Research Insight Open Access

AI-Assisted Genomic Prediction Models in Cotton Breeding

Jinhua Cheng, Mengting Luo

Institute of Life Science, Jiyang College of Zhejiang A&F University, Zhuji, 311800, China

Corresponding email: mengting.luo@jicat.org

Cotton Genomics and Genetics, 2025, Vol.16, No.3 doi: 10.5376/cgg.2025.16.0015

Received: 28 Apr., 2025

Accepted: 07 Jun., 2025

Published: 29 Jun., 2025

Preferred citation for this article:

Cheng J.H., and Luo M.T., AI-Assisted genomic prediction models in cotton breeding, Cotton Genomics and Genetics, 16(3): 137-147 (doi: 10.5376/cgg.2025.16.0015)

Abstract Cotton is an important economic crop related to the national economy and people's livelihood, but traditional breeding faces challenges such as long cycle, low efficiency and difficulty in improving yield and quality simultaneously. As a new technology of molecular breeding, genomic selection (GS) improves breeding accuracy and efficiency by utilizing whole genome marker information, and shows great potential in crop breeding. In recent years, the rapid development of artificial intelligence (AI) technology has injected new impetus into agricultural breeding. The application of machine learning and deep learning to crop genome big data analysis is expected to accelerate the breeding process of crops such as cotton. This study reviews the current status and challenges of cotton breeding, the basic principles of genomic prediction breeding, and the application progress of artificial intelligence algorithms in cotton breeding. The research progress of genomic prediction of major cotton traits such as yield, stress resistance and fiber quality is discussed in detail. Typical cases in Australia, the United States and China are cited to analyze the practice of cotton intelligent breeding. The current challenges in data quality and model generalization ability, multi-omics data integration, model interpretability, etc. are analyzed, and the future development direction of the integration of artificial intelligence and genomic prediction is prospected. This study hopes to break through the bottleneck of traditional breeding, improve the efficiency and accuracy of cotton breeding, and cultivate new varieties with high yield, high quality and multi-resistance.

Keywords Cotton breeding; Genomic selection; Phenotypic prediction; Deep learning; Intelligent breeding

1 Introduction

Cotton (*Gossypium hirsutum* Linn.) is one of the most important natural fiber crops in the world, and it is also an important cash crop and textile industry raw material in my country. Its yield and quality directly affect the textile industry and farmers' income. After years of development, cotton breeding has made remarkable progress, especially after the promotion of insect-resistant transgenic cotton, my country has become the second major country after the United States to have transgenic cotton varieties with independent intellectual property rights. However, cotton breeding still faces many challenges. Traditional breeding mainly relies on phenotypic selection and experience accumulation, with a long breeding cycle and low efficiency, and it is difficult to respond to new challenges of climate change and pests and diseases in a timely manner. At the same time, cotton yield and fiber quality are often negatively correlated, and it is extremely difficult to maintain or improve quality while increasing yield. For example, in the past, it was difficult to take into account both fiber length and strength in breeding, which once became a technical bottleneck. Conventional breeding has limited improvement in soil salinity, drought and other adversity resistance, and cotton production is still deeply affected by drought, salinity and disease. These factors have led to severe challenges in my country's high-quality and high-yield cotton breeding, which requires new technical means to break through (Sun et al., 2022).

The development of molecular breeding has provided new ideas for cotton breeding. Genomic selection (GS) was proposed by Meuwissen et al. in 2001. It has developed significantly in the past two decades and has been verified in crops such as wheat and corn to improve selection accuracy and accelerate the breeding process. GS predicts individual breeding values by weighted estimation of high-density markers across the whole genome, overcoming the limitation of traditional marker-assisted selection that only uses a few major QTLs, and has shown great potential in improving crop yield, stress resistance and quality (Budhlakoti et al., 2018; 2022). With the decline in the cost of high-throughput sequencing and genotyping, cotton genome sequencing and variation map construction have been rapidly advanced, providing a data basis for the implementation of GS. In particular, the high-quality reference genome and pan-genome mapping of cotton completed in recent years have helped to discover key genes that control important traits such as yield and quality. On this basis, the rise of artificial intelligence technology has given breeding wings. Machine learning and deep learning methods can automatically extract complex patterns from massive multi-omics data for trait prediction and decision support. This makes it possible for crop breeding to shift from experience-driven to data-driven. For example, intelligent breeding systems that integrate genetic genotype, environment and phenotypic big data have emerged in some studies, which can accurately predict offspring traits, screen excellent genes and improve breeding efficiency in the early stages of breeding (Yan and Wang, 2022). It can be foreseen that the combination of artificial intelligence and genomic technology will lead the future "breeding 5.0" era and accelerate the cultivation of new crop varieties that meet future needs (Wu et al., 2024).

This study focuses on "AI-assisted cotton genomic prediction breeding", sorts out the basic principles and application status of genomic selection in cotton breeding, as well as the practical application and research progress of AI algorithms such as machine learning and deep learning in cotton breeding, introduces the background status and technical needs of cotton breeding, and explains the concept and methodological basis of genomic prediction breeding, including genotype data acquisition, genetic variation analysis and prediction model establishment. It focuses on reviewing the research progress of artificial intelligence methods (such as random forests, support vector machines, neural networks, etc.) in the prediction of important cotton traits (yield, stress resistance and fiber quality). Through cases such as Australia's CSIRO breeding program, the US public breeding project and China's intelligent breeding practice, the actual application effect of AI in cotton breeding is analyzed, and suggestions for promoting cotton intelligent breeding are put forward. This study hopes to provide useful references for scientific researchers and breeders, and accelerate the cultivation of new high-yield, high-quality and stress-resistant cotton varieties. This is of great significance for ensuring the supply of textile raw materials, improving the competitiveness of the cotton industry and the sustainable development of agriculture.

2 Basic Principles of Genomic Prediction in Cotton

2.1 Concepts of genomic selection and phenotypic prediction

Genomic selection is a breeding method that uses genome-wide molecular markers to predict the genetic potential of individuals. Unlike traditional breeding that relies on measured phenotypes, GS builds a prediction model by estimating marker effects in a training population, and directly predicts the genetic breeding value of candidate individuals that have not been phenotyped, thereby accelerating the selection process (Viana et al., 2016). The core of GS is to capture the genetic control information of quantitative traits using a large number of SNP markers across the genome. As long as the molecular marker coverage is dense enough, even if the effect of a single marker is small, the accumulation of thousands of markers can accurately predict complex traits. This strategy of "pre-selecting phenotypes with genomes" is regarded as a key step in modern crop breeding, which can improve selection accuracy, shorten generation cycles, and increase genetic gain. In cotton, GS is particularly suitable for breeding of typical quantitative traits such as yield, fiber quality, and stress resistance. It is reported that the prediction accuracy of GS for cotton fiber length and strength can reach a high level of 0.65-0.76, showing an effect superior to traditional phenotypic selection. Phenotypic prediction is the goal of GS, that is, to predict the phenotypic performance or breeding value of an individual through genotypic data. In addition to classic statistical methods such as GBLUP (genomic best linear unbiased prediction), machine learning algorithms have also been gradually applied to GS models in recent years to improve prediction accuracy (Billings et al., 2022).

2.2 Acquisition and quality control of genotypic data

The premise for implementing genomic predictive breeding is high-quality genotypic data. The cotton genome is large (2n=4x=52) and highly repetitive, but the development of sequencing technology in recent years has made high-density typing possible. Commonly used genotype acquisition methods include SNP chips and resequencing. For example, the US CSIRO breeding project constructed a high-density chip containing 12 296 polymorphic SNP sites and genotyped 1 385 cotton materials. With the reduction of high-throughput sequencing costs, whole genome resequencing has become increasingly popular in cotton, and millions of marker variants can be detected at one time (Sun et al., 2022). However, strict quality control (QC) is crucial for both chips and sequencing data. Common QC steps include: removing markers with high missing and error rates, filtering low allele frequency (MAF) markers to reduce background noise, and removing samples with abnormal heterozygosity or duplicate identity. High-quality data can ensure the reliability of model training. It is reported that in cotton GS research, only thousands to tens of thousands of SNPs with high quality are usually selected for modeling. It is also necessary to pay attention to the impact of homologous fragments and structural variations unique to polyploid cotton on typing. The recently constructed cotton genome variation database (CottonGVD) integrates a large amount of SNP, Indel and structural variation information of different cotton species. Using these resources, breeders can obtain the genotype of the target material more comprehensively. In essence, the quality of genotype data will directly affect the accuracy of the prediction model. Only by conducting strict QC on the basis of fully understanding the genetic variation and diversity structure of cotton can a solid foundation be laid for subsequent GS modeling (Peng et al., 2021).

2.3 Association analysis methods between genetic variation and prediction models

The association analysis between genetic variation and trait phenotype is a key link in building genomic prediction models. In cotton breeding, traditional genome-wide association analysis (GWAS) and quantitative trait mapping (QTL mapping) have been widely used to discover gene loci that affect yield, quality, and resistance (Tan et al., 2024). However, GWAS can usually only detect a few significant main effect loci, and it is difficult to capture most of the minor effect genes for complex quantitative traits. The genomic prediction model improves the utilization efficiency of minor effect genes by integrating the effects of all markers. Commonly used GS statistical models include: GBLUP, RR-BLUP, BayesA/B/C, LASSO, etc. These methods are similar to multivariate regression in principle, but the difference lies in the different prior assumptions about marker effects. In practice, the model can be selected according to the genetic architecture of the trait. For example, for traits such as fiber quality that may be controlled by fewer large-effect QTLs, the Bayesian model sometimes performs better; while for traits such as yield that are highly controlled by multiple genes, RR-BLUP and other models that assume uniform minor effects are more robust (Budhlakoti et al., 2022).

Machine learning algorithms such as random forests (RF) and support vector machines (SVM) do not need to assume linear additivity and can capture nonlinear interactions between markers. They are introduced into GS to improve prediction capabilities. For example, a study used a machine learning model to successfully predict the response genes of cotton under low temperature stress, with an accuracy significantly higher than that of the traditional linear model. In the process of GS model training, cross-validation or independent validation sets are usually used to evaluate the prediction accuracy (such as correlation coefficient or root mean square error). It is worth noting that cotton is an allotetraploid, and the interaction between its A and D subgenomes may affect the prediction model. Homologous genes and linkage disequilibrium structure should be fully considered. Based on association analysis and combined with genetic parameters (such as marker variance and heritability), the training set selection and weighted modeling strategies can also be optimized (Billings et al., 2022). The association discovery and modeling of genetic variation and phenotype is a process of continuous iterative optimization. Reasonable selection of analysis methods and adjustment in combination with cotton genetic characteristics can improve the performance of genomic prediction models and promote the successful implementation of predictive breeding.

3 Application of AI Algorithms in Cotton Breeding

3.1 Machine learning algorithms

Machine learning, with its powerful nonlinear modeling capabilities, is increasingly becoming a powerful tool for cotton breeding data analysis. Classic machine learning algorithms such as random forest (RF), support vector machine (SVM), gradient boosting tree, etc. have been used in cotton genome prediction and trait mining research. For example, the random forest algorithm can evaluate the relative importance of each gene marker to the trait by integrating a large number of decision trees, thereby achieving accurate prediction of complex quantitative traits. Dhaliwal et al. (2022) used random forest combined with long-term field trial data to successfully predict the yield performance of cotton under different conservation tillage measures. The model not only gave high-precision predictions, but also provided an explanation of the factors affecting yield, which has reference value in actual agronomic decision-making. Support vector machines have outstanding performance in small sample modeling and high-dimensional data processing. Studies have used SVM combined with cotton gene expression data to predict disease resistance genes with a high accuracy rate. In addition, clustering algorithms can be used for genetic diversity analysis and kinship division of cotton germplasm resources, providing a basis for combining parents. It should be pointed out that machine learning models often have many hyperparameters, which need to be optimized through methods such as cross-validation to prevent overfitting and improve generalization ability. In cotton genomic selection, the introduction of machine learning algorithms helps to capture non-additive interaction effects between markers and improve prediction accuracy. For example, Zhao et al. (2023) integrated machine learning methods into gene regulatory network analysis to identify key control genes affecting cottonseed yield. This shows that machine learning can not only be used to predict trait values, but also to discover important breeding factors. In breeding practice, machine learning models can also integrate phenotypic imaging data to achieve automatic measurement and evaluation of cotton agronomic traits.

3.2 Construction and optimization of deep learning models

As a subfield of machine learning, deep learning is characterized by multi-layer neural networks and can automatically extract complex data features. It is emerging in cotton genetic improvement research. Compared with traditional machine learning, deep learning models (such as convolutional neural networks CNN, recurrent neural networks RNN, graph neural networks GNN, etc.) have end-to-end learning capabilities and are particularly suitable for processing large-scale, high-dimensional genomic and phenotypic data. In cotton breeding, a typical application of deep learning is to combine high-throughput phenotypic imaging for trait prediction. For example, Li et al. (2024) used a deep convolutional neural network to analyze high-throughput image data of cotton fruit branch angles, extracted phenotypic characteristics related to genotypes, and then combined GWAS to locate key genes affecting fruit branch angles, greatly improving the efficiency of trait genetic analysis. This combination of "deep phenotype+genome" provides a new paradigm for quantitative trait improvement. In addition, deep learning can also be directly used to build genomic prediction models. Budhlakoti et al. (2022) developed the DeepGS model, which inputs the whole genome SNP sequence into a multi-layer neural network to capture the complex relationship between markers and phenotypes in a nonlinear way, showing better prediction accuracy than GBLUP. In cotton, Zhao et al. (2024)'s research is unique: they used a combination of convolutional neural networks and Transformers to develop the DeepFDML deep model, which specifically predicts functional methylation sites in the cotton genome. DeepFDML trained a CNN-Transformer hybrid network on thousands of known functional methylation sites, and ultimately increased the model's area under the ROC curve from 0.65 to 0.82, significantly outperforming traditional methods. This shows that deep learning has unique advantages in mining complex "gene-epigene-phenotype" relationships. Of course, deep models often require massive data support, and the training process is computationally intensive, requiring the use of high-performance computing resources such as GPUs (Yan et al., 2024). In terms of model optimization, regularization, Dropout and other techniques can alleviate overfitting, and hyperparameter adjustment and architecture improvements (such as adding attention mechanisms, autoencoder pre-training, etc.) can improve model performance.

3.3 Integration and analysis of multi-environment data

In the actual breeding process, different ecological environments have a significant impact on the performance of cotton traits, so how to integrate multi-environment data to improve the generalization ability of the model is an important topic. In traditional breeding experiments, multi-point ring tests are often used to evaluate the adaptability and stability of varieties. However, incorporating environmental factors into genomic prediction models remains challenging because environmental variables are often difficult to quantify and there is an interaction between genes and the environment (G×E). Current research shows that genomic predictions in multiple environments can make progress by combining statistical models and machine learning. For example, the so-called “reaction paradigm model” adds environmental covariates to GS or constructs environmental principal components to explain G×E variation in genomic prediction (Budhlakoti et al., 2022). In cotton, methods such as Jarquín have been used to improve predictions of yields in different locations, but their application is not yet widespread.

Zhang et al. (2025) recently conducted a breakthrough in genome-wide association and prediction analysis of resistance to Verticillium wilt using data from natural cotton populations from different test sites in Xinjiang that have been identified for many years. As a result, 10 disease resistance QTL loci that are stable in multiple environments were identified, and a genomic selection breeding model was established based on these loci, which was verified in the offspring population for its good predictive ability for disease resistance phenotypes. This case proves that the multi-environment GS model can identify robust favorable allele combinations and achieve effective improvement of traits in complex environments. The integration of environmental data also includes the quantification of factors such as climate and soil. For example, meteorological indicators at the test site can be obtained through remote sensing, or measured environmental parameters can be used as covariates to add to the model. Some scholars have also proposed using the multimodal capabilities of deep learning to simultaneously input environmental and genomic data into the neural network, allowing the model to autonomously learn the interaction between the two. At present, for crops such as cotton, a major bottleneck in the integration of multi-environment data is adaptability: the model performs well in one region, but the accuracy may decrease when it is transferred to another region, so a larger range of training data and more physically explanatory environmental representations are needed (Gapare et al., 2018). With the construction of cotton experimental station networks and big data platforms in various countries, more abundant multi-environment genotype-phenotype data will be available in the future, creating conditions for the development of robust cross-environment prediction models. Multi-environment genomic prediction is expected to improve the reliability of breeding selection and screen out new cotton varieties that are both high-yielding and widely adaptable, which is of practical significance for responding to climate change and heterogeneous environmental challenges.

4 Research Progress on Prediction of Major Cotton Traits

4.1 Prediction models for yield traits

Increasing cotton yield has always been the primary goal of breeding, and genomic prediction provides a new way to accelerate the selection of high-yield varieties. Cotton yield traits include seed cotton yield and its components (such as boll weight, boll number per plant, lint percentage, etc.), which are controlled by quantitative genes and easily affected by the environment. Early gene mapping studies have identified many QTLs related to yield, but the effect of a single locus is limited (Sun et al., 2022). Genomic selection predicts yield performance by integrating whole genome information. The Australian CSIRO study was the first to verify the prediction effect of GS on yield in a large-scale cotton breeding population: Li et al. (2022) conducted genotyping and two-season field trials on cotton of 1,385 breeding lines and established multiple prediction models. The results showed that the Bayesian model combining genomic markers and pedigrees had a correlation coefficient of 0.64 for the prediction of lint yield, and could accurately distinguish high-yield materials in lines that had not been field tested. This suggests that GS can help eliminate low-yield genotypes in the early breeding generation and improve selection efficiency. In the US public breeding program, researchers have also evaluated the feasibility of GS for yield traits. Billings et al. (2022) pointed out that the accuracy of genomic prediction of cotton yield and related agronomic traits is currently slightly lower than that of fiber quality and other traits, but is comparable to traditional phenotypic selection. With the optimization of models and the improvement of phenotypic accuracy, there is room for further improvement. They suggested that GS could be implemented on quality traits first, and then gradually expanded to complex traits such as yield after accumulating experience. In addition to directly predicting yield, the combination of high-throughput phenotyping and GS is also a direction of progress. For example, the extraction of cotton field canopy characteristics through remote sensing images and the use of machine learning models to predict final yield have been successful at the regional scale (Dhaliwal et al., 2022). In China, some scholars have used whole genome association analysis to identify key loci that affect yield composition and used them for marker-assisted selection (MAS), but the application of GS is still in its infancy. With the mapping of genetic variation maps of my country's cotton core germplasm and the accumulation of breeding big data, the role of genomic prediction in high-yield cotton breeding will gradually emerge. It is worth mentioning that AI technology can also help analyze the complex mechanism of yield formation. For example, Zhao et al. (2023) integrated transcriptome and machine learning and discovered multiple major regulatory genes that affect boll number and boll weight at the same time. These findings provide gene targets for directly improving yield through gene editing and other means. In summary, the research on genomic prediction breeding in cotton yield improvement has made initial progress. In the future, with more perfect models and more training data, it is expected to achieve high-precision prediction and efficient selection of yield traits.

4.2 Prediction studies on stress resistance traits

Common adversities in cotton production include diseases (such as Verticillium wilt, Fusarium wilt), drought, high salt, etc. Breeding new stress-resistant varieties is the key to ensuring stable yield. Traditional stress-resistant breeding is often time-consuming and labor-intensive, and genomic prediction is expected to accelerate this process. In terms of disease-resistant breeding, genomic selection technology has shown feasibility (Li, 2024). Zhang et al. (2025) used multi-environment Verticillium wilt resistance identification data of 1 152 upland cotton germplasms to construct a GS model for disease resistance traits. Based on the multi-year stable QTL information training model, they predicted the disease resistance of an $F_{2:3}$ segregating population, and the correlation coefficient was above 0.5, which was significantly better than phenotypic screening alone. More importantly, the model predicted that the selected materials showed higher disease resistance in the field, proving that GS is practical in cotton disease resistance breeding. This study also located 10 major disease resistance QTLs in combination with GWAS and aggregated them in breeding materials, reflecting the power of association analysis combined with GS (Figure 1). For drought resistance, salt and alkali resistance and other stresses, molecular biological methods are more often used at home and abroad to clone functional genes or create transgenic materials. There are few reports on the application of GS in drought-resistant breeding. The reasons are that drought resistance phenotypes are difficult to obtain and that environmental interactions are strong, making predictions complicated. However, some indirect traits such as drought-related physiological indicators can be used as alternative phenotypes to apply GS models. At present, studies have summarized the physiological and molecular regulatory mechanisms of cotton drought and salt tolerance, providing candidate markers and genes for subsequent genome predictions (Ma et al., 2021). For example, the cloned GhCBL1-GhCIPK signaling pathway genes are involved in the regulation of cotton drought resistance. If corresponding markers can be developed, they can be incorporated into the GS model to improve the prediction accuracy of drought resistance. Epigenetic information has also been shown to be related to stress resistance. Zhao et al. (2024) found that 36% of the expression variations of resistance-related genes were associated with DNA methylation variations, but not with conventional genetic variations. These epigenetic markers independent of DNA sequences can also be used to assist prediction models, thereby improving the ability to capture stress resistance. It can be expected that with the development of multi-omics technology, stress-resistant breeding will shift from single gene engineering to whole genome integration optimization. Combining genomic selection with high-pressure screening (such as artificial inoculation of pathogens and simulated drought stress) is expected to quickly select stress-resistant superior plants from a large number of offspring. In general, the research on genomic prediction of cotton stress resistance traits has just started but has broad prospects. By constructing an intelligent model that comprehensively considers genomic, epigenetic and environmental factors, we can expect to achieve early prediction of disease resistance and stress resistance potential and accelerate the breeding process of new stress-resistant cotton varieties.

4.3 Advances in predicting fiber quality traits

Fiber quality (including fiber length, strength, fineness, etc.) is the core indicator for measuring the value of cotton, and is also a trait that has a trade-off with yield in breeding. Genomic prediction is of special significance in improving fiber quality, because improving quality in traditional breeding often comes at the expense of yield. Through GS, it is expected to discover gene combinations that increase yield without reducing quality, and achieve synergistic improvement of the two. At present, cotton fiber quality is one of the most significant areas of GS research. CSIRO's experiments have shown that the accuracy of GS prediction of fiber quality is much higher than that of yield: in its study of 1 385 materials, the prediction accuracy of the average length of the upper half of the fiber and the specific breaking strength reached 0.76 and 0.65 respectively. This means that long fibers and high-strength materials can be reliably distinguished based on genotype alone, which provides the possibility for quality-oriented selection. The reason for this phenomenon may be that the genetic control of fiber quality is relatively simple, the main effect gene plays a greater role, and the impact of the environment is relatively small relative to yield, so the GS model is more effective. In the analysis of the US public breeding program, it was also found that fiber quality traits are more suitable for GS implementation to speed up the screening of new lines and reduce the workload of field fiber testing (Billings et al., 2022).

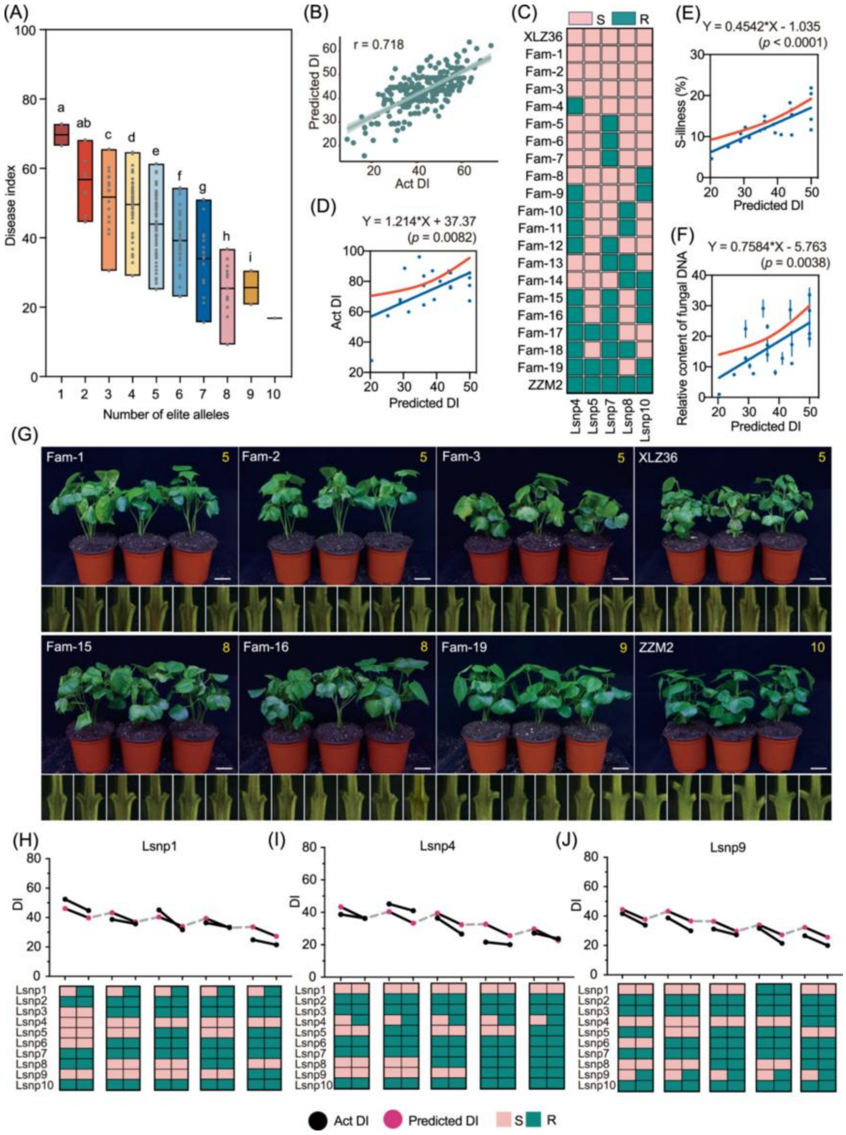

Figure 1 Pyramiding effect of 10 Lsnp^{R}s. (A) Distribution of DI in cotton accessions carrying different numbers of Lsnp^R. The x-axis represents cotton accessions carrying 1-10 Lsnp^R. The y-axis represents All-b disease index (DI) across all environments. The letters indicate the statistical test after the *t*-test ($p < 0.001$). (B) The correlation between the DI predicted by the MDIC and the actual DI. The x-axis represents predicted DI; the y-axis represents the actual DI of the corresponding genotype materials. (C) The genotypes of 19 F$_2$ individuals at 5 Lsnps. ZZM2 and XLZ36 are resistant and susceptible parents, respectively. (D) Correlation between actual DI and predicted DI in 19 F$_{2:3}$ lines. (E) Correlation between extracted proportion of diseased area in stem sections and predicted DI in 19 F$_{2:3}$ lines. (F) Correlation between actual vascular pathogen content and predicted DI in 19 F$_{2:3}$ lines. D-F are all simple linear regressions, and the p-value represents the hypothesis that the slope is non-zero. (G) Six extreme F$_{2:3}$ lines were selected from the predicted DI, with images showing the disease phenotype and cross-sections of stems at leaf nodes. Photographed at 13 days post-inoculation. The yellow number in the top right corner indicates the number of Lsnp^R carried by the line. Fam1-Fam3 are lines with the highest predicted DI. Fam15, 16, and 19 are lines with the lowest predicted DI. Scale bar, 3 cm. (H-J) The effect of enhancing VW resistance in existing cotton materials after transformation from Lsnp^S to Lsnp^R. (H-J) Represents the DI distribution between the Lsnp^S series cotton varieties and Lsnp^R series cotton varieties at Lsnp1, 4, and 9 with low frequency. The genotypic map (below x-axis) shows haplotypes across accessions, with the upper line graph indicating mean DI per haplotype. The black dots represent the actual DI, and the magenta dots represent the predicted DI by the MDIC. BLUE, best linear unbiased estimates; MDIC, molecular disease index calculator (Adopted from Zhang et al., 2025)

In addition to high prediction accuracy, GS can also help break the genetic bottleneck of quality improvement. The "subgenome modular design breeding" proposed by Chinese scholars is an innovative idea: by comparing the differences in the contribution of the two subgenomes of tetraploid upland cotton to fiber development, the genetic modules that restrict quality are identified, and molecular methods are used to recombinantly design the key gene combinations therein, thereby breaking through the limitations of quality improvement. This concept has made progress in preliminary experiments and is expected to be verified in future breeding practices and combined with GS methods (Zhang and Wang, 2024). At the same time, the research of the Xinjiang Production and Construction Corps used the excellent allele variation of sea island cotton to introgress into the upland cotton background, significantly improving the latter's fiber length and strength (Sun et al., 2022). These excellent alleles can also be incorporated into the GS model through marker development to improve the prediction and selection efficiency of quality. In recent years, with the in-depth study of high-quality cotton germplasm resources, several important genes affecting fiber development have been cloned or located, such as GhPAP and other fiber cell wall synthesis-related genes, whose downregulation will lead to a significant decrease in fiber strength. Incorporating these gene loci information into GS can further improve the biological interpretability and effectiveness of the model. At the breeding practice level, India, the United States and other countries have used GS to screen intermediate materials with excellent fiber quality, accelerating the introduction of new lines (Islam et al., 2019). In my country, there are still few studies on genomic prediction of cotton fiber quality, but the relevant foundation has been established: cotton genome sequencing and variation identification have revealed multiple structural variations and candidate genes related to quality; the National Cotton Improvement Center has constructed a series of recombinant inbred line populations with improved quality, which can provide training sets for GS. It can be foreseen that in the near future, breeders will be able to use AI models to pre-evaluate the fiber quality and yield of hundreds of recombinant offspring at the same time, and select excellent individuals that "have both fish and bear's paw". This will greatly improve the efficiency and success rate of breeding of high-quality cotton varieties in my country, and promote fiber quality breeding into a new stage of intelligence.

5 Case Studies: Practical Applications of AI in Cotton Breeding

5.1 Genomic prediction practices in CSIRO's cotton breeding program in Australia

The cotton breeding project of CSIRO, Australia is one of the examples of the successful application of genomic prediction technology in crop breeding. Faced with the dual goals of improving fiber quality and maintaining high yield, the breeding team of CSIRO began to try to incorporate GS into its breeding program in the 2010s. They collected multi-season field phenotypic data of thousands of breeding materials and performed high-density genotyping on them. In the process of improving the germplasm material resistant to two-spotted spider mites through backcrossing, despite the continuous advancement of backcrossing generations (BC generations), the mite resistance trait score of the selected material was always significantly better than that of the susceptible parent Sicot 714B3F and remained stable. This resistance stability reflects that CSIRO has effectively retained the target traits through GS and improved the disease and insect resistance without sacrificing the main agronomic traits. The picture shows the significant differences between resistant and susceptible varieties under natural infection at the phenotypic level, providing visual verification for the superior individuals selected by the GS model (Figure 2) (Conaty et al., 2022). Li et al. (2022) reported in detail the results of CSIRO's implementation of genomic prediction on 1,385 lines: the Bayesian LASSO model was combined with genomic SNP and pedigree data to predict traits such as fiber length, strength and yield, achieving remarkable accuracy (length 0.76, strength 0.65, yield 0.64). Of particular note, they found that the fusion of whole genome marker information with conventional pedigree data can effectively improve prediction accuracy, indicating that genomic data has formed a beneficial supplement to traditional breeding information. In practice, CSIRO has used GS to assist in the screening of early-generation materials: for new combinations that have not been field tested, the fiber quality and yield potential are predicted by genotype, and combinations with poor prediction values are eliminated, thereby reducing the workload of field trials and accelerating the generation process. It is reported that after applying GS, the breeding cycle of its new varieties was shortened by about 2 years, and the resource utilization efficiency was significantly improved (Li et al., 2022).

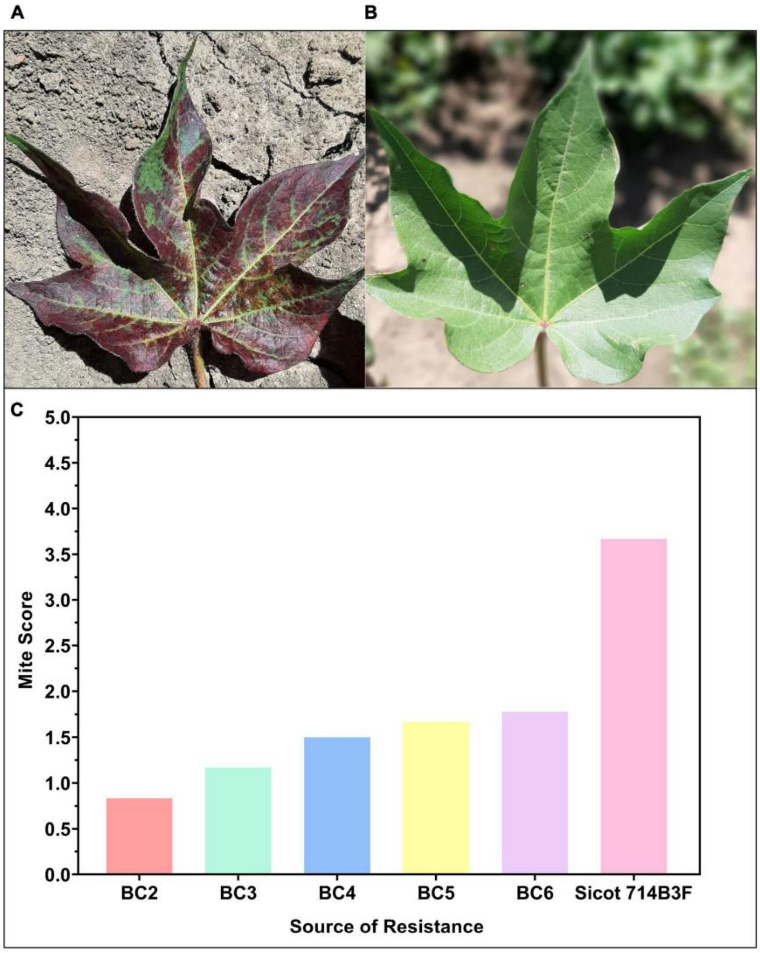

Figure 2 (A) Susceptible (Sicot 714B3F recurrent parent) and (B) resistant two-spotted spider mite cotton germplasm from the CSIRO cotton breeding program (Photos: Lucy Egan). (C) Progress of breeding mite resistant germplasm showing that as backcross (BC) generation number increases mite resistance scores have remained lower than the susceptible recurrent parent, Sicot 714B3F, and relatively stable. Data from C. Trapero, used with permission (Adopted from Conaty et al., 2022)

CSIRO has also developed a new parent selection strategy in combination with GS. For example, in response to the negative correlation between yield and quality that has been troubled in the past, they used the prediction model to select hybrids with better fiber quality without significantly reducing yield. Today, CSIRO's cotton varieties enjoy a reputation in the international market for their excellent quality. This successful case shows that integrating AI-driven genomic prediction into the traditional breeding process can achieve a "win-win" in breeding efficiency and breeding effect. CSIRO's experience also provides a reference for other crop breeding, that is, it is necessary to establish a high-quality phenotype-genotype database, continuously optimize the prediction model, and gradually expand the application scope of GS in actual decision-making. It can be expected that CSIRO will further try to integrate environmental data, phenotypic images and other information into predictions in the future, and build a more intelligent breeding decision support system to maintain its international leading position in cotton breeding.

5.2 Comparison of genomic selection methods in U.S. public cotton breeding programs

In the United States, since commercial cotton breeding is mainly dominated by private enterprises, public breeding units are particularly active in exploring new technologies. In recent years, the United States Department of Agriculture (USDA) has conducted feasibility studies on cotton genomic selection in collaboration with several universities. The goal is to evaluate the effects of GS on different traits and provide decision-making basis for public breeding programs. Billings et al. (2022) collected a large amount of phenotypic data from regional trials of cotton in the United States, including yield, quality and disease resistance traits, and used existing cotton high-density SNP chips to perform typing analysis on these varieties. They used multiple statistical models for comparison and found that: for quality traits such as fiber length and strength, the genomic prediction accuracy is high, and GS can be fully implemented in early-generation screening; for traits such as yield and maturity, the prediction accuracy is relatively low but still comparable to the efficiency of traditional family selection. Based on this, they suggested that public breeding projects can be divided into "two steps": first, introduce GS in fiber quality improvement to accelerate the cultivation of lines that meet industrial high-end requirements; then, with the accumulation of more environmental data and model improvement, expand GS to complex traits such as yield and stress resistance (Billings et al., 2022). It is worth mentioning that the study also compared the effects of different algorithms, including G-BLUP, BayesC, and random forests, and found that the stability of traditional linear models was slightly better when the amount of data was limited. But they also pointed out that machine learning and deep learning may have greater potential in the future. Some American breeders have begun to try to use simple artificial neural network models to simulate the combining ability of cotton hybrid combinations to predict which parent combinations are more likely to produce excellent offspring (Patil et al., 2023). These explorations of public projects have laid the foundation for the promotion of GS in cotton breeding.

The experience of the United States also emphasizes the importance of model interpretability by breeders: they expect the prediction model to not only give results, but also indicate which markers or genes are most important in trait control, so as to verify with traditional genetic knowledge. Therefore, the US team often combines GS analysis with GWAS, incorporating significant markers into the model or annotating the markers with the highest model weight. This practice has increased breeders' trust in AI models and increased the adoption rate of GS results in practice. Overall, the US public breeding department has clarified the advantages and disadvantages of GS and its application boundaries through comparative analysis, providing a scientific basis for the implementation of technology. At present, they are promoting the establishment of a cotton breeding big data platform to integrate scattered historical breeding data and create conditions for the large-scale implementation of artificial intelligence-assisted breeding in the next step.

5.3 Intelligent design breeding systems in China's cotton breeding

Compared with Europe and the United States, my country's practice of artificial intelligence-enabled cotton breeding started later but progressed rapidly. On the one hand, national scientific research institutions and universities are actively carrying out relevant research; on the other hand, enterprises and new R&D institutions have also joined in to develop intelligent breeding platforms. In recent years, the Cotton Research Institute of the Chinese Academy of Agricultural Sciences has laid out the "smart breeding" research direction, using its own cotton germplasm resource bank, phenotyping platform and molecular laboratory to explore the application of AI technology in the entire process of cotton breeding (Si et al., 2022). One of the representative achievements is the construction of a cotton whole genome selection breeding platform. A joint team from Zhejiang University and the Chinese Academy of Agricultural Sciences reported in 2023 that they integrated 32.5 Tb of multi-omics and phenotypic data, developed a central database for breeders to query gene expression, gene networks and epigenetic information, and established a cotton trait prediction model and decision support system on this basis. The platform is figuratively called the breeding "central kitchen". Breeders only need to input the genotypes of candidate parents, and the system can give the predicted performance and optimal selection plan of the hybrid combination on the target trait. Although the platform is still in the trial stage, it has shown great potential to shorten the breeding cycle (Zhao et al., 2024).

Another eye-catching case is the cooperation between Lakeside Laboratory and Xinjiang Academy of Agricultural Sciences to use AI to crack the genetic mechanism of cotton stress resistance and apply it to breeding. They constructed a genome-wide DNA methylation map of 207 cotton varieties, identified 287 million single methylation polymorphic sites, and developed a deep learning model DeepFDML to predict which methylation variations affect gene expression, thereby discovering 43 key eQTM genes potentially involved in fiber development. More importantly, they successfully increased cotton fiber length after editing one of the genes through CRISPR. This achievement shows that AI can not only assist in selection, but also guide the discovery of gene editing targets to achieve "design breeding". In terms of the development of intelligent breeding systems, some domestic agricultural technology companies have also invested in it. Of course, the construction of my country's cotton intelligent breeding system is still in its early stages, but under the dual promotion of the government and the market, the relevant infrastructure is gradually being improved. The national level has deployed a major "digital seed industry" project to support the research and development of crop intelligent breeding technology, among which cotton is one of the key targets. It can be foreseen that in the near future, a phenotypic big data network covering major cotton-growing test stations, a database integrating the genome information of major domestic cotton germplasm, and a set of open and shared AI breeding tools will be established to empower the majority of breeders. China's intelligent breeding system will give full play to its latecomer advantage, learn from foreign experience and integrate local massive data to achieve leapfrog development. This will effectively promote the improvement of the breeding level of new cotton varieties in my country and ensure the sustainable and healthy development of the national cotton industry.

6. Challenges and Future Directions

6.1 Limitations in data quality and model generalization

Although artificial intelligence has shown great prospects in cotton breeding, it still faces many challenges. The first is the problem of data quality. High-precision genotype and phenotypic data are the basis for establishing reliable prediction models, but it is not easy to obtain large-scale, high-quality data in actual breeding. For example, field management and measurement errors at different test sites will cause phenotypic data noise, and there may also be missed variants or typing errors in genotyping (Ma et al., 2021). These data noises will directly affect the training effect and prediction accuracy of the model. Therefore, it is necessary to improve data quality through repeated experiments, standardized measurements and strict data cleaning. As an allopolyploid, cotton has a complex genome that makes accurate typing difficult, and some structural variants and homologous fragments may not be detected or correctly located (Sun et al., 2022). This lack of information will weaken the model's explanatory power for traits. The second is the limited generalization ability of the model. A model trained in a specific population and environment is often difficult to directly apply to materials with large differences in genetic basis or under different ecological conditions. For example, the drought prediction model established on Xinjiang data may not be applicable to varieties in the cotton region of the Yellow River Basin. Therefore, the model needs to have certain transfer learning and adaptive capabilities. Many current GS models will perform significantly worse outside the training set, which is one of the practical problems facing cotton AI breeding (Liu and Huang, 2022). To improve the generalization of the model, we can consider: increasing the diversity of training data to cover more genetic backgrounds and environments; introducing hierarchical models to embed population division or environmental classification into the model structure; and using ensemble learning to improve robustness by fusing multiple models. High-dimensional labeled data can easily lead to model overfitting, and feature selection or regularization methods are needed to constrain model complexity. Another problem is that AI models are highly dependent on input data. When the genotype of the new material has allelic variation that does not appear in the training set, the model may not be effectively used. This suggests that we should continuously update the training data and model parameters to keep them synchronized with the latest genetic diversity information. Although the data and model challenges are significant, they are not insurmountable. With the promotion of the construction of agricultural big data platforms at the national level, cotton breeders will be able to share richer data resources in the future. At the same time, the development of machine learning is also providing new algorithms to improve the ability of small sample learning and cross-domain generalization. As long as we face these shortcomings and actively improve them, artificial intelligence will surely serve cotton breeding practice more maturely.

6.2 Challenges in integrating and analyzing multi-omics data

The formation of cotton traits involves multi-level information such as genome, transcriptome, epigenome, metabolome and environmental factors. How to effectively integrate multi-omics data to improve breeding models is a frontier topic in current artificial intelligence breeding. On the one hand, multi-omics data is huge and of different types. For example, a cotton variety may have hundreds of millions of DNA methylation sites, tens of thousands of expressed genes and thousands of metabolites. These data dimensions far exceed traditional genotypes and phenotypes, and fusion analysis requires powerful computing power and new algorithms. On the other hand, there are complex associations and redundancies between multi-omics. For example, some key gene mutations can cause chain changes in transcription and metabolism, resulting in highly correlated data; for example, epigenetic changes are sometimes independent of DNA sequences and have additional contributions to traits. Simply splicing different omics features into the model may cause information noise and overfitting. Therefore, it is necessary to design special methods to extract key signals from each group and establish hierarchical connections between them and phenotypes.

At present, some studies have initially shown the value of multi-omics fusion in cotton breeding. The work of Zhao et al. (2024) proved that combining genomic and epigenomic data can reveal many functional variations that cannot be identified by genomic data alone, providing a new perspective for trait prediction. By constructing a multi-omics regulatory network, they locked 43 core genes for cotton fiber development, which would not be found if they were based on traditional GWAS alone. However, the challenges that need to be solved in multi-omics integration include: data standardization and alignment, different omics measurement scales are different, and normalization processing is required; feature selection, multi-omics data dimensions are extremely high, and how to screen out features that are truly associated with traits is difficult; model complexity, multi-omics fusion models often contain a large number of parameters, and the risk is more prone to overfitting, so more stringent regularization and cross-validation strategies need to be introduced (Guo et al., 2022). Deep learning provides a powerful tool for multi-omics fusion, and its multimodal network can automatically learn associations between different data modes. However, such models usually lack biological interpretability, which is also an aspect that needs to be weighed. The cost of acquiring multi-omics data is high. For example, whole-genome methylation sequencing of hundreds of cotton materials generates tens of TB of data. Therefore, economic costs and computational overheads must be considered in practical applications. Nevertheless, with the innovation of sequencing and detection technologies, multi-omics data will become increasingly abundant and accessible. We have reason to believe that by developing smarter data fusion algorithms (such as neural networks, Transformers, etc.) and making full use of cloud computing and high-performance computing clusters, multi-omics-driven cotton intelligent breeding will become possible. It will enable breeders to understand the formation mechanism of excellent traits from all aspects of genes, transcription, and epigenetics, and make more targeted designs and selections on this basis.

6.3 Combining model interpretability with practical applications

Artificial intelligence models, especially deep learning models, are often regarded as "black boxes", which may affect their promotion and application in the field of breeding. What breeders want to know more is: Why does the model give such a prediction? Which genes or markers play a key role in it? If the model is difficult to explain, its prediction results are often difficult to be directly adopted by breeding decisions. Therefore, improving the biological interpretability of the model is a problem that AI breeding must face. One approach is to combine traditional genetic knowledge to analyze the model output. For example, the largest number of markers in the GS model can be counted and compared with known QTLs or genes to verify whether the model captures reasonable genetic signals. If the markers emphasized by the model are connected to important functional genes, the credibility of the results is increased (Billings et al., 2022).

Another approach is to use specialized interpretation algorithms, such as SHAP values and sensitivity analysis, to quantify the impact of each input marker on the prediction, thereby identifying the gene regions that the model "values". A recent study explained the random forest model for cotton yield prediction and found that some of the markers that the model assigned the highest weights were precisely the areas near the previously reported yield QTLs. This shows that AI models can reproduce the judgment of human experts to a certain extent, thereby enhancing the trust of breeders.

In order to facilitate practical application, the model needs to be closely integrated with the breeding process. For example, develop a friendly user interface so that breeders can input data and obtain prediction results without knowing programming; embed models into breeding management systems to achieve real-time prediction and decision support; provide model uncertainty indicators to remind users to be cautious when the prediction confidence is low. These are all the work being promoted in the current application of AI breeding. In terms of promotion and application, it is also important to train breeders to master basic data analysis and model interpretation skills. Only when breeders understand the model can they better use it to guide practical work such as hybrid combination design and generation advancement. China has already taken some actions in this regard, such as organizing national cotton breeding backbones to participate in the "Digital Seed Industry and Intelligent Breeding" training course to share AI breeding cases and experiences. It can be foreseen that in the next few years, artificial intelligence breeding tools will be gradually implemented in front-line breeding units and continuously improved based on feedback. At that time, the model will no longer be a mysterious black box, but will become a daily auxiliary tool similar to soil testing and disease diagnosis, and will be skillfully used by breeders.

7 Concluding Remarks

Artificial intelligence technology is gradually being integrated into cotton breeding research and has achieved initial results. Predictive breeding models such as genomic selection have made useful attempts to improve cotton yield, fiber quality and stress resistance traits: in terms of yield, high-yield genotypes are predicted through whole genome markers, which improves the selection efficiency of early breeding generations; in terms of fiber quality, the GS model achieves high-precision prediction of indicators such as length and strength, promoting the selection of high-quality new lines; in terms of stress resistance, the prediction model combined with machine learning successfully identified multi-environmentally stable disease-resistant QTLs, accelerating the screening of disease-resistant varieties. Internationally, Australia's CSIRO took the lead in integrating genomic prediction into the breeding process, significantly shortening the breeding cycle and cultivating high-quality and high-yield new varieties; the US public breeding department systematically evaluated the effects and limitations of GS, laying the foundation for further promotion and application. Domestic scientific research institutions have also actively deployed intelligent breeding research, developed a cotton intelligent breeding platform, and used AI technology to crack the genetic mechanism of some complex cotton traits. It can be said that artificial intelligence is helping breeders break through the bottleneck of traditional breeding and realize the transformation of breeding decisions from experience-driven to data-driven. Although the application of AI-assisted breeding in cotton is still in its infancy, the existing results have proved its great potential and bright prospects.

Looking to the future, the deep integration of artificial intelligence and genome prediction will lead cotton breeding into a new era. On the one hand, with the advancement of sequencing and phenotyping technologies, breeding will obtain exponentially growing multidimensional data to provide fuel for AI models. Whole genome selection will be combined with new technologies such as gene editing and epigenetic regulation to form integrated innovation in breeding technology. Intelligent algorithms will be able to handle more complex breeding goals, such as improving yield, quality and multi-resistance at the same time, and realizing the optimal design of comprehensive traits. On the other hand, the emergence of a new generation of AI models (such as neural networks, generative AI, etc.) is expected to further improve the accuracy and breadth of breeding predictions. In the future, cotton breeding decisions may be generated by AI with countless breeding schemes, taking into account gene combinations, environmental adaptability and market demand, and selecting the best scheme for breeders' reference. The breeding cycle will also be greatly shortened due to assisted generation prediction. In theory, a breeder is expected to experience a complete iteration of multiple breeding cycles in his career, which was unimaginable in the past. Of course, we also need to realize that on the road to "breeding 5.0", there are still many scientific problems to be solved, such as how to accurately simulate the impact of gene interactions on traits, how to integrate evolution and niche theory in the model, etc. These require in-depth cross-disciplinary and cooperation between genetics and artificial intelligence. But what is certain is that artificial intelligence will serve as a powerful new engine to drive cotton breeding forward and contribute to the safety of textile raw materials and sustainable agriculture.

In order to better integrate artificial intelligence technology into cotton breeding practice, we put forward the following suggestions: First, establish a standardized cotton breeding big data system. Including unified phenotypic measurement specifications, building a national joint breeding database, and improving the genotype information sharing platform, etc., to provide high-quality training data for AI models. Secondly, strengthen the training of interdisciplinary talents, encourage breeding experts to learn data science methods, and cultivate more compound talents who understand both genetic breeding and AI algorithms to open up the "last mile" of technology application. Thirdly, AI breeding should be promoted step by step: starting with easy-to-predict traits such as fiber quality, accumulating experience in practice, and then gradually expanding to complex goals such as yield and stress resistance, step by step, and taking the lead. Fourth, scientific research and production should be closely integrated to strengthen demonstration and application. Select several advantageous cotton-producing areas to establish "intelligent breeding demonstration stations" to actually test the selection effect and economic benefits of AI models, and win the trust of the breeding community with example verification. Finally, formulate relevant standards and specifications to ensure the reliability and repeatability of AI breeding software tools, and avoid waste of resources caused by improper application. In short, we believe that with the joint efforts of all parties in industry, academia and research, the road to AI-enabled cotton breeding will become wider and wider. In the future, the cultivation of new varieties of "super cotton" with high yield, high quality and multi-resistance will no longer rely entirely on the intuition and experience of breeders, but will be accurately obtained through scientific data analysis and prediction with the assistance of artificial intelligence. This will greatly improve breeding efficiency, reduce breeding costs, and promote the quality improvement, efficiency increase and sustainable development of the cotton industry in my country and even the world. The tide of the times is rolling forward, and the integrated development of artificial intelligence and cotton breeding has become a general trend. We have reason to be confident in its bright prospects.

Acknowledgments

We would like to express my gratitude to the reviewers for their valuable feedback, which helped improve the manuscript.

Conflict of Interest Disclosure

The authors affirm that this research was conducted without any commercial or financial relationships that could be construed as a potential conflict of interest.

References

Billings G.T., Jones M.A., Rustgi S., Bridges W.C., Holland J.B., HulseKemp A.M., Campbell B.T., 2022, Outlook for implementation of genomics-based selection in public cotton breeding programs, Plants, 11(11): 1446.
https://doi.org/10.3390/PLANTS11111446

Budhlakoti N., Kushwaha A.K., Rai A., Chaturvedi K.K., Kumar A., Pradhan A.K., Kumar U., Kumar R.R., Juliana P., Mishra D.C., and Kumar S., 2022, Genomic selection: a tool for accelerating the efficiency of molecular breeding for development of climate-resilient crops, Frontiers in Genetics, 13: 832153.
https://doi.org/10.3389/fgene.2022.832153

Budhlakoti N., Mishra D., Arora D., and Kumar R., 2018, Application of genomic selection in agriculture, Bhartiya Krishi Anusandhan Patrika, 33(4): 295-297.
https://doi.org/10.18805/BKAP138

Conaty W.C., Broughton K.J., Egan L.M., Li X.Q., Li Z.T., Liu S.M., Llewellyn D.J., MacMillan C.P., Moncuquet P., Rolland V., Ross B., Sargent D., Zhu Q.H., Pettolino F.A., and Stiller W.N., 2022, Cotton breeding in Australia: meeting the challenges of the 21st century, Frontiers in Plant Science, 13: 904131.
https://doi.org/10.3389/fpls.2022.904131

Dhaliwal J.K., Panday D., Saha D., Lee J., Jagadamma S., Schaeffer S., and Mengistu A., 2022, Predicting and interpreting cotton yield and its determinants under long-term conservation management practices using machine learning, Computers and Electronics in Agriculture, 199: 107107.
https://doi.org/10.1016/J.COMPAG.2022.107107

Gapare W., Liu S., Conaty W., Zhu Q., Gillespie V., Llewellyn D., Stiller W., and Wilson I., 2018, Historical datasets support genomic selection models for the prediction of cotton fiber quality phenotypes across multiple environments, G3: Genes, Genomes, Genetics, 8(5): 1721-1732.
https://doi.org/10.1534/g3.118.200140

Guo W.J., Li D.W., Xie S., Yang L.W., Li C., Tian J., Pu L., and Gu X.F., 2022, Artificial intelligence accelerates epigenetics and plant breeding, Journal of Agricultural Science and Technology, 24(12): 90-100.
https://doi.org/10.13304/j.nykjdb.2022.1031

Islam M., Fang D., Jenkins J., Guo J., McCarty J., and Jones D., 2019, Evaluation of genomic selection methods for predicting fiber quality traits in Upland cotton, Molecular Genetics and Genomics, 295(1): 67-79.
https://doi.org/10.1007/s00438-019-01599-z

Li Z., 2024, *Fusarium* boll rot in cotton: pathogen dynamics and control options, Molecular Pathogens, 15(4): 200-208.
https://doi.org/10.5376/mp.2024.15.0019

Li L., Chang H., Zhao S., Liu R., Yan M., Li F., El-Sheery N.I., Feng Z., and Yu S., 2024, Combining high-throughput deep learning phenotyping and GWAS to reveal genetic variants of fruit branch angle in upland cotton, Industrial Crops and Products, 220: 119180.
https://doi.org/10.1016/j.indcrop.2024.119180

Li Z., Liu S., Conaty W., Zhu Q., Moncuquet P., Stiller W., and Wilson I., 2022, Genomic prediction of cotton fibre quality and yield traits using Bayesian regression methods, Heredity, 129(2): 103-112.
https://doi.org/10.1038/s41437-022-00537-x

Liu J.J., and Huang F.J., 2022, Current situation and problems and countermeasures of cotton production in Xinjiang, Cotton Sciences, 44(5): 15-19.

Ma P.P., Zhao Z.Q., Zhu J.B., and Sun G.Q., 2021, Physiological and molecular mechanisms of drought and salt tolerance in cotton, Journal of Agricultural Science and Technology, 23(2): 27-36.
https://doi.org/10.13304/j.nykjdb.2019.0718

Patil A.E., Deosarkar D., Khatri N., and Ubale A.B., 2023, A comprehensive investigation of Genotype-Environment interaction effects on seed cotton yield contributing traits in *Gossypium hirsutum* L. Using multivariate analysis and artificial neural network, Computers and Electronics in Agriculture, 211: 107966.
https://doi.org/10.1016/j.compag.2023.107966

Peng Z., Li H., Sun G., Dai P., Geng X., Wang X., Zhang X., Wang Z.Z., Jia Y., Pan Z., Chen B.J., Du X., and He S., 2021, CottonGVD: a comprehensive genomic variation database for cultivated cottons, Frontiers in Plant Science, 12: 803736.
https://doi.org/10.3389/fpls.2021.803736

Si Z.F., Jin S.K., Li J.Y., Han Z.G., Li Y.Q., Wu X.N., Ge Y.X., Fang L., Zhang T.Z., and Hu Y., 2022, The design, validation, and utility of the "ZJU CottonSNP40K" liquid chip through genotyping by target sequencing, Industrial Crops and Products, 188(Part A): 115629.
https://doi.org/10.1016/j.indcrop.2022.115629

Sun Z.W., Gu Q.S., Zhang Y., Wang X.F., and Ma Z.Y., 2022, Research progress on cotton gene discovery and molecular breeding, Journal of Agricultural Science and Technology, 24(7): 32-38.

Tan H., Tang B., Sun M., Yin Q., Ma Y., Li J., Wang P., Li Z., Zhao G., Wang M., Zhang X., You C., and Tu L., 2024, Identification of new cotton fiber-quality QTL by multiple genomic analyses and development of markers for genomic breeding, The Crop Journal, 12(3): 866-879.
https://doi.org/10.1016/j.cj.2024.03.014

Viana J., Piepho H., and Silva F.F., 2016, Quantitative genetics theory for genomic selection and efficiency of breeding value prediction in open-pollinated populations, Scientia Agricola, 73(3): 243-251.
https://doi.org/10.1590/0103-9016-2014-0383

Wu C., Luo J., and Xiao Y., 2024, Multi-omics assists genomic prediction of maize yield with machine learning approaches, Molecular Breeding,44(2): 14.
https://doi.org/10.1007/s11032-024-01454-z

Yan C., Li J., Feng Q., Luo J., and Luo H., 2024, ResDeepGS: a deep learning-based method for crop phenotype prediction, In: Proceedings of the 4th international conference on computational agriculture and bioinformatics, Springer Nature Singapore, Singapore, pp.470-481.
https://doi.org/10.1007/978-981-97-5131-0_40

Yan J., and Wang X., 2022, Machine learning bridges omics sciences and plant breeding, Trends in Plant Science, 28(2): 199-210.
https://doi.org/10.1016/j.tplants.2022.08.018

Zhang Q., and Wang Y., 2024, AI in biology: transforming genomic research with machine learning, Computational Molecular Biology, 14(3): 106-114.
https://doi.org/10.5376/cmb.2024.14.0013

Zhang X.J., Liu S.M., Wu P., Xu W.Y., Yang D.Y., Ming Y.Q., Xiao S.H., Wang W.R., Ma J., Nie X.H., Gao Z., Lv J.Y., Wu F., Yang Z.G., Zheng B.X., Du P., Wang J.M., Ding H., Kong J., Aierxi A., Yu Y., Gao W., Lin Z.X., You C.Y., Lindsey K., Štajner N., Wang M.J., Wu J.H., Jin S.X., Zhang X.L., and Zhu L.F., 2025, A panoramic view of cotton resistance to *Verticillium dahliae*: from genetic architectures to precision genomic selection, iMeta, 4(3): e70029.
https://doi.org/10.1002/imt2.70029

Zhao T., Guan X.Y., Hu Y., Zhang Z.Q., Yang H., Shi X.W., Han J., Mei H., Wang L.Y., Shao L., Wu H.Y., Chen Q.Q., Zhao Y.Y., Pan J.Y., Hao Y.P., Dong Z.Y., Long X., Deng Q., Zhao S.J., Zhang M.K., Zhu Y.M., Ma X.W., Chen Z.Q., Deng Y.Y., Si Z.F., Li X., Zhang T.Z., Gu F., Gu X.F., and Fang L., 2024, Population-wide DNA methylation polymorphisms at single-nucleotide resolution in 207 cotton accessions reveal epigenomic contributions to complex traits, Cell Research, 34(12): 859-872.
https://doi.org/10.1038/S41422-024-01027-X

Zhao T., Wu H.Y., Wang X.T., Zhao Y.Y., Wang L.Y., Pan J.Y., Mei H., Han J., Wang S.Y., Lu K.N., Li M.L., Gao M.T., Cao Z.Y., Zhang H.L., Wan K., Li J., Fang L., Zhang T.Z., and Guan X.Y., 2023, Integration of eQTL and machine learning to dissect causal genes with pleiotropic effects in genetic regulation networks of seed cotton yield, Cell Reports, 42(9): 113111.
https://doi.org/10.1016/j.celrep.2023.113111

CropSci Publisher©

Disclaimer/Publisher's Note

The statements, opinions, and data contained in all publications are solely those of the individual authors and contributors and do not represent the views of the publishing house and/or its editors. The publisher and/or its editors disclaim all responsibility for any harm or damage to persons or property that may result from the application of ideas, methods, instructions, or products discussed in the content. Publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.