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

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

[1]State Key Laboratory of Molecular Developmental Biology, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing, China

[2]State Key Laboratory of Plant Genomics, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing, China

[3]State Key Laboratory of Plant Cell and Chromosome Engineering, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing, China

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

* Correspondence (Tel: +86-10-64806367; Email: hanrch@genetics.ac.cn (R.H.); Tel: +86-1064806362; Email: yqjiang@genetics.ac.cn (Y.J.))

Keywords: Genomic Prediction, Deep Learning, Complex Traits

Abstract

Genomic selection (GS), which integrates genomic markers with phenotypic data, has emerged as a powerful breeding strategy to predict phenotypes and breeding values in candidate populations. The effectiveness of GS largely depends on the accuracy of genomic prediction (GP) models. However, conventional GP models-primarily based on linear assumptions-often fail to capture the intricate nonlinear and epistatic interactions underlying complex traits, thus limiting their predictive power. To address these limitations, we introduce the contrastive learning and chromosome-aware network (CLCNet), a novel deep learning framework that combines contrastive learning with a chromosome-aware network architecture. CLCNet incorporates two critical components: (i) a contrastive learning module that enhances the model's ability to distinguish phenotypic variation, and (ii) a chromosome-aware module designed to account for local and global epistasis. We benchmarked CLCNet across four diverse crop species-maize (*Zea mays*), cotton (*Gossypium hirsutum*), rapeseed (*Brassica napus*), and soybean (*Glycine max*)-and compared its performance to three traditional models (rrBLUP, Bayesian Ridge, Bayesian Lasso), two machine learning methods (LightGBM, SVR), as well as two deep learning approaches (DNNGP, DeepGS). CLCNet consistently outperformed all comparator models, yielding higher Pearson correlation coefficients (PCC) and demonstrating enhanced predictive accuracy, generalizability, and scalability across high-dimensional genomic datasets. 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) represents a transformative approach in modern plant and animal breeding, enabling the prediction of genomic estimated breeding values (GEBVs) through the integration of genome-wide molecular markers and phenotypic data (Meuwissen et al., 2001). In contrast to traditional marker-assisted selection (MAS), which typically targets loci of large effect, GS captures the cumulative influence of genome-wide loci, including those with small effects, thereby facilitating more accurate and timely selection decisions (Crossa et al., 2017; Xu et al., 2020; Budhlakoti et al., 2022). Despite its growing adoption, GS continues to face key challenges-most notably, the inherent imbalance between the high dimensionality of single nucleotide polymorphism (SNP) data (ranging from thousands to millions of SNPs) and the relatively limited availability of phenotypic records (Wang et al., 2018). To mitigate this issue and improve predictive performance, a range of statistical and machine learning-based genomic prediction (GP) models have been proposed, including rrBLUP, GBLUP, BayesianA, BayesianB, BayesianC, BayesianCπ, 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). While these models have contributed to the progress of GS, their limited ability to capture complex, nonlinear genetic interactions-such as epistasis and gene-by-environment effects-restricts their ability to fully exploit genomic information (Liu et al., 2019a). Additionally, their performance often deteriorates in the presence of genomic noise or sparse phenotypic data (Habier et al., 2007 ; Tong and Nikoloski, 2021). These limitations underscore the need for more advanced, biologically-informed modeling strategies capable of uncovering complex genotype–phenotype relationships to further enhance the predictive accuracy and utility of GS in plant breeding.

Deep learning (DL) has emerged as a promising alternative, offering the capacity to learn complex genotype-phenotype relationships through hierarchical neural architectures (Liu et al., 2020; Wang et al., 2020; Gao et al., 2025). A major advantage of DL is its ability to model non-linear patterns without predefined assumptions, making it particularly well-suited for traits governed by intricate genetic architectures (González-Camacho et al., 2016; Abdollahi-Arpanahi et al., 2020; Wu et al., 2023). DL frameworks have been successfully applied across diverse modeling paradigms, including single-task, multi-task, multi-modal, and transfer learning approaches (Eraslan et al., 2019; Washburn et al. 2021).

Several DL-based models have demonstrated superior performance in GP. For instance, DeepGS, employing convolutional neural networks, outperformed RR-BLUP in wheat yield prediction (Ma et al., 2018), while DLGWAS improved trait prediction in soybean through residual and parallel convolutional architectures (Liu et al., 2019b). DNNGP addressed data sparsity and noise by incorporating principal component analysis, enhancing prediction in wheat, maize, and tomato (Wang et al., 2023). Recent innovations, such as Cropformer and SoyDNGP, have further expanded the scope of DL by integrating self-attention mechanisms, MIC-based feature selection, and genomic image transformation techniques (Gao et al., 2023; Wang et al., 2025).

Multi-modal DL models have also made significant strides by integrating multi-omics data. For instance, DEM fused genomic, transcriptomic, and epigenomic data to improve cross-species prediction via self-attention (Ren et al., 2024), while DeepCCR enhanced rice yield prediction by integrating SNP and planting location information using a bidirectional LSTM (Ma et al., 2024). Similarly, GEFormer integrated SNP and environmental data using a gating multilayer perceptron and a linear attention mechanism to model genotype–environment interactions (Yao et al., 2025). Transfer

learning models, such as TrG2P, have leveraged knowledge from non-yield traits to improve yield prediction by capturing pleiotropic effects (Li et al., 2024).

Despite these advances, multi-task learning (MTL)-a strategy that leverages shared information across traits-remains underexplored in GS. MTL holds promise for improving generalization, especially in data-limited contexts, but challenges such as weak inter-task correlation modeling, data heterogeneity, and computational demands have hindered its widespread adoption. Innovative MTL architectures and optimization strategies are therefore needed to fully realize its potential (Eraslan et al., 2019).

Herein, we present CLCNet (contrastive learning and chromosome-aware network), a novel MTL-based deep learning framework for genomic prediction. CLCNet integrates two core components: a contrastive learning module that enhances discrimination of phenotypic variation, and a chromosome-aware module that captures both local and global epistatic effects. Benchmark evaluations across four major crops demonstrated that CLCNet consistently outperformed existing classical, machine learning, and deep learning models, achieving higher prediction accuracy as indicated by Pearson correlation coefficients. These findings position CLCNet as a robust and scalable tool for accelerating GS in modern plant breeding.

Results

1. The architectural design of CLCNet

As shown in Figure 1a, CLCNet is a multi-task model designed to predict phenotypic traits from SNP data as its primary objective, while simultaneously generating low-dimensional SNP feature representations as an auxiliary task. The model architecture features a dual-pathway backbone, consisting of one path with three nonlinear layers and another with a single linear layer, integrated via a residual connection. The resulting

outputs are directed into distinct branches tailored for the primary and auxiliary tasks, respectively.

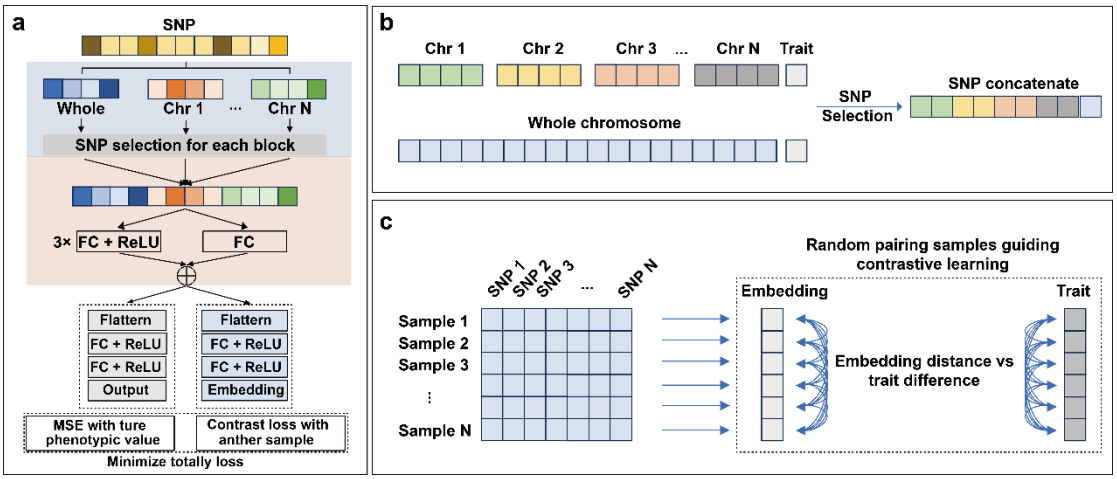

Figure 1. CLCNet Model Framework. (a) CLCNet Model Architecture. CLCNet is a multi-task model that predicts phenotypic traits from SNP data while also extracting compact feature representations. Its architecture features a dual-path backbone that combines nonlinear and linear transformations through a residual connection, allowing it to efficiently handle both the main prediction task and auxiliary objectives. (b) Chromosome-aware Module. SNPs are organized based on their respective chromosomes, and feature selection is performed using LightGBM. The module conducts both local selection-focusing on important SNPs within each chromosome and whole selection-identifying influential SNPs across the genome. Only those SNPs with positive information gain (IG > 0) are preserved. The selected SNPs from both selection strategies are concatenated to form the final input. (c) Contrastive Learning Module. In this 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. Whole, whole-chromosome SNP. Chr, chromosome-wise SNP. FC, fully connected layer. ReLU, the rectified linear unit. SNP selection, feature selection of SNP using LightGBM. Output, predicted phenotypic value. Embedding, low-dimensional SNP feature representation. Trait, true phenotypic value. Embedding distance, Euclidean distance between SNP embeddings of

paired samples. Trait difference, absolute difference in phenotypic values of paired samples.

The framework integrates two key innovations, as shown in Figures 1b and 1c. 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 at both the local (per-chromosome) and global (whole-chromosome) levels. Only the most relevant SNPs are retained and concatenated to construct the model input. Second, a contrastive learning module is introduced to enhance the training process. This module utilizes a random pairing strategy during each training epoch and optimizes the model using a composite loss function that combines phenotypic prediction loss with a custom contrastive loss. By leveraging pairwise sample relationships, the model emphasizes phenotypic differences and learns more discriminative, biologically meaningful representations.

2. Superior performance of CLCNet over conventional prediction models

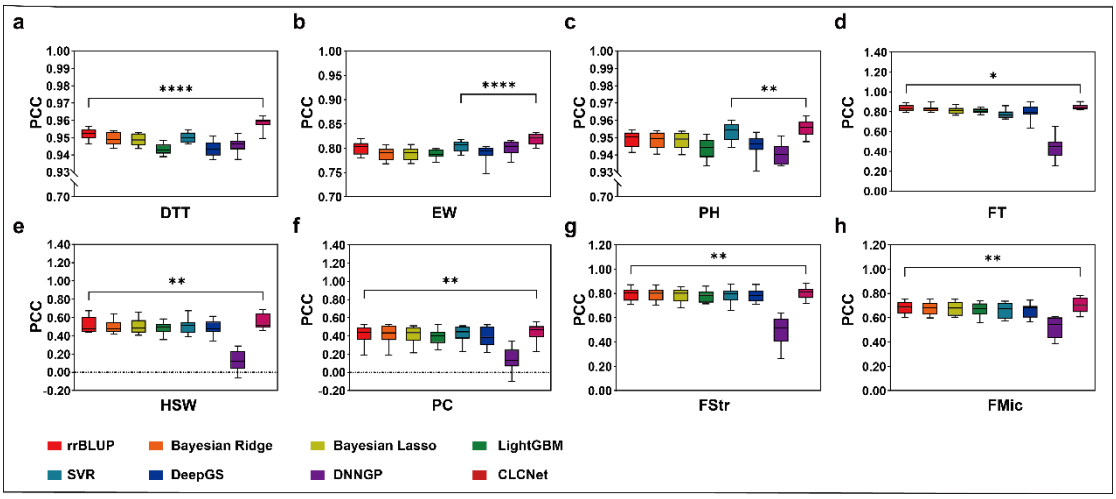

Figure 2. Prediction performance of CLCNet on eight phenotypic traits across multiple crop datasets: maize DTT (a), EW (b), and PH (c); rapeseed FT (d); soybean HSW (e) and PC (f); and cotton FStr (g) and FMic (h). Trait abbreviations correspond to: DTT - days to tasseling, EW- ear weight, PH - plant height, FT - flowering time, FStr - fiber strength, FMic - fiber micronaire, HSW - hundred seed weight, and PC - protein content. Statistically significant

differences between CLCNet and the best-performing baseline model are denoted by asterisks. Significance levels: *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, ****$p < 0.0001$.

As illustrated in Figure 2, CLCNet consistently outperformed seven benchmark models in phenotypic prediction tasks across multiple crop datasets. 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 emerged as the most competitive baseline for six traits, achieving notable predictive performance-for instance, a PCC of 95.20% for maize days to tasseling (DTT), 83.70% for rapeseed flowering time (FT), 51.74% for soybean hundred seed weight (HSW), 41.40% for soybean protein content (PC), 79.15% for cotton fiber strength (FStr), and 68.09% for cotton micronaire (FMic). SVR showed the strongest performance in predicting maize ear weight (EW, 80.01% PCC) and plant height (PH, 94.89% PCC).

Notably, while rrBLUP performing best on six traits and SVR excelling in two, CLCNet consistently statistically significant outperformed both, with two traits showing highly significant improvements (****$p < 0.0001$). Specifically, as shown in Figure 2, CLCNet improved PCC by 0.65% for DTT, 1.54% for EW, and 0.18% for PH in maize; 0.92% for FT in rapeseed; 2.09% for FMic and 1.07% for FStr in cotton; and 3.51% for HSW and 2.98% for PC in soybean datasets.

3. Demonstrating the effectiveness of two independently designed modules in CLCNet

To assess the individual contributions of the two core components of CLCNet, we performed extensive ablation studies across all phenotypic prediction tasks by evaluating two model variants: CLCNet_1 and CLCNet_2. In CLCNet_1, the local (per-chromosome) SNP feature selection was removed, retaining only the global (whole-genome) feature selection. Building upon CLCNet_1, the contrastive learning

module was further excluded to yield CLCNet_2, effectively converting the model from a multi-task to a single-task architecture. The structural distinctions among the three models are shown in Figures 3a-c.

Performance comparisons revealed that CLCNet significantly outperformed CLCNet_1 across all traits, achieving an average PCC of 76.00% versus 70.84% ($p < 0.001$), as shown in Figure 3d. Additionally, CLCNet_1 outperformed CLCNet_2 (70.84% vs. 68.50%, $p < 0.001$). The 5.16% performance decline observed between CLCNet and CLCNet_1 underscores the importance of modeling local epistatic interactions, specifically through SNPs selected at the individual chromosome level. This suggests that global feature selection alone is insufficient and may substantially limit predictive accuracy.

To further explore this, genome-wide SNPs were partitioned based on chromosomal location: SNPs selected from individual chromosomes reflect *intra-chromosomal* interactions, while those derived from the global selection process capture *inter-chromosomal* interactions (Figure 3e). Here some overlap exists between the two SNP sets, they are not identical-indicating that both capture complementary dimensions of the underlying genetic architecture, as evidenced across all four crop datasets (maize, cotton, rapeseed, and soybean; Figures 3f–i). Furthermore, the 2.34% performance reduction from CLCNet_1 to CLCNet_2 highlights the added value of the contrastive learning module in enhancing feature representation.

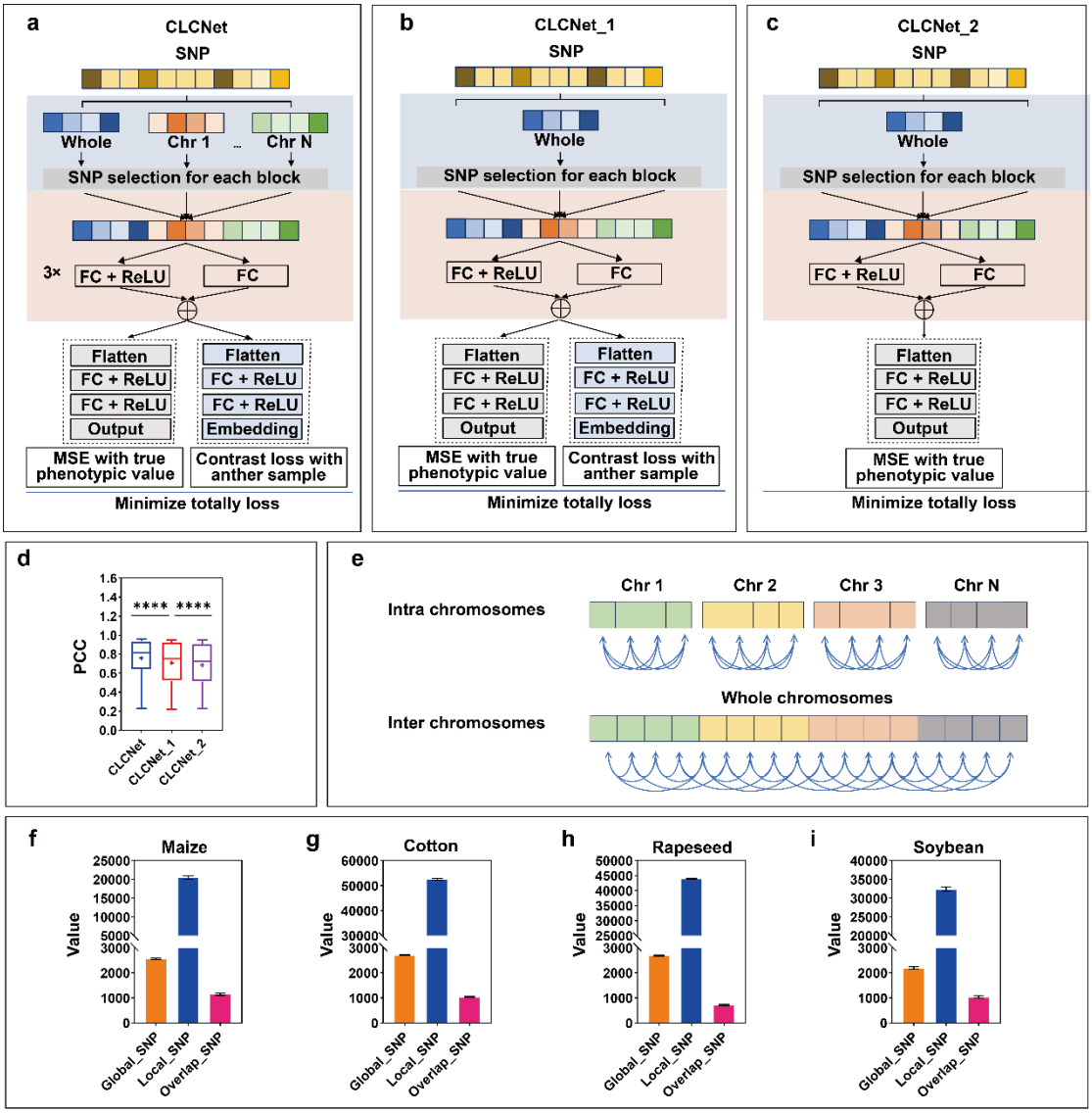

Figure 3. Analysis of CLCNet's core modules. Model architecture diagrams: (a) Full CLCNet model, (b) CLCNet_1 (without per-chromosome feature selection), and (c) CLCNet_2 (without both local feature selection and contrastive learning). (d) Predictive performance comparison of CLCNet and its ablated variants across all prediction tasks. The x-axis represents the model variants, and the y-axis shows Pearson correlation coefficient (PCC) values. Statistical significance levels are indicated as follows: $p < 0.05$ (*), $p < 0.01$ (**), $p < 0.001$ (***), and $p < 0.0001$ (****). (e) Conceptual diagram of SNP selection strategies illustrating the distinction between global and local epistatic feature selection. Global epistasis involves interactions among SNPs across the whole-chromosome, while local epistasis focuses on interactions within individual chromosomes. (f-i) Distribution of selected SNPs across four crop species: (f) maize, (g) cotton, (h) rapeseed, and (i) soybean. Global_SNP refers to SNPs selected via whole-

chromosome feature selection; Local_SNP represents those identified through chromosome-level selection; Overlap_SNP indicates SNPs shared by both strategies. The y-axis indicates the number of selected SNPs.

Together, these results collectively confirm that both components are integral to CLCNet's superior performance, with the chromosome-aware module exerting a more pronounced impact. These findings provide strong empirical validation for the architectural design of the CLCNet framework.

4. CLCNet outperforms other models in detecting chromosome-level disturbances

We simulated perturbations to SNP positions, which randomized epistatic interactions and linkage disequilibrium (LD) at the single-chromosome level by disrupting the original associations among SNPs. This allowed to evaluate whether different GP models can capture chromosome-level structural changes. Consequently, such positional perturbations may alter the outcome of chromosome-aware module, leading to distinct sets of informative SNPs being identified before and after perturbation (Figure 4a). When SNP positions are perturbed, the newly selected SNPs may exhibit reduced explanatory power, thereby compromising predictive performance. This highlights the importance of a model’s ability to detect and adapt to chromosome-level SNP positional shifts to maintain robustness.

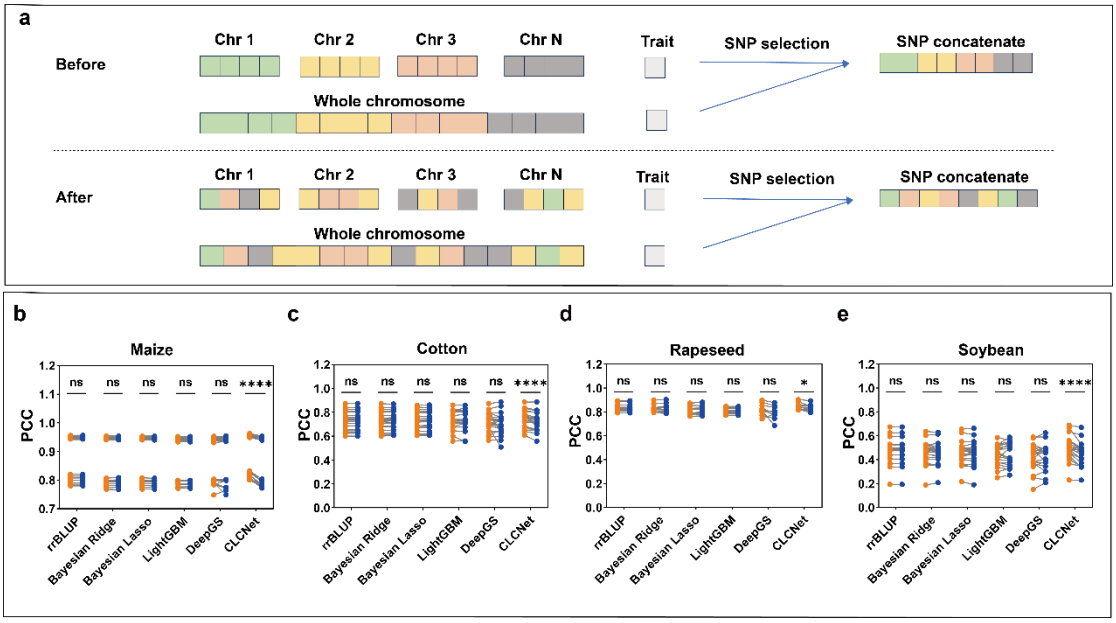

Figure 4. Model performance under SNP position perturbation. (a) The top panel shows the original feature selection approach (chromosome-aware module) prior to SNP position perturbation. SNPs are grouped by their chromosome locations into individual-chromosome and whole-chromosome sets. LightGBM is applied independently to each group for feature selection, retaining SNPs with an information gain (IG) greater than 0. The selected SNPs are then concatenated for downstream prediction (denoted as "before"). The bottom panel depicts the strategy after SNP position perturbation, where SNP positions are randomly shuffled. For the individual chromosome grouping, SNPs are re-sampled using a sliding window that corresponds to the original chromosome size. The chromosome-aware process is then applied, and the selected SNPs are concatenated as "after." (b-e) PCC values of different models before and after SNP position perturbation for phenotypic prediction tasks in maize (b), cotton (c), rapeseed (d), and soybean (e). The x-axis represents model groups: rrBLUP, Bayesian Ridge, Bayesian Lasso, LightGBM, DeepGS, and CLCNet. Orange dots represent PCCs from 10-fold cross-validation before shuffling, and blue dots after shuffling. The y-axis indicates PCC values. Significance levels are indicated as: $p < 0.05$ (*), $p < 0.01$ (**), $p < 0.001$ (***), $p < 0.0001$ (****).

CLCNet demonstrates a notable sensitivity to chromosome-level changes in SNP positional structure. Across all four crop datasets-maize, cotton, rapeseed, and soybean-

traditional approaches such as rrBLUP, Bayesian Ridge, Bayesian Lasso, LightGBM, and DeepGS showed no significant differences in PCC values before and after random SNP position perturbation in whole-chromosome. In contrast, CLCNet exhibited a significant decline in predictive accuracy, reflecting its unique ability to model and respond to chromosome-sensitive genomic features. This is attributed to the design of CLCNet's chromosome-aware module, which performs feature selection within individual chromosomes. Specifically, CLCNet's PCC dropped from 91.11% to 89.39% in maize, 75.20% to 72.92% in cotton, 84.63% to 83.46% in rapeseed, and 49.82% to 45.04% in soybean (Figures 4b–e). Notably, although LightGBM serves as the feature selection algorithm within CLCNet's chromosome-aware module, it alone did not exhibit significant PCC changes under whole-chromosome perturbation when evaluated independently. This indicates that SNP positional changes at the whole chromosome level do not severely impair LightGBM's performance. The performance drop observed in CLCNet is primarily due to the disruption of SNP–SNP interactions within individual chromosomes, which are explicitly modeled by the chromosome-aware design.

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"-dimensionality reduction is critical for enhancing both computational efficiency and predictive accuracy (Pudjihartono et al., 2022). Feature selection, in particular, mitigated noise and redundancy by identifying the most informative and representative SNPs (Ho et al., 2019; Urbanowicz et al., 2019). While traditional GWAS-based feature selection approaches are useful, they typically rely on linear models that evaluate

associations between individual SNPs and phenotypes, thereby overlooking potential SNP–SNP interactions (Liu et al., 2019a). Other GP feature selection methods, such as PCA (Wang et al., 2023), MIC (Wang et al., 2025), and PCS (Zhao et al., 2025), often fail to consider local epistasis within individual chromosomes. In contrast, LightGBM—a tree-based ensemble learning method—can inherently capture non-linear relationships and epistatic interactions during feature selection. This enables more accurate and sensitive detection of phenotype-associated SNPs (Yan et al., 2021). We implemented LightGBM with an explicit chromosome-based grouping strategy: SNPs from individual chromosomes were used to model local epistatic interactions, while whole-chromosome SNPs were incorporated to capture global epistasis. The resulting local and global SNP subsets were then merged into a unified set of representative features, with each type complementing the other. Similar to other methods, the chromosome-aware module can also be readily integrated into different GP models, enhancing its versatility and practical utility. This adaptability makes it as a powerful tool for data preprocessing across a broad spectrum of genomic prediction tasks.

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 enhance performance across related tasks by leveraging shared information, highlighting its potential as a promising direction in DL (Zhang et al., 2018). In this study, we incorporate contrastive learning into the multi-task framework of CLCNet, enhancing 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, the application of a multi-task learning framework based on contrastive learning in GP models is a first of its kind approach. Future research will focus on supplementing the random pairing method with exhaustive pairing to further improve the performance of CLCNet. In this work, CLCNet randomly pairs samples for contrastive loss calculation during each training iteration. While effective, random pairing typically results in fewer unique pairings (e.g., with 100 samples, fewer than 100 unique pairings out of a theoretical 4950 combinations). Higher performance of CLCNet can be expected through incorporating a broader variety of sample combinations via exhaustive pairing.

In conclusion, we have developed CLCNet, a novel deep learning model for plant GP. By integrating chromosome-aware and contrastive learning modules, CLCNet outperforms traditional linear models, machine learning models, and other DL approaches in terms of prediction accuracy. The model effectively captures both local and global epistasis, and optimizes molecular marker selection, thereby reducing data noise and enhancing overall performance. Additionally, CLCNet's multi-task modeling strategy, which includes predicting phenotypic differences as an auxiliary task, promotes improved inter-task information sharing, leading to increased prediction accuracy. Overall, CLCNet is a reliable, efficient, and powerful tool for GS, excelling with high-dimensional datasets and offering substantial potential for plant breeding applications.

Materials and methods

Data description and preprocessing

This study utilized the CropGS-Hub database, which consolidates genotypic and phenotypic data from 14 populations of 7 major crops, comprising over 30,000 individuals and 224 billion genotypic records (Chen et al., 2024). From these, 4 datasets were selected. 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 8 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 $n \times m$ 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. $\|\text{embedding}_1^i - \text{embedding}_2^i\|$ 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 partitions. CLCNet hyperparameters—learning rate, batch size, and epoch—were tuned using the Optuna framework (Akiba et al., 2019), with search spaces of [0.01, 0.001, 0.0001], [16, 24, 32, 64], and [50, 80, 100, 200], respectively. The optimized settings were applied to the benchmark dataset, as detailed in Table S2.

SNP position perturbation

The input is an n ×m matrix, where n represents the number of samples and m denotes the number of markers, as previously described and defined. The order of m reflected the genomic positions of the markers. To perform SNP position perturbation, the order of m was randomly shuffled across all n samples, disrupting the original SNP positions. After shuffling, the input matrix shape remains unchanged. Subsequently, rrBLUP, Bayesian Ridge, Bayesian Lasso, LightGBM, and DeepGS were used to calculate the PCC before and after the SNP position shuffle. 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, Assoc. Prof. Shusong Zheng, and Dr. Jiachao Xu, from the Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, for their professional and insightful discussions. We would also like to express our gratitude to Dr. Shihao Shao from Peking University

Health Science Center for his assistance with the deep learning model. In addition, we are also grateful to Prof. Jianxiao Liu and Prof. Xiaolei Liu from Huazhong Agricultural University for their guidance.

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

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.

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.

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.

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.

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.

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.

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.