomes were randomized to break the native patterns of linkage disequilibrium and epistatic interactions, thereby destroying the positional context critical for chromosome-aware modeling. This allowed an assessment of whether GP models could detect and adapt to structural shifts.

As expected, the output of the CA module was altered by these positional perturbations, resulting in different sets of selected SNPs before and after shuffling (Figure S4 E). Moreover, the explanatory power of the newly selected SNPs was reduced, leading to a measurable decline in predictive performance. These results underscore the importance of explicitly modeling positional structure and validate the CA module's capability to encode chromosomal information essential for robust GP.

Compared to the original CLCNet, the CLCNet_Shuffle variant-trained on the identical data but with randomized SNP positions-exhibited a consistent decline in prediction accuracy across nearly all traits (Figure 6). For example, the PCC decreased on key traits such as PC (-4.92%) and FMic (-3.45%), indicating that disrupting positional information impaired the model’s ability to capture biologically meaningful interactions.

This trend is further corroborated by the MSE results, where higher errors were

observed after positional perturbation (Supplementary Figure 6). Notably, traits like FT (+4.7%), PC (+1.5%), and FMic (+2.7%) showed measurable increases in error, indicating reduced predictive precision. Even for morphological traits like DTT and PH, modest MSE increases (e.g., DTT: +8.8%) reflect a consistent negative impact.

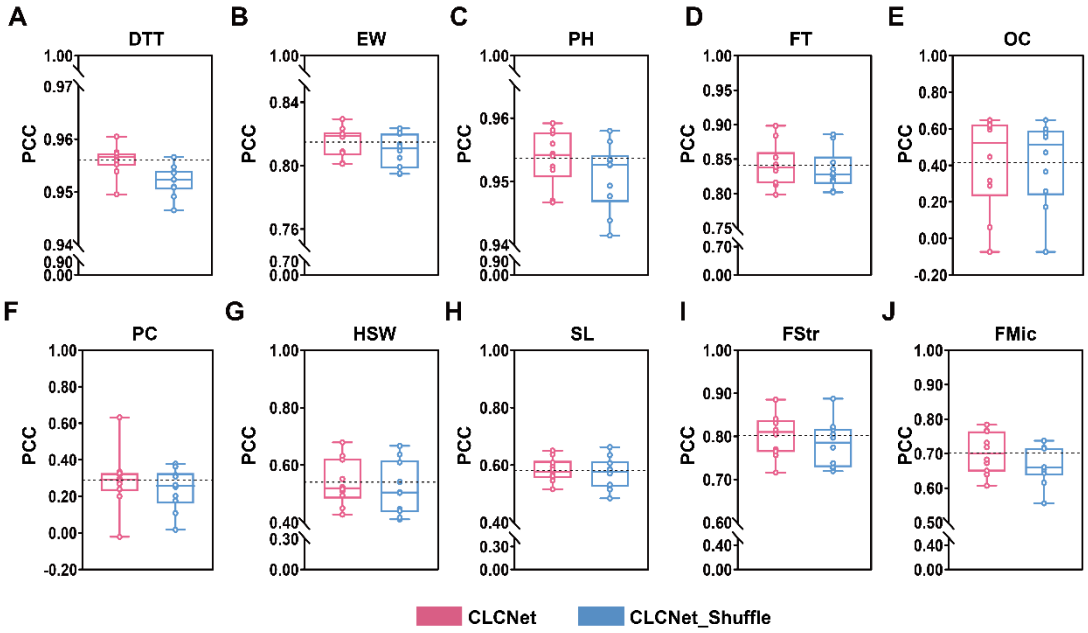

Figure 6. Model PCC performance under SNP position perturbation. (A) maize DTT, (B) maize EW, (C) maize PH, (D) rapeseed FT, (E) rapeseed OC, (F) rapeseed PC, (G) soybean HSW, (H) soybean SL, (I) cotton FStr and (J) cotton FMic.

Together, these results highlight the critical role of SNP positional information in the effectiveness of chromosome-aware modeling. The performance degradation under random perturbation demonstrates that CLCNet successfully captures structured genomic signals, and that preserving chromosomal SNP architecture is essential for maintaining predictive robustness.

DISCUSSION

The chromosome-aware module developed in this study serves as an effective feature selection strategy. In genomic data, where the number of features far exceeds the number of samples, a challenge commonly referred to as the “curse of dimensionality”-

391 dimensionality reduction is critical for enhancing both computational efficiency and
392 predictive accuracy (Pudjihartono et al., 2022). Feature selection, in particular,
393 mitigated noise and redundancy by identifying the most informative and representative
394 SNPs (Ho et al., 2019; Urbanowicz et al., 2019). Although traditional GWAS-based
395 feature selection approaches are useful, they usually rely on linear models to evaluate
396 the association between individual SNPs and phenotypes, thereby overlooking potential
397 non-linear interactions (Liu et al., 2019a). Other GP feature selection methods, such as
398 Principal Component Analysis (PCA), Maximal Information Coefficient (MIC), and
399 Pearson-Collinearity Selection (PCS), often fail to account for both inter- and intra-
400 chromosomal interactions (Wang et al., 2023; Wang et al., 2025; Zhao et al., 2025). To
401 overcome these limitations, an explicit chromosome-based grouping strategy was
402 implemented in this work: SNPs within individual chromosomes were used to model
403 local interactions, while SNPs across whole chromosomes were incorporated to capture
404 global interactions. Since feature selection was performed using non-linear models,
405 many of these global interactions likely reflect epistatic effects, which have been
406 previously reported to play important roles in complex trait prediction (Yan et al., 2021).
407 Non-linear models such as Random Forests (RF), LightGBM, convolutional neural
408 networks (CNNs), and recurrent neural networks (RNNs) vary in complexity and
409 should be selected according to the characteristics of the data and computational
410 constraints. The resulting local and global SNP subsets were then merged into a unified
411 set of representative features, with each type complementing the other. Like other
412 methods, the chromosome-aware module can also be readily integrated into different
413 GP models, enhancing its versatility and practical utility. This adaptability makes it as
414 a powerful tool for data preprocessing across a broad spectrum of GP tasks.
415 Furthermore, CLCNet adopts a multi-task learning framework, distinguishing it from

conventional single-task, multi-modal, and transfer learning methods. Previous studies have demonstrated that multi-task learning can improve performance across related tasks by leveraging shared information, highlighting its potential as an exciting area of research in DL (Zhang et al., 2018). In this study, contrastive learning was incorporated into the multi-task framework of CLCNet, thereby improving both phenotypic value prediction and the extraction of low-dimensional SNP representations. Contrastive loss is computed by comparing the Euclidean distance between the SNP embeddings of randomly paired samples with the absolute difference in their phenotypic values. This design reinforces inter-task correlations, introduces complementary information into the shared learning process, and promotes continuous refinement of model parameters. To the best of our knowledge, this is the first approach to apply a multi-task learning framework based on contrastive learning to GP models.

In conclusion, we have developed CLCNet, a novel deep learning model for plant GP. The integration of chromosome-aware and contrastive learning modules into the CLCNet model has been demonstrated to yield superior prediction accuracy in comparison to traditional linear models, machine learning models, and other deep learning approaches. The model effectively captures both local and global epistasis, and optimises molecular marker selection, thereby reducing data noise and enhancing overall performance. Furthermore, the multi-task modelling strategy employed by CLCNet, which incorporates the prediction of phenotypic differences as an auxiliary task, has been shown to promote enhanced inter-task information sharing, thereby resulting in increased prediction accuracy. It is evident that, on the whole, CLCNet is a reliable, efficient, and powerful tool for GS. It is particularly effective when dealing with high-dimensional datasets and offers considerable potential for plant breeding applications.

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) for linkage disequilibrium (LD) pruning (*--indep-pairphase 100 10 0.2*). The final datasets, representing maize (*Zea mays*), cotton (*Gossypium hirsutum*), rapeseed (*Brassica napus*), and soybean (*Glycine max*), include 10 complex traits. Preprocessed datasets were summarized in Table S1.

The architecture of CLCNet

We developed CLCNet using PyTorch (v2.1.0, https://pytorch.org/). CLCNet is a multi-task deep learning model that takes genomic single-nucleotide polymorphism (SNP) data as input and outputs phenotypic predictions (primary task) alongside low-dimensional feature vectors (auxiliary task). The input receives an n × m matrix, where n represents the sample number and m represents the number of markers. The 0/1/2 genotype encoding was used. The model's backbone network employs a dual-pathway architecture: one pathway comprises three non-linear fully connected layers (ReLU activation; dimensions: 4096→2048→1024), while the other consists of a single linear fully connected layer (dimension: 1024). The outputs of the two pathways are summed at the 1024-dimensional level and subsequently passed to the primary and auxiliary task branches. Each branch is composed of three non-linear fully connected layers, with dimensions mapped as 4096→2048→1 for the primary task and 4096→2048→1024 for the auxiliary task.

The chromosome-aware module is essential for preprocessing genomic data. SNPs are grouped into chromosomal categories based on their genomic locations, followed

by local feature selection within each group to identify key SNPs (local SNPs). Global feature selection is then applied across the whole genome to capture comprehensive features (global SNPs). Local and global SNPs are concatenated to form the input features for the model. Feature selection is performed using the LightGBM algorithm, a tree-based approach to model non-linear relationships among SNPs, facilitating the identification of key loci (Yan et al., 2021). It identifies key SNP by retaining SNPs with positive information gain (IG > 0) and discarding those with IG = 0.

The contrastive learning module introduces a customized training strategy. During each epoch, when loading each sample, a random pairing mechanism selects another sample for pairing. Both samples are fed into the model simultaneously, and their phenotypic values and low-dimensional feature vectors are predicted. Model optimization is guided by three loss functions: the mean squared error (MSE) between the predicted and true phenotypic values for each of the two input samples (equation 1), and a custom contrastive loss based on the Euclidean distance between the feature vectors of the two samples and the difference in their phenotypic values (equation 2). The model aims to minimize the total loss, which is the sum of the MSE for both samples and the contrastive loss. The server was equipped with two Xeon Xeon(R) Gold 6330 CPU, 512 GB of memory, and an NVIDIA GeForce RTX3090 GPU.

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.

Model comparison and cross-validation

The predictive performance of CLCNet was evaluated against seven baseline models on benchmark datasets. These included three classical statistical methods (Ridge Regression Best Linear Unbiased Prediction, rrBLUP; Bayesian Ridge; Bayesian Lasso), two machine learning models (Light Gradient Boosting Machine, LightGBM; Support Vector Regression, SVR), and two deep learning models (Deep Neural Network Genomic Prediction, DNNGP; Deep Learning Genomic Selection, DeepGS). 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 (Ma et al., 2018). 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 (Wang et al., 2023).

Predictive accuracy was evaluated using the Pearson correlation coefficient (PCC). To assess the significance of performance differences between CLCNet and the best-performing baseline models, paired t-tests were conducted under a 10-fold cross-validation scheme. To avoid data leakage, phenotype values in the test set were z-normalized using the mean and variance calculated from the training set. All models were trained exclusively on the training data and evaluated on the corresponding test

514 partitions. CLCNet hyperparameters-learning rate, batch size, and epoch-were tuned
515 using the Optuna framework (Akiba et al., 2019), with search spaces of [0.01, 0.001,
516 0.0001], [16, 24, 32, 64], and [50, 80, 100, 200], respectively. The hyperparameters
517 settings of all models were applied to the benchmark dataset, as detailed in Table S2.

518 **Feature selection comparison**

519 To benchmark the effectiveness of our chromosome-aware module, we compared it
520 with two widely used feature selection strategies: GWAS-based p-value filtering and
521 principal component analysis (PCA). For GWAS-based filtering, SNPs were scored
522 using a mixed linear model implemented via GCTA (Yang et al., 2011), and markers
523 with significant associations ($p < 0.01$) were retained for model input. For PCA,
524 dimensionality reduction was performed using the *sklearn.decomposition.PCA*
525 function in Python. Consistent with DNNGP, we retained principal components that
526 collectively explained more than 98% of the total variance in the genotype data. Each
527 of the baseline models—rrBLUP, Bayesian Ridge, Bayesian Lasso, LightGBM, SVR,
528 DNNGP and DeepGS, was evaluated using features selected by all three strategies:
529 GWAS, PCA, and our chromosome-aware (CA) method. Model performance was
530 quantified by the Pearson correlation coefficient (PCC) and Mean Squared Error (MSE),
531 enabling a systematic comparison of feature selection impacts across statistical,
532 machine learning, and deep learning paradigms.

533 **SNP position perturbation**

534 The input is an n × m matrix, where n represents the number of samples and m denotes
535 the number of markers, as previously described and defined. The order of m reflected
536 the genomic positions of the markers. To perform SNP position perturbation, the order
537 of m was randomly shuffled across all n samples, disrupting the original SNP positions.
538 After shuffling, the input matrix shape remains unchanged. For CLCNet, chromosome-

level perturbation was handled with additional consideration. Specifically, prior to perturbation, SNPs were split into individual chromosomes based on their original lengths. After shuffling the entire marker set, segments corresponding to the original chromosome lengths were reassigned as pseudo-chromosomes in the perturbed dataset. Both the perturbed single-chromosome inputs and the whole shuffled genome were then used as inputs to CLCNet for prediction.

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).

Code availability

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

Table S1. Benchmark dataset.

Table S2. Model hyperparameter settings.

Figure S1. Prediction performance of CLCNet on ten traits across multiple crop datasets.

Figure S2. Comparison the MSE of CA (chromosome-aware module), GWAS (genome wide association study) and PCA (principal component analysis) feature selection methods on eight different models cross ten traits.

Figure S3. MSE rank distribution of feature selection strategies in eight models across ten traits.

Figure S4. Model architecture diagrams.

Figure S5. Analysis of CLCNet's core modules on MSE.

Figure S6. Model MSE performance under SNP position perturbation.

References

Abdollahi-Arpanahi, R., Gianola, D., and Peñagaricano, F. (2020). Deep learning versus parametric and ensemble methods for genomic prediction of complex phenotypes. Genet. Sel. Evol. 52:1-15.

Akiba, T., Sano, S., and Yanase, T. (2019). Optuna: A next-generation hyperparameter optimization framework. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 2623-2631.

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.

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.

614 Habier, D., Fernando, R.L., Kizilkaya, K., and Garrick, D.J. (2011). Extension of the
615 Bayesian alphabet for genomic selection. BMC Bioinformatics 12: 186.
616 Heslot, N., Yang, H.-P., Sorrells, M. E., and Jannink, J.-L. (2012). Genomic selection
617 in plant breeding: A comparison of models. Crop Science, 52(1), 146-160.
618 Ho, D. S. W., Schierding, W., Wake, M., Saffery, R., and O'Sullivan, J. (2019). Machine
619 learning SNP based prediction for precision medicine. Front. Genet.10, 267.
620 Li, J., Zhang, D., Yang, F., Yang, J., Wang, K., and Zhao, C. (2024). TrG2P: A transfer-
621 learning-based tool integrating multi-trait data for accurate prediction of crop yield.
622 Plant Commun. 5: 100975.
623 Liu, H. J., and Yan, J. (2019a). Crop genome-wide association study: A harvest of
624 biological relevance. Plant J. 97: 8-18.
625 Liu, J., Li, J., Wang, H., and Yan, J. (2020). Application of deep learning in genomics.
626 Science China Life Sciences, 63, 1860-1878.
627 Liu, Y., Wang, D., He, F., Wang, J., Joshi, T., and Xu, D. (2019b). Phenotype prediction
628 and genome-wide association study using deep convolutional neural network of
629 soybean. Front. Genet. 10: 1091.
630 Ma, W., Qiu, Z., Song, J., Li, J., Cheng, Q., Zhai, J., and Ma, C. (2018). A deep
631 convolutional neural network approach for predicting phenotypes from genotypes.
632 Planta 248: 1307-1318.
633 Ma, X., Wang, H., Wu, S., Han, B., Cui, D., Liu, J., Zhang, Q., Xia, X., Song, P., Tang,
634 C., Geng, L., Yang, Y., Yan, S., Zhou, K., and Han, L. (2024). DeepCCR: Large-
635 scale genomics-based deep learning method for improving rice breeding. Plant
636 Biotechnol. J. 1-3.
637 Maenhout, S., De Baets, B., Haesaert, G., and Van Bockstaele, E. (2007). Support
638 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.

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.

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.

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.