nature portfolio

Peer Review File

Robustly Enhancing Crop Genomic Prediction Accuracy Through Ensemble Learning and Iterative Optimization

Corresponding Author: Dr Jianxiao Liu

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
Referee Report – Nature / Nature Communications
Robustly Enhancing Crop Genomic Prediction Accuracy Through Ensemble Learning and Iterative Optimization

Overall Evaluation
This manuscript introduces GEG2P, an ensemble genomic prediction (GP) framework integrating twenty heterogeneous base learners (statistical, machine learning, and deep learning models). The ensemble weights are optimized using a genetic algorithm (GA) combined with a three-step iterative optimization designed to improve stability and cross-species generalization. The study evaluates 36 traits across five major crops and claims up to ~15% gains in predictive accuracy relative to single models and existing deep-learning methods (SoyDNGP, CropFormer, WheatGP, EGGPT). The unique contribution of this work lies in its comprehensive approach to ensemble learning in genomic prediction, which has the potential to enhance the accuracy and applicability of genomic prediction models significantly.
Despite the apparent novelty and scale, I cannot recommend acceptance in its current form. Important structural, methodological, and conceptual issues must be resolved before reconsideration. I would classify this submission as Rejection due to a lack of statistical robustness, methodological transparency, and over-interpretation of biological findings.

1. Conceptual and Statistical Framework
The article lacks a formal statistical foundation connecting ensemble learning to quantitative genetics. The authors treat ensemble learning as a 'black box' that aggregates heterogeneous predictors, but did not establish its statistical equivalence or contrast it with established frameworks (GBLUP, RKHS, multi-kernel GBLUP, Bayesian model averaging). Without a model of variance decomposition (additive, dominance, and residual), it is not possible to evaluate whether reported gains are due to improved modeling of additive effects, non-linear interactions, or random noise. The claim that GEG2P 'improves stability and generalization across species' is unsupported by any formal measure of variance or uncertainty. A Nature paper in genomic prediction must quantify uncertainty, cross-trait variance, and bias–variance tradeoffs.

2. Cross-Validation and Information Leakage
It appears that the GA-based optimization and hyperparameter tuning were conducted globally before model evaluation, not within a nested cross-validation. If so, the reported improvements in PCC are statistically inflated and cannot be interpreted as generalization performance. There is no indication that environmental or population structure stratification was maintained during fold splitting, which is critical in genomic data.

3. Statistical Evaluation and Metrics
The authors report only Pearson correlation coefficients (PCC), which are insufficient and can exaggerate apparent gains when the variance in the target variable differs among traits. Provide RMSE, MAE, and R^2 adjusted for trait heritability. Quantify uncertainty (SD or SE) across CV folds and use paired statistical tests to demonstrate the significance of differences vs. baselines. Report heritability (h^2) for each trait and contextualize results relative to expected genetic variance.

4. Algorithmic Transparency and Reproducibility
The genetic algorithm parameters (population size, crossover/mutation rates, selection scheme, and stopping rule) are not

well described. The three-step iterative optimization is vaguely defined, lacking pseudocode, convergence analysis, and justification for the number of iterations. Computational requirements and open-source code are lacking.

5. Benchmarking and Comparative Framework
Although the proposed GEG2P is compared with other methods, these benchmarks are not evaluated under equivalent conditions. Hyperparameter optimization is mentioned but not described in detail. It is unclear whether each model was retrained using identical cross-validation splits. No statistical significance testing or discussion of computational cost vs. predictive gain is provided.

6. Overextended Biological Interpretation
The SHAP-based interpretability analysis is impressive in scope but scientifically unfocused and overinterpreted. Hundreds of genes and pathways are listed without synthesis. The claim that 'GEG2P identifies broader functional genes' is not substantiated with quantitative validation. The mapping of epistatic networks remains purely computational, with no biological validation or replication.

7. Writing, Structure, and Length
Many sections repeat results for different species. Figures are overcrowded, and labels are hard to read. English syntax requires professional editing, and the introduction should be refocused on theoretical advances rather than a literature listing.

8. Impact and Theoretical Value
Despite computational scale, the proposed method does not provide new theoretical insights into genomic prediction. Ensemble learning is not novel; the work primarily scales up existing ideas without clear theoretical advancements. No ablation or sensitivity analysis identifies which components actually drive improvement.

Minor and Editorial Points
1. Abstract: summarize quantitative improvements once and avoid repetition.
2. Clarify Figure 1 (GEG2P v1–v3 distinctions).
3. Verify that all cited works are accessible publications or preprints.
4. Move redundant supplementary figures (S11–S16) to supporting files.
5. Include complete Data and Code Availability statements per Nature policy.
Final Recommendation
Recommendation: Reject

While the manuscript presents an ambitious and computationally heavy approach, it does not yet meet the standards of statistical rigor, reproducibility, or interpretability required for publication in Nature. To reach that level, the authors must:
1. Establish a clear statistical and genetic framework for ensemble integration.
2. Redesign cross-validation and quantify uncertainty.
3. Report complete algorithmic parameters and release code.
4. Condense and reframe interpretability analyses into biologically meaningful insights.
5. Reorganize and improve clarity and precision throughout.

(Remarks on code availability)
no applicable

Reviewer #2

(Remarks to the Author)
The manuscript "GEG2P: Robustly Enhancing Crop Genomic Prediction Accuracy Through Ensemble Learning and Iterative Optimization" proposes a comprehensive ensemble learning framework integrating 20 base learners across statistical, machine learning, and deep learning domains. The authors introduce a three-step iterative strategy and a Genetic Algorithm for weight optimization, demonstrating improved prediction accuracy across five crop species.
I acknowledge the significant engineering effort and the extensive benchmarking performed in this study. The proposed framework shows potential practical value in breeding programs where accuracy is paramount. However, the complexity of the ensemble approach raises concerns regarding computational efficiency, stability, and the depth of interpretability. I recommend the authors address the following points to strengthen the manuscript.

1. The framework integrates 20 heterogeneous models across R and Python environments, creating a complex computational pipeline. While accuracy improves, the manuscript lacks a detailed assessment of computational resources , such as training and inference time, hardware requirements. I suggest adding a comparison of computational costs between GEG2P and single SOTA models like Cropformer to discuss whether the increased computational burden is cost-effective for practical breeding applications.
2. The weight optimization relies on a GA, which inherently introduces stochasticity. I recommend verifying the robustness of the optimized weights by running the algorithm multiple times with different random seeds to ensure consistency. This validation is necessary to prove that the performance gains stem from capturing stable biological patterns rather than random fluctuations or overfitting to the validation set.
3. The current interpretability analysis, relying largely on SHAP values and the overlap of Top 10% SNPs, is somewhat

superficial for such a complex ensemble. SHAP explains feature importance but not the mechanism of ensemble correction. I suggest providing a specific case study demonstrating how the ensemble resolves conflicts via weight allocation when different base learners such as linear or others that generate contradictory predictions.
4. In the third step of iterative optimization, the algorithm progressively removes underperforming base learners. I suggest further analyzing the patterns of these removed models, for instance, whether specific model types , such as overfitting-prone deep learning models, are more likely to be discarded in certain species or traits. Discussing the patterns behind this selection process would provide more insight than simply describing the screening procedure.
5. The results indicate that XGBoost consistently dominates the weights for yield-related traits across multiple species. If a single model contributes the vast majority of predictive power, the necessity of ensembling many other low-weight models becomes questionable. The authors need to clarify whether the low-weight models contribute essential complementary signals or merely add system complexity without substantial benefit.
6. Given that 20 weight parameters are optimized through multiple iterative rounds, there is a risk of overfitting to the validation set. Please clarify whether the dataset used for weight optimization is completely independent of the test set used for reporting final accuracies. If there is overlap, the performance assessment might be inflated, and a strict independent test set or outer-loop cross-validation is recommended to confirm the results.
7. The paper constructs epistasis networks based on significant SNP interactions, but this appears to be a post-hoc statistical result based on ANOVA rather than a structure directly learned by the ensemble model. Please clarify the logical link between the model's predictive mechanism and these epistasis networks; specifically, does the ensemble's accuracy gain truly stem from implicitly capturing these interactions detected by ANOVA?
8. Since the integrated base learners have different requirements for input data formats (e.g., deep learning often requires One-hot encoding, while statistical models use 0/1/2 numerical encoding), please detail the data preprocessing pipeline in the Methods section. Clarifying whether a unified encoding was used or if specific processing was applied for different model categories is crucial for engineering reproducibility.
9. When comparing GEG2P with SOTA methods like Cropformer and WheatGP, Optuna was used for hyperparameter tuning. Please confirm that the computational budget (e.g., number of trials) allocated to these baseline methods was comparable to the extensive resources used for GEG2P. It is essential to ensure that the baselines were evaluated at their optimal performance to rule out artificial performance gaps caused by under-tuning.
10. Given the complex cross-language environment dependencies , reproducing this framework is challenging. To truly demonstrate the engineering value of this study, I strongly recommend that the authors provide a containerized environment or a detailed Conda environment configuration file in addition to the GitHub code to resolve potential dependency conflicts.
In conclusion, although the ensemble model lacks theoretical elegance, the manuscript demonstrates significant engineering value and practical utility. Therefore, I recommend a minor revision.

(Remarks on code availability)
Given the complex cross-language environment dependencies , reproducing this framework is challenging. To truly demonstrate the engineering value of this study, I strongly recommend that the authors provide a containerized environment or a detailed Conda environment configuration file in addition to the GitHub code to resolve potential dependency conflicts.

Version 1:

Reviewer comments:

Reviewer #2

(Remarks to the Author)
I have carefully reviewed the authors' responses to my initial review comments and examined their revised manuscript. I am satisfied with how they have addressed each of my concerns. I recommend acceptance of this manuscript for publication in Nature Communications.

(Remarks on code availability)
The code repository is publicly accessible and well organized, with a README file providing sufficient instructions for installation and execution. I was able to run the main workflow successfully, and the provided code appears sufficient to reproduce the major results reported in the manuscript.

Reviewer #3

(Remarks to the Author)
We commend the authors for the substantial engineering effort and extensive benchmarking invested in developing the GEG2P framework. Integrating twenty heterogeneous base learners (spanning statistical, machine learning, and deep learning models) into a unified genomic prediction tool represents a highly ambitious endeavor. Evaluating this framework across 36 traits in five major crop species illustrates the extensive scale of this study and highlights its relevance to applied breeding programs. The core concept of combining diverse algorithms offers a promising direction for genomic prediction. However, to fully realize the scientific value of an undertaking of this magnitude, it is necessary to move beyond reporting experimental prediction accuracies and establish robust statistical equivalence and unified interpretability standards.
In its current form, the revised manuscript has not fully articulated the mechanistic necessity and theoretical underpinnings of the proposed framework. We find that several core methodological concerns raised by Reviewer #1 during the initial review remain inadequately resolved. Specifically, the rebuttal predominantly relies on reporting raw prediction accuracies and providing descriptive defenses, rather than establishing the rigorous statistical baselines, convergence proofs, and computational clarifications necessary for a multi-algorithm fusion architecture. To elevate this manuscript to the requisite

standard and clearly elucidate its theoretical contribution, we strongly recommend addressing the following unresolved points.

1. Statistical Equivalence and Model Synergy (Ref. R1 Comment 1)

Although the authors compared their method with established frameworks, they did not establish statistical equivalence. We infer that Reviewer #1 expected a comparison at this level, rather than simply reporting experimental accuracy metrics. For a multi-algorithm fusion architecture, evaluating statistical equivalence (even empirically) constitutes its primary scientific value. Establishing this equivalence provides the necessary framework to demonstrate mechanistically how Linear Mixed Models (LMMs), machine learning, and deep learning capture complementary effects, a claim the authors currently assert without structural proof. As an actionable alternative, a straightforward residual analysis on the test set could yield rich insights into these synergistic relationships. While we certainly encourage the authors to employ other rigorous statistical approaches if they prefer, providing this depth of analysis will offer substantially more mechanistic value than supplementary accuracy tables.

2. Convergence Analysis and Iteration Ablation (Ref. R1 Comments 4 & 8)

The three-step iterative optimization is presented as a core algorithmic contribution. Structurally, however, this process resembles recursive re-weighting. To solidify this as a genuine algorithmic breakthrough (addressing the "Theoretical Value" and "Ablation" requested in Comment 8) rather than the natural smoothing effect of repeated averaging, a rigorous convergence analysis is required. The authors should provide actual convergence curves and conduct an ablation study that explicitly contrasts the three-step iteration against a direct, single-step optimization computed until full mathematical convergence. Demonstrating the mechanistic superiority of the three-step approach over a fully converged single step is essential to establish algorithmic rigor.

3. Rigor in Statistical Significance Testing (Ref. R1 Comment 5)

Integrating 20 highly diverse models introduces immense mathematical complexity, making traditional analytical P-values highly susceptible to assumption violations (e.g., normality, independence) in downstream testing. To ensure statistical rigor in a methodological paper, the authors must detail the foundational parameters of their significance tests (i.e., data splits, null and alternative hypotheses, specific test methods, and assumed distributions). Furthermore, in such a highly parameterized scenario, we strongly recommend deriving empirical P-values to validate the statistical significance of the findings.

4. Unified Interpretability Baseline and SHAP Implementation (Ref. R1 Comment 6)

The expanded biological validations and citations supporting SHAP's general effectiveness are noted. However, R1's underlying theoretical concern points to the implementation scale. Applying perturbation-based feature attribution to genomic data with tens or hundreds of thousands of SNPs is computationally exorbitant, particularly when scaled across 20 diverse methods. The authors should elucidate two key theoretical components: 1) how a unified interpretability baseline was established across these heterogeneous models, and 2) what approximation algorithms or computational strategies were utilized to make SHAP computationally feasible at this massive scale.

5. Minor Methodological Clarifications (Terminology and Formula Notation)

Upon reviewing the newly added and revised Methods sections, a few specific details require adjustment to prevent ambiguity:

(1) Page 31, Line 972 (Section 4.2.2 Genetic algorithm): Standard algorithmic terms such as "population" and "individual" are used. In the context of genomic prediction, these terms correspond to biological entities (e.g., breeding populations and individual accessions). We recommend adjusting the wording (e.g., using "candidate solutions") to explicitly distinguish the algorithmic parameters from biological datasets for interdisciplinary readers.

(2) Page 32, Line 992 (Equation 4): We recommend directly denoting this metric as RMSE and removing the explicit notation where j represents the sample number. This will prevent further conceptual confusion between biological individuals in a breeding population and algorithmic individuals in the genetic algorithm.

(3) Page 36, Line 1135 (Section 4.2.7 Genetic variance decomposition): Equation (14) presents a variance decomposition. This mathematical decomposition relies on specific underlying statistical conditions. Without explicitly stating these necessary conditions, the decomposition is technically incomplete. Please supplement the formula with its required statistical assumptions.

(4) Page 1, Line 21 (Abstract): The manuscript reports a 10.79% performance improvement compared against "single base learners." It is currently ambiguous whether this baseline represents the average performance of all 20 individual models or the performance of the single best constituent model. Evaluating an optimized ensemble against an average that incorporates poorly performing models risks artificially inflating the perceived gain. We request that the authors explicitly define this comparative baseline to ensure the reported metrics accurately reflect the true algorithmic advancement.

(Remarks on code availability)

Reviewer #4

(Remarks to the Author)

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

Dear Editor,

I conducted a structural and logical review of the provided source code, focusing specifically on the implementation of the genetic algorithm and the three-step iterative optimization modules. Due to time constraints, I did not locally execute the scripts to reproduce the results.

I examined the codebase primarily to resolve conceptual ambiguities in the manuscript's Methodology section. Specifically, the overlapping use of standard algorithmic terminology (e.g., "population" and "individual") and biological entities caused significant confusion. While inspecting the code clarified these mathematical implementations and confirmed that the underlying logic aligns with the proposed framework, the manuscript requires substantial revision. As detailed in my main comments, the authors must rewrite the text to ensure that interdisciplinary readers can fully comprehend the algorithmic design without relying on the source code.

Best regards,

Yuexin Ma

Version 2:

Reviewer comments:

Reviewer #3

(Remarks to the Author)
The authors have adequately addressed my previous concerns. The revised manuscript has been significantly improved, and the responses to the reviewers are clear, thorough, and satisfactory.

I appreciate the authors' efforts in revising the manuscript and providing detailed explanations for the issues raised during the review process. The additional analyses and clarifications have strengthened the work and improved its overall presentation.

I have no further major concerns and believe that the manuscript is now suitable for publication in Nature Communications.

(Remarks on code availability)
I have also reviewed the provided code and confirmed that it runs successfully.

Reviewer #4

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)
I have reviewed the provided code and confirmed that it runs successfully.

Response to Reviewers:

Dear Mr./Ms. reviewers:

We appreciate your valuable comments, and we tried our best to address them in the revised version of our paper. For the reviewers' comments, we have made the following modifications.

Reviewer: 1

Referee Report – Nature / Nature Communications

Robustly Enhancing Crop Genomic Prediction Accuracy Through Ensemble Learning and Iterative Optimization

Overall Evaluation

This manuscript introduces GEG2P, an ensemble genomic prediction (GP) framework integrating twenty heterogeneous base learners (statistical, machine learning, and deep learning models). The ensemble weights are optimized using a genetic algorithm (GA) combined with a three-step iterative optimization designed to improve stability and cross-species generalization. The study evaluates 36 traits across five major crops and claims up to ~15% gains in predictive accuracy relative to single models and existing deep-learning methods (SoyDNGP, CropFormer, WheatGP, EGGPT). The unique contribution of this work lies in its comprehensive approach to ensemble learning in genomic prediction, which has the potential to enhance the accuracy and applicability of genomic prediction models significantly.

Despite the apparent novelty and scale, I cannot recommend acceptance in its current form. Important structural, methodological, and conceptual issues must be resolved before reconsideration. I would classify this submission as Rejection due to a lack of statistical robustness, methodological transparency, and over-interpretation of biological findings.

(1) Conceptual and Statistical Framework

The article lacks a formal statistical foundation connecting ensemble learning to quantitative

genetics. The authors treat ensemble learning as a 'black box' that aggregates heterogeneous predictors, but did not establish its statistical equivalence or contrast it with established frameworks (GBLUP, RKHS, multi-kernel GBLUP, Bayesian model averaging). Without a model of variance decomposition (additive, dominance, and residual), it is not possible to evaluate whether reported gains are due to improved modeling of additive effects, non-linear interactions, or random noise. The claim that GEG2P 'improves stability and generalization across species' is unsupported by any formal measure of variance or uncertainty. A Nature paper in genomic prediction must quantify uncertainty, cross-trait variance, and bias–variance tradeoffs.

Response: Thank you for the suggestion. According to your suggestion, we further added experiments to evaluate the RKHS and Multi-kernel GBLUP methods and compared their prediction accuracies with that of GEG2P across 36 traits from 5 species (Supplemental Data 6). The results indicated that GEG2P outperformed both methods in prediction accuracy. Specifically, compared with RKHS, the average prediction accuracy of GEG2P increased by 5.94%, 5.53%, 3.48%, 4.30%, and 6.15% in maize, wheat, rice, chickpea, and soybean, respectively. In addition, compared with Multi-kernel GBLUP, the corresponding improvements were 5.52%, 8.30%, 4.40%, 3.76%, and 3.82%, respectively. We have added the analysis of the results of RKHS and Multi-kernel GBLUP in Section 2.5, as shown in the following:

“To further validate the phenotype prediction performance of GEG2P, we compared it with six state-of-the-art genomic prediction methods (SoyDNGP, Cropformer, WheatGP, EGGPT, RKHS, and Multi-kernel GBLUP) in maize, rice, wheat, chickpea, and soybean.”

“In maize, the average prediction accuracy of GEG2P is 111.26%, 11.03%, 77.25%, 4.18%, 5.94%, and 5.52% higher than that of SoyDNGP, CropFormer, WheatGP, EGGPT, RKHS, and Multi-kernel GBLUP, respectively (Figure 5A).”

“In wheat, the average prediction accuracy of GEG2P improved by 17.26%, 11.43%, 11.46%, 16.91%, 5.33%, and 8.30% compared with the six methods, respectively (Figure S3).”

“In rice, the prediction accuracy of GEG2P is also higher than that of the six state-of-the-art methods, with the average PCC improved by 14.84%, 9.30%, 11.82%, 1.89%, 3.48%, and 4.40%, respectively (Figure 5B).”

“In chickpea, the average PCC of GEG2P increased by 10.67%, 14.83%, 15.61%, 4.91%, 4.30%, and 3.76% compared with the six methods, respectively (Figure 5C).”

“In soybean, GEG2P also achieved the highest prediction accuracy, with the average PCC improved by 18.89%, 6.38%, 3.07%, 3.34%, 6.15%, and 3.82% compared with the six methods, respectively (Figure 5D).”

We referred to Endelman (2011), who demonstrated that rrBLUP and GBLUP are theoretically equivalent. In our study, GEG2P was already compared with rrBLUP in Section 2.2, where its prediction accuracy was 4.97% higher than that of rrBLUP. For Bayesian methods, we included six representative Bayesian methods in this study, namely BayesA, BayesB, BayesC, BL, BRR, and BRRN. The results showed that the average prediction accuracy of GEG2P was 5.00%, 4.80%, 5.04%, 5.43%, 5.26%, and 19.96% higher than that of these six methods, respectively (Section 2.2).

In addition, genetic variance decomposition was incorporated to further quantify differences in the genetic features identified by different base learners. Specifically, for the maize traits DTA, PH, and EW, the five base learners with the highest weights for each trait (Weight-Top-5) were selected, and variance decomposition was performed using SNPs ranked in the top 10% according to absolute SHAP values (hereafter denoted as Top10% SNPs). Notably, as the maize dataset used in this study consisted entirely of inbred lines, only additive, epistatic, and residual effects were considered in the variance decomposition model. The results showed that phenotypic variation in these traits was predominantly explained by genetic effects, with additive and epistatic variance accounting on average for 67.20% and 22.81%, respectively, whereas residual variance accounted for only 9.99% (Figure S19A~C). These results indicate that the improved prediction performance achieved by integrating multiple base learners primarily arises from complementary genetic effects rather than random noise. We have added the analysis of genetic variance decomposition to Section 3.4, as shown in the following:

“Similarly, variance decomposition analyses further demonstrated that Top10% SNPs identified by different base learners consistently exhibited substantial genetic effects. For maize DTA, Top10% SNPs in DTA-Weight-Top-5 explained on average 72.69% and 11.20% of additive and epistatic variance, respectively, with residual variance accounting for only 16.10% (Figure S19A). Similar patterns were observed for the PH and EW. For PH, the Top10% SNPs in PH-Weight-Top-5 explained on average 60.05%, 36.99%, and 2.95% of additive, epistatic, and residual variance, respectively (Figure S19B). For EW, the corresponding proportions were

68.86%, 20.24%, and 10.90%, respectively (Figure S19C). These results indicate that the key loci identified consistently exert potential effects on the phenotype despite variation in the genetic features captured by different base learners."

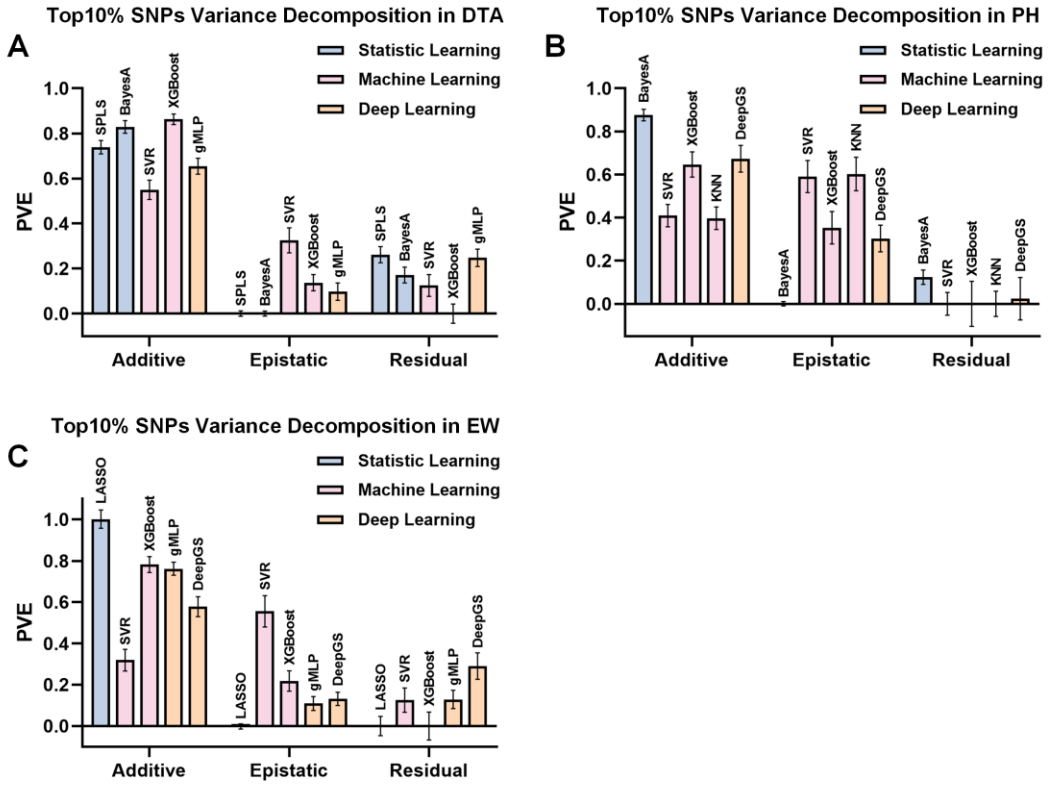

Figure S19. Variance decomposition of genetic effects for the Top10% loci in DTA-Weight-Top-5 (A), PH-Weight-Top-5 (B), and EW-Weight-Top-5 (C).

Furthermore, a description of the genetic variance decomposition method has been added in Section 4.2.7, as shown in the following:

"4.2.7 Genetic variance decomposition

In this study, the sommer was used to perform variance decomposition of the genetic effects for Top10% SNPs identified by different base learners. With the maize dataset used in this study consisting entirely of inbred lines, only additive, epistatic, and residual effects were considered in the variance decomposition model. Specifically, the *A.mat* and *E.mat* functions were first used to construct relationship matrices for additive (GRM_a) and epistatic (GRM_e) effects based on Top10% SNPs. Subsequently, a mixed linear model was fitted by incorporating GRM_a, GRM_e, and the residual as random effects. To improve numerical stability during matrix inversion, the tolerance

parameter (tolParInv) was set to 5×10^{-4}. The mixed linear model is shown in Eq. (14).

$$y = \mu + G_a + G_e + \varepsilon \tag{14}$$

In Eq. (14), y is an $n\times1$ vector of phenotypes, n denotes the sample size. μ is the overall mean. $G_a \sim N(0, GRM_a\sigma_a^2)$, $G_e \sim N(0, GRM_e\sigma_e^2)$ are vectors of additive and additive epistatic genotypic effects respectively. σ_a^2 is the additive genetic variance. σ_e^2 is the epistatic genetic variance. ε is the model residual. Accordingly, the proportions of additive variance, epistatic variance, and residual variance were estimated using Eq. (15), (16), and (17), respectively.

$$PVE_a = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2 + \sigma_\varepsilon^2} \tag{15}$$

$$PVE_e = \frac{\sigma_e^2}{\sigma_a^2 + \sigma_e^2 + \sigma_\varepsilon^2} \tag{16}$$

$$PVE_\varepsilon = \frac{\sigma_\varepsilon^2}{\sigma_a^2 + \sigma_e^2 + \sigma_\varepsilon^2} \tag{17}$$

In Eq. (15), (16), and (17), σ_a^2, σ_e^2, σ_ε^2 denote the additive genetic variance, epistatic genetic variance, and residual variance, respectively."

(2) It appears that the GA-based optimization and hyperparameter tuning were conducted globally before model evaluation, not within a nested cross-validation. If so, the reported improvements in PCC are statistically inflated and cannot be interpreted as generalization performance. There is no indication that environmental or population structure stratification was maintained during fold splitting, which is critical in genomic data.

Response: Thank you for the suggestion. We carefully checked the experiments and codes, with particular attention to the data partitioning strategy used during the genetic algorithm optimization and hyperparameter tuning procedures. We confirmed that the same sample split used during model training was strictly maintained throughout the genetic algorithm optimization and hyperparameter tuning processes, thereby avoiding any potential data leakage. In addition, all comparison methods were evaluated using the same data partition, ensuring that all methods were compared under the same experimental settings. Therefore, the performance improvement

reported in this study reflects a fair comparison rather than an artifact caused by data leakage. We have added a description of the data partitioning procedure in Section 4.3, as shown in the following:

"During optimization of base learner weights using the genetic algorithm, the sample partitioning strategy from the training phase is strictly followed to avoid data leakage."

In addition, the code of GEG2P has been made publicly available on GitHub (https://github.com/Deep-Breeding/GEG2P) to further enhance the transparency and reproducibility of the experimental results.

According to your suggestion, we further added experiments to evaluate the performance of GEG2P using a data partitioning scheme stratified by population structure. According to Liu *et al*. (2020), we used Plink and Admixture software to divide CUBIC1404 into a reference population and four sub-populations (L1/L2/L3/L4), as shown in Figure 2.1 and Supplemental Data 3. The reference population includes 1140 materials and the sub-populations of L1, L2, L3, L4 include 67, 64, 63 and 70 materials, respectively. There are large differences in the number of materials of the reference population and sub-populations, and the number of materials of the four sub-populations is about the same. The genetic distances between the four sub-populations and the reference population are L4, L3, L2, and L1, respectively.

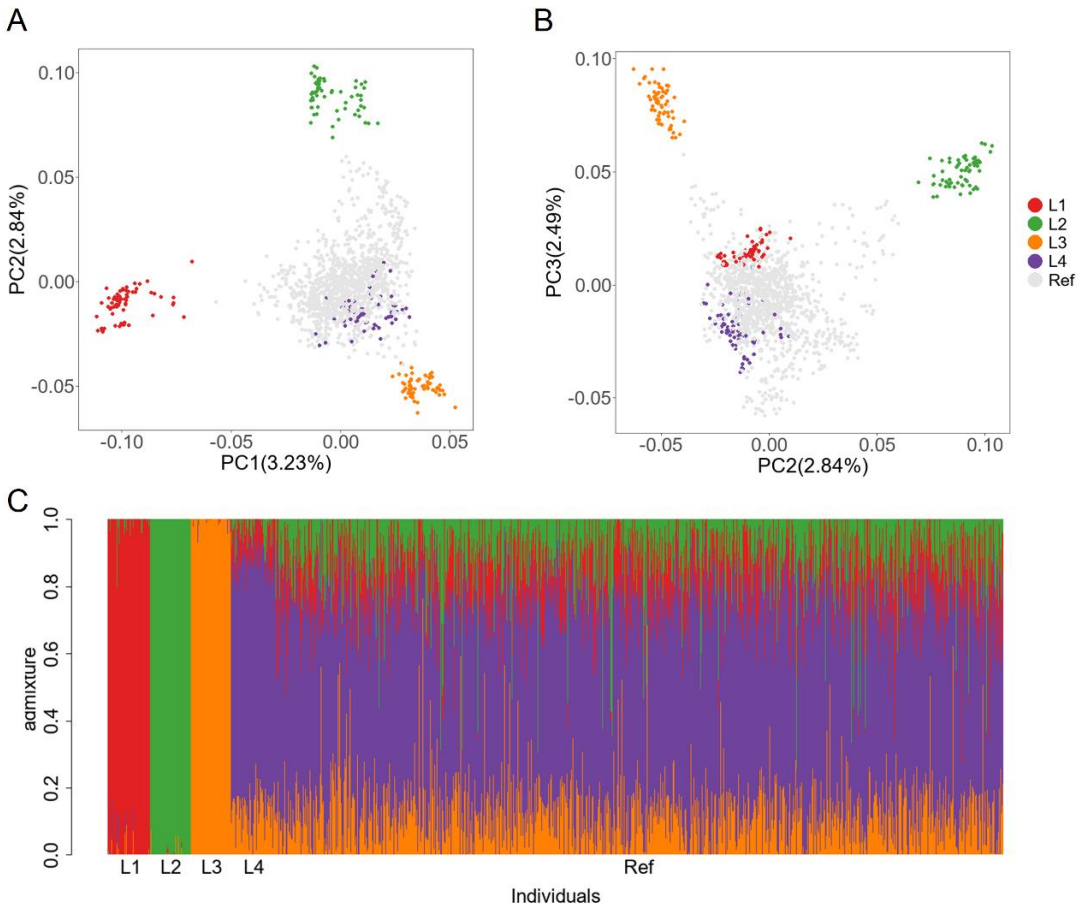

Figure 8. The population genetic structure of CUBIC1404 population. (A, B) The genetic structure of CUBIC1404 population using PCA. (C) The genetic structure of CUBIC1404 population using STRUCTURE. Gray denotes the reference population, red denotes the L1 subpopulation, green denotes the L2 subpopulation, yellow denotes the L3 subpopulation, and purple denotes the L4 subpopulation.

Furthermore, we used each sub-population as a test population and employed GEG2P models constructed from other populations to predict the phenotype of each sub-population. For DTA, PH, and EW, the prediction accuracy of the sub-populations of L4, L3, L2, and L1 was overall positively correlated with the distance of the subpopulation with the reference population (Supplemental Data 3). In other words, prediction accuracy decreased as the distance increased. In addition, we used rrBLUP to perform the same experiments, and the result was consistent with that of GEG2P. The results showed that GEG2P achieved higher prediction accuracy than rrBLUP for all traits, indicating its strong robustness. We have added the result of dividing data according to population structure in Supplemental Information, as shown in the following:

"We further evaluated the performance of GEG2P under a data partitioning scheme stratified

by population structure, using rrBLUP as the comparison method. Following Liu *et al*. (2020), we divided the CUBIC population into one reference population (Reference) and four subpopulations (L1, L2, L3, and L4), containing 1,140, 67, 64, 63, and 70 lines, respectively. The genetic distances between the four subpopulations and the reference population, ranked from greatest to smallest, were L4, L3, L2, and L1. Each subpopulation was sequentially used as the test set, while the remaining populations were used as the training set to evaluate the prediction accuracy of GEG2P and rrBLUP for DTA, PH, and EW (Supplemental Data 3). The results showed that prediction accuracy of the L4, L3, L2, and L1 subpopulations was positively correlated with their genetic relatedness to the reference population. In addition, for all traits, GEG2P achieved higher prediction accuracy than rrBLUP and showed stronger stability."

Supplemental Data 3. The prediction accuracy of GEG2P and rrBLUP in the CUBIC1404 population.

Method	**Phenotype**	**Reference**	**L4**	**L3**	**L2**	**L1**
GEG2P	DTA	0.2030	0.5817	0.3620	0.2589	0.5299
	EW	0.1987	0.0609	0.1922	0.5037	0.5621
	PH	0.3650	0.6191	0.4071	0.5592	0.6488
rrBLUP	DTA	0.1594	0.4892	0.3090	0.2120	0.3946
	EW	0.1527	-0.035	0.1424	0.4901	0.5386
	PH	0.3143	0.6033	0.4029	0.4887	0.6133

In addition, the key references discussed above are listed, as shown in the following:

[1] Liu H, Wang X, Xiao Y, *et al*. (2020). CUBIC: an atlas of genetic architecture promises directed maize improvement. *Genome Biology*, 21(1): 20.

(3) The authors report only Pearson correlation coefficients (PCC), which are insufficient and can exaggerate apparent gains when the variance in the target variable differs among traits. Provide RMSE, MAE, and R^2 adjusted for trait heritability. Quantify uncertainty (SD or SE) across CV folds and use paired statistical tests to demonstrate the significance of differences vs. baselines. Report heritability (h^2) for each trait and contextualize results relative to expected genetic

variance.

Response: Thank you for your suggestion. We added experiments to evaluate the performance of GEG2P and six SOTA methods across different traits and species using three additional evaluation metrics, including RMSE, MAE, and R^2, as shown in Supplemental Tables S3-S5. Specifically, the average RMSE of GEG2P across the five species was 67.43%, 88.34%, 35.26%, 55.91%, 4.09%, and 4.08% lower than that of Cropformer, WheatGP, SoyDNGP, EGGPT, RKHS, and Multi-kernel GBLUP, respectively, indicating that GEG2P achieved lower RMSE. For MAE, GEG2P also showed the best performance, with an average MAE across the five species that was 72.32%, 90.72%, 37.86%, 1.83%, 3.62%, and 3.95% lower than that of these six methods, respectively. In addition, the average R^2 of GEG2P was higher than that of all six comparison methods.

According to your suggestion, we further added experiments using the standard error (SE) to quantify the uncertainty across different folds in the 10-fold cross-validation, and the results have been added to Supplemental Tables S3-S5. We have also added the relevant description in Section 2.5.

Additionally, we have added experiments to statistically compare the prediction performance of GEG2P with six state-of-the-art (SOTA) methods across multiple species and traits. For the 23 maize traits, GEG2P achieved significantly higher 10-fold cross-validated PCCs than SoyDNGP, Cropformer, WheatGP, RKHS, and Multi-kernel GBLUP (Figure S2). For the remaining four species (13 phenotypes × 6 SOTA methods, for a total of 78 comparisons), GEG2P exhibited significantly higher 10-fold cross-validated PCCs in more than half of the comparisons (Supplemental Table S2). These results indicate that GEG2P consistently outperforms the SOTA methods in predicting diverse traits across multiple species. The statistical analysis of GEG2P and SOTA performance has been added to Section 2.5, as shown in the following:

“In maize, the average prediction accuracy of GEG2P is 111.26%, 11.03%, 77.25%, 4.18%, 5.94%, and 5.52% higher than that of SoyDNGP, CropFormer, WheatGP, EGGPT, RKHS, and Multi-kernel GBLUP, respectively (Figure 5A). All improvements are statistically significant except for the comparison with EGGPT (Figure S2).”

“For phenotypes in wheat, rice, chickpea, and soybean, we further conducted statistical comparisons of prediction accuracy between GEG2P and SOTA methods. Among all 78

combinations derived from 13 phenotypes and six methods, GEG2P achieved significantly higher 10-fold cross-validated PCC than SOTA methods in more than half of the comparisons (Supplemental Table S2).”

Furthermore, we have added experiments to evaluate the correlations between prediction metrics of GEG2P and narrow-sense heritability of the phenotypes. The results showed that PCC and R^2 across all 36 phenotypes were significantly correlated with narrow-sense heritability (*P*-values = 2.9×10^{-7} and 2.13×10^{-6}), with correlation coefficients of 0.737 and 0.699, respectively (Figure S5A~B). In contrast, RMSE and MAE exhibited no significant correlation with heritability (*P*-values = 0.78 and 0.77, respectively) (Figure S5C~D). This can be attributed to RMSE and MAE primarily capturing the numerical deviation between predicted and observed values, rather than genetic effects. The correlation analysis between the predictive metrics of GEG2P and narrow-sense heritability has been added to Section 2.5, as shown in the following:

“Moreover, correlations between narrow-sense heritability (Figure S4) and all evaluated prediction metrics of GEG2P were assessed. PCC and R^2 across all 36 phenotypes showed significantly correlations with narrow-sense heritability (*P*-values = 2.9×10^{-7} and 2.13×10^{-6}), with correlation coefficients of 0.737 and 0.699, respectively (Figure S5A~B). In contrast, RMSE and MAE showed no significant correlations with narrow-sense heritability (Figure S5C~D).”

We sincerely appreciate your valuable comments on this study, which have enhanced the comprehensiveness and rigor of the experimental results.

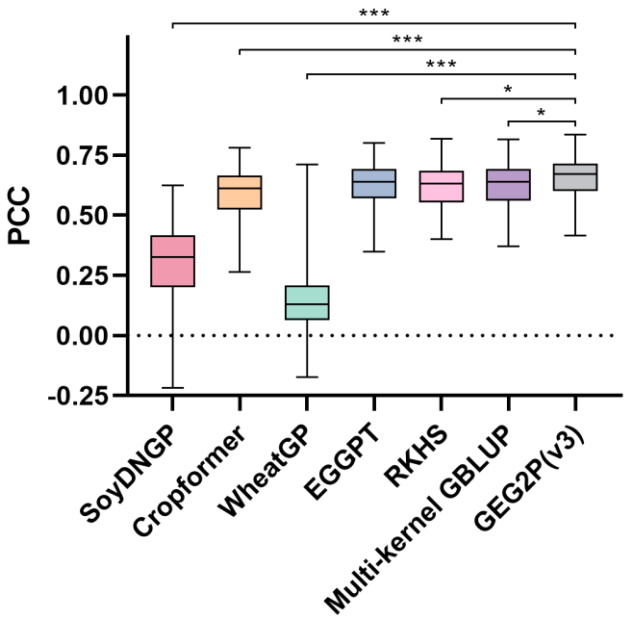

Figure S2. Statistical comparison results of prediction accuracy between GEG2P and six SOTA methods across 23 maize traits. Asterisks indicate significance (* *P*-value < 0.05; ** *P*-value < 0.01; *** *P*-value < 0.001; *t*-test).

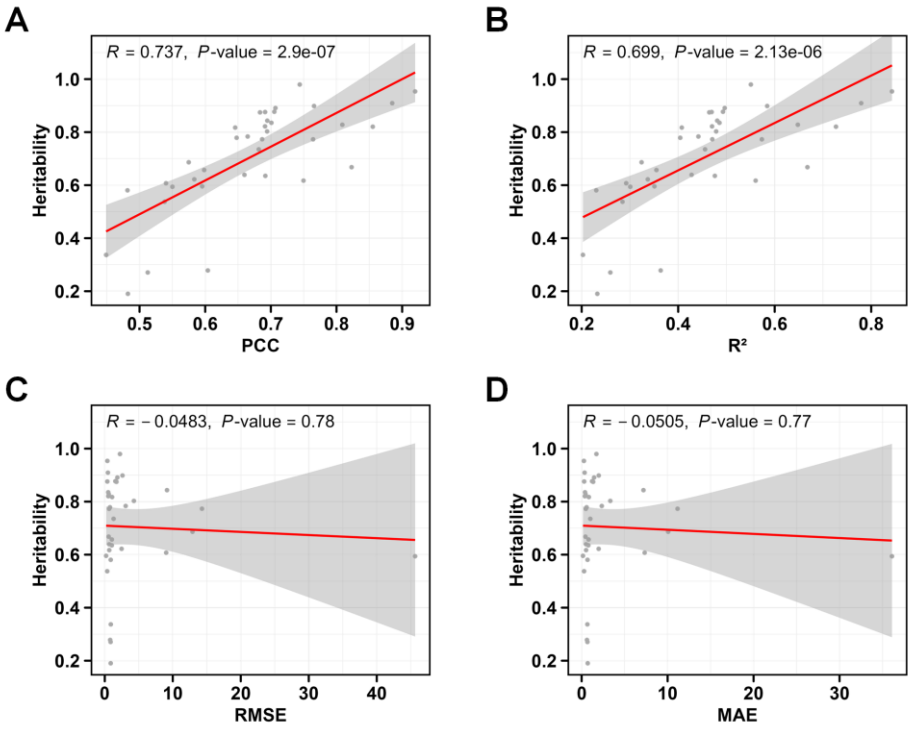

Figure S5. Correlation analysis results between the PCC, R^2, RMSE, and MAE of GEG2P and narrow-sense heritability across 36 traits. Correlation analyses were conducted using Pearson's correlation test. Each point represents one trait.

Supplemental Table S3. RMSE and SE of GEG2P and six SOTA methods across 36 traits in five species

Crop	Trait	SoyDNGP		Cropformer		WheatGP		RKHS		Multi-kernel GBLUP		EGGPT		GEG2P(v3)	
		RMSE	**SE**	**RMSE**	**SE**	**RMSE**	**SE**	**RMSE**	**SE**	**RMSE**	**SE**	**RMSE**	**SE**	**RMSE**	**SE**
Maize	DTA	4.8991	1.3962	20.3718	0.7670	67.2067	1.2907	1.6176	0.0340	1.6254	0.0333	1.6292	0.0455	1.5562	0.0339
	DTS	2.5092	0.1010	20.0833	0.5766	66.9404	1.0946	1.7942	0.0335	1.7956	0.0345	1.7838	0.0453	1.7207	0.0358
	DTT	2.5930	0.0737	19.1178	0.6098	62.8806	0.5599	1.9380	0.0541	1.9322	0.0510	1.9390	0.0451	1.8519	0.0526
	ATI	0.9410	0.0423	1.5003	0.0643	0.7136	0.0232	0.6879	0.0139	0.6875	0.0121	0.6698	0.0210	0.6583	0.0162
	SAI	1.4220	0.0461	1.6182	0.0628	1.0502	0.0389	1.0881	0.0280	1.0813	0.0278	1.0621	0.0207	1.0404	0.0262
	STI	1.9820	0.1396	2.8541	0.1228	1.4062	0.0432	1.3601	0.0290	1.3566	0.0276	1.3250	0.0211	1.2963	0.0257
	EH	16.7767	0.5655	27.0564	1.0873	80.5860	1.3330	9.5448	0.1670	9.4092	0.1803	9.5237	0.2002	9.1542	0.1668
	ELL	9.2322	2.2093	23.8750	0.8467	65.7323	1.3028	4.4724	0.0825	4.3925	0.0859	4.3890	0.0708	4.2871	0.0831
	ELW	0.8695	0.0229	3.3839	0.2035	0.8363	0.0177	0.6282	0.0148	0.6275	0.0150	0.6147	0.0100	0.5966	0.0153
	LNAE	0.5682	0.0107	2.2956	0.1446	0.4949	0.0127	0.3946	0.0071	0.3941	0.0080	0.3862	0.0076	0.3797	0.0090
	LNBE	1.1092	0.0933	2.9660	0.1075	0.5991	0.0142	0.5258	0.0114	0.5191	0.0115	0.5211	0.0124	0.5039	0.0121
	PH	28.9476	3.7135	67.6098	0.7670	198.5436	1.7890	15.0959	0.4366	14.8242	0.3507	14.8349	0.2295	14.2929	0.3587
	TBN	3.7509	0.1998	5.0274	0.2768	3.1146	0.0816	2.3083	0.0772	2.3558	0.0712	2.3316	0.0582	2.2343	0.0733
	TL	4.1725	0.0653	12.0505	0.3347	27.9236	0.2157	2.7308	0.0694	2.7346	0.0704	2.7400	0.0731	2.6076	0.0722
	EL	2.4040	0.2247	5.2644	0.2762	8.9468	0.0275	1.1333	0.0214	1.1280	0.0152	1.1121	0.0170	1.0791	0.0169
	ERN	1.4487	0.0322	5.3815	0.1623	9.8867	0.0286	1.0906	0.0280	1.0825	0.0252	1.0704	0.0357	1.0359	0.0263
	EW	19.3036	0.6543	31.0057	0.9724	81.9478	0.2184	13.5539	0.2523	13.6159	0.2338	171.4819	0.3293	12.9066	0.2335

	KNPE	61.7357	2.8098	81.6254	1.7542	307.6776	1.0077	47.1291	1.0807	47.4919	0.9815	45.7470	1.1459	45.6432	1.0377
	KNPR	3.0470	0.0342	8.8618	0.2628	6.3107	0.0197	2.6001	0.0307	2.6098	0.0327	2.4923	0.0564	2.4747	0.0282
	KWPE	12.8832	0.5770	25.8765	0.9312	68.2869	0.1666	9.4836	0.1699	9.4554	0.2279	9.0225	0.2204	9.0317	0.1899
	LBT	0.4598	0.0143	0.5930	0.0353	0.3360	0.0086	0.3780	0.0060	0.3765	0.0061	0.3643	0.0109	0.3647	0.0066
	CW	4.1728	0.0884	8.2806	0.3042	16.3787	0.0677	3.2021	0.0731	3.2044	0.0779	3.1653	0.1084	3.0745	0.0718
	ED	0.3670	0.0238	1.6056	0.0496	0.2094	0.0073	0.2111	0.0039	0.2061	0.0039	0.2044	0.0051	0.1996	0.0036
Chickpea	Days_to_0.5_flowering	0.9073	0.0100	1.0033	0.0211	0.9416	0.0333	0.8760	0.0107	0.8723	0.0104	0.8774	0.0099	0.8578	0.0066
	Plant_height	0.8285	0.0083	0.9699	0.0199	0.8891	0.0101	0.8137	0.0137	0.8117	0.0144	0.8142	0.0119	0.7960	0.0069
	Yield	0.8998	0.0035	1.1185	0.0084	0.9007	0.0252	0.8862	0.0136	0.8838	0.0142	0.8844	0.0164	0.8740	0.0035
Rice	BLUP_Heading_date	0.6059	0.0117	0.5849	0.0295	0.6551	0.0895	0.4141	0.0147	0.4419	0.0152	0.4240	0.0168	0.3864	0.0152
	BLUP_Height	0.5995	0.0141	0.8957	0.1051	0.7975	0.0745	0.4911	0.0111	0.5059	0.0102	0.4807	0.0078	0.4653	0.0100
	BLUP_Yield_per_plant	0.9380	0.0335	0.9861	0.0369	0.9501	0.0321	0.9133	0.0301	0.9118	0.0297	0.8996	0.0314	0.8881	0.0309
Wheat	Maturity	0.7828	0.0154	0.7914	0.0334	0.7253	0.0199	0.6634	0.0060	0.6853	0.0050	0.7444	0.0063	0.5760	0.0104
	Grain_Yield	0.9199	0.0147	0.9454	0.0364	0.9237	0.0234	0.8856	0.0068	0.8976	0.0071	0.9108	0.0073	0.8767	0.0149
	Grain_Size_TKW	0.8006	0.0131	0.7270	0.0383	0.7503	0.0157	0.6825	0.0034	0.6871	0.0039	0.7427	0.0061	0.6626	0.0140
Soybean	BBD_BLUP	0.6858	0.0209	0.9454	0.1087	0.8648	0.0693	0.5633	0.0163	0.5522	0.0144	0.5534	0.0185	0.5187	0.0142
	PLH_BLUP	0.7746	0.0133	0.7441	0.0241	0.7400	0.0669	0.6377	0.0191	0.6370	0.0189	0.6272	0.0156	0.5887	0.0184
	SL_BLUP	0.8985	0.0331	0.8929	0.0257	0.8148	0.0201	0.8211	0.0252	0.8123	0.0225	0.7799	0.0163	0.7635	0.0218
	ST_BLUP	0.7594	0.0303	0.7770	0.0172	0.7921	0.0516	0.6888	0.0236	0.6772	0.0193	0.6954	0.0153	0.6411	0.0201

Supplemental Table S4. MAE and SE of GEG2P and six SOTA methods across 36 traits in five species

Crop	Trait	SoyDNGP		Cropformer		WheatGP		RKHS		Multi-kernel GBLUP		EGGPT		GEG2P(v3)	
		MAE	SE	MAE	SE	MAE	SE	MAE	SE	MAE	SE	MAE	SE	MAE	SE
Maize	DTA	4.5179	1.4312	20.2962	0.7693	67.1680	1.2941	1.2570	0.0251	1.2655	0.0235	1.2817	0.0305	1.2146	0.0233
	DTS	1.9564	0.0948	19.9933	0.5787	66.9024	1.0969	1.4052	0.0273	1.3998	0.0254	1.4017	0.0326	1.3379	0.0322
	DTT	2.0512	0.0569	18.9948	0.6152	62.8272	0.5627	1.5156	0.0380	1.5168	0.0347	1.5182	0.0411	1.4494	0.0331
	ATI	0.7169	0.0329	1.2914	0.0698	0.5503	0.0174	0.5351	0.0120	0.5340	0.0106	0.5152	0.0133	0.5100	0.0135
	SAI	1.0871	0.0360	1.2992	0.0607	0.8204	0.0296	0.8434	0.0203	0.8398	0.0191	0.8093	0.0118	0.8048	0.0172
	STI	1.5444	0.1301	2.4260	0.1293	1.0775	0.0347	1.0593	0.0249	1.0616	0.0246	1.0210	0.0157	1.0094	0.0200
	EH	13.7120	0.5101	25.0040	1.2058	79.4644	1.4253	7.4337	0.1185	7.3154	0.1263	7.4143	0.1525	7.1925	0.1301
	ELL	8.0724	2.2577	23.3417	0.8546	65.4168	1.3152	3.4535	0.0638	3.4176	0.0626	3.3959	0.0676	3.3351	0.0525
	ELW	0.6862	0.0187	3.2977	0.2083	0.6276	0.0151	0.4817	0.0111	0.4795	0.0121	0.4628	0.0081	0.4525	0.0118
	LNAE	0.4515	0.0110	2.2467	0.1463	0.3948	0.0122	0.3085	0.0062	0.3102	0.0065	0.3031	0.0070	0.2978	0.0076
	LNBE	0.9593	0.0922	2.8985	0.1103	0.4662	0.0123	0.4083	0.0091	0.4028	0.0084	0.4031	0.0088	0.3905	0.0095
	PH	24.9444	3.8055	65.4165	0.7134	197.5275	1.8212	11.7573	0.3922	11.4967	0.3395	11.3846	0.1678	11.1584	0.3254
	TBN	2.9094	0.1712	4.3301	0.2875	2.4644	0.0618	1.7758	0.0408	1.8134	0.0353	1.7691	0.0280	1.7182	0.0374
	TL	3.2445	0.0576	11.5870	0.3345	27.6674	0.2195	2.0841	0.0432	2.0846	0.0433	2.1809	0.0540	1.9900	0.0498
	EL	2.0926	0.2306	5.0933	0.2869	8.8390	0.0191	0.8859	0.0121	0.8744	0.0110	0.8860	0.0113	0.8389	0.0106
	ERN	1.0957	0.0284	5.2161	0.1681	9.7988	0.0245	0.8176	0.0205	0.8113	0.0184	0.8085	0.0211	0.7804	0.0176
	EW	15.5263	0.5675	27.6816	0.9690	80.4617	0.2234	10.5252	0.1837	10.6830	0.1929	10.2143	0.2679	10.0671	0.1658
	KNPE	49.7186	2.3857	68.9353	1.7360	302.9557	0.6687	37.0323	0.8828	37.4786	0.8416	36.1283	1.0277	36.1005	0.8777

	KNPR	2.4307	0.0340	8.2898	0.2743	5.6048	0.0252	2.0775	0.0300	2.0962	0.0286	2.0014	0.0480	1.9811	0.0260
	KWPE	10.6213	0.5196	23.5884	0.9464	67.5668	0.1546	7.6199	0.1348	7.7005	0.1713	7.2555	0.1908	7.3335	0.1398
	LBT	0.3566	0.0089	0.4977	0.0331	0.2529	0.0081	0.2945	0.0044	0.2946	0.0044	0.2861	0.0093	0.2867	0.0054
	CW	3.2225	0.0659	7.4410	0.3428	15.8796	0.0563	2.4321	0.0462	2.4410	0.0495	2.3911	0.0807	2.3433	0.0503
	ED	0.3119	0.0246	1.5891	0.0498	0.1725	0.0057	0.1628	0.0031	0.1598	0.0031	0.1569	0.0032	0.1549	0.0029
Chickpea	Days_to_0.5_flowering	0.7101	0.0082	0.7963	0.0201	0.7405	0.0349	0.6896	0.0112	0.6852	0.0111	0.7112	0.0105	0.6737	0.0053
	Plant_height	0.6445	0.0056	0.7683	0.0166	0.7002	0.0066	0.6335	0.0083	0.6316	0.0092	0.6290	0.0084	0.6196	0.0038
	Yield	0.7289	0.0042	0.8971	0.0077	0.7278	0.0229	0.7200	0.0124	0.7157	0.0120	0.7714	0.0116	0.7084	0.0036
Rice	BLUP_Heading_date	0.4305	0.0075	0.4405	0.0216	0.4822	0.0727	0.2824	0.0086	0.3065	0.0093	0.2890	0.0086	0.2665	0.0084
	BLUP_Height	0.4580	0.0094	0.7479	0.1037	0.6462	0.0682	0.3638	0.0076	0.3771	0.0085	0.3592	0.0068	0.3470	0.0076
	BLUP_Yield_per_plant	0.6984	0.0184	0.7622	0.0266	0.7202	0.0230	0.6872	0.0167	0.6849	0.0171	0.6898	0.0203	0.6681	0.0194
Wheat	Maturity	0.6393	0.0133	0.6142	0.0309	0.5561	0.0190	0.5072	0.0045	0.5314	0.0038	0.6095	0.0037	0.4416	0.0110
	Grain_Yield	0.7208	0.0098	0.7433	0.0336	0.7240	0.0254	0.6929	0.0043	0.7039	0.0039	0.7308	0.0058	0.6865	0.0100
	Grain_Size_TKW	0.6355	0.0125	0.5734	0.0308	0.5846	0.0133	0.5341	0.0040	0.5386	0.0044	0.6245	0.0054	0.5194	0.0124
Soybean	BBD_BLUP	0.4930	0.0119	0.7727	0.1109	0.6656	0.0608	0.4009	0.0107	0.3965	0.0099	0.4168	0.0102	0.3709	0.0085
	PLH_BLUP	0.6053	0.0119	0.5808	0.0204	0.5815	0.0607	0.4893	0.0153	0.4892	0.0136	0.4771	0.0090	0.4512	0.0133
	SL_BLUP	0.6748	0.0181	0.6951	0.0181	0.6250	0.0129	0.6171	0.0135	0.6156	0.0122	0.6036	0.0118	0.5799	0.0122
	ST_BLUP	0.5727	0.0189	0.5962	0.0124	0.5992	0.0370	0.5183	0.0151	0.5102	0.0126	0.5189	0.0094	0.4829	0.0124

Supplemental Table S5. R^2 and SE of GEG2P and six SOTA methods across 36 traits in five species

Crop	Trait	SoyDNGP		Cropformer		WheatGP		RKHS		Multi-kernel GBLUP		EGGPT		GEG2P(v3)	
		R^2	SE	R^2	SE	R^2	SE	R^2	SE	R^2	SE	R^2	SE	R^2	SE
Maize	DTA	-7.5263	3.8486	-86.8896	7.0496	-866.8021	16.4578	0.4510	0.0291	0.4455	0.0296	0.4504	0.0209	0.4928	0.0249
	DTS	-0.1750	0.1416	-73.0731	5.2954	-881.7530	13.5014	0.4171	0.0181	0.4162	0.0183	0.4253	0.0151	0.4644	0.0160
	DTT	0.0256	0.0095	-53.6376	4.4105	-587.6933	6.8118	0.4485	0.0298	0.4510	0.0302	0.4601	0.0140	0.4964	0.0272
	ATI	-0.1690	0.0686	-2.0567	0.3017	0.4460	0.0079	0.3735	0.0187	0.3726	0.0220	0.4159	0.0250	0.4278	0.0146
	SAI	-0.2128	0.0678	-0.6366	0.1832	0.4351	0.0049	0.2939	0.0233	0.3031	0.0206	0.3370	0.0217	0.3549	0.0185
	STI	-0.2831	0.1511	-1.6556	0.2010	0.4306	0.0034	0.4011	0.0251	0.4040	0.0255	0.4375	0.0240	0.4557	0.0231
	EH	-0.7622	0.1291	-3.6587	0.4715	-35.1764	1.0399	0.4353	0.0191	0.4514	0.0191	0.4327	0.0282	0.4810	0.0170
	ELL	-2.5513	1.7228	-15.3112	1.2719	-103.4377	2.4662	0.4310	0.0259	0.4517	0.0249	0.4559	0.0135	0.4787	0.0210
	ELW	-0.1212	0.0268	-16.6779	2.1075	0.0486	0.0035	0.4112	0.0239	0.4129	0.0228	0.4378	0.0178	0.4702	0.0191
	LNAE	-0.1918	0.0388	-19.4345	2.7400	0.1293	0.0047	0.4257	0.0160	0.4276	0.0157	0.4492	0.0237	0.4693	0.0151
	LNBE	-1.6913	0.4385	-16.9730	1.3006	0.1926	0.0078	0.4389	0.0264	0.4533	0.0254	0.4539	0.0135	0.4856	0.0231
	PH	-1.5074	0.7103	-10.8632	0.4487	-96.9547	3.4764	0.4084	0.0320	0.4297	0.0278	0.4222	0.0363	0.4707	0.0242
	TBN	-0.2819	0.1211	-1.3683	0.3106	0.0336	0.0039	0.5199	0.0240	0.5000	0.0229	0.5005	0.0260	0.5506	0.0211
	TL	-0.0625	0.0372	-7.9154	0.5180	-53.7475	0.1465	0.5442	0.0208	0.5433	0.0202	0.5395	0.0321	0.5845	0.0198
	EL	-2.0886	0.5146	-13.6824	1.6649	-40.7291	0.0023	0.3459	0.0190	0.3520	0.0145	0.3644	0.0236	0.4075	0.0122
	ERN	-0.0141	0.0202	-13.3518	1.0968	-55.4557	0.0042	0.4202	0.0328	0.4283	0.0318	0.4464	0.0183	0.4760	0.0319
	EW	-0.5136	0.0862	-2.9030	0.2162	-26.8243	0.0017	0.2541	0.0319	0.2482	0.0289	0.3110	0.0200	0.3241	0.0272
	KNPE	-0.2892	0.1017	-1.2478	0.1028	-31.8317	0.0051	0.2545	0.0235	0.2428	0.0226	0.2976	0.0193	0.3007	0.0221

	KNPR	-0.0038	0.0150	-7.5431	0.4637	-3.4429	0.0041	0.2680	0.0178	0.2627	0.0174	0.3241	0.0210	0.3370	0.0155
	KWPE	-0.4613	0.1225	-4.8934	0.4584	-46.6619	0.0059	0.2181	0.0300	0.2245	0.0306	0.2963	0.0177	0.2923	0.0255
	LBT	-0.1367	0.0513	-0.9844	0.2595	0.3805	0.0330	0.2317	0.0207	0.2369	0.0247	0.2825	0.0279	0.2846	0.0218
	CW	-0.0175	0.0058	-3.1199	0.3874	-15.6393	0.0033	0.3945	0.0314	0.3919	0.0360	0.4139	0.0162	0.4419	0.0290
	ED	-1.3229	0.3300	-41.2988	2.6686	0.3454	0.0477	0.2732	0.0273	0.3074	0.0255	0.3241	0.0207	0.3508	0.0221
Chickpea	Days_to_0.5_flowering	0.1724	0.0107	-0.0325	0.0344	0.1052	0.0525	0.2278	0.0103	0.2341	0.0114	0.2276	0.0101	0.2594	0.0051
	Plant_height	0.3108	0.0057	0.0396	0.0367	0.1991	0.0180	0.3348	0.0150	0.3381	0.0158	0.3340	0.0107	0.3637	0.0060
	Yield	0.1867	0.0070	-0.2727	0.0139	0.1848	0.0483	0.2113	0.0116	0.2153	0.0139	0.2148	0.0184	0.2324	0.0043
Rice	BLUP_Heading_date	0.6250	0.0148	0.6431	0.0360	0.4753	0.1355	0.8209	0.0158	0.7958	0.0181	0.8153	0.0125	0.8431	0.0152
	BLUP_Height	0.6355	0.0151	0.0563	0.2805	0.2896	0.1211	0.7541	0.0130	0.7397	0.0112	0.7631	0.0109	0.7792	0.0113
	BLUP_Yield_per_plant	0.1117	0.0188	0.0173	0.0292	0.0853	0.0279	0.1565	0.0162	0.1588	0.0181	0.1804	0.0200	0.2027	0.0189
Wheat	Maturity	0.3864	0.0121	0.3691	0.0919	0.4655	0.0337	0.5594	0.0051	0.5297	0.0047	0.4445	0.0074	0.6679	0.0109
	Grain_Yield	0.1527	0.0146	0.1022	0.0816	0.1446	0.0465	0.2144	0.0066	0.1930	0.0055	0.1685	0.0092	0.2302	0.0155
	Grain_Size_TKW	0.3582	0.0126	0.4701	0.0919	0.4283	0.0178	0.5337	0.0044	0.5274	0.0047	0.4473	0.0069	0.5605	0.0149
Soybean	BBD_BLUP	0.5246	0.0104	-0.0128	0.2685	0.1997	0.1060	0.6784	0.0095	0.6904	0.0098	0.6848	0.0213	0.7270	0.0080
	PLH_BLUP	0.3884	0.0259	0.4316	0.0407	0.4185	0.0978	0.5864	0.0217	0.5880	0.0199	0.6011	0.0137	0.6480	0.0183
	SL_BLUP	0.1775	0.0221	0.1837	0.0234	0.3151	0.0313	0.3119	0.0123	0.3232	0.0228	0.3730	0.0292	0.4043	0.0122
	ST_BLUP	0.4062	0.0215	0.3642	0.0408	0.3333	0.0795	0.5077	0.0236	0.5235	0.0201	0.5089	0.0165	0.5734	0.0186

(4) Algorithmic Transparency and Reproducibility. The genetic algorithm parameters (population size, crossover/mutation rates, selection scheme, and stopping rule) are not well described. The three-step iterative optimization is vaguely defined, lacking pseudocode, convergence analysis, and justification for the number of iterations. Computational requirements and open-source code are lacking.

Response: Thank you for your suggestion. First, we have provided a clearer description of the key parameters used in the genetic algorithm. In GEG2P, the genetic algorithm was used to optimize the weights of the base learners, with a population size of 50, a crossover probability of 0.7, and a mutation probability of 0.2. For the selection operation, we adopted a tournament selection strategy. The number of generations was set to 40, and the algorithm terminated after reaching this limit. These parameter settings have been added to Supplemental Data 26 and are also described in Section 4.2.2 of the manuscript.

Supplemental Data 26. Key parameters of the genetic algorithm

Parameter Name	**Value**
Population size	50
Cross Probability	0.7
Mutation Probability	0.2
Evolution Rounds	40
Select Method	Tournament

Second, the three-step iterative optimization is a core algorithmic design of GEG2P for identifying the optimal combination of base learners, rather than a tunable hyperparameter. This strategy consists of three sequential optimization stages. In the first step, base learners from different categories with complementary feature extraction capabilities are directly combined to overcome category boundaries. In the second step, the models generated using different ensemble strategies are further integrated as new base learners. In the third step, base learners with lower PCC are gradually removed to retain those that contribute effectively to prediction performance. Through this three-step iterative optimization strategy, GEG2P can overcome the category boundaries of base learners, further increase the diversity of base learners, and identify the optimal combination of base learners dynamically. According to your suggestion, we have added the

pseudocode for this three-step iterative optimization strategy in Supplemental Information. In addition, the code for GEG2P has been made publicly available on GitHub (https://github.com/Deep-Breeding/GEG2P).

Algorithm 1. Three-step iterative optimization algorithm of GEG2P

Input: Base Learner Set *B*, GenoType Data *G*, Trait Name *T*, Genomic Algorithm Parameters *C*, Phenotype Data *P*, Base Learner Labels *L*

Output: GEG2Pv3Model *M*, Saved Base Learners *S*, Prediction Values *R*, Weights of Base Learners *W*

 # Step 1: Constructing GEG2P(v1) to break category boundary.

BLModels ← TrainBaseLearners(*B*, *G*, *P*, *T*)
GEG2Pv1Model, BLWeights ← CreateGEG2PwithGA(BLModels, *G*, *P*, *T*, *C*)
Predv1 ← EnsemblePredictions(GEG2Pv1Model, BLWeights, *G*, *P*, *T*)

 # Step 2: Constructing GEG2P(v2) to Use composite base learners.

DL ← SelectDeepLearningBL(*B*, *L*)
ML ← SelectMachineLearningBL(*B*, *L*)
SS ← SelectStatisticalBL(*B*, *L*)
DLModels ← TrainBaseLearners(DL, *G*, *P*, *T*)
GEG2PDLModel, DLWeights ← CreateGEG2PwithGA(DLModels, *G*, *P*, *T*, *C*)
MLModels ← TrainBaseLearners(ML, *G*, *P*, *T*)
GEG2PMLModel, MLWeights ← CreateGEG2PwithGA(MLModels, *G*, *P*, *T*, *C*)
SSModels ← TrainBaseLearners(SS, *G*, *P*, *T*)
GEG2PSSModel, SSWeights ← CreateGEG2PwithGA(SSModels, *G*, *P*, *T*, *C*)
BLModelsv2 ← { BLModels, GEG2Pv1Model, GEG2PDLModel, GEG2PMLModel, GEG2PSSModel}
GEG2Pv2Model, BLWeightsv2 ← CreateGEG2PwithGA(BLModelsv2, *G*, *P*, *T*, *C*)
Predv2 ← EnsemblePredictions(GEG2Pv2Model, BLWeightsv2, *G*, *P*, *T*)

 # Step 3: Constructing GEG2P(v2) to remove base learners with poor accuracy.

CandidateBL ← { BLModels, GEG2Pv1Model, GEG2PDLModel, GEG2PMLModel, GEG2PSSModel}
BestScore ← -∞; SavedBL ← CandidateBL; isImproved ← True
while isImproved and |SavedBL| ⩾ 2 **do**
    GEG2Pv3Model, BLWeightsv3 ← CreateGEG2PwithGA(SavedBL, *G*, *P*, *T*, *C*)
    Predv3 ← EnsemblePredictions(GEG2Pv3Model, BLWeightsv3, *G*, *P*, *T*)
    Score ← Metric(Predv3, *P*)

```
        if Score > BestScore then
            BestScore ← Score
            SavedBL ← CandidateBL
        else
            isImproved ← False
            break
        end if
        Contrib ← EvaluateSingleBL(CandidateBL, G, P, T)
        WorstBL ← SelectWorstBL(Contrib)
        CandidateBL ← RemoveBL(SavedBL, WorstBL)
end while
GEG2Pv3Model, BLWeightsv3 ← CreateGEG2PwithGA(SavedBL, G, P, T, C)
Predv3 ← EnsemblePredictions(GEG2Pv3Model, BLWeightsv3, G, P, T)
M ← GEG2Pv3Model; S ← SavedBL;  R ← Predv3;  W ← BLWeightsv3
return M, S, R, W
```

Finally, for computational requirements, GEG2P can be run in a standard server environment without specialized hardware. GEG2P can be deployed in a standard server environment without specialized hardware. To balance resource consumption and runtime efficiency, two versions of GEG2P are available. GEG2P-Serial is suitable for routine applications with relatively low computational cost, whereas GEG2P-Parallel is designed for scenarios requiring higher runtime efficiency. By adopting parallel computing, GEG2P-Parallel substantially reduces runtime, although it requires greater computational resources. We added experiments using the demo dataset to evaluate the computational resource consumption of GEG2P, and the results are shown in Table 4.1. The results have also been made publicly available on GitHub (https://github.com/Deep-Breeding/GEG2P/tree/main/01GEG2P_code).

Table 1. Computing resource requirements for GEG2P

Item	GEG2P-Serial	GEG2P-Parallel
Operating System	Ubuntu 22.04	Ubuntu 22.04
RAM Memory	2354 MB	12.25 GB
GPU Memory	600 MB	4 GB
CPU Core	2 cores	40 cores
Computing Time (s)	1853	914

(5) Benchmarking and Comparative Framework. Although the proposed GEG2P is compared with other methods, these benchmarks are not evaluated under equivalent conditions. Hyperparameter optimization is mentioned but not described in detail. It is unclear whether each model was retrained using identical cross-validation splits. No statistical significance testing or discussion of computational cost vs. predictive gain is provided.

Response: Thank you for your suggestion. We carefully reviewed all comparative experiments. First, we confirmed that all experiments were conducted under the same hardware environment. Second, to address your concern about hyperparameter settings, we reviewed the hyperparameter optimization strategies and search ranges used for all methods. We have summarized the hyperparameter tuning ranges in Supplemental Table S1. These settings allowed all methods to be optimized within a reasonable range. We also reviewed the data partitioning strategy used in cross-validation and confirmed that the same partitioning was applied to all methods. Therefore, all comparisons were conducted under the same conditions. We have added the corresponding description in Section 2.5, as shown in the following:

"The adjustment range of key parameters for all methods can be found in Supplemental Table S1."

"During optimization of base learner weights using the genetic algorithm, the sample partitioning strategy from the training phase is strictly followed to avoid data leakage."

Supplemental Table S1. The tuning range of key parameters.

Method	**Optuna Trial Times**	**Parameter**		**Range**
GEG2P	50	Deep Learning Base Learners	Epoch	20~100
			Early Stop Patience	2~10
			Learning Rate	1e-5~1e-2
			Batch Size	8~64
	20	KNN	n_neighbors	3~15
		Random Forest	max_depth	10~50
			n_estimators	50~200
		XGBoost	max_depth	3~10
			n_estimators	50~200
		SVR	C	1e-3~1e2
		MLP	hidden_size1	10~100

			hidden_size2	10~100
			alpha	1e-5~1e-1
Cropformer	50		Epoch	20~100
			Early Stop Patience	2~10
			Learning Rate	1e-5~1e-2
			Batch Size	8~64
WheatGP	50		Epoch	20~100
			Early Stop Patience	2~10
			Learning Rate	1e-5~1e-2
			Batch Size	8~64
SoyDNGP	50		Epoch	20~100
			Early Stop Patience	2~10
			Learning Rate	1e-5~1e-2
			Batch Size	8~64
RKHS	50		Gemma	1e-3~1e2
			L2	1e-6~1
EGGPT	50	MLP	Learning Rate	1e-5~1e-1
			batch_size	32~512
			hidden_units	64~1024
		Random Forest	min_samples_split	2~11
			n_estimators	50~1000
			min_samples_leaf	1~6
			max_depth	10~50
		SVR	C	5e-2~100
			epsilon	1e-3~10
		LightGBM	min_data_in_leaf	1~6
			n_estimators	50~1000
			max_depth	10~50
			Learning Rate	1e-5~1e-1
		DCNN	C	1e-3~100
			Learning Rate	1e-5~1e-1
			Batch Size	32~512

We added experiments to evaluate the computational resources required by GEG2P-Serial and GEG2P-Parallel. The results are provided in our response to Comment (2).

As shown in the response to Comment 3 in Review 1, we have added experiments to statistically compare the prediction performance of GEG2P with six SOTA methods across

multiple species and traits. The results indicated that, in the majority of comparisons, GEG2P achieved significantly higher 10-fold cross-validated PCCs than the SOTA methods (Figure S2, Supplemental Table S2). The statistical analysis comparing the performance of GEG2P with that of the SOTA methods has been added to Section 2.5, as shown in the following:

"In maize, the average prediction accuracy of GEG2P is 111.26%, 11.03%, 77.25%, 4.18%, 5.94%, and 5.52% higher than that of SoyDNGP, CropFormer, WheatGP, EGGPT, RKHS, and Multi-kernel GBLUP, respectively (Figure 5A). All improvements are statistically significant except for the comparison with EGGPT (Figure S2)."

"For phenotypes in wheat, rice, chickpea, and soybean, we further conducted statistical comparisons of prediction accuracy between GEG2P and SOTA methods. Among all 78 combinations derived from 13 phenotypes and six methods, GEG2P achieved significantly higher 10-fold cross-validated PCC than SOTA methods in more than half of the comparisons (Supplemental Table S2)."

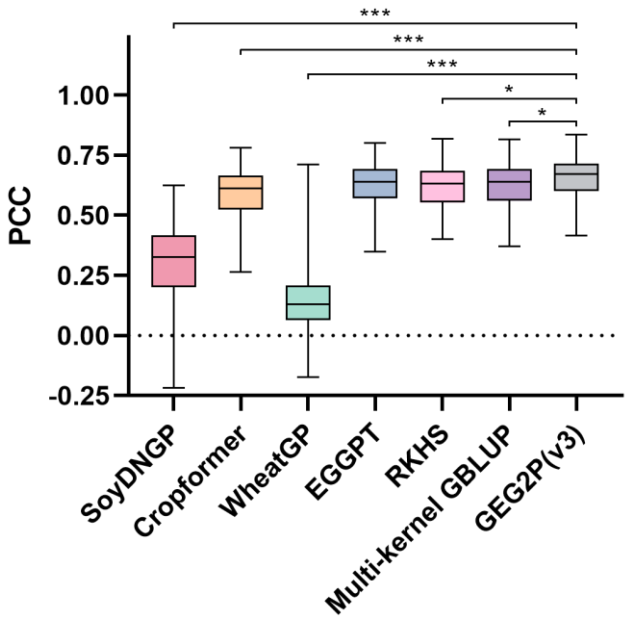

Figure S2. Statistical comparison results of prediction accuracy between GEG2P and six SOTA methods across 23 maize traits. Asterisks indicate significance (* *P*-value < 0.05; ** *P*-value < 0.01; *** *P*-value < 0.001; *t*-test).

(6) Overextended Biological Interpretation. The SHAP-based interpretability analysis is impressive in scope but scientifically unfocused and overinterpreted. Hundreds of genes and pathways are listed without synthesis. The claim that 'GEG2P identifies broader functional genes' is not substantiated with quantitative validation. The mapping of epistatic networks remains purely computational, with no biological validation or replication.

Response: Thank you for your suggestion. After reviewing relevant studies and re-evaluating the evidence, we determined that current results still support the conclusions of this study.

Firstly, a major challenge in current genomic selection (GS) research lies in the interpretability of prediction results, particularly in crop breeding. In this context, explainable artificial intelligence tools have been widely applied in genomic prediction studies. Among them, SHapley Additive exPlanations (SHAP) has shown strong capability for feature interpretation in numerous studies. For example, Wang *et al*. (2024), in a study published in *Nature Communications*, predicted flowering time phenotypes in *Arabidopsis thaliana* using transcriptomic data and applied SHAP to identify key genes regulating flowering time, successfully revealing differential expression patterns among accessions with distinct flowering times. In addition, Wang *et al*. (2025) combined Cropformer with SHAP to interpret potential quantitative trait loci. He *et al*. (2025) applied SHAP to successfully identify key loci with stable regulatory effects on phenotypes across multiple environments. Sun *et al*. (2026) further integrated SHAP with GWAS, addressing limitations of GWAS in capturing nonlinear genetic effects. Based on the strong feature interpretability of SHAP, we used this method for the interpretability analysis in this study. We have further detailed the reason for using SHAP in Section 3.3, as shown in the following:

“However, the mechanisms underlying these improvements in prediction accuracy remain unclear. In particular, the interpretability of prediction results represents a major challenge in current genomic selection (GS) research, especially in the context of crop breeding.

To this end, explainable artificial intelligence tools have been widely applied in genomic prediction studies. Among them, SHapley Additive exPlanations (SHAP) has demonstrated strong capability in feature interpretation across numerous studies. For example, Wang *et al*. (2024) used SHAP to identify key genes regulating flowering time in *Arabidopsis thaliana*, successfully revealing differential expression patterns among accessions with distinct flowering times. Wang *et*

al. (2025b) combined Cropformer with SHAP to interpret potential quantitative trait loci. He *et al*. (2025) applied SHAP to identify key loci with stable regulatory effects on phenotypes across multiple environments. Sun *et al*. (2026) further integrated SHAP with GWAS, addressing limitations of GWAS in capturing nonlinear genetic effects. Collectively, these studies highlight the strong capability of SHAP in feature interpretability."

Second, to validate the association between the candidate genes listed in the manuscript and phenotypes, Top10% SNPs identified by each base learner were compared with GWAS results from the same population (Liu *et al*., 2020). As described in Section 2.6, a subset of Top10% SNPs from each base learner co-localized with phenotype-associated QTLs. GO enrichment analysis results also indicated that the key loci identified by SHAP are functionally associated with phenotypes.

Additionally, the primary objective of this study is to improve the stability of genomic prediction through an ensemble learning approach. Genes identified from Top10% SNPs are regarded as candidate genes with potential relevance to the phenotypes, providing a reference for future investigations. Moreover, due to experimental and resource constraints, functional validation of a large set of candidate genes remains infeasible. This limitation is widely observed in relevant studies. For example, Han *et al*. (2023), in a study published in *Nature Genetics*, integrated multi-omics data to identify 2,651 candidate genes associated with maize flowering time, and subsequently validated 20 of these genes. In this context, Top10% SNPs from each base learner were compared with phenotype-associated cloned genes, and several known functional genes were found to co-locate, thereby reinforcing the biological validity of the findings. We further listed multiple well-known genes associated with the phenotypes that could be mapped by Top10% SNPs in Section 2.6, as shown in the following:

"For example, chr10.s_94433094 (ranked 658 by SHAP) identified by XGBoost is located within *ZmCCT10* (*Zm00001d024909*). Previous studies have reported that *ZmCCT10* regulates the proliferation of archesporial and tapetum cells and plays a key role in early anther development (Li *et al*., 2023)."

"For example, multiple SNPs in Top10% of PH-Weight-Top-5 are located within *Br2* (*Zm00001d031871*), including chr1.s_204750415, chr1.s_204751737, and chr1.s_204752205. *Br2* regulates the polar transport of auxin and thereby affects plant height in maize (Multani *et al*.,

2003). chr6.s_94192268 identified by both SVR and LASSO in EW-Weight-Top-5 is located within *KNR6* (*Zm00001d036602*). Previous studies have reported that *KNR6* regulates floret number per ear and kernel row number in maize, and overexpressing *KNR6* can significantly increase yield (Jia *et al*., 2020)."

The remaining cloned genes mapped by Top10% SNPs identified by each base learner are systematically listed in Supplemental Data 17. Taken together, these results suggest that GEG2P, by integrating multiple base learners, can effectively identify candidate genes associated with phenotypes.

Additionally, loci involved in significant epistatic interaction pairs identified by ANOVA were mapped to genes. We have added experiments that integrate RNA-seq data from the same population (Liu *et al*., 2020) to identify potential transcriptional regulatory relationships, thereby supporting the reliability of the putative epistatic network. Examples of potential regulatory relationships between selected genes have been added to Section 2.6, as shown in the following:

"In addition, integrating RNA-seq data from 391 accessions of the same population (Liu *et al*., 2020) revealed that the expression value of *Zm00001d013683*, mapped by chr5.s_17285732 among the 1,077 interacting SNPs, was significantly correlated with *ZMM15* expression (*P*-value = 3.8×10^{-8}). And earlier flowering was significantly associated with increased expression of *Zm00001d013683* (*P*-value = 1.5×10^{-4}) (Figure S17A). These results provide evidence supporting potential interactions among the Top10% SNPs identified by different base learners from the perspective of gene expression and phenotype association."

"In addition, the expression value of *Zm00001d032283*, mapped by chr1.s_220370448 among the 401 interacting SNPs, was significantly correlated with *BRD1* expression (*P*-value = 3.4×10^{-8}). Moreover, the dwarf phenotype was significantly influenced by the expression value of *BRD1* (*P*-value = 0.017) (Figure S17B). These results reflect potential interactions among genes mapped by the Top10% SNPs."

"Furthermore, the expression value of *Zm00001d012822*, mapped by chr5.s_922298 among the 268 interacting SNPs, was significantly correlated with *ZmACO2* expression (*P*-value = 1.9×10^{-8}). In addition, the expression value of *ZmACO2* showed a significant effect on ear weight (*P*-value = 2.7×10^{-4}) (Figure S17C). These results further provide evidence for potential interactions among the Top10% SNPs identified by different base learners."

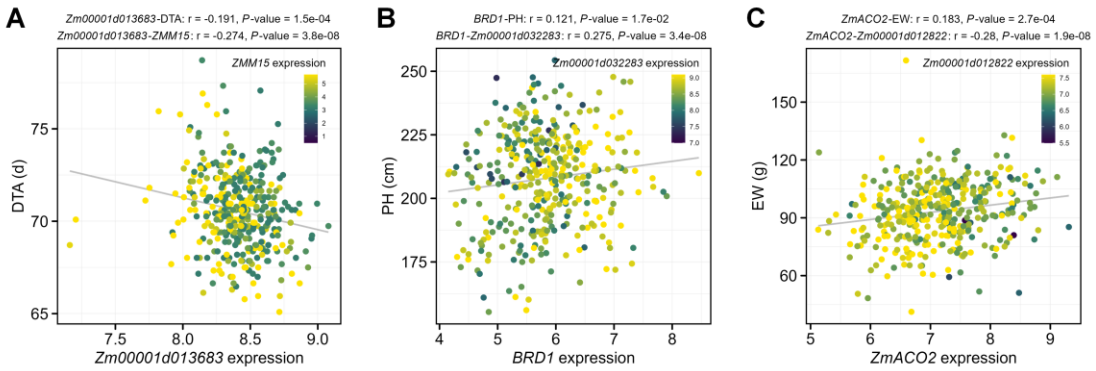

Figure S17. Correlation between expression values of interacting gene pairs and phenotypic variation. (A) Pearson correlation between the expression value of *Zm00001d013683* (*x*-axis) and DTA (*y*-axis). Color scale indicates *ZMM15* expression value. (B) Pearson correlation between the expression value of *BRD1* (*x*-axis) and PH (*y*-axis). Color scale indicates expression value of *Zm00001d032283*. (C) Pearson correlation between the expression value of *ZmACO2* (*x*-axis) and EW (*y*-axis). Color scale indicates expression value of *Zm00001d012822*.

In addition, the key references discussed above are shown in the following:

[1] Wang P, Lehti-Shiu M, Lotreck S, *et al*. (2024). Prediction of plant complex traits via integration of multi-omics data. *Nature Communications*, 15(1): 6856.

[2] Wang H, Yan S, Wang W, *et al*. (2025). Cropformer: An interpretable deep learning framework for crop genomic prediction. *Plant Communications*, 6(3): 101223.

[3] He K, Yu T, Gao S, *et al*. (2025). Leveraging automated machine learning for environmental data-driven genetic analysis and genomic prediction in maize hybrids. *Advanced Science*, 12(17): e2412423.

[4] Sun J, Zhang X, You X, *et al*. (2026). Bayesian neural networks for genomic prediction: uncertainty quantification and SNP interpretation with SHAP and GWAS. *Theoretical and Applied Genetics*, 139(1): 29.

[5] Han L, Zhong W, Qian J, *et al*. (2023). A multi-omics integrative network map of maize. *Nature Genetics*, 55(1): 144-153.

[6] Li B, Wang Z, Jiang H, *et al*. (2023). *ZmCCT10*-relayed photoperiod sensitivity regulates natural variation in the arithmetical formation of male germinal cells in maize. *New Phytologist*, 237(2): 585-600.

[7] Multani D, Briggs S, Chamberlin M, *et al*. (2003). Loss of an MDR transporter in compact

stalks of maize *br2* and sorghum *dw3* mutants. *Science*, 302(5642): 81-84.

[8] Jia H, Li M, Li W, *et al*. (2020). A serine/threonine protein kinase encoding gene *KERNEL NUMBER PER ROW6* regulates maize grain yield. *Nature Communications*, 11(1): 988.

[9] Liu H, Wang X, Xiao Y, *et al*. (2020). CUBIC: an atlas of genetic architecture promises directed maize improvement. *Genome Biology*, 21(1): 20.

(7) Writing, Structure, and Length. Many sections repeat results for different species. Figures are overcrowded, and labels are hard to read. English syntax requires professional editing, and the introduction should be refocused on theoretical advances rather than a literature listing.

Response: Thank you for the suggestion. First, the motivation of GEG2P was to develop a genomic prediction method with stronger generalization ability across different species. We estimated the heritability of different species and traits, and the results are shown in Figure S4. The results showed substantial variation in heritability across species and traits. The prediction performance of GEG2P also varied across species and traits. Therefore, showing results for only a single species would not fully reflect the overall performance and generalization ability of GEG2P. Based on these considerations, we believe that reporting prediction results across multiple species is necessary to accurately demonstrate the applicability and advantages of GEG2P. According to your suggestion, we have further streamlined and integrated the descriptions of the results for different species in Section 2.3, 2.4, and 2.5. We hope this revision makes the presentation clearer and more accurate while avoiding unnecessary repetition, as shown below:

"For flowering-related traits, the average PCC of GEG2P(v1) was higher than that of SVR, GEG2P(DL), GEG2P(ML), and GEG2P(SS), with an average improvement of 3.12% (Figure 3A, Supplemental Data 4). For plant architecture-related traits, the average PCC of GEG2P(v1) was improved by 2.54% on average (Figure 3B, Supplemental Data 4). For yield-related traits, the average PCC of GEG2P(v1) was improved by 4.15% on average (Figure 3C, Supplemental Data 4)."

"For flowering-related traits, GEG2P(v3) achieved an average PCC of 0.6726, representing improvements of 0.16% over GEG2P(v2), and an average improvement of 16.14% over all single base learners (Figure 3A). For plant architecture-related traits, the average PCC of GEG2P(v3)

was 0.7086, which was 0.06% higher than that of GEG2P(v2), and on average 12.41% higher than the average PCC of each single base learner across all traits (Figure 3B). For yield-related traits, the average PCC of GEG2P(v3) was 0.11% higher than those of GEG2P(v2), and on average 17.67% higher than the average PCC of each single base learner across all traits (Figure 3C)."

"In rice, the average PCC of GEG2P(v3) was on average 7.40% higher than that of single base learners. Across the three trait categories, the average PCC of GEG2P(v3) for flowering time-related traits (Figure 4C), plant architecture-related traits (Figure S1, Supplemental Data 5), and yield-related traits (Figure 4D) was on average 3.78%, 4.46%, and 24.72% higher than those of the single base learners, respectively. In chickpea, the average PCC of GEG2P(v3) was on average 8.05% higher than that of the single base learners. For flowering time-related traits (Figure 4E), plant architecture-related traits (Figure S1), and yield-related traits (Figure 4F), the average PCC of GEG2P(v3) was on average 9.24%, 6.39%, and 9.01% higher than those of the single base learners, respectively. In soybean, the average PCC of GEG2P(v3) was on average 8.60% higher than that of the single base learners. For flowering time-related traits (Figure 4G), plant architecture-related traits (Figure S1), and yield-related traits (Figure 4H), the average PCC of GEG2P(v3) was on average 4.40%, 8.92%, and 11.54% higher than those of the single base learners, respectively."

"For flowering time-, plant architecture-, and yield-related traits, the average PCC of GEG2P is at least 1.01%, 3.76%, and 7.15% higher than that of the six comparison methods, respectively."

"Specifically, GEG2P improved the PCC of flowering time-related traits by at least 9.84% (Figure S3A) and yield-related traits by at least 2.84% (Figure S3B)."

"From the perspective of trait categories, GEG2P improved the prediction accuracy of flowering time-related traits by at least 1.26%, plant architecture-related traits by at least 0.84%, and yield-related traits by at least 4.46%, respectively."

"Specifically, GEG2P showed consistent performance gains, with average PCC improvements of at least 4.61%, 3.45%, and 3.23% for flowering time-related, plant architecture-related, and yield-related traits, respectively."

"From the perspective of trait categories, GEG2P improved the prediction accuracy of flowering time-related traits by at least 2.65%, plant architecture-related traits by at least 3.52%, and yield-related traits by at least 6.40%, respectively."

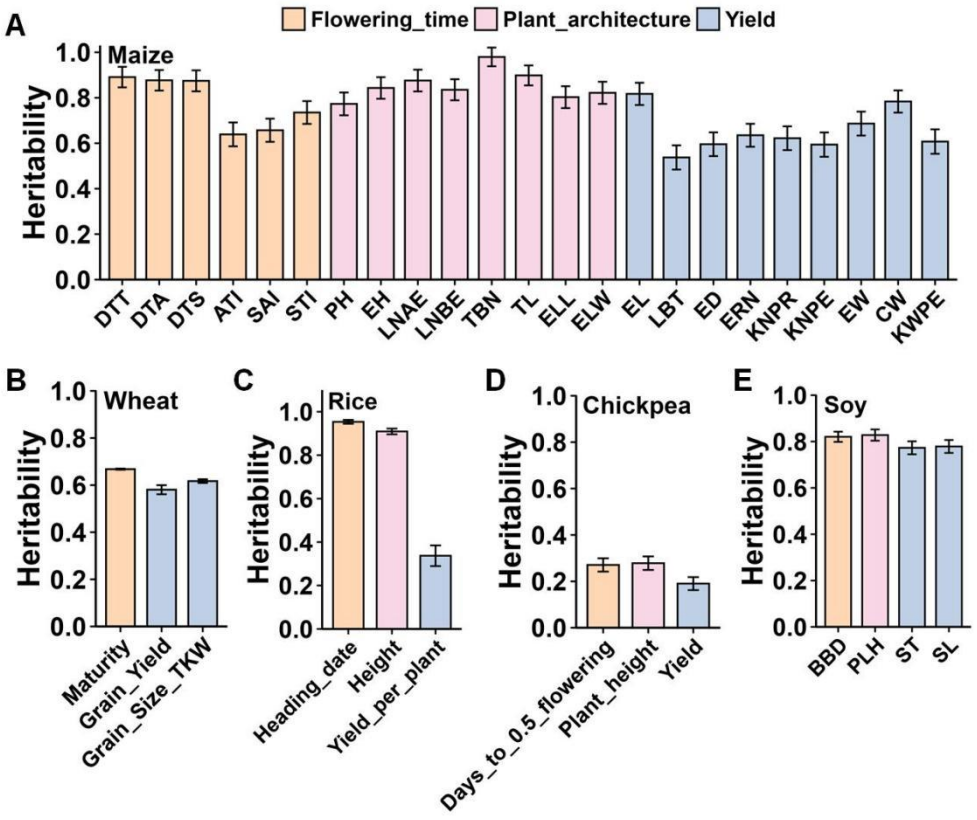

Figure S4. Heritability of 36 traits in maize, wheat, rice, chickpea, and soybean. (A) Heritability of 23 traits related to flowering time, plant architecture, and yield in maize. (B) Heritability of 3 traits related to flowering time and yield in wheat. (C) Heritability of traits related to flowering time, plant architecture, and yield in rice. (D) Heritability of traits related to flowering time, plant architecture, and yield in chickpea. (E) Heritability of 4 traits related to flowering time, plant architecture, and yield in soybean. Error bars indicate the standard errors of the heritability estimation value.

Second, regarding the figures in the manuscript, each figure summarizes model performance by trait category rather than showing the prediction performance for each single trait, in order to limit the number of figures. According to your suggestion, we adjusted the font sizes in the figures to improve their clarity and readability.

In addition, we revised the Introduction according to your suggestions. In Section 1, we further strengthened the theoretical background and highlighted recent advances in deep learning-based genomic prediction to clarify the motivation and innovation of this study, as shown

in the following:

"Deep learning-based models commonly employ convolutional neural networks (CNNs) to extract genotype features (Ma *et al*., 2017; Liu *et al*., 2019). To address the limitation of CNNs in primarily focusing on local patterns, multi-head attention has been introduced to enhance global modeling capability (Wang *et al*., 2024; Ma *et al*., 2024). Meanwhile, researchers have attempted to use multi-omics data for phenotype prediction (Wang *et al*., 2023). In addition, growing attention has been paid to the associations among different phenotypes, motivating the exploration of transfer learning strategies to further improve prediction performance (Li *et al*., 2024b)."

Finally, we carefully reviewed and revised the entire manuscript to improve its grammatical accuracy and overall clarity. We sincerely thank you again for your valuable suggestion.

(8) Impact and Theoretical Value. Despite computational scale, the proposed method does not provide new theoretical insights into genomic prediction. Ensemble learning is not novel; the work primarily scales up existing ideas without clear theoretical advancements. No ablation or sensitivity analysis identifies which components actually drive improvement.

Response: Thanks for your suggestions. We fully understand your concern regarding theoretical insights. Genomic prediction is widely recognized as an interdisciplinary field at the intersection of computer science and quantitative genetics, where methodological innovation is often reflected in the adaptation and optimization of existing machine learning frameworks for complex genetic data and practical breeding tasks. For example, Wang *et al*. (2023) used convolutional neural networks in genomic prediction, thereby improving the model's ability to extract genotype features. Ma *et al*. (2024) adopted LSTM in genomic prediction to enhance the model's ability to capture long-range dependencies. Wang *et al*. (2024) used the random forest algorithm in genomic prediction and improved prediction accuracy. He *et al*. (2025) further improved genomic prediction accuracy by averaging the predictions of four machine learning algorithms. These studies were all built on existing machine learning frameworks, and their innovation lies in addressing practical problems under complex genetic backgrounds.

Consistent with recent advances in genomic prediction, this study was motivated by the practical needs of crop breeding and aimed to improve prediction performance more stably across multiple species and traits. Recent studies have shown that the performance of existing genomic prediction methods is often unstable across species and traits. For example, MultiGS evaluated 24 genomic prediction methods across different crops and traits and showed that their performance varied substantially, which is consistent with our findings (You *et al*., 2026). Given the practical challenges in breeding, ensemble learning provides an effective way to integrate the strengths of different genomic prediction methods and improve prediction accuracy more stably. As a key step in genomic selection, higher prediction accuracy is important for the early identification of superior agronomic traits and can accelerate crop breeding. This study focused on improving genomic prediction accuracy through ensemble learning, with particular attention to the limited stability of current methods across species and traits. To account for genetic differences among species and traits, we designed a three-step iterative optimization strategy to dynamically identify the optimal combination of base learners for each trait. This strategy improves model flexibility and enables stable prediction across different species and traits under complex genetic backgrounds. In addition, we used the genetic algorithm to globally optimize the weights of the base learners, thereby introducing an evolutionary search process into genomic prediction. Although previous studies have also explored ensemble learning, they included only a limited number of base learners and did not systematically investigate ensemble strategies or model interpretability (Wu *et al*., 2024; Zhang *et al*., 2025). In this study, we integrated 20 base learners covering statistical, machine learning, and deep learning methods and systematically evaluated their performance across 36 traits in five species. Overall, we developed GEG2P, a method that achieves stable and accurate prediction across different species and traits under complex genetic backgrounds, thereby providing methodological support for addressing practical breeding challenges using artificial intelligence.

Second, regarding the ablation and sensitivity analyses of GEG2P, we have provided a clearer description of the relevant experimental results in the revised manuscript. In fact, the existing experiments effectively serve as progressive ablation experiments. Specifically, in Section 2.2, we constructed GEG2P(ML), GEG2P(SS), and GEG2P(DL) by integrating base learners according to their categories. The results showed that these models achieved higher prediction accuracy than

single base learners, supporting the effectiveness of the ensemble learning strategy. In Section 2.3, we further compared the prediction performance of GEG2P(v1), GEG2P(v2), and GEG2P(v3), which were progressively constructed using the three-step iterative optimization strategy. These results also supported the effectiveness of the three-step iterative optimization strategy. According to your suggestion, we further improved the presentation of these experiments in Sections 2.2 and 2.3, as shown in the following:

"Therefore, the ensemble strategy implemented in GEG2P achieves higher prediction accuracy than single base learners by integrating deep learning, machine learning, and statistical genomic prediction methods."

"Therefore, the three-step iterative optimization strategy was demonstrated to be effective, leading to consistently improved genomic prediction accuracy of GEG2P across multiple maize traits."

Thank you again for your valuable suggestions. In addition, the key references discussed above are shown in the following:

[1] Wang K, Abid M, Rasheed A, *et al*. (2023). DNNGP, a deep neural network-based method for genomic prediction using multi-omics data in plants. *Molecular Plant*, 16(1): 279-293.

[2] Ma X, Wang H, Wu S, *et al*. (2024). DeepCCR: large-scale genomics-based deep learning method for improving rice breeding. *Plant Biotechnology Journal*, 22(10): 2691-2693.

[3] Wang P, Lehti-Shiu M D, Lotreck S, *et al*. (2024) Prediction of plant complex traits via integration of multi-omics data. *Nature Communications*, 15(1): 6856.

[4] He K, Yu T, Gao S, *et al*. (2025) Leveraging automated machine learning for environmental data‐driven genetic analysis and genomic prediction in maize hybrids. *Advanced Science*, 12(17): 2412423.

[5] You F, Zheng C, Daniel J J Z, et al. (2026). MultiGS: A comprehensive and user-friendly genomic prediction platform integrating statistical, machine learning, and deep learning models for breeders. *Crop Breed Genet Genom*, 8(1):e260004.

[6] Wu J, Wang Y, Yu R, et al. (2024). EGGPT: an extensible and growing genomic predi ction technology. Preprint in Research Square. Available at: https://doi.org/10.21203/rs.3.rs-4581596/v1.

[7] Zhang C, Liang Q, Yu Y, *et al*. (2025). MFMGP: an integrated machine learning fusion model

for genomic prediction. *Plant Biotechnology Journal*, 2025, 23(3): 712.

Minor and Editorial Points

1. Abstract: summarize quantitative improvements once and avoid repetition.

Response: Thanks for your suggestions. According to your suggestion, we have revised the abstract to summarize quantitative improvements once, as shown in the following:

"Compared with the single base learners, GEG2P improved the average prediction accuracy of multiple phenotypes by 10.79% on average across maize, wheat, rice, chickpea, and soybean."

2. Clarify Figure 1 (GEG2P v1–v3 distinctions).

Response: Thank you for your suggestions. GEG2P(v1), GEG2P(v2), and GEG2P(v3) denote the GEG2P models constructed after applying the first, second, and third steps of the iterative optimization strategy, respectively. According to your suggestion, we have provided a clearer description of how GEG2P(v1), GEG2P(v2), and GEG2P(v3) were constructed using the three-step iterative optimization strategy in Section 2.1, as shown in the following:

"In the first step, we broke the category boundaries among base learners to enhance feature extraction capability and constructed the initial ensemble model GEG2P(v1). In the second step, four intermediate ensemble models including GEG2P(DL), GEG2P(ML), GEG2P(SS), and GEG2P(v1) were added to construct GEG2P(v2), which enhances the diversity of base learners. Among them, GEG2P(DL), GEG2P(ML), and GEG2P(SS) were obtained from deep learning, machine learning, and statistical methods, respectively, while GEG2P(v1) was retained as the integrated model from the previous iteration. In the third step, we gradually removed base learners with lower prediction accuracy in GEG2P(v2) and dynamically explored to obtain the optimal combination of base learners."

3. Verify that all cited works are accessible publications or preprints.

Response: Thank you for your suggestion. We carefully checked all references cited in this

manuscript. We confirm that all cited references are accessible publications or preprints. Thank you again for your valuable suggestion.

4. Move redundant supplementary figures (S11–S16) to supporting files.

Response: Thank you for your suggestion. We have reorganized the figures in the Supplemental Information and moved Figures S11-S16 to the Supplemental Files.

5. Include complete Data and Code Availability statements per Nature policy.

Response: Thank you for your suggestion. We have added the Data Availability and Code Availability statements required by *Nature* to the manuscript.

While the manuscript presents an ambitious and computationally heavy approach, it does not yet meet the standards of statistical rigor, reproducibility, or interpretability required for publication in Nature. To reach that level, the authors must:

1. Establish a clear statistical and genetic framework for ensemble integration.

Response: Thank you for your suggestion. According to your suggestion, we performed additional statistical analyses in this study. Firstly, we have added experiments to conduct variance decomposition on Top10% SNPs to quantify differences in genetic features identified by each base learner, the results are shown in the response to Comment 1 in Review 1. Secondly, in the comparison between GEG2P and SOTA methods, the uncertainty of prediction accuracy was assessed using SE, and we have added experiments to further evaluate model performance using R^2, RMSE, and MAE, the results are shown in the response to Comment 3 in Review 1. Finally, we reviewed the dataset partitioning and the Optuna parameter search ranges to ensure the accuracy and reliability of the reported performance improvements of GEG2P, the results are shown in the response to Comment 9 in Review 2.

2. Redesign cross-validation and quantify uncertainty.

Response: Thanks for your suggestion. According to your suggestion, we evaluated the performance of GEG2P when the data were partitioned by stratification based on population structure. The results have been provided in our response to Comment 2. In addition, we used standard error (SE) to quantify the uncertainty of prediction accuracy and added error bars to the result figures. The details have been provided in our response to Comment 3 of Review 1.

3. Report complete algorithmic parameters and release code.

Response: Thanks for your suggestion. We have made the scripts used in this study publicly available on GitHub (https://github.com/Deep-Breeding/GEG2P). We have also added the key parameters of GEG2P and their optimization ranges to Supplemental Table S1.

4. Condense and reframe interpretability analyses into biologically meaningful insights.

Response: Thank you for the suggestion. According to your suggestion, the results of the interpretability analysis have been condensed. In addition, biological evidence has been added to further support the reliability of results reported in the manuscript, as described in our response to Comment 6 of Review 1.

5. Reorganize and improve clarity and precision throughout.

Response: Thank you very much for your suggestions. According to your suggestions, we revised the manuscript to improve its clarity, accuracy, and overall readability. We analyzed the variance of different genetic effects, used standard error (SE) to quantify the uncertainty of prediction accuracy, and added error bars to the result figures. We also included R^2, RMSE, and MAE to provide a more comprehensive evaluation of model performance. In addition, we carefully checked the dataset partitioning and reviewed the search ranges of hyperparameters to ensure that the performance improvement of GEG2P was accurate and

reliable. Furthermore, we revised the Introduction section and streamlined the Results section. We believe that these revisions have substantially improved the quality of the manuscript.

Reviewer: 2

The manuscript “GEG2P: Robustly Enhancing Crop Genomic Prediction Accuracy Through Ensemble Learning and Iterative Optimization” proposes a comprehensive ensemble learning framework integrating 20 base learners across statistical, machine learning, and deep learning domains. The authors introduce a three-step iterative strategy and a Genetic Algorithm for weight optimization, demonstrating improved prediction accuracy across five crop species.

I acknowledge the significant engineering effort and the extensive benchmarking performed in this study. The proposed framework shows potential practical value in breeding programs where accuracy is paramount. However, the complexity of the ensemble approach raises concerns regarding computational efficiency, stability, and the depth of interpretability. I recommend the authors address the following points to strengthen the manuscript.

(1) The framework integrates 20 heterogeneous models across R and Python environments, creating a complex computational pipeline. While accuracy improves, the manuscript lacks a detailed assessment of computational resources, such as training and inference time, hardware requirements. I suggest adding a comparison of computational costs between GEG2P and single SOTA models like Cropformer to discuss whether the increased computational burden is cost-effective for practical breeding applications.

Response: Thank you for your valuable suggestion. In response to your concern about computational cost and resource consumption, we added a corresponding evaluation in the revised manuscript. Using the EW in maize as an example, we compared the computation time and computational resource requirements of GEG2P with those of the comparison methods, and the relevant results have been summarized in Supplemental Data 25.

Supplemental Data 25. Computational resources and running time of different methods in

predicting EW of maize.

Item	GPU Memory (MB)	RAM Memory (MB)	CPU cores	Time (s)	PCC
GEG2P-Serial	13844	4420	2	36929	0.5749
GEG2P-Parallel	19546	27918	40	14236	0.5708
Cropformer	680	2847	1	482	0.4771
EGGPT	2112	5106	8	90180	0.5613
SoyDNGP	1878	1790	1	376	0.2646
WheatGP	6386	1814	1	2747	0.2551
RKHS	0	2315	112	3.31	0.5223
Multi-kernel GBLUP	0	1979	124	601	0.5173

The results showed that GEG2P requires more computational time than a single model due to the integration of multiple heterogeneous base learners. However, GEG2P improved prediction accuracy by 2.42% over the second-best method, EGGPT, and by an average of 47.79% over all comparison methods, indicating a substantial performance advantage. In genomic selection, where prediction accuracy is the primary objective, we believe that this increase in computational cost is acceptable. Notably, most of the computational cost of GEG2P arises from the feature extraction stage of the base learners. In practice, this part is highly reusable. Once the base learners have been trained, they can be directly used in the subsequent ensemble and optimization process. Therefore, for long-term and large-scale applications, GEG2P provides a reasonable balance between computational cost and prediction performance.

Meanwhile, to better meet practical demands for computational efficiency, we implemented and added experiments to evaluate a parallel version of GEG2P. The results showed that the parallel implementation reduced the overall runtime of GEG2P by approximately 50%, substantially reducing the overhead introduced by multi-model integration. These results indicate that GEG2P has good parallel scalability and can effectively utilize multi-core CPU environments. This parallel design improves the scalability of GEG2P in large-scale breeding scenarios while maintaining prediction accuracy and enhancing computational efficiency. Specifically, Section 3.6 has been added to discuss the impact of genomic prediction on computational resource

consumption. As shown in the following:

"Computational resource consumption is an important consideration in this study (Supplemental Data 25). The ensemble learning strategy increases the computational cost of GEG2P, resulting in longer runtime and higher computational resource requirements than those of a single method. This observation is consistent with the inherent characteristics of ensemble learning. Previous studies have shown that ensemble learning generally leads to higher computational cost, whether in ensembles of deep learning models (Proskura *et al*., 2022) or in prediction tasks that combine multiple machine learning and deep learning models (Reddy *et al*., 2022). Notably, the computational resource consumption of GEG2P is mainly concentrated in the feature extraction stage of the base learners. Once the base learners are trained, they can be directly used in subsequent ensemble and optimization processes. Therefore, from a practical application perspective, GEG2P achieves a reasonable balance between computational resource consumption and predictive performance."

"In genomic selection, where prediction accuracy is the primary objective, an increase in computational cost to improve predictive accuracy is generally acceptable. To meet the computational efficiency requirements in practical applications, we implemented and evaluated a parallelized version of GEG2P. Experimental results show that the parallelized implementation reduces the overall runtime of GEG2P by approximately 50%, substantially alleviating the overhead introduced by integrating multiple base learners. These results indicate that GEG2P has good scalability for large-scale breeding scenarios, maintaining high prediction accuracy while ensuring computational efficiency."

(2) The weight optimization relies on a GA, which inherently introduces stochasticity. I recommend verifying the robustness of the optimized weights by running the algorithm multiple times with different random seeds to ensure consistency. This validation is necessary to prove that the performance gains stem from capturing stable biological patterns rather than random fluctuations or overfitting to the validation set

Response: Thanks for your suggestion. According to your suggestion, we added experiments in which the genetic algorithm-based optimization was repeated three times with different random

seeds to evaluate the prediction accuracy of GEG2P(v3). The corresponding results have been added to Supplemental Data 23. We found that the prediction accuracy of GEG2P(v3) was highly consistent across the three runs. For all traits, the standard deviation was below 0.01, and the coefficient of variation (CV) was below 0.50%. These results indicate that the improvement in prediction accuracy achieved by GEG2P(v3) across different species and traits is stable. We have added the corresponding description to Section 3.2, as shown in the following:

"The genetic algorithm-based optimization was repeated three times with different random seeds to evaluate the prediction accuracy of GEG2P(v3). The results confirmed that the performance of GEG2P (v3) remained stable across multiple runs (Supplemental Data 23)."

Supplemental Data 23. Prediction accuracy of GEG2P(v3) with different random seeds in the genetic algorithm.

Crop	Trait	Replicate 1	Replicate 2	Replicate 3	Mean	SD	CV
Maize	DTA	0.7053	0.7051	0.7050	0.7051	0.0002	0.02%
	DTS	0.6835	0.6837	0.6832	0.6835	0.0003	0.04%
	DTT	0.7070	0.7068	0.7061	0.7066	0.0004	0.06%
	ATI	0.6598	0.6599	0.6595	0.6597	0.0002	0.03%
	SAI	0.5985	0.5995	0.5990	0.5990	0.0005	0.08%
	STI	0.6816	0.6812	0.6812	0.6813	0.0002	0.03%
	EH	0.6946	0.6942	0.6942	0.6943	0.0002	0.03%
	ELL	0.6943	0.6942	0.6942	0.6942	0.0001	0.01%
	ELW	0.6911	0.6914	0.6913	0.6913	0.0001	0.02%
	LNAE	0.6912	0.6920	0.6914	0.6915	0.0004	0.06%
	LNBE	0.7008	0.7008	0.7013	0.7010	0.0003	0.04%
	PH	0.6868	0.6885	0.6864	0.6872	0.0011	0.17%
	TBN	0.7440	0.7439	0.7432	0.7437	0.0005	0.06%
	TL	0.7658	0.7658	0.7661	0.7659	0.0002	0.02%
	EL	0.6457	0.6456	0.6458	0.6457	0.0001	0.02%
	ERN	0.6916	0.6913	0.6910	0.6913	0.0003	0.04%
	EW	0.5749	0.5764	0.5740	0.5751	0.0012	0.21%
	KNPE	0.5501	0.5506	0.5509	0.5505	0.0004	0.07%
	KNPR	0.5834	0.5831	0.5821	0.5828	0.0007	0.12%
	KWPE	0.5403	0.5412	0.5402	0.5406	0.0006	0.10%
	LBT	0.5382	0.5381	0.5390	0.5385	0.0005	0.09%
	CW	0.6647	0.6640	0.6648	0.6645	0.0004	0.06%
	ED	0.5954	0.5959	0.5960	0.5957	0.0003	0.05%

Chickpea	Days_to_0.5_flowering	0.5125	0.5124	0.5125	0.5125	0.0000	0.01%
	Plant_height	0.6041	0.6042	0.6041	0.6041	0.0001	0.02%
	Yield	0.4823	0.4823	0.4826	0.4824	0.0002	0.04%
Rice	BLUP_Heading_date	0.9198	0.9199	0.9198	0.9198	0.0001	0.01%
	BLUP_Height	0.8849	0.8849	0.8849	0.8849	0.0000	0.00%
	BLUP_Yield_per_plant	0.4491	0.4495	0.4503	0.4496	0.0006	0.13%
Wheat	Glaucousness	0.7608	0.7608	0.7609	0.7608	0.0000	0.00%
	Maturity	0.8230	0.8230	0.8233	0.8231	0.0002	0.02%
	Grain_Yield	0.4817	0.4823	0.4823	0.4821	0.0004	0.08%
	Grain_Size_TKW	0.7497	0.7500	0.7499	0.7499	0.0001	0.02%
Soybean	BBD_BLUP	0.8552	0.8550	0.8551	0.8551	0.0001	0.01%
	PLH_BLUP	0.8089	0.8091	0.8089	0.8090	0.0002	0.02%
	SL_BLUP	0.6479	0.6482	0.6498	0.6486	0.0010	0.15%
	ST_BLUP	0.7643	0.7640	0.7646	0.7643	0.0003	0.04%

(3) The current interpretability analysis, relying largely on SHAP values and the overlap of Top 10% SNPs, is somewhat superficial for such a complex ensemble. SHAP explains feature importance but not the mechanism of ensemble correction. I suggest providing a specific case study demonstrating how the ensemble resolves conflicts via weight allocation when different base learners such as linear or others that generate contradictory predictions.

Response: Thank you for the suggestion. In this study, Top10% SNPs were designated as key phenotype-associated genetic variants, following methodologies used in relevant studies. For example, Gamba *et al*. (2024) identified potential functional genes from the 100 SNPs with the smallest *P*-values based on GWAS of flowering time in *Arabidopsis*. Wang *et al*. (2024) used SHAP to select the 20 genes with the highest contributions to genomic prediction and effectively clustered flowering time across different accessions based on their expression patterns. Sun *et al*. (2026) analyzed the chromosomal distribution of the top 100 SNPs ranked by combined SHAP and GWAS contributions, revealing differences in the genetic architecture among upland cotton phenotypes. These studies systematically screened key loci or genes across the whole genome for subsequent analyses. Based on this framework, the present study designated Top10% SNPs from each base learner as key phenotype-associated genetic features, with detailed descriptions provided in Section 3.3, as shown in the following:

"Traditional genomics studies generally assume that a limited number of major-effect genes play dominant roles in determining phenotypic variation. Guided by this concept, current bioinformatics studies often identify phenotype-associated functional genes based on a small set of key genetic features. For example, Gamba *et al*. (2024) identified key genes from 100 SNPs showing the most significant associations with flowering time in *Arabidopsis* based on GWAS. Wang *et al*. (2024) applied SHAP to select top 20 genes contributing to genomic prediction and achieved clustering of flowering time among different accessions based on their expression patterns. In addition, Sun *et al*. (2026) analyzed the chromosomal distribution of the top 100 SNPs ranked by combined SHAP and GWAS contributions, revealing differences in the genetic architecture among upland cotton phenotypes.

Following the analytical framework of previous studies, we also used SHAP to quantify the contribution of individual SNPs to phenotypic prediction. The Top10% SNPs ranked by contribution within each base learner were subsequently selected for further analysis."

In addition, we have added experiments to analyze the ensemble correction mechanism of GEG2P. For the prediction of maize DTA, PH, and EW, we first calculated the differences between the observed phenotypic values and the predictions from individual base learners. The results indicated that most base learners achieved high prediction accuracy, although some predictions consistently deviated from the observed values. For example, LCNN predictions were generally higher than the observed values for three phenotypes, whereas DNNGP predictions were lower for DTA and PH but higher for EW (Figure S18A, C, E). Based on this, we further integrated the phenotypic prediction deviations with the corresponding base learner weights (Figure S18B, D, F). Base learners exhibiting larger deviations were consistently assigned lower weights. For example, LCNN weights were 0, 0.020, and 0, and DNNGP weights were 0, 0.024, and 0 for DTA, PH, and EW, respectively. These results indicate that GEG2P improves predictive performance by assigning lower weights to poorly performing base learners, thereby reducing their influence. We have supplemented the discussion of ensemble correction mechanism in Section 3.2, as shown in the following:

"Taking the maize DTA, PH, and EW phenotypes as examples, most base learners achieve high prediction accuracy although the predictions of a few base learners deviate substantially from the observed values (Figure S18A, C, E). These base learners are assigned lower weights or excluded from the final model (Figure S18B, D, F), thereby reducing their influence on the overall predictions and improving the performance of GEG2P."

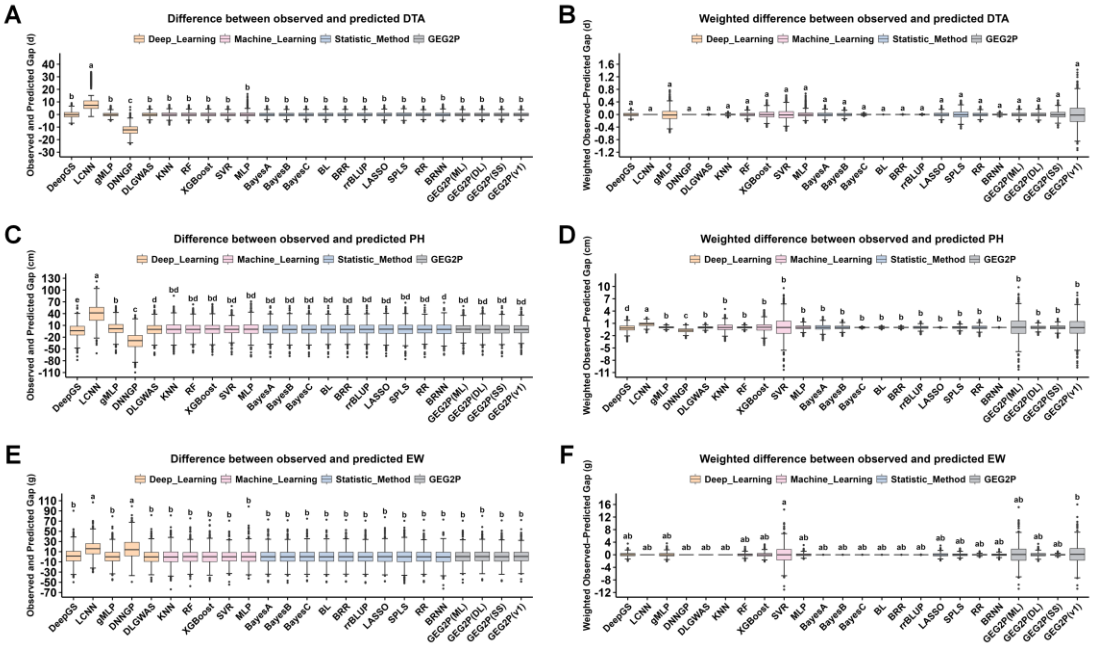

Figure S18. Weight-based ensemble correction among prediction results of different base learners. (A, C, E) Prediction deviations (observed minus predicted values) of base learners for the DTA, PH, and EW, respectively. (B, D, F) The weighted results of these deviations derived from base learner weights. Distinct letters denote statistically significant differences at *P*-value < 0.05 (Tukey's HSD test).

In addition, the key references discussed above are shown in the following:

[1] Gamba D, Lorts C, Haile A, *et al*. (2024). The genomics and physiology of abiotic stressors associated with global elevational gradients in *Arabidopsis thaliana*. *New Phytologist*, 244(5): 2062-2077.

[2] Wang P, Lehti-Shiu M, Lotreck S, *et al*. (2024). Prediction of plant complex traits via integration of multi-omics data. *Nature Communications*, 15(1): 6856.

[3] Sun J, Zhang X, You X, *et al*. (2026). Bayesian neural networks for genomic prediction: uncertainty quantification and SNP interpretation with SHAP and GWAS. *Theoretical and Applied Genetics*, 139(1): 29.

(4) In the third step of iterative optimization, the algorithm progressively removes underperforming base learners. I suggest further analyzing the patterns of these removed models, for instance, whether specific model types, such as overfitting-prone deep learning models, are more likely to be discarded in certain species or traits. Discussing the patterns behind this selection process would provide more insight than simply describing the screening procedure.

Response: Thank you for the suggestion. According to your suggestion, we summarized the base learners removed from GEG2P(v3) together with their methodological categories across different species and traits. The corresponding results have been added to Supplemental Data 24. The results showed that base learners from all methodological categories were removed in some cases, indicating that the performance of different base learners varied across species and traits. These results further confirm that single base learners fail to maintain stable prediction performance across different species and traits, highlighting the need for an ensemble strategy that integrates base learners with different feature extraction capabilities. We have added the analysis

of the number and categories of removed base learners in Section 3.2, as shown in the following:

"We also analyzed the number and types of base learners removed in different species and traits (Supplemental Data 24). We found that at least one base learner from each method category was removed in some cases. These findings further confirm that a single genomic prediction method is unlikely to achieve consistently stable predictive performance in multiple species and traits."

In addition, among the removed base learners, deep learning-based models were removed more frequently than those from other methodological categories. This pattern may be related to the greater sensitivity of deep learning models to hyperparameter settings, which can lead to larger variation in prediction performance under different data distributions and parameter configurations. We also found that DNNGP and LCNN were removed more frequently, which is consistent with their relatively lower overall prediction accuracy for these traits.

Supplemental Data 24. Base learners removed in different crops and traits

Crop	Trait	Removed Base Learners	Deep Learning	Machine Learning	Statistical Methods
Maize	DTA	DNNGP, LCNN	2	0	0
	DTS	DNNGP, LCNN	2	0	0
	DTT	DNNGP, LCNN	2	0	0
	ATI	DNNGP	1	0	0
	SAI	-	0	0	0
	STI	KNN	0	1	0
	EH	-	0	0	0
	ELL	LCNN	1	0	0
	ELW	DNNGP, MLP	1	1	0
	LNAE	DNNGP	1	0	0
	LNBE	-	0	0	0
	PH	-	0	0	0
	TBN	DNNGP	1	0	0
	TL	-	0	0	0
	EL	-	0	0	0
	ERN	DNNGP, LCNN, MLP, SPLS	2	1	1
	EW	DNNGP, LCNN, KNN	2	1	0

	KNPE	-	0	0	0
	KNPR	DNNGP, KNN, BRNN	1	1	1
	KWPE	-	0	0	0
	LBT	XGBOOST	0	1	0
	CW	-	0	0	0
	ED	-	0	0	0
	Maturity	KNN, SVR	0	2	0
Wheat	Grain_Yield	-	0	0	0
	Grain_Size_TKW	-	0	0	0
	BLUP_Heading_date	DeepGS	1	0	0
Rice	BLUP_Height	-	0	0	0
	BLUP_Yield_per_plant	-	0	0	0
Chickpea	Days_to_0.5_flowering	MLP, SVR	0	2	0
	Plant_height	MLP	0	1	0
	Yield	-	0	0	0
	BBD_BLUP	-	0	0	0
Soybean	PLH_BLUP	-	0	0	0
	SL_BLUP	-	0	0	0
	ST_BLUP	KNN	0	1	0

(5) The results indicate that XGBoost consistently dominates the weights for yield-related traits across multiple species. If a single model contributes the vast majority of predictive power, the necessity of ensembling many other low-weight models becomes questionable. The authors need to clarify whether the low-weight models contribute essential complementary signals or merely add system complexity without substantial benefit.

Response: Thank you for the suggestion. In Section 3.5, most base learners were unable to consistently achieve high weights. Only XGBoost was consistently assigned relatively high weights for yield-related traits across different species, indicating its comparatively stable

predictive performance. It should be noted that this result reflects the relative stability of XGBoost compared with other learners, rather than its superiority in predictive accuracy.

Furthermore, integrating additional base learners beyond XGBoost remains necessary. We have added experiments to compare the 10-fold cross-validated PCCs of XGBoost, GEG2P (ML), and GEG2P (v3) for yield-related trait prediction across different species. The results showed that integrating more base learners significantly improved prediction accuracy in the majority of cases (Figure S20A~G), indicating that the integration of multiple base learners provides critical complementary features for GEG2P. We have clarified the necessity of integrating additional base learners beyond XGBoost in Section 3.5, as shown in the following:

"Although XGBoost exhibited relatively stable performance in predicting yield-related traits across species, further integration of additional base learners could still improve prediction accuracy. A comparison of 10-fold cross-validated PCCs of XGBoost, GEG2P(ML), and GEG2P(v3) for yield-related trait prediction across species showed that GEG2P(ML) and GEG2P(v3) achieved significantly higher prediction accuracy than XGBoost in most scenarios (Figure S20A~G). These results indicate that the integration of multiple base learners provides critical complementary features for the model, further underscoring the necessity of the ensemble strategy."

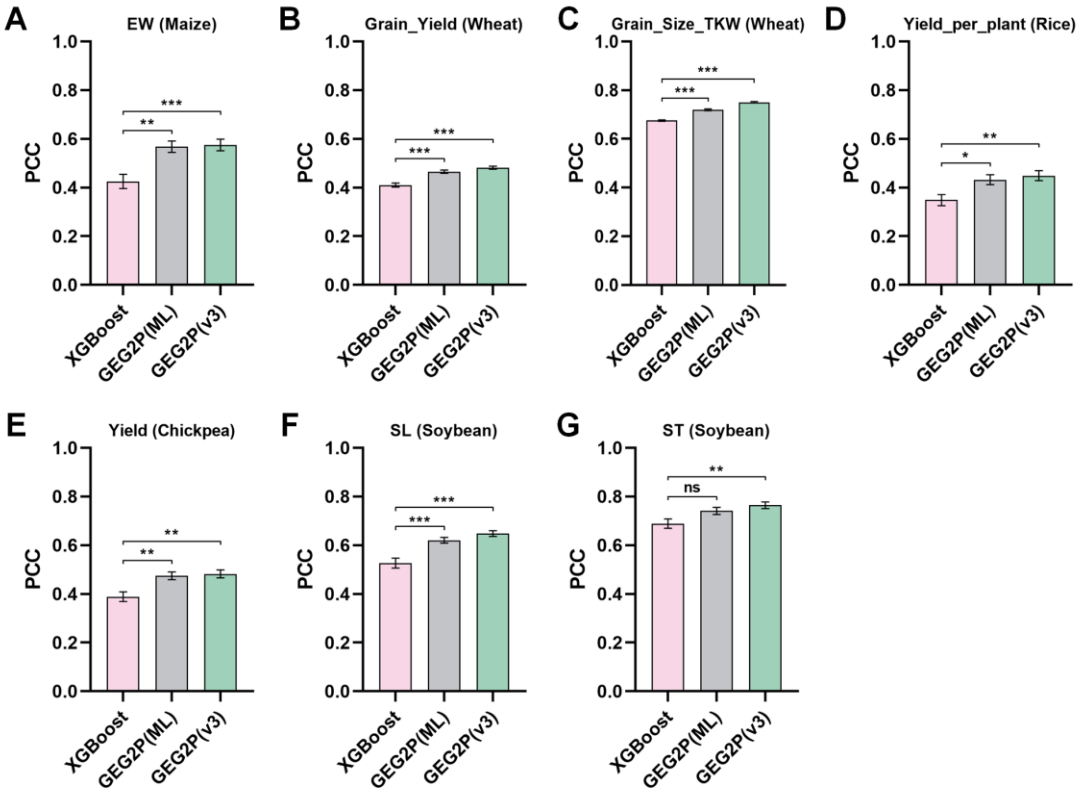

Figure S20. Prediction accuracy of maize EW (A), wheat Grain_Yield (B), wheat Grain_Size_TKW (C), rice Yield_per_plant (D), chickpea Yield (E), soybean SL (F) and soybean ST (G) using XGBoost, GEG2P (ML), and GEG2P (v3). Error bars indicate the standard error and asterisks indicate significance (* *P*-value < 0.05; ** *P*-value < 0.01; *** *P*-value < 0.001; ns, not significant; *t*-test).

(6) Given that 20 weight parameters are optimized through multiple iterative rounds, there is a risk of overfitting to the validation set. Please clarify whether the dataset used for weight optimization is completely independent of the test set used for reporting final accuracies. If there is overlap, the performance assessment might be inflated, and a strict independent test set or outer-loop cross-validation is recommended to confirm the results.

Response: Thanks for your suggestion. We re-examined the entire experimental design, with particular attention to the data partitioning strategy used during weight optimization. We confirmed that the same sample partitioning scheme used in the base-learner training stage was strictly maintained during weight optimization process, ensuring that no data leakage occurred.

Therefore, the improvement in PCC reported in this study resulted from a fair comparison and was not artificially inflated by inappropriate data partitioning. According to your suggestion, we have clarified the data partitioning strategy in Section 4.3, as shown in the following:

"During optimization of base learner weights using the genetic algorithm, the sample partitioning strategy from the training phase is strictly followed to avoid data leakage."

(7) The paper constructs epistasis networks based on significant SNP interactions, but this appears to be a post-hoc statistical result based on ANOVA rather than a structure directly learned by the ensemble model. Please clarify the logical link between the model's predictive mechanism and these epistasis networks; specifically, does the ensemble's accuracy gain truly stem from implicitly capturing these interactions detected by ANOVA?

Response: Currently, genomic prediction studies increasingly incorporate SHAP to facilitate the identification of key genetic features. For example, Wang *et al*. (2024) used SHAP to identify genes regulating flowering time in *Arabidopsis thaliana*. Wang *et al*. (2025) combined Cropformer with SHAP to interpret potential quantitative trait loci. He *et al*. (2025) applied SHAP to identify key loci with stable regulatory effects on phenotypes across multiple environments. Sun *et al*. (2026) integrated SHAP with GWAS to overcome the limitations of GWAS in capturing nonlinear genetic effects. However, existing studies have primarily focused on the direct interpretation of feature contribution effects derived from SHAP, while further statistical analysis of these effects remains limited. In this context, the present study provides a modest extension to existing analytical approaches. Specifically, SHAP was first used to quantify the contribution of each SNP to phenotype prediction, followed by ANOVA to assess interaction effects among key SNPs and their associations with phenotypes. This post hoc strategy based on SHAP results aims to elucidate potential genetic effects from a statistical perspective.

As shown in the response to Comment 6 in Review 1, SNPs involved in significant epistatic interaction pairs identified by ANOVA were mapped to genes. We have added experiments that integrate RNA-seq data from the same population (Liu *et al*., 2020) to demonstrate the reliability of the potential epistatic network constructed based on ANOVA. Examples of potential regulatory relationships between selected genes have been added to Section 2.6, as shown in the following:

"In addition, integrating RNA-seq data from 391 accessions of the same population (Liu *et al*., 2020) revealed that the expression value of *Zm00001d013683*, mapped by chr5.s_17285732 among the 1,077 interacting SNPs, was significantly correlated with *ZMM15* expression (*P*-value = 3.8×10^{-8}). And earlier flowering was significantly associated with increased expression of *Zm00001d013683* (*P*-value = 1.5×10^{-4}) (Figure S17A). These results provide evidence supporting potential interactions among the Top10% SNPs identified by different base learners from the perspective of gene expression and phenotype association."

"In addition, the expression value of *Zm00001d032283*, mapped by chr1.s_220370448 among the 401 interacting SNPs, was significantly correlated with *BRD1* expression (*P*-value = 3.4×10^{-8}). Moreover, the dwarf phenotype was significantly influenced by the expression value of *BRD1* (*P*-value = 0.017) (Figure S17B). These results reflect potential interactions among genes mapped by the Top10% SNPs."

"Furthermore, the expression value of *Zm00001d012822*, mapped by chr5.s_922298 among the 268 interacting SNPs, was significantly correlated with *ZmACO2* expression (*P*-value = 1.9×10^{-8}). In addition, the expression value of *ZmACO2* showed a significant effect on ear weight (*P*-value = 2.7×10^{-4}) (Figure S17C). These results further provide evidence for potential interactions among the Top10% SNPs identified by different base learners."

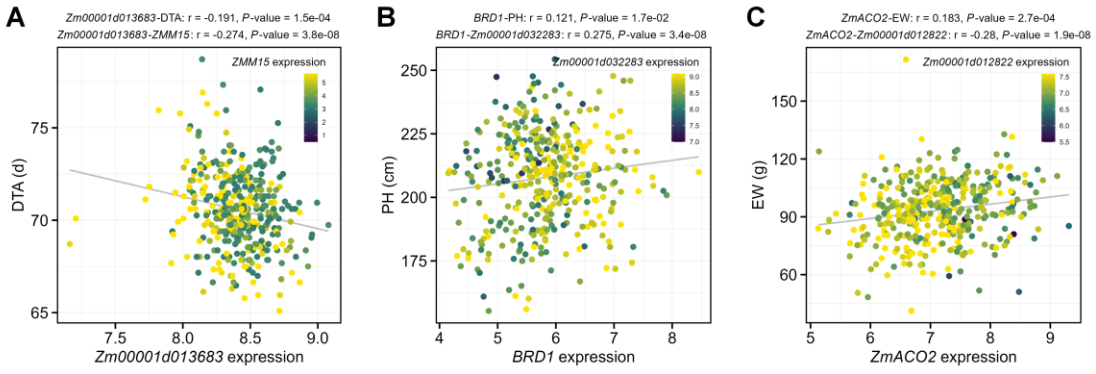

Figure S17. Correlation between expression values of interacting gene pairs and phenotypic variation. (A) Pearson correlation between the expression value of *Zm00001d013683* (*x*-axis) and DTA (*y*-axis). Color scale indicates *ZMM15* expression value. (B) Pearson correlation between the expression value of *BRD1* (*x*-axis) and PH (*y*-axis). Color scale indicates expression value of *Zm00001d032283*. (C) Pearson correlation between the expression value of *ZmACO2* (*x*-axis) and EW (*y*-axis). Color scale indicates expression value of *Zm00001d012822*.

In addition, the key references discussed above are shown in the following:

[1] Wang P, Lehti-Shiu M, Lotreck S, *et al*. (2024). Prediction of plant complex traits via integration of multi-omics data. *Nature Communications*, 15(1): 6856.

[2] Wang H, Yan S, Wang W, *et al*. (2025). Cropformer: An interpretable deep learning framework for crop genomic prediction. *Plant Communications*, 6(3): 101223.

[3] He K, Yu T, Gao S, *et al*. (2025). Leveraging automated machine learning for environmental data-driven genetic analysis and genomic prediction in maize hybrids. *Advanced Science*, 12(17): e2412423.

[4] Sun J, Zhang X, You X, *et al*. (2026). Bayesian neural networks for genomic prediction: uncertainty quantification and SNP interpretation with SHAP and GWAS. *Theoretical and Applied Genetics*, 139(1): 29.

[5] Liu H, Wang X, Xiao Y, *et al*. (2020). CUBIC: an atlas of genetic architecture promises directed maize improvement. *Genome Biology*, 21(1): 20.

(8) Since the integrated base learners have different requirements for input data formats (e.g., deep learning often requires one-hot encoding, while statistical models use 0/1/2 numerical encoding), please detail the data preprocessing pipeline in the Methods section. Clarifying whether a unified encoding was used or if specific processing was applied for different model categories is crucial for engineering reproducibility.

Response: Thank you for your suggestion. For the 20 base learners integrated into GEG2P, we uniformly used a numerical encoding scheme of 0/1/2. Specifically, the homozygous reference genotype was encoded as 0, the heterozygous genotype as 1, and the homozygous alternate genotype as 2. As you noted, statistical methods commonly use this 0/1/2 encoding scheme. In addition, DeepGS, DNNGP, and DLGWAS also use the same encoding. Therefore, to ensure a consistent input format across all base learners and to simplify data processing and model integration, we applied the same 0/1/2 encoding scheme to all base learners in GEG2P. We have revised the description of data processing in Section 2.1 accordingly, as shown in the following:

“In the stage of data processing, genotype data were encoded using the 0/1/2 scheme to ensure a unified input format for all base learners.”

(9) When comparing GEG2P with SOTA methods like Cropformer and WheatGP, Optuna was used for hyperparameter tuning. Please confirm that the computational budget (e.g., number of trials) allocated to these baseline methods was comparable to the extensive resources used for GEG2P. It is essential to ensure that the baselines were evaluated at their optimal performance to rule out artificial performance gaps caused by under-tuning.

Response: Thank you for your suggestion. Following your suggestion, we checked and summarized the hyperparameter search ranges for all comparison methods. The relevant information has been added to Supplemental Table S1. To ensure a fair comparison, all comparison methods were allocated 50 trials during hyperparameter tuning. These settings ensured that each method was optimized within a reasonable parameter space. In addition, all comparison methods and GEG2P were trained on the same hardware configuration to further ensure fairness. A detailed response has been provided in our reply to Reviewer 1, Comment 3.

(10) Given the complex cross-language environment dependencies, reproducing this framework is challenging. To truly demonstrate the engineering value of this study, I strongly recommend that the authors provide a containerized environment or a detailed Conda environment configuration file in addition to the GitHub code to resolve potential dependency conflicts.

Response: Thank you for your suggestion. According to your suggestion, we reviewed the dependencies of GEG2P and constructed a Conda environment configuration file. We uploaded this file to GitHub (https://github.com/Deep-Breeding/GEG2P/blob/main/01GEG2P_code/restore_env.sh) and detailed instructions for creating the conda environment using this file have been added to the user guide. Users can now automatically create the Conda environment and install all dependencies required for GEG2P with a single command. In addition, we developed a Docker image for GEG2P (https://hub.docker.com/r/coder02lq/geg2p) and prepared a user manual. This Docker image includes all dependencies required to run GEG2P. Users can run GEG2P directly from the Docker image without any additional environment setup.

Best regards,

Jianxiao Liu

Response to Reviewers:

Dear Mr./Ms. reviewers:

We appreciate your valuable comments, and we tried our best to address them in the revised version of our manuscript. For the reviewers' comments, we have made the following modifications.

Reviewer #2 (Remarks to the Author):

I have carefully reviewed the authors' responses to my initial review comments and examined their revised manuscript. I am satisfied with how they have addressed each of my concerns. I recommend acceptance of this manuscript for publication in Nature Communications.

Response: Thank you very much.

Reviewer #2 (Remarks on code availability):

The code repository is publicly accessible and well organized, with a README file providing sufficient instructions for installation and execution. I was able to run the main workflow successfully, and the provided code appears sufficient to reproduce the major results reported in the manuscript.

Response: Thank you very much.

Reviewer #3 (Remarks to the Author):

We commend the authors for the substantial engineering effort and extensive benchmarking invested in developing the GEG2P framework. Integrating twenty heterogeneous base learners (spanning statistical, machine learning, and deep learning models) into a unified genomic prediction tool represents a highly ambitious endeavor. Evaluating this framework across 36 traits in five major crop species illustrates the extensive scale of this study and highlights its relevance to applied breeding programs. The core concept of combining diverse

algorithms offers a promising direction for genomic prediction. However, to fully realize the scientific value of an undertaking of this magnitude, it is necessary to move beyond reporting experimental prediction accuracies and establish robust statistical equivalence and unified interpretability standards.

In its current form, the revised manuscript has not fully articulated the mechanistic necessity and theoretical underpinnings of the proposed framework. We find that several core methodological concerns raised by Reviewer #1 during the initial review remain inadequately resolved. Specifically, the rebuttal predominantly relies on reporting raw prediction accuracies and providing descriptive defenses, rather than establishing the rigorous statistical baselines, convergence proofs, and computational clarifications necessary for a multi-algorithm fusion architecture. To elevate this manuscript to the requisite standard and clearly elucidate its theoretical contribution, we strongly recommend addressing the following unresolved points.

1. Statistical Equivalence and Model Synergy (Ref. R1 Comment 1)

Although the authors compared their method with established frameworks, they did not establish statistical equivalence. We infer that Reviewer #1 expected a comparison at this level, rather than simply reporting experimental accuracy metrics. For a multi-algorithm fusion architecture, evaluating statistical equivalence (even empirically) constitutes its primary scientific value. Establishing this equivalence provides the necessary framework to demonstrate mechanistically how Linear Mixed Models (LMMs), machine learning, and deep learning capture complementary effects, a claim the authors currently assert without structural proof. As an actionable alternative, a straightforward residual analysis on the test set could yield rich insights into these synergistic relationships. While we certainly encourage the authors to employ other rigorous statistical approaches if they prefer, providing this depth of analysis will offer substantially more mechanistic value than supplementary accuracy tables.

Response: Thank you for the suggestion. In Section 2.6, we sequentially integrated the Top10% SNPs of Weight-Top-5 and estimated the narrow-sense heritability of the integrated SNP set. The results indicated that the additive genetic effect explained by the union of

Top10% SNPs progressively rose as the number of integrated base learners increased. According to your suggestion, we further applied the BGLR package (Version 1.1.4, Pérez and de Los, 2014; Gogna *et al*., 2026) to perform genetic variance decomposition on the union of Top10% SNPs of Weight-Top-5, and contrasted the results with those derived from Top10% SNPs of individual base learners. For maize DTA, PH, and EW, the variance decomposition model based on the union of Weight-Top-5 SNPs exhibited lower residual variance in most scenarios (Figure S19A~C). These results further indicate that different base learners capture complementary genetic effects. We have added the genetic variance decomposition results for the union of Top10% SNPs of Weight-Top-5 in Section 3.4, as shown in the following:

"Similarly, genetic variance decomposition analysis results (see the "Methods" section) further demonstrated that Top10% SNPs identified by different base learners consistently exhibited substantial genetic effects. For maize DTA, the additive and epistatic genetic effects of Top10% SNPs in DTA-Weight-Top-5 explained a mean of 70.12% and 6.32% of the phenotypic variance, respectively (Figure S19A). Similarly, the genetic effects of Top10% SNPs in PH-Weight-Top-5 and EW-Weight-Top-5 explained a mean of 75.77% and 76.42% of the phenotypic variance, respectively (Figure S19B~C). These results indicate that Top10% SNPs consistently explained the majority of the phenotypic variation despite variation in the genetic features captured by different base learners. Moreover, the variance decomposition model based on the union of Weight-Top-5 SNPs exhibited lower residual variance in most scenarios of maize DTA, PH, and EW (Figure S19A~C). These results indicate that integrating Top10% SNPs from multiple base learners explain a greater proportion of phenotypic variance, reflecting the complementarity of the genetic features captured by different base learners."

Moreover, we have modified the description of genetic variance decomposition method in Section 4.2.7, specifically clarifying the R package employed and the composition of genetic markers, as shown in the following: "In this study, the BGLR (Version 1.1.4, Pérez and de los Campos, 2014) R package was used to perform genetic variance decomposition for both Top10% SNPs of individual base learners and the union of Top10% SNPs of Weight-Top-5."

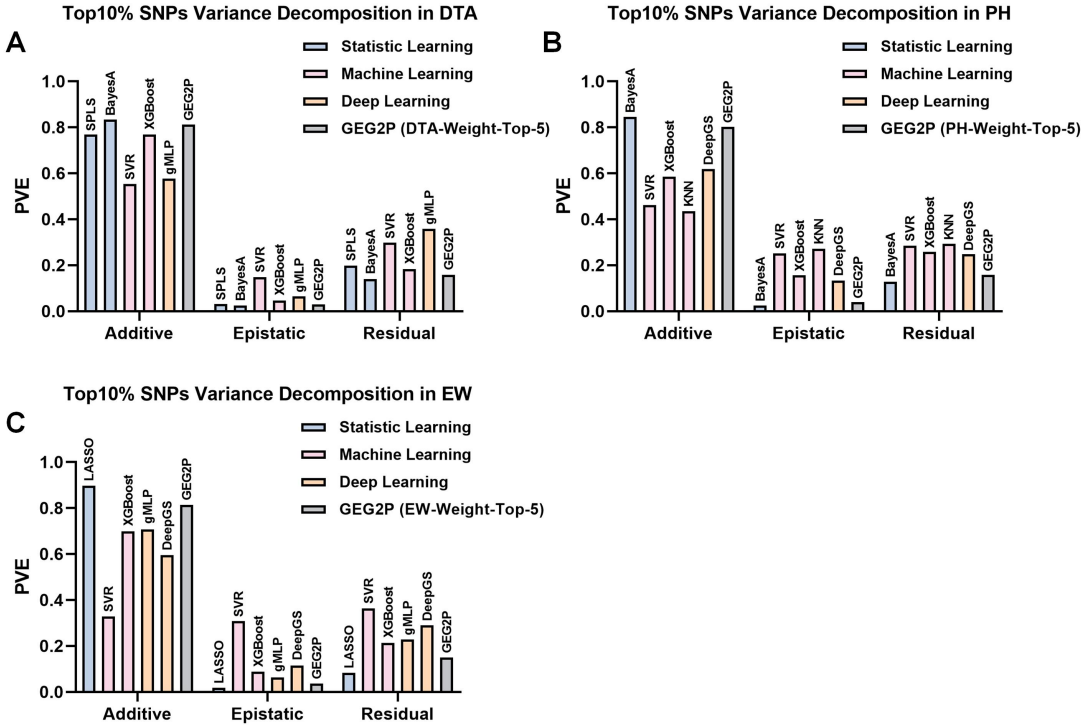

Figure S19. Variance decomposition of genetic effects for the Top10% SNPs identified by individual base learners and their union in DTA-Weight-Top-5 (A), PH-Weight-Top-5 (B), and EW-Weight-Top-5 (C).

In addition, all the references mentioned above are shown in the following:

[1] Gogna A, Kamali B, Wimmer V, *et al*. (2026). Predicting enviromically adapted varieties with big data. *Genome Biology*, 27(1): 3.

[2] Pérez P, de los Campos G. (2014). Genome-wide regression and prediction with the BGLR statistical package. *Genetics*, 198(2): 483-495.

2. Convergence Analysis and Iteration Ablation (Ref. R1 Comments 4 & 8). The three-step iterative optimization is presented as a core algorithmic contribution. Structurally, however, this process resembles recursive re-weighting. To solidify this as a genuine algorithmic breakthrough (addressing the "Theoretical Value" and "Ablation" requested in Comment 8) rather than the natural smoothing effect of repeated averaging, a rigorous convergence analysis is required. The authors should provide actual convergence curves and conduct an ablation study that explicitly contrasts the three-step iteration against a direct, single-step

optimization computed until full mathematical convergence. Demonstrating the mechanistic superiority of the three-step approach over a fully converged single step is essential to establish algorithmic rigor.

Response: Thank you for the suggestion. We would like to clarify that the three-step iterative optimization strategy proposed in this study is distinct from recursive re-weighting. Recursive re-weighting generally refers to the repeated adjustment of weights within a fixed solution space. In contrast, the key feature of our strategy is that the candidate set of base learners is dynamically updated at each step, rather than repeatedly adjusting the weights of a fixed set of base learners. At each step, the weights of the base learners are globally optimized based on the current candidate set, and the optimized weights are not inherited by subsequent steps. In the first step, base learners are no longer separated by their categories. Instead, all base learners are placed into a unified candidate set. This allows GEG2P(v1) to remove the boundaries of base learners and directly integrate different types of base learners with different feature extraction capabilities. GEG2P(DL), GEG2P(ML), GEG2P(SS), and GEG2P(v1) are ensemble models constructed by different integration strategies. These models integrate diverse base learners from different perspectives, and provide broader feature extraction capabilities. In the second step, these four models are treated as new base learners to expand the candidate set and construct GEG2P(v2). Therefore, this step is performed on a newly expanded candidate set, rather than simply re-optimizing the weights of the base learners included in GEG2P(v1). In the third step, we design a gradual elimination strategy to further update the set of candidate base learners. In each iteration, we remove the worst-performing base learners and evaluate the prediction accuracy of the ensemble model based on the updated candidate set. This process continues until the prediction accuracy stabilizes, and finally construct GEG2P(v3). GEG2P(v3) retains the base learners that contribute most to the improvement of prediction performance and reduces the interference of redundant base learners. Therefore, the core of the three-step iterative optimization strategy is to optimize the candidate set of base learners, rather than recursively updating the weights of a fixed candidate set.

According to your suggestion, we added experiments to evaluate the convergence of GEG2P under the three-step iterative optimization strategy. GEG2P(v1) does not use the

complete iterative optimization strategy, but directly integrates all base learners. Therefore, GEG2P(v1) can be regarded as a control model for direct single-step optimization. In contrast, GEG2P(v3) is constructed using the complete three-step iterative optimization strategy. We conducted experiments using the EW, PH, and DTA in the maize dataset, and analyzed convergence curves of PCC and RMSE during weight optimization for the two models (Figure S21A and Figure S21B). The results showed that both GEG2P(v1) and GEG2P(v3) exhibited stable convergence trends within 40 rounds of the genetic algorithm. As the number of optimization rounds increased, the RMSE values of both models generally decreased, whereas their PCC values generally increased. Both RMSE and PCC gradually stabilized in the later rounds. These results indicate that the three-step iterative optimization strategy can converge effectively, and that the improvement in model performance is unlikely to be caused by random fluctuations. As shown in Figure S21A and Figure S21B, GEG2P(v1) showed stable convergence trends for the EW, PH, and DTA. However, GEG2P(v3) achieved further performance improvement after the complete three-step iterative optimization strategy was used, and this improvement was consistent for both PCC and RMSE. In Figure S21C and Figure S21D, GEG2P(v3) showed higher PCC values and lower RMSE values than GEG2P(v1) for EW, PH, and DTA. The above results demonstrate that the three-step iterative optimization strategy can further improve model performance, even when the model obtained by direct single-step optimization has already tended to converge. These results further confirm the effectiveness of the proposed strategy.

In summary, the core of the three-step iterative optimization strategy proposed in this study is to progressively update the candidate set of base learners, rather than recursively re-weighting a fixed candidate set. In GEG2P(v1), the category boundaries among base learners are removed, and all base learners are placed into a unified candidate set for weight optimization. In GEG2P(v2), the candidate set is expanded by introducing ensemble learners constructed using different integration strategies. In GEG2P(v3), a more effective candidate set is further identified through a gradual elimination strategy. Because the candidate set of base learners is updated at each optimization step, the performance improvement of GEG2P(v3) is not attributable to the smoothing effect of repeated weighting or averaging. Instead, it mainly results from the updated candidate set, which better captures

complementary predictive information from different base learners and reduces the influence of redundant learners. According to your suggestion, we have modified the corresponding content in Section 2.3, as shown in the following:

"We further analyzed the prediction performance of GEG2P(v1) and GEG2P(v3) across different numbers of optimization rounds in genetic algorithm. The results showed that both models exhibited stable convergence as the number of optimization rounds increased (Figure S21). Overall, RMSE showed a decreasing trend, whereas PCC showed an increasing trend. Compared with GEG2P(v1), GEG2P(v3) achieved higher PCC values and lower RMSE values for EW, PH, and DTA in maize. These results indicate that the three-step iterative optimization strategy can further improve model performance, even when the single-step model has already converged. This finding further confirms the effectiveness of the proposed three-step iterative optimization strategy."

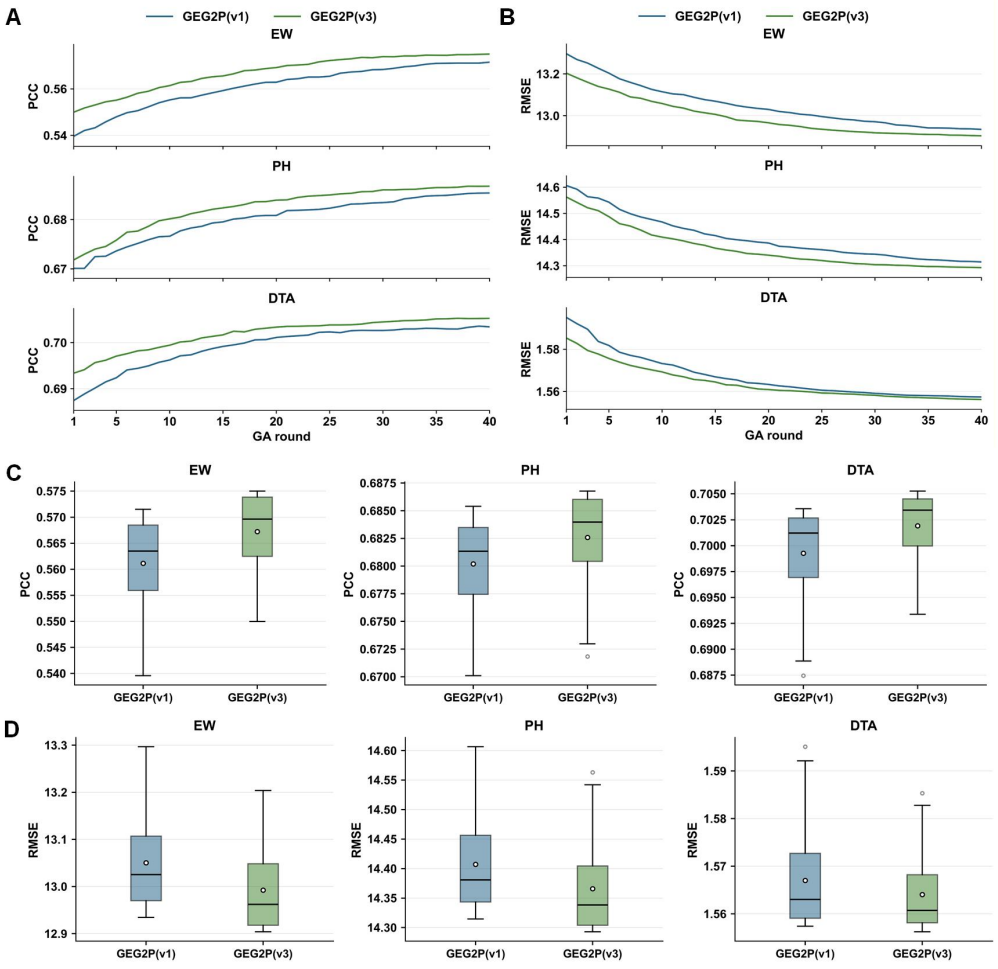

Figure S21 Comparison of the predictive performance of GEG2P(v1) and GEG2P(v3) across different iteration rounds of genetic algorithm. (A) Trends in PCC of GEG2P(v1) and

GEG2P(v3) across iteration rounds for EW, PH, and DTA in maize. (B) Trends in RMSE of GEG2P(v1) and GEG2P(v3) across iteration rounds for EW, PH, and DTA in maize. (C) PCC distributions of GEG2P(v1) and GEG2P(v3) for EW, PH, and DTA in maize. (D) RMSE distributions of GEG2P(v1) and GEG2P(v3) for EW, PH, and DTA in maize.

3. Rigor in Statistical Significance Testing (Ref. R1 Comment 5)

Integrating 20 highly diverse models introduces immense mathematical complexity, making traditional analytical P-values highly susceptible to assumption violations (e.g., normality, independence) in downstream testing. To ensure statistical rigor in a methodological paper, the authors must detail the foundational parameters of their significance tests (i.e., data splits, null and alternative hypotheses, specific test methods, and assumed distributions). Furthermore, in such a highly parameterized scenario, we strongly recommend deriving empirical P-values to validate the statistical significance of the findings.

Response: Thank you for the suggestion. We added experiments to evaluate the distribution of predicted phenotype values for maize DTA, PH, and EW. The results showed that the predicted values by GEG2P followed an approximately normal distribution (Figure 3.1A~C). Similarly, the state-of-the-art (SOTA) ensemble learning model EGGPT (Wu *et al*., 2024) exhibited the same distributional pattern (Figure 3.1D~F). These results indicate that integrating multiple base learners has a limited impact on the distributional complexity of the predicted values.

According to your suggestion, we modified the significance testing approach for comparing the prediction performance of GEG2P with six SOTA methods. Specifically, we originally used independent *t*-tests to compare the prediction performance (PCC) of GEG2P and the SOTA methods. However, this statistical test requires the samples to follow a normal distribution, whereas the distribution of cross-validation accuracies is unlikely to conform to this assumption. Moreover, because the same partitioning strategy was applied to all methods, the predictive accuracies for different methods on the same fold are not independent. Therefore, we re-evaluated the statistical significance of the predictive accuracy between GEG2P and the six SOTA methods using the Wilcoxon signed-rank test (Wilcoxon, 1945). As

a non-parametric paired test independent of normality assumptions, the Wilcoxon signed-rank test offers a robust framework for accurately comparing prediction performance across different models.

In this study, the null hypothesis of the Wilcoxon signed-rank test states that the median difference in PCC between GEG2P and the SOTA methods across cross-validation folds is zero, while the alternative hypothesis is that the median difference is non-zero. Statistical testing results revealed that the ten-fold PCC of GEG2P was significantly higher than SOTA methods in most scenarios (Figure S2). These results demonstrate that GEG2P exhibits superior performance across diverse phenotypes in multiple species. We have updated the statistical analysis results of prediction accuracy for GEG2P and the SOTA methods in Section 2.5, as shown in the following:

"Furthermore, we conducted the statistical comparisons of prediction accuracy between GEG2P and six SOTA methods across all the traits of five species using the Wilcoxon signed-rank test (Wilcoxon, 1945). The results indicated that GEG2P achieved significantly higher 10-fold cross-validated PCC than SOTA methods for most traits (Figure S2A~E)."

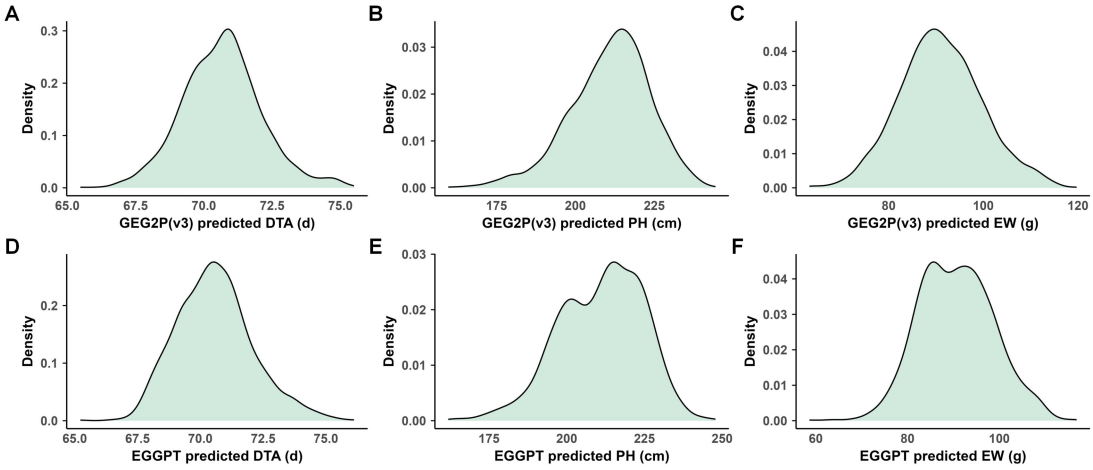

Figure 3.1. Density distributions of the predicted phenotypes of maize. (A)~(C) DTA, PH, and EW predicted by GEG2P, and (D)~(F) DTA, PH, and EW predicted by DGGPT.

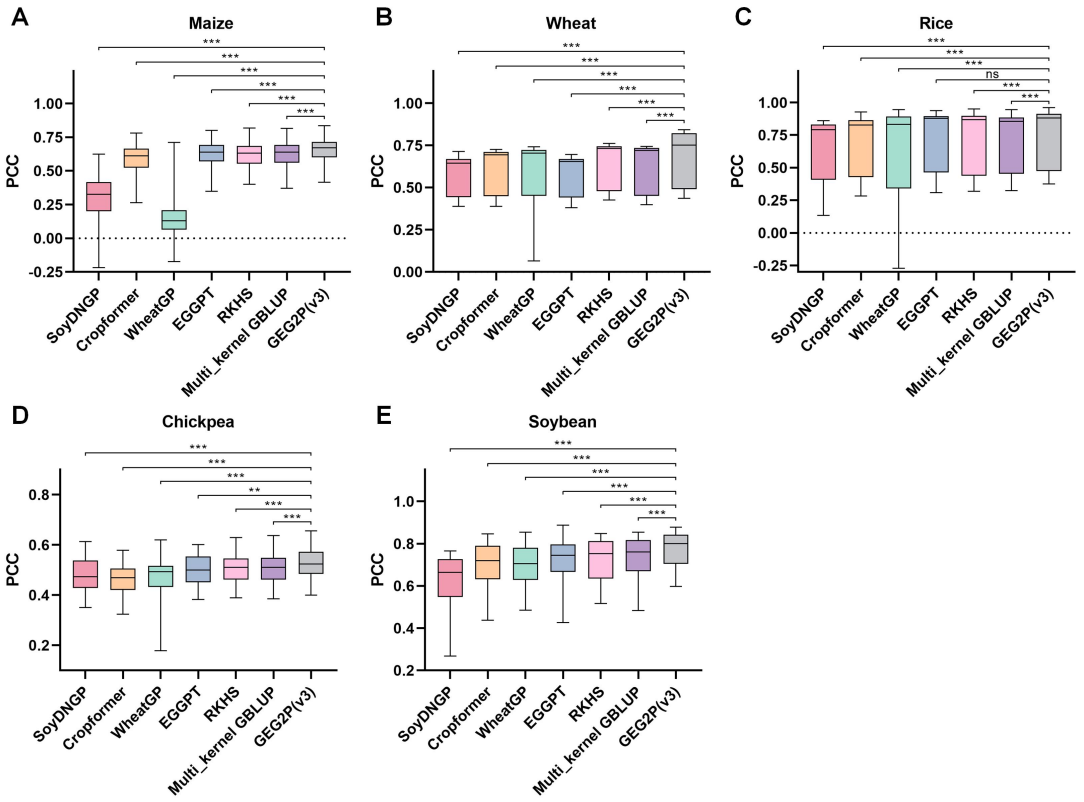

Figure S2. Statistical comparison results of prediction accuracy between GEG2P and six SOTA methods across all the traits of five species. (A) Maize. (B) Wheat. (C) Rice. (D) Chickpea. (E) Soybean. Asterisks indicate significance (* *P*-value < 0.05; ** *P*-value < 0.01; *** *P*-value < 0.001; ns, not significant; Wilcoxon signed-rank test).

In addition, all the references mentioned above are shown in the following:

[1] Wu J, Wang Y, Yu R, *et al*. (2024). EGGPT: an extensible and growing genomic prediction technology. Preprint in Research Square. Available at: https://doi.org/10.21203/rs.3.rs-4581596/v1.

[2] Wilcoxon F (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6): 80-83.

4. Unified Interpretability Baseline and SHAP Implementation (Ref. R1 Comment 6)

The expanded biological validations and citations supporting SHAP's general effectiveness are noted. However, R1's underlying theoretical concern points to the implementation scale. Applying perturbation-based feature attribution to genomic data with tens or hundreds of

thousands of SNPs is computationally exorbitant, particularly when scaled across 20 diverse methods. The authors should elucidate two key theoretical components: 1) how a unified interpretability baseline was established across these heterogeneous models, and 2) what approximation algorithms or computational strategies were utilized to make SHAP computationally feasible at this massive scale.

Response: Thanks for your suggestion. To establish a unified interpretability baseline, the contribution of each SNP to phenotype prediction was quantified via SHAP for all base learners. Moreover, this study primarily analyzed the rankings of SNP SHAP values rather than their absolute magnitudes, ensuring that the interpretability results from different models could be compared on a consistent scale.

From a computational perspective, the exact calculation of SHAP values requires evaluating all possible feature subsets, and the computational complexity increases exponentially with the number of SNPs. Therefore, exact SHAP computation is computationally demanding for high-dimensional genotype data. To reduce the computational burden, we selected suitable SHAP value calculation methods based on the model structure of each base learner. Previous studies have shown that SHAP has been widely used in genomic prediction for feature contribution assessment and candidate loci identification (Wang *et al*., 2024; Wang *et al*., 2025; Wei *et al*., 2026). For the deep learning-based base learners, we used GradientExplainer to calculate the SHAP values of SNPs. GradientExplainer uses representative background samples to approximate the expected model output and estimates the contribution of each SNP to the prediction based on the expected gradients between the input samples and the background samples. This is consistent with the usage of SHAP analysis for feature attribution in previous deep learning-based genomic prediction studies (Sun *et al*., 2026; Li *et al*., 2026). For the machine learning-based base learners, we used KernelExplainer and TreeExplainer for attribution analysis. KernelExplainer was used for non-tree-based models, where the contribution of each SNP to the model output was approximately estimated through feature perturbation and sampling. TreeExplainer was used for tree-based models, where the tree structure was used to improve the computational efficiency of SHAP value estimation. This strategy is consistent with the previous studies using SHAP value calculation for tree-based models (Yu *et al*., 2025). For the statistical base

learners, SNP contributions were estimated using the *iml* package. All attribution analyses were performed using the trained models without retraining. The overall importance of each SNP locus was evaluated using the mean SHAP value across samples, which was used to identify key SNP loci (Wang *et al*., 2024; Yu *et al*., 2025; Li *et al*., 2026). In addition, we used a high-performance server to calculate SHAP values for large-scale SNP loci. The server was equipped with dual Intel Xeon 8480C 2.00 GHz processors, providing 224 logical cores. It also has 2 TB of high-speed DDR5 memory and 220 TB of storage. In addition, the server was equipped with five NVIDIA A100 GPUs, providing 400 GB of total GPU memory. These computing resources supported parallel computing and ensured the feasibility and efficiency of SHAP value calculation for large-scale of SNP loci.

According to your suggestion, we have modified the corresponding content in Section 2.6 and Section 3.3, as shown in the following:

"To mitigate the impact of scale differences in SHAP values across different models on comparative analysis, this study primarily analyzed the relative ranking of SHAP values of SNPs for each base learner."

"To evaluate the contribution of SNPs to model predictions in high-dimensional genotype data, we used appropriate SHAP value calculation methods based on the structure of the base learner (Wang *et al*., 2024; Wang *et al*., 2025; Wei *et al*., 2026; Li *et al*., 2026). For the deep learning-based base learners, we used GradientExplainer to calculate SHAP values of SNPs. For the machine learning-based base learners, we used KernelExplainer and TreeExplainer to calculate SHAP values. Specifically, KernelExplainer was used for non-tree-based models, whereas TreeExplainer was used for tree-based models. For the statistical base learners, we used the *iml* package to calculate the SHAP values of SNPs."

In addition, all the references mentioned above are shown in the following:

[1] Li J, Yu L, Li M, *et al*. (2026). Leveraging weighted embedding and Transformer architecture to improve phenotype prediction of complex traits for crops. *Nature Communications*, 17: 4427.

[2] Sun J, Zhang X, You X, *et al*. (2026). Bayesian neural networks for genomic prediction: uncertainty quantification and SNP interpretation with SHAP and GWAS. *Theoretical and Applied Genetics*, 139(1): 29.

[3] Wang P, Lehti-Shiu M D, Lotreck S, *et al*. (2024). Prediction of plant complex traits via integration of multi-omics data. *Nature Communications*, 15(1): 6856.

[4] Wei L, Jiang Z, Fan B, *et al*. (2026). Automated interpretable artificial intelligence genomic prediction with AIGP. *Genome Research*, 36(4): 814-826.

[5] Wang H, Yan S, Wang W, *et al*. (2025). Cropformer: An interpretable deep learning framework for crop genomic prediction. *Plant Communications*, 6(3).

[6] Yu T, Zhang H, Chen S, *et al*. (2025). EXGEP: a framework for predicting genotype-by-environment interactions using ensembles of explainable machine-learning models. *Briefings in Bioinformatics*, 26(4): bbaf414.

5. Minor Methodological Clarifications (Terminology and Formula Notation)

Upon reviewing the newly added and revised Methods sections, a few specific details require adjustment to prevent ambiguity:

(1) Page 31, Line 972 (Section 4.2.2 Genetic algorithm): Standard algorithmic terms such as "population" and "individual" are used. In the context of genomic prediction, these terms correspond to biological entities (e.g., breeding populations and individual accessions). We recommend adjusting the wording (e.g., using "candidate solutions") to explicitly distinguish the algorithmic parameters from biological datasets for interdisciplinary readers.

Response: Thank you for the suggestion. Yes, the original algorithm terms may cause confusion. In Section 4.2.2, we have revised the algorithm description by replacing “individual” with “candidate solution” and “population” with “candidate solution set”, as shown in the following:

“In GA, an individual represents a candidate solution, and a population consists of multiple candidate solutions. GEG2P uses GA to determine the optimal weights of the base learners. Specifically, GEG2P takes the weight set of the base learners as a candidate solution, and takes multiple sets of weights as a candidate solution set. GA is used to optimize the candidate solution set to determine the optimal weight of each base learner. The GA in GEG2P mainly includes initialization of the candidate solution set, RMSE-based evaluation, selection and evolutionary operations.

Initialization of the candidate solution set

The candidate solution set in GEG2P is shown in Eq. (2).

$$S = \{s_1, s_2, ..., s_n\} \tag{2}$$

In Eq. (2), s_i represents the i-th candidate solution, namely a set of weights assigned to the base learners. n represents the number of candidate solutions, which was set to 50 in GEG2P (Supplemental Data 26). The weights in each candidate solution s_i were initialized according to Eq. (3).

$$s_i = \{w_1, w_2, ..., w_m\} \tag{3}$$

"To simulate the natural selection mechanism in biological evolution, this study used the tournament selection strategy to select candidate solutions with lower RMSE from the candidate solution set for the subsequent evolution. Specifically, GEG2P randomly selected k candidate solutions from the candidate solution set to form a tournament group and then selected the candidate solution with the lowest RMSE as a parent solution. This process was repeated until the predetermined number of parent solutions was obtained. The selected candidate solutions were then used in the subsequent crossover and mutation operations to generate a new candidate solution set. The selection operation is defined in Eq. (5)."

"In Eq. (5), S represents the candidate solution set, and T_i denotes the tournament group formed during the i-th selection. k denotes the number of candidate solutions in the tournament group, s represents the candidate solution in T_i, $RMSE(s)$ represents the prediction error of candidate solution s, and $Parent_i$ represents the parent candidate solution selected for the i-th tournament."

"In order to effectively explore the solution space and maintain the diversity of the candidate solution set, this study adopts a combination strategy of mixed crossover and polynomial mutation. In the mixed crossover operation, the corresponding base learner weights in the parent candidate solutions are averaged in a weighted manner to generate new offspring candidate solutions."

"In Eq. (6), w_{1k} and w_{2k} represent the weights of the k-th base learner in the two parent candidate solutions, respectively. w_k^c represents the weight of the k-th base learner of the offspring candidate solution, and α represents the crossover coefficient."

"In the polynomial mutation operation, this study perturbed the weights assigned to the base learners in the offspring candidate solution within a specified range. For the weight w_k^c of the k-th base learner in the offspring candidate solution, the mutation operation is shown in Eq. (7)."

"To preserve the relative stability of base learner weights in offspring candidate solutions, this study applied a mutation probability to determine whether these weights were perturbed."

"In each generation of the candidate solution set, candidate solutions underwent the operations of selection, crossover, mutation, and their performance was evaluated according to RMSE until the maximum number of generations was reached."

"During the iteration process, GEG2P gradually optimized the weight distribution, ultimately obtaining the weight combination with the lowest RMSE and the optimal phenotype prediction model."

(2) Page 32, Line 992 (Equation 4): We recommend directly denoting this metric as RMSE and removing the explicit notation where j represents the sample number. This will prevent further conceptual confusion between biological individuals in a breeding population and algorithmic individuals in the genetic algorithm.

Response: Thank you for the suggestion. According to your suggestion, we denote the evaluation metric as RMSE and remove the explicit annotation of the sample number j in the formula. We have modified the corresponding formula and related text in the manuscript, as shown in the following: "

RMSE-based evaluation

In GEG2P, RMSE-based evaluation was used to evaluate the quality of different candidate solutions, where each candidate solution represented a set of base learner weights. After integrating various base learners based on the weights in a candidate solution, we used the Root Mean Square Error (RMSE) between the predicted phenotype values by GEG2P and the true phenotype values as the evaluation metric, as shown in Eq. (4). A lower RMSE indicates that the weight set represented by the candidate solution is more effective.

$$RMSE(s) = \sqrt{\frac{1}{M}\left\| \boldsymbol{y} - \sum_{k=1}^{K} \boldsymbol{w}_k \boldsymbol{r}_k \right\|_2^2} \tag{4}$$

In Eq. (4), s denotes a candidate solution, namely a set of base learner weights. M denotes the number of samples in the training set, and K denotes the number of base learners. $\boldsymbol{w}_k$ denotes the weight of the k-th base learner in candidate solution s. $\boldsymbol{r}_k$ denotes the vector of predicted phenotype values generated by the k-th base learner for the training set, and $\boldsymbol{y}$ denotes the vector of true phenotype values in the training set. RMSE measures the prediction error after integrating the base learners according to the weights in candidate solution s. A lower RMSE indicates that the predicted phenotype values are closer to the true phenotype values, suggesting that the weight combination represented by the candidate solution is more effective."

(3) Page 36, Line 1135 (Section 4.2.7 Genetic variance decomposition): Equation (14) presents a variance decomposition. This mathematical decomposition relies on specific underlying statistical conditions. Without explicitly stating these necessary conditions, the decomposition is technically incomplete. Please supplement the formula with its required statistical assumptions.

Response: Thank you for the suggestion. As shown in the response to Comment 1 of Review 3, we employed the BGLR R package to perform genetic variance decomposition. According to your suggestion, we have supplemented the specific underlying statistical conditions in Section 4.2.7, as shown in the following:

"**4.2.7 Genetic variance decomposition**

In this study, the BGLR (Version 1.1.4, Pérez and de los Campos, 2014) R package was used to perform genetic variance decomposition for both Top10% SNPs of individual base learners and the union of Top10% SNPs of Weight-Top-5. Since the maize dataset consists entirely of inbred lines, only additive (G_a) and epistatic (G_e) effects were modeled in the variance decomposition. The mixed linear model is shown in Eq. (14).

$$y = \mu + G_a + G_e + \varepsilon \tag{14}$$

In Eq. (14), y is a vector of phenotypes assumed to follow a normal distribution. μ is the overall mean. $G_a \sim N(0, GRM_a\sigma_a^2)$, $G_e \sim N(0, GRM_e\sigma_e^2)$ are vectors of additive and additive epistatic genotypic effects, respectively. $\mathcal{E} \sim N(0, I_\varepsilon\sigma_\varepsilon^2)$ denotes the model residual. I_ε denotes the corresponding identity matrix. G_a, G_e and $\mathcal{E}$ are assumed to be mutually independent random effects following multivariate normal distributions. GRM_a and GRM_e are the covariance matrices for the additive and additive epistatic effects, which are defined in Eq. (15) and Eq. (16), respectively.

$$GRM_a = \frac{X_{ij}X_{ij}^T}{2\sum_{k=1}^{p} p_k(1-p_k)'} \tag{15}$$

$$GRM_e = GRM_a \odot GRM_a \tag{16}$$

In Eq. (15), $X_{ij} = (x_{ij} - 2p_j)$, where x_{ij} represents the genotype coding of the i^{th} genotype at the j^{th} SNP marker, and p_j is the reference allele frequency at the j^{th} SNP marker. X_{ij}^T is the transpose matrix of X_{ij}. Accordingly, the covariance matrices for additive epistatic effects in Eq. (16) were approximated using the Hadamard product of GRM_a.

Following the analytical framework of Gogna *et al*. (2026), additive (σ_a^2), additive epistatic (σ_e^2), and residual ($\sigma\varepsilon^2$) variances were estimated based on the Bayesian regression framework implemented in BGLR package. The proportions of additive, additive epistatic, and residual variance were estimated using Eq. (17), (18), and (19).

$$PVE_a = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2 + \sigma_\varepsilon^2} \tag{17}$$

$$PVE_e = \frac{\sigma_e^2}{\sigma_a^2 + \sigma_e^2 + \sigma_\varepsilon^2} \tag{18}$$

$$PVE_\varepsilon = \frac{\sigma_\varepsilon^2}{\sigma_a^2 + \sigma_e^2 + \sigma_\varepsilon^2} \tag{19}$$

”

In addition, all the references mentioned above are shown in the following:

[1] Gogna A, Kamali B, Wimmer V, *et al*. (2026). Predicting enviromically adapted varieties with big data. *Genome Biology*, 27(1): 3.

[2] Pérez P, de los Campos G. (2014). Genome-wide regression and prediction with the

BGLR statistical package. *Genetics*, 198(2): 483-495.

(4) Page 1, Line 21 (Abstract): The manuscript reports a 10.79% performance improvement compared against "single base learners." It is currently ambiguous whether this baseline represents the average performance of all 20 individual models or the performance of the single best constituent model. Evaluating an optimized ensemble against an average that incorporates poorly performing models risks artificially inflating the perceived gain. We request that the authors explicitly define this comparative baseline to ensure the reported metrics accurately reflect the true algorithmic advancement.

Response: Thank you for the suggestion. We have redefined the comparison baseline as the single base learner with the highest prediction accuracy. Based on this revised baseline, we calculated the performance improvement of GEG2P. The results showed that GEG2P achieved an average improvement of 4.02% in prediction accuracy across the maize, rice, wheat, chickpea, and soybean datasets compared with the best-performing single base learner. We have modified the description in the abstract, as shown in the following:

“Compared with the best-performing single base learners, GEG2P achieved an average improvement of 4.02% in prediction accuracy across maize, wheat, rice, chickpea, and soybean.”

Reviewer #4 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Response: Thank you very much.