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, Contrastive Learning, Complex Traits

Abstract

Genomic selection (GS) is a sophisticated breeding technology that utilizes genomic markers in conjunction with phenotypic data to predict breeding values and phenotypes of candidate populations. The foundation of GS is the construction of precise genomic prediction (GP) models. Traditional GP models, which are typically linear in nature, encounter difficulties in capturing non-linear relationships within genetic data, thereby constraining their capacity to adequately characterize complex genetic architectures. To address this limitation, we have developed a novel deep learning framework, namely contrastive learning and chromosome-aware network (CLCNet), which has been specifically tailored for plant GP. The architecture of CLCNet comprises two key modules: a contrastive learning module designed to capture phenotypic variations, and a chromosome-aware module that accounts for the effects of linkage, as well as local and global epistasis. The performance of CLCNet was evaluated across eight datasets representing seven plant species: maize (*Zea mays*), rice (*Oryza sativa*), cotton (*Gossypium hirsutum*), millet (*Setaria italica*), chickpea (*Cicer arietinum*), rapeseed (*Brassica napus*), and soybean (*Glycine max*). In comparative assessments against three classical models (rrBLUP, Bayesian Ridge, Bayesian Lasso), two machine learning approaches (LightGBM, SVR), and two other deep learning models (DNNGP, DeepGS), CLCNet exhibited superior performance, achieving higher Pearson correlation coefficients (PCC) and lower root mean squared errors (RMSE). Due to its superior prediction accuracy, generalization capability and scalability in high-dimensional datasets, CLCNet is expected to be a powerful tool for future GS applications in the field of plant breeding.

Introduction

As an advanced plant and animal breeding technology, genomic selection (GS) utilizes genome-wide markers and phenotypic data to predict genomic estimated breeding values (GEBVs) (Meuwissen et al., 2001). In contrast to the marker-assisted selection (MAS) method, GS captures the collective effects of numerous loci, including those with minor contributions, enabling earlier and more efficient selection (Crossa et al., 2017; Xu et al., 2020). A prevalent issue in GS is the discrepancy between the large number of markers and the relatively small number of phenotypic observations. To address this, various models and algorithms, such as GBLUP, Bayesian approaches, and machine learning techniques, have been developed to enhance prediction accuracy (Habier et al., 2007; Maenhout et al., 2007; Crossa et al., 2010; Heslot et al., 2012; Wang et al., 2018). Nevertheless, the efficacy of these methodologies is often impeded by their inability to adequately capture nonlinear interactions and manage noise in genomic data.

Deep learning (DL) has emerged as a powerful alternative, capable of modelling complex, high-dimensional datasets through multi-layer neural networks for genomics (Liu et al., 2020,). The efficacy of DL in capturing non-linear genotype-phenotype relationships without the necessity for predefined assumptions renders it a particularly advantageous approach for complex traits (Abdollahi-Arpanahi et al., 2020; Wang et al., 2020). Furthermore, DL supports a range of modelling paradigms, encompassing single-task, multi-task, multi-modal, and transfer learning (Eraslan et al., 2019). These paradigms enhance performance by focusing on specific phenotypes, leveraging shared trait information, integrating multi-omics data, and utilizing pre-trained models for data-scarce scenarios.

DL-based models have demonstrated notable success in genomic prediction (GP). For instance, DeepGS, which employed convolutional layers, outperformed RR-BLUP in wheat predictions (Ma et al., 2018). Furthermore, DLGWAS has been shown to enhance soybean phenotypic predictions through the utilization of residual and parallel convolutional layers (Liu et al., 2019). Addressing the issues of denoising and dimensionality, DNNGP combined PCA with DL for robust predictions across wheat, maize, and tomato (Wang et al., 2023). SoyDNGP further innovated by transforming genomic matrices into RGB-like channels for soybean GS (Gao et al., 2023).

Multi-modal DL models have also advanced GS by integrating multi-omics data. For instance, DEM combined genomic, transcriptomic, and epigenomic data using self-attention mechanisms, thereby improving predictions for various species (Ren et al., 2024). DeepCCR enhanced rice by incorporating SNP and environmental data with a Bi-LSTM module (Ma et al., 2024). Transfer learning, as demonstrated by TrG2P, employed pre-trained models on non-yield traits to improve yield predictions, leveraging pleiotropic relationships (Li et al., 2024).

Despite these above advances, multi-task learning in GS remains underexplored. A multi-task learning based GS strategy that leverages shared information across traits has the potential to improve generalization and robustness, particularly in data-limited scenarios (Eraslan et al., 2019). However, challenges such as inadequate modelling of inter-task correlations, variability in phenotype data quality leading to negative transfer, and doubled complexity in model training limit its application. The advancement of GS through the implementation of innovative multi-task architectures and optimization strategies represents a promising avenue for future research endeavors. Consequently, there is an urgent necessity to develop innovative multitasking architectures and optimization strategies to propel further advancements in this domain.

Herein, we report a novel deep learning framework, named as contrastive learning and chromosome-aware network (CLCNet), which has been specifically tailored for plant GP. The architecture of CLCNet comprises two fundamental modules: a contrastive learning module for capturing phenotypic variations, and a chromosome-aware module that accounts for linkage, as well as both local and global epistasis. The efficacy of CLCNet was evaluated across eight datasets from seven plant species (maize (*Zea mays*), rice (*Oryza sativa*), cotton (*Gossypium hirsutum*), millet (*Setaria italica*), chickpea (*Cicer arietinum*), rapeseed (*Brassica napus*), and soybean (*Glycine max*)). The results indicated that CLCNet outperformed three classical models (rrBLUP, Bayesian Ridge, Bayesian Lasso), two machine learning approaches (LightGBM, SVR), and two deep learning models (DNNGP, DeepGS), achieving higher Pearson correlation coefficients (PCC) and lower root mean squared errors (RMSE). These results underscore the superior accuracy, generalization, and scalability of CLCNet, indicating it as a powerful tool for future GS applications in plant breeding.

Results

1. The architectural design of CLCNet

As shown in Figure 1a, CLCNet is a multi-task model that uses single-nucleotide polymorphism (SNP) data to predict phenotypic values as the primary task, whilst concurrently generating low-dimensional feature representations as an auxiliary task. The model incorporates a dual-pathway backbone, comprising one pathway with three non-linear layers and another with a linear layer, combined through a residual structure. Outputs are processed through separate branches for the primary and auxiliary tasks.

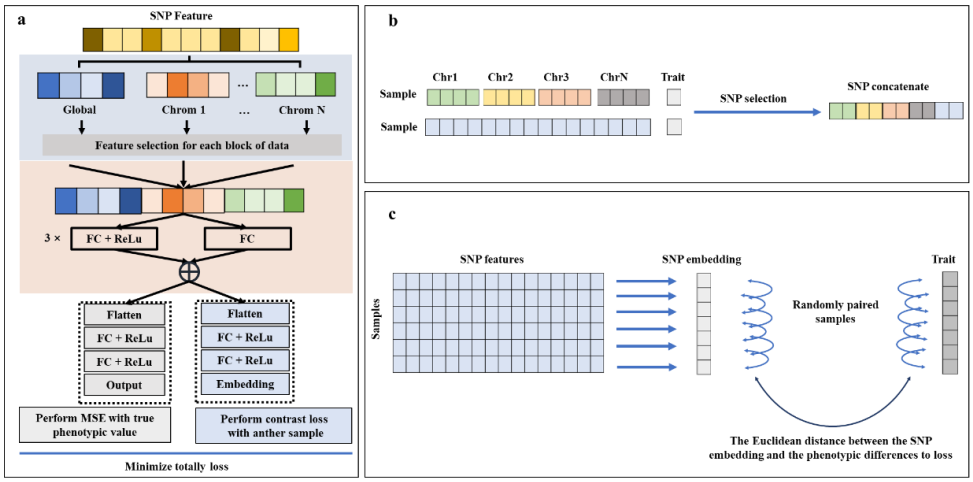

Figure 1. CLCNet model framework. (a) CLCNet model architecture. CLCNet is a multi-task model designed to predict phenotypic values from SNP data while generating low-dimensional feature representations. It incorporates a dual-pathway backbone that combines non-linear and linear layers through a residual structure, enabling simultaneous support for primary and auxiliary tasks. (b) Chromosome-aware module. SNPs are grouped by chromosome, and LightGBM is employed for feature selection. Local selection identifies key SNPs within individual chromosomes, while global selection captures key SNPs across the entire genome. Only SNPs with positive information gain (IG > 0) are retained. The selected local and global SNPs are concatenated to form the final input features. (c) Contrastive learning module. This module utilizes a random pairing strategy per epoch, processing two samples simultaneously to predict phenotypic values and generate feature vectors. Optimization is achieved through three loss functions: mean squared error (MSE) for phenotypic predictions and a contrastive loss that links feature vector distances to phenotypic differences.

As demonstrated in Figure 1b, the framework incorporates two innovations. Firstly, it includes a chromosome-aware module for the preprocessing of SNPs. Secondly, it incorporates a contrastive learning module for the training process. SNPs are organized by chromosome, and key loci are selected both locally and globally using a feature selection algorithm that efficiently identifies important loci. These selected features are then concatenated to form the model input. The contrastive learning module incorporates a random pairing strategy during training to optimize the model. This module minimizes a combined loss function that includes both phenotypic prediction loss and contrastive loss, leveraging pairwise relationships between samples (Figure 1c).

2. Performance of CLCNet and comparison with other methods

CLCNet demonstrated notable advantages in the domain of phenotypic prediction tasks. Out of 27 tasks, CLCNet significantly outperformed all other methods in 21 instances. This superiority was characterized by an enhanced Pearson's correlation coefficient (PCC) in 21 tasks, as demonstrated in Figure 2a, and a reduction in root mean square error (RMSE) in 20 tasks, as shown in Figure 2b. Notably, for one phenotype (DTT in GSTP001), CLCNet showed a significantly higher PCC, while the RMSE difference was not significant. This indicated that CLCNet's performance for DTT remains robust despite dataset limitations, as its PCC consistently exceeded 95% across other methods (Supplementary Figures 1).

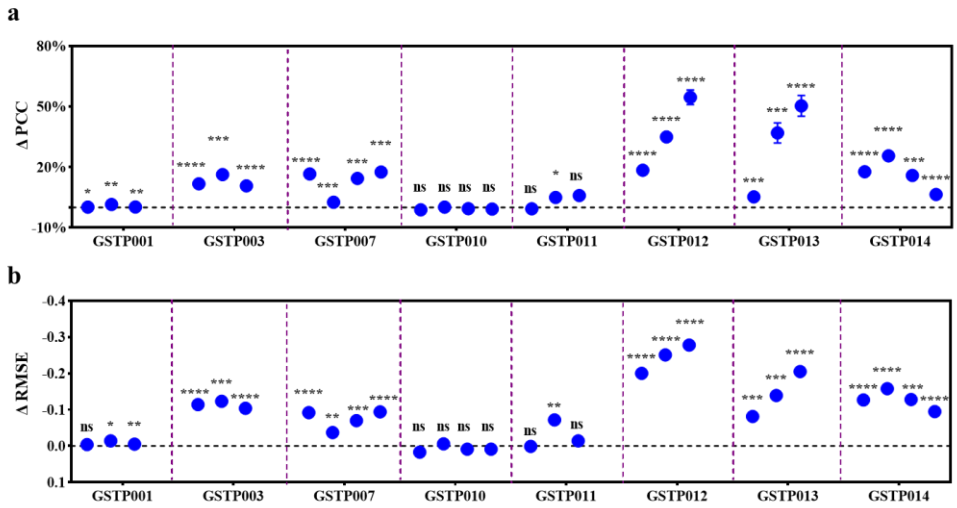

Figure 2. CLCNet overall prediction performance. (a) Prediction accuracy performance of CLCNet on eight datasets. (b) Prediction error performance of CLCNet on eight datasets. ΔPCC and ΔRMSE measure the improvement of CLCNet over the best baseline model specific to each task. Significant differences between CLCNet and the best baseline model are indicated by asterisks, and non-significant differences are indicated by ns. Significant level, *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$, ****: $p < 0.0001$.

However, for six phenotypes, CLCNet did not show performance improvements, with both PCC and RMSE being statistically comparable to the best baseline method. These phenotypes included fiber elongation (Felo), fiber length (Flen), micronaire value (FMic), fiber strength (FStr), hundred seed weight (HSW), and grain weight per

plant (PGW), which were distributed across the GSTP010 (cotton) and GSTP011 (millet) datasets (Figure 2a; Figure 2b; Supplementary Figures 4-5). Among the 21 tasks where CLCNet achieved significant PCC improvements, it outperformed rrBLUP 16 times, SVR 4 times, and Bayesian Lasso once (Supplementary Figures 1-8). These results underscore the superiority of CLCNet in predictive accuracy and robustness across diverse phenotypic traits.

Table 1. Benchmark dataset

ID	Species	Sample type	chromosomes	Sample number	Raw SNP	SNP filter	SNP number	Traits	Δ PCC $_{mean}$	Δ RMSE $_{mean}$
GSTP001	Maize	Hybrid Line	2n = 2x = 20	8,652	4,549,828	LD	125,820	EW, DTT, PH	0.599%	0.008
GSTP003	Maize	Inbred Line	2n = 2x = 20	1,458	4,549,828	LD	132,584	EW, DTT, PH	12.897%	-0.114
GSTP007	Rice	Hybrid Line	2n = 2x = 24	1,495	1,651,507	LD	87,363	YPP_HZ, YPP_SY, GW, PN	12.779%	-0.073
GSTP010	Cotton	Inbred Line	2n = 4x = 52	1,245	1,122,352	LD; MAF	14,201	FElo, FLen, FMic, FStr	-0.600%	0.007
GSTP011	Millet	Inbred Line	2n = 2x = 18	827	161,562	LD	20,860	MSW, HSW, PGW	3.387%	-0.028
GSTP012	Chickpea	Inbred Line	2n = 2x = 16	2,921	2,418,866	LD; MAF	289,653	HSW, YPP, SPP	36.015%	-0.243
GSTP013	Rapeseed	Inbred Line	2n = 4x = 38	991	4,286,896	LD	666,221	FT, OC, PC	31.866%	-0.385
GSTP014	Soybean	Inbred Line	2n = 2x = 40	2,795	29,599,017	LD	173,511	SL, VBN, HSW, PC	16.375%	-0.127

The LD and MAF commands are *--indep-pairphase 100 10 0.2* and *--maf 0.05*, respectively. EW, DTT, PH, YPP_HZ, YPP_SY, GW, PN, Felo, Flen, Fmic, FStr, MSW, HSW, PGW, YPP, SPP, FT, OC, PC, SL, and VBN denote ear weight, days to tasseling, plant height, yield per plant in Hangzhou and Sanya, grain weight, panicle number, fiber elongation rate, fiber length, fiber micronaire, fiber strength, main stem width, hundred seed weight, per plant grain weight, yield per plant, seeds per pod, flowering time, oil content, protein content, seed length, and valid branch number, respectively. ΔPCC and ΔRMSE indicate the improvement of CLCNet over the task-specific best baseline model, while ΔPCC mean and ΔRMSE mean represent the average ΔPCC and ΔRMSE across all tasks in one dataset.

CLCNet exhibited outstanding performance in PCC improvements across multiple datasets (Table 1, Supplementary Table 1). In the GSTP012 (chickpea) dataset, the ΔPCC $_{mean}$ increase reached 36.015%, with the largest improvement observed in yield per plant (YPP) at 54.603% (Supplementary Table 1). In GSTP013 (rapeseed), the ΔPCC $_{mean}$ increase was 31.866%, with protein content (PC) showing the greatest improvement at 50.389% (Supplementary Table 1). Similarly, in GSTP014 (soybean), the ΔPCC $_{mean}$ improvement was 16.375%, with protein content (PC) again exhibiting

the largest increase at 25.651% (Supplementary Table 1). In contrast, ΔPCC_{mean} in GSTP001 (hybrid maize) was relatively small, at only 0.599%. This modest improvement may be attributed to the performance plateau of the dataset, where the prediction accuracy for traits such as days to tassel (DTT), ear weight (EW), and plant height (PH) had already exceeded 95%, 80%, and 94%, respectively, with the other seven methods (Supplementary Figures1).

Regarding $\Delta RMSE_{mean}$, CLCNet performed best in GSTP013 (rapeseed), GSTP012 (chickpea), and GSTP014 (soybean), with reductions of -0.385, -0.243, and -0.127, respectively. Conversely, the smallest ΔRMSE reduction was observed in GSTP001 (hybrid maize), at just -0.008.

Collectively, CLCNet exhibited significant predictive advantages across most phenotypic prediction tasks, underscoring its superior potential value and applicability.

3. SNP analysis of the chromosome-aware module

An analysis was conducted of the characteristics of SNPs in the chromosome-aware module across eight datasets. The following proportions were calculated: (a) the proportion of SNPs specific to the global chromosome, (b) the proportion of SNPs specific to the local chromosome, (c) the proportion of SNPs that overlap, and (d) the proportion of SNPs in quantity. These metrics were defined based on global SNPs and local SNPs. The proportion of global-specific SNPs represents the percentage of global SNPs not included in the local SNP set, relative to the total global SNPs. The proportion of local-specific SNPs is defined as the percentage of local SNPs not included in the global SNP set, relative to the total local SNPs. The proportion of overlapping SNPs represents the percentage of overlapping SNPs in the global and local SNP sets relative to their union minus the overlap. Finally, the proportion of SNP quantity represents the

incremental count of local SNPs compared to global SNPs, relative to their union minus the overlap.

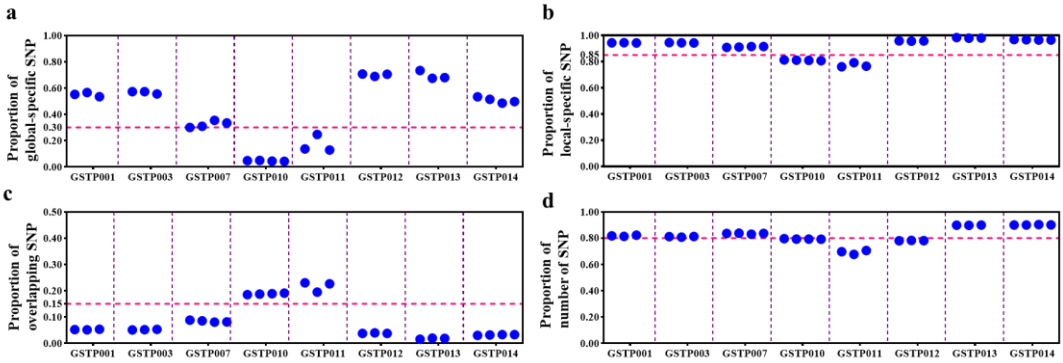

Figure 3. Analysis of SNP selected by the chromosome-aware module for each of the eight datasets. (a) Proportion of global-specific SNP for each of the eight datasets. It calculated as the percentage of global SNPs not included in the local SNP set relative to the total global SNPs. (b) Proportion of local-specific SNP for each of the eight datasets. It the percentage of local SNPs not included in the global SNP set relative to the total local SNPs. (c) Proportion of overlapping SNP for each of the eight datasets. It calculated as the percentage of shared SNPs between global and local SNP sets relative to their union minus the overlap. (d) Proportion of number of SNP for each of the eight datasets. It representing the incremental count of local SNPs compared to global SNPs relative to their union minus the overlap.

The results (Figures 3a, 3b, 3c and 3d) demonstrated that GSTP010 (cotton) and GSTP011 (millet) exhibited significant outlier characteristics in all four proportions. Specifically, they had the lowest proportion of global-specific SNPs, proportion of local-specific SNPs, and proportion of SNP quantity, but the highest proportion of overlapping SNPs. Despite differences in SNP counts, sample sizes, and chromosome numbers across datasets, the outlier features of GSTP010 and GSTP011 aligned with their SNP counts, which were the lowest among all datasets (Supplementary Figures 9a-c).

Notably, ΔPCC and ΔRMSE for these two datasets showed no significant differences (Figure 2, Table 1, Supplementary Figures 4 and 5). These findings suggest that the quantity of SNPs may have a substantial impact on the predictive performance of CLCNet.

4. Impact of SNP number on CLCNet's predictive performance

The LD pruning and MAF filtering steps were removed from the preprocessing of the GSTP010 (cotton) and GSTP011 (millet) datasets. Consequently, the number of SNPs for GSTP010 increased substantially from 14,201 to 1,122,352, and for GSTP011 from 20,860 to 161,562. Subsequent SNP analysis through the chromosome-aware module revealed increases in the proportions of global-specific SNPs, local-specific SNPs, and overlapping SNPs, while the proportion of SNP quantity decreased (Supplementary Figures 9d-g).

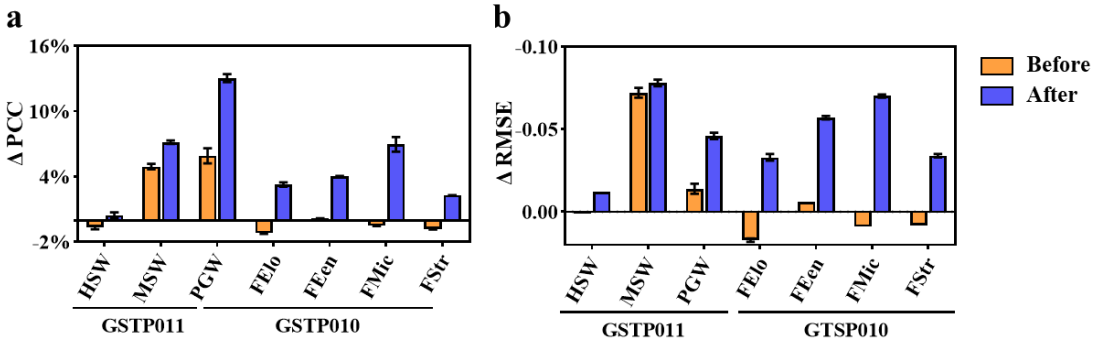

Figure 4. Analysis of GSTP10 and GSTP011 before and after increasing the number of SNPs. (a) ΔPCC of GSTP10 and GSTP011 before and after increasing the number of SNPs. (b) ΔRMSE of GSTP10 and GSTP011 before and after increasing the number of SNPs. ΔPCC and ΔRMSE reflect the improvement of CLCNet over the best-performing baseline model for each specific task.

Subsequently, a comparison was made of the performance of CLCNet and the other seven methods on the GSTP010 (cotton) and GSTP011 (millet) datasets after increasing the SNP count. For GSTP010, among its four traits, three traits that initially showed no significant difference between CLCNet and the best-performing method demonstrated significant differences after the SNP count was increased, while fiber length (FElo) remained non-significant (Supplementary Figures 4 and 10). In GSTP011, the PCC for hundred seed weight (HSW) remained non-significant in comparison to other methods; however, all RMSE improved from non-significant to significant. For plant grain

weight (PGW), both PCC and RMSE shifted from non-significant to significant differences. Additionally, for main stem width (MSW), the difference improved from significant to highly significant (Supplementary Figures 5 and 11).

Further analysis revealed that both ΔPCC and ΔRMSE were significantly better after the SNP count increased (Figure 4a; Figure 4b), indicating that the performance improvement of CLCNet surpassed that of other methods. These results demonstrate that increasing SNP count can substantially enhance the predictive performance of CLCNet.

5. Analysis of the Importance of CLCNet's Core Module

To ascertain the impact of the two core modules of CLCNet on the model's performance, a series of modified models were constructed in which the two core modules of CLCNet were removed. The performance of these models was then compared with that of the complete CLCNet model. This comparison was made across 27 phenotypic prediction tasks in terms of PCC and RMSE. Upon the removal of the contrastive learning module, the model transitioned from multi-task learning to single-task learning, and this transition is referred to as CLCNet_1. The third model, named as CLCNet_2, was characterized by the omission of single-chromosome feature selection, whilst retaining feature selection for the entire chromosome group. The architecture of the three models is detailed in Supplementary Figure 12.

The results indicate that, with respect to the ΔPCC, CLCNet demonstrated the optimal performance, with CLCNet_1 and CLCNet_2 exhibiting the lowest predictive accuracy. The ΔPCC for CLCNet_2 decreased to a negative value, indicating that its performance inferior to that of traditional methods (Figure 5a). A parallel tendency was observed in terms of the ΔRMSE (Figure 5b). Specifically, CLCNet demonstrated the lowest error, followed by CLCNet_1, while CLCNet_2 exhibited the highest error,

exceeding the baseline, indicating that its error was greater than that of traditional methods (Figure 5b).

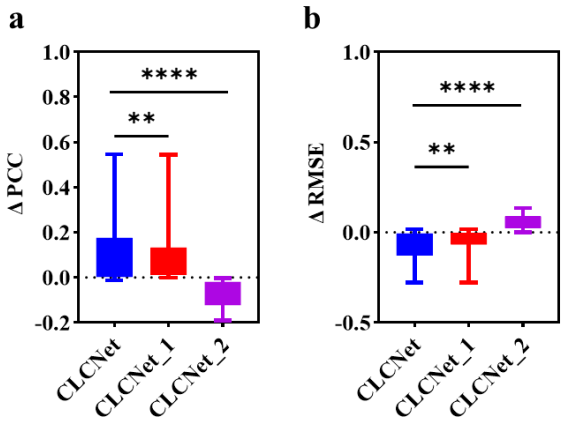

Figure 5. Comparison of CLCNet and its variants across 27 phenotype prediction tasks from eight datasets. (a) ΔPCC comparison between CLCNet and its variants. (b) ΔRMSE comparison between CLCNet and its variants. ΔPCC and ΔRMSE represent the performance improvement of CLCNet relative to the best-performing baseline model for each individual task.

The comprehensive analysis of this study indicates that the chromosome-aware module is crucial for improving prediction accuracy and reducing error, and its removal results in a significant decline in model performance. The contrastive learning module validates the effectiveness of multi-task modelling, optimizing errors and enhancing accuracy through weight sharing and loss-driven mechanisms. It is only when both modules are used in combination that the model performs optimally. This emphasizes the key role of the chromosome-aware module and the contrastive learning module and their synergistic effect in complex phenotypic prediction tasks.

6. Comparison of algorithm performance under SNP position perturbation

As shown in Figure 1b, the chromosome-aware module employs a two-tiered approach for the selection of representative SNPs, categorizing them into local and global groups. Local SNPs are identified through local selection, while global SNPs are obtained

through genome-wide selection. This methodology accounts for both local and global epistasis, *i.e.* non-linear SNP relationships within chromosome groups and across the genome. It is important to note that local and global selections are distinct processes; local-specific SNPs are not identifiable through global selection, and vice versa. The efficacy of the chromosome-aware module is evidenced by its integration of both local and global selections, a feature absented in CLCNet_2, which relies solely on global epistasis selection. This finding serves to confirm the limitations of global-only SNP selection (Figures 5a and 5b).

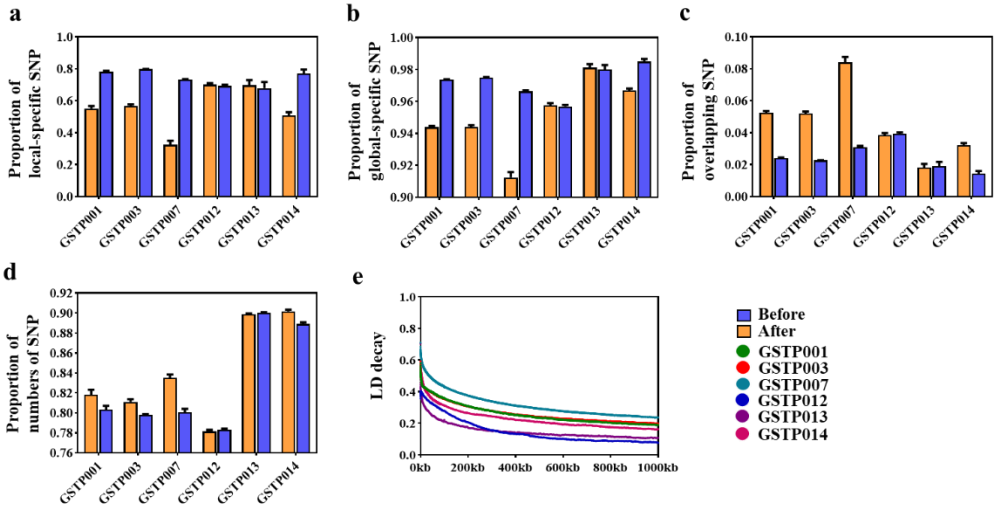

Figure 6. Analysis of SNP selected by the chromosome-aware module for each of the eight datasets before and after position disruption and linkage disequilibrium decay analysis for each of the eight datasets. (a) Proportion of global-specific SNP during positional shuffling. (b) Proportion of local-specific SNP during positional shuffling. (c) Proportion of overlapping SNP during position shuffling. (d) Proportion of numbers of SNP during position shuffling. (e) Linkage disequilibrium decay analysis.

These proportions were previously defined and calculated based on the global and local SNPs for each dataset, to compared before and after the disruption of SNP positions.

It is hypothesized that the chromosome-aware module is responsible for the linkage effects observed within chromosomes, thereby enhancing the model's performance. To verify this hypothesis, an experiment was conducted in which the positions of SNP were

randomized across the genome. The performance of the model was then compared before and after the randomization. This disruption of linkage effects introduced more redundant SNPs, leading to a subsequent decrease in prediction accuracy. The analysis revealed that in most datasets (GSTP001 for hybrid maize, GSTP003 for inbred maize, GSTP007 for rice, and GSTP014 for soybean), the shuffle increased the proportion of both global-specific and local-specific SNPs. Global-specific SNPs increased by an average of 20%, while local-specific SNPs increased by only 2%. Furthermore, overlapping SNPs decreased, thereby supporting our hypothesis that linkage dispersion causes the model to favour global-specific SNPs (Figure 6a-d). It is hypothesised that GSTP012 (chickpea) and GSTP013 (rapeseed) exhibit smaller linkage effects, a hypothesis that is confirmed by linkage decay distance calculations showing lower linkage effects in these datasets (Figure 6e).

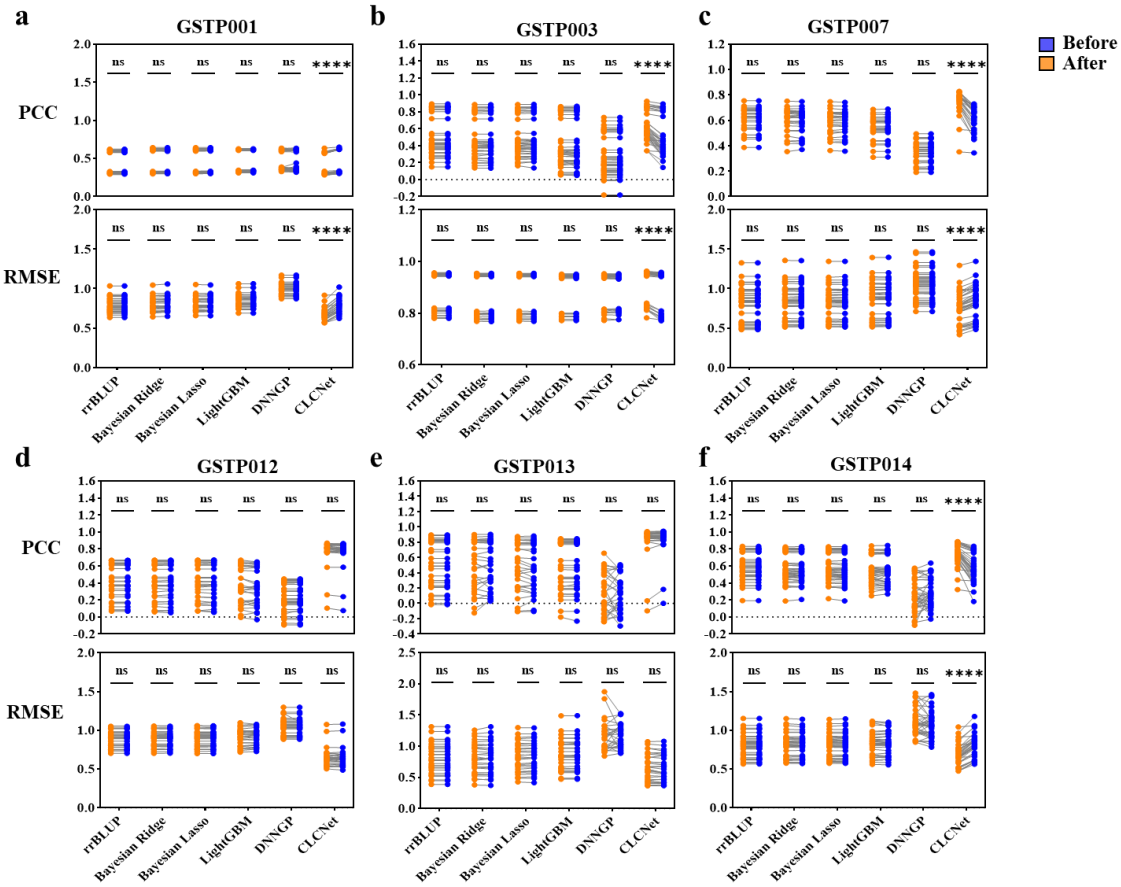

Figure 7. Comparison of PCC and RMSE of six models on six datasets before and after

position shuffling.

(a-f) Performance changes in PCC and RMSE on each dataset (GSTP001, GSTP003, GSTP007, GSTP012, GSTP013, and GSTP014) before and after SNP position shuffling.

For datasets from hybrid maize (GSTP001), inbred maize (GSTP003), rice (GSTP007), and soybean (GSTP014), following the shuffling of the SNP positions, the PCC of CLCNet significantly decreased and the RMSE significantly increased. In contrast, methods such as rrBLUP, Bayesian Ridge, Bayesian Lasso, LightGBM, and DNNGP demonstrated no substantial alterations in performance, suggesting that they are unable to account for linkage effects. Furthermore, for the chickpea (GSTP012) and rapeseed (GSTP013) dataset, CLCNet's PCC and RMSE did not exhibit significant changes before and after the shuffle, likely attributable to the reduced linkage effects observed in these species (Figure 6e). Collectively, these findings validate the hypothesis that the process of shuffling the positions of SNPs leads to the dispersal of linkage effects, thereby introducing redundant SNPs that result in a substantial decline in model performance.

Discussion

CLCNet employs a multi-task learning framework, a distinction that sets it apart from single-task, multi-modal, and transfer learning approaches. In comparison to single-task models, multi-task models leverage auxiliary tasks' losses to guide the learning process, often leading to superior performance (Eraslan et al., 2019). In this study, we have integrated contrastive learning within the multi-task framework of CLCNet. Within this framework, the primary task is the prediction of phenotypic value, while the prediction of low-dimensional SNP representation serving as the auxiliary task. The contrastive loss is calculated by comparing the Euclidean distance between the low-dimensional

SNP representations of two randomly paired samples with the absolute difference in their phenotypic values. This design has been shown to strengthen inter-task correlations, ensure diverse information sources for shared learning, as well as facilitate continuous model weight updates.

To the best of our knowledge, the application of a multitasking framework based on comparative learning to genomic prediction (GP) models is the first of its kind, although there is potential for further optimization. One such potential enhancement could be achieved with the random pairing method, which could be supplemented by exhaustive pairing. Specifically, in each training iteration, CLCNet randomly pairs a sample with another from the dataset to compute contrastive loss. While this random pairing method has proven to be effective, it typically results in a smaller number of unique pairings (e.g., with 100 samples, fewer than 100 unique pairings was made out of a theoretical 4950 possible combinations). The implementation of exhaustive pairing could enhance model performance by leveraging a broader diversity of sample combinations. To balance computational efficiency and performance, we propose increasing the sampling rate of pairwise combinations moderately.

The chromosome-aware module we designed in this study is an effective feature selection method. In the context of genomic data, where sample sizes are limited and feature dimensions are substantial, dimensionality reduction is paramount for achieving computational efficiency. Feature selection facilitates dimensionality reduction and noise reduction by identifying representative SNPs. While traditional methods combining GWAS and GS are useful for feature selection, GWAS typically employs linear models to assess associations between individual SNPs and phenotypes, overlooking SNP interactions (Liu et al., 2019). In contrast, LightGBM accounts for SNP interactions during feature selection, offering greater sensitivity and accuracy in

identifying relevant SNPs (Yan et al., 2021). In our approach, we employ LightGBM to capture SNP interactions, applying chromosome grouping for local SNPs and genome-wide selection for global SNPs. The selected local and global SNPs are then combined to form the final set of representative SNPs. These local and global SNPs complement each other and cannot replace one another.

In addition, the chromosome-aware module, another distinctive feature of CLCNet, accounts for linkage effects, a special form of local interaction in genomic data. The integration of local, global, and linkage effects enhances the model's performance, setting CLCNet apart from other models. Furthermore, the chromosome-aware module is flexible and can be conveniently introduced to a variety of GP models, including linear methods, machine learning approaches, and deep learning frameworks. This versatility positions the chromosome-aware module as a potent instrument for data preprocessing across a range of GP tasks.

CLCNet demonstrated no substantial enhancement in performance when compared to the optimal approach for GSTP10 (cotton) and GSTP11 (millet), primarily due to the restricted number of SNPs, which was impacted by LD pruning and MAF preprocessing. Following the elimination of these preprocessing steps, there was a substantial increase in the SNP number (Table 1). Upon re-evaluation, CLCNet exhibited superior performance, surpassing the most effective baseline methods (Figure 4a; Figure 4b). A similar observation was made in majority of tasks that had hitherto exhibited no improvement, which demonstrated significant gains, while those with significant improvements exhibited even higher levels of enhancement (Supplementary Figures 4 and 10 for GSTP10; Supplementary Figures 5 and 11 for GSTP11). Notably, this was accomplished without altering sample sizes, highlighting the strong potential of CLCNet for cost-effective planting scenarios by increasing sequencing depth.

Conclusion

We developed a novel deep learning model, the CLCNet, for plant GP. The integration of chromosome-aware and contrastive learning modules has resulted in CLCNet demonstrating superior performance in terms of prediction accuracy when compared to traditional linear models, machine learning models, and other deep learning approaches. The model effectively captures linkage effects, local and global epistasis, and optimizes molecular marker selection, thereby reducing data noise and improving overall performance. Furthermore, the multi-task modelling strategy of CLCNet, which incorporates the prediction of phenotypic difference as an auxiliary task, facilitates enhanced inter-task information sharing and a consequent increase in prediction accuracy. In summary, CLCNet is a reliable, efficient, and powerful tool for GS, excelling in high-dimensional datasets and holding significant 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, 8 datasets were selected as benchmarks based on sample size, SNP loci, and population type. Data dimensionality was reduced using PLINK (v1.9) for linkage disequilibrium (LD) pruning (*--indep-pairphase 100 10 0.2*) and minor allele frequency (MAF) filtering (*--maf 0.05* for GSTP012 and GSTP014). The final datasets, representing maize (*Zea*

mays), rice (*Oryza sativa*), cotton (*Gossypium hirsutum*), foxtail millet (*Setaria italica*), chickpea (*Cicer arietinum*), rapeseed (*Brassica napus*), and soybean (*Glycine max*), include 27 complex traits. Preprocessed datasets were summarized in Table 1.

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 transformed into one-hot encoding. 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 entire 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 loci 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 measured using the Pearson Correlation Coefficient (PCC), and prediction error was assessed using the Root Mean Squared Error (RMSE). To compare CLCNet with the best-performing baseline models, we calculated the PCC_{CLCNet} and PCC_{Best} (PCC values for CLCNet and the best baseline, respectively), $RMSE_{CLCNet}$ and $RMSE_{Best}$ (RMSE values for CLCNet and the best baseline, respectively), and ΔPCC and ΔRMSE (differences in PCC and RMSE between CLCNet and the best baseline).

The significance of performance differences was assessed using paired t-tests under a 10-fold cross-validation. Phenotype values of test set were z-normalized using the mean and variance from the training set to prevent data leakage. All the models were trained on the training set, and their performance was evaluated on the test set.

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, CLCNet and other methods were used to calculate the changes in PCC and RMSE before and after the SNP position shuffle.

Linkage disequilibrium decay analysis

SNPs with a minor allele frequency (MAF) < 0.05 and a missing rate > 10% were filtered using VCFtools (v0.1.15). Pairwise linkage disequilibrium (LD) was calculated as r^2 for the remaining SNPs within a 1 MB window.

Acknowledgements

We would like to express our gratitude to Dr. Jiachao Xu, a postdoctoral researcher from the Institute of Genetics and Developmental Biology of the Chinese Academy of Sciences, for his professional guidance on server construction. We also appreciate Prof. Fei Lu (the Institute of Genetics and Developmental Biology of the Chinese Academy of Sciences) for the insightful discussions. We thanks to Dr. Shihao Shao from Peking University Health Science Center for his valuable teaching and constructive suggestions on deep learning.

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

Supplemental Figure 1

Performance comparison of eight models on the GSTP001 dataset

Supplemental Figure 2

Performance comparison of eight models on the GSTP003 dataset

Supplemental Figure 3

Performance comparison of eight models on the GSTP007 dataset

Supplemental Figure 4

Performance comparison of eight models on the GSTP010 dataset

Supplemental Figure 5

Performance comparison of eight models on the GSTP011 dataset

Supplemental Figure 6

Performance comparison of eight models on the GSTP012 dataset

Supplemental Figure 7

Performance comparison of eight models on the GSTP013 dataset

Supplemental Figure 8

Performance comparison of eight models on the GSTP014 dataset

Supplemental Figure 9

he relationship between dataset characteristics and SNPs selected by the chromosome-aware module

Supplemental Figure 10

Performance comparison of eight models on the GSTP010 dataset after increasing the

number of SNPs

Supplemental Figure 11

Performance comparison of eight models on the GSTP011 dataset after increasing number of SNPs

Supplemental Figure 12

The CLCNet and its variant model architectures

Supplemental Table S1

A comparison of the results obtained by the CLCNet in relation to 27 trait predictions, with the results of the most effective method among the other methods.

References

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

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(2), 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.

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.

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.

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.

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. (2019). 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. (2019). 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.

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.

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, 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.

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.