Chao *et al. Plant Methods* (2025) 21:12
https://doi.org/10.1186/s13007-024-01321-0

Plant Methods

RESEARCH Open Access

MtCro: multi-task deep learning framework improves multi-trait genomic prediction of crops

Dian Chao[1†], Hao Wang[2†], Fengqiang Wan[1], Shen Yan[2*], Wei Fang[2*] and Yang Yang[1*]

Abstract

Genomic Selection (GS) predicts traits using genome-wide markers, speeding up genetic progress and enhancing breeding efficiency. Recent emphasis has been placed on deep learning models to enhance prediction accuracy. However, current deep learning models focus on learning specific phenotypes for the given task, overlooking the inter-correlations among different phenotypes. In response, we introduce MtCro, a multi-task learning approach that simultaneously captures diverse plant phenotypes within a shared parameter space. Extensive experiments reveal that MtCro outperforms mainstream models, including DNNGP and SoyDNGP, with performance gains of 1-9% on the Wheat2000 dataset, 1-8% on Wheat599, and 1-3% on Maize8652. Furthermore, comparative analysis shows a consistent 2-3% improvement in multi-phenotype predictions, emphasizing the impact of inter-phenotype correlations on accuracy. By leveraging multi-task learning, MtCro efficiently captures diverse plant phenotypes, enhancing both model training efficiency and prediction accuracy, ultimately accelerating the progress of plant genetic breeding. Our code is available on https://github.com/chaodian12/mtcro.

Key points

We revealed a strong correlation among plant phenotypes, providing a new perspective on the interaction between phenotypes in the field of gene-phenotype prediction.

We developed the MtCro model based on the concept of multitask learning, incorporating task-shared parameter networks and task-specific networks. This enables the model to simultaneously learn multiple phenotypes of plants.

Through extensive experiments on all phenotypes in different datasets, MtCro consistently outperforms traditional models, saving training resources while improving prediction accuracy. This validates the effectiveness of MtCro in handling plant phenotype prediction tasks.

†Dian Chao and Hao Wang contributed equally to this work.

*Correspondence:
Shen Yan
yanshen@caas.cn
Wei Fang
fangwei@caas.cn
Yang Yang
yyang@njust.edu.cn

Full list of author information is available at the end of the article

We compared the performance of single-tower MtCro and multi-tower MtCro with the same model architecture on identical phenotypes. Our analysis reveals that highly correlated phenotypes significantly promote the prediction results of the current phenotype.

Keywords Deep learning, Genomic prediction, Multi-task learning, Crop breeding

Introduction

Given the future challenges posed by population growth and climate change [1, 2], the annual increase in production needs to surpass the historical annual growth trends in yields [3]. Plant breeding technologies represent a viable partial solution to face the challenges ahead [4, 5]. As an emerging technology of this century, Genomic Selection (GS) has been increasingly applied in plant breeding [6]. GS employs markers spread across the entire genome for the purpose of genomic prediction [7], accelerates genetic advancements by reducing the duration of the breeding cycle, enabling rapid selection, and improving the efficiency of field evaluations [8]. Leveraging these advantages, GS has gained widespread recognition in numerous studies over recent years, and has been recently illustrated by a variety of plant breeding studies for crops like rice [9, 10], wheat [11, 12], and maize [13, 14].

To enhance the crucial accuracy of predictions, a wide array of methods have been developed for constructing GS prediction models [15–18]. By integrating techniques such as gBLUP [19], CropGBM [20], Cropformer [21], and DNNGP [22], which draw on principles of linear statistics and machine learning, this diverse toolkit offers a robust strategy for improving prediction accuracy within genomic studies. However, these prediction models reported could be further refined for multi-trait prediction tasks to align more closely with the practical application strategies of GS. Generally, breeders must balance multiple objectives, including optimizing yield, grain quality, and disease resistance, to ensure the productivity of a crop variety [21]. Therefore, breeders meticulously compile different types of phenotype data from the training population and execute genome sequencing, culminating in the establishment of a consolidated Genome-(multi)Trait dataset for prediction models. Current models typically split the dataset into several Genome-(single)Trait subsets, carrying out a full "train-test-predict" pipeline independently on each [20, 22, 23, 24]. In some cases, hundreds of models are built for varying phenotypes within the same population [25], adding substantial complexity to the task. Additionally, this approach overlooks the genetic correlations between different phenotypes, which could be beneficial in prediction efforts [26]. Moreover, dividing the dataset into multiple subsets reduces the efficiency of utilizing the training data. Recent studies have explored multi-trait prediction, but they often apply identical model weights across all phenotypes, overlooking trait-specific characteristics [27–31]. This generalization can limit performance compared to single-trait models, as it fails to capture the unique predictive patterns essential for each phenotype.

In this study, we introduce MtCro, incorporating the concept of multi-task learning in deep learning, to concurrently learn multiple phenotypes within a single plant. MtCro consists of a shared-bottom network and several task-specific tower networks. The shared-bottom network is responsible for learning the correlations between phenotypes, while the task-specific networks focus on capturing the specific features of individual phenotypes. This allows the MtCro to capture inter-phenotype correlations within a shared parameter space, enhancing predictive performance across various phenotypes. We conducted experiments on the Maize8652, Wheat2000, and Wheat599 datasets. Experimental results demonstrate that MtCro outperforms mainstream models in terms of Pearson correlation coefficients. Meanwhile, MtCro's efficiency is evident as it requires only one training session for different phenotypes within different datasets, saving considerable parameter tuning time. Finally, we conduct a comparative analysis of MtCro's performance on single-phenotype and multi-phenotype predictions under the same architecture. Results indicate a consistent 2-3% improvement in multi-phenotype predictions, especially for highly correlated traits.

Materials and methods

Datasets

This paper employs three datasets, encompassing two crops, namely wheat and maize. It conducts analyses on various phenotypes of these two crops, as well as the performance of the same phenotype in different environments. The datasets are Maize8652, Wheat2000 and Wheat599.

Maize8652 consists of 8652 samples of *F1 hybrid maize*, with recorded phenotypic measurements for days to tasseling (DTT), plant height (PH), and ear weight (EW). Produced through the crossings of a maternal pool and a panel of 30 paternal testers using a North Carolina-II design (as described in the Methods) [32], Maize8652 originated from a maternal pool known as CUBIC (Complete-diallel design plus Unbalanced Breeding-like Inter-Cross) [33]. This maternal population comprised 1428 inbred lines, derived from 24 elite founder lines that represented local-adaptive alleles. The paternal pool

comprised 30 tester lines representing six major heterotic groups, primarily consisting of improved overseas germplasms carrying advantageous foreign alleles. Consequently, the population is structured into thirty sets of paternal half-sibling subpopulations, abbreviated as *F_1s*, showcasing diverse patterns of heterosis effects in hybrid maize. Following the handling of data anomalies, including missing values, this study retained 27,379 genotype-phenotypes pairs. MtCro uses the coding scheme 0–9 to represent all forms of the genotype as follows: AA (0), AT (1), TA (1), AC (2), CA (2), AG (3), GA (3), TT (4), TC (5), CT (5), TG (6), GT (6), CC (7), CG (8), GC (8), GG (9). Following the encoding process, we utilized Principal Component Analysis (PCA [34] to reduce the dimensionality of the genotype data to 2,000 dimensions.

Wheat2000 includes 2000 *Iranian bread wheat* (Triticum aestivum) landraces sourced from the CIMMYT wheat gene bank [35]. Genotyping of these landraces was performed using 33,709 DArT markers. Individual alleles were coded as either 1 (present) or 0 (absent) in the recorded accessions. After implementing dimensionality reduction via PCA, a reduced set of principal components was retained for subsequent analysis. The dataset involves the evaluation of six agronomic traits: thousand kernel weight (TKW), test weight (TW), grain length (GL), grain width (GW), grain height (GH), and grain protein (GP).

Wheat599, a product of the Global Wheat Program at the International Maize and Wheat Improvement Center [11], is constituted by 599 *historical wheat* lines. The genotyping process involved the use of 1447 DArT (Diversity Array Technology) markers, which were generated by Triticarte Pty [36]. Ltd. based in Canberra, Australia (https://www.diversityarrays.com). Eliminat ion of markers with allele frequencies below 0.05 was performed, and any missing genotypes were imputed through the utilization of samples from the marginal distribution of marker genotypes. Following a stringent quality control process, a total of 1279 markers were retained. The average yield phenotype of the same genotype across four environments in this dataset is treated as four distinct phenotypes in this study. In the experimental section, we utilized the processed Wheat2000 and Wheat599 datasets, which were provided by the DNNGP [22].

MtCro

The fundamental idea of multi-task models involves parallel backpropagation through multiple outputs. Due to the sharing of a common hidden layer among the outputs, knowledge learned from different tasks can be shared while training tasks in parallel. MtCro introduces a mixed expert mechanism, dividing the shared layer into multiple expert groups. Multiple tasks share the learning from these expert groups, and a mixed network dynamically determines the weights of the expert groups corresponding to each task. This design enables the model to both share and differentiate specific knowledge among tasks.

The architecture of MtCro is depicted in Fig. 1.c, consisting of an input layer, multiple expert groups, gating networks, and tower networks. We annotated the mutation information in the SNPs, recording mutations as "1" (mutation occurred) and "0" (mutation did not occur). Subsequently, these mutation data were subjected to dimensionality reduction using PCA [34]. The reduced-dimensional data were then input into the model's input layer, with the data processed in batches according to the specified batch size. The expert groups, embedded within the shared bottom of the designed MtCro, consist of six specialized units. Each of these expert groups is structured with six layers, wherein each layer comprises a linear function, a batch normalization operation, a Rectified Linear Unit (ReLU) activation function, and a dropout mechanism. The linear function captures linear relationships, batch normalization standardizes inputs, and dropout regularizes the model. The choice of ReLU as the activation function is motivated by its ability to introduce non-linearity, aiding in learning complex patterns efficiently. Additionally, ReLU helps mitigate the vanishing gradient problem, ensuring effective training of the deep neural network. The sparse activation induced by ReLU enhances interpretability, contributing to the model's overall robustness and performance. The gating network dynamically allocates weights based on the traits that need to be predicted at the moment. Ultimately, the output layer generates the final prediction results.

Specifically, the output of the gating network layer is:

$$f^k(x) = \sum_{i=1}^{n} g^k(x)_i f_i(x) \tag{1}$$

In Eq. (1), $g^k(x)_i$ represents the output of the i_{th} gating network, indicating the weight of the i_{th} expert network for the k_{th} task, with $\sum_{i=1}^{n} g(x)_i = 1$. The term $f_i(x)$ denotes the i_{th} expert network. The primary objective of this step is to achieve conditional computation, where for each input example, the model selectively involves a subset of expert networks determined by the gating network. $f^k(x)$ represents the mixed output of the k_{th} task through the collaboration of expert and gating networks.

The computation method for the gating network $g^k(x)$ is:

$$g^k(x) = \mathrm{softmax}(W_{gk}x) \tag{2}$$

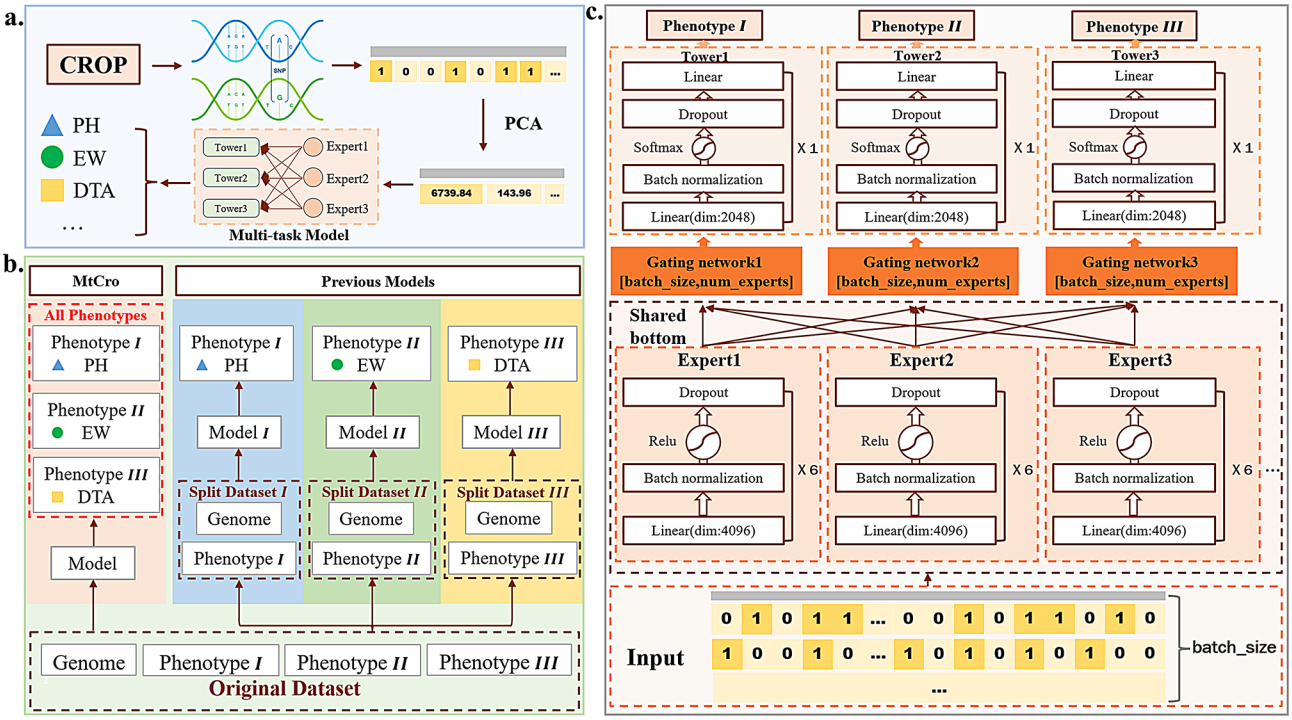

Fig. 1 **a**: Overall flowchart of the MtCro model. **b**: Comparison of the training processes between the MtCro model and other models. Traditional models require dataset splitting based on different objectives during training and the use of different models for training different phenotypes. In contrast, the MtCro model only requires a single model to simultaneously predict multiple phenotypes. **c**: Detailed architecture of the MtCro model, including a shared expert network layer composed of n experts and k task-specific tower network layers. Each tower network learns a phenotype

The matrix parameter $W_{gk} \in \mathcal{R}^{n \times d}$ in the formula is a learnable matrix, where n represents the number of expert networks, and d is the dimensionality of features. Each gating network for a task learns how to select a subset of expert networks. Each gating network effectively linearly partitions the input space into n regions, with each region corresponding to one expert.

Finally, the output for the k_{th} task is obtained through the tower network layer h^k:

$$y_k = h^k\left(f^k(x)\right) \tag{3}$$

The structure of the tower network closely resembles that of the expert network in the MtCro model. Unlike the expert network, the tower network adopts a simplified single-layer architecture. This streamlined design is intentional, focusing the tower network on the integration of expert network outputs without the added complexity of multiple layers. The activation function in the tower network is distinctive, as it opts for softmax, introducing normalization at the output layer. This operation facilitates the activation and recovery of task-specific features that may have been attenuated or lost in the expert network layers. Furthermore, a linear layer is strategically incorporated into the final layer of the tower network to map its output effectively to the ultimate prediction space for the task. The combination of Softmax and Linear layers enables the model to adaptively adjust task features across multiple tasks, allowing unique features that may have been diminished to be re-emphasized and reactivated through Softmax.

As shown in Fig. 1.b, the multitask model, compared to traditional models, can efficiently utilize a single model to learn the SNP information of plants and predict multiple traits. In contrast, traditional models require the use of multiple models to learn these diverse traits, highlighting the inherent advantages of multitask models. As a result, multitask models exhibit significant advantages in terms of model predictive capability, training time, and model parameter count when compared to traditional models.

Genetic prediction inputs typically involve features with several thousand dimensions. To address this issue, we undertook a redesign of the MtCro architecture, particularly adjusting parameters related to expert groups and towers to adapt to the larger-scale input parameters in genetic prediction tasks. Specifically, we dynamically adjusted the number of expert groups and towers during training based on the varying lengths of SNP groups. This design modification takes into account the specificity of genetic prediction tasks, ensuring that MtCro can better leverage its strengths when dealing with large-scale SNP data, thereby enhancing the model's performance.

Methods of comparison

This study extensively investigated the performance of different models in genomic selection by comparing multiple single-task models. In the realm of traditional models, the study first examined GBLUP (Genomic Best Linear Unbiased Prediction model) [19], a classical statistical model designed specifically for genomic selection. Based on a linear model, GBLUP analyzes individual genomic information to predict their performance and traits, widely applied in livestock and agricultural breeding. Secondly, a comparison was made with LightGBM [20], a decision-tree-based gradient boosting model known for its outstanding performance in handling large-scale data and high-dimensional feature scenarios. LightGBM exhibits efficient training speed and robust generalization capabilities, mainly employed for solving regression and classification problems. Finally, a comparison was drawn with SVR (Support Vector Regression model) [37], a well-known machine learning method that accepts various types of data as input and can be paired with multiple kernel functions for handling classification and regression problems.

In the realm of deep learning models, DeepGS [38], a deep learning genomic selection model based on deep convolutional neural networks. DeepGS utilizes deep convolutional neural networks to jointly represent features in the genotype through hidden variables, employing strategies such as convolution and sampling to reduce the complexity of high-dimensional genomic data. Secondly, DLGWAS [39] was compared, a whole-genome association analysis model based on a dual-CNN flow. DLGWAS uses convolutional neural networks to predict the quantity traits of SNPs, studies the contribution of genotypes to traits significantly, and treats missing SNPs as new genotypes in the deep learning model. DNNGP [22], a deep neural network model for genomic prediction. Then, DNNGP employs a multi-layered hierarchical structure and applies normalization, early stopping, rectified linear activation functions, etc., to prevent overfitting, making it adaptable to various omics data for phenotype prediction. The introduction of these deep learning models aims to explore more complex genotype-phenotype associations, providing more accurate and comprehensive predictions for genomic selection. SoyDNGP [24], inspired by the input structure of images, transforms SNP sequences into a multi-channel input, resembling image-like representations. Employing a distinctive convolutional neural network architecture, SoyDNGP has demonstrated high predictive capabilities, particularly on soybean and similar varieties. Lastly, MTUE [31] employs a fully connected architecture within a deep learning model to address genomic selection tasks across multiple phenotypes and environments. Experimental results demonstrate that MTUE achieves significantly higher predictive accuracy compared to single-trait methods in a single-environment context, highlighting the advantages of leveraging multi-trait and multi-environment data for genomic prediction.

Evaluation metrics utilized

We use the Pearson correlation coefficient to assess the relationship between predicted values and actual values. The Pearson correlation coefficient is a statistical metric that measures the strength and direction of a linear relationship between two variables. The calculation involves the covariance and the standard deviations of the two variables. Firstly, the covariance of the two variables is calculated, and then it is divided by the product of their respective standard deviations. The resulting value ranges from − 1 to 1, where 1 indicates a perfect positive correlation, -1 indicates a perfect negative correlation, and 0 indicates no linear relationship. The calculation of the Pearson correlation coefficient provides a quantitative measure to assess the strength and direction of the linear relationship between variables.

$$\gamma = \frac{\mathrm{cov}(X, Y)}{\sigma_X * \sigma_Y} \tag{4}$$

Equation 4 outlines the calculation method for the Pearson correlation coefficient. In the formula, cov(X, Y) represents the covariance between variables X and Y, while σ_X and σ_Y denote the standard deviations of variables X and Y, respectively.

Results

Phenotypic correlation analysis

In our study, we employed the Pearson correlation coefficient [40] as a statistical metric to measure the correlation between different phenotypes, and visually represented these relationships by creating a heatmap. The Pearson correlation coefficient is widely used for assessing the linear relationship between two variables, with values ranging from − 1 to 1. Specifically, a value of 1 indicates a perfect positive correlation, -1 signifies a perfect negative correlation, and 0 indicates no linear relationship. For each pair of phenotypes, we calculated the Pearson correlation coefficient, obtaining the strength and direction of the linear relationship between them through the computation of covariance and standard deviation. Given that the Pearson correlation coefficient can be positive or negative, and recognizing that negative correlations also indicate associations between phenotypes, we chose to take the absolute value during the creation of the heatmap. This operation mapped all correlation values to the range of 0 to 1, providing a clearer representation of the relationship between the two sets of phenotypes.

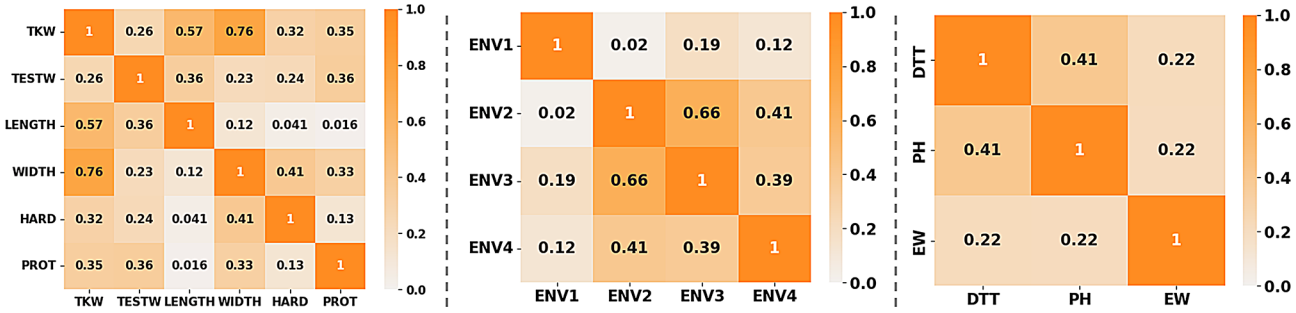

Fig. 2 This heatmap illustrates the Pearson correlation coefficients among all phenotypes in the Wheat2000/Wheat599/Maize8652 dataset. Each row and column represent a phenotype, and the color intensity in each cell reflects the Pearson correlation coefficient between the corresponding pair of phenotypes

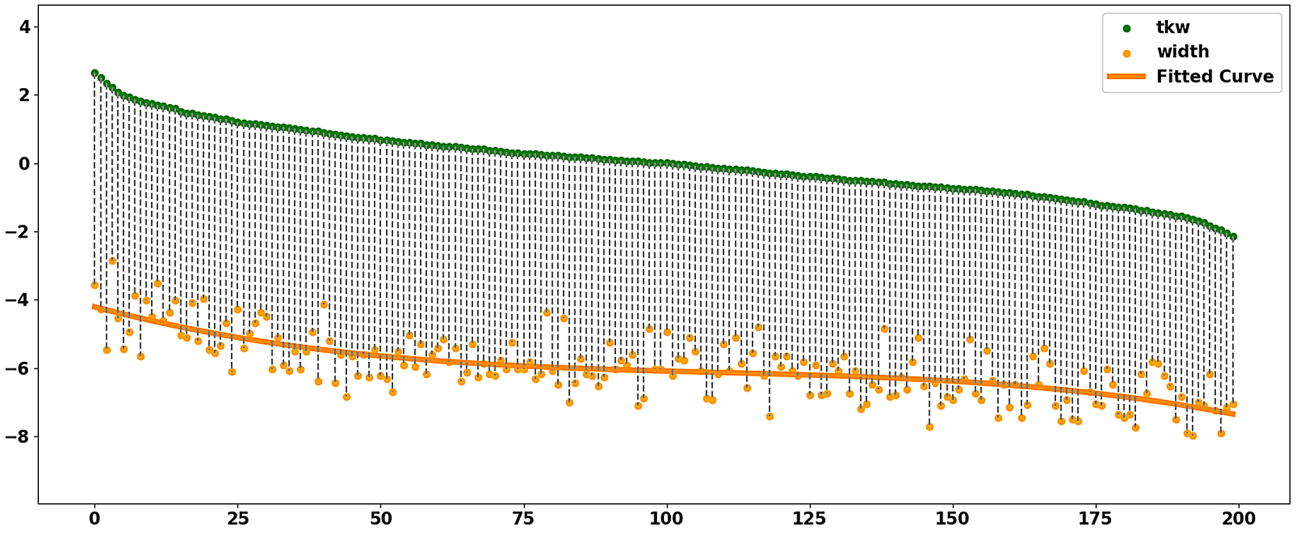

Fig. 3 Distribution of TKW and WIDTH Phenotype Data in the Wheat2000 Dataset. The processed phenotypic pairs are arranged in descending order based on TKW values. The horizontal axis represents different samples (selected every 10 samples), while the vertical axis depicts the values of TKW and WIDTH

As shown in Fig. 2, we conducted a detailed analysis of the phenotypic correlations within each of the three datasets, revealing strong relationships present in each dataset. In the Wheat2000 dataset, the correlation between the TKW and WIDTH phenotypes reached a significant 0.76. Further examination in the Wheat599 dataset focused on the correlation of the same phenotype across different environments. The results indicated that the phenotype in ENV1 exhibited lower correlations with phenotypes in the other three environments, while the correlation between phenotypes in ENV2 and ENV3 reached a high 0.66. Within the Maize8652 dataset, we observed a correlation of 0.44 between DTT and PH, and a correlation of 0.22 between EW and both DTT and PH. This finding suggests that different phenotypes and the same phenotype in different environmental conditions exhibit significant correlations. These results provide crucial insights for a deeper understanding of the relationships between different traits and their variations across diverse environmental settings.

In order to provide a clearer illustration of the specific relationship between highly correlated phenotypes in the Wheat2000 dataset, this study conducted an in-depth analysis of the TKW and WIDTH phenotypes using Fig. 3. Firstly, mean centering was applied to both TKW and WIDTH phenotypes. This involved calculating the average values for each phenotype in the dataset and then subtracting these averages from the corresponding measurement values. This mean centering process was employed to eliminate overall offsets between the two phenotypes, ensuring a more accurate and interpretable analysis. To spatially separate TKW and WIDTH for a clearer observation of the data distribution, WIDTH values were globally reduced by 6 units, shifting the data distribution to around −6. Finally, linear fitting was performed on the adjusted WIDTH values. The visualization results indicate a high level of consistency in the data distribution between TKW and WIDTH.

Parameter settings and experimental validation of MtCro

This study implemented the MtCro model based on the PyTorch framework. The model was configured with a learning rate of 0.0001, utilizing the Pearson distribution loss function. During the model training process, the batch size was set to 32 to enhance training accuracy, and the number of training epochs was set to 100. The model training configuration also incorporates a weight decay rate and an early stopping criterion to enhance generalization and efficiency. The weight decay rate was set to 0.00001, which helps to regularize the model by constraining the magnitude of the weights and thereby reducing the likelihood of overfitting. This small decay rate ensures that the model's parameters remain controlled, contributing to improved generalization on unseen data. Additionally, an early stopping patience parameter of 30 epochs was established, allowing the training to halt if there is no improvement in performance within this specified number of epochs. This early stopping mechanism reduces unnecessary computation, ensuring an optimal balance between training duration and model performance.

In the model configuration, each expert network within the shared bottom structure consists of a 6-layer Multilayer Perceptron (MLP). We conducted experiments to examine the impact of varying the number of MLP layers on model performance. The results, presented in Table 1, demonstrate the influence of MLP depth on the model's effectiveness and provide insights into selecting an optimal layer configuration for improved performance. We conducted experiments on the Wheat2000 dataset, keeping all parameter settings consistent except for the number of layers in the expert networks. The results indicate that as the number of layers increases, the model's overall performance initially improves but then declines. This trend suggests an optimal range for the number of layers, beyond which additional depth may lead to overfitting or diminished returns in performance.

MtCro compared with other mainstream methods

We conducted a comparative analysis between MtCro and six mainstream single-task learning methods, including GBLUP, LightGBM, SVR, DeepGS, DLGWAS, DNNGP, and SoyDNGP. All experiments used fixed random seeds, selecting 10% of the samples as the test set, and training the models on the remaining 90% of the samples. During training, 20% of the training samples were used as a validation set to select the best-performing epoch based on validation performance. Initial training and testing were performed on the Wheat2000 dataset, which is characterized by a relatively extensive set of phenotypes. The experimental results indicate that MtCro outperforms other models across all phenotypes. Specifically, it surpasses the best-performing single-task model by 1% in LENGTH and WIDTH phenotypes, and by 2–3% in TKW, TESTW, and HARD phenotypes (Fig. 4). Notably, its performance is particularly strong in the PROT phenotype, exceeding the best single-task model by 8%. Upon analysis, predictions for other phenotypes are consistently above 0.66, while PROT's predictions are relatively lower. This study attributes this to the use of a multi-task model, enabling the model to indirectly learn the poorly performing PROT phenotype by leveraging accurate predictions from other phenotypes. As a result, the model demonstrates improved predictions on the challenging PROT phenotype.

Next, we conducted experiments on the Wheat599 dataset, characterized by a more limited dataset and fewer phenotypes. Remarkably, the MtCro model showcased optimal efficacy across all phenotypes in this context in Fig. (5). Specifically, it surpassed the best-performing model by 1% in ENV1 and exhibited a superior margin of over 3% in the remaining three environments. Upon scrutinizing Fig. 3, it is discerned that the phenotypic correlations in ENV1 are comparably diminished, resulting in a relatively modest enhancement in the model's prediction of average yield for this environment. In contrast, the model demonstrated notable proficiency

Table 1 The effect of varying the number of layers in the expert networks on model performance, with all experiments conducted on the Wheat2000 dataset. The "Expert Layer" column represents different layer configurations, and the results are reported as Pearson correlation coefficients

Expert Layer	TKW	TESTW	LENGTH	WIDTH	HARD	PROT
1	0.62	0.58	0.68	0.75	0.61	0.51
2	0.62	0.59	0.68	0.75	0.62	0.52
3	0.64	0.60	0.69	0.76	0.64	0.54
4	0.65	0.60	0.70	0.76	0.65	0.55
5	0.68	0.63	0.72	0.76	0.68	0.56
6	0.70	0.66	0.75	0.78	0.71	0.58
7	0.69	0.65	0.76	0.79	0.67	0.57
8	0.66	0.63	0.73	0.76	0.64	0.54
9	0.65	0.60	0.70	0.76	0.64	0.53
10	0.64	0.60	0.69	0.76	0.63	0.52

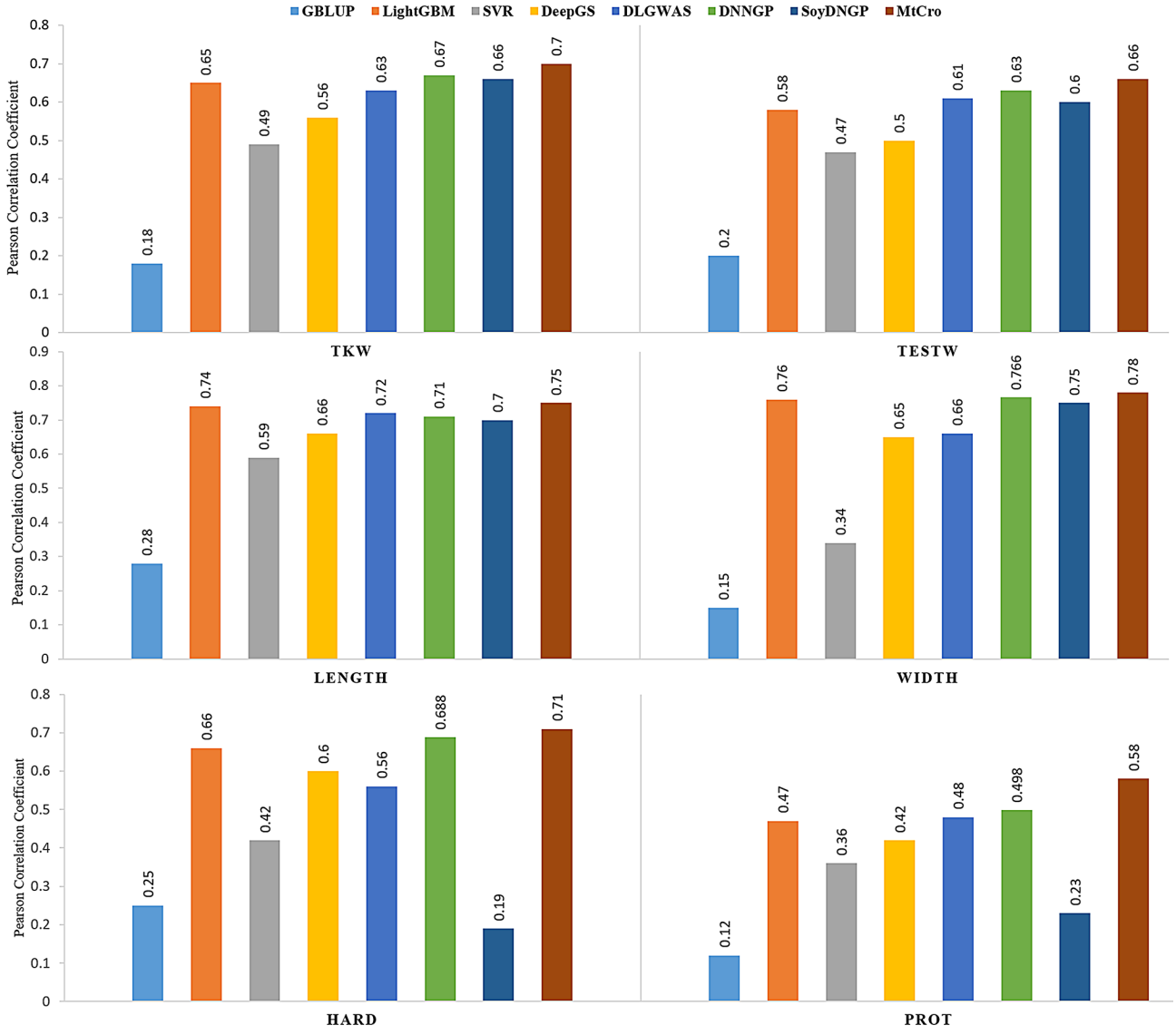

Fig. 4 Pearson Correlation Coefficients between Model Predictions and Ground Truth on the Wheat2000 Dataset

in predicting average yield across the other three environments, where phenotypic correlations are more robust. We observed a relatively poorer performance of the DeepGS model on the Wheat599 dataset compared to the Wheat2000 dataset. This can be attributed to the limited sample size in the Wheat599 dataset, which is not conducive to the deep hierarchical network architecture employed by DeepGS.

Then, leveraging the Maize8652 dataset with a more comprehensive representation of genotypes, a meticulous comparative analysis was conducted among models exhibiting superior performance on the Wheat2000 and Wheat599 datasets—specifically, LightGBM, DNNGP, SoyDNGP, MTUE and MtCro (in Fig. 6). The assessment encompassed a thorough examination of Pearson correlation coefficients and Mean Squared Error (MSE) for each model. The findings reveal that, for the Ear Weight (EW) phenotype with the lowest phenotypic correlation, MtCro surpassed the second-best LightGBM model by 1%. In the instances of the strongly correlated Days to Tasseling (DTT) and Plant Height (PH) phenotypes, MtCro outperformed the second-best models by 2 and 3%, respectively. In terms of MSE calculations, LightGBM exhibited a marginally superior result of 0.003 compared to MtCro. However, when scrutinizing the MSE for the other two phenotypes, LightGBM displayed more significant mean squared losses, signifying a diminished stability in its quantitative prediction results compared to MtCro.

Additionally, we observed that multi-task models can simultaneously learn multiple phenotypes to be predicted in a single training session, whereas other approaches

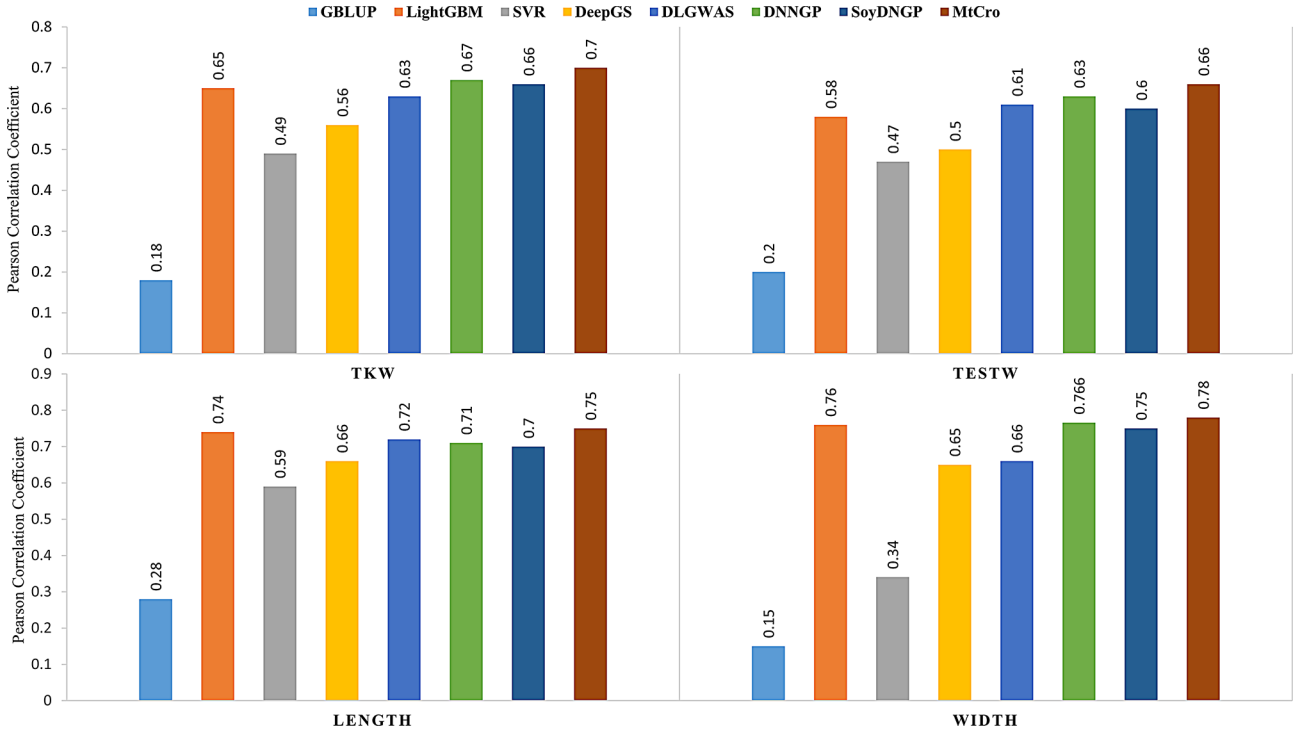

Fig. 5 Pearson Correlation Coefficients between Model Predictions and Ground Truth on the Wheat599 Dataset

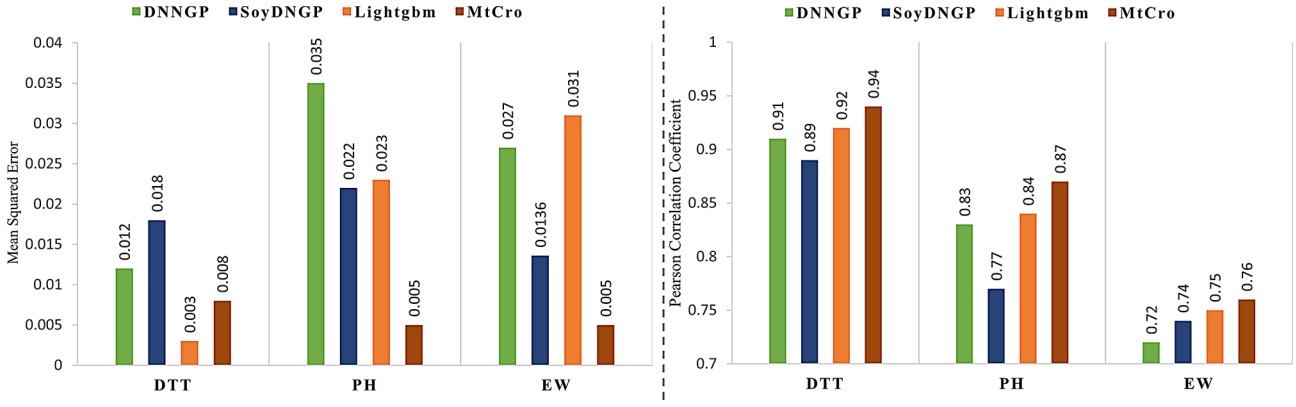

Fig. 6 Performance Evaluation of Various Models on the Maize8652 Dataset. The right figure displays the MSE for each model, while the left figure illustrates Pearson correlations

require training multiple models based on the number of phenotypes predicted (as shown in Fig. 1.b). Deep learning models also involve tuning various parameters during training, such as batch size and learning rate, resulting in significantly higher training costs for multiple models. Multi-task models not only enhance the predictive capabilities across various phenotypes but also save training costs.

Unveiling the benefits of multi-tasking by revitalizing MtCro

To further dissect the advantages brought by inter-phenotypic associations to the models, we devised a Single-Tower MtCro model tailored for single-phenotype prediction, contrasting it with the Multi-Tower MtCro model designed for multi-phenotype prediction. Specifically, the Single-Tower MtCro comprises only the Tower network layer and Gating Network dedicated to the current task, excluding Tower Network layers and Gating Networks for other tasks. The structure and number of Experts in the Shared Bottom remain consistent with the Multi-Tower MtCro. To ensure that the Single-Tower MtCro model does not incorporate information from other phenotypes, all shared bottom modules within the Single-Tower MtCro model were trained from scratch. This design allows us to scrutinize the gains in predictive

Fig. 7 Comparison of Results between Single-Tower and Multi-Tower MtCro Models on the Wheat2000 Dataset. The Single-Tower model takes only the phenotype of the current task as input, resembling a single-task model. In contrast, the Multi-Tower model incorporates inputs involving all phenotypes within the current dataset, essentially representing a multi-task model with the same architecture as the Single-Tower MtCro

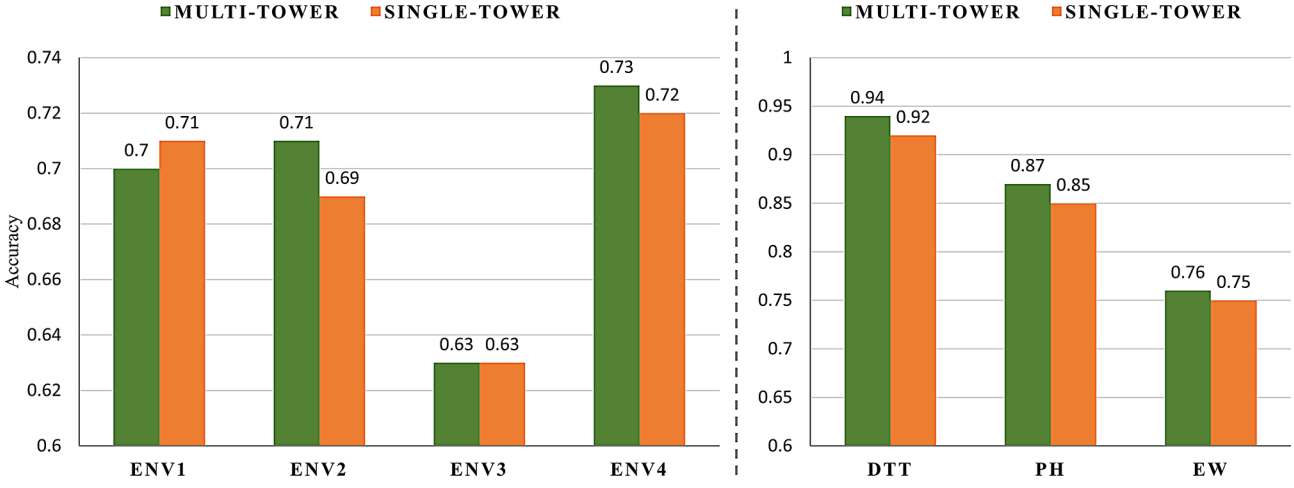

Fig. 8 Comparison of Results between Single-Tower and Multi-Tower MtCro Models on the Wheat599(left) and Maize8652(right)

performance attributed to the consideration of other phenotypes when focusing on a particular phenotype.

We observed a consistent enhancement in the predictive performance of the current phenotype when considering other phenotypes, as shown in Fig. 7. From the heatmap analysis in Fig. 2, it is evident that there is a strong correlation between phenotypes in this dataset. As a result, the Multi-Tower MtCro model outperforms the Single-Tower MtCro model across all phenotypes. Specifically, there was a 1% improvement for the LENGTH phenotype, a 2-3% enhancement for TKW, WIDTH, and TESTW phenotypes, and the most substantial improvement of 4–5% for the HARD and PROT phenotypes. This supports our hypothesis that phenotypes with high predictive accuracy can significantly enhance the predictive capabilities of phenotypes with lower accuracy.

Figure 8 illustrates the outcomes on the Wheat599 and Maize8652 datasets, where the heatmap analysis reveals weaker correlations between phenotypes compared to the Wheat2000 dataset. As a result, the improvement of the Multi-Tower MtCro model over the Single-Tower MtCro model is relatively modest, with a 1–2% improvement across all phenotypes in Maize8652. However, in the Wheat599 dataset, there is a 1% negative impact on the ENV1 phenotype. Analysis revealed a correlation coefficient of only 0.1 between the ENV1 phenotype

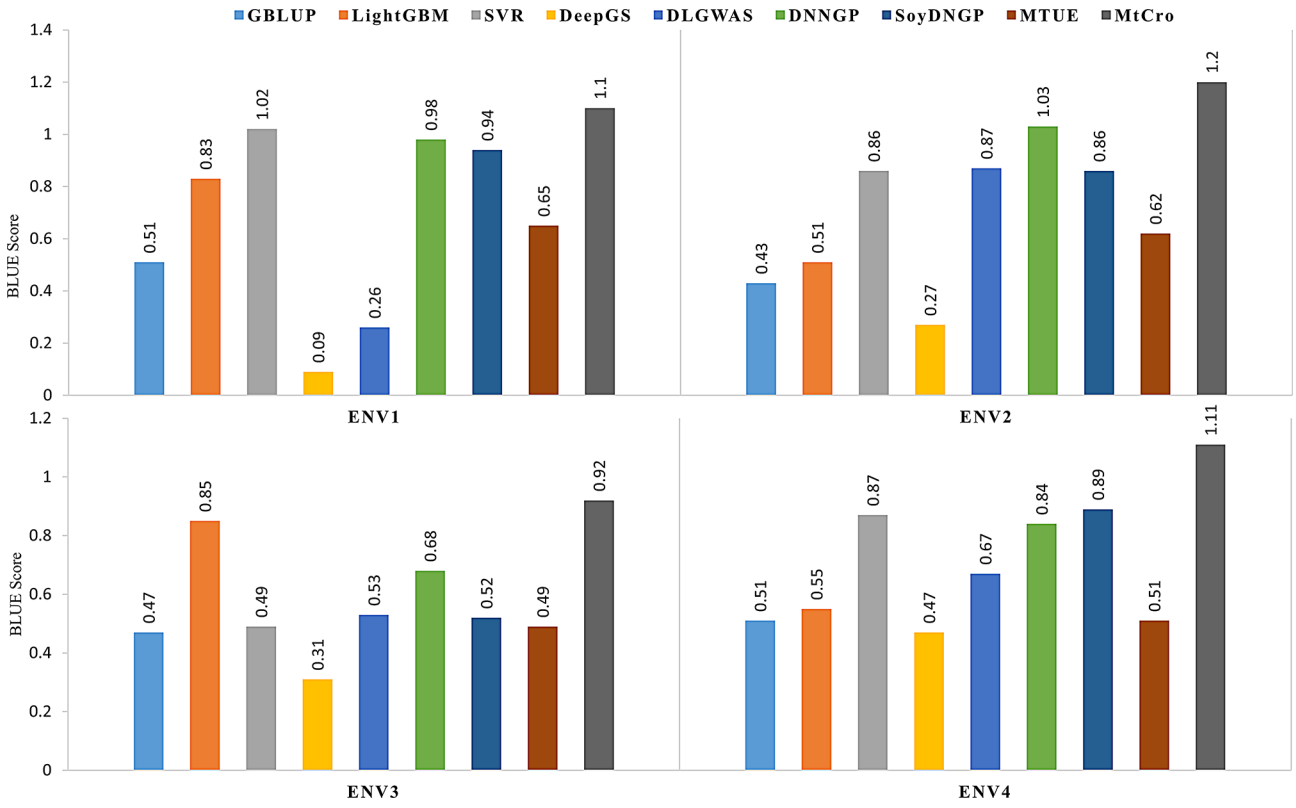

Fig. 9 BLUE Score between Model Predictions and Ground Truth on the Wheat599 Dataset

Table 2 Standard Deviations of Various Models under Cross-validation on the Maize8652 dataset. "SD(MSE)" refers to the standard deviation of the MSE metric, and "SD(Pearson)" refers to the standard deviation of the Pearson correlation coefficient

SD(MSE)	DTT	PH	EW	SD(Pearson)	DTT	PH	EW
DNNGP	0.00098	0.00155	0.00265	DNNGP	0.004	0.005	0.005
SoyDNGP	0.00073	0.00249	0.00431	SoyDNGP	0.008	0.009	0.011
Lightgbm	0.00087	0.00215	0.00514	Lightgbm	0.011	0.007	0.008
MTUE	0.00081	0.00107	0.00239	MTUE	0.006	0.004	0.007
MtCro	0.00069	0.00083	0.00074	MtCro	0.003	0.003	0.002

and those in the other three environments within the Wheat599 dataset. We hypothesize that, in scenarios of extremely low correlation, these phenotypes might negatively affect the predictive performance of the current phenotype.

Comprehensive evaluation of model performance and robustness

Since the Wheat599 dataset contains data for a single trait across four different environments, the genomic prediction accuracy is likely to be more precisely evaluated by using Best Linear Unbiased Estimates (BLUEs) for these environments. Therefore, we used BLUE to fit the predicted and actual values and computed the regression coefficients between them to illustrate their linear relationship. The specific results are shown in Fig. 9. The MtCro model achieved optimal prediction performance across all environments, surpassing the multi-phenotype prediction model MTUE by 0.6 points in ENV4. Moreover, it outperformed the current best model by 0.07 to 0.22 points across all environments.

To further assess the robustness of the models, we conducted a five-fold cross-validation and evaluated the performance of various deep learning models under different parameter initializations on the Maize8652 dataset. The models evaluated in this study include LightGBM, DNNGP, SoyDNGP, MTUE, and MtCro, all of which have been previously validated on the Maize8652 dataset. Table 2 presents the standard deviations of these five models under five-fold cross-validation. The results indicate that, although the LightGBM model exhibits superior performance on the Maize8652 dataset, as shown in Fig. 6, it demonstrates a larger standard deviation when the dataset undergoes shifts. This suggests that its robustness is inferior compared to the other four deep learning models. In contrast, both DNNGP and MTUE

Table 3 Standard Deviations of Models with different parameter initializations on the Maize8652 dataset

SD(MSE)	DTT	PH	EW	SD(Pearson)	DTT	PH	EW
DNNGP	0.00066	0.00092	0.00094	DNNGP	0.003	0.005	0.004
SoyDNGP	0.00084	0.00102	0.00135	SoyDNGP	0.007	0.006	0.007
MTUE	0.00070	0.00089	0.00067	MTUE	0.004	0.003	0.003
MtCro	0.00057	0.00079	0.00062	MtCro	0.002	0.003	0.001

demonstrate smaller standard deviations, reflecting their greater robustness in the face of such shifts. Notably, MtCro achieves the smallest standard deviations across all phenotypes, highlighting its strong robustness due to its integration of multi-phenotype features alongside phenotype-specific characteristics. This enables MtCro to maintain stable performance even when data shifts occur, showcasing its superior robustness in dynamic environments.

In Table 3, we evaluate the performance of four deep learning models under different random seed initializations, focusing on the phenotypic performance metrics. The experimental indicators are the standard deviations of the MSE and Pearson correlation coefficient for three phenotypes on the Maize8652 dataset. The experimental results demonstrate that, despite variations in random seed initializations, MtCro consistently maintains the highest robustness among the four models.

Discussion

Genomic selection enables breeders to identify varieties with significantly improved agronomic performance, thereby addressing global food security challenges. In this work, we propose a multi-task deep learning approach to predict crop phenotypes by utilizing the genetic information contained in genomic data. The results demonstrate that MtCro can effectively integrate multiple tasks, such as simultaneous prediction of multiple traits in crops. The independent test set showed that MtCro could achieve similar and better performance than the single-task learning approach in both tasks. Overall, the success of MtCro in crop multi-tasking genome prediction provides a novel framework for breeders to help in the selection of superior breeding lines and accelerates the breeding cycle.

Plant breeders are routinely interested in multiple traits, which can complicate the progress of genomic selection. Typical predictive models are univariate (i.e., one trait), which fails to take full advantage of potential correlations between different traits in the genomic data, and the process is time-consuming. Furthermore, current multi-trait prediction models are outperformed by single-trait models due to the absence of task-specific layers. In this work, MtCro effectively utilizes a single model to learn SNP information from crop and predict multiple traits, with significant advantages in model prediction ability, training time, and number of model parameters. Unlike traditional architectures, we developed a distinctive MLP-based multitasking model with embedded input layers, expert components, gated networks, and tower networks, enabling dynamic weight allocation for enhanced robustness. The architecture is tailored to meet the specific needs of genetic prediction tasks, guaranteeing the model's superior performance in processing extensive SNP data. We compared MtCro with other methods for predicting complex traits in three different datasets and achieved excellent performance, demonstrating that MtCro can help in predicting crop multi-trait potentials.

Although our MtCro model achieves improved performance in multi-task prediction of crop phenotypes, some limitations remain. Crops are affected by both internal and external environments. However, our model does not incorporate environmental phenotypic variation and is unable to capture the complex genotype-environment type relationships [41]. To better understand crop phenotypes, environmental factors need to be fully explored to improve predictive breeding. Second, the current version of MtCro only considers genomic data of crops as input learning. In practice, incorporating and fusing multimodal information such as transcriptomic data, structural variation (SV) and proteomics data have the potential to improve model performance. While MtCro provides a new framework for genomic prediction, the model still needs to be tuned and trained before being used. Consistent with all deep learning models, the performance of the trained model is heavily dependent on the selected training model hyperparameters. In addition, exploring how multi-task learning can improve the predictive performance of models, how to reduce randomness in neural network training, and how to integrate more tasks by designing new loss functions will be important directions for future research.

In conclusion, this study proposes a multi-task neural network genome prediction method. To our knowledge, this is the first time that multitask learning has been used for genome prediction. MtCro can serve as a community resource that promises to accelerate crop breeding through targeted selection of superior germplasm for multiple traits. In future work, we plan to extend MtCro to support multi-phenotype prediction for more crops.

Acknowledgements
The author would like to thank National Key RD Program of China for providing valuable microbiological related resources.

Author contributions
Y.Y., W.F., and S.Y. conceived and designed the study. D.C. did most of the bioinformatics analysis. H.W.were helpful for materials collection. D.C., H.W., and F.Q.W wrote the paper. All authors read and approved the manuscript.

Funding
This work is partially supported by the National Key RD Program of China (2022YFF0712100), NSFC (62276131), the Fundamental Research Funds for the Central Universities (NO.NJ2022028, No.30922010317,No.30923011007) and the National Nature Scientific Foundation of China (32371996).

Data availability
No datasets were generated or analysed during the current study.

Declarations

Ethical approval
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China
[2]Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing 100081, China

Received: 7 May 2024 / Accepted: 26 December 2024
Published online: 05 February 2025

References

1. U.N.D. o.I. Economic, World population prospects, UN1978.
2. Lobell DB, Gourdji SM. The influence of climate change on global crop productivity. Plant Physiol. 2012;160(4):1686–97.
3. Ray DK, Mueller ND, West PC, Foley JA. Yield trends are insufficient to double global crop production by 2050. PLoS ONE. 2013;8(6):e66428.
4. Lenaerts B, Collard BC, Demont M. Improving global food security through accelerated plant breeding. Plant Sci. 2019;287:110207.
5. Zaidi SS-e-A, Vanderschuren H, Qaim M, Mahfouz MM, Kohli A, Mansoor S, Tester M. New plant breeding technologies for food security. Science. 2019;363(6434):1390–1.
6. McGowan M, Wang J, Dong H, Liu X, Jia Y, Wang X, Iwata H, Li Y, Lipka AE, Zhang Z. Ideas in genomic selection with the potential to transform plant molecular breeding: a review. Plant Breed Reviews. 2021;45:273–319.
7. Meuwissen TH, Hayes BJ, Goddard M. Prediction of total genetic value using genome-wide dense marker maps, genetics 157(4) (2001) 1819–1829.
8. Crossa J, Pérez-Rodríguez P, Cuevas J, Montesinos-López O, Jarquín D, De Los Campos G, Burgueño J, González-Camacho JM, Pérez-Elizalde S, Beyene Y. Genomic selection in plant breeding: methods, models, and perspectives. Trends Plant Sci. 2017;22(11):961–75.
9. Wang X, Li L, Yang Z, Zheng X, Yu S, Xu C, Hu Z. Predicting rice hybrid performance using univariate and multivariate GBLUP models based on North Carolina mating design II. Heredity. 2017;118(3):302–10.
10. Xu S, Zhu D, Zhang Q. Predicting hybrid performance in rice using genomic best linear unbiased prediction. Proc Natl Acad Sci. 2014;111(34):12456–61.
11. Daetwyler HD, Bansal UK, Bariana HS, Hayden MJ, Hayes BJ. Genomic prediction for rust resistance in diverse wheat landraces. Theor Appl Genet. 2014;127:1795–803.
12. Montesinos-Lopez OA, Montesinos-Lopez JC, Salazar E, Barron JA, Montesinos-Lopez A, Buenrostro-Mariscal R, Crossa J. Application of a Poisson deep neural network model for the prediction of count data in genome-based prediction. Plant Genome. 2021;14(3):e20118.
13. Wang B, Lin Z, Li X, Zhao Y, Zhao B, Wu G, Ma X, Wang H, Xie Y, Li Q. Genome-wide selection and genetic improvement during modern maize breeding. Nat Genet. 2020;52(6):565–71.
14. Kadam DC, Potts SM, Bohn MO, Lipka AE, Lorenz AJ. Genomic prediction of single crosses in the early stages of a maize hybrid breeding pipeline. G3: Genes Genomes Genet. 2016;6(11):3443–53.
15. Spindel J, Iwata H. Genomic selection in rice breeding, Rice genomics, genetics and breeding (2018) 473–96.
16. Ma X, Wang H, Wu S, Han B, Cui D, Liu J, Zhang Q, Xia X, Song P, Tang C. DeepCCR: large-scale genomics-based deep learning method for improving rice breeding. Plant Biotechnol J (2024) 1–3.
17. Yu S, Liu L, Wang H, Yan S, Zheng S, Ning J, Luo R, Fu X, Deng X. AtML: an Arabidopsis thaliana root cell identity recognition tool for medicinal ingredient accumulation. Methods. 2024;231:61–9.
18. Wang H, Lin Y-N, Yan S, Hong J-P, Tan J-R, Chen Y-Q, Cao Y-S, Fang W. NRT-Predictor: identifying rice root cell state in single-cell RNA-seq via ensemble learning. Plant Methods. 2023;19(1):119.
19. Clark SA, van der Werf J. Genomic best linear unbiased prediction (gBLUP) for the estimation of genomic breeding values, genome-wide association studies and genomic prediction (2013) 321–30.
20. Yan J, Xu Y, Cheng Q, Jiang S, Wang Q, Xiao Y, Ma C, Yan J, Wang X. LightGBM: accelerated genomically designed crop breeding through ensemble learning. Genome Biol. 2021;22:1–24.
21. Moeinizade S, Kusmec A, Hu G, Wang L, Schnable PS. Multi-trait genomic selection methods for crop improvement. Genetics. 2020;215(4):931–45.
22. Wang K, Abid MA, Rasheed A, Crossa J, Hearne S, Li H. DNNGP, a deep neural network-based method for genomic prediction using multi-omics data in plants. Mol Plant. 2023;16(1):279–93.
23. Ma W, Qiu Z, Song J, Li J, Cheng Q, Zhai J, Ma C. A deep convolutional neural network approach for predicting phenotypes from genotypes. Planta. 2018;248:1307–18.
24. Gao P, Zhao H, Luo Z, Lin Y, Feng W, Li Y, Kong F, Li X, Fang C, Wang X. SoyDNGP: a web-accessible deep learning framework for genomic prediction in soybean breeding. Brief Bioinform. 2023;24(6):bbad349.
25. He Q, Tang S, Zhi H, Chen J, Zhang J, Liang H, Alam O, Li H, Zhang H, Xing L. A graph-based genome and pan-genome variation of the model plant Setaria. Nat Genet (2023) 1–11.
26. Fernandes SB, Dias KO, Ferreira DF, Brown PJ. Efficiency of multi-trait, indirect, and trait-assisted genomic selection for improvement of biomass sorghum. Theor Appl Genet. 2018;131:747–55.
27. Montesinos-López O.A., Martín-Vallejo J., Crossa J., Gianola D., Hernández-Suárez C.M., Montesinos-López A., Juliana P., Singh R. New deep learning genomic-based prediction model for multiple traits with binary, ordinal, and continuous phenotypes. G3: Genes Genomes Genet. 2019;9(5):1545–56.
28. Montesinos-López OA, Montesinos-López A, Bernal Sandoval DA, Mosqueda-Gonzalez BA, Valenzo-Jiménez MA, Crossa J. Multi-trait genome prediction of new environments with partial least squares. Front Genet. 2022;13:966775.
29. Montesinos-López O.A., Montesinos-López A, Crossa J, Cuevas J, Montesinos-López JC, Gutiérrez ZS, Lillemo M, Philomin J, Singh R. A bayesian genomic multi-output regressor stacking model for predicting multi-trait multi-environment plant breeding data. G3: Genes Genomes Genet. 2019;9(10):3381–93.
30. Montesinos-López O.A., Montesinos-López A, Crossa J, Gianola D, Hernández-Suárez CM, Martín-Vallejo J. Multi-trait, multi-environment deep learning modeling for genomic-enabled prediction of plant traits. G3: Genes Genomes Genet. 2018;8(12):3829–40.
31. Mora-Poblete F, Maldonado C, Henrique L, Uhdre R, Scapim CA, Mangolim CA. Multi-trait and multi-environment genomic prediction for flowering traits in maize: a deep learning approach. Front Plant Sci. 2023;14:1153040.
32. Xiao Y, Jiang S, Cheng Q, Wang X, Yan J, Zhang R, Qiao F, Ma C, Luo J, Li W. The genetic mechanism of heterosis utilization in maize improvement. Genome Biol. 2021;22(1):1–29.
33. Liu H-J, Wang X, Xiao Y, Luo J, Qiao F, Yang W, Zhang R, Meng Y, Sun J, Yan S. CUBIC: an atlas of genetic architecture promises directed maize improvement. Genome Biol. 2020;21:1–17.
34. Abdi H, Williams LJ. Principal component analysis. Wiley Interdisciplinary Reviews: Comput Stat. 2010;2(4):433–59.
35. Ortiz R, Braun H-J, Crossa J, Crouch JH, Davenport G, Dixon J, Dreisigacker S, Duveiller E, He Z, Huerta J. Wheat genetic resources enhancement by the International Maize and Wheat Improvement Center (CIMMYT). Genet Resour Crop Evol. 2008;55:1095–140.
36. Crossa J, Jarquín D, Franco J, Pérez-Rodríguez P, Burgueño J, Saint-Pierre C, Vikram P, Sansaloni C, Petroli C, Akdemir D. Genomic prediction of gene bank wheat landraces, G3: genes, genomes. Genetics. 2016;6(7):1819–34.
37. Caraka RE, Chen RC, Toharudin T, Tahmid M, Pardamean B, Putra RM. Evaluation performance of SVR genetic algorithm and hybrid PSO in rainfall forecasting. ICIC Express Lett Part B Appl. 2020;11(7):631–9.

38. Ma W, Qiu Z, Song J, Cheng Q, Ma C. DeepGS: Predicting phenotypes from genotypes using Deep Learning, BioRxiv (2017) 241414.
39. Liu Y, Wang D, He F, Wang J, Joshi T, Xu D. Phenotype prediction and genome-wide association study using deep convolutional neural network of soybean. Front Genet. 2019;10:1091.
40. Cohen I, Huang Y, Chen J, Benesty J, Benesty J, Chen J, Huang Y, Cohen I. Pearson correlation coefficient, Noise reduction in speech processing (2009) 1–4.
41. Xu Y, Zhang X, Li H, Zheng H, Zhang J, Olsen MS, Varshney RK, Prasanna BM, Qian Q. Smart breeding driven by big data, artificial intelligence, and integrated genomic-enviromic prediction. Mol Plant. 2022;15(11):1664–95.

Publisher's note