Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

KANG2P enables interpretable genotype-to-phenotype prediction through genetically anchored molecular-state learning

Xiao Luo

xluo@hnu.edu.cn

Hunan University https://orcid.org/0000-0002-3914-359X

Lei Ling

Hunan Research Center of the Basic Discipline for Cell Signaling, College of Biology, Hunan University, Changsha, China

Yang Liu

Hunan Research Center of the Basic Discipline for Cell Signaling, College of Biology, Hunan University, Changsha, China

Pingzhou Wang

Hunan Research Center of the Basic Discipline for Cell Signaling, College of Biology, Hunan University, Changsha, China

Fengxin Li

Hunan Research Center of the Basic Discipline for Cell Signaling, College of Biology, Hunan University, Changsha, China

Hang Ning

College of Forestry and Horticulture, Hubei Minzu University, Enshi, China

Zhiming Zhang

Hunan Research Center of the Basic Discipline for Cell Signaling, College of Biology, Hunan University, Changsha, China

Cheng Yang

College of Computer Science and Electronic Engineering, Hunan University, Changsha, China

Liang Qiao

Taichu (Wuxi) Electronic Technology Co., Ltd., Wuxi, China

Si Qin

Laboratory of Food Function and Nutrigenomics, College of Food Science and Technology, Hunan Agricultural University, Changsha, China

Alexander Schönhuth

Bielefeld University https://orcid.org/0000-0003-3529-0856

Taihong Wu

Hunan University https://orcid.org/0000-0002-9760-6978

Yuansheng Liu

Hunan University https://orcid.org/0000-0002-7680-3155

Article

Keywords:

Posted Date: October 6th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10772215/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

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

KANG2P enables interpretable genotype-to-phenotype prediction through genetically anchored molecular-state learning

Lei Ling[1,†], Yang Liu[1,†], Pingzhou Wang[1,†] Fengxin Li[1],
Hang Ning[4], Zhimin Zhang[1], Cheng Yang[2,3], Liang Qiao[5,6], Si Qin[7],
Alexander Schönhuth[8], Taihong Wu[1,*], Yuansheng Liu[2,3,*], Xiao Luo[1,*]

[1] Hunan Research Center of the Basic Discipline for Cell Signaling, College of Biology, Hunan University, Changsha, China
[2] College of Computer Science and Electronic Engineering, Hunan University, Changsha, China
[3] Yuelushan Laboratory, Changsha, China
[4] College of Forestry and Horticulture, Hubei Minzu University, Enshi, China
[5] Taichu (Wuxi) Electronic Technology Co., Ltd., Wuxi, China
[6] Department of Computer Science and Technology, Tsinghua University, Beijing, China
[7] Laboratory of Food Function and Nutrigenomics, College of Food Science and Technology, Hunan Agricultural University, Changsha, China
[8] Faculty of Technology, Bielefeld University, Bielefeld, Germany
[†]These authors contributed equally to the work.
[*]To whom correspondence should be addressed.
Email: taihongwu@hnu.edu.cn
Email: yuanshengliu@hnu.edu.cn
Email: xluo@hnu.edu.cn

Abstract

Genotype-to-phenotype (G2P) mapping represents a central scientific challenge in genetics, necessitating advancements in both prediction performance and mechanistic interpretability. The complexity of biological systems creates a substantial gap between genotypic variations and phenotype. While multi-omics integration strategies strive to bridge this gap, they face limitations including prohibitive costs and challenges in acquiring large-scale cohorts. To address these constraints, we present KANG2P, an interpretable genotype-to-phenotype framework that bridges these settings by transferring genetic regulatory information learned from external reference panels to target cohorts. Modality-specific autoencoders learn compact representations of genotype and genetically predicted molecular profiles, which are subsequently integrated by a Kolmogorov–Arnold Network to model the nonlinear path from genetic variation through molecular regulation to phenotype. Across human neurodegenerative diseases and crop agronomic traits, KANG2P outperformed conventional and deep-learning G2P models. Notably, the benefit of genetically predicted molecular profiles extended beyond KANG2P itself, as their integration also improved established G2P models. Beyond predictive accuracy, KANG2P recovered established amyotrophic lateral sclerosis genes and newly prioritized *SLC1A2* as the leading genome-wide candidate. Functional analysis of its conserved *C. elegans* orthologue, *glt-6*, revealed locomotor impairment and heightened paraquat sensitivity, providing in vivo support for the disease relevance of the *SLC1A2–glt-6* axis. Together, these findings establish KANG2P as a scalable, accurate and experimentally grounded framework for interpretable G2P mapping across species.

Introduction

Establishing genotype-to-phenotype (G2P) relationships is a central goal of genetics, with broad implications for understanding disease susceptibility, precision medicine and crop improvement [1, 2, 3]. Classical genomic prediction approaches, including genomic best linear unbiased prediction (GBLUP) [4], ridge-regression BLUP (rrBLUP) [5] and Bayesian models [6, 7], have provided powerful and scalable frameworks for estimating genetic contributions to complex traits. However, these approaches largely rely on additive assumptions and may not fully capture the nonlinear genetic architectures, epistatic interactions and regulatory processes that shape complex phenotypes.

To capture potential nonlinear relationships between genotype and complex traits, deep-learning approaches have increasingly been explored in G2P prediction [8]. By learning hierarchical representations from high-dimensional genome-wide data, these models provide a flexible alternative to conventional linear genomic prediction frameworks [9, 10, 11]. Models such as DeepGS [12], DNNGP [13] and Cropformer [14] have demonstrated that deep-learning models can improve prediction by capturing complex patterns beyond conventional statistical models. Nevertheless, most deep-learning G2P models remain difficult to interpret, limiting their ability to reveal biological mechanisms, prioritize candidate genes and generate experimentally testable hypotheses [15, 16]. This lack of interpretability represents a major barrier for translating predictive models into biological discovery. DiseaseCapsule [3] addressed part of this limitation by introducing a capsule-network architecture for interpretable disease-risk prediction. However, DiseaseCapsule relies on genotype profiles alone, leaving the molecular intermediates that connect genetic variation to phenotype largely unmodelled.

A complementary challenge is that genetic variation influences phenotypes through intermediate molecular states, including transcriptomic and proteomic regulation. Multi-omics integration provides a natural strategy to bridge genotype and phenotype by incorporating these regulatory layers, and recent approaches have demonstrated improved phenotype prediction through the integration of multiple molecular modalities [13, 17, 18, 19]. However, most existing multi-omics prediction frameworks require experimentally measured molecular profiles from the same individuals used for phenotype prediction. Such matched multi-omics datasets remain costly, difficult to scale and often unavailable for biologically important tissues, such as the human brain.

Genetically predicted molecular states provide an opportunity to overcome this limitation by reconstructing molecular phenotypes from genotype data alone. Inspired by the PrediXcan framework and related genotype-based molecular profiles prediction approaches [20, 21], genetically regulated transcriptomic and proteomic profiles can serve as intermediate representations linking DNA variation to phenotype. However, effectively exploiting these molecular states requires models capable of capturing nonlinear interactions while maintaining interpretability.

Kolmogorov–Arnold Networks (KANs) provide a promising foundation for this challenge. Unlike conventional multilayer perceptrons that rely on fixed activation functions and learnable linear weights, KANs parameterize learnable functions on network edges, enabling flexible approximation of complex multivariate relationships while retaining interpretable functional representations [22]. This property makes KANs particularly attractive for modelling biological systems, where phenotypes emerge from nonlinear interactions between genetic variation, molecular regulation and environmental context. Beyond their nonlinear modelling capacity, KANs offer an inherent interpretability advantage: the learned edge functions explicitly describe how individual input variables are transformed and contribute to the final prediction, providing a more direct link between model decisions and underlying biological mechanisms than conventional black-box neural networks. Such functional transparency is particularly valuable for G2P mapping, where the ultimate goal is not only accurate prediction but also the identification of biologically meaningful genetic and molecular determinants.

Here we introduce KANG2P, a genetically anchored molecular-state learning framework powered by Kolmogorov–Arnold Networks for interpretable genotype-to-phenotype mapping. KANG2P first reconstructs individualized genetically regulated molecular states from genotype data using external genotype–omics reference panels, extending this strategy from transcriptomic to proteomic prediction. These molecular states are subsequently integrated with genotype through modality-specific representation learning and gated fusion, and the resulting representation is modelled by a KAN-based predictor to

capture nonlinear genotype–molecular state–phenotype relationships. Across human neurodegenerative diseases and crop agronomic traits, KANG2P achieves accurate prediction while retaining gene-level interpretability for candidate discovery and experimental validation.

Results

Workflow of KANG2P

KANG2P was designed as a three-module framework for genetically anchored molecular-state learning and phenotype prediction (Fig. 1a). First, genotype–omics relationships learned from reference panels were transferred to target cohorts to reconstruct individualized genetically regulated molecular states. Depending on data availability, these states comprised predicted transcriptomic or proteomic profiles. Second, genotype and predicted molecular profiles were processed by separate autoencoders to learn compact modality-specific representations. Third, the resulting latent representations were integrated by a Kolmogorov–Arnold Network for disease classification or quantitative-trait regression. Gene-level attribution scores derived from the fitted model were subsequently used for biological interpretation and candidate-gene prioritization.

Module 1: Reconstruction of genetically regulated molecular states

Molecular-state reconstruction in human disease cohorts In the first module, individualized genetically regulated molecular states were reconstructed from genotype data. For transcriptomic-state reconstruction, we used pre-trained gene-expression prediction models from version 8 of the Genotype–Tissue Expression (GTEx) Project [23], which are publicly available through the PredictDB repository PredictDB. These elastic-net models were trained using matched genotype and multi-tissue transcriptomic data and estimate the *cis*-genetically regulated component of gene expression from local SNPs. Although deep-learning approaches have been developed for individual-level expression prediction, recent benchmarks have not demonstrated a consistent advantage over regularized linear models for *cis*-genetic prediction [24, 25]. We therefore used the established GTEx v8 elastic-net models to obtain robust and transferable representations of genetically regulated expression.

Given the focus on neurodegenerative diseases, the GTEx v8 predictors were applied to genotype data from the target cohorts to generate individualized expression profiles across 13 brain tissues. Tissue-specific gene selection was incorporated into the nested cross-validation procedure. Within the relevant training partition of each fold, Wilcoxon rank-sum tests were first used to identify genes whose predicted expression differed between cases and controls. The minimum-redundancy maximum-relevance (mRMR) algorithm [26] was subsequently applied to retain genes that were informative for disease status while minimizing redundancy among selected features. The resulting tissue-specific gene sets were then applied unchanged to the corresponding validation or held-out test samples (Fig. 1b).

For disease analyses in which a proteomic layer was included, published genotype-to-protein abundance prediction models were applied to target-cohort genotype data to generate individualized genetically predicted protein-abundance profiles [27]. These models were used without retraining, enabling proteomic-state reconstruction without requiring matched protein measurements in the target cohort.

Molecular-state reconstruction in crop populations Because broadly transferable pre-trained expression-prediction models were not available for the crop cohorts, gene-expression predictors were trained directly in species-specific reference populations containing matched genotype and transcriptome data. For each gene, elastic-net regression was used to model expression from local *cis*-SNPs, thereby learning the genetically regulated component of expression within the reference population. Predictive performance was evaluated by cross-validation within the reference panel, and genes within the upper decile of cross-validated prediction performance were retained for downstream analysis.

The fitted prediction models were subsequently applied, without retraining, to independent rice and maize cohorts for which genotype data were available. This transfer generated individualized genetically

predicted expression profiles in the target samples while avoiding the need for matched transcriptomic measurements. The resulting profiles represented the crop-specific genetically regulated molecular state and were used as the molecular input branch for downstream agronomic-trait prediction (Fig. 1c).

Module 2: Modality-specific representation learning

Genotype and genetically predicted molecular profiles were encoded in parallel by modality-specific autoencoders. Each encoder used linear transformations, LeakyReLU activation, layer normalization, dropout and residual connections to learn compact latent representations, while modality-specific reconstruction objectives regularized the resulting embeddings. Rather than directly concatenating the two representations, KANG2P used a learnable sigmoid gate to adaptively balance the contributions of genotype and molecular-state information. The resulting fused representation therefore retained complementary signals while suppressing redundant or noisy features before being passed to the phenotype-prediction module (Fig. 1d).

Module 3: KAN-based phenotype prediction and interpretation

The gated multimodal representation $\mathbf{z}_F$ was passed to a FourierKAN predictor for task-specific phenotype estimation (Fig. 1d). The predictor comprised two FourierKAN layers interleaved with LeakyReLU activation, layer normalization and dropout. The first layer transformed $\mathbf{z}_F$ into a nonlinear hidden representation, whereas the second generated either a disease-risk logit for classification or a continuous phenotype value for agronomic-trait regression.

Unlike spline-based KAN implementations, FourierKAN parameterizes each learnable edge function using a truncated Fourier expansion. This formulation replaces conventional affine transformations and fixed activation functions with adaptive univariate functions expressed in a Fourier basis, enabling flexible modelling of nonlinear relationships within the fused genotype and molecular-state representation. Model-attribution analysis was subsequently used to quantify the contribution of molecular features to each phenotype. Feature-level attribution scores were aggregated at the gene level to support candidate-gene prioritization, biological interpretation and downstream experimental assessment.

KANG2P achieves state-of-the-art performance for ALS risk prediction

We first evaluated KANG2P for amyotrophic lateral sclerosis (ALS) risk prediction using datasets from previous large-scale genetic studies [28, 29, 30]. KANG2P was benchmarked against DiseaseCapsule, a published capsule-network model for genotype-based disease prediction [3], and four conventional machine-learning methods: logistic regression (LR), support vector machine (SVM), random forest (RF) and AdaBoost. Models were assessed using genotype (G), genetically predicted expression (PE) and their integrated representation (G+PE) (Fig. 2).

KANG2P delivered the best performance in genotype only was input settings. Using genotype alone, KANG2P achieved an F1 score of 0.850 ± 0.007, significantly higher than LR (0.815 ± 0.012), the strongest classical classifier in this setting ($P = 0.03$, Extended Data Fig. 1a). KANG2P also outperformed the previously reported DiseaseCapsule model in the same genotype setting (DiseaseCapsule, 0.791 ± 0.023), corresponding to an absolute F1 gain of 5.9%. The advantage became larger after integrating predicted expression with genotype. In the G+PE setting, KANG2P achieved the highest overall ALS performance, with an F1 score of 0.871 ± 0.012 and an accuracy of 0.868 ± 0.009. Compared with LR (0.816 ± 0.013), KANG2P (G+PE) improved F1 by 5.5% ($P = 0.03$, Extended Data Fig. 1a), and compared with DiseaseCapsule (0.785 ± 0.029), the improvement was 8.6% units.

When used alone, PE provided a measurable but incomplete representation of disease risk, with KANG2P achieving an F1 score of 0.694 ± 0.063 (Extended Data Fig. 1a). Integrating PE with G increased KANG2P performance from 0.850 to 0.871 in F1 score and from 0.845 to 0.868 in accuracy. The G+PE model also achieved the highest recall (0.891 ± 0.034) among all models, suggesting improved sensitivity to ALS cases. Together, these results show that predicted gene expression is not merely a

stand-alone predictor, but provides complementary molecular information that enhances genotype-based ALS risk prediction when integrated within KANG2P.

KANG2P generalizes to AD and benefits from genetically predicted molecular profiles

We next examined whether the predictive advantage of KANG2P extended to Alzheimer's disease (AD) dataset [31], a genetically complex neurodegenerative disorder for which both transcriptomic and proteomic dysregulation are implicated. In contrast to ALS, where genotype and predicted expression were evaluated, the AD analysis further incorporated predicted proteomic profiles (PP), allowing us to assess whether multiple genetically predicted molecular layers could improve disease-risk prediction.

Across all AD feature settings, KANG2P achieved the highest mean F1 score, demonstrating a consistent advantage over both conventional machine-learning models and the DiseaseCapsule benchmark (3). The improvement was most pronounced in the genotype, predicted-expression and G+PE settings. Using genotype alone, KANG2P increased F1 score from 0.679 ± 0.018 for LR, the strongest non-KANG2P comparator, to 0.720 ± 0.015 ($P = 0.03$, Extended Data Fig. 1b). KANG2P also markedly outperformed DiseaseCapsule in the same setting (0.605 ± 0.018), corresponding to an absolute F1 gain of 11.5%. A similar pattern was observed for PE alone, where KANG2P achieved an F1 score of 0.680 ± 0.003, compared with 0.597 ± 0.017 for RF, the second-best method ($P = 0.03$,Extended Data Fig. 1b), and 0.540 ± 0.069 for DiseaseCapsule. When predicted expression was integrated with genotype, KANG2P again outperformed the strongest comparator, improving F1 from 0.672 ± 0.012 for LR to 0.719 ± 0.017 ($P = 0.03$), and exceeding DiseaseCapsule by 10.5% F1 units (DiseaseCapsule, 0.614 ± 0.021).

For the predicted proteome modality, we used the published genotype-to-protein prediction weights [27] to project individual genotype profiles into genetically regulated protein-abundance profiles (see Methods), thereby enabling disease-risk prediction from an imputed proteomic layer. Across 1,761 proteins with available FUSION [21] prediction models, protein-expression heritability showed a right-skewed distribution, with a median h^2 of 0.102, an upper quartile of 0.202 and a 90th percentile of 0.350 (Extended Data Fig. 2a), indicating widespread but heterogeneous genetic regulation of protein abundance. Cross-validated predictability increased with heritability across the five FUSION models, and 410 proteins passed both the upper-quartile heritability threshold and a cross-validated $R^2 \geq 0.1$, defining a set of genetically regulated proteins that could be robustly imputed from genotype for downstream disease-risk prediction (Extended Data Fig. 2b).

Adding predicted proteomic profiles to genotype yielded a smaller margin over the second-best model in the G+PP setting (KANG2P, 0.718 ± 0.014; second-best, 0.716 ± 0.010; $P = 0.41$), indicating that PP alone did not substantially increase separability beyond genotype in this configuration. However, the full G+PE+PP representation achieved the highest F1 score among all AD models (0.724 ± 0.014), with a trend towards improvement over the second-best comparator (0.711 ± 0.005; $P = 0.06$). Notably, this trimodal model also increased recall relative to genotype-only KANG2P (0.839 ± 0.034 versus 0.800 ± 0.036), suggesting improved sensitivity to AD cases. Thus, while genotype remained a strong predictor of AD risk, integrating genetically predicted expression and proteomic profiles provided additional molecular information that enhanced case detection and yielded the best overall KANG2P performance.

KANG2P demonstrates excellent performance in the Rice18K population

Gene-expression predictors were trained on the Rice277 reference panel and then transferred to the Rice18K population to yield genome-wide imputed transcriptomes. The distribution of *cis* SNP counts within the gene expression level prediction models is illustrated in Supplementary Fig. 1, while the distribution of prediction accuracy for these gene expression level prediction models is presented in Supplementary Fig. 2.

We next evaluated KANG2P on the Rice18K dataset [32] to determine whether its predictive advantage generalizes to quantitative agronomic traits. We implemented BayesC, Bayesian ridge regression (BRR), Bayesian LASSO (BL), reproducing kernel Hilbert space (RKHS) regression, sparse partial least squares

(SPLS), random forest regression (RFR), support vector regression (SVR) and LASSO using the G2P R package (`https://github.com/G2P-env/G2P`). However, nested cross-validation of these R-based implementations on a dataset of approximately 18,000 samples proved computationally prohibitive and could not be completed within a practical time frame under the same evaluation protocol. We therefore focused the large-scale comparison on GPU-accelerated deep-learning models, including DeepGS, Cropformer and DNNGP, while retaining ridge-regression best linear unbiased prediction (RR-BLUP) as a conventional linear baseline. Five agronomic traits were analysed, including plant height, culm length, grain length, grain width and grain yield, using genotype (G), predicted expression (PE) and the integrated representation (G+PE). KANG2P retained the strongest overall performance across the three evaluation metrics (Fig. 4). Across traits and outer folds, KANG2P achieved the highest R^2 and Pearson correlation coefficient (PCC), together with the lowest mean squared error (MSE), in all three feature settings. With genotype alone, KANG2P reached an R^2 of 0.663 ± 0.174, a PCC of 0.807 ± 0.115 (Supplementary Fig. 3) and an MSE of 0.006 ± 0.002 (Supplementary Fig. 4), outperforming RR-BLUP, the strongest non-KANG2P comparator, across all three metrics. The improvement over RRBLUP was 1.4% in R^2 and 0.009 in PCC, accompanied by a 4.6% reduction in MSE.

Predicted expression also carried substantial phenotype-relevant signal. When PE was used alone, KANG2P achieved an R^2 of 0.638 ± 0.168 and a PCC of 0.793 ± 0.112, while reducing MSE to 0.007 ± 0.002. This exceeded the strongest PE-based comparator, DeepGS, by 2.5% in R^2 and 1.3% in PCC, and reduced MSE by 7.6%. Thus, predicted expression was informative even without direct genotype input, indicating that the genetically regulated transcriptomic layer captures quantitative trait variation in rice.

The best overall performance was obtained by integrating genotype and predicted expression. In the G+PE setting, KANG2P achieved the highest aggregate R^2 (0.665 ± 0.168) and PCC (0.809 ± 0.111), and the lowest MSE (0.006 ± 0.002). Compared with the strongest non-KANG2P model in the same setting, RR-BLUP, KANG2P improved R^2 by 1.3% and PCC by 0.9%, while reducing MSE by 4.3%. Importantly, G+PE improved over genotype-only KANG2P in four of the five traits across all three metrics, with the largest gains observed for grain yield, the most difficult trait to predict. Together, these results show that KANG2P provides robust and consistently superior prediction of Rice18K quantitative traits, and that predicted expression contributes complementary information that is most effectively exploited when integrated with genotype.

Gene-level attribution analysis placed *D3*, *Ghd7* and *GW2* among the highest-scoring genes in the KANG2P model for grain yield prediction (Extended Data Fig. 3). These loci represent largely distinct components of rice yield architecture: *D3* regulates tillering and the balance between shoot branching and productive panicle formation through strigolactone signalling [33, 34, 35]; *Ghd7* pleiotropically coordinates heading date, plant height and grain number per panicle [36, 37, 38, 39]; and *GW2* restricts spikelet-hull growth and grain weight through ubiquitin-mediated control of cell proliferation and expansion [40, 41, 42]. Thus, the model recovered established regulators of panicle number, grains per panicle and grain mass—the principal multiplicative determinants of rice yield—supporting the biological validity of its gene-level attribution landscape.

KANG2P remains advantageous in sample-limited crop populations

The classification and regression tasks described above were conducted on large-scale cohorts comprising tens of thousands of samples. In crop genomic prediction, evaluating model performance under small-sample scenarios is essential, primarily because the scale of G2P datasets typically amounts to approximately 1,000 samples in practical breeding programs—rendering small-sample conditions the common real-world context rather than an exception. We evaluated the predictive performance of our proposed KANG2P model against 12 popular benchmark methods for agronomic traits in the Rice1495 [43] and Maize1404 dataset [44]. We benchmarked KANG2P against a broad panel of established G2P predictors, including linear and Bayesian regression models, kernel- and ensemble-based methods, sparse partial least squares and three deep-learning architectures, DeepGS, DNNGP and Cropformer. The methodology behind all benchmark models, including their algorithmic principles and implementation, is outlined in the Methods section within the main text.

We trained gene-expression prediction models in the Maize CUBIC reference panel and applied the predictors to the remaining 1,010 genotyped samples, generating genome-wide imputed transcriptomes for downstream phenotype prediction. The distribution of *cis* SNP counts within the gene expression level prediction models is illustrated in Supplementary Fig. 5. For the 15,625 genes of gene expression level prediction models on the Maize CUBIC population, the 90th quantiles of R^2 and spearman correlation coefficient were 0.446 and 0.624, respectively, while the 10th quantile of MSE was 0.085 (Supplementary Fig. 6).

KANG2P showed a highly consistent advantage across these smaller cohorts (Fig. 5 and 6). Considering the four traits, three input representations and three evaluation metrics in each species, KANG2P ranked first in 70 comparisons, including all PE-based and G+PE-based comparisons (Fig. 5 and 6, Supplementary Fig. 7-10). The advantage was strongest in Rice1495, where KANG2P ranked first in all 36 comparisons. In Maize1404, KANG2P ranked first in 34 of 36 comparisons, with the only exceptions restricted to genotype-only prediction of Kernel weight per ear. When aggregated across traits and folds, KANG2P outperformed the strongest non-KANG2P comparator in every dataset–modality–metric setting. In Rice1495, the average improvement over the best non-KANG2P model was 2.1% in R^2 and 1.5% in Pearson correlation coefficient (PCC), accompanied by a 4.7% reduction in mean squared error (MSE). In Maize1404, the corresponding gains were 2.7% in R^2 and 2.0% in PCC, together with a 3.2% reduction in MSE. These results indicate that KANG2P remains the most robust predictor even when training sample size is reduced to approximately one thousand individuals.

The integration of genetically predicted expression further improved prediction, particularly in Rice1495. Within KANG2P, G+PE increased average R^2 from 0.432 to 0.471 and PCC from 0.611 to 0.642, while reducing MSE by 8.6% relative to genotype alone. The gain was observed across all four rice traits, with the largest improvement for heading date, where G+PE increased R^2 by 7.5% and PCC by 4.5%, and reduced MSE by 23.7%.

Importantly, the gain from predicted expression was not specific to KANG2P, but was also observed across established G2P models. With all fold-level results retained, G+PE improved the fold-averaged performance of the 12 non-KANG2P models in 85 of 96 model–trait comparisons for R^2, 86 of 96 comparisons for PCC and 85 of 96 comparisons for MSE. This effect was particularly consistent in Rice1495, where G+PE improved every paired comparison across traits, models and metrics.

Among conventional G2P models, the largest benefit was observed for sparse partial least square (SPLS). Across the two small crop cohorts, integrating PE with genotype increased SPLS performance by 19.5% in R^2 and 5.7% in PCC, while reducing MSE by 18.3%. The improvement was observed for every trait in both Rice1495 and Maize1404, and was especially pronounced in Rice1495, where SPLS gained 23.9% in R^2 and 8.2% in PCC, with a 21.9% reduction in MSE. Other conventional models also benefited substantially, including SVM (ΔR^2 = 7.2% ΔPCC = 3.4%; MSE reduction, 11.7%) and LASSO (ΔR^2 = 4.7%; ΔPCC = 4.9%; MSE reduction, 9.2%). The three deep-learning G2P baselines showed the same trend, with the largest gain observed for Cropformer (ΔR^2 = 9.7%; ΔPCC = 3.2%; MSE reduction, 12.7%), followed by DeepGS (ΔR^2 = 4.1%; ΔPCC = 3.5%; MSE reduction, 6.5%) and DNNGP (ΔR^2 = 3.0%; ΔPCC = 2.2%; MSE reduction, 6.9%). Thus, genetically predicted expression acts as a broadly useful molecular layer, improving not only KANG2P but also both conventional and deep-learning genotype-to-phenotype prediction models.

Integrated prioritization nominates the conserved *SLC1A2–glt-6* axis for functional validation

We first used dual-branch KANG2P model to generate an interpretable gene-level attribution profile across the human genome. Gradient-based attribution scores were assigned to 18,265 genes, producing a model-derived ranking of genes contributing to the prediction task (Methods).

To further assess whether the KANG2P attribution ranking recovered established ALS biology, we examined curated definitive ALS genes from ALSoD database https://alsod.ac.uk/ within the top-ranked 150 genes. Ten definitive ALS genes were present in the top 150, including *UNC13A* (rank 2), *C9orf72* (rank 11), *PFN1* (rank 19), *VAPB* (rank 34), *FUS* (rank 36), *TARDBP* (rank 57), *NEK1* (rank 62) and *TBK1* (rank 85), *SOD1* (rank 122)and *OPTN* (rank 125) (Fig. 7a).

We therefore asked whether the KANG2P-derived ranking contained coherent disease-relevant biological signal beyond individual high-ranking genes. Gene set enrichment analysis (GSEA, Methods) of the dual-branch gradient prior-aware gene priority scores identified *superoxide metabolic process* (GO:0006801) as the most significantly enriched pathway (NES = 1.43, $P = 1.1 \times 10^{-6}$, FDR = 0.006; Fig. 7b). The leading-edge subset comprised 20 genes concentrated within the top 11% of the ranked list, spanning ranks 17–1,949 of 17,377 protein-coding genes, and included the established ALS gene *SOD1* (rank 122) together with *SOD2*, *NOS3*, *NOX3*, *NCF2*, *GSTP1* and *CYB5R4*, all of which encode enzymes directly involved in reactive oxygen species generation or detoxification. This result is consistent with the known role of *SOD1* in familial ALS and with oxidative stress as a mechanism implicated in motor neuron degeneration [45].

The enrichment signal was further supported by over-representation analysis (Methods) of the top 500 ranked genes, which recovered superoxide-related GO terms as significantly enriched (FDR < 0.01; Supplementary Fig. 11). Beyond oxidative stress, GSEA also highlighted pathways related to mRNA catabolism and post-transcriptional regulation (NES = 1.20–1.34, FDR = 0.07–0.21; Supplementary Fig. 12), consistent with ALS-associated RNA-binding protein dysfunction, including the presence of *ELAVL1* in the leading-edge subset. Neural tube patterning was also enriched at a nominal level (NES = 1.41, FDR = 0.24), suggesting that the KANG2P-derived ranking may additionally capture neurodevelopmental processes relevant to motor neuron specification.

We next asked whether KANG2P-prioritized ALS candidate genes could nominate potential drug-associated targets. Using DeepDTAGen [46], we evaluated the predicted affinities of two approved ALS drugs, riluzole and edaravone, against proteins encoded by the top 500 KANG2P-ranked genes (Extended Data Fig. 4a). Both drugs yielded affinity predictions for 409 mappable proteins. The two compounds showed partially overlapping but distinct high-affinity candidate profiles, with *CALM3*, *OR56A1*, *GPR26* and *SLC1A1* ranking highest for riluzole, and *OR56A1*, *SLC1A1*, *CALM3* and *GDF7* ranking highest for edaravone (Extended Data Fig. 4a). To provide an orthogonal in silico assessment, we further performed AutoDock docking for the top DeepDTAGen-ranked drug–target pairs. Several pairs showed favourable docking scores of approximately $-6, \mathrm{kcal}, \mathrm{mol}^{-1}$ or lower, including edaravone–*SLC1A1*, riluzole–*OR56A1*, riluzole–*SLC1A1* and edaravone–*OR56A1* (Extended Data Fig. 4b,c). Integrating predicted affinity with docking support highlighted *SLC1A1* as a convergent candidate target for both ALS drugs, with consistent computational binding evidence across the two analyses (Extended Data Fig. 4d,e).

To nominate candidates for experimental follow-up, we moved beyond the KANG2P attribution ranking and constructed a weighted evidence-integration framework (Methods). This strategy reflected the premise that no single ranking criterion can fully capture disease relevance. To identify high-confidence ALS candidates, we integrated KANG2P-derived attribution scores with independent lines of evidence, including gene-prioritization results from complementary approaches [3, 47], ALS-specific evidence categories curated in the ALSoD database (`https://alsod.ac.uk/`) and transcriptome evidence [48]. Under this integrated scoring scheme, *SLC1A2* emerged as the highest-ranked human candidate, with the top 20 prioritized genes shown in Fig. 7c. We subsequently focused functional assessment on prioritized genes for which the corresponding *C. elegans* orthologues and mutant strains were experimentally accessible. These included *glt* family mutants corresponding to *SLC1A2*, *amph-1* corresponding to *AMPH*, *nex-3* corresponding to *ANXA1*, and *anoh-1* corresponding to *ANO2*.

Because motor impairment is a defining feature of ALS, we first assessed locomotor behaviour in *C. elegans* strains carrying mutations in orthologues of the prioritized human genes. Mutants of *amph-1*, *anoh-1*, *glt-6* and *nex-3* exhibited significant locomotor deficits relative to the corresponding controls (Fig. 7d), providing functional evidence that these conserved genes contribute to the maintenance of motor function.

Oxidative stress has been widely implicated in ALS pathogenesis, and mutations in sod-1 represent one of the most extensively studied genetic causes of familial ALS [49]. To further investigate whether *amph-1*, *anoh-1*, *glt-6* and *nex-3* are involved in oxidative stress-related neurodegeneration, we exposed animals to paraquat-induced oxidative stress and assessed locomotor performance and the *sod-1*(rt451[sod-1G85R]) strain was included as a positive control. Under oxidative stress conditions, the locomotor defect of

glt-6 animals was significantly exacerbated compared with untreated controls (Fig. 7e), indicating an increased susceptibility to oxidative damage.

Survival assays further separated candidates with general locomotor effects from those conferring oxidative-stress vulnerability. Mutations in *sod-1*, *glt-6*, *amph-1* and *anoh-1* increased sensitivity to paraquat-induced oxidative injury, whereas *nex-3* mutants did not show a clear survival defect under the same condition (Fig. 7f). Among the tested genes, *glt-6* was supported by multiple lines of evidence: it corresponded to the top-ranked human candidate *SLC1A2* in the integrated prioritization framework, showed locomotor impairment in the worm screen, displayed stress-dependent reduction in crawling speed, and overlapped with ALS-associated transcriptomic changes. Together, these results support the conserved *SLC1A2*–*glt-6* axis as an experimentally validated candidate linking interpretable genome-wide prediction, SOD1-associated oxidative-stress biology and organism-level motor dysfunction.

Discussion

We present KANG2P, a genetically anchored molecular-state learning framework for interpretable genotype-to-phenotype mapping. Rather than requiring matched multi-omics measurements in every target cohort, KANG2P transfers regulatory information learned from external genotype–omics reference panels to reconstruct individualized, genetically regulated molecular states from genotype data alone. In the first stage, genotype-based molecular prediction models generate genetically predicted transcriptomic or proteomic profiles, providing biologically meaningful representations of the regulatory processes that connect DNA variation to phenotype. In the second stage, modality-specific autoencoders learn compact representations of genotype and predicted molecular profiles, which are subsequently integrated by a Kolmogorov–Arnold Network to model the nonlinear path from genetic variation through molecular regulation to phenotypic outcome. Across amyotrophic lateral sclerosis, Alzheimer's disease and agronomic traits in rice and maize, KANG2P consistently outperformed conventional machine-learning and deep-learning G2P models. Importantly, the value of genetically predicted molecular states extended beyond KANG2P itself: incorporating predicted expression also broadly improved established crop G2P models, indicating that these profiles capture regulatory information that is only indirectly represented by genotype. KANG2P further connected prediction with mechanism through gene-level attribution, prioritizing experimentally actionable candidates, including *SLC1A2* in ALS, whose conserved *C. elegans* orthologue *glt-6* was functionally supported in motor and oxidative-stress assays. Together, these findings establish KANG2P as a scalable and experimentally grounded framework that transforms reference-panel molecular information into transferable, interpretable representations for phenotype prediction across species.

The superior performance of KANG2P over existing genomic prediction methods stems primarily from the unique properties of Kolmogorov-Arnold Networks (KANs) in modeling the complex genetic regulatory relationships underlying phenotypic variation, coupled with its rational two-stage integration strategy. Unlike linear models that assume additive genetic effects and fail to capture nonlinear SNP-gene and gene-phenotype interactions, KANs are theoretically grounded in the Kolmogorov-Arnold representation theorem, which enables them to decompose and approximate high-dimensional nonlinear functions through the superposition of univariate functions. This intrinsic capability allows KANs to precisely model high-order epistatic interactions that are ubiquitous in crop genetic regulation but remain unaddressed by traditional linear frameworks. Compared to deep learning methods (DeepGS, Cropformer), KANs offer a more parsimonious solution to nonlinear modeling. While CNNs and transformers rely on large numbers of parameters and require massive training data to avoid overfitting—limiting their applicability in typical crop breeding scenarios with 300–1,000 samples, KANs achieve comparable or superior nonlinear fitting with fewer parameters. This efficiency arises from their structure: instead of using dense layers or self-attention mechanisms, KANs leverage interpretable univariate piecewise functions to model feature contributions, making them more robust to small-sample constraints common in crop research.

A central advantage of KANG2P is that it decouples direct molecular profiling in target cohorts from the use of molecular information for phenotype prediction. Genotype–omics relationships are learned in external reference panels and subsequently transferred to genotyped target cohorts, where they are

used to reconstruct individualized, genetically regulated molecular states without requiring matched transcriptomic or proteomic measurements. Building on the genotype-based expression-prediction principle introduced by PrediXcan [20], KANG2P extends this concept into a multi-layer G2P framework in which reference-panel-derived molecular states serve as intermediates between genetic variation and phenotype. The resulting gain was not confined to KANG2P, as predicted expression also strengthened established G2P models, indicating that these molecular representations capture regulatory information that is incompletely expressed in genotype-only feature spaces. This cross-sample strategy therefore offers a practical route to multi-omics-informed prediction without requiring matched molecular profiling in every study participant, an advantage that is particularly relevant for large cohorts and inaccessible tissues. At the same time, because the predicted profiles capture predominantly genetically regulated variation, they complement rather than substitute for molecular measurements that reflect environmental, developmental or disease-dependent states.

Beyond the transcriptomic and proteomic layers evaluated here, the same cross-sample integration strategy could be extended to other molecular intermediates, such as epigenomic states, chromatin accessibility, DNA methylation or metabolomic profiles. As increasingly large reference multi-omics resources become available, predictors trained in these panels could be used to impute additional molecular layers in genotyped target cohorts, further expanding the biological resolution of scalable G2P mapping. This scalable framework enables exploration of regulatory hierarchies across biological scales without requiring complete multi-omics profiling of all individuals, representing a practical advancement in multi-omics-driven phenotypic prediction.

KANG2P advances our understanding of the genotype-phenotype map by considering the contribution of gene expression regulation. Importantly, KANG2P linked prediction to mechanism: model-attribution analysis identified *SLC1A2* as a ALS candidate gene, and experimental interrogation of its conserved *C. elegans* orthologue, *glt-6*, provided functional evidence connecting this axis to motor impairment and oxidative-stress vulnerability. This capability to pinpoint functionally relevant genes represents a key advancement over "black-box" predictive models, as it bridges the gap between accurate prediction and biological insight. This finding supports the hypothesis that transcriptional regulation plays a pivotal role in shaping phenotypic diversity and underscores the value of incorporating intermediate molecular phenotypes in predictive models. For translational research, KANG2P's interpretability-driven gene prioritization offers a practical tool for identifying candidate genes and regulatory regions associated with complex traits, potentially accelerating the development of personalized medicine strategies by enabling targeted functional validation and therapeutic discovery.

Despite its strengths, this study highlights certain limitations. The first limitation does not relate to KANG2P itself, but to the still insufficient availability of data resources—of course, this limitation gets alleviated with the foreseeable accumulation of relevant data in the near future. Namely, there is still a scarcity of crop population datasets that simultaneously encompass both gene expression profiles and genotypic data, as well as a relatively narrow range of tissue types represented in available plant multi-omics datasets. In contrast to the rich and diverse data resources available for human studies, this paucity of crop data does not allow to explore the full range of robust gene expression prediction models that one can establish from scratch. Consequently, the accuracy of predicted gene expression profiles in crop may be compromised, which could in turn affect the reliability of phenotype predictions in general, which is an integrative part of KANG2P. The second limitation refers to the number of polymorphic sites involved in G2P cohorts, which emerges as a critical factor. For datasets with a small number of polymorphic sites, the ability to predict precise gene expression profiles is rather limited. The paucity of polymorphic sites reduces the power to capture the intricate regulatory relationships that govern gene expression, leading to increased uncertainty in the predicted expression levels. Third, while we have demonstrated that predicted gene expression contributes to the explanation of phenotypic variance, a more rigorous and quantitative assessment of the specific impact of the accuracy of predicted expression profiles on subsequent performance in phenotype prediction is required. A systematic evaluation of how errors in the prediction of expression levels propagate to accuracy in phenotype prediction would provide a clearer understanding of the framework's dependencies and limitations.

Looking forward, KANG2P lays the groundwork for several promising directions. Expanding the

framework to incorporate epigenetic modifications and environmental factors could further improve prediction accuracy and capture gene-environment interactions. Furthermore, integrating single-cell RNA sequencing data into KANG2P may enable cell-type-specific phenotype predictions, offering new avenues for studying tissue-specific traits. Finally, advances in explainable AI techniques could enhance the interpretability of KANG2P, allowing researchers to identify key regulatory pathways and genetic variants driving phenotypic variation. As population genomic datasets continue to grow in size and complexity, KANG2P represents a significant step toward more accurate and biologically informed phenotype prediction.

Methods

We provide comprehensive methodological details and describe the corresponding datasets for each step outlined in Figure 1.

Datasets

We analyzed data from two neurodegenerative diseases (ALS and AD) and two plant species (*Zea mays* and *Oryza sativa*) representing various population sizes and different reproductive systems. The published datasets used in this study are available from websites or literature sources: (1) the ALS dataset, Project Mine, dbGaP Study Accession: phs003146.v1.p1 [28, 29, 30]; (2) the AD dataset, dbGaP Study Accession:phs000168.v2.p2; [31]); (3) the rice 18K dataset [32]; (4) the rice1495 hybrid line dataset [43], available at https://iagr.genomics.cn/CropGS/; and (5) the maize CUBIC dataset [44].

ALS dataset

The ALS cohort consisted of 10,405 participants, including 7,213 healthy controls and 3,192 individuals with ALS; 5,208 participants were female and 5,197 were male. All participants were genotyped using the Illumina 2.5M single-nucleotide polymorphism (SNP) array. Genotype data were processed following the protocol described by [3]. Model evaluation was performed using five-fold outer cross-validation, with a three-fold inner cross-validation loop used for model selection and hyperparameter tuning (Supplementary Fig. 14). Cross-validation splits were generated before phenotype-informed feature selection, and all preprocessing steps were fitted within the corresponding training partition.

Within each training split, PLINK [50] was used to perform GWAS-based SNP filtering, retaining SNPs with ($P < 0.05$). SNPs were annotated to genes using ANNOVAR (version 20191024) [51] based on the hg19 human reference genome. Gene-level principal component analysis was fitted using the training samples, and the resulting transformations were applied to the corresponding validation or test samples. Feature scaling, resampling, hyperparameter tuning and decision-threshold selection were conducted within the inner cross-validation loop. After model selection, the final preprocessing and modelling pipeline was refitted on the full outer training split and evaluated on the held-out outer test fold.

After annotation, the genotype data comprised 505,333 SNPs mapped to 18,265 genes. Fold-specific gene-level PCA yielded a median of 74,634 Gene-PCA dimensions across the five outer folds for downstream prediction.

AD dataset

The Alzheimer's disease (AD) dataset comprised 4,686 individuals in a binary case–control setting, including 2,442 AD cases and 2,244 controls. For each individual, we assembled three complementary molecular feature modalities: genotype-derived gene-level principal-component features (G; 11,548 features), genetically predicted transcriptomic features (PE; 635 features), and standardized predicted protein-expression features (PP; 1,473 features). The combined G+PE representation was constructed by concatenating the genotype-derived and predicted-expression matrices after confirming one-to-one

sample correspondence, whereas PP features were aligned to the same sample order using individual identifiers.

Rice18K dataset

The genotypic matrix of Rice18K is downloaded from https://figshare.com/s/12978737918eecb74903. Given the distinct linkage-disequilibrium patterns in crop genomes relative to humans, we did not perform Gene-PCA for genotypic data dimensionality reduction. Instead, we employed a dimensionality reduction approach consistent with that used by [14] to select core SNP set from the initial 2,929,530 SNPs spanning 0.78% of the Nipponbare reference genome (MSU v7.0). The core SNP set of Rice18K dataset was screened using PLINK [50]. SNPs were removed by linkage disequilibrium pruning, a total of 56,992 SNPs were retained through the application of a 1 kb window size, a 100-SNP window step, and an r^2 threshold of 0.1. The 18,421 Rice18K samples were evaluated using nested cross-validation. Samples were randomly divided into five outer folds; in each iteration, one fold was held out as the test set and the remaining folds were used for training and inner-loop model selection. Five agronomic traits in the Rice18K dataset, including culm length, grain length , grain width, plant height, and grain yield was analysed. The phenotypic values of all traits were normalized to the 0-1 range.

Rice1495 dataset

The Rice1495 dataset was downloaded from CropGS database (https://iagr.genomics.cn/CropGS/). The core SNP set was screened using PLINK [50]. SNPs were removed by linkage disequilibrium pruning, using a window size of 1 kb, a window step of 100 SNPs, and an r^2 threshold of 0.1, yielding 20,556 SNPs. The 1,495 rice samples were evaluated using a nested cross-validation framework with five outer folds and three inner folds. In each outer iteration, one-fifth of the samples was held out for testing, while the remaining samples were used for training and three-fold inner cross-validation for model selection and hyperparameter tuning. Four agronomic traits in this dataset, including grain width, heading date, seed setting rate, and grain yield was analysed. The phenotypic values of all traits were normalized to the 0-1 range. The Rice277 reference panel, which includes both genotype and matched gene expression profiles, was downloaded from the NetGP repository (https://figshare.com/articles/dataset/NetGP/26495956/4) and used to train additive models of gene expression levels based on elastic net. Trained models was used to predict transcriptome for the Rice1495 population.

Maize CUBIC dataset

The maize CUBIC dataset consisted of 1,404 inbred offspring derived from 24 founds [44],among which 394 samples contain both RNA-seq data and genotype profiling data. Four phenotypic traits, kernel number per ear, kernel weight per ear, plant height, and date to anther, were measured in 1,404 samples. SNPs were retained only if they met the criteria of minor allele frequency (MAF) of at least 0.05. Subsequently, imputation of the filtered genotype data was conducted using Beagle [52], resulting in a final dataset containing 10,098,584 SNPs—accounting for 0.488% of the total length of maize B73 reference genome (v3.25). The gene expression matrix of the 394 samples was derived from the authors of DEM [18]. We excluded 394 samples containing RNA-seq, because these samples were used to build gene expression level prediction models, ultimately retaining 1,010 samples. The 1,010 maize samples were analyzed using 5-fold outer and 3-fold inner nested cross-validation. For each outer fold, one subset was held out as the test set, and the remaining samples were used for training and inner-loop model selection. We directly downloaded the genotype matrix from the R data within the G2P R package https://github.com/G2P-env/G2P/tree/main/data [53].

Predicting gene expression levels into genotyped samples

To predict gene expression levels from genetic variants into genotyped samples, we utilized MetaXcan [20], which leverages transcriptome prediction models provided by the PredictDB database (`https://predictdb.org/`). Specifically, genotype data was input into MetaXcan to obtain predicted gene

expression values across 13 brain tissues, reflecting tissues relevant to neurodegenerative pathologies. For each brain tissue, differentially expressed genes were identified in a fold-specific manner. In each of the five outer folds of nested cross-validation, Wilcoxon rank-sum tests were performed using only the predicted expression profiles from the corresponding training samples (Supplementary Table 1). Next, feature selection was performed within the nested cross-validation framework. After tissue-specific gene screening, the minimum-redundancy maximum-relevance (mRMR) algorithm was applied using only the training samples in each fold to define a compact and non-redundant gene set, which was subsequently used for prediction in the corresponding validation or held-out test samples.

Unlike humans, plants lack pre-trained models for predicting gene expression levels. A reference population including genotype data and gene expression levels was needed to train gene expression level prediction models based on elastic net (Fig. 1c).

$$\hat{\beta} \equiv \arg\min_{\beta} \left(\|y - X\beta\|^2 + \lambda_2 \|\beta\|^2 + \lambda_1 \|\beta\|_1 \right). \tag{1}$$

where y is the vector of gene expresssion level, X is SNPs in the vicinity of each gene (within 50 kb of the gene start or end), β is the effect size of SNPs for gene. Specifically, λ_2 is the regularization parameter governing the L_2 ridge penalty, which shrinks the magnitude of regression coefficients toward zero and stabilizes estimates by trading variance for bias. λ_1 controls the L_1 lasso penalty, enforcing sparsity by driving the values of small-magnitude coefficients exactly to zero and thus performing automatic feature selection.

Using genetic variation data from the GWAS and pretrained gene expression level model, we impute expression levels for the whole transcriptome in the genomic prediction population. The top 10% of models, ranked by prediction accuracy, were used to predict expression profiles, which were then integrated with genotype data using as input features in KANG2P workflow.

Predicting protein abundance into genotyped samples

Genetically predicted protein abundance was inferred from genotype data using the pre-trained protein prediction weights generated by [27], with the protein weights available at https://doi.org/10.7303/syn23627957.

The architecture of KANG2P

Modality-specific autoencoders

To learn robust and disentangled latent representations from high-dimensional omics data, we employ a Modality-specific autoencoder architecture. Each omics modality is independently encoded into a low-dimensional latent space and subsequently reconstructed to preserve biological information.

Encoder For each input modality $m \in \{1, \ldots, M\}$, a modality-specific encoder $\mathcal{E}_m$ maps $\mathbf{x}^{(m)} \in \mathbb{R}^{D_m}$ to a latent representation $\mathbf{z}^{(m)} \in \mathbb{R}^{d_m}$. Depending on the analysis, $M = 2$ for genotype combined with either genetically predicted transcriptomic or proteomic profiles, and $M = 3$ when all three modalities are jointly modelled. Each encoder comprises a fully connected layer with LeakyReLU activation, layer normalization and dropout, a residual block for nonlinear feature refinement, and a final linear projection to the modality-specific latent space. The encoders share the same architectural template but have independent parameters, allowing heterogeneous modalities to be represented separately before gated fusion.

Decoder Each modality-specific decoder reconstructed its corresponding input from the latent representation. The decoders shared the same architectural template but had independent parameters, comprising a feed-forward expansion with LeakyReLU activation and layer normalization, a self-attention module, a residual refinement block and a final linear projection to the original input dimension. Modality-specific reconstruction objectives were used to regularize the latent representations before gated multimodal fusion.

Reconstruction Objective The modality-specific autoencoders were jointly trained by minimizing the mean reconstruction error across the M input modalities:

$$\mathcal{L}_{\text{recon}} = \frac{1}{M} \sum_{m=1}^{M} \frac{1}{D_m} \left\| \mathbf{x}^{(m)} - \hat{\mathbf{x}}^{(m)} \right\|_2^2. \tag{2}$$

This objective preserves information from each modality in its corresponding latent representation while preventing differences in feature dimensionality from biasing the optimization towards a single input branch.

Gated multimodal fusion

Rather than concatenating the modality-specific embeddings, KANG2P used a learnable gated-fusion module to adaptively balance their contributions. Each latent representation was first projected into a shared d_F-dimensional space:

$$\mathbf{h}^{(m)} = \mathbf{W}_{\text{p}}^{(m)} \mathbf{z}^{(m)} + \mathbf{b}_{\text{p}}^{(m)}, \qquad m = 1, \ldots, M. \tag{3}$$

Feature-wise gating weights were then inferred jointly from all projected representations:

$$\left[\boldsymbol{\alpha}^{(1)}, \ldots, \boldsymbol{\alpha}^{(M)} \right] = \text{softmax}_m \left(\mathbf{W}_{\text{g}} \left[\mathbf{h}^{(1)} \| \cdots \| \mathbf{h}^{(M)} \right] + \mathbf{b}_{\text{g}} \right), \tag{4}$$

where $\|$ denotes temporary concatenation for gate estimation and the softmax is applied across modalities. The fused representation was computed as

$$\mathbf{z}_F = \sum_{m=1}^{M} \boldsymbol{\alpha}^{(m)} \odot \mathbf{h}^{(m)}, \tag{5}$$

where $\odot$ denotes element-wise multiplication. This allows the model to emphasize informative modality-specific signals while attenuating redundant or noisy features. The resulting representation $\mathbf{z}_F$ was passed to the FourierKAN phenotype-prediction module.

Prediction head: FourierKAN

The gated multimodal representation $\mathbf{z}_F$ was passed to a two-layer FourierKAN prediction head. In each FourierKAN layer, the learnable edge functions were parameterized using truncated Fourier expansions, enabling adaptive nonlinear transformation of the fused latent representation. The first layer projected $\mathbf{z}_F$ into a hidden representation, followed by LeakyReLU activation, layer normalization and dropout:

$$\mathbf{h} = \text{Dropout}\left[\text{LayerNorm}\left(\text{LeakyReLU}\left(\text{FourierKAN}_1(\mathbf{z}_F) \right) \right) \right]. \tag{6}$$

A second FourierKAN layer mapped $\mathbf{h}$ to the task-specific output:

$$o = \text{FourierKAN}_2(\mathbf{h}). \tag{7}$$

For binary disease prediction, o was converted to a disease probability using a sigmoid function, whereas for agronomic-trait prediction, o was used directly as the predicted continuous phenotype. This unified prediction head allowed the same architecture to support both classification and regression tasks while modelling nonlinear relationships within the integrated genotype and molecular-state representation.

Model training and evaluation

We trained the KANG2P model in an end-to-end manner using a objective combining classification and reconstruction losses. The model was implemented in PyTorch, with KAN layers replacing conventional linear transformations in the prediction head to enable interpretable non-linear decision boundaries.

Model performance was assessed using a nested cross-validation framework (Supplementary Fig. 13) to ensure robust evaluation and prevent information leakage. Specifically, we employed a 5-fold outer cross-validation for performance estimation, with an inner 3-fold cross-validation used for hyperparameter tuning.

KANG2P was trained end-to-end by jointly optimizing phenotype prediction and modality-specific reconstruction:

$$\mathcal{L} = \lambda_{\text{pred}}\mathcal{L}_{\text{pred}} + \lambda_{\text{rec}}\frac{1}{M}\sum_{m=1}^{M}\mathcal{L}_{\text{rec}}^{(m)}, \tag{8}$$

where M denotes the number of input modalities and $\mathcal{L}_{\text{rec}}^{(m)}$ is the mean squared reconstruction error for modality m. The prediction loss $\mathcal{L}_{\text{pred}}$ was defined as binary cross-entropy for disease classification and mean squared error for quantitative agronomic-trait prediction. The reconstruction term regularizes each modality-specific latent representation, whereas the prediction term encourages the fused representation to retain phenotype-relevant information. Unless otherwise specified, λ_{pred} and λ_{rec} were both set to 1.0.

Within each outer fold, model parameters were optimized using the Adam optimizer with weight decay regularization, which improves generalization performance. Training was performed for 30 epochs. The learning rate followed an exponential decay schedule with decay factor $\gamma = 0.8$ applied at each epoch.

Hyperparameter optimization was conducted using Optuna (https://github.com/optuna/optuna) within the inner cross-validation loop.

Gradient-based interpretation of the KANG2P model

To interpret the trained G+PE KANG2P model at the gene level, we quantified the contribution of genotype-derived features to the ALS-class output using gradient-based attribution. Feature-level importance was subsequently aggregated to obtain gene-level scores. The full procedure is described below.

Gene expression and genotype-derived principal component (PC) features were processed through the G and PE branches of the DualBranch architecture, respectively. During interpretation, both branches were retained to ensure that gradient attributions for the G branch were computed in the full multimodal context of the predictive model.

Input features for both branches were standardized using the mean and scale parameters stored in the trained model checkpoint, ensuring consistency between interpretation and model training.

For each individual, a forward pass was performed through the trained model, and gradients of the ALS-class logit with respect to standardized genotype-branch features were computed via backpropagation.

Feature-level importance was defined as the mean absolute gradient across individuals:

$$a_j = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\partial f_1(x_i^G, x_i^{PE})}{\partial x_{ij}^G}\right| \tag{9}$$

where $f_1(\cdot)$ denotes the ALS-class logit and x_{ij}^G represents the j-th genotype-branch feature of individual i.

Because each gene may be represented by multiple genotype-derived PCA components, feature-level attributions were aggregated to gene-level scores using the root mean square (RMS):

$$A_g = \sqrt{\frac{1}{|J_g|}\sum_{j\in J_g} a_j^2} \tag{10}$$

where J_g denotes the set of genotype-branch features mapped to gene g. This procedure yielded a gradient-based ranking score K_g across 18,265 genes.

Drug–target affinity prediction and docking-assisted computational assessment

To explore whether KANG2P-prioritized ALS candidate genes could nominate potential drug-associated targets, we performed an in silico drug–target prioritization analysis. The top 500 KANG2P-ranked candidate genes were mapped to UniProt protein entries, and genes with valid amino acid sequences were retained for downstream analysis. In total, 409 mappable proteins were included. We evaluated two approved small-molecule ALS drugs, riluzole (DrugBank ID: DB00740) and edaravone (DrugBank ID: DB12243). Tofersen was not included because it is an antisense oligonucleotide and is therefore outside the small-molecule SMILES-based input setting of DeepDTAGen [46].

Drug–target affinity prediction was performed using the BindingDB-pretrained DeepDTAGen model [46]. For each drug–protein pair, the model input consisted of the drug SMILES string and the candidate protein amino acid sequence. DeepDTAGen encodes molecular structures and protein sequences through graph-based representation modules and outputs a predicted affinity score for each pair. Predicted scores were used for within-drug ranking of candidate proteins and were not interpreted as experimentally measured binding affinities. Candidate proteins were ranked separately for riluzole and edaravone, and the top-ranked drug–target pairs were selected for docking-assisted computational assessment.

Molecular docking was performed using AutoDock Vina 1.1.2 [54] to provide an orthogonal structural assessment of high-scoring DeepDTAGen pairs. Ligand structures were obtained from PubChem in SDF format, and three-dimensional conformers were generated and energy-minimized using RDKit when required. Protein receptor structures were obtained from experimentally resolved PDB structures when available; otherwise, AlphaFold DB v6 monomer models were used. Receptor structures were prepared by removing water molecules and non-relevant ligands in PyMOL. Ligand and receptor files were converted to PDBQT format using AutoDockTools 1.5.6. Because experimentally validated binding pockets were not available for all candidate proteins, docking was performed in a blind-docking setting, with the search box defined to cover the receptor structure and capped at a maximum box size of (60,Å). AutoDock Vina was run with an exhaustiveness of 8, nine output modes and a random seed of 20260706. For each drug–target pair, the best Vina binding score was retained.

DeepDTAGen predicted affinity scores and Vina docking scores were jointly used for visualization and computational prioritization of candidate drug–target interactions. Docking poses were visualized using PyMOL 2.6.0.

Weighted integrative evidence score

To integrate KANG2P model-derived evidence with external ALS gene annotations, we constructed a weighted composite scoring framework. The KANG2P evidence component (K_g) was defined as the gene-level rank-derived score obtained from the gradient-based DualBranch attribution analysis after adjustment with the curated ALS gene prior.

Four independent external evidence sources were incorporated. (i) DiseaseCapsule [3] membership (D_g): genes present in a curated set of 922 ALS-associated genes were assigned a value of 1, and 0 otherwise. (ii) RefMap [47] ALS evidence (R_g): genes annotated as both GWAS-associated and curated "associated" in the RefMap sheet were assigned a normalized score in [0,1] based on their curated ordering, and 0 otherwise. (iii) ALSoD https://alsod.ac.uk/database evidence strength (A_g): categorical evidence levels were mapped to ordinal scores (Definitive, Strong evidence, Clinical modifier, Moderate evidence, Tenuous, and Unassigned corresponding to 1.00, 0.85, 0.75, 0.60, 0.35, and 0.20, respectively). (IV) transcriptome evidence T_g from GEO:GSE122649 [48].

$$T_g = 0.50 \times \min\left(\frac{-\log_{10}(\mathrm{FDR})}{5},\, 1\right) + 0.50 \times \min\left(\frac{|\log_2 \mathrm{FC}|}{1.5},\, 1\right) \tag{11}$$

Here, FDR and FC were derived from differential expression analysis of GSE122649 [48], comparing ALS or sporadic ALS with frontotemporal dementia (sALS-FTD) cases ($n = 13$) with non-neurological disease controls ($n = 6$). FDR denotes the multiple-testing-adjusted significance level for each gene, whereas FC denotes the expression fold change in ALS/sALS-FTD cases relative to controls.

The final Weighted integrative evidence score was defined as:

$$S_g = 0.2 \times (K_g + D_g + R_g + A_g + T_g) \tag{12}$$

where K_g, D_g, R_g, and A_g and T_g denote KANG2P evidence, DiseaseCaps membership, RefMap/curated ALS evidence, ALSoD evidence strength and transcriptome evidence, respectively. All genes were ranked in descending order of S_g to obtain the final integrative prioritization.

Evaluation metrics

Model performance was evaluated on independent test datasets. For binary classification tasks, we computed accuracy, precision, recall, and the F1-score. These metrics are defined in terms of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{13}$$

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{14}$$

$$\text{Recall} = \frac{TP}{TP + FN}, \tag{15}$$

$$F1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{16}$$

Accuracy quantifies the overall correctness of predictions, precision measures the proportion of correctly identified positive samples among all predicted positives, recall measures the proportion of actual positive samples correctly identified, and the F1-score provides a balanced harmonic mean of precision and recall, which is particularly informative under class imbalance.

For quantitative traits, model performance was evaluated using the Pearson correlation coefficient (PCC), coefficient of determination (R^2), and mean squared error (MSE). The PCC is defined as:

$$\text{PCC} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \overline{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{n}(\hat{y}_i - \overline{\hat{y}})^2}}, \tag{17}$$

where y_i and $\hat{y}_i$ denote the observed and predicted phenotype values, respectively, $\bar{y}$ and $\overline{\hat{y}}$ are their respective means, and n is the number of samples.

The coefficient of determination (R^2) is defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{18}$$

and the mean squared error (MSE) is given by:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2. \tag{19}$$

Methods used for comparison

Regression task

For regression task, the predictive performance of KANG2P was compared with the performance of 13 popular methods: Ridge Regression Best Linear Unbiased Predictor (RR-BLUP), BayesC, Bayesian LASSO (BL), LASSO, Bayesian Ridge Regression (BRR), Reproducing Kernel Hilbert Space (RKHS),

Sparse Partial Least Squares (SPLS), Random Forest (RF), Support vector machine (SVM), DeepGS, DNNGP, Cropfomer and NetGP.

RR-BLUP was employed as a baseline genomic prediction method [5]. In RR-BLUP, the phenotype vector $\mathbf{y}$ is modeled as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}, \tag{20}$$

where $\mathbf{X}$ is the design matrix for fixed effects with coefficients $\boldsymbol{\beta}$, $\mathbf{Z}$ is the marker genotype matrix, $\mathbf{u}$ is the vector of marker effects, and $\boldsymbol{\epsilon}$ represents the residual error. The marker effects are assumed to follow a normal distribution $\mathbf{u} \sim N(0, \sigma_u^2\mathbf{I})$, imposing a ridge penalty that shrinks all marker effects toward zero. This formulation yields best linear unbiased predictions of genetic values while controlling overfitting by uniformly shrinking marker effects.

BayesC is a Bayesian regression method widely used in genomic prediction, which extends the BayesB framework by allowing a proportion of markers to have zero effect [7]. Specifically, the phenotype vector is modeled as

$$y_i = \mu + \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i, \tag{21}$$

where μ is the overall mean, x_{ij} is the genotype covariate of individual i at marker j, β_j is the marker effect, and $\epsilon_i \sim N(0, \sigma_e^2)$. In BayesC, each marker effect β_j is assumed to follow a mixture prior distribution:

$$\beta_j \sim (1-\pi)\, N(0, \sigma_\beta^2) + \pi\, \delta_0, \tag{22}$$

where π is the prior probability that a marker has no effect, σ_β^2 is the variance of nonzero marker effects, and δ_0 denotes a point mass at zero. This prior induces variable selection by shrinking a subset of marker effects exactly to zero, while estimating the remaining effects from a normal distribution. Gibbs sampling is typically used to infer the posterior distribution of marker effects and variance components.

BL was employed for genomic prediction [55]. In this approach, the phenotype vector is modeled as $y = X\beta + \epsilon$, where X is the marker matrix, β is the vector of marker effects, and $\epsilon \sim N(0, \sigma_e^2 I)$. A double-exponential (Laplace) prior is assigned to each marker effect β_j, which induces shrinkage of coefficients towards zero and allows variable selection. Posterior inference is conducted via Markov Chain Monte Carlo (MCMC), providing both estimates of marker effects and measures of their uncertainty. In Bayesian Ridge Regression (BRR), all marker effects are assumed to follow a common normal prior distribution $\beta_j \sim N(0, \sigma_\beta^2)$, which imposes homogeneous shrinkage across markers.

RKHS regression is a semi-parametric method for genomic prediction that captures both linear and non-linear genetic effects through kernel functions [56]. The phenotype of the i-th individual is modeled as

$$y_i = \mu + f(\mathbf{x}_i) + \epsilon_i, \tag{23}$$

where μ is the overall mean, $\mathbf{x}_i$ is the genotype vector, $f(\cdot)$ is an unknown function belonging to a reproducing kernel Hilbert space, and $\epsilon_i \sim N(0, \sigma_e^2)$. By the representer theorem, $f(\mathbf{x}_i)$ can be expressed in terms of kernel evaluations:

$$f(\mathbf{x}_i) = \sum_{j=1}^{n} \alpha_j K(\mathbf{x}_i, \mathbf{x}_j), \tag{24}$$

where $K(\cdot, \cdot)$ is a positive definite kernel function (e.g., Gaussian kernel) measuring genetic similarity between individuals, and α_j are coefficients estimated from the data.

SPLS was employed to perform dimension reduction and variable selection in high-dimensional genomic data [57]. SPLS extends the conventional Partial Least Squares (PLS) regression by introducing sparsity constraints on the weight vectors, thereby selecting a subset of informative predictors while projecting both predictors and responses into a lower-dimensional latent space. Formally, SPLS seeks latent components that maximize the covariance between linear combinations of predictors and responses, subject to an L_1-norm penalty that enforces sparsity on the loading vectors. This approach enables simultaneous feature extraction and variable selection, making it particularly suitable for genomic prediction where the number of markers greatly exceeds the number of samples.

BayesC, BRR, BL, RR-BLUP, RKHS, SPLS, RFR, SVR, and LASSO was implemented in G2P software `https://github.com/G2P-env/G2P`.

DeepGS is a deep learning model implemented with convolutional neural network (CNN) layers (Ma et al., 2018) for genomic prediction [12]. Its architecture comprises one input layer, a convolutional layer, a sampling layer, three dropout layers, two fully connected layers, and a final output layer. DeepGS was implemented using the official code repository https://github.com/cma2015/DeepGS.

Deep Neural Network Genomic Prediction (DNNGP) is a deep learning–based method designed to enhance the accuracy of genomic predictions by effectively integrating multi-omics data in plants. The architecture of DNNGP comprises multiple layers, including convolutional layers, one batch normalization layer, two dropout layers, one flattening layer, one dense layer, and one output [13]. DNNGP was implemented using the official code repository `https://github.com/AIBreeding/DNNGP`.

Cropformer is an advanced deep learning framework designed to enhance genomic selection (GS) by improving the prediction of complex phenotypic traits in crops. It integrates convolutional neural networks (CNNs) with multiple self-attention mechanisms to effectively capture both local and global features from genomic data [14]. Cropformer was implemented using the official code repository `https://github.com/jiekesen/Cropformer`.

Classification task

For classification task, the predictive performance of KANG2P was compared with the performance of 5 popular methods: DiseaseCapsule [3], Logistic regression [58] , Random Forest [59], Support vector machine [60], and AdaBoost [61]. Logistic regression, Random Forest, Support Vector Machine, and AdaBoost were implemented using the scikit-learn library `https://scikit-learn.org/stable/`.

DiseaseCapsule [3] is a deep learning framework for genomic prediction that employs capsule networks to capture hierarchical and non-linear relationships between genetic variants and disease phenotypes. Unlike conventional linear models, which primarily assume additive marker effects, DiseaseCapsule leverages the dynamic routing mechanism of capsule networks to represent higher-order feature interactions, thereby improving the modeling of complex genetic architectures. DiseaseCapsule was implemented using the official python code `https://github.com/HaploKit/DiseaseCapsule`.

Enrichment analysis

Gene set enrichment analysis (GSEA) was performed using the fgseaMultilevel algorithm as implemented through clusterProfiler [62] v.4.14.0. A total of 17,377 protein-coding genes were ranked according to their dual-branch gradient scores. Gene symbols were converted to Entrez Gene identifiers using org.Hs.eg.db v.3.20.0. To increase the dynamic range of the ranking metric while preserving rank order, priority scores were cubed before enrichment analysis.

Gene sets were obtained from the Molecular Signatures Database (MSigDB v.2026.1), including the Hallmark collection (50 gene sets), C5 Gene Ontology Biological Process collection (7,538 gene sets) and C2 CP:KEGG Legacy collection (186 gene sets). Gene sets containing fewer than 10 or more than 500 genes after mapping were excluded. Normalized enrichment scores (NESs) and nominal two-sided P values were estimated using fgseaMultilevel, and multiple testing was controlled using the Benjamini–Hochberg false discovery rate (FDR). Pathways with FDR < 0.25 were considered enriched for exploratory interpretation, whereas pathways with FDR < 0.05 were considered statistically significant.

Over-representation analysis (ORA) was performed on the top 500 ranked genes against the same MSigDB collections. All 17,377 scored protein-coding genes were used as the background universe. Enrichment P values were adjusted using the Benjamini–Hochberg procedure, and gene sets with FDR < 0.05 were considered significant.

Locomotion analysis

Day-1 young adult worms cultured by standard methods were transferred onto an area in NGM plates (diameter 6 cm) with a thin layer of OP50 and allowed them to move freely for 3 min. Worm locomotion

was recorded under a Zeiss Discovery V8 stereomicroscope (Carl Zeiss MicroImaging GmbH, Göttingen, Germany). The image sequences were captured with an Andor iXonEM+ DV885K EMCCD camera (Andor Technology plc., Springvale Business Park, Belfast, UK) at 10 frame per second. The worm locomotion speed was analyzed by use of Worm Lab software.

Paraquat Concentration Optimization

To determine an appropriate paraquat concentration for oxidative stress induction, worms were exposed to 2.5, 5, or 10 mM paraquat for 4 h. Paraquat-containing OP50 suspensions were seeded onto NGM plates and incubated overnight at 20°C to allow complete diffusion of paraquat into the agar medium. Day 1 adult worms were transferred to the treatment plates and maintained at 20°C during exposure. Locomotion behavior was subsequently recorded, and average locomotor speed was quantified using WormLab software. Untreated worms served as controls.

Paraquat-Induced Oxidative Stress Treatment

Oxidative stress was induced by paraquat treatment with minor modifications from the method described by [63]. Briefly, paraquat was mixed with concentrated Escherichia coli OP50 culture and 100 μL of the mixture was evenly spread onto 35-mm NGM plates to achieve a final paraquat concentration of 10 mM. The plates were incubated overnight at 20°C to allow complete diffusion of paraquat into the agar medium. Day 1 adult worms of the indicated strains were randomly selected to transfer onto paraquat-containing plates. Twenty healthy and actively moving worms were placed on each plate and incubated at 20°C in the dark for 4 h. Following paraquat exposure, locomotor activity was assessed by measuring crawling speed as described above Locomotion analysis section. Worms maintained on OP50 plates without paraquat treatment were used as controls and analyzed in parallel.

Paraquat-Induced Oxidative Stress Survival Assay

Day 1 adult worms were transferred into 96-well plates (10 worms per well) containing 50 μL M9 buffer supplemented with 200 mM paraquat. For each genotype, 3–5 replicate wells were analyzed. Animals were maintained at 20°C and scored for survival every 30 min. Worms that failed to respond to gentle touch with a platinum wire were considered dead. Survival percentages were calculated at each time point.

Maintenance of *C. elegans* strains

All *C. elegans* strains were cultured on nematode growth media (NGM) plates at 20°C using *E. coli* bacteria OP50 as food by standard procedures [64]. The strains were obtained from the CGC http://www.cbs.umn.edu/CGC/ or the National Bio-Resources Project (NBRP, http://www.shigen.nig.ac.jp /c.elegans/indices.jsp). The strains used in this study are listed in the Supplementary Table 3.

Data availability

All data used in this study are publicly available. The ALS data used in this study is from the Dutch cohort of Project MinE [28, 29, 30]. The ALS data are publicly available through the Database of Genotypes and Phenotypes (dbGaP) under study accession number phs003146.v1.p1. Amyotrophic Lateral Sclerosis online Database (ALSoD) is available at `https://alsod.ac.uk/`. The data of AD was downloaded from dbGaP Study Accession: phs000168.v2.p2 [31]. The rice 18K dataset used in this study is from [32]. The genotypic matrix of Rice18K is downloaded from https://figshare.com/s/12978737918eecb74903. The phenotypic data for the rice 18K population were obtained from the supplementary materials of the [32] 's publication. The rice1495 hybrid line dataset is from [43]. The genotype and phenotype data was downloaded from `https://iagr.genomics.cn/CropGS/`. The maize CUBIC dataset is from [44].

Genotype and phenotype data of CUBIC was downloaded from the R data within the G2P R package https://github.com/G2P-env/G2P/tree/main/data [53].

Code availability

The source code of KANG2P is publicly available at `https://github.com/LuoGroup2023/KANG2P/`.

Acknowledgements

XL is supported by the National Natural Science Foundation of China (Grant No. 32400506), the Natural Science Foundation of Hunan Province (Grant No. 2024JJ4008), Fundamental Research Funds for the Central Universities (Grant No. 541109030062) and Hunan Science and Technology Innovation Plan 2025ZYJ003. YuanshengLiu is supported by the National Natural Science Foundation of China (Grant No. 62372159), and Yuelushan Laboratory Breeding Program (No. YLS-2026-ZY01001). TW was funded by the National Natural Science Foundation of China (32571198), and the Fundamental Research Funds for the Central Universities(531119200243).

Author contributions

Xiao Luo, Yuansheng Liu, and Taihong Wu conceived this study. Xiao Luo, Lei Ling and Yang Liu designed the method. Lei Ling implemented the code. Pingzhou Wang and Zhimin Zhang performed the *C. elegans* functional assays and contributed to experimental validation of candidate genes. Lei Ling, Pingzhou Wang, Alexander Schönhuth, Taihong Wu, Yuansheng Liu and Xiao Luo wrote and revised the manuscript. Lei Ling, Yang Liu, Hang Ning, Fengxin Li, and Cheng Yang conducted the data analysis. Liang Qiao, Si Qin and Alexander Schönhuth provided critical feedback and assisted with data interpretation. All authors read and approved the final version of the manuscript.

Competing interests

We, the authors, have a patent application (No. 2025113496060) related to this work, and we confirm that there are no patents held by immediate family members that may potentially conflict with the content of this paper. The authors declare that they have no any other competing interests.

References

[1] Gad Abraham, Aki S Havulinna, Oneil G Bhalala, Sean G Byars, Alysha M De Livera, Laxman Yetukuri, Emmi Tikkanen, Markus Perola, Heribert Schunkert, Eric J Sijbrands, et al. Genomic prediction of coronary heart disease. *European heart journal*, 37(43):3267–3278, 2016.

[2] Stephen-John Sammut, Mireia Crispin-Ortuzar, Suet-Feung Chin, Elena Provenzano, Helen A Bardwell, Wenxin Ma, Wei Cope, Ali Dariush, Sarah-Jane Dawson, Jean E Abraham, et al. Multi-omic machine learning predictor of breast cancer therapy response. *Nature*, 601(7894):623–629, 2022.

[3] Xiao Luo, Xiongbin Kang, and Alexander Schönhuth. Predicting the prevalence of complex genetic diseases from individual genotype profiles using capsule networks. *Nature Machine Intelligence*, 5(2):114–125, 2023.

[4] Paul M VanRaden. Efficient methods to compute genomic predictions. *Journal of dairy science*, 91(11):4414–4423, 2008.

[5] Jeffrey B Endelman. Ridge regression and other kernels for genomic selection with r package rrblup. *The plant genome*, 4 (3), 2011.

[6] Theo HE Meuwissen, Ben J Hayes, and ME1461589 Goddard. Prediction of total genetic value using genome-wide dense marker maps. *genetics*, 157(4):1819–1829, 2001.

[7] David Habier, Rohan L Fernando, Kadir Kizilkaya, and Dorian J Garrick. Extension of the bayesian alphabet for genomic selection. *BMC bioinformatics*, 12(1):186, 2011.

[8] Shang Gao, Tingxi Yu, Awais Rasheed, Jiankang Wang, Jose Crossa, Sarah Hearne, and Huihui Li. Fast-forwarding plant breeding with deep learning-based genomic prediction. *Journal of Integrative Plant Biology*, 2025.

[9] Christina B Azodi, Emily Bolger, Andrew McCarren, Mark Roantree, Gustavo de Los Campos, and Shin-Han Shiu. Benchmarking parametric and machine learning models for genomic prediction of complex traits. *G3: Genes, Genomes, Genetics*, 9(11):3691–3702, 2019.

[10] Osval A Montesinos-Lopez, Moises Chavira-Flores, Kismiantini, Leo Crespo-Herrera, Carolina Saint Piere, Huihui Li, Roberto Fritsche-Neto, Khalid Al-Nowibet, Abelardo Montesinos-López, and Jose Crossa. A review of multimodal deep learning methods for genomic-enabled prediction in plant breeding. *Genetics*, 228(4):iyae161, 2024.

[11] Osval Antonio Montesinos-López, Abelardo Montesinos-López, Paulino Pérez-Rodríguez, José Alberto Barrón-López, Johannes WR Martini, Silvia Berenice Fajardo-Flores, Laura S Gaytan-Lugo, Pedro C Santana-Mancilla, and José Crossa. A review of deep learning applications for genomic selection. *BMC genomics*, 22(1):19, 2021.

[12] Wenlong Ma, Zhixu Qiu, Jie Song, Qian Cheng, and Chuang Ma. A deep convolutional neural network approach for predicting phenotypes from genotypes. *Planta*, 248(5):1307–1318, 2018. doi: 10.1007/s00425-018-2976-9.

[13] Kelin Wang, Muhammad Ali Abid, Awais Rasheed, Jose Crossa, Sarah Hearne, and Huihui Li. DNNGP, a deep neural network-based method for genomic prediction using multi-omics data in plants. *Molecular Plant*, 16(1):279–293, 2023.

[14] Hao Wang, Shen Yan, Wenxi Wang, Yongming Chen, Jingpeng Hong, Qiang He, Xianmin Diao, Yunan Lin, Yanqing Chen, Yongsheng Cao, et al. Cropformer: An interpretable deep learning framework for crop genomic prediction. *Plant Communications*, 6(3), 2025.

[15] Christina B Azodi, Jiliang Tang, and Shin-Han Shiu. Opening the black box: interpretable machine learning for geneticists. *Trends in genetics*, 36(6):442–455, 2020.

[16] Gherman Novakovsky, Nick Dexter, Maxwell W Libbrecht, Wyeth W Wasserman, and Sara Mostafavi. Obtaining genetics insights from deep learning via explainable artificial intelligence. *Nature Reviews Genetics*, 24(2):125–137, 2023.

[17] Hakim Benkirane, Yoann Pradat, Stefan Michiels, and Paul-Henry Cournède. CustOmics: A versatile deep-learning based strategy for multi-omics integration. *PLOS Computational Biology*, 19(3):e1010921, 2023.

[18] Yanlin Ren, Chenhua Wu, He Zhou, Xiaona Hu, and Zhenyan Miao. Dual-extraction modeling: A multi-modal deep-learning architecture for phenotypic prediction and functional gene mining of complex traits. *Plant Communications*, 5(9), 2024.

[19] Peipei Wang, Melissa D Lehti-Shiu, Serena Lotreck, Kenia Segura Abá, Patrick J Krysan, and Shin-Han Shiu. Prediction of plant complex traits via integration of multi-omics data. *Nature Communications*, 15(1):6856, 2024.

[20] Eric R Gamazon, Heather E Wheeler, Kaanan P Shah, Sahar V Mozaffari, Keston Aquino-Michaels, Robert J Carroll, Anne E Eyler, Joshua C Denny, GTEx Consortium, Dan L Nicolae, et al. A gene-based association method for mapping traits using reference transcriptome data. *Nature genetics*, 47(9):1091–1098, 2015.

[21] Alexander Gusev, Arthur Ko, Huwenbo Shi, Gaurav Bhatia, Wonil Chung, Brenda WJH Penninx, Rick Jansen, Eco JC De Geus, Dorret I Boomsma, Fred A Wright, et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nature genetics*, 48(3):245–252, 2016.

[22] Ziming Liu, Yixuan Wang, Sachin Vaidya, Fabian Ruehle, James Halverson, Marin Soljačić, Thomas Y Hou, and Max Tegmark. Kan: Kolmogorov-arnold networks. *arXiv preprint arXiv:2404.19756*, 2024.

[23] John Lonsdale, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, Gary Walters, Fernando Garcia, Nancy Young, et al. The genotype-tissue expression (gtex) project. *Nature genetics*, 45(6):580–585, 2013.

[24] Shiron Drusinsky, Sean Whalen, and Katherine S Pollard. Deep-learning prediction of gene expression from personal genomes. *Genome Biology*, 2026.

[25] Viksar Dubey and Li Shen. Personalized gene expression prediction in the era of deep learning: a review. *Briefings in Bioinformatics*, 27(1), 2026.

[26] Hanchuan Peng, Fuhui Long, and Chris Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence*, 27(8):1226–1238, 2005.

[27] Aliza P Wingo, Yue Liu, Ekaterina S Gerasimov, Jake Gockley, Benjamin A Logsdon, Duc M Duong, Eric B Dammer, Chloe Robins, Thomas G Beach, Eric M Reiman, et al. Integrating human brain proteomes with genome-wide association data implicates new proteins in alzheimer's disease pathogenesis. *Nature genetics*, 53(2):143–146, 2021.

[28] Wouter Van Rheenen, Aleksey Shatunov, Annelot M Dekker, Russell L McLaughlin, Frank P Diekstra, Sara L Pulit, Rick AA Van Der Spek, Urmo Võsa, Simone De Jong, Matthew R Robinson, et al. Genome-wide association analyses identify new risk variants and the genetic architecture of amyotrophic lateral sclerosis. *Nature genetics*, 48(9):1043–1048, 2016.

[29] Project MinE ALS Sequencing Consortium. Project MinE: study design and pilot analyses of a large-scale whole-genome sequencing study in amyotrophic lateral sclerosis. *European Journal of Human Genetics*, 26(10):1537–1546, 2018.

[30] Wouter Van Rheenen, Rick AA Van Der Spek, Mark K Bakker, Joke JFA Van Vugt, Paul J Hop, Ramona AJ Zwamborn, Niek De Klein, Harm-Jan Westra, Olivier B Bakker, Patrick Deelen, et al. Common and rare variant association analyses in amyotrophic lateral sclerosis identify 15 risk loci with distinct genetic architectures and neuron-specific biology. *Nature genetics*, 53(12):1636–1648, 2021.

[31] Joseph H Lee, Rong Cheng, Neill Graff-Radford, Tatiana Foroud, Richard Mayeux, National Institute on Aging Late-Onset Alzheimer's Disease Family Study Group, et al. Analyses of the national institute on aging late-onset alzheimer's disease family study: implication of additional loci. *Archives of neurology*, 65(11):1518–1526, 2008.

[32] Xin Wei, Mengjiao Chen, Qi Zhang, Junyi Gong, Jie Liu, Kaicheng Yong, Qin Wang, Jiongjiong Fan, Suhui Chen, Hua Hua, et al. Genomic investigation of 18,421 lines reveals the genetic architecture of rice. *Science*, 385(6704):eadm8762, 2024.

[33] Jinfeng Zhao, Tao Wang, Minxia Wang, Yuanyuan Liu, Shoujiang Yuan, Yanan Gao, Liang Yin, Wei Sun, Lixiang Peng, Wenhui Zhang, et al. Dwarf3 participates in an scf complex and associates with dwarf14 to suppress rice shoot branching. *Plant and Cell Physiology*, 55(6):1096–1109, 2014.

[34] Yusuke Yamada, Mami Otake, Takuma Furukawa, Masato Shindo, Koichiro Shimomura, Shinjiro Yamaguchi, and Mikihisa Umehara. Effects of strigolactones on grain yield and seed development in rice. *Journal of Plant Growth Regulation*, 38(3): 753–764, 2019.

[35] Juan Zhao, Xing Liu, Mei Wang, Lingjuan Xie, Zhengxin Wu, Jiuming Yu, Yuchen Wang, Zhiqiao Zhang, Yufang Jia, and Qingpo Liu. The mir528-d3 module regulates plant height in rice by modulating the gibberellin and abscisic acid metabolisms. *Rice*, 15(1):27, 2022.

[36] Weiya Xue, Yongzhong Xing, Xiaoyu Weng, Yu Zhao, Weijiang Tang, Lei Wang, Hongju Zhou, Sibin Yu, Caiguo Xu, Xianghua Li, et al. Natural variation in ghd7 is an important regulator of heading date and yield potential in rice. *Nature genetics*, 40(6):761–767, 2008.

[37] Manjari Mishra, Ray Singh Rathore, Jayram Bagri, Rajeev Nayan Bahuguna, Ashwani Pareek, and Sneh Lata Singla-Pareek. The key contribution of osghd7 in controlling flowering time, grain yield, and abiotic stress tolerance in photoperiod-insensitive rice. *Physiologia Plantarum*, 176(6):e14632, 2024.

[38] Kangli Sun, Wubei Zong, Dongdong Xiao, Zeqiang Wu, Xiaotong Guo, Fuquan Li, Yingang Song, Shengting Li, Guangliang Wei, Yu Hao, et al. Effects of the core heading date genes hd1, ghd7, dth8, and prr37 on yield-related traits in rice. *Theoretical and Applied Genetics*, 136(11):227, 2023.

[39] Qingmei Su, Rongsheng Wang, Xiaomin Feng, Mengyue Zhao, Guanlin Zhu, Qing Wang, Fang Zhang, Shaoyang Lin, Yuhang Zhang, Li Zhu, et al. A new strategy of molecular breeding for optimal heading date and grain yield in rice by modulating elite allelic combinations of ghd7, hd3a, rft1 and gn1a. *The Crop Journal*, 13(2):490–500, 2025.

[40] Xian-Jun Song, Wei Huang, Min Shi, Mei-Zhen Zhu, and Hong-Xuan Lin. A qtl for rice grain width and weight encodes a previously unknown ring-type e3 ubiquitin ligase. *Nature genetics*, 39(5):623–630, 2007.

[41] Beom Seok Choi, Yeon Jeong Kim, Kesavan Markkandan, Yeon Jong Koo, Jong Tae Song, and Hak Soo Seo. Gw2 functions as an e3 ubiquitin ligase for rice expansin-like 1. *International Journal of Molecular Sciences*, 19(7):1904, 2018.

[42] V Mohan Murali Achary and Malireddy K Reddy. Crispr-cas9 mediated mutation in grain width and weight2 (gw2) locus improves aleurone layer and grain nutritional quality in rice. *Scientific Reports*, 11(1):21941, 2021.

[43] Xuehui Huang, Shihua Yang, Junyi Gong, Yan Zhao, Qi Feng, Hao Gong, Wenjun Li, Qilin Zhan, Benyi Cheng, Junhui Xia, et al. Genomic analysis of hybrid rice varieties reveals numerous superior alleles that contribute to heterosis. *Nature communications*, 6(1):6258, 2015.

[44] Jingyun Luo, Chengcheng Wei, Haijun Liu, Shikun Cheng, Yingjie Xiao, Xiaqing Wang, Jianbing Yan, and Jianxiao Liu. MaizeCUBIC: a comprehensive variation database for a maize synthetic population. *Database*, 2020:baaa044, 2020.

[45] Laura Ferraiuolo, Janine Kirby, Andrew J Grierson, Michael Sendtner, and Pamela J Shaw. Molecular pathways of motor neuron injury in amyotrophic lateral sclerosis. *Nature Reviews Neurology*, 7(11):616–630, 2011.

[46] Pir Masoom Shah, Huimin Zhu, Zhangli Lu, Kaili Wang, Jing Tang, and Min Li. Deepdtagen: a multitask deep learning framework for drug-target affinity prediction and target-aware drugs generation. *Nature Communications*, 16(1):5021, 2025.

[47] Sai Zhang, Johnathan Cooper-Knock, Annika K Weimer, Minyi Shi, Tobias Moll, Jack NG Marshall, Calum Harvey, Helia Ghahremani Nezhad, John Franklin, Cleide dos Santos Souza, et al. Genome-wide identification of the genetic basis of amyotrophic lateral sclerosis. *Neuron*, 110(6):992–1008, 2022.

[48] Oliver H Tam, Nikolay V Rozhkov, Regina Shaw, Duyang Kim, Isabel Hubbard, Samantha Fennessey, Nadia Propp, Hemali Phatnani, Justin Kwan, Dhruv Sareen, et al. Postmortem cortex samples identify distinct molecular subtypes of als: retrotransposon activation, oxidative stress, and activated glia. *Cell reports*, 29(5):1164–1177, 2019.

[49] Elena Abati, Nereo Bresolin, Giacomo Comi, and Stefania Corti. Silence superoxide dismutase 1 (sod1): a promising therapeutic target for amyotrophic lateral sclerosis (als). *Expert opinion on therapeutic targets*, 24(4):295–310, 2020.

[50] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel AR Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul IW De Bakker, Mark J Daly, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American journal of human genetics*, 81(3):559–575, 2007.

[51] Kai Wang, Mingyao Li, and Hakon Hakonarson. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research*, 38(16):e164–e164, 2010.

[52] Brian L Browning, Ying Zhou, and Sharon R Browning. A one-penny imputed genome from next-generation reference panels. *The American journal of human genetics*, 103(3):338–348, 2018.

[53] Qian Wang, Shan Jiang, Tong Li, Zhixu Qiu, Jun Yan, Ran Fu, Chuang Ma, Xiangfeng Wang, Shuqin Jiang, and Qian Cheng. G2p provides an integrative environment for multi-model genomic selection analysis to improve genotype-to-phenotype prediction. *Frontiers in Plant Science*, 14:1207139, 2023.

[54] Oleg Trott and Arthur J Olson. Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry*, 31(2):455–461, 2010.

[55] Trevor Park and George Casella. The bayesian lasso. *Journal of the american statistical association*, 103(482):681–686, 2008.

[56] Gustavo de los Campos, Daniel Gianola, and Guilherme JM Rosa. Reproducing kernel hilbert spaces regression: a general framework for genetic evaluation. *Journal of animal science*, 87(6):1883–1887, 2009.

[57] Hyonho Chun and Sündüz Keleş. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 72(1):3–25, 2010.

[58] Michael P LaValley. Logistic regression. *Circulation*, 117(18):2395–2399, 2008.

[59] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[60] Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28, 1998.

[61] Cao Ying, Miao Qi-Guang, Liu Jia-Chen, and Gao Lin. Advance and prospects of adaboost algorithm. *Acta Automatica Sinica*, 39(6):745–758, 2013.

[62] Tianzhi Wu, Erqiang Hu, Shuangbin Xu, Meijun Chen, Pingfan Guo, Zehan Dai, Tingze Feng, Lang Zhou, Wenli Tang, LI Zhan, et al. clusterprofiler 4.0: A universal enrichment tool for interpreting omics data. *The innovation*, 2(3), 2021.

[63] Saba N Baskoylu, Jill Yersak, Patrick O'Hern, Sarah Grosser, Jonah Simon, Sarah Kim, Kelsey Schuch, Maria Dimitriadi, Katherine S Yanagi, Jeremy Lins, et al. Single copy/knock-in models of als sod1 in c. elegans suggest loss and gain of function have different contributions to cholinergic and glutamatergic neurodegeneration. *PLoS genetics*, 14(10):e1007682, 2018.

[64] Sydney Brenner. The genetics of caenorhabditis elegans. *Genetics*, 77(1):71–94, 1974.

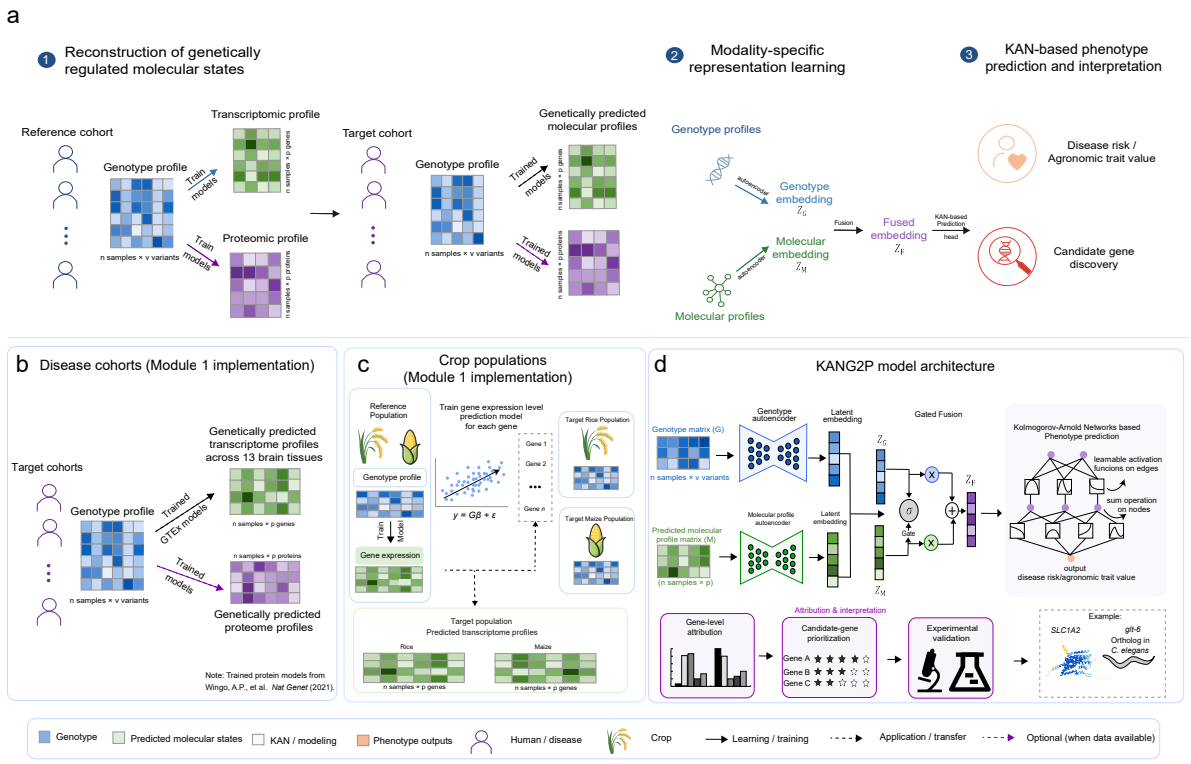


Figure 1. KANG2P couples genetically regulated molecular-state reconstruction with nonlinear phenotype prediction. **a,** Overview of the KANG2P framework. Genotype–omics relationships learned in external reference panels are first used to reconstruct individualized genetically regulated transcriptomic or proteomic states in target cohorts. Genotype and predicted molecular profiles are then encoded by modality-specific autoencoders, and the resulting latent representations are integrated by a Kolmogorov–Arnold Network for disease classification or quantitative-trait prediction. Gene-level attribution supports candidate-gene prioritization and experimental follow-up. **b,** Disease-specific implementation of molecular-state reconstruction. Genotype–Tissue Expression (GTEx) Project v8 brain-tissue models were used to predict expression across 13 brain tissues, and published genotype-to-protein prediction models were used where proteomic integration was evaluated. **c,** Crop-specific implementation. Gene-expression prediction models were trained in reference populations with matched genotype and transcriptome data and applied to target rice and maize genotype cohorts. **d,** KANG2P prediction architecture. Genotype and predicted molecular profiles are processed by separate autoencoders, and their latent representations are fused and passed to a Kolmogorov–Arnold Network. The output corresponds to disease probability for human cohorts or quantitative trait values for crop cohorts.

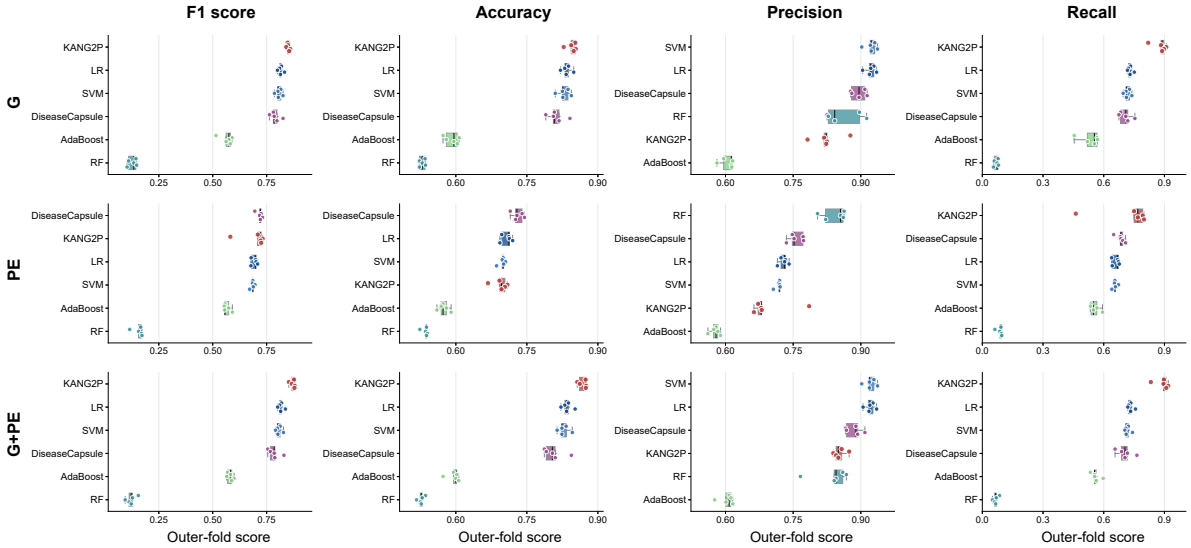


Figure 2. Comparison of predictive performance across six classification methods on ALS datasets. Box plots show the outer-fold performance of six classification models evaluated by nested five-fold cross-validation on the ALS dataset, stratified by input feature set and evaluation metric. Rows correspond to three input settings: genotype only (G), predicted expression profile only (PE), and the integrated genotype plus predicted expression profile (G+PE). Columns show accuracy (ACC), F1 score, precision and recall. For each box, the centre line denotes the median, the box bounds indicate the 25th and 75th percentiles, and whiskers extend to 1.5× the interquartile range. Each box summarizes five outer-fold estimates from nested cross-validation. LR, logistic regression; RF, random forest; SVM, support vector machine.

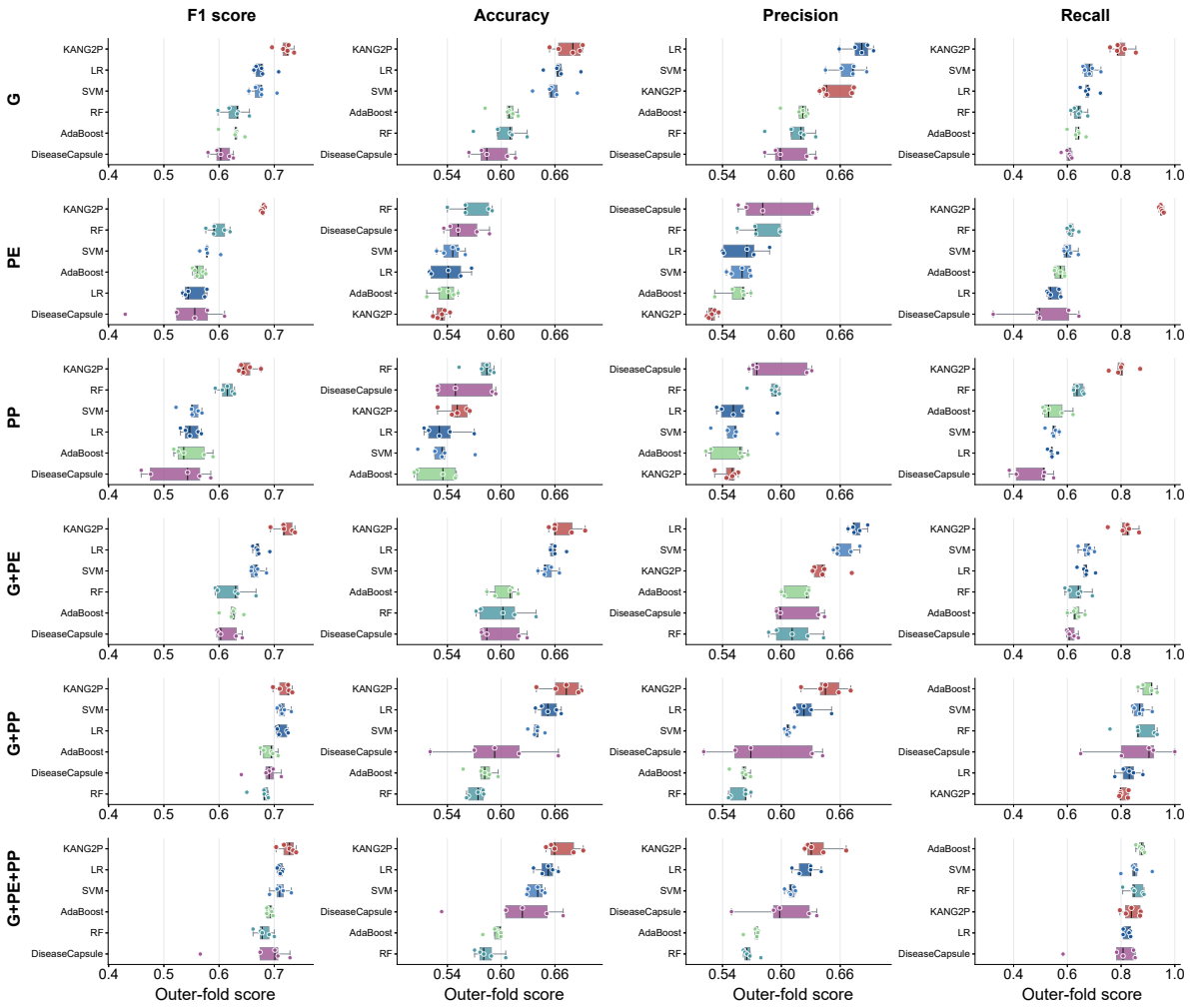


Figure 3. Comparison of predictive performance across six classification methods on AD datasets. Box plots show the outer-fold performance of six classification models evaluated by nested five-fold cross-validation on the ALS dataset, stratified by input feature set and evaluation metric. Rows denote the six input settings: genotype (G), predicted expression profile (PE), predicted protein abundance (PP), and their integrated combinations (G+PE, G+PP and G+PE+PP). Columns show accuracy , F1 score, precision and recall. For each box, the centre line denotes the median, the box bounds indicate the 25th and 75th percentiles, and whiskers extend to 1.5× the interquartile range. Each box summarizes five outer-fold estimates from nested cross-validation. LR, logistic regression; RF, random forest; SVM, support vector machine.

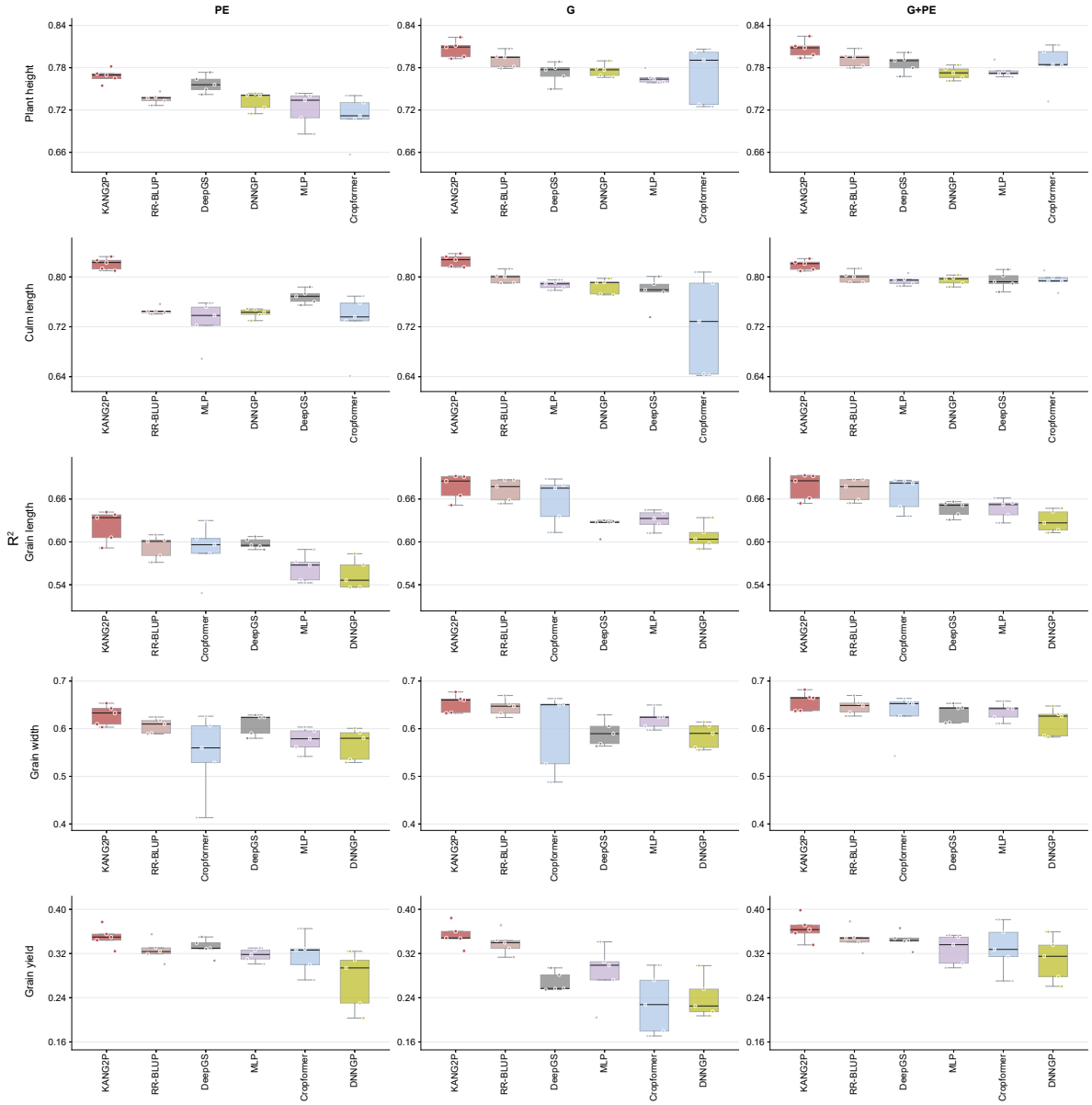


Figure 4. Predictive performance comparison of six regression models across Rice18K agronomic traits. Box plots show outer-fold R^2 values from nested cross-validation for six regression models across five Rice18K agronomic traits. Rows indicate traits and columns indicate input feature settings: genotype only (G), predicted expression profile only (PE) and the integrated combinations (G+PE). Each point represents one outer-fold estimate ($n = 5$). For each box, the centre line denotes the median, box bounds indicate the 25th and 75th percentiles, and whiskers extend to 1.5 times the interquartile range. MLP, multilayer perceptron.

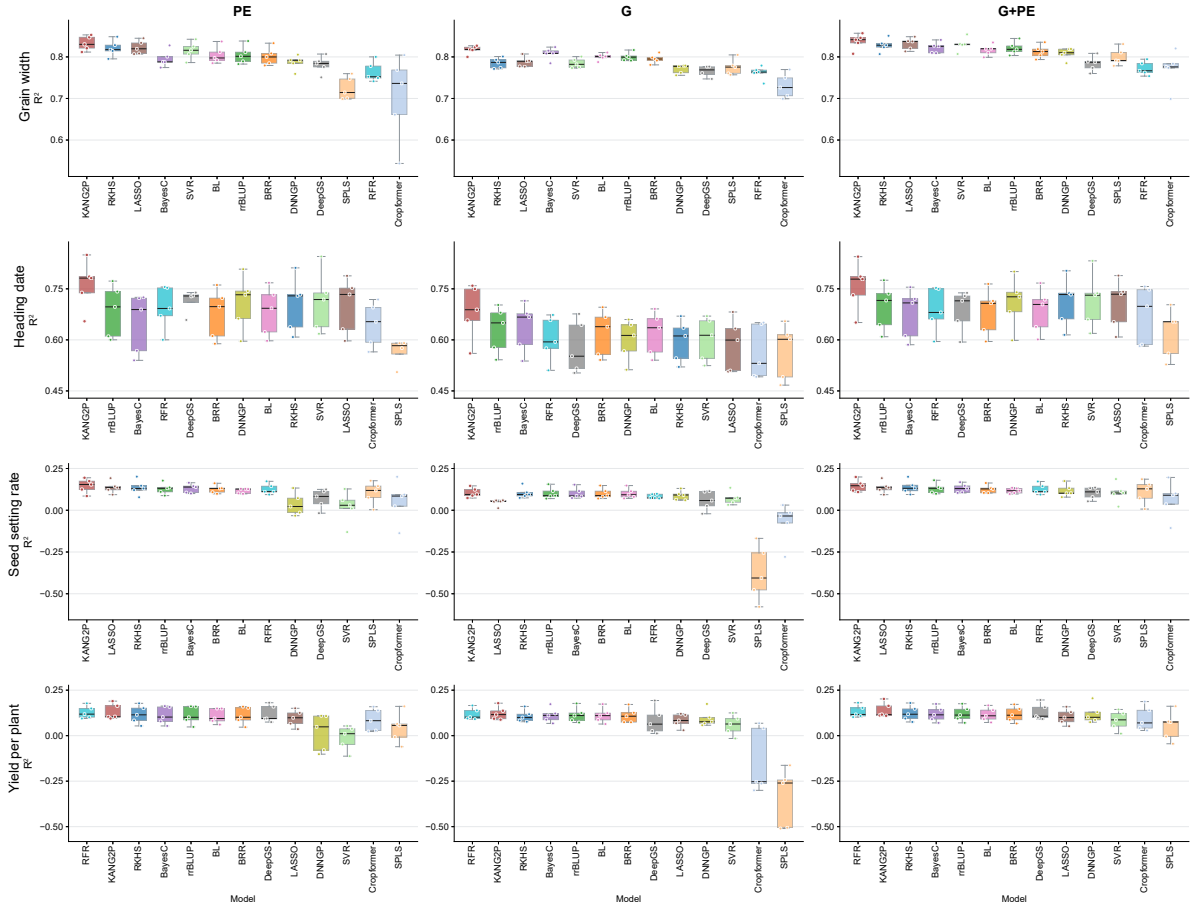


Figure 5. Benchmarking 13 regression models across agronomic traits in the Rice1495 population. Box plots show outer-fold R^2 values from nested cross-validation for 13 regression models across four agronomic traits. Rows indicate traits and columns indicate input feature settings: genotype only (G), predicted expression profile only (PE) and the integrated combinations (G+PE). Each point represents one outer-fold estimate ($n = 5$). For each box, the centre line denotes the median, box bounds indicate the 25th and 75th percentiles, and whiskers extend to 1.5 times the interquartile range. RFR, Random Forest regressor. BL, Bayes LASSO. BRR, Bayes ridge regression. RKHS, Reproducing kernel Hilbert space. SPLS, sparse partial least squares. SVR, support vector machine regressor. RFR, Random Forest regressor.

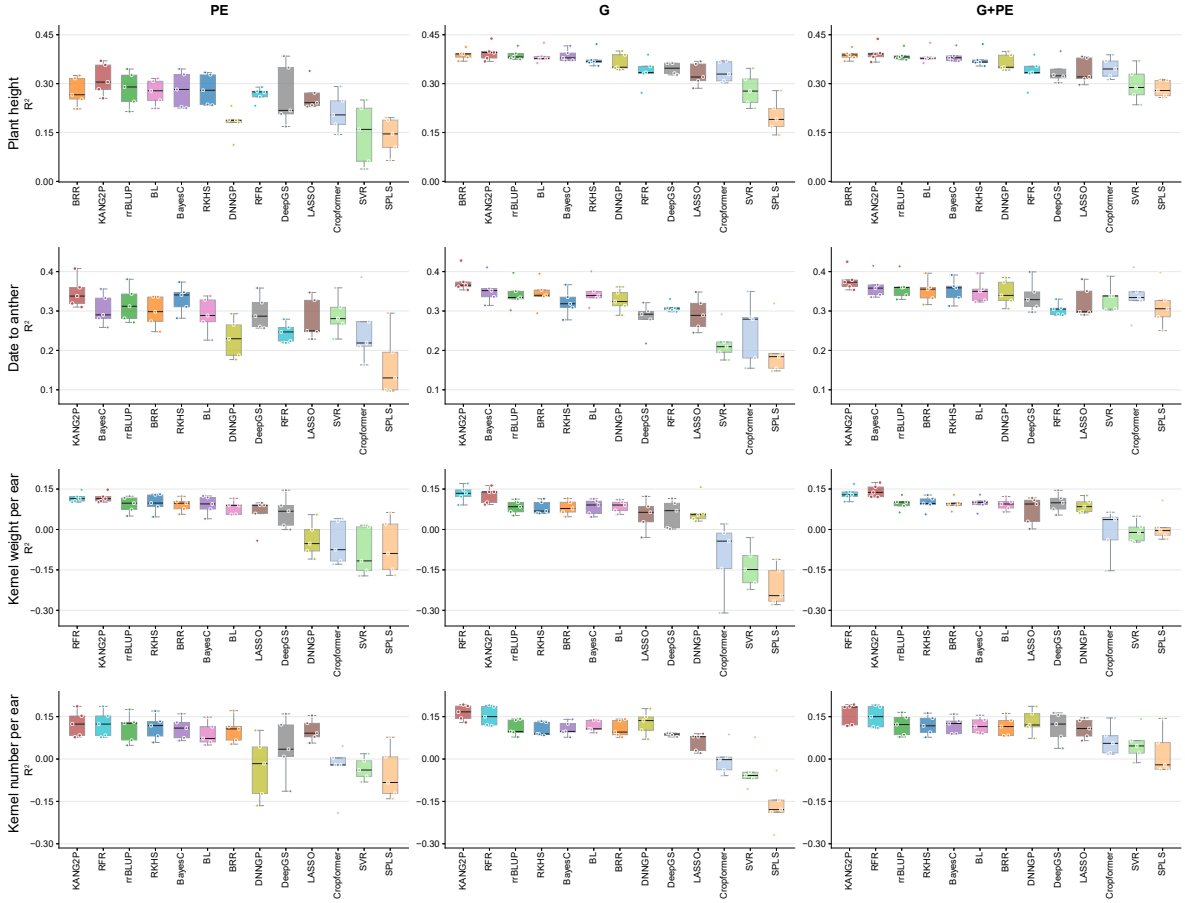


Figure 6. Benchmarking 13 regression models across agronomic traits in the maize CUBIC population.. Box plots show outer-fold R^2 values from nested cross-validation for 13 regression models across four agronomic traits. Rows indicate traits and columns indicate input feature settings: genotype only (G), predicted expression profile only (PE) and the integrated combinations (G+PE). Each point represents one outer-fold estimate ($n = 5$). For each box, the centre line denotes the median, box bounds indicate the 25th and 75th percentiles, and whiskers extend to 1.5 times the interquartile range. RFR, Random Forest regressor. BL, Bayes LASSO. BRR, Bayes ridge regression. RKHS, Reproducing kernel Hilbert space. SPLS, sparse partial least squares. SVR, support vector machine regressor. RFR, Random Forest regressor.

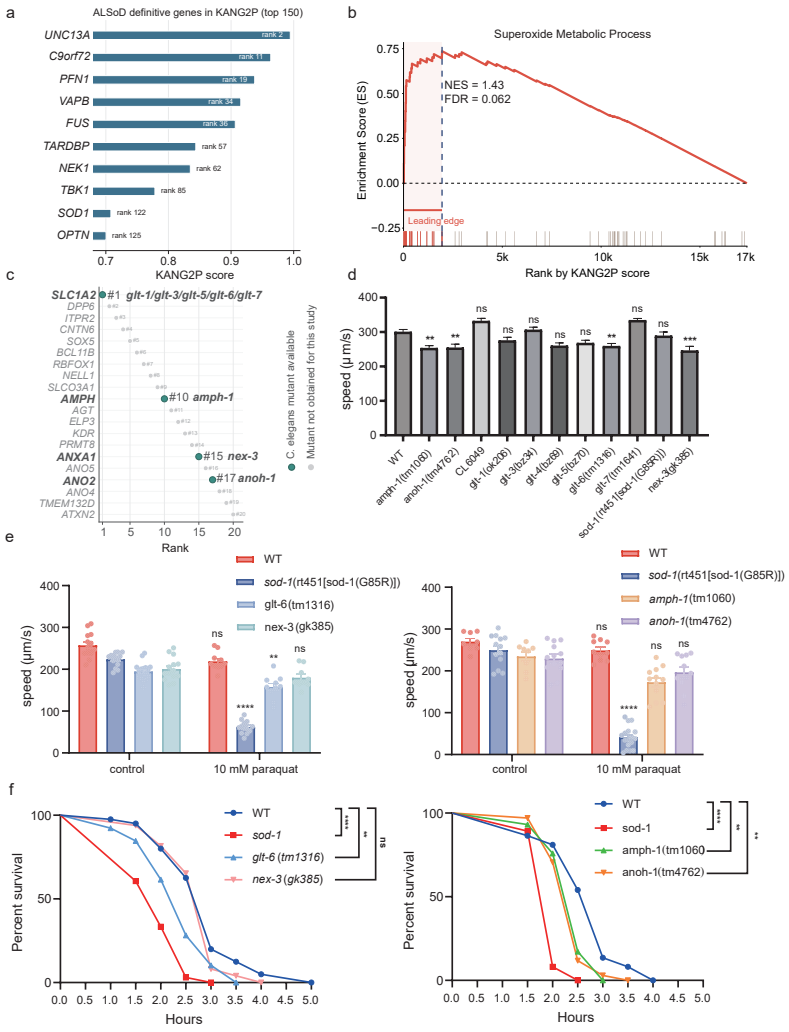
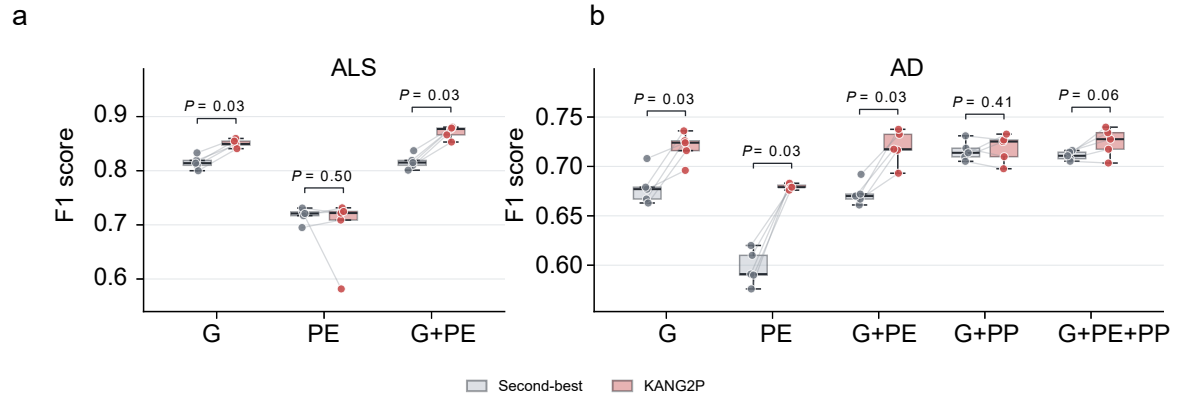
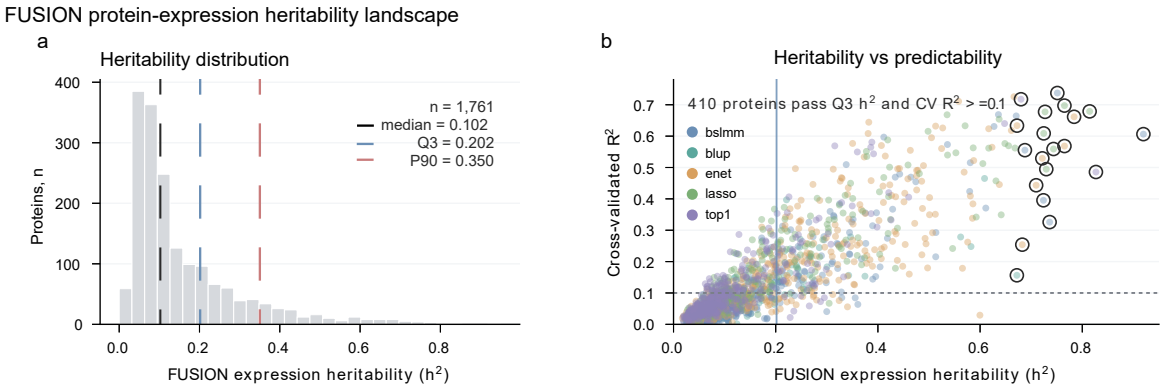


Figure 7. KANG2P-guided interpretability and integrative prioritization nominate ALS candidate genes for functional validation. a, Preranked gene set enrichment analysis (GSEA) of human genes ordered by KANG2P-derived attribution scores. The GO Biological Process term *superoxide metabolic process* was enriched among genes with higher KANG2P scores, linking the model-derived ranking to oxidative-stress biology. The running enrichment score, gene-set hits and ranked metric are shown. NES, normalized enrichment score; FDR, false discovery rate. **b,** KANG2P attribution ranking of the top 100 human genes, with curated definitive ALS genes and *SLC1A2* annotated. *SLC1A2* ranked 92nd in the raw KANG2P gene-level ranking, indicating a detectable model-derived signal but not a top-ranked hit by KANG2P alone. **c,** Top 20 genes ranked by the multi-method weighted prioritization score. The integrated score combined KANG2P-derived attribution with complementary evidence layers described in the Methods. **d,** Basal locomotion speed of control and indicated mutant animals in the hTDP-43 transgenic background. The control strain CL6049 carries *snb-1p*::hTDP-43/3' UTR and *mtl-2p*::GFP. Each group included $n \geq 15$ animals. Statistical significance was assessed by one-way analysis of variance (ANOVA) followed by Dunnett's multiple-comparisons test against the control group. **e,** Paraquat-induced locomotion phenotypes. Dose-response analysis in control and *sod-1* mutant animals was used to select 10 mM paraquat for subsequent oxidative-stress validation, under which locomotion speed was measured in control and indicated mutant animals. Statistical significance was assessed by two-way ANOVA followed by Tukey's multiple-comparisons test. **f,** Lifespan of control and indicated mutant animals under paraquat-induced oxidative stress. Survival distributions were compared using the log-rank Mantel–Cox test. For all panels, ns, not significant; $^{**}P < 0.01$, $^{***}P < 0.001$, $^{****}P < 0.0001$. Exact P values and sample sizes are provided in the Source Data.

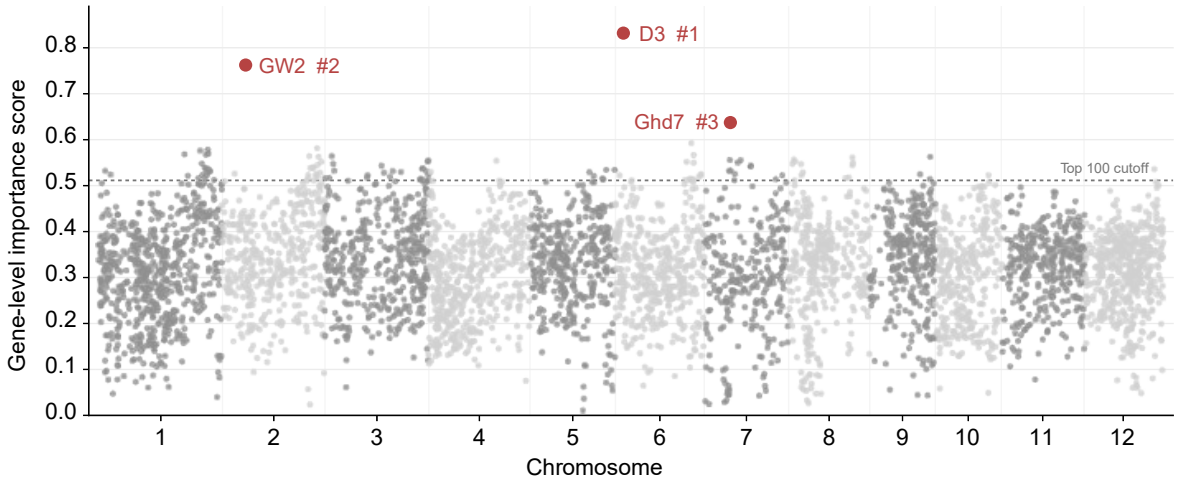
Extended data



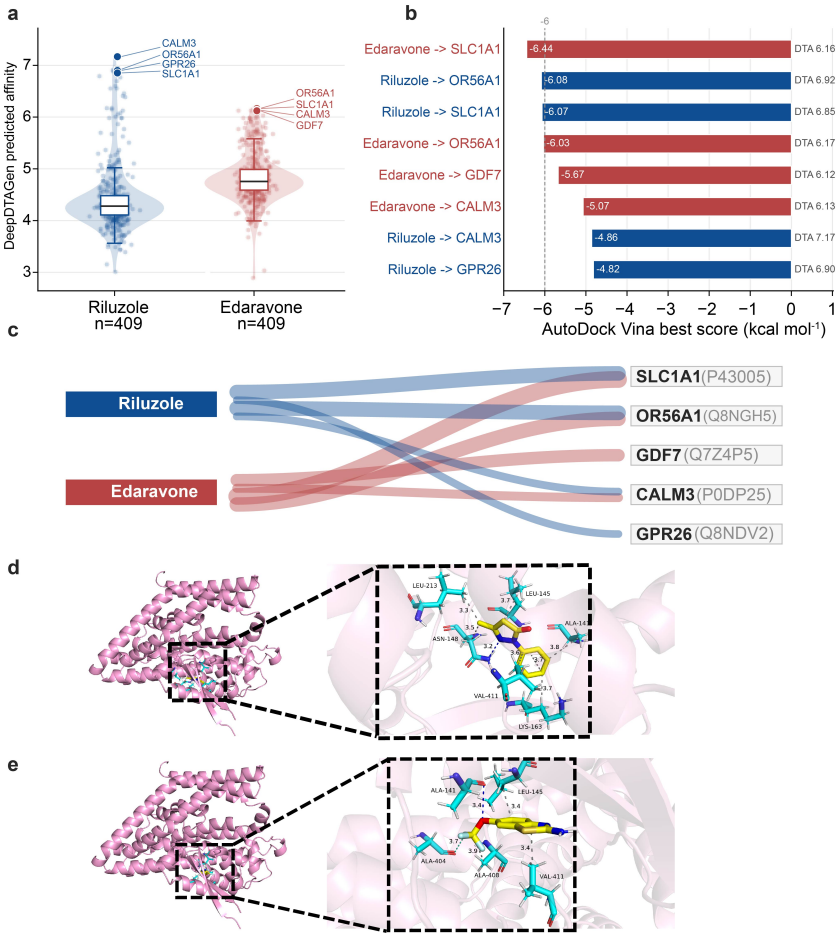
Extended Data Fig. 1. KANG2P improves disease-risk prediction across neurodegenerative diseases. **a,b,** F1 scores of KANG2P compared with the strongest non-KANG2P model for each feature setting in amyotrophic lateral sclerosis (ALS; a) and Alzheimer's disease (AD; b). Models were evaluated using genotype (G), predicted expression (PE), predicted proteome (PP) and their integrated representations. Each point represents one outer fold from nested cross-validation, with five outer folds in total for each comparison. Grey lines connect paired results from the same outer fold. Boxes summarize the distribution of F1 scores across folds, with centre lines indicating the median. Second-best denotes the best-performing non-KANG2P method under the corresponding disease and modality. *P* values were calculated using a one-sided paired Wilcoxon signed-rank test comparing KANG2P with the second-best comparator within each feature setting.



Extended Data Fig. 2. Cis-genetic architecture of protein expression in ROSMAP. **a**, Frequency distribution of cis-genetic heritability estimates for protein expression traits in the Religious Orders Study and Rush Memory and Aging Project (ROSMAP). **b**, Scatter plot showing the relationship between cis-genetic heritability and prediction accuracy of cis-protein expression models. Each point represents one protein.



Extended Data Fig. 3. Gene-level interpretability landscape of the KANG2P model for rice grain yield prediction. Manhattan plot showing gene-level attribution scores derived from the KANG2P model trained for rice yield trait prediction. Each point represents an annotated gene, arranged according to its physical position across the rice genome. The y axis denotes the gene-level KANG2P interpretability score, reflecting the relative contribution of each gene to the model prediction. Genes with high attribution scores form genomic peaks and were prioritized as candidate genes potentially associated with yield-related variation.



Extended Data Fig. 4. Computational prioritization of candidate targets of approved ALS drugs among KANG2P-ranked genes. a, DeepDTAGen-predicted affinity scores for riluzole and edaravone across proteins encoded by the top 500 KANG2P-prioritized ALS candidate genes. A total of 409 genes could be mapped to valid protein sequences and were evaluated for each drug. Each point represents one drug–protein pair; violin plots show the score distributions, and box plots indicate the median and interquartile range. The four highest-scoring candidate proteins for each drug are annotated. **b,** AutoDock Vina docking scores for selected high-ranking DeepDTAGen drug–target pairs. Bars show the best docking score obtained for each pair, with riluzole and edaravone shown in blue and red, respectively; the corresponding DeepDTAGen affinity scores are indicated on the right. The dashed vertical line marks a docking score of $-6\,\mathrm{kcal\,mol^{-1}}$. **c,** Sankey representation of the partially overlapping and drug-specific candidate-target profiles of riluzole and edaravone. UniProt accessions are shown for the corresponding proteins. **d,e,** Representative docking poses of edaravone (**d**; best Vina score, $-6.44\,\mathrm{kcal\,mol^{-1}}$) and riluzole (**e**; best Vina score, $-6.07\,\mathrm{kcal\,mol^{-1}}$) with *SLC1A1* (UniProt accession P43005). Complete receptor structures and enlarged views of the predicted binding sites are shown. Ligands are displayed as yellow sticks, interacting residues as cyan sticks and *SLC1A1* as a pink cartoon; annotated distances are given in Å. These docking analyses provide computational support for candidate prioritization.

Supplementary Files

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

- KANG2PSupplementaryinformation.pdf