Deep learning-based phenotype prediction analysis of genotype-environment interactions and mining of environmentally stable germplasm and elite loci in cotton

Xiongming Du

duxiongming@caas.cn

State Key Laboratory of Cotton Biology, Institute of Cotton Research of the Chinese Academy of Agricultural Sciences https://orcid.org/0000-0001-8431-5592

Jiale Li

Agricultural information Institute of the Chinese Academy of Agricultural Sciences, Nanfan Research Institute of the Chinese Academy of Agricultural Sciences

Zikang He

Agricultural information Institute of the Chinese Academy of Agricultural Sciences, Nanfan Research Institute of the Chinese Academy of Agricultural Sciences

Zhen Peng

National Key Laboratory of Cotton Bio-breeding and Integrated Utilization, Institute of Cotton Research, Chinese Academy of Agricultural Sciences

Xiuxiu Wang

Institute of Cotton Research of the Chinese Academy of Agricultural Sciences

Longyu Huang

Institute of Cotton Research of the Chinese Academy of Agricultural Sciences, Nanfan Research Institute of the Chinese Academy of Agricultural Sciences

Hongge Li

National Key Laboratory of Cotton Bio-breeding and Integrated Utilization, Institute of Cotton Research, Chinese Academy of Agricultural Sciences

Shoupu He

National Key Laboratory of Cotton Bio-breeding and Integrated Utilization, Institute of Cotton Research, Chinese Academy of Agricultural Sciences https://orcid.org/0000-0001-8434-5484

Guomin Zhou

Nanfan Research Institute of the Chinese Academy of Agricultural Sciences,Nanjing Institute of Agricultural Mechanization, Ministry of Agriculture and Rural Affairs

Jianhua Zhang

Agricultural information Institute of the Chinese Academy of Agricultural Sciences, Nanfan Research Institute of the Chinese Academy of Agricultural Sciences

---

Article

Keywords: cotton, genomic prediction, deep learning, genome loci, environmental adaptation

Posted Date: September 11th, 2025

DOI: https://doi.org/10.21203/rs.3.rs-7483914/v1

Additional Declarations: There is **NO** Competing Interest.

Deep learning-based phenotype prediction analysis of genotype-environment interactions and mining of environmentally stable germplasm and elite loci in cotton

Abstract: This study investigates the complex regulatory mechanisms of genotype-environment interactions (GEI) in cotton phenotype formation and explores the genetic basis of environmental adaptation through an integrated analytical approach. The research methodology encompasses four key components: (1) deep learning model construction, (2) phenotypic plasticity analysis, (3) environmental adaptation assessment, and (4) genome-wide association study (GWAS). Based on the multi-head self-attention mechanism and the deep feature interaction, we constructed the AttGEI-Net deep learning framework. The model demonstrates remarkable predictive performance with an average accuracy of 0.96 in fixed environments, though this decreases to 0.39-0.44 in novel environments, revealing fundamental differences between genotype-dominated and environment-dominated prediction scenarios. A total of 10,215 significant SNP loci is identified by GWAS, including 2,705 Main-SNPs, 41 phenotype plasticity loci (PP-SNPs), and 9,022 environmental adaptation loci (EvA-SNPs). The regulation of phenotypes by these loci has a distinct hierarchical character: the basic genetic architecture (Main-SNPs) maintains the basic expression of traits, the PP-SNPs mediates the immediate response of phenotypes to environmental changes, and the EvA-SNPs constitutes the highest-level adaptive regulatory network that coordinates the expression of multiple traits by integrating environmental signals. Shared loci of interpretability analyses of model and GWAS may be the key genetic basis adapting to different environments. Broadly adapted varieties in the Yellow River basin (e.g., F096, L090, etc.) can be used as the backbone parents for suitability breeding.

Keywords: cotton; genomic prediction; deep learning; genome loci; environmental adaptation.

Introduction

Cotton, as one of the globally important cash crops, its yield and quality directly affect the textile industry and agricultural economy, of which upland cotton has become an important natural fiber worldwide [1]. The trend of global warming has led to more complex environmental factors for crop growth in the future [2], and the breeding of high-yielding and environmentally adapted varieties is an important goal of breeding. Phenotype is the result of genotype-environment interaction (GEI) [3]: the expression of genes determines the phenotypic variation of varieties in different environments, and environmental factors in turn also affect gene expression, and the two together shape the phenotype of the crops. Through this GEI we

can select genotypes that are adapted to the environmental variability and thus screen out superior varieties to meet the needs [4]. Phenotypic variation of varieties is affected by multiple factors such as phenotypic plasticity and environmental adaptation. Current research focuses on the study of genotype-based phenotypic regulation, but the interaction between the genotypes and environment exhibits a complex relationship at multiple levels. Elucidating these complex relationships through genomic and genetic analyses is essential for establishing a robust foundation for genome-wide selective breeding and precision breeding strategies.

Environmental adaptation (EaV) refers to the ability of a genotype to be unaffected by changes in environmental conditions and is closely related to GEI [5,6,7]. Phenotypic plasticity (PP) refers to the ability of an individual to change its phenotype in response to environmental changes [8,9]. Methods such as the AMMI model and the GGE biscale map reveal GEI patterns to some extent, thus identifying ideal genotypes for multiple environmental adaptations [10,11]. Clarifying the intrinsic mechanisms by which EaV, PP, and GEI affect phenotypic variation and resolving the formation process of environmental adaptation in crops is a guarantee for coping with future environmental changes. Based on this, many studies have found that incorporating environmental data for GWAS can accurately identify environmental adaptation loci [12,13], and have identified some environmental adaptation-related loci using environmental variables and multi-year, multi-location trait data [14,15,16]. However, there are a limited number of related studies in the crop field [17], with even fewer studies dedicated to revealing the genetic basis of environmental adaptation in upland cotton [18]. The traits analyzed by association are not comprehensive enough, and there is a scarcity of methodologies focused on the identification of loci related to environmental adaptation. Furthermore, the differential regulatory roles of environmental adaptation loci versus main effector loci in phenotypic variation have yet to be thoroughly analyzed.

Integrating environmental factors such as light, temperature and precipitation can accurately predict the phenotypes of genotypes in target environments [19], but the uncertainty of environmental conditions makes it difficult to select superior varieties through past experience [22]. And screening out environmentally adapted varieties also requires the conduct of multi-year, multi-location field experiments, which are carried out across regions and years in order to increase the complexity of trait assessment [23]. While these extensive experiments generate a large amount of data that increase the burden and cost of research, they simultaneously provide an engine for the development of data-driven phenotypic prediction models. Previous models have predicted phenotypes based on genotypes [24.25] and lacked consideration of the effects of environmental factors, and recent studies have begun to incorporate environmental data to mine for factors affecting phenotypic variation. Statistical analysis models including GBLUP, BayesA, and BayesB are able to resolve the GEI effect

[26.27.28], but rely on linear assumptions and have difficulty capturing complex nonlinear relationships, e.g., Gustavo de los Campos et al [29] built a phenotypic prediction simulation platform covering basic linear mixed models and genotype-environment interactions, and predicted wheat yields using a large-scale dataset with an accuracy of up to 0.58. Machine learning methods such as LightGBM, XGBoost, and Random Forest [30,31,32] can handle high-dimensional data, but have limited feature interaction modelling capability and reduced generalisation performance when predicting across environments, e.g., Kunhui He et al [33] integrated multiple machine learning methods to combine environmental parameters across multiple developmental stages, and predicted maize plant height with an accuracy of 0.88. Deep learning through deep neural networks can effectively integrate time-series environmental and genetic data. It automatically learn higher-order features of environmental and genomic data and capture nonlinear interactions effects between genotypes and environments [34.35], with unique advantages in integrating genomic selection models with environmental data [36], e.g., Qixiang Zou et al. [37] who proposed a multimodal fusion framework, by modelling time series information of different environmental variables, the prediction accuracy of wheat plant height in multiple scenarios reached 0.90, 0.86, and 0.66, respectively. Although these methods have achieved some success in predicting phenotypes based on GEI effects, there are still shortcomings and challenges. Firstly, the heterogeneity of environmental data (e.g., temperature, precipitation, etc.) and genomic data leads to the difficulty of traditional methods in effectively integrating multi-source data, and the high-noise and multi-scale characteristics of multi-source data put forward a higher requirement on the feature extraction ability of the model. Secondly, the limited interpretability of the models makes it difficult to quantify the contribution of genetic and environmental factors to the phenotype and lacks the analysis of the intrinsic mechanism of GEI for important traits. Finally, there is a lack of mining of environmental adaptation loci.

Therefore, in this study, based on the principle of attention mechanism and deep feature interaction, we constructed the AttGEI-Net deep learning framework, which achieves the deep fusion and dynamic association modelling of genotype-environment heterogeneous data through the enhanced encoder, the multi-head self-attention mechanism, and the bilinear interaction layer with the three core functions of multilevel feature extraction, interpretable interaction modelling, and environmentally-adapted loci mining. The experimental dataset contains 1720 upland cotton genotypes planted in multiple locations over many years, which are divided into three datasets of 1201, 419, and 100 [48], which were used for the 13 traits of model training, testing and validation, respectively. This study seeks to overcome the constraints of conventional models in integrating heterogeneous genotype-environment data by employing deep learning techniques to investigate the underlying mechanisms of GEI. Furthermore, it aims to elucidate the hierarchical genetic architecture of regulatory phenotypes

through multi-parameter genome-wide association studies (GWAS). By integrating GWAS with model interpretability to mine key environmentally adapted loci and screen broadly adapted cotton varieties through environmental adaptation assessment. In conclusion, this study provides a reliable tool and theoretical basis for integrating environmental data into genomic selection and further reveals the intrinsic mechanism of GEI in influencing phenotypes, and explores the elite loci, which provides key targets and data-driven scientific methods for cotton breeding.

Results

This study consists of two parts, one is the construction of a cotton phenotype prediction deep learning model through integrating genotype and environmental data (Fig. 1A, B, C, D, E, H), and the other is the mining of environmentally stable germplasm and elite loci (Fig. 1F, G, H).

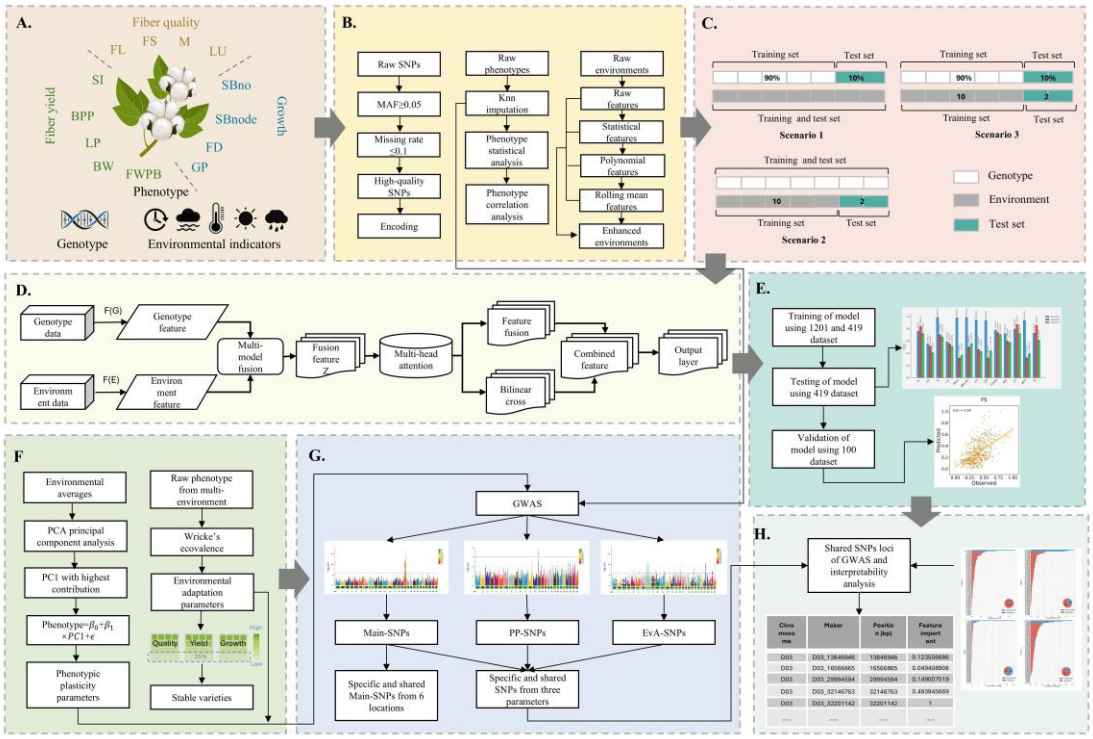


Fig. 1 Workflow of this study. (A) Dataset. The dataset contains 1720 cotton genotypes and was divided into three subsets: (i) The training samples (Dataset 1, N=1201; Dataset 2, N=419); (ii) The testing and genome-wide association study (GWAS) samples (Dataset 2, N=419); (iii) The validation samples (Dataset 3, N=100). **(B) Data analysis and preprocessing.** This section contains methods and procedures for quality control of SNPs, phenotype data, and environmental data processing. **(C) Data split.** The testing samples (Dataset 2, N=419) are divided under three scenarios. **(D) Architecture of AttGEI-Net Model.** The general framework of the model is shown. **(E) Assessment and validation of**

AttGEI-Net Model. Results of model testing and validation are shown. **(F) Phenotypic plasticity analysis and environmental adaptation assessment.** Processes for quantifying phenotypic plasticity and environmental adaptation, and screening for broadly adapted varieties. **(G) Genome-wide association study.** Analysis of SNPs loci identified by GWAS based on three parameters (phenotype values, phenotypic plasticity parameters, environmental adaptation parameters), the three parameters identified SNPs loci are Main-SNPs, PP-SNPs, EvA-SNPs, respectively. (**H) Interpretation analysis.** Interpretability analysis of the AttGEI-Net model demonstrates the extent of genetic and environmental contributions to the phenotype and identifies SNPs loci shared with GWAS.

Diversity of dataset. The experiment dataset contained a total of 1720 upland cotton genotypes, and this data was divided into three datasets, 1201, 419 and 100, which were used for model training, testing and validation, respectively. The 1201 dataset consisted of 1201 genotypes that were planted at Anyang (AY) in Henan Province in 2018 and was used as the training set. The 419 dataset consisted of 419 genotypes that were planted at six different locations in 2014 and 2015 (12 environments in total) and was used as a training and testing set and GWAS. The six locations were Anyang (AY) in Henan Province, Cangzhou (CZ) in Hebei Province, Yancheng (YC) in Jiangsu Province, Jingzhou (JZ) in Hubei Province, Dunhuang (DH) in Gansu Province, Alaer in the Xinjiang (XJ) autonomous region. The 100 dataset includes 100 genotypes that were planted at three different locations in 2018 and 2019 and was used as validation set. The three locations were Anyang (AY) in Henan Province, Alaer (XJ) and Kuitun (KT) in the Xinjiang autonomous region. Based on the distribution of phenotype values of the traits (Supplementary Fig. 1A~C), 13 traits were selected in this study for model training, test and validation as well as GWAS, which were classified into three major categories: fiber quality-related traits, growth period-related traits, and fiber yield-related traits. (Fig. 1A, Supplementary Table 1, and Methods). The environmental data of all locations are described in Methods.

Given that 419 upland cotton contains phenotype data from 6 locations over 2 years (12 environments in total), it is better able to study genotype-environment interactions (GEI) and therefore was selected to carry out data diversity analyses.

Analysis of Ma Zhiying et al.'s study found [48] that 419 upland cotton was divided into three genetic groups, G1, G2 and G3 (Supplementary Fig. 1D), which included 65, 268 and 86 samples, respectively, and that G1 was mainly cultivated in China's Yangtze River Basin, G2 was mainly cultivated in the Yellow River Basin of China, and G3 mainly consisted of varieties from the northwestern inland cotton region of China.

Analyses of the coefficient of variation (CV) of 419 upland cotton phenotypes revealed

the variability characteristics of important traits and the diversity of their responses to the environment. The study showed that the pattern of phenotype variation in single environments showed clear category specificity (Supplementary Fig. 1E): fiber yield-related traits (e.g. BPP and FWPB) exhibited the highest degree of phenotype variation, with single-environment CV values of 15-30%, while fiber quality-related traits (FL, FS, LU) showed greater stability (single-environment CV <10%). Cross-environmental analyses further showed (Supplementary Table 2) that BPP showed the most significant variation (cross-environmental CV = 44.09% ± 2.77), followed by growth period-related traits (CV = 10 ~ 40%), whereas most of the cross-environmental CVs of fiber quality-related traits were lower than 10%, reflecting the differential pattern of different traits in their response to the environment. Notably, all traits exhibited a wide span of phenotype variation (e.g., BPP range of 2.40~35.70), which, combined with high environmental coefficients of variation (CVs), confirmed that 419 upland cotton possesses both abundant genetic variation and responsiveness to environmental changes. These characteristics make 419 upland cotton an ideal material for analyzing genotype-environment interactions (GEI), which is particularly suitable for identifying environmental adaptation loci and selecting broadly adapted varieties.

Based on the correlation analysis of 13 traits in 419 upland cotton (Supplementary Fig. 2), we found significant associations in same trait category, with fiber quality-related traits (FL, FS, LU) showing highly significant positive correlations ($p<0.001$) at all six locations, and fiber yield-related traits FWPB and LP, BW and LP also showing highly significant positive correlations ($p<0.001$), while LP and SI showed highly significant negative correlations ($p<0.001$). In addition, significant associations were also found between traits across categories, with the fiber quality-related trait FL showing a highly significant positive correlation ($p<0.001$) with the fiber yield-related trait LP and the fiber quality-related trait LU showing a highly significant positive correlation ($p<0.001$) with the fiber yield-related traits FWPB and LP in all six locations. These association patterns provide important clues for resolving the genetic basis of synergistic variation in traits (e.g., pleiotropic loci or genes).

Architecture of the AttGEI-Net model. In this study, based on the principle of attention mechanism and deep feature interaction, the AttGEI-Net neural network framework is designed (Fig. 2), adopting enhanced genotype-environment encoding, multi-head self-attention mechanism and bilinear interaction strategy, and proposing dynamic association modelling and explicit multiplicative interactions capture method, which optimizes the feature extraction capability, training stability and high-dimensional data processing efficiency. Specific innovations are reflected in: (1) enhancing the robustness of the feature representation by combining the Layer Normalization with a three-layer deep network; (2) using the 12-head self-

attention mechanism to achieve fine-grained feature interaction modeling; (3) designing a bilinear interaction layer to explicitly resolve genotype-environment multiplicative effects; (4) optimizing gradient flow paths using residual linkage; (5) expanding the dimension of the hidden layer to 768 to accommodate high-dimensional genotype data, which systematically improves the model's performance and interpretability in the GEI phenotype prediction task.

The overall architecture of AttGEI-Net model consists of an input layer, a feature extract layer, an attention interaction layer, a feature interaction layer, a feature fusion layer and an output layer (Fig. 2 and Methods). The AttGEI-Net model adopts the parallel input of genotype data and environmental data, which are pre-processed in the input layer and then mapped to the high-dimensional feature space and preliminarily fused through the feature extraction layer, respectively. Subsequently, the attention interaction layer captures the preliminary association features between genotype and the environment through an adaptive weight allocation mechanism and passes them to the feature interaction layer. This layer adopts a dual-branch structure, extracts feature interaction information at different layers through the bilinear interaction module and the deep interaction module respectively and uses the gating mechanism to dynamically filter effective features. Finally, the feature fusion layer integrates multi-layer feature representations to generate highly accurate phenotype prediction results (Fig. 1D).

Three advantages can be achieved through the above model design: (1) hierarchical feature extraction, achieved through enhanced encoders that progressively captures feature representations from lower to higher layers; (2) comprehensive interaction modeling, which integrates the attention mechanism (soft interaction), the bilinear layer (explicit interaction) and the deep interaction layer (higher-order nonlinear interaction) to effectively characterize the GEI effect; (3) dynamic regulation of information flow. The gating mechanism is used to filter effective features and keep the gradient stable by residual connection. Consequently, AttGEI-Net can more precisely uncover the complex association among phenotype, genotype, and environment, thereby significantly improve the accuracy of phenotype prediction.

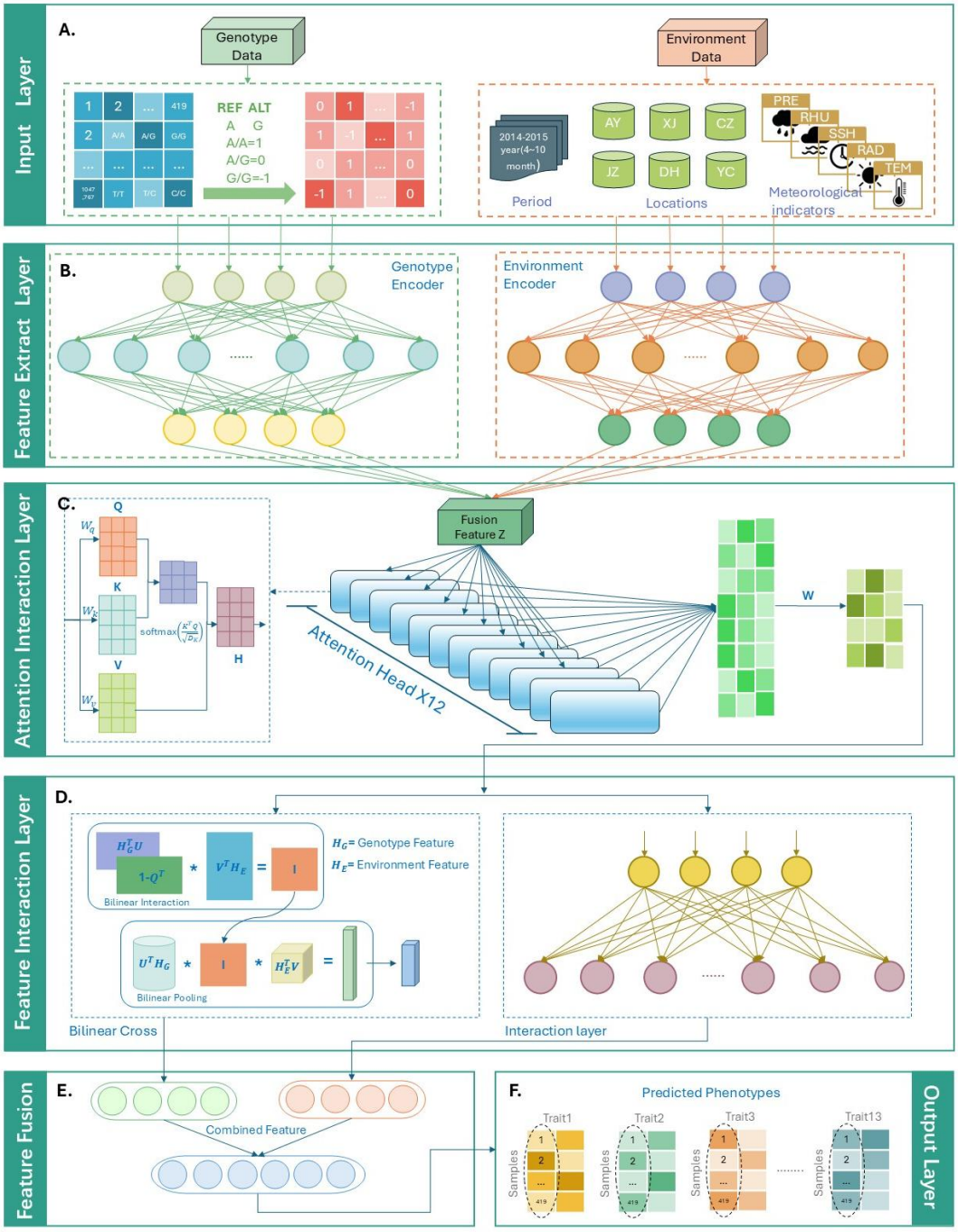


Fig. 2 Architecture of the AttGEI-Net model. (A) Input layer. Genotype data and environmental data are preprocessed in this layer before being input in parallel to the next layer. (**B) Feature extract layer.** This layer utilizes enhanced genotype encoders and environmental encoders to perform high-dimensional feature mapping on the input genotype data and environmental data, respectively. **(C) Attention interaction layer.** The layer is set up with 12 attention heads to capture initial association features between genotypes and the environment. **(D) Feature interaction layer.** This layer consists of a bilinear interaction layer and a deep interaction layer, modeling different levels of feature interaction, respectively. **(E) Feature fusion.** This layer uses a gating mechanism to dynamically regulate the

information flow, integrating multiple features from the feature interaction layer to the output layer. **(F) Output layer.** This layer outputs the final phenotype predictions for 13 traits of upland cotton.

Multi-scenario prediction performance evaluation of the AttGEI-Net model. The 419 dataset contains phenotype data of 419 genotypes for 13 traits at 6 locations in 2 years (a total of 419*6*2*13=65364 data), and the 1201 dataset contains phenotype data of 1201 genotypes for 13 traits at a single location in a single year (a total of 1201*1*1*13=15613 data). Based on the complementary characteristics of these two datasets in terms of genotype diversity and environmental diversity (the 419 dataset is rich in environmental variation but has fewer genotypes, and the 1201 dataset is diverse in genotypes but limited in environmental variation), we therefore integrated them to construct the training set and test set under three scenarios (Fig. 1C and Methods). Pearson correlation coefficients (PCC) were used as the main assessment metrics, and the results showed significant trait dependence and scenario specificity (Fig. 3, Supplementary Table 3). The overall model performance was optimal in scenario 1, where the prediction accuracies ($r > 0.9$) for traits M, SBno, SBnode, FD, GP, and BPP were significantly higher than in the other scenarios (Fig. 3, Supplementary Fig. 4).

Analyzed in terms of genotype-environment interactions (GEI), different traits showed significantly different predicted results. Genotype-dominant traits including fiber quality-related traits FL (h^2=0.94), FS (h^2=0.75), and LU (h^2=0.85), and fiber yield-related traits FWPB (h^2=0.88), BW (h^2=0.74), LP (h^2=0.96), and SI (h^2=0.93) maintained stable predictive performances in all three scenarios (Fig. 3, Supplementary Table 3), which is consistent with the high heritability trait (Supplementary Fig. 1G, Supplementary Table 6), suggesting that the phenotype variance was mainly controlled by additive genetic effects. Environment-sensitive traits including grow period-related traits (SBno, h^2 = 0.09; SBnode, h^2 = 0.32; FD, h^2 = 0.32; GP, h^2 = 0.29) as well as the fiber yield-related trait BPP (h^2 = 0.43) reached an average prediction accuracy of 0.96 in a fixed environment (Scenario 1), whereas in the new environments (Scenarios 2 and 3) average prediction accuracies of 0.39 and 0.44, a decrease of 50-60% (Fig. 3, Supplementary Table 3), showing significant environmental dependence, which is consistent with the ANOVA and heritability results (Supplementary Fig. 1F, G, Supplementary Table 6), reflecting the dependence of their phenotype expression on certain environmental signals.

Analyzed from the perspective of location specificity, different cultivation regions showed significantly different prediction results. The stable light and temperature conditions ($\geqslant$10°C cumulative temperature of 4200°C) in the Yellow River basin cotton region AY, a traditional high-quality cotton production area, enabled the full expression of the genetic potential of fiber-quality-related traits (FL, M, and LU) (Supplementary Figs. 5~6), and in particular, M in

scenario 3 showed an improved prediction accuracy in AY compared with the other locations (Supplementary Fig. 6). Saline soil conditions (pH 8.2-8.5) in the CZ of the North China Plain cotton region were correlated with the prediction performance of fiber yield-related traits, and the prediction accuracies of FWPB, BW, LP, and SI were improved at this location compared with the other locations (Supplementary Figs. 5~6).

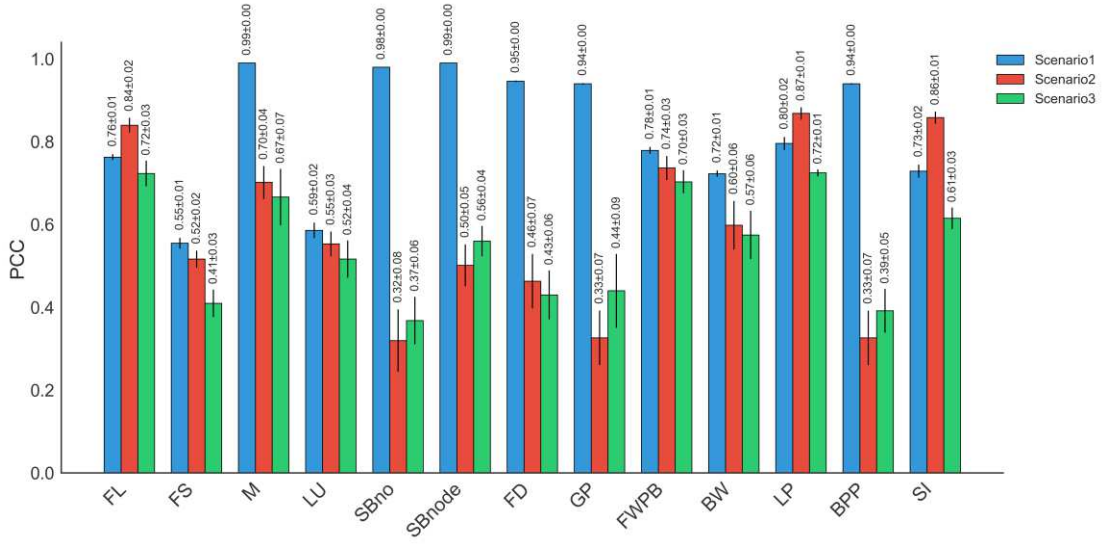


Fig. 3 Prediction accuracy of AttGEI-Net model based on test set in three scenarios. The blue columns represent scenario 1, the red columns represent scenario 2, and the green columns represent scenario 3.

Validation analysis of the AttGEI-Net model. In order to verify the generalization ability of the AttGEI-Net model to new datasets, the phenotype data of 100 new upland cotton genotypes in 5 new environments were selected as an independent validation set in this study, and the prediction performance of 13 traits was systematically evaluated, and the results showed that the prediction effectiveness of different trait categories presented significant differences (Fig. 4). Among fiber quality-related traits, FS showed the highest prediction accuracy ($r=0.54$, $p<0.001$), while M reached a significant level ($p<0.05$) although the correlation was lower ($r=0.10$). Grow period-related traits performed well overall, with GP in particular having the best prediction accuracy ($r=0.65$, $p<0.001$) and SBnode also showing moderate but highly significant correlation ($r=0.37$, $p<0.001$), which provides an important reference for fertility regulation. Among fiber yield-related traits, BPP and SI had more prominent prediction accuracy ($r=0.46$ and 0.40, $p<0.001$). Although the absolute values of the correlation coefficients for some traits such as FL ($r=0.27$) and LP ($r=0.21$) were not high, their extreme significance ($p<0.001$) indicated that these predictions were statistically significant.

In addition, by analyzing the phenotype prediction results of the 13 traits of upland cotton in the validation set versus the test set, we found that the prediction results of different traits were closely related to their degree of influence by the environment (Fig. 4, Fig. 3, Supplementary Table 3). Specifically, the fiber yield-related trait SI showed high prediction accuracy in both the validation and test sets, indicating that its phenotype was less disturbed by the environment. In contrast, the prediction results of the grow period-related trait GP and the yield-related trait BPP differed between the validation and test sets: the prediction accuracy of GP was higher in the validation set (r = 0.65) and decreased in the test set (r = 0.44); the same trend was observed for BPP (r = 0.46 in the validation set and r = 0.39 in the test set). This instability in predictive performance may stem from the high sensitivity of these traits to environmental changes, leading to inconsistency in their predictive results. The above results further confirm that genotype-environment interactions (GEI) jointly determine the stability of phenotype prediction models.

The validation results showed that although the prediction accuracy of the AttGEI-Net model decreased under the new genotypes and new environments, its prediction performance for important traits was still better, indicating that the model was able to effectively capture the key features of GEI to a certain extent, which proved that the model has practical application value in breeding practice and can provide reliable reference of the performance of cotton varieties under untested environments.

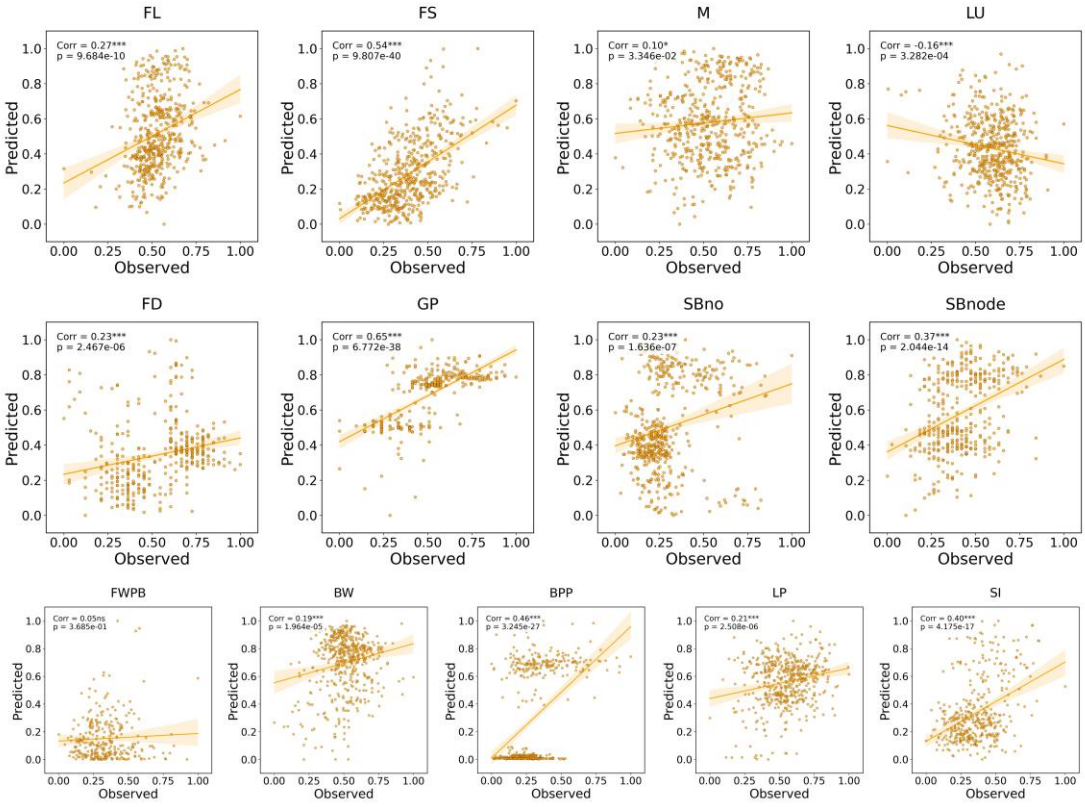


Fig. 4 Prediction result of AttGEI-Net model based on validation set. Pearson correlation analysis between predicted and observed values for 13 traits of 100 upland cotton, with Pearson correlation coefficients and significance labelled in the top left corner, *P-values<0.05, **P-values<0.01, ***P-values<0.001, ns P-values≥0.05(not significant).

The results of model testing and validation mapped the discrepancy of different traits to be affected by the environment. Next, we used the 419 upland cotton to further analyze the degree of these traits affected by the environment by adopting phenotypic plasticity analysis and environmental adaptation assessment, and to screen the broadly adapted varieties and mine elite loci through the phenotype values, phenotypic plasticity parameters, and environmental adaptation parameters to investigate how genotypes regulate the phenotypes and thus resolve the multi-level intrinsic mechanism of GEI affect phenotypes.

Phenotypic plasticity analysis. Phenotypic plasticity (PP) refers to the ability of an individual to change its phenotype in response to environmental changes [46], which is defined in this paper as the immediate adaptability of an individual to environmental changes by quantifying the magnitude of the phenotype response of the genotype in different environments. To deeply resolve the drivers of phenotype variation in upland cotton, we systematically assessed the PP of the 419 population. The results showed that all 13 traits exhibited different degrees of PP (Supplementary Fig. 7), but the pattern of variation differed significantly among traits. The phenotype values of the grow period-related traits SBno, SBnode, FD, GP, and the fiber yield-related trait BPP varied significantly among locations ($P < 0.05$) with a wide range of fluctuations, indicating that the phenotype expression of these traits was highly dependent on environmental conditions. ANOVA further confirmed the dominant contribution of environmental factors to these traits (Supplementary Fig. 1F, Supplementary Table 6), consistent with the low accuracy of scenarios 2 and 3 in the model test results (Fig. 3). Phenotype values of fiber quality-related traits FL and FS and fiber yield-related traits FWPB, LP, and SI were relatively stable among locations, but differed significantly among genotypes within the same location, suggesting that their phenotype variation was mainly controlled by genetic factors, which was consistent with the high heritability and stable prediction accuracy in the three scenarios (Supplementary Fig. 1G, Fig. 3).

Environmental adaptation assessment and broadly adapted varieties screening. We quantified the environmental adaptation (EvA) as Wricke's Ecovalence (W_i), which was originally proposed by Wricke in 1962, and it reveals the genotype-specific adaptive capacity

to the environment [47]. A low W_i value indicates a consistent response of the genotype to environmental changes. Based on this concept, EvA is defined in this paper as an adaptability formed during long-term crop environmental adaptation by integrating multi-environmental phenotype variation with environmental factors. EvA assessment based on the 419 population found (Supplementary Fig. 8) that all 13 traits showed a right-skewed distribution, reflecting the generally strong environmental adaptation of the 419 population at the single-trait level. W_i was mainly concentrated in the 0-0.4 interval, with a much narrower distribution (0-0.2) for fiber FS and LP, showing higher stability. This is in line with the test results of the model, which showed that the prediction accuracy was more stable in multiple scenarios (Figure 3). The 25% quartile was set as a threshold to screen for broadly adapted varieties (Supplementary Fig. 9), e.g., the thresholds for FL and FS were 0.07 and 0.02, respectively, and varieties with W_i lower than this value were considered as broadly adapted varieties.

To resolve the comprehensive adaptive capacity of varieties in multiple environments, we analyzed the intersection patterns of stable varieties in multiple traits. The results showed that there were a limited number of varieties that could be stable in multiple traits at the same time (Supplementary Fig. 10, Supplementary Table 7), among which F096, L090, D039, L051, B058, and D025 were from the Yellow Rive basin cotton region (G2 population), and these varieties had a clean plant shape, medium-size and thickness of leaf blades, and good ventilation and light penetration, which were common characterize that might have helped the varieties to reduce the environmental changes on phenotypes. In the 3 traits category analyses, fiber quality-related traits (FL-FS-M combinations) encompassed 9 stable varieties (Fig. 5B), grow period-related traits (SBno-FD-GP combinations) covered 10 stable varieties (Fig. 5C), and fiber yield-related traits (BW-LP-SI combinations) involved 11 stable varieties (Fig. 5A). In addition, D010, D076 and B020 in the fiber quality category, F034, F033, D018, D019, L055 and D101 in the growth period category and D002 in the fiber yield category, these varieties performed environmentally stable for all traits in this category. These results reveal that there are a limited number of varieties that can remain stable for all traits in a single category, and that even traits in the same category may be affected by different genetic regulatory networks. In addition, the relatively large number of stable varieties (11) for fiber yield-related traits may reflect the fact that fiber yield-related traits have been subjected to stronger selective pressures during the evolutionary process, resulting in the development of a more stable genetic structure.

The results of the intersection pattern of multiple environmentally unstable varieties showed (Supplementary Fig. 11, Supplementary Table 7) that combinations that were unstable across trait categories (e.g., M-SBno-BW and FL-M-FD-SI) were mainly concentrated in some genotypes, such as D063, L098, and L123, which may carry key genetic variants affecting the environmental stability of multiple traits. There were differences in the performance of the

number of varieties that were unstable on all the relevant traits in the same category (Supplementary Fig. 12), with the lowest number of yield traits (6), reflecting differences in the sensitivity of different trait categories to environmental changes. The phenotypic variation in these unstable varieties may originate from the environment-dependent expression of key regulatory genes, providing important materials for studying genotype-environment interactions (GEI).

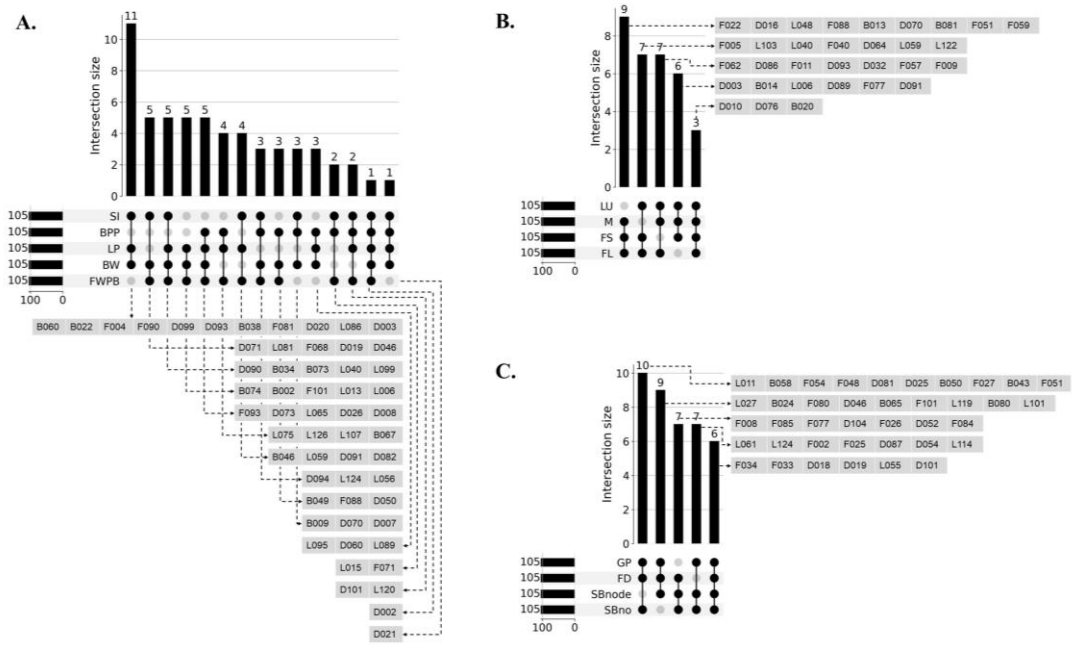


Fig. 5 Upset plot of multi-trait, multi-environmentally stable varieties of 3 trait categories of 419 upland cotton. (A) Multi-environmentally stable varieties of fiber yield-related traits. Each set of intersections is labelled with the stable variety of fiber yield-related traits. **(B) Multi-environmentally stable varieties of fiber quality-related traits.** Each set of intersections is labelled with the stable variety of fiber quality-related traits. **(C) Multi-environmentally stable varieties of growth period-related traits.** Each set of intersections is labelled with the stable variety of growth period-related traits.

Mining of Main-SNPs loci in multiple environments. Based on the fact that the 419 population has abundant environmental and phenotype variation, we carried out GWAS using phenotype values of 419 upland cotton at 6 locations in 2 years and identified a total of 2,705 Main-SNPs significantly associated with 13 traits (Supplementary Table 4), and these loci showed obvious trait category-specific distribution characteristics (Supplementary Table 8). 902 Main-SNPs associated with fiber quality-related traits (FL, FS, M, LU) mainly enriched in D11: 20.75~71.07Mb. 1,076 Main-SNPs associated with growth period-related traits (GP, FD, SBno, SBnode) mainly concentrated on D03:6.33~43.87Mb. 752 Main-SNPs associated with

fiber yield-related traits (FWPB, BW, LP, SI, BPP) mainly concentrated on A08: 23.11-78.15 Mb. This genomic distribution pattern implied that different trait categories were controlled by independent genetic regulatory networks. In addition, cross-trait category association analysis identified 25 Main-SNPs with pleiotropic effects (Supplementary Table 9), of which 24 Main-SNPs regulated both fiber yield and quality traits, mainly in A07: 88.74-90.59Mb.

Environment-specific Main-SNP analyses showed that the AY region exhibited a unique genetic architecture for grow period-related traits (Supplementary Fig. 13, Supplementary Table 10), and a total of 116 SBnode, 928 FD, and 503 GP-specific Main-SNPs were identified to exhibit dense distributions on chromosome D03, which were concentrated in 6.47-41.12 Mb, 9.61-43.62 Mb, and 6.47~41.57Mb, respectively, where the SBnode and GP association intervals were highly overlapping, while the number of Main-SNPs in FD was significantly higher than that of other traits, which might be regulated by a more complex genetic network. In addition, the climatic characteristics of AY (e.g. seasonal temperature variation and light duration) may have exerted stronger selection pressure on cotton growth, leading to adaptive divergence in allele frequencies of growth period-related genes on chromosome D03. We identified 270 FS-specific Main-SNPs in the CZ region (Supplemental Fig. 14, Supplemental Table 10), which were mainly concentrated in D11: 24.51-24.72 Mb, and 306 LP-specific Main-SNPs, which were concentrated in A08: 23.14-78.15 Mb. A total of 153 M- and 147 FWPB-specific Main-SNPs were identified in the YC region and were distributed in D11:28.95~30.78Mb and D11:28.25~50.47Mb, respectively (Supplementary Fig. 15, Supplementary Table 10), and there was an overlap region of 12.22Mb between them, indicating the presence of pleiotropic regulatory elements in this region. These results suggest that traits are regulated in specific environments by main effector loci and related genes in specific regions of the chromosome.

Through multi-environmental stable loci analysis (Supplementary Figure 16, Supplementary Table 10), we detected 192 and 98 multi-environmental shared Main-SNPs associated with FL and FS, enriched in D11: 24.51-24.72 Mb and A07: 90.33-90.59 Mb, respectively, which suggests that these loci may be environmental stability loci, explaining why these two traits showed strong cross-environmental consistency in model predictions (Figure 3) and both CVs were below 10% (Supplementary Table 2). Further analyses revealed that the Main-SNPs associated with M, SBno, GP, FWPB, and BPP appeared only at individual locations (Supplementary Fig. 16), and these loci may have environment-specific expression, explaining why these traits showed strong cross-environmental fluctuations in model predictions (Fig. 3).

Mining of elite loci under multiple parameters. Based on the GWAS of 419 population under

three parameters (phenotype values, phenotypic plasticity parameters, environmental parameters), we identified a total of 10,215 significant SNPs, including 2,705 Main-SNPs, 41 phenotypic plasticity loci (PP-SNPs) and 9,022 environmental adaptation loci (EvA-SNPs) (Table 1). Notably, EvA-SNPs accounted for 88.3% (9,022/10,215) of the significant loci and contained 7,495 novel loci not detected by other parameters (Fig. 6A), which was significantly higher than the proportion of newly discovered loci for PP-SNPs (15/41). This result suggests that environmental adaptation parameters have unique advantages to reveal potential genetic variation and are able to detect environmental adaptation loci that are difficultly detected by traditional phenotypic analyses. In addition, most of the associated SNPs across trait categories were EvA-SNPs (Supplementary Table 11), e.g. 83.5% of the SNPs associating both fiber quality and yield traits were specific EvA-SNPs, which suggests that pleiotropic regulation mainly occurs at environmental adaptation loci, suggesting that synergistic variation among traits may be achieved through environmentally responsive pathways.

Among the 192 and 98 multi-environmental shared Main-SNPs associated with FL and FS (Supplementary Table 10), the 192 Main-SNPs associated with FL did not intersect with EvA-SNPs, while 74.49% of the 98 Main-SNPs associated with FS intersected with EvA-SNPs. These loci have important molecular breeding value as they are associated with both phenotypic expression and environmental adaptation and can be used for genome-wide selection and marker selection of broadly adapted varieties.

Table 1. The numbers of SNPs associated with 13 traits of 419 upland cotton using three parameters for GWAS. Main-SNPs represent the significant SNPs using phenotype values for GWAS. PP-SNPs represent the significant SNPs using phenotype plasticity parameters for GWAS. EvA-SNPs represent the significant SNPs using environmental adaptation parameters for GWAS.

Trait categories	Traits	Numbers	Main-SNPs	PP-SNPs	EvA-SNPs	Top 3 chromosome
Fiber quality related-traits	FL	1,824	492	0	1,332	1,018/D06, 463/D11, 106/A02
	FS	1,416	467	0	1,070	293/A07, 198/D03, 173/D02
	M	654	167	0	501	178/D11, 124/D05, 68/D13
	LU	356	68	1	289	113/A04, 58/D04, 57/D05
		3,933				

Growth period related-traits	SBno	2,667	788	11	0	777	744/A11, 8/A05, 8/D02
	SBnode		1,136	121	18	1,117	34/D11, 22/D09, 18/D10,
	FD		1,220	1,062	0	961	1,066/D03, 40/D10, 18/A11
Fiber yield related-traits	GP	3,910	1,626	505	3	1,620	130/A11, 110/D11, 56/A02
	FWPB		1,164	202	0	974	486/A02, 138/D11, 69/A08
	BW		916	1	1	914	637/D10, 83/A09, 48/A08
	LP		1,804	516	0	1,518	534/A08, 477/D10, 246/A02
	BPP		132	16	18	106	42/A05, 20/A12, 19/D01
	SI		292	57	0	235	55/D12, 52/D05, 42/A07
Summary		10,215		2,705	41	9,022	

By analyzing the significant SNPs shared by the three parameters (Figure 6A, Supplementary Table 12), we found that: (1) the 18 SNPs jointly detected by the three parameters were all related to FD, and centrally located in D03: 20.08-39.50 Mb, which was consistent with existing studies, further confirming the central position of this region in the regulation of cotton flowering, and may contain flowering-regulated genes so that they play a stable genetic regulatory role under different environmental conditions; (2) the 8 SNPs shared by the phenotype value and phenotypic plasticity parameters were all located in the narrow interval of D01: 2.09-2.13 Mb and were significantly associated with BPP; (3) among the 1,509 SNPs shared by phenotype values and environmental adaptability parameters, 1035 (68.6%) were associated with FD (D03:6.41-43.62Mb), 273 (18.1%) were associated with LP (A08:23.11-78.15 Mb). The combined analyses showed that although cotton phenotypes are

mainly controlled by genetic factors, environmental factors have important effects on trait expression through the mechanisms of phenotypic plasticity and environmental adaptation. The 8 BPP-associated loci shared by phenotypic plasticity parameters and phenotype values may contain key “plasticity hub genes”, while the 1509 SNPs (mostly FD-associated loci) shared by phenotype values and environmental adaptation parameters constitute the main genetic basis for environmental adaptation in cotton.

Further analysis of the specific loci identified by the phenotypic plasticity parameters and the environmental adaptability parameters revealed that among the 15 specific PP-SNPs, 10 loci (66.7%) were concentrated in D01: 2.09-4.37 Mb, all of which were significantly associated with the BPP (Fig. 6B, Supplemental Fig. 17, Supplemental Table 12), suggesting that this region may play an important role in regulating cotton yield plasticity. In contrast, 7,495 specific EvA-SNPs were detected by environmental adaptation parameters (Fig. 6 C~F, Supplemental Fig. 18~20, Supplemental Table 12). The specific EvA-SNPs associated with fiber quality-related traits were mainly enriched in D06: 0.40-8.52 Mb (1,018 FL-associated specific EvA-SNPs), the specific EvA-SNPs associated with grow period-related traits were concentrated in A11:0.47-93.25Mb (744 SBno-associated specific EvA-SNPs), while the specific EvA-SNPs associated with fiber yield-related traits were mainly located in chromosomes D10 and A02, where BW and FWPB-associated specific EvA-SNPs were concentrated in D10:3.66-47.65Mb and A02:9.80-10.77 Mb, respectively.

The comprehensive analysis showed that although the number of specific loci varied significantly (only 15 specific PP-SNPs vs 7,495 specific EvA-SNPs), the sparse number of specific PP-SNPs may coordinate the phenotype response in multiple environments through the regulation of a limited number of “plasticity hub genes”. In contrast, the discovery of a large number of specific EvA-SNPs (e.g., in the D06 fiber quality region and the A11 grow period region) revealed important environmental response modules missed by the traditional phenotypic-value-based GWAS (which only detected the D11 fiber quality region and the D03 grow period region), and these loci provide a rich resource of candidate genes for breeding across different ecological zones, which demonstrates that the environmental adaptation parameters can effectively expand the mining scope of important trait-associated loci. Therefore, modern crop breeding requires the integration of Main-SNPs, PP-SNPs and EvA-SNPs to maximize genetic gain.

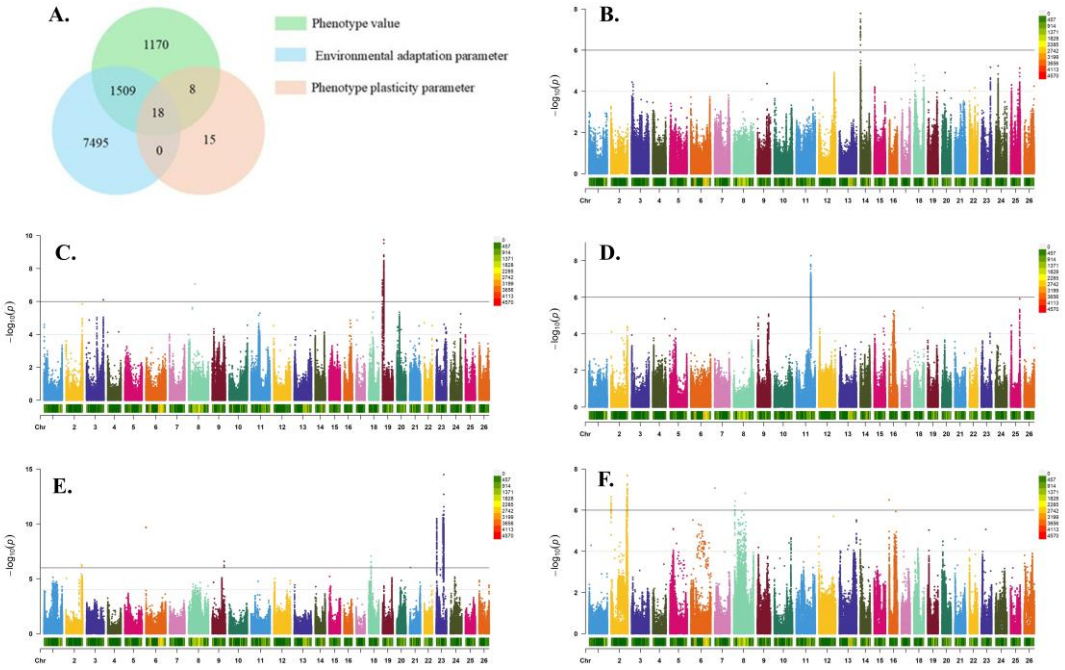


Fig. 6 The specific and shared SNPs associated with traits of 419 upland cotton from GWAS based on three parameters. (A) Venn diagram of significant SNPs loci from GWAS based on three parameters. Green represents significant SNPs loci from GWAS based on phenotype values. Blue represents significant SNPs loci from GWAS based on environmental adaptation parameters. Orange represents significant SNPs loci from GWAS based on phenotypic plasticity parameters. **(B) Manhattan plots related to BPP from GWAS based on phenotype plasticity parameters. (C) Manhattan plots related to FL from GWAS based on environmental adaptation parameters of 2015. (D) Manhattan plots related to SBno from GWAS based on environmental adaptation parameters of YC. (E) Manhattan plots related to BW from GWAS based on environmental adaptation parameters of 2015. (F) Manhattan plots related to FWPB from GWAS based on environmental adaptation parameters of 2014 and 2015.** The genome-wide significant P-value threshold (10^{-6}) is indicated by the black line. BPP: bolls per plant, FL: fiber length, SBno: sympodial branch number, BW: boll weight, FWPB: fiber weight per boll, YC: Yancheng.

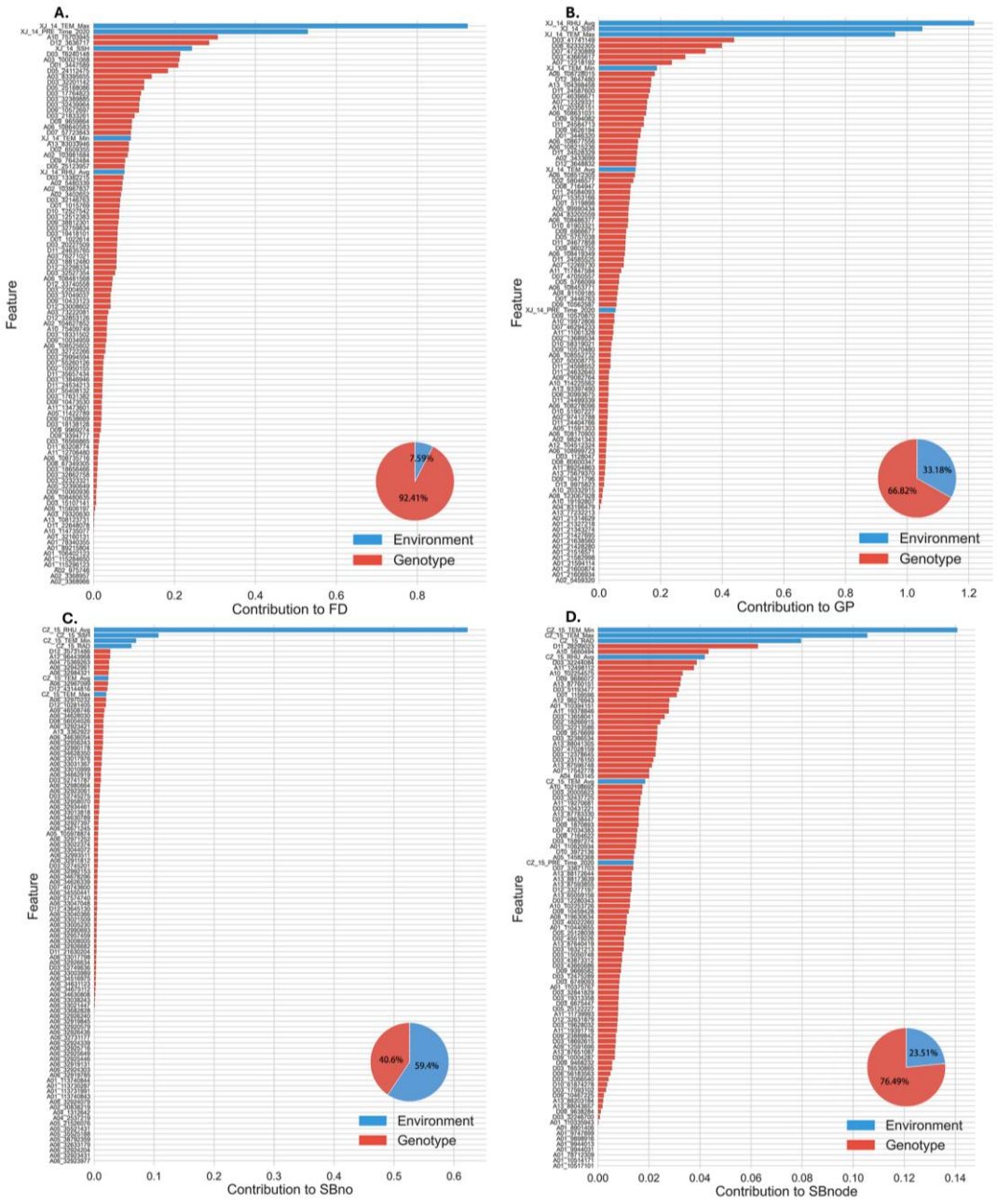
Hierarchical networks of genotype-environment interactions regulating phenotype variation in cotton. This study revealed a complex hierarchical network of genotype-environment interactions (GEI) regulating cotton phenotype variation. The genetic basis of upland cotton phenotypes showed parameter specificity, and the sparse number of PP-SNPs (41) indicated that phenotypic plasticity was regulated by a small number of main effector genes, whereas the large number of specific EvA-SNPs (7,495) reflected the importance of

environmental adaptive analyses in mining new genetic loci, exposing specific loci not found in previous studies [15,18,47]. The EvA-SNPs not only accounted for the absolute dominance of significant loci in terms of number (88.3%), but also showed unique pleiotropic regulatory characteristics. This multi-parameter regulatory network has obvious hierarchical features: the basic genetic architecture (Main-SNPs) maintains the basic expression of traits, the PP-SNPs mediate the immediate response of phenotypes to environmental changes, and the EvA-SNPs constitute the highest-level adaptive regulatory network that coordinates the expression of multiple traits by integrating environmental signals. These findings profoundly reveal that cotton phenotype variation is the result of long-term interactions between genotypes and the environment, in which EvA-SNPs play an irreplaceable role in shaping environmentally adapted phenotypes as an important loci for environmental response.

Interpretability analysis of the AttGEI-Net model. The result of interpretable analysis of the AttGEI-Net model showed (Fig. 7, Supplementary Figs. 21-22) that although the contribution of genetic features was higher than that of environmental features for the most traits, different trait categories showed significant regulatory differences, among which the overall contribution of environmental features for three growth period-related traits, GP, SBno, SBnode, and one fiber yield-related trait BPP, accounted for 33.18%, 59.4%, 23.51% and 30.67%, respectively, showing a strong environmental dependence, which was consistent with the prediction results of the model (Fig. 3). Further analysis of the important environmental features for influencing growth period-related traits revealed (Fig. 7, Supplementary Table 5) that GP was mainly influenced by humidity conditions (29.83-60.3%), sunshine hours (8.31-10.71 h), and daily maximum temperature (21.85-33.54°C) at XJ in 2014, and that these environmental factors might influence photosynthetic efficiency and metabolic rate to regulate growth period length. In addition, the environmental conditions of CZ in 2015 showed significant effects on fruiting branch development (SBno, SBnode), where moderate relative humidity (53.71-72.22%), sufficient solar radiation (14.35-21.68 MJ/m2day), hours of sunshine (5.33-9.42h), and the diurnal temperature difference (daily minimum temperature 9.28-22.44°C and daily maximum temperature 21.03-32.65°C) together constitute the optimal environmental combination to promote fruiting branch differentiation. Especially importantly, the formation of BPP, a key yield trait, significantly depended on a similar combination of environmental parameters of CZ in 2015, suggesting that the synergistic effect of temperature and light water resources is decisive for cotton reproductive development.

Notably, we identified shared loci of interpretable analyses and GWAS in all trait categories (Supplementary Table 15), a finding with multiple important implications. First, interpretability analysis, a popular deep learning method, and the agreement of its identification

results with traditional GWAS fully validate the reliability and biological reasonability of the AttGEI-Net model, suggesting that the two independent methods reveal the real genetic regulatory mechanisms from different perspectives. Secondly, among these shared loci, 14 FS-associated shared loci were enriched on chromosomes A07, D09 and D03, 8 FD-associated shared loci were clustered on chromosome D03, and 8 LP-associated shared loci were distributed on chromosome A08, and most of these shared loci were EvA-SNPs. This phenomenon suggests that the repeated validation of association loci for environmental adaptation as a characterization of GEI not only indicates the stable contribution of environmental adaptation loci to phenotype variation but also suggests that they may be the key genetic basis for cotton adaptation to different environments. These findings confirm the effectiveness of deep learning models in genetic analysis and provide an important reference for screening environmental adaptation loci in molecular design breeding.



580

581 **Fig. 7 Feature importance ranking of growth period-related traits calculated using the**
582 **interpretability analysis of the AttGEI-Net model.** Horizontal coordinates represent the contribution
583 value of the features to traits, vertical coordinates list the names of the top 100 important features, blue
584 represents the environmental features (PRE, RHU, RAD, SSH, TEM), red represents the genetic
585 features (SNPs loci), and the pie chart in the bottom right corner indicates the percentage of importance
586 of the two features. PRE: Precipitation, RHU: Relative humidity, RAD: Radiation, SSH: Sunshine,
587 TEM: Temperature. **(A~D) Interpretability analysis of AttGEI-Net models for FD, GP, SBno,**
588 **SBnode.** FD: flowering date, GP: growth period, SBno: sympodial branch number, SBnode: the first
589 Sympodial branch node.

Discussion

In this study, an AttGEI-Net deep learning model was constructed based on the attention mechanism to systematically analyze the genotype-environment interactions (GEI) in cotton phenotype formation. It was found that the model showed significant differences in three prediction scenarios, with an average prediction accuracy of 0.96 for phenotypes in fixed environments, whereas the decrease in prediction results in new environments revealed the complexity of GEI affecting phenotypes. From the perspective of genetic mechanism, high heritability traits (e.g. FL, LP, SI, etc.) maintained stable prediction performance in all three scenarios, indicating that their phenotype variation was mainly dominated by additive genetic effects, while low heritability traits (e.g. SBno, GP, BPP, etc.) showed a decrease in prediction accuracy by 50-60% in response to the environmental changes, reflecting the key regulatory role of environmental signals on their phenotype expression. In addition, location-specific analyses indicated that environmental factors such as temperature gradient and soil salinity may affect prediction accuracy through specific gene pathways. The inconsistency between the prediction results of the test set and the validation set mainly stemmed from the differences in trait responses to the environment. Based on the cross-environmental phenotypic coefficient of variation (CV) analysis (Supplementary Table 2), we established a system to assess the reliability of the model prediction results: (1) high-variance traits (CV>20%, such as SBno, BPP, etc.) were significantly affected by environmental fluctuations, and the prediction results were only used as a reference; (2) moderately-variant traits (CV 10-20%, e.g., GP, BW, etc.) have relatively stable genetic expression patterns and have higher prediction confidence; (3) low variance traits (CV<10%, e.g., FL, FS, etc.) are dominated by genetic factors and have optimal prediction results. Therefore, for low-variance traits, model predictions can directly guide seed selection; while for high-variance traits, a comprehensive assessment is recommended in combination with multi-environmental test data.

Based on the GWAS of multi-environmental phenotype values, this study revealed the genetic basis of multi-environmental variation regulating important traits in cotton. The results revealed that different trait categories have specific genomic distribution patterns, with fiber quality-associated loci enriched on chromosome D11, grow period-associated loci concentrated on chromosome D03, and fiber yield-associated loci mainly distributed on chromosome A08. In addition, there were a large number of cross-environmentally shared loci associated with FL and FS (192 and 98, respectively), of which 74.49% of the 98 loci associated with FS were associated with both phenotype expression and environmental adaptations. Whereas the loci associated with growth period-related traits (SBno, GP) and fiber yield-related traits (FWPB, BPP) were environmentally specific, which reflects that traits are regulated by main effect loci and related genes in specific intervals of the chromosome under specific environments,

revealing the importance of environmental factors in shaping the genetic architecture. For example, saline conditions in the CZ may have screened key regulatory loci associated with FS on chromosome D11, and the seasonal climate in AY may have driven directional selection of allele frequencies for growth period-related traits (FD, GP) on chromosome D03.

By integrating interpretability analyses of model with GWAS based on phenotype values, phenotypic plasticity parameters, and environmental adaptation parameters, this study systematically elucidated the hierarchical regulatory network of phenotype variation in cotton. A total of 9,022 environmental adaptation loci (EvA-SNPs) were identified by multiparameter GWAS, accounting for 88.3% of the total, significantly expanding the detection range of traditional analysis, while only 41 phenotypic plasticity loci (PP-SNPs) were identified. The sparse number of specific PP-SNPs (15) may coordinate phenotype responses in multiple environments by regulating a limited number of "plasticity hub genes". In contrast, the discovery of a large number of specific EvA-SNPs (7,495) revealed important environmental response modules missed by traditional GWAS based on phenotype values. Such differential results suggest a clear hierarchical characteristic of GEI. In addition, the high contribution loci screened by interpretability analysis was consistent with the GWAS results and most of them were EvA-SNPs. This validation not only confirmed the reliability of the AttGEI-Net deep learning model but also provided key targets to analyze the environmental adaptation mechanism of cotton traits, which is of great value for the realization of precision molecular design breeding.

Conclusion

In this study, we resolved the regulatory mechanism of GEI on phenotype variation in upland cotton based on a deep learning and revealed the important role of multi-hierarchical genetic networks in environmental adaptation. The results show that the AttGEI-Net model has an average prediction accuracy of 0.96 for phenotypes in fixed environments and shows good prediction ability for important traits in the validation set. It indicates that the AttGEI-Net model is not only applicable in multiple scenarios but also analyzes the intrinsic mechanism of GEI through the differences in prediction performance, which provides an effective tool for cotton phenotype prediction and precision breeding.

A total of 10,215 significant SNP loci were identified by multiparameter GWAS, including 2,705 Main-SNPs, 41 PP-SNPs, and 9,022 EvA-SNPs, which indicates that environmental adaptation parameters can effectively detect genetic variation that is difficult to be detected by traditional methods, and reveals that crops respond to environmental changes through a system of "basic genetic architecture - plasticity regulation - environmental adaptation network". This

regulatory system has obvious hierarchical characteristics: the basic genetic architecture (Main-SNPs) maintains the basic expression of traits, the phenotypic plasticity loci (PP-SNPs) mediate the immediate response of phenotypes to environmental changes, and the environmental adaptation loci (EvA-SNPs) constitute the highest-level adaptive regulatory network, which coordinates the expression of multiple traits by integrating environmental signals. In addition, interpretability analyses further validated the reliability of the AttGEI-Net model and provided key targets for resolving the environmental adaptation mechanisms of cotton traits. The broadly adapted varieties F096, L090, D039, L051, B058, and D025 remained stable in multiple environments for multiple traits, and their common characterizes such as clear plant shape, medium leaf size and thickness, and good ventilation and lightness may help the varieties to reduce the impact of the environment.

The three-level regulatory network proposed in this study provides a new perspective on the GEI mechanism of complex traits, and the "prediction model-genetic analysis-mechanism elucidation" paradigm established in this study is of great value in guiding the molecular design and breeding of crops. Nevertheless, there are still limitations in this study on the complex regulatory network of phenotype-genotype-environment. In the future, we aim to integrate multi-omics data and construct multi-dimensional interactions network to establish the whole chain model of phenotype-genotype-environment. This approach will enable the analysis of the entire process from molecular mechanism to field performance through multi-environmental field trials. Such advancements are expected to offer a more comprehensive theoretical basis and technological support for the precision design and breeding of crops.

Methods

Dataset and data processing. Among 13 traits of 3 trait categories, fiber quality-related traits included fiber length (FL, mm), fiber length (FS, cN/tex), micronaire (M), and length uniformity (LU, %). Grow period-related traits included sympodial branch number (SBno), the first Sympodial branch node (SBnode), flowering date (FD, d), and growth period (GP, d). Fiber yield-related traits included fiber weight per boll (FWPB, g), boll weight (BW, g), lint percentage (LP, %), bolls per plant (BPP), and seed index (SI, %) (Supplementary Table 1). Phenotypic missing values were interpolated using the knn algorithm [38] (Fig. 1B).

2356,312, 1047,767, and 2356,321 high-quality SNPs were screened after quality control (MAF $\geqslant$ 0.05, missing rate $\leqslant$ 0.1) of the genotype data in three datasets, 1201, 419, and 100, respectively. Referring to Ma, Wenlong et al.'s study [39], the 0/0, 0/1, 1/1, missing of genotype data were coded 1, 0, -1, and 2, respectively, where 1, 0, and -1 represent pure no mutation, heterozygous mutation, and pure mutation, respectively. Due to the high

dimensionality of the genotype data and the large amount of data, the model training and testing was slow, so we used p-values (p = 0.1, 0.01, 0.001, 0.0001, and 1e-05) to test the accuracy of the model[40], and chose the threshold p = 0.001 with the best model testing effect to screen the genotype data as the final input for model testing and validation (Supplementary Fig. 3).

The environmental data are obtained from the Resource and environmental science data platform (RESP), a dataset that records latitude (°), longitude (°), altitude (m), and five meteorological indicators precipitation (PRE, mm), relative humidity (RHU, %), sunshine hours (SSH, h), radiation (RAD, MJ/m2day), temperature (TEM, ℃) day-by-day data of all planting locations and years. To improve the efficiency of the model in utilizing the environmental features, we enhanced the environmental features by retaining the original features, adding statistical and polynomial features, and calculating the sliding average features, which improved the representativeness of the environmental data and enabled the model to better capture the effects of environmental factors on crop traits (Fig. 1B).

Scenarios split. To systematically evaluate the prediction performance of the model in different application scenarios, we designed three test scenarios (Fig. 1C): scenario 1 was used to evaluate the prediction ability of the model for untrained genotypes in the trained environment, scenario 2 was used to evaluate the prediction ability of the model for trained genotypes in the untrained environment, and scenario 3 was used to evaluate the prediction ability of the model for untrained genotypes in the untrained environment.

Architecture of the AttGEI-Net model. The feature extract layer consists of an enhanced genotype encoder and an environment encoder, which perform high-dimensional feature mapping on the input genotype data (g∈ R^{d_g}) and environment data (e∈ R^{d_e}), respectively (Fig. 2B). Both encoders use a three-layer fully connected (FC) neural network, where the hidden layer dimension is extended to 768 to enhance the expression ability of the model. To avoid data dependency during small batch training, Layer Normalization (LayerNorm) is used instead of the traditional BatchNorm to enhance the training stability. LeakyReLU (α=0.1) is chosen for the activation function to enhance the nonlinear modelling ability while mitigating the gradient vanishing. Eventually, genotype and environmental features are converted to high-dimensional representations, respectively.

The attention interaction layer uses the multi-head self-attention (MSA) mechanism, which sets up 12 attention heads with inputs as a splice of genotype feature (h_g) and environmental features (h_e) (h=[h_g; h_g] ∈ R^{d_h}) (Fig. 2C). Each attention head computes a weighted interaction of query (Q), key (K) and value (V):

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

Weighted summation of V values for each attention head based on attention scores:

$$head_i = AttentionScore_i V_i \quad (2)$$

Splicing the output of all the attention heads gives the final multi-head attention output:

$$MultiHead(Q,K,V) = Concat(head_1, \dots, head_h)W^O \quad (3)$$

Where Q, K and V denote the Query, Key and Value matrices respectively, d_k denotes the dimensionality of the key vector, and W^O denotes the weight matrix.

The feature interaction layer consists of a bilinear interaction layer and a deep interaction layer, modelling different levels of feature interaction respectively (Fig. 2D). The bilinear interaction layer explicitly captures the GEI effect, calculated as follows:

$$h_{bilinear} = h_g W h_e^T + b, W \in R^{d_g \times d_e} \quad (4)$$

W is the learnable weight matrix and b is the bias term, which computes the second-order interaction between genotype and environmental feature.

The deep interaction layer uses a two-layer fully connected network with input spliced attention features (h_{attn}) to learn higher-order feature combinations:

$$h_{deep} = FC\left(LeakyReLU\left(FC(h_{attn})\right)\right) \quad (5)$$

FC is fully connected layer and LeakyReLU is used to enhance nonlinear modelling capabilities.

The feature fusion layer integrates multiple features ($h_{bilinear}$和h_{deep}) from the feature interaction layer and dynamically regulates the information flow using a gating mechanism (Fig. 2E). The gating weights ($g \in [0,1]$) are computed by a Sigmoid function:

$$g = \sigma\left(W_g\left[h_{bilinear}; h_{deep}\right] + b_g\right) \quad (6)$$

where W_g and b_g is the gating weights and bias, respectively and σ is the Sigmoid activation function for generating feature selection weights.

Residual linking of filtered features to the original input to mitigate the problem of vanishing gradients:

$$h_{fused} = g \cdot h_{bilinear} + (1-g) \cdot h_{deep} + h_{skip} \quad (7)$$

where h_{skip} is a jump-connected input feature for stable training. Finally, the fused features are mapped to the predicted output via the fully connected layer (Figure 2F) to achieve highly accurate phenotype prediction.

Interpretability analysis of the AttGEI-Net model. This study conducts an interpretability analysis of model based on the Shapley Value to quantify the contribution of different features to the prediction of complex traits [41]. Specifically, given the input sample x and the model prediction f(x), the Shapley Value $\emptyset_i$ for each feature i can be expressed as [42]:

$$\emptyset_i(f,x) = \sum_{S \in N \setminus \{i\}} \frac{|S|!\,(|N| - |S| - 1)!}{|N|!} \big(f(S \cup \{i\}) - f(S)\big) \tag{8}$$

where N is the set of all features, S is the subset without feature i, and $f(S)$ denotes the predicted value of the model based only on the subset S.

Operating environment. The experimental environments in this study were configured as follows: a Linux operating system (Ubuntu 20.04 LTS or CentOS 7) was used for the training phase, while the test environment was compatible with Windows 10 and Linux systems. The experiments were built based on Python 3.8 and above and utilized CUDA 11.7 with cuDNN 8.4.1 for GPU-accelerated computation. The main deep learning framework relied on was PyTorch 1.13.1 (CUDA version 11.7), while the following scientific computing and visualization toolchains were integrated: NumPy (⩾1.23.5) for efficient numerical operations, Pandas (⩾1.5.3) for data preprocessing, scikit-learn (⩾1.2.2) to provide machine learning model evaluation and data processing algorithms, and matplotlib (⩾3.7.1) and seaborn (⩾ 0.12.2) for data visualization. The combination of this software stack ensures efficient model training and reliable experimental results.

Variance component analysis and heritability. The formula for the complete variance component analysis in this study is as follows [43]. Where G is a random variable representing 419 genotypes, E is a random variable representing 12 environments, and G × E is the interaction term between genotypes and environments. The AMMI model of the Genstat [44] program was used for the ANOVA, and the linear mixed model of the Genstat program, REML, was used for the variance component estimation.

$$Variance = G + E + G \times E \tag{9}$$

The generalized heritability of the traits was calculated as follows. V_G represents genetic variance, V_E represents environmental variance, and $V_{G\times E}$ represents genotype-environment interaction variance.

$$h^2 = \frac{V_G}{V_G + {V_E}/{12} + {V_{G\times E}}/{12}} \tag{10}$$

GWAS and screening of broadly adapted varieties. This study used the phenotype values, phenotypic plasticity parameters, and environmental adaptation parameters of the 419 population as GWAS inputs, respectively (Supplementary Tables 13-14), and performed GWAS using GEMMA under Linux [45], setting p-value = 1e-6 as the threshold to screen three types of significant loci Main-SNPs, PP-SNPs, and EvA-SNPs.

Of the above three inputs, the phenotype value was the mean value of 419 upland cotton at 6 locations in 2 years, and phenotypic plasticity parameters were quantified as the slope (β_1) of the regression analysis of phenotype values against environmental indices with the following formula:

$$Phenotype = \beta_0 + \beta_1 \times PC1 + \epsilon \tag{11}$$

Where β_0 is the interception, β_1 is the slope, PC1 is the environmental index, and ϵ is the error term. Calculation of environmental indices: the average values of 5 meteorological indicators (PRE, RHU, SSH, RAD, TEM) of 12 environments from April to October were calculated respectively as the environmental variables under the environment, and principal component analysis was performed on the environmental variables using R language, and the PC1 with the highest contribution rate was selected as the environmental index.

We used Genstat to calculate the environmental adaptation parameters (W_i) of 419 upland cotton in 2014, 2015, 2014 and 2105, Anyang (AY) in Henan Province, Cangzhou (CZ) in Hebei Province, Yancheng (YC) in Jiangsu Province, Jingzhou (JZ) in Hubei Province, Dunhuang (DH) in Gansu Province, Alaer in the Xinjiang (XJ) autonomous region, respectively, which were used as the input for GWAS, and normalized the W_i to a range of 0 to 1 range, then set the 25% quartile as a threshold to screen for varieties with high environmental adaptability and the 75% quartile as a threshold to screen for varieties with poor environmental adaptability. The W_i calculation formula was as follows:

$$W_i = \sum_{j=1}^{n} \left(Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..}\right)^2 \tag{12}$$

where Y_{ij} represents the phenotype value of genotype i in environment j, $\bar{Y}_{i.}$ represents the average phenotype value of genotype i across all environments, $\bar{Y}_{.j}$ represents the average phenotype value of all genotypes in environment j, and $\bar{Y}_{..}$ represents the overall average phenotype value.

Statistical analysis. The formulas for the phenotype standard deviation and coefficient of variation for this study are shown below:

$$SD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2} \tag{13}$$

$$CV(\%) = \left(\frac{SD}{\bar{x}}\right) \times 100 \tag{14}$$

x_i: phenotype value of the i variety, $\bar{x}$: average phenotype value of the trait across all varieties, N: number of varieties.

t-distribution of significance tests (obeying a degree of freedom of n-2):

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \tag{15}$$

r: Pearson's correlation coefficient, n: sample size.

Data availability

All the genomic sequence datasets for the experiment of model and GWAS have been deposited in the NCBI Sequence Read Archive under accession numbers PRJNA399050, and PRJNA605345 and NGDC under PRJCA010331. Other relevant data is available from the corresponding author on request.

Code availability

The code of our model has been made available at https://github.com/hezikang-git/AttGEI-Net, which can be used to evaluate the performance of the model and carry out interpretable analysis.

References

[1] Zhang, Tianzhen, et al. "Sequencing of allotetraploid cotton (Gossypium hirsutum L. acc. TM-1) provides a resource for fiber improvement." *Nature biotechnology* 33.5 (2015): 531-537.

[2] Lesk, Corey, Pedram Rowhani, and Navin Ramankutty. "Influence of extreme weather disasters on global crop production." *Nature* 529.7584 (2016): 84-87.

[3] Cornelius, P. L., J. Crossa, and M. S. Seyedsadr. "Statistical tests and estimators of multiplicative models for genotype-by-environment interaction." *Genotype-by-environment interaction. CRC Press, Boca Raton, FL* (1996): 199-234.

[4] Cooper, Mark, and Graeme L. Hammer, eds. Plant adaptation and crop improvement. IRRI, 1996.

[5] Macholdt, Janna, et al. "Long-term analysis from a cropping system perspective: yield stability, environmental adaptability, and production risk of winter barley." *European Journal of Agronomy* 117 (2020): 126056.

[6] Zhang, He, Jens D. Berger, and Chris Herrmann. "Yield stability and adaptability of canola (Brassica napus L.) in multiple environment trials." Euphytica 213.7 (2017): 155.

[7] Egea-Gilabert, Catalina, Mario A. Pagnotta, and Pasquale Tripodi. Genotype× environment interactions in crop breeding. *Agronomy* 11.8 (2021): 1644.

[8] Fusco, Giuseppe, and Alessandro Minelli. "Phenotypic plasticity in development and evolution: facts and concepts." *Philosophical Transactions of the Royal Society B: Biological Sciences* 365.1540 (2010): 547-556.

[9] Huang, Yuefan, et al. "Genetic interrogation of phenotypic plasticity informs genome-enabled breeding in cotton." *Journal of Genetics and Genomics* 50.12 (2023): 971-982.

[10] Gauch Jr, Hugh G., and Richard W. Zobel. "Identifying mega‐environments and targeting genotypes." *Crop science* 37.2 (1997): 311-326.

[11] de Figueiredo, André Gradowski, et al. "Application of mixed models for evaluating stability and adaptability of maize using unbalanced data." *Euphytica* 202.3 (2015): 393-409.

[12] Li, J, et al. "Identifying loci with breeding potential across temperate and tropical adaptation via EigenGWAS and EnvGWAS." *Molecular ecology* 28.15 (2019): 3544-3560.

[13] Wang, J, et al. "A major locus controls local adaptation and adaptive life history variation in a perennial plant." *Genome biology* 19.1 (2018): 72.

[14] Lasky, Jesse R., et al. "Genome-environment associations in sorghum landraces predict adaptive traits." *Science advances* 1.6 (2015): e1400218.

[15] Dai, Panhong, et al. "Extensive haplotypes are associated with population differentiation and environmental adaptability in Upland cotton (Gossypium hirsutum)." *Theoretical*

882 *and Applied Genetics* 133.12 (2020): 3273-3285.
883 [16] Christmas, Matthew J., et al. "Chromosomal inversions associated with environmental
884 adaptation in honeybees." *Molecular ecology* 28.6 (2019): 1358-1374.
885 [17] Fang, Zhou, et al. "Megabase-scale inversion polymorphism in the wild ancestor of
886 maize." *Genetics* 191.3 (2012): 883-894.
887 [18] Fang, Lei, et al. "Genomic analyses in cotton identify signatures of selection and loci
888 associated with fiber quality and yield traits." *Nature genetics* 49.7 (2017): 1089-1098.
889 [19] Xu, Yunbi, et al. "Smart breeding driven by big data, artificial intelligence, and
890 integrated genomic-enviromic prediction." *Molecular Plant* 15.11 (2022): 1664-1695.
891 [20] Tomar, Vipin, et al. "Increased predictive accuracy of multi-environment genomic
892 prediction model for yield and related traits in spring wheat (Triticum aestivum
893 L.)." *Frontiers in Plant Science* 12 (2021): 720123.
894 [21] Sandhu, Karansher S., et al. "Multi-trait multi-environment genomic prediction for end-
895 use quality traits in winter wheat." *Frontiers in Genetics* 13 (2022): 831020.
896 [22] Yan, Weikai, Kirby T. Nilsen, and Aaron Beattie. "Mega-environment analysis and
897 breeding for specific adaptation." *Crop Science* 63.2 (2023): 480-494.
898 [23] Saludares, Rica Amor, et al. "Multi-trait multi-environment genomic prediction of
899 preliminary yield trial in pulse crop." *The Plant Genome* 17.3 (2024): e20496.
900 [24] Habyarimana, Ephrem, Bruno Parisi, and Giuseppe Mandolino. "Genomic prediction for
901 yields, processing and nutritional quality traits in cultivated potato (Solanum tuberosum
902 L.)." *Plant Breeding* 136.2 (2017): 245-252.
903 [25] Wang, Weixuan, et al. "Integration of high-throughput phenotyping, GWAS, and
904 predictive models reveals the genetic architecture of plant height in maize." *Molecular*
905 *Plant* 16.2 (2023): 354-373.
906 [26] Wang, jingxin, et al. "Accurate genomic prediction for grain yield and grain moisture
907 content of maize hybrids using multi-environment data." *Journal of Integrative Plant*
908 *Biology* 67.5 (2025): 1379-1394.
909 [27] Saludares, Rica Amor, et al. "Multi‐trait multi‐environment genomic prediction of
910 preliminary yield trial in pulse crop." *The Plant Genome* 17.3 (2024): e20496.
911 [28] Costa-Neto, Germano, Jose Crossa, and Roberto Fritsche-Neto. "Enviromic assembly
912 increases accuracy and reduces costs of the genomic prediction for yield plasticity in
913 maize." *Frontiers in Plant Science* 12 (2021): 717552.
914 [29] de Los Campos, Gustavo, et al. "A data-driven simulation platform to predict cultivars'
915 performances under uncertain weather conditions." *Nature communications* 11.1 (2020):
916 4876.
917 [30] Chengxiu Wu, Jingyun Luo, and Yingjie xiao. "Multi-omics assists genomic prediction
918 of maize yield with machine learning approaches." *Molecular Breeding* 44.2 (2024): 14.

[31] Fernandes, Igor K., et al. "Using machine learning to combine genetic and environmental data for maize grain yield predictions across multi-environment trials." *Theoretical and Applied Genetics* 137.8 (2024): 189.

[32] Jighly, Abdulqader, Matthew Hayden, and Hans Daetwyler. "Integrating genomic selection with a genotype plus genotype x environment (GGE) model improves prediction accuracy and computational efficiency." *Plant, Cell & Environment* 44.10 (2021): 3459-3470.

[33] He, Kunhui, et al. "Leveraging automated machine learning for environmental data-driven genetic analysis and genomic prediction in maize hybrids." *Advanced Science* 12.17 (2025): 2412423.

[34] Novakovsky, Gherman, et al. "Obtaining genetics insights from deep learning via explainable artificial intelligence." *Nature Reviews Genetics* 24.2 (2023): 125-137.

[35] Gao, Pengfei, et al. "SoyDNGP: a web-accessible deep learning framework for genomic prediction in soybean breeding." *Briefings in bioinformatics* 24.6 (2023): bbad349.

[36] Montesinos-López, Abelardo, et al. "Multi-environment genomic prediction of plant traits using deep learners with dense architecture." *G3: Genes, Genomes, Genetics* 8.12 (2018): 3813-3828.

[37] Zou, Qixiang, et al. "Large-scale crop dataset and deep learning-based multi-modal fusion framework for more accurate G× E genomic prediction." *Computers and Electronics in Agriculture* 230 (2025): 109833.

[38] Peterson, Leif E. "K-nearest neighbor." *Scholarpedia* 4.2 (2009): 1883.

[39] Ma, Wenlong, et al. "A deep convolutional neural network approach for predicting phenotypes from genotypes." *Planta* 248.5 (2018): 1307-1318.

[40] Wang, Kelin, et al. "DNNGP, a deep neural network-based method for genomic prediction using multi-omics data in plants." *Molecular Plant* 16.1 (2023): 279-293.

[41] Wang, Peipei, et al. "Prediction of plant complex traits via integration of multi-omics data." *Nature Communications* 15.1 (2024): 6856.

[42] Sundararajan, Mukund, and Amir Najmi. "The many Shapley values for model explanation." *International conference on machine learning*. PMLR, (2020): 9269-9278.

[43] Almasy, Laura, and John Blangero. "Variance component methods for analysis of complex phenotypes." *Cold Spring Harbor Protocols* 2010.5 (2010): pdb-top77.

[44] Payne, Roger W. "GenStat." *Wiley Interdisciplinary Reviews: Computational Statistics* 1.2 (2009): 255-258.

[45] Mohammed, Jamal, et al. "A GWAS identified loci and candidate genes associated with fiber quality traits in a new cotton MAGIC population." *Theoretical and Applied Genetics* 138.1 (2025): 10.

[46] Ran Fu, and xiangfeng Wang. "Modeling the influence of phenotypic plasticity on maize

hybrid performance." *Plant Communications* 4.3 (2023):100548.
[47] Piepho, H. P. "A comparison of the ecovalence and the variance of relative yield as measures of stability." *Journal of Agronomy and Crop Science* 173.1 (1994): 1-4.
[48] Ma, Zhiying, et al. "Resequencing a core collection of upland cotton identifies genomic variation and loci influencing fiber quality and yield." *Nature genetics* 50.6 (2018): 803-813.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTable7.Multitraitmultienvironmentallystableandunstablevarietiesof419uplandcotton.xlsx
- SupplementaryTable6.Analysisofvarianceandheritabilityfor13traitsof419uplandcotton.xlsx
- SupplementaryTable9.ThesharedMainSNPsassociatedwithcrosstraitcategoriesusingphenotypevaluesof419uplandcottonforGWAS.xlsx
- SupplementaryTable13.Thephenotypicplasticityparametersof13traitsof419uplandcotton.xlsx
- SupplementaryTable11.ThesharedSNPsassociatedwithcrosstraitcategoriesusingthreeparametersof419uplandcottonforGWAS.xlsx
- SupplementaryTable12.ThespecificandsharedSNPsassociatedwithtraitsof419uplandcottonfromGWASbasedonthreeparameters.xlsx
- SupplementaryTable5.Thegeometeorologicalcharacteristicsofthesixlocationsin2014and2015.xlsx
- SupplementaryTable14.Theenvironmentaladaptabilityparametersof13traitsof419uplandcotton.xlsx
- SupplementaryTable8.TheMainSNPsassociatedwith3traitcatagoriesusingphenotypevaluesof419uplandcottonforGWAS.xlsx
- SupplementaryTable10.TheenvironmentspecificandsharedMainSNPsusingphenotypevaluesof419uplandcottonfromsixlocationsforGWAS.xlsx
- SupplementaryTable15.ThesharedSNPlociofGWASandinterpretabilityanalysisofAttGEINetmodel.xlsx
- Supplementaryinformation.pdf