Ensembles of Graph Neural Networks Supervised by Genotype-to-Phenotype Structures Improved Genomic Prediction Performance

Shunichiro Tomura[1,2,*], Owen Powell[1,2,3], Melanie J. Wilkinson[1,2,3], Mark Cooper[1,2,3]

[1]Queensland Alliance for Agriculture and Food Innovation (QAAFI), Centre for Crop Science, the University of Queensland, St Lucia, QLD 4072, Australia
[2]ARC Centre of Excellence for Plant Success in Nature and Agriculture, the University of Queensland, St Lucia, QLD 4072, Australia
[3]ARC Training Centre in Predictive Breeding, the University of Queensland, St Lucia, QLD 4072, Australia

*Corresponding author: s.tomura@uq.edu.au

ORCID
Shunichiro Tomura: 0009-0005-9050-9963
Owen Powell: 0000-0001-5131-4286
Melanie J. Wilkinson: 0000-0003-1400-3176
Mark Cooper: 0000-0002-9418-3359

Abstract

Accurate selection of favourable crop genotypes has motivated the exploration of diverse prediction algorithms for crop breeding applications. One genomic prediction method that has not been fully explored is graph attention networks (GAT). By directly analysing graphical data with the attention mechanism, GAT can incorporate the genotype-to-phenotype (G2P) structure to regularise predictions. There are many possible graph structures informed by knowledge of trait genetic architecture. The graphical representation of the inferred trait genetic architecture can be used as prior knowledge for GAT to effectively learn key features of trait genetic architecture and enhance the prediction patterns. Here, we investigated whether incorporating prior knowledge into GAT improved performance compared to GAT models representing a continuum of G2P structures, ranging from infinitesimal to fully connected. Applying the Diversity Prediction Theorem, we also combined these diverse G2P structures into an ensemble of GAT genomic prediction models to integrate complementary strengths of multiple models. The results for flowering time traits in two maize nested association mapping datasets showed a lack of consistent performance improvement in the prior knowledge GAT model. However, consistent outperformance was observed for the ensemble of GAT models. Improved predictions from the ensemble model may be driven by its ability to capture a more complete representation of the trait genetic architecture through the integration of information from diverse G2P structures, as is proposed by the Diversity Prediction Theorem. Our results show that an ensemble of GAT models can enhance prediction performance.

Introduction

Improving the accuracy of genomic prediction models (Meuwissen et al., 2001) has been a primary research target in crop breeding, motivated by the need for rapid and precise selection of heritable variation to accelerate genetic gain (Voss-Fels et al., 2019; Escamilla et al., 2025). This research target has promoted the continuous exploration of novel algorithms. From the early stages of genomic prediction research, parametric approaches, such as genomic best linear unbiased prediction (GBLUP) and Bayesian regression models, have been the primary target of investigations (Azodi et al., 2019). While the parametric models are still widely employed in crop breeding today, the emergence of non-parametric models, such as machine learning and deep learning models, has diversified genomic prediction models, aiming to capture the complex genotype-by-genotype interactions underlying genomic markers (Cooper et al., 2025; Crossa et al., 2025).

Applications of graph neural networks (Scarselli et al., 2008) provide opportunities to emulate the networks of genes that contribute to the genetic architecture of traits. The algorithmic and technical advancement in deep learning models has enabled the analysis of graphically structured datasets using graph neural networks (Bhatti et al., 2023; Vatter et al., 2023). One considerable advantage of graph neural networks is the direct incorporation of topology connecting nodes (storing a collection of feature values) grounded in underlying relationships that are analysed by deep learning (Scarselli et al., 2008). In particular, graph attention network (GAT; Veličković et al., 2018) models have been a widely utilised graph neural network model due to the flexibility of the attention mechanism enhancing performance by allocating higher weights to critical predictive information (Brody et al., 2021). GAT has improved the prediction performance in various tasks and been used as a benchmark model for the evaluation of new prediction models mentioned above due to its high prediction performance. Graph neural networks, including GAT, have already been applied to diverse research areas in biology, including drug discovery (Abate et al., 2023; Wang et al., 2025a), gene sequence analysis (Wang and Zhang, 2021; Wang et al., 2023) and disease discovery (Hernández-Lorenzo et al., 2022; Duan et al., 2023). While graph neural networks have been employed in several cases in crop breeding, such as crop yield prediction by converting the information from farming plots into nodes (Gupta and Singh, 2023) and suitability evaluation for crops (Zhang et al., 2022), the possibility of using graph neural networks for genomic prediction remains largely unexplored.

In genomic prediction for crop breeding, graph neural networks can predict target traits while considering the possible relationship between genomic markers and target phenotypes that represent the genotype-to-phenotype (G2P) structure. One possible representation of a G2P structure can be a gene network that can be used as prior knowledge, when evidence indicates traits are regulated by complex genomic marker-by-marker interactions (Cooper et al., 2005; Powell et al., 2022). The direct relationship between genomic markers and traits can represent additive effects, whereas the interactions between the genomic markers can be used to represent non-additive effects. Prior knowledge provides prediction models with problem-specific external information, expected to guide the prediction models to efficiently capture key features that will have the greatest influence on prediction outcomes (Von Rueden et al., 2021). The benefit of prior knowledge inclusion is especially emphasised when the training set size is small (Von Rueden et al., 2021), which is true for many crop breeding programs that can only collect a small scale of phenotype data (Mascher et al., 2024). Consequently, the integration of such gene network information as prior knowledge into genomic prediction models may help capture key additive and non-additive effects, expected to improve prediction performance (Zhao et al., 2025).

While graph neural networks with prior knowledge can be leveraged to improve prediction performance, other possible G2P structures have already been proposed to represent the biological relationship between genomic markers and traits. For instance, infinitesimal structures assume no explicit interactions among genomic markers (Meuwissen et al. 2001; Cooper et al., 2021). In contrast, fully connected graph structures may detect active non-additive interactions by allowing genomic markers to directly interact with others. Prior knowledge graph structure lies between the infinitesimal and fully connected structure, expected to capture more accurate trait genetic architecture by removing non-additive interactions that hardly influence the traits. This continuum of different G2P structures (Figure 1; the infinitesimal, prior knowledge and fully connected) captures different views of trait genetic architecture that is usually constructed by the mixture of additive and non-additive effects, especially for complex traits. In such cases, the prediction performance of graph neural networks based on respective G2P structures can potentially capture distinct prediction patterns. This has not explicitly been investigated in genomic prediction studies for crop breeding despite its potential for influencing prediction performance.

An ensemble of multiple graph neural networks grounded in different G2P structures could be another approach to improve prediction performance, as proposed by the Diversity Prediction Theorem (Page, 2018; Messina et al., 2025). The theorem suggests that an ensemble of diverse individual prediction models can be constructed to lower the prediction error and improve prediction accuracy. Experimental results have demonstrated that ensembles of multiple genomic prediction models have consistently improved prediction performance by increasing prediction accuracy and reducing prediction error in comparison to the individual genomic prediction models (McCormick et al., 2021; Kick and Washburn, 2023; Tomura et al., 2025b). Hence, the ensemble of genomic prediction models, guided by diverse G2P graph structures, can also be evaluated for potential improvements beyond individual genomic prediction models.

Here, we evaluated the prediction performance of graph neural networks with three G2P graph structures (infinitesimal, fully connected and prior knowledge from an inferred gene network) and an ensemble of these graph neural networks. This study investigated the following three objectives. Firstly, the prediction performance of GAT models designed to utilise prior gene network knowledge was compared with other GAT models based on G2P graph structures that did not use the prior gene network knowledge. The prediction performance of the ensemble GAT models was also evaluated by comparing with each GAT model. Secondly, the performance decay of the GAT models with decreasing training set size was compared to analyse which GAT model is relatively less responsive to reduced training set size. Thirdly, the trait genetic architecture from each GAT model was compared to understand the effect of diversity in genomic marker effects from each model on the results achieved by the ensemble model.

Materials and Methods

1. Datasets

We leveraged two maize (*Zea Mays*) nested association mapping (NAM) datasets in this study, the TeoNAM (Chen et al., 2019) and MaizeNAM (Buckler et al., 2009). The TeoNAM dataset is based on samples of recombinant inbred lines (RILs) from five populations. The maize inbred line W22 was crossed with five teosinte types, four from *Z. mays ssp. parviglumis* and one from *Z. mays ssp. mexicana*. In each population,

RILs were developed by controlled self-pollination following backcrossing W22 with one of the teosinte types after the initial F1 cross. All TeoNAM experiments were conducted at the University of Wisconsin West Madison Agricultural Research Station using a randomised complete block design. The TeoNAM dataset was generated by evaluating each set of RILs in two environments: the W22TIL01, W22TIL03 and W22TIL11 populations were evaluated in the summer of 2015 and 2016. The W22TIL14 population was evaluated in the summer of 2016 and 2017. The W22TIL25 population was evaluated in two environments within the 2017 summer.

The MaizeNAM dataset has 25 populations of RILs from crosses between the maize inbred B73 and 25 inbreds sampled from temperate and tropical regions. The 25 RIL populations were developed by advancing the F1 crosses to the F5 generation by controlled self-pollination. Each RIL population was evaluated in eight environments with a randomised design in the United States. Two experiments were conducted at each location in Aurora (New York), Clayton (North Carolina), Columbia (Missouri) and Urbana (Illinois). The total number of genomic markers and RILs from each dataset is summarised in Table S1.

The primary reason for selecting these two datasets is their contrast in genetic diversity. The TeoNAM dataset crossed a domesticated maize line with teosinte, an ancestral species of maize, leading to variation in allelic frequency for many genomic regions between the crosses. In contrast, the MaizeNAM contains crosses among domesticated inbred lines developed by breeding programs, thereby driving higher levels of fixation of alleles through selective breeding in the MaizeNAM populations relative to the TeoNAM populations. Using these two NAM datasets with different levels of genomic marker diversity, we can investigate the prediction performance of the proposed approaches under various prediction scenarios.

The two flowering time-related traits, days to anthesis (DTA) and anthesis to silking interval (ASI), were chosen for investigation in this study. For DTA, the genetic control of flowering time in maize has previously been investigated with the identification of gene pathways (Dong et al., 2012), extended by further investigation of the genetic basis for response to selection (Wisser et al., 2019). Key gene regions and regulatory pathways identified from the previous studies can be compared with the extracted trait genetic architecture from GAT models in this study. ASI is a secondary trait representing the difference between DTA and days to silking (DTS) (Leng et al., 2022), and hence the gene network might be regulated by both DTA

142 and DTS and other mechanisms involved in the synchrony of development of the male and female flowers.
143 This potential greater complexity in the ASI network relative to the days to flowering traits might affect the
144 prediction performance. Using these two distinctive traits, we can investigate the predictive behaviour of the
145 GAT and ensemble models in detail for a range of trait scenarios. For the TeoNAM dataset, the reported (Chen
146 et al. 2019) phenotype values from two environments were used as input for the genomic prediction models.
147 For the MaizeNAM dataset, the reported (Buckler et al. 2009) Best linear unbiased predictors (BLUPs) of
148 each RIL were used as phenotype values.

149

150 2. Data preprocessing

151 Prior to implementing the genomic prediction models, three data preprocessing methods were applied to the
152 datasets: (1) imputing missing genomic marker alleles, (2) removing missing phenotypes and (3) reducing the
153 total genomic markers. For imputing missing genomic marker alleles, the flanking marker imputation
154 approach was utilised for both datasets. This approach imputes the missing allele information with the allele
155 of the closest flanking marker. For the TeoNAM dataset, the flanking marker imputation approach is detailed
156 in Tomura et al. (2025b) and for the MaizeNAM dataset, the approach is detailed in Buckler et al. (2009). For
157 the genomic marker reduction, linkage disequilibrium (LD) was leveraged to filter genomic markers. If LD
158 was greater than 0.8, then one of the genomic markers was removed. The LD-filtering was implemented by
159 PLINK (v1.9) (Chang et al., 2015) with a 30kb window and step size of 5 genomic markers, for both the
160 TeoNAM and MaizeNAM datasets.

161 To allow for genotype-by-environment effects in the TeoNAM dataset, we combined the data from the
162 two environments. We concatenated the two environments with a factor variable explicitly representing each
163 environment. This enabled us to evaluate whether the genomic prediction models can predict phenotypes with
164 higher performance using available environmental information.

165

166 3. Genomic prediction models

In this study, all GAT (Veličković et al., 2018) models and interpretative analyses were implemented using the computational tool, Ensemble AnalySis with Interpretable Genomic Prediction (EasiGP; Tomura et al., 2025a) in Python (v3.11.10). GAT is a graph neural network that can directly analyse graphically structured data to predict target values. Graph neural networks extract key information for predictions from nodes and edges in graphs as embeddings, concentrating the information into vectors of numerical format. Hence, the extraction of embeddings containing key predictive information is a critical aspect of graph neural networks.

One significant characteristic of the GAT model is the application of a self-attention mechanism (Brody et al., 2021), assigning a weight to each piece of information from graph components based on the level of their relevance towards prediction targets. Assigning a heavier weight to highly relevant information during the embedding process helps the GAT model precisely capture key information from graphs. The attention mechanism can be written as below (Brody et al., 2021):

$$\alpha_{ij} = \frac{exp\left(LeakyReLU\left(\boldsymbol{a}^{\top}\left[\boldsymbol{W}\boldsymbol{h}_{\boldsymbol{i}}||\boldsymbol{W}\boldsymbol{h}_{j}\right]\right)\right)}{\sum_{j'\in N_i} exp\left(LeakyReLU\left(\boldsymbol{a}^{\top}\left[\boldsymbol{W}\boldsymbol{h}_{i}||\boldsymbol{W}\boldsymbol{h}_{j'}\right]\right)\right)} \tag{1}$$

where $\boldsymbol{a}^{\top} \in R^{2d'}$ is the transformed weight vector, d' is the total number of attributes in each node, $\boldsymbol{W}$ is the weight matrix, i is the target node, $j \in N_i$ is a neighbouring node of the node i, $\boldsymbol{h}_j = \{h_1, h_2, \dots, h_N\}$ is the attribute set of the neighbouring nodes and $||$ concatenates vectors. The Leaky Rectified Linear Unit (LeakyReLU; Maas et al., 2013) activation function nonlinearly converts attention values. The attention value is normalised at the end. This calculation process can be iterated K times as respective attention layers to stabilise the normalised attention values (multi-head attention mechanism) (Veličković et al., 2018):

$$\boldsymbol{h}_i' = ||_{k=1}^{K}\sigma\left(\sum_{j\in N_i}\alpha_{ij}^{k}\boldsymbol{W}^{k}\boldsymbol{h}_j\right) \tag{2}$$

where σ is the nonlinear activation function and $\boldsymbol{h}_i'$ is the updated attributes (embeddings) of the target node. At the final layer, the mean embedding values were calculated from the embedding values of each attention layer.

Three distinctive GAT models (infinitesimal model, fully connected model and prior knowledge model) were developed based on three forms of biological interaction structures (Figure 1). The infinitesimal model (Figure 1a) assigned no edges between the genomic marker nodes to explicitly represent the independent contributions of the additive genomic marker effects. Thus, for the infinitesimal GAT model, there are no non-additive marker effects allowed. The fully connected model (Figure 1b) predicted phenotypes using a graph with genomic marker nodes connected to all the other nodes with undirected edges. In this case, all genomic markers are assumed to interact explicitly with other genomic markers to capture the potential for non-additive marker effects between all combinations of markers. The prior knowledge model (Figure 1c) is intermediate between the infinitesimal and fully connected extremes. The implementation of the prior knowledge model considered here was supervised by a gene network inferred from a random forest (RF; Breiman, 2001) equipped with Shapley scores (Shapley, 1953; Lundberg and Lee, 2017) as the prior knowledge (the calculation method is discussed in the next subsection). Shapley scores were returned to every pair of genomic markers, and the top 20% of the genomic marker pairs with the highest Shapley scores were selected to draw undirected edges for the prior knowledge graph. The extracted gene network only leverages the information of key genomic marker-by-marker interactions identified by the Shapley scores, thereby expected to remove complexity and noise generated from connecting all genomic marker nodes with edges as observed in the fully connected model. The prediction performance of RF was used as a reference for the prior knowledge GAT model since both predict phenotypes based on the same inferred trait genetic architecture. The prediction performance of the prior knowledge GAT model was expected to be higher than RF due to the incorporation of the attention mechanism.

For the TeoNAM dataset, one hidden layer with 50 neurons was added with a dropout rate of 0. The Exponential Linear Unit (ELU; Clevert et al., 2015) function was leveraged as an activation function. The Adaptive Moment Estimation (Adam; Kingma, 2014) with a learning rate of 0.005 and a weight decay rate of 0.0004 was used for the optimisation. The number of heads was 1. The number of epochs was set as 200 for the infinitesimal model and 50 for the remaining models. For the MaizeNAM dataset, the same hyperparameter settings in the TeoNAM dataset were used for the number of hidden layers, neurons, activation function, optimisation function and head number. The dropout rate and epoch numbers were set as 0.5 and 200,

respectively, across the GAT models. To extract key gene interactions for the prior knowledge model, RF was implemented with a total tree number of 1,000, while the default settings were used for the other hyperparameters.

Following the framework of the Diversity Prediction Theorem (Page, 2018), discussed further below, the naïve ensemble approach of Tomura et al. (2025b) was applied to the GAT models, which calculated the arithmetic mean of the predicted phenotypes from each GAT model with equivalent weight. Four distinctive naïve ensemble-average models were constructed based on all possible combinations of GAT models to compare prediction performance.

4. Genomic marker and marker-by-marker interaction effect estimation

From the GAT models, the genomic marker effects were estimated using Integrated Gradients (Sundararajan et al., 2017) from the genomic markers as a part of the pipeline in EasiGP:

$$IG(x_i) \approx (x_i - x_i') * \sum_{k=1}^{m} \frac{\partial F\left(x_i' + \frac{k}{m} * (x_i - x_i')\right)}{\partial x_i} * \frac{1}{m} \tag{4}$$

where x is the attribute (SNP) value, x' is the baseline value of the attribute x and m is the total interpolation steps of the integral. Integrated Gradients measure the influence of each attribute using the cumulative values of integrals between the predicted values for individual RILs including and excluding the target attribute. Integrated Gradients calculates genomic marker effects element-wise, and thus the final genomic marker effect was calculated by summing the genomic marker effects of 50 randomly chosen individuals (RILs) from the test set. The mean values of genomic marker effects from each prediction scenario were calculated as the final genomic marker effects in the end.

For RF, the impurity-based importance of each genomic marker was used as the genomic marker effects. The pairwise Shapley score (Shapley, 1953; Lundberg and Lee, 2017) was calculated to estimate the genomic marker-by-marker interactions:

$$\Phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} \left[f_{S \cup \{i\}}\left(x_{S \cup \{i\}}\right) - f_s(x_s) \right] \tag{5}$$

where Φ_i is the Shapley score of the target attribute (SNP) i, S is a subset of the attributes (SNPs) F, $f_{S\cup\{i\}}$ is the predicted value from the prediction model with an attribute subset including i, f_s is the predicted value from the prediction model without the attribute i. Since the Shapley scores were also calculated element-wise, 30 individuals were randomly selected from the test sets and averaged throughout the prediction scenarios. Since the Shapley scores were calculated from possible combinations with other attributes, the scores can be decomposed at the pairwise level to calculate the pairwise Shapley scores to be leveraged as genomic marker-by-marker interaction effects.

5. Inference of the trait genetic architecture

The extracted genomic marker effects from the GAT models were projected to their corresponding genomic marker regions, accompanied by the selected genomic marker-by-marker interaction effects from RF. This trait genetic architecture was visualised in the form of circos plots (Krzywinski et al., 2009) as the final phase in EasiGP (Tomura et al., 2025a). The mean value of the genomic marker effects was leveraged to infer the genetic architecture from the naïve ensemble-average model. The captured genetic architectures were compared against the key gene regulators identified in previous studies: QTL information from Chen et al. (2019) for the TeoNAM dataset, Buckler et al. (2009) for the MaizeNAM dataset and the maize adaptation study in flowering time by Wisser et al. (2019).

6. Genomic prediction model evaluation

The prediction accuracy of the genomic prediction models was measured by Pearson correlation, measuring the accordance level of ranking between the observed and predicted phenotype values. A Pearson correlation of 1 indicates that both the predicted and observed phenotype values are completely accordant and hence perfectly accurate. The prediction error was measured by mean square error (MSE). An MSE of 0 indicates no difference between the observed and predicted values and thus no squared prediction error.

The Diversity Prediction Theorem (Page, 2018) was used to measure the level of diversity in predicted phenotypes from each GAT model in the naïve ensemble-average model:

$$(\bar{M} - V)^2 = \sum_{i=1}^{N} \frac{(M_i - V)^2}{N} - \sum_{i=1}^{N} \frac{(M_i - \bar{M})^2}{N} \quad (3)$$

where M_i is the set of predicted values for the RILs from the prediction model (a GAT model) i, $\bar{M}$ is the set of mean predicted RIL values from the i GAT models, V is the set of true RIL values and N is the total number of individual prediction models. In this study, observed phenotypes were used as V. The many-model ensemble error (first term) is calculated by subtracting the prediction diversity (third term) from the average error of the individual prediction models (second term). Hence, Equation (3) shows that the ensemble many-model error decreases as the prediction values from the individual prediction models become more diverse.

The prediction performance of the genomic prediction models was iteratively evaluated under multiple prediction scenarios. For the TeoNAM, data were uniquely split into training and test sets 500 times per population and trait. Three training-test set ratios (0.8-0.2, 0.65-0.35 and 0.5-0.5) were leveraged to investigate the effect of training set size. Hence, 7,500 prediction scenarios (5 populations * 3 ratios * 500 samplings) were generated per trait. For the MaizeNAM, data was sampled 50 times per population with the same three training-test set ratios. In total, 3,750 prediction scenarios (25 populations * 3 ratios * 50 samplings) were generated per trait.

Results

1. Prior knowledge GAT models did not always improve overall prediction performance

Our results were not consistent with our hypothesis that the prior knowledge model would always outperform the other models across traits and datasets (Figure 2). While the prior knowledge model reached the highest median prediction accuracy for DTA and the lowest median prediction error for both DTA and ASI in the MaizeNAM dataset within the GAT models, its outperformance was not consistent in other prediction scenarios and metrics. The extracted prior knowledge based on the Shapley scores generated for the RF graph models in this study did not always improve the prediction performance of the GAT models.

For DTA, the GAT model with the highest prediction performance was heavily dependent on the dataset (Table 1). In the TeoNAM dataset, the infinitesimal model reached the highest median prediction accuracy (Pearson correlation) and the lowest prediction error (MSE). In the MaizeNAM dataset, the prior

knowledge model reached the highest median prediction accuracy and lowest prediction error. The prediction performance of the prior knowledge model was the highest or the second highest throughout DTA.

The level of prediction performance improvement using prior knowledge was also dataset dependent in ASI (Table 1). In the TeoNAM dataset, the highest median prediction accuracy and lowest prediction error were achieved by the infinitesimal model within the GAT models. This outperformance of the infinitesimal model was also observed in prediction accuracy, while the prior knowledge model reached the lowest prediction error within the GAT models in the MaizeNAM dataset. These prediction results indicate that the use of prior knowledge did not consistently improve the prediction performance of the GAT models for ASI.

2. Non-infinitesimal models showed greater robustness to limited training data

Prediction performance deterioration by reducing the training set size was less acute for the non-infinitesimal models compared to the other GAT models (Figure 3). For the TeoNAM dataset, the performance reduction rate of the median prediction accuracy increased as the size of the training set reduced for the infinitesimal model. A similar trend was observed for the median prediction error in the TeoNAM dataset as the size of the training set became smaller. The infinitesimal models showed a higher rate of error increase. In the MaizeNAM dataset, a similar trend of rapid performance decrease was also observed as the training set size reduced for ASI. These results indicate that although the training set size generally reduced the prediction performance of the GAT models, explicitly specifying interactions between genomic markers reduced the decay of prediction performance when the training set size reduces for DTA.

3. Ensembles of GAT models improved prediction performance

The ensembles of GAT models consistently outperformed or were comparable to the best GAT model (Figure 2; Table 1). In the TeoNAM dataset, the median prediction performance of the naïve ensemble-average model was higher than the GAT model with the highest prediction performance for both DTA and ASI. Similarly, the outperformance of the naïve ensemble-average model was observed for both DTA and ASI in the MaizeNAM. These prediction performance results indicate that the naïve ensemble-average model can consistently improve prediction performance relative to the individual GAT models.

4. Diversity of predictive information associated with increased prediction performance

The increased diversity in the predicted phenotypes improved the prediction performance of the naïve ensemble-average model (Figure 4, Table S2). As the total number of GAT models included in the naïve ensemble average model increased, the ensemble error and MSE became lower and the Pearson correlation became higher in both DTA and ASI, especially for the TeoNAM dataset. There was a strong positive correlation in the ranking of the ensemble models with different GAT combinations between the ensemble error and both Pearson correlation and MSE (Table 2, S2). This positive association of the ranking between the metrics indicates that the addition of the diverse GAT models into the naïve ensemble-average model tends to amplify the prediction performance of the naïve ensemble-average model.

The diversity in the captured information from each GAT model was also observed at the genome level (Figure 5). The patterns of the colour intensity in each ring corresponding to the magnitude of genomic marker effects varied for each GAT model. Variation in the marker effects indicates that the GAT models captured features of the trait genetic architecture differently and the ensemble generated a unique view of the genetic architecture by integrating the distinctively estimated additive and non-additive genomic marker effects from the GAT models.

5. GAT models identified several key gene regulators

Highlighted genomic marker regions from the GAT models overlapped with several key gene regulators for DTA that had been identified in previous studies (Figure 5). For instance, the genomic marker region between 35 cM and 50 cM in chromosome 10 was identified across genomic prediction models for DTA in the TeoNAM dataset. This region is in close proximity to *ZmCCT10* (55 cM), known to be one of the most significant maize flowering time genes through its control of the photoperiod pathway (Dong et al., 2012; Chen et al., 2019; Wisser et al., 2019). This highlighted region in chromosome 10 showed heavy interactions with another region containing the gene *ZCN8* (70 cM in chromosome 8), captured by all the GAT models. *ZCN8* is another gene that controls the photoperiod pathway regulating the transition process from the vegetative to reproductive stage (Meng et al., 2011; Dong et al., 2012; Guo et al., 2018; Aiyesa et al., 2025). For the MaizeNAM dataset,

346 the highlighted genomic marker region overlapped with the gene region homologue of *Ghd7* (85 cM in
347 chromosome 1) in rice (Xue et al., 2008; Buckler et al., 2009). The GAT models also highlighted the genomic
348 marker region close to *ZCN8*, as observed in the TeoNAM dataset, and further recognised the marker region
349 close to another photoperiod pathway gene, *ZmCCT9* (55cM in chromosome 9). The photoperiod pathway
350 gene *ZmCCT9* regulates flowering time through responses to the changes in day length associated with levels
351 of latitude by controlling the expression of *ZCN8* (Huang et al., 2018). Such overlaps were frequently observed,
352 with known gene regulators reported in the literature aligning with markers assigned large effects in the GAT
353 genomic prediction models (Figure 5), indicating that the genomic prediction models included genomic
354 regions containing genes with an established influence on target traits for their prediction.
355

356 Discussion

357 1. More comprehensive prior knowledge may improve the prediction performance of GAT models

358 The primary objective of this study was to investigate the possibility of performance improvement in genomic
359 prediction, over the prediction performance achieved applying an infinitesimal model, through the integration
360 of gene networks as prior knowledge utilising GAT models. A lack of consistency in prediction performance
361 improvement has remained a critical issue despite the introduction of numerous prediction algorithms,
362 attributed to the effect of the No Free Lunch Theorem (Wolpert and Macready 1997). The integration of prior
363 knowledge was expected to overcome this problem by customising the genomic prediction models into
364 problem-specific ones.

365 Contrary to the expected outcome, there was still a lack of consistency in the overall prediction
366 performance improvement of the GAT models using prior knowledge (Figure 2). There are a number of
367 possible reasons. This inconsistent performance improvement using prior knowledge might have been
368 attributed to the potential imprecision in the captured genomic marker-by-marker interactions. Shapley scores
369 were estimated using SHAP (Lundberg and Lee, 2017) incorporated in EasiGP, approximating Shapley values
370 as "scores" instead of calculating the actual values to overcome the prohibitive computational time by
371 leveraging conditional expectations and a perturbation approach (Janzing et al., 2020; Molnar et al., 2020;
372 Contreras et al., 2024). A series of approximation formulas, such as kernel-based approaches, allows the

373 estimation of Shapley values in a computationally efficient manner (Lundberg and Lee, 2017). However, such
374 an approximation approach may calculate Shapley scores that may not be aligned with actual Shapley values,
375 especially when a number of irrelevant attributes were included in the datasets (Huang and Marques-Silva,
376 2024). The approximation formulas can be affected by the irrelevant attributes that can distort Shapley scores
377 as noise. The leveraged perturbation approach may also affect this imprecision by considering the attribute
378 value combinations that do not reflect real-world conditions (Contreras et al., 2024). Hence, Shapley scores
379 may not represent the actual influential size of each attribute (genomic marker). Inaccurate Shapley scores
380 may compromise the effectiveness of the supervision of the GAT models in this study. Alternative approaches
381 for computing Shapley scores are an area for further investigation indicated by this study.

382 The complexity of genomic marker-by-marker interactions might also have affected the
383 informativeness of the extracted prior knowledge. ASI is a secondary trait (Leng et al., 2022) calculated by
384 the difference between DTA and DTS and thus can be regulated by the gene network of both quantitative traits
385 and other crop development traits (Messina et al. 2019). Consequently, ASI might be influenced by the
386 combination of two or more distinct gene networks, thereby increasing the complexity of the genetic
387 architecture of the trait. The current inferred gene network likely did not capture such complex interactions
388 between genomic markers in ASI due to the lower prediction performance of RF used to calculate the Shapley
389 scores. Hence, the integrated prior knowledge may not have improved the prediction performance of GAT for
390 ASI regulated by potentially more complex gene networks.

391 To capture more informative prior knowledge, genomic marker-by-marker interactions can be inferred
392 as epistasis using conditional dependencies between the markers as an alternative strategy (Behrouzi and Wit,
393 2019). This approach only extracts genomic marker pairs that are directly conditionally dependent, removing
394 noncausal relationships created by any genomic marker pair that is conditionally independent but correlated
395 through another genomic marker that is highly correlated with the two. This alternative approach may help
396 capture more precise relationships among genomic markers. Exclusion of noncausal connections was not
397 explicitly implemented in the current Shapley score-based method. Another approach for estimating genomic
398 marker-by-marker interactions is BShap (Sundararajan and Najmi, 2020), which extends Shapley scores using
399 a baseline. The baseline returns a prediction value excluding the effect of a target attribute (genomic marker).

The contribution of the target attribute is calculated as the difference between the prediction values including and excluding the target attribute, as employed in Integrated Gradients. These potential approaches may vary in how they weight genomic marker-by-marker interaction effects, consequently generating diverse views of the trait genetic architecture. Thus, integration of all these approaches into an ensemble might enhance prediction performance by providing more comprehensive prior knowledge to assist the GAT models.

2. An ensemble of GAT models can consistently improve prediction performance

While the methods discussed above to improve the performance of the GAT models using prior knowledge, based on the Diversity Prediction Theorem expectations, they may still not outperform an ensemble-based approach. The prediction performance of the ensemble was consistently higher than or equivalent to the best GAT models. This consistent prediction performance increase by the naïve ensemble-average can contribute to improving prediction performance across many prediction scenarios of crop breeding programs. Such an approach is especially useful given the difficulty in selecting the optimal genomic prediction model when prediction performance is heavily scenario-dependent (Voss-Fels et al., 2019; Escamilla et al., 2025). Consequently, it has been a challenge for breeders to achieve high prediction performance across all of their prediction scenarios. Our results suggest that the risk of selecting an unsuitable genomic prediction model can be mitigated by an ensemble of GAT models, which consistently match or outperform the individual models.

Such performance improvement by the ensemble was likely attributed to diversity in the phenotypic predictions as elucidated in the Diversity Prediction Theorem (Page, 2018; Messina et al., 2025). When diversity increased, by adding more GAT models into the ensemble, the prediction performance of the ensemble improved with increases in prediction accuracy and reductions in error (Figure 4, Table S2). This positive association between diversity and prediction performance of the ensemble was also observed in other studies (Dietterich, 2000a; Melville et al., 2005; Kick and Washburn, 2023), showing a downward trend in prediction error as the number of individual prediction models included in the ensemble increased. The aggregation of diverse prediction models can offset prediction errors and weaknesses of each, improving the prediction performance of the ensemble approach (Kick and Washburn, 2023; Washburn et al., 2025). The diversity in the ensemble might support reaching the global optimum rather than convergence at the

suboptimal local optimum (Bourel et al., 2024). Complex prediction models, including GAT, aim to solve optimisation problems (reaching the highest possible prediction performance) using a training set in a non-convex way that contains numerous local optima. There is a risk that the prediction performance of such complex prediction models can be optimised locally, and the model parameters may not be further adjusted for performance improvement (Audhkhasi et al., 2012). GAT may stop a drastic adjustment of trainable weights once it starts reaching a convergence in the prediction performance improvement, but the global optima can be achieved by another set of weights that is far from the current local optima. This problem can also be described as overfitting, insufficiently generalising the predictive power (Karystinos et al., 2000). The stagnation at the local optima can be mitigated by combining diverse local optima through the ensemble approach (Dietterich, 2000b). Hence, the ensemble of the multiple GAT models improved prediction performance by approaching the global optima.

The performance improvement using diverse predictive information in the ensemble was also demonstrated at the genome level by each GAT model assigning a different effect size in each genomic marker (Figure 5). By combining the distinct genetic architecture from each GAT model through the ensemble, a more complete view of the trait genetic architecture was developed to enhance prediction performance (Cooper et al., 2025; Messina et al., 2025; Tomura et al., 2025b). The interpretable genomic prediction models enabled such investigation at the genome level in this study, which has not been well-investigated in the realm of genomic prediction for crop breeding. The significance of accurately capturing the genetic architecture was aligned with the conclusion of Morgante et al. (2018), showing that there was a prediction performance improvement when the true trait genetic architecture was well-captured in simulated data. In contrast, prediction performance remained low when the true genetic architecture was not clearly identified and heavy interactions between genomic markers (fully connected-like structure) were included. Hence, capturing a comprehensive view of the genetic architecture that closely reflects the true genetic architecture is a key step in improving the predictive performance of genomic prediction models.

3. Genomic marker-by-marker interactions can supplement information loss for small training sets

The prediction performance of the infinitesimal model was more reactive to the size of the training set. The reduction rate of prediction performance was larger than that of the fully connected or prior knowledge models when the size of the training set became smaller for DTA in the TeoNAM dataset and ASI in the MaizeNAM dataset (Figure 3). Maintaining prediction performance in the non-infinitesimal models might have been due to the simplicity of the provided graph structures in the GAT infinitesimal model. It might have been relatively easy for the infinitesimal model to learn key prediction patterns with a large training set due to the absence of explicit complex interactions between genomic markers in their graph structures. Consequently, the infinitesimal model outperformed the other models. However, the simplicity in the graph structure reduced the effectiveness of the infinitesimal models in learning key predictive patterns when the allocated training set was small. There is a tendency that a star-shaped (infinitesimal) graphical structure can cause overfitting as the training set becomes smaller (Bechler-Speicher et al., 2024). In contrast, the explicit marker-by-marker interactions might have supplemented the information loss derived from the diminishing size of the training set for the non-infinitesimal models. These explicit marker-by-marker interactions might have improved the prediction performance of the non-infinitesimal model when the training set size became smaller.

Among the non-infinitesimal models in this study, the integration of prior knowledge showed higher prediction performance as an overall trend. Unnecessary explicit interactions between genomic markers can introduce noise, thereby negatively affecting prediction performance. Selecting genomic marker-by-marker interactions using RF with Shapley scores might have excluded such unnecessary explicit interactions. Hence, prior knowledge can provide more accurate marker-by-marker interaction information into graphs, contributing to performance improvement especially when the training set is small (Schapire et al., 2002). In crop breeding programs, it can be a challenge to conduct large experiments based on a large number of individuals with their genotypic and phenotypic information recorded (Ibba et al., 2020; Sneller et al., 2021; Jines et al., 2025). Therefore, using prior knowledge information for the GAT models can be a preferable approach compared to the fully connected GAT models.

4. Interpretability of genomic prediction models can accelerate genetic gain

479 The interpretability of the GAT models and the ensemble approach can identify genomic marker regions that
480 are highly influential on target traits. Identifying preferred allele combinations in the highlighted genomic
481 marker regions using the interpretable GAT models helps breeders select individuals with desirable traits. The
482 accordance rate of their alleles in those genomic marker regions can be used as a selection criterion, in contrast
483 to considering all alleles across the genome. This potential selection criterion can be used simultaneously with
484 other criteria, such as the ranking from predicted phenotypes and genetic diversity. The combination of
485 multiple selection criteria can improve the selection performance of individuals with more desirable traits.
486 Pook et al. (2025) accelerated the genetic gain of simulated maize breeding programs by developing a formula
487 that combines several selection criteria from the view of short-term, long-term genetic gain and the level of
488 genetic diversity. Hence, the combination of diverse selection criteria can also positively affect the speed of
489 genetic gain.

490 The interpretable genomic prediction models also identified several new genomic marker regions. This
491 interpretability-oriented genomic prediction approach has been applied to infer the genetic architecture in
492 other fields such as detecting genes influencing body heights and schizophrenia in humans (Askland et al.,
493 2021; van Hilten et al., 2021; Hequet et al., 2025), muscle growth in livestock (Guo et al., 2024) and fungus
494 tolerance in ash trees (Doonan et al., 2025). Interpretable machine learning models have started to be leveraged
495 in crop breeding using interpretable RF and deep learning models (Novielli et al., 2024; Wang et al., 2025b),
496 but no other studies have collectively utilised interpretable genomic prediction models as an ensemble.
497 Although validation of new genomic marker regions identified through the ensemble approach is still required,
498 this strategy has the potential to help breeders focus on the identified gene regions rather than the
499 comprehensive investigation of the whole genome to suppress cost and time. It is encouraging that multiple
500 regions containing previously identified key genes involved in controlling flowering time in maize were
501 consistently identified across two divergent datasets, the TeoNAM and MaizeNAM, in the applications of
502 GAT and their ensembles.

503

504 Conclusion

In this study, the prediction performance was evaluated for three GAT models incorporating different G2P structures relevant to genomic prediction for crop breeding, and ensembles of GAT models. The prediction results from our study showed that while using the GAT models with prior knowledge did not always increase the prediction performance, the ensemble of multiple GAT models, including the one with prior knowledge, consistently improved prediction performance. Capturing a more complete view of the genetic architecture using the ensemble was likely a critical component for prediction performance improvement. Hence, the use of the ensemble approach in genomic prediction for crop breeding programs can improve selection accuracy, leading to the acceleration of genetic gain.

Data and code availability

The TeoNAM dataset (Chen et al., 2019) used in this study is located at https://datacommons.cyverse.org/browse/iplant/home/shared/panzea/genotypes/GBS/TeosinteNAM for the genotypes and https://gsajournals.figshare.com/articles/dataset/Supplemental_Material_for_Chen_et_al_2019/9250682 for the phenotypes. The MaizeNAM dataset (Buckler et al., 2009) used in this study is located at https://cbsusrv04.tc.cornell.edu/users/panzea/download.aspx?filegroupid=10 for the genotypes and phenotypes. The computational tool, EasiGP, is located at https://github.com/ShunichiroT/EasiGP.

Acknowledgments

We thank the National Computing Infrastructure (NCI) and the Research Computing Centre (RCC) at the University of Queensland for providing access to the High Performance Computing (HPC) machines.

Funding

This study was funded by the Australian Research Council through the support of the Australian Research Council Centre of Excellence for Plant Success in Nature and Agriculture (CE200100015).

Conflicts of interest

The authors declare no conflicts of interest.

References

Abate C, Decherchi S, Cavalli A. 2023. Graph neural networks for conditional de novo drug design. Wiley Interdisciplinary Reviews: Computational Molecular Science. 13:e1651. https://doi.org/10.1002/wcms.1651.

Aiyesa LV, Beissinger T, Scholten S, Link W, Zumbach B, Kaufmann D. 2025. Individual plant genetics reveal the control of local adaption in european maize landraces. BMC Biology. 23. https://doi.org/10.1186/s12915-025-02241-8.

Askland KD, Strong D, Wright MN, Moore JH. 2021. The translational machine: A novel machine-learning approach to illuminate complex genetic architectures. Genetic Epidemiology. 45:485–536. https://doi.org/10.1002/gepi.22383.

Audhkhasi K, Sethy A, Ramabhadran B, Narayanan SS. 2012. Creating ensemble of diverse maximum entropy models. pp. 4845–4848. IEEE. https://doi.org/10.1109/ICASSP.2012.6289004.

Azodi CB, Bolger E, McCarren A, Roantree M, de Los Campos G, Shiu SH. 2019. Benchmarking parametric and machine learning models for genomic prediction of complex traits. G3: Genes, Genomes, Genetics. 9:3691–3702. https://doi.org/10.1534/g3.119.400498.

Bechler-Speicher M, Amos I, Gilad-Bachrach R, Globerson A. 2024. Graph neural networks use graphs when they shouldn't. Forty-first International Conference on Machine Learning.

Behrouzi P, Wit EC. 2019. Detecting epistatic selection with partially observed genotype data by using copula graphical models. Journal of the Royal Statistical Society Series C: Applied Statistics. 68:141–160. https://doi.org/10.1111/rssc.12287.

Bhatti UA, Tang H, Wu G, Marjan S, Hussain A. 2023. Deep learning with graph convolutional networks: An overview and latest applications in computational intelligence. International Journal of Intelligent Systems. 2023:8342104. https://doi.org/10.1155/2023/8342104.

Bourel M, Cugliari J, Goude Y, Poggi JM. 2024. Boosting diversity in regression ensembles. Statistical Analysis and Data Mining: The ASA Data Science Journal. 17:e11654. https://doi.org/10.1002/sam.11654.

Breiman L. 2001. Random forests. Machine Learning. 45:5–32. https://doi.org/10.1023/A:1010933404324.

Brody S, Alon U, Yahav E. 2021. How attentive are graph attention networks? arXiv preprint arXiv:2105.14491. https://doi.org/10.48550/arXiv.2105.14491.

Buckler ES, Holland JB, Bradbury PJ, Acharya CB, Brown PJ, Browne C, Ersoz E, Flint-Garcia S, Garcia A, Glaubitz JC *et al*. 2009. The genetic architecture of maize flowering time. Science. 325:714–718. https://doi.org/10.1126/science.1174276.

Chen Q, Yang CJ, York AM, Xue W, Daskalska LL, DeValk CA, Krueger KW, Lawton SB, Spiegelberg BG, Schnell JM *et al*. 2019. Teonam: A nested association mapping population for domestication and agronomic trait analysis in maize. Genetics. 213:1065–1078. https://doi.org/10.1534/genetics.119.302594.

Clevert DA, Unterthiner T, Hochreiter S. 2015. Fast and accurate deep network learning by exponential linear units (elus). arXiv preprint arXiv:1511.07289. https://doi.org/10.48550/arXiv.1511.07289.

Contreras J, Winterfeld A, Popp J, Bocklitz T. 2024. Spectral zones-based shap/lime: Enhancing interpretability in spectral deep learning models through grouped feature analysis. Analytical Chemistry. 96:15588–15597. https://doi.org/10.1021/acs.analchem.4c02329.

Cooper M, Podlich DW, Smith OS. 2005. Gene-to-phenotype models and complex trait genetics. Australian Journal of Agricultural Research. 56:895–918. https://doi.org/10.1071/AR05154.

Cooper M, Powell O, Voss-Fels K, Messina C, Gho C, Podlich D, Technow F, Chapman S, Beveridge C, Ortiz-Barrientos D *et al*. 2021. Modelling selection response in plant-breeding programs using crop models as mechanistic gene-to-phenotype (cgm-g2p) multi-trait link functions. in silico Plants. 3:diaa016. https://doi.org/10.1093/insilicoplants/diaa016.

Cooper M, Tomura S, Wilkinson MJ, Powell O, Messina CD. 2025. Breeding perspectives on tackling trait genome-to-phenome (g2p) dimensionality using ensemble-based genomic prediction. Theoretical and Applied Genetics. https://doi.org/10.1007/s00122-025-04960-6.

Crossa J, Montesinos-Lopez OA, Costa-Neto G, Vitale P, Martini JW, Runcie D, Fritsche-Neto R, Montesinos-Lopez A, Pérez-Rodríguez P, Gerard G et al. 2025. Machine learning algorithms translate big data into predictive breeding accuracy. Trends in Plant Science. https://doi.org/10.1016/j.tplants.2024.09.011.

Dietterich TG. 2000a. Ensemble methods in machine learning. pp. 1–15. Springer. https://doi.org/10.1007/3-540-45014-91.

Dietterich TG. 2000b. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. Machine learning. 40:139–157. https://doi.org/10.1023/A:1007607513941.

Dong Z, Danilevskaya O, Abadie T, Messina C, Coles N, Cooper M. 2012. A gene regulatory network model for floral transition of the shoot apex in maize and its dynamic modeling. Plos One. https://doi.org/10.1371/journal.pone.0043450.

Doonan JM, Budde KB, Kosawang C, Lobo A, Verbylaite R, Brealey JC, Martin MD, Pliura A, Thomas K, Konrad H *et al*. 2025. Multiple, single trait gwas and supervised machine learning reveal the genetic architecture of fraxinus excelsior tolerance to ash dieback in europe. Plant, Cell & Environment. https://doi.org/10.1111/pce.15361.

Duan Z, Lee C, Zhang J *et al*. 2023. Exad-gnn: explainable graph neural network for alzheimer's disease state prediction from single-cell data. APSIPA Transactions on Signal and Information Processing. 12. http://dx.doi.org/10.1561/116.00000239.

Escamilla DM, Li D, Negus KL, Kappelmann KL, Kusmec A, Vanous AE, Schnable PS, Li X, Yu J. 2025. Genomic selection: Essence, applications, and prospects. The Plant Genome 18:e70053. https://doi.org/10.1002/tpg2.70053.

Guo L, Wang X, Zhao M, Huang C, Li C, Li D, Yang CJ, York AM, Xue W, Xu G *et al*. 2018. Stepwise cis-regulatory changes in zcn8 contribute to maize flowering-time adaptation. Current biology. 28:3005–3015. https://doi.org/10.1016/j.cub.2018.07.029.

Gabur I, Simioniuc DP, Snowdon RJ, Cristea D. 2022. Machine learning applied to the search for nonlinear features in breeding populations. Frontiers. https://doi.org/10.3389/frai.2022.876578.

Guo Y, Li S, Na R, Guo L, Huo C, Zhu L, Shi C, Na R, Gu M, Zhang W. 2024. Comparative transcriptome analysis of bovine, porcine, and sheep muscle using interpretable machine learning models. Animals: an Open Access Journal from MDPI. 14:2947. https://doi.org/10.3390/ani14202947.

Gupta A, Singh A. 2023. Agri-gnn: A novel genotypic-topological graph neural network framework built on graphsage for optimized yield prediction. arXiv preprint arXiv:2310.13037. https://doi.org/10.48550/arXiv.2310.13037.

Hequet CC, Gaggiotti O, Parachini S, Bochukova E, Ye J. 2025. Biologically-informed interpretable deep learning framework for phenotype prediction and gene interaction detection. bioRxiv. pp. 2025–04. https://doi.org/10.1101/2025.04.11.648325.

Hernández-Lorenzo L, Hoffmann M, Scheibling E, List M, Matías-Guiu JA, Ayala JL. 2022. On the limits of graph neural networks for the early diagnosis of alzheimer's disease. Scientific Reports. 12:17632. https://doi.org/10.1038/s41598-022-21491-y.

Huang C, Sun H, Xu D, Chen Q, Liang Y, Wang X, Xu G, Tian J, Wang C, Li D *et al*. 2018. Zmcct9 enhances maize adaptation to higher latitudes. Proceedings of the National Academy of Sciences. 115:E334–E341. https://doi.org/10.1073/pnas.1718058115.

Huang X, Marques-Silva J. 2024. On the failings of shapley values for explainability. International Journal of Approximate Reasoning. 171:109112. https://doi.org/10.1016/j.ijar.2023.109112.

Ibba MI, Crossa J, Montesinos-López OA, Montesinos-López A, Juliana P, Guzman C, Delorean E, Dreisigacker S, Poland J. 2020. Genome-based prediction of multiple wheat quality traits in multiple years. The Plant Genome. 13. https://doi.org/10.1002/tpg2.20034.

Janzing D, Minorics L, Blöbaum P. 2020. Feature relevance quantification in explainable ai: A causal problem. pp. 2907–2916. PMLR.

Jines M, Chandler M, Podlich D, Baumgarten A, Wright D, Jacobson A, Zhao H, Gogerty J, Regennitter M, Hung H-Y, McDowell R, Tang T, Diepenbrock C, Helmi H, Messina C, Ross A, Henke G, Spangler L, Caldwell M, Stirling L, Reyes A, Gho C, Cooper M, Arbuckle J, Smalley M, Milach S, Graham G, LR Totir. 2025. The Big BIT maize experiment: A large multi-location, multi-year, multi-tester, multi-population predictive breeding validation study. The Plant Genome 18:e70117. https://doi.org/10.1002/tpg2.70117.

Karystinos GN, Pados DA. 2000. On overfitting, generalization, and randomly expanded training sets. IEEE Transactions on Neural Networks. 11:1050–1057. https://doi.org/10.1109/72.870038.

Kick DR, Washburn JD. 2023. Ensemble of best linear unbiased predictor, machine learning, and deep learning models predict maize yield better than each model alone. in silico Plants. p. diad015. https://doi.org/10.1093/insilicoplants/diad015.

Kingma DP. 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980. https://doi.org/10.48550/arXiv.1412.6980.

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. 2009. Circos: an information aesthetic for comparative genomics. Genome research. 19:1639–1645. https://doi.org/10.1101/gr.092759.109.

Leng P, Khan SU, Zhang D, Zhou G, Zhang X, Zheng Y, Wang T, Zhao J. 2022. Linkage mapping reveals qtl for flowering time-related traits under multiple abiotic stress conditions in maize. International Journal of Molecular Sciences. 23:8410. https://doi.org/10.3390/ijms23158410.

Lundberg SM, Lee SI. 2017. A unified approach to interpreting model predictions. Advances in neural information processing systems. 30.

Maas AL, Hannun AY, Ng AY *et al*. 2013. Rectifier nonlinearities improve neural network acoustic models. volume 30. p. 3. Atlanta, GA.

Mascher M, Jayakodi M, Shim H, Stein N. 2024. Promises and challenges of crop translational genomics. Nature. 636:585–593. https://doi.org/10.1038/s41586-024-07713-5.

McCormick RF, Truong SK, Rotundo J, Gaspar AP, Kyle D, Van Eeuwijk F, Messina CD. 2021. Intercontinental prediction of soybean phenology via hybrid ensemble of knowledge-based and data-driven models. in silico Plants. 3:diab004. https://doi.org/10.1093/insilicoplants/diab004.

Melville P, Mooney RJ. 2005. Creating diversity in ensembles using artificial data. Information Fusion. 6:99–111. https://doi.org/10.1016/j.inffus.2004.04.001.

Meng X, Muszynski MG, Danilevskaya ON. 2011. The ft-like zcn8 gene functions as a floral activator and is involved in photoperiod sensitivity in maize. The Plant Cell. 23:942–960. https://doi.org/10.1105/tpc.110.081406.

Messina C, Garcia-Abadillo J, Powell O, Tomura S, Zare A, Ganapathysubramanian B, Cooper M. 2025. Toward a general framework for ai-enabled prediction in crop improvement. Theoretical and Applied Genetics. 138:1–15. https://doi.org/10.1007/s00122-025-04928-6.

Messina CD, Hammer GL, McLean G, Cooper M, van Oosterom EJ, Tardieu F, Chapman SC, Doherty A, Gho C (2019) On the dynamic determinants of reproductive failure under drought in maize. in silico Plants 1: diz003. https://doi.org/10.1093/insilicoplants/diz003.

Meuwissen TH, Hayes BJ, Goddard M. 2001. Prediction of total genetic value using genome-wide dense marker maps. genetics. 157:1819–1829. https://doi.org/10.1093/genetics/157.4.1819.

Molnar C, Casalicchio G, Bischl B. 2020. Interpretable machine learning–a brief history, state-of-the-art and challenges. pp. 417–431. Springer. https://doi.org/10.1007/978-3-030-65965-3$_2$8.

Morgante F, Huang W, Maltecca C, Mackay TF. 2018. Effect of genetic architecture on the prediction accuracy of quantitative traits in samples of unrelated individuals. Heredity. 120:500–514. https://doi.org/10.1038/s41437-017-0043-0.

Novielli P, Romano D, Pavan S, Losciale P, Stellacci AM, Diacono D, Bellotti R, Tangaro S. 2024. Explainable artificial intelligence for genotype-to-phenotype prediction in plant breeding: a case study with a dataset from an almond germplasm collection. Frontiers in Plant Science. 15:1434229. https://doi.org/10.3389/fpls.2024.1434229.

Page SE. 2018. *The model thinker: What you need to know to make data work for you*. Hachette UK.

Pook T, Hassanpour A, Niehoff TA, Calus MP. 2025. Strategies to improve selection compared to selection based on estimated breeding values. bioRxiv. pp. 2025–01. https://doi.org/10.1101/2025.01.09.632197.

Powell OM, Barbier F, Voss-Fels KP, Beveridge C, Cooper M. 2022. Investigations into the emergent properties of gene-to- phenotype networks across cycles of selection: a case study of shoot branching in plants. in silico Plants. 4:1–9. https://doi.org/10.1093/insilicoplants/diac006.

Scarselli F, Gori M, Tsoi AC, Hagenbuchner M, Monfardini G. 2008. The graph neural network model. IEEE transactions on neural networks. 20:61–80. https://doi.org/10.1109/TNN.2008.2005605.

Schapire RE, Rochery M, Rahim M, Gupta N. 2002. Incorporating prior knowledge into boosting. volume 2. pp. 538–545.

Sneller C, Ignacio C, Ward B, Rutkoski J, Mohammad M. 2021. Using genomic selection to leverage resources among breeding programs: Consortium-based breeding. Agronomy. 11. https://doi.org/10.3390/agronomy11081555.

Sundararajan M, Najmi A. 2020. The many shapley values for model explanation. pp. 9269–9278. PMLR.

Sundararajan M, Taly A, Yan Q. 2017. Axiomatic attribution for deep networks. pp. 3319–3328. PMLR.

Tomura S, Wilkinson MJ, Powell O, Cooper M. 2025a. Ensemble AnalySis with Interpretable Genomic Prediction (EasiGP): Computational Tool for Interpreting Ensembles of Genomic Prediction Models. The Plant Genome. https://doi.org/10.1002/tpg2.70138.

Tomura S, Wilkinson MJ, CooperM, Powell O. 2025b. Improved genomic prediction performance with ensembles of diverse models. G3: Genes, Genomes, Genetics. https://doi.org/10.1093/g3journal/jkaf048.

van Hilten A, Kushner SA, Kayser M, Ikram MA, Adams HH, Klaver CC, Niessen WJ, Roshchupkin GV. 2021. Gennet framework: interpretable deep learning for predicting phenotypes from genetic data. Communications biology. 4:1094. https://doi.org/10.1038/s42003-021-02622-z.

Vatter J, Mayer R, Jacobsen HA. 2023. The evolution of distributed systems for graph neural networks and their origin in graph processing and deep learning: A survey. ACM Computing Surveys. 56:1–37. https://doi.org/10.1145/3597428.

Veličković P, Cucurull G, Casanova A, Romero A, Liò P, Bengio Y. 2018. Graph attention networks. arXiv:1710.10903. https://doi.org/10.48550/arXiv.1710.10903.

Von Rueden L, Mayer S, Beckh K, Georgiev B, Giesselbach S, Heese R, Kirsch B, Pfrommer J, Pick A, Ramamurthy R *et al*. 2021. Informed machine learning–a taxonomy and survey of integrating prior knowledge into learning systems. IEEE Transactions on Knowledge and Data Engineering. 35:614–633. https://doi.org/10.1109/TKDE.2021.3079836.

Voss-Fels KP, Cooper M, Hayes BJ. 2019. Accelerating crop genetic gains with genomic selection. Theoretical and Applied Genetics. 132:669–686. https://doi.org/10.1007/s00122-018-3270-8.

Wang C, Kumar GA, Rajapakse JC. 2025a. Drug discovery and mechanism prediction with explainable graph neural networks. Scientific Reports. 15:179. https://doi.org/10.1038/s41598-024-83090-3.

Wang D, Zhang M. 2021. Artificial intelligence in optical communications: from machine learning to deep learning. Frontiers in Communications and Networks. 2:656786. https://doi.org/10.3389/frcmn.2021.656786.

Wang H, Yan S, Wang W, Chen Y, Hong J, He Q, Diao X, Lin Y, Chen Y, Cao Y *et al*. 2025b. Cropformer: An interpretable deep learning framework for crop genomic prediction. Plant Communications. 6. https://doi.org/10.1016/j.xplc.2024.101223.

Wang Z, Liang S, Liu S, Meng Z, Wang J, Liang S. 2023. Sequence pre-training-based graph neural network for predicting lncrna-mirna associations. Briefings in Bioinformatics. 24:bbad317. https://doi.org/10.1093/bib/bbad317.

Washburn JD, Varela JI, Xavier A, Chen Q, Ertl D, Gage JL, Holland JB, Lima DC, Romay MC, Lopez-Cruz M *et al*. 2025. Global genotype by environment prediction competition reveals that diverse modeling strategies can deliver satisfactory maize yield estimates. Genetics. 229:iyae195. https://doi.org/10.1093/genetics/iyae195.

Wisser RJ, Fang Z, Holland JB, Teixeira JE, Dougherty J, Weldekidan T, de Leon N, Flint-Garcia S, Lauter N, Murray SC *et al*. 2019. The genomic basis for short-term evolution of environmental adaptation in maize. Genetics. 213:1479–1494. https://doi.org/10.1534/genetics.119.302780.

Wolpert DH, Macready WG. 1997. No free lunch theorems for optimization. IEEE transactions on evolutionary computation. 1:67–82. https://doi.org/10.1109/4235.585893.

Xue W, Xing Y, Weng X, Zhao Y, Tang W, Wang L, Zhou H, Yu S, Xu C, Li X *et al*. 2008. Natural variation in ghd7 is an important regulator of heading date and yield potential in rice. Nature genetics. 40:761–767. https://doi.org/10.1038/ng.143.

Zhang Q, Li B, Zhang Y, Wang S. 2022. Suitability evaluation of crop variety via graph neural network. Computational Intelligence and Neuroscience. 2022:5614974. https://doi.org/10.1155/2022/5614974.

Zhao L, Tang P, Luo J, Liu J, Peng X, Shen M, Wang C, Zhao J, Zhou D, Fan Z *et al*. 2025. Genomic prediction with netgp based on gene network and multi-omics data in plants. Plant Biotechnology Journal. https://doi.org/10.1111/pbi.14577.

Tables and Figures

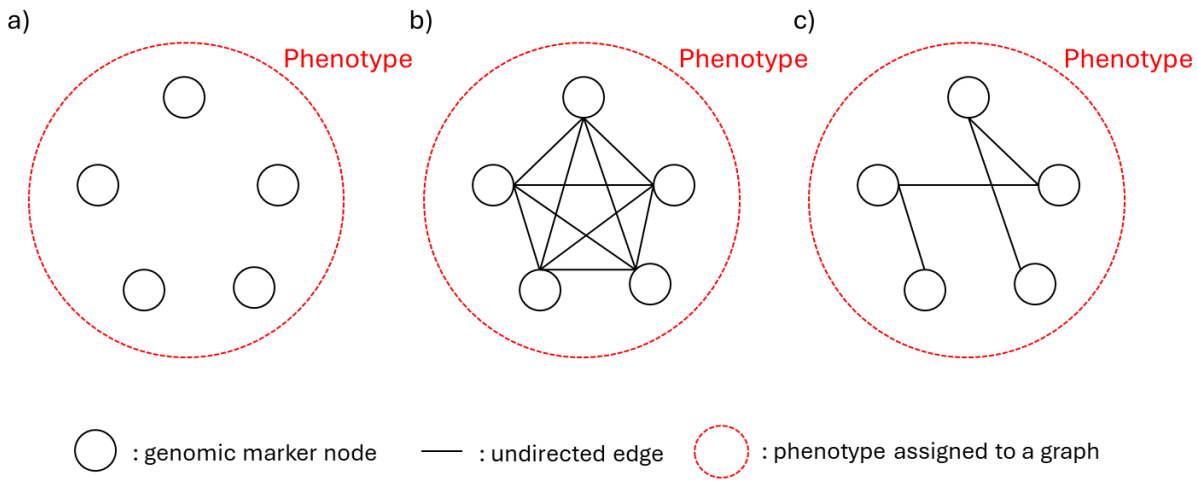

Figure 1: Three graph attention network (GAT) models that leverage possible genotype-to-phenotype (G2P) relationships: a) infinitesimal model, b) fully connected model and c) prior knowledge model. The infinitesimal model assumes no explicit connections between genomic marker nodes. The fully connected model allows all genomic marker nodes to be connected to others. The prior knowledge model only allows certain connections between genomic marker nodes based on the inferred trait genetic architecture.

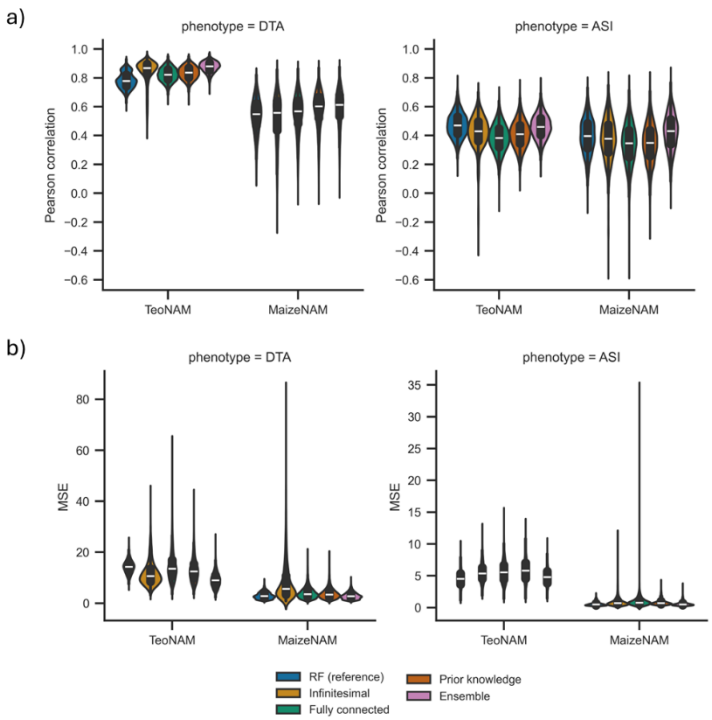

Figure 2: The comparison of prediction performance for infinitesimal, fully connected, prior knowledge and the naïve ensemble-average model (Ensemble) for days to anthesis (DTA) and anthesis to silking interval (ASI) traits in the TeoNAM and MaizeNAM datasets. Random forest was implemented as a reference (RF (reference)) used to infer the genetic architecture as prior knowledge for the Prior knowledge model. The prediction performance was measured in a) Pearson correlation for prediction correlation and b) mean square error (MSE) for prediction error. For each trait, the prediction performance was measured in 7,500 prediction scenarios for the TeoNAM dataset and 3,750 prediction scenarios for the MaizeNAM dataset. The prediction scenarios were generated by the combination of the three training-test ratios (0.8-0.2, 0.65-0.35 and 0.5-0.5), populations (5 and 25 for the TeoNAM and MaizeNAM datasets, respectively) and the total number of random dataset splits (500 and 50 for the TeoNAM and MaizeNAM datasets, respectively). The violin width represents the distribution of performance metrics. The white horizontal lines on the black box plots indicate the median value for each metric. The whiskers extend 1.5 times the interquartile range.

Table 1: The prediction performance of each GAT model measured in the median prediction accuracy (Pearson correlation) and median prediction error (mean square error (MSE)) for the days to anthesis (DTA) and anthesis to silking interval (ASI) in the TeoNAM and MaizeNAM datasets. The highlighted values in black indicate the highest prediction performance in each scenario.

Metric	Dataset	Trait	Infinitesimal	Fully connected	Prior knowledge	Ensemble
Pearson correlation	TeoNAM	DTA	0.845	0.822	0.834	**0.876**
		ASI	0.421	0.383	0.408	**0.461**
	MaizeNAM	DTA	0.523	0.551	0.581	**0.594**
		ASI	0.372	0.338	0.346	**0.425**
MSE	TeoNAM	DTA	11.667	14.494	13.038	**9.151**
		ASI	5.477	5.625	5.848	**4.815**
	MaizeNAM	DTA	8.548	4.031	3.794	**2.900**
		ASI	0.912	1.001	0.841	**0.609**

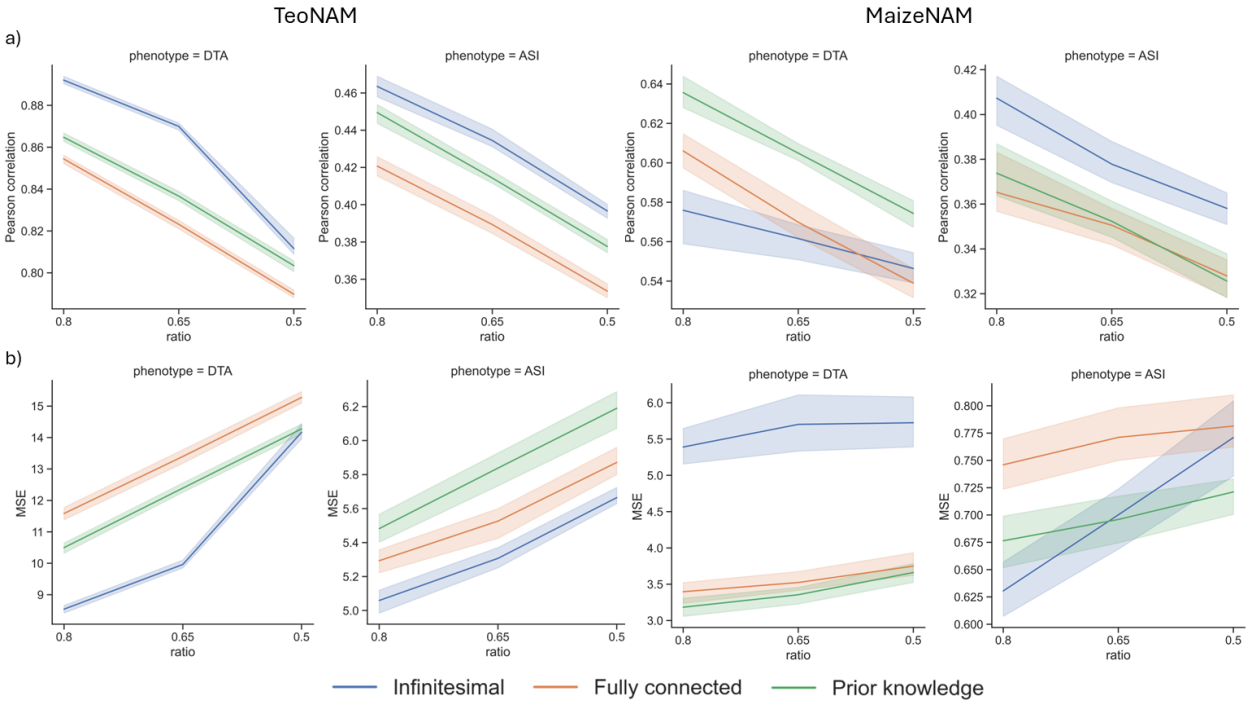

Figure 3: Prediction performance transition of the three GAT models (infinitesimal, fully connected and prior knowledge) when the ratio of the training set became smaller (0.8, 0.65 and 0.5), measured by a) Pearson correlation and b) mean square error (MSE) for the TeoNAM and MaizeNAM datasets. The upper and lower boundaries accompanying each line show the metric value range with standard error based on 2,500 and 1,250 prediction scenarios in each ratio for the TeoNAM and MaizeNAM datasets, respectively.

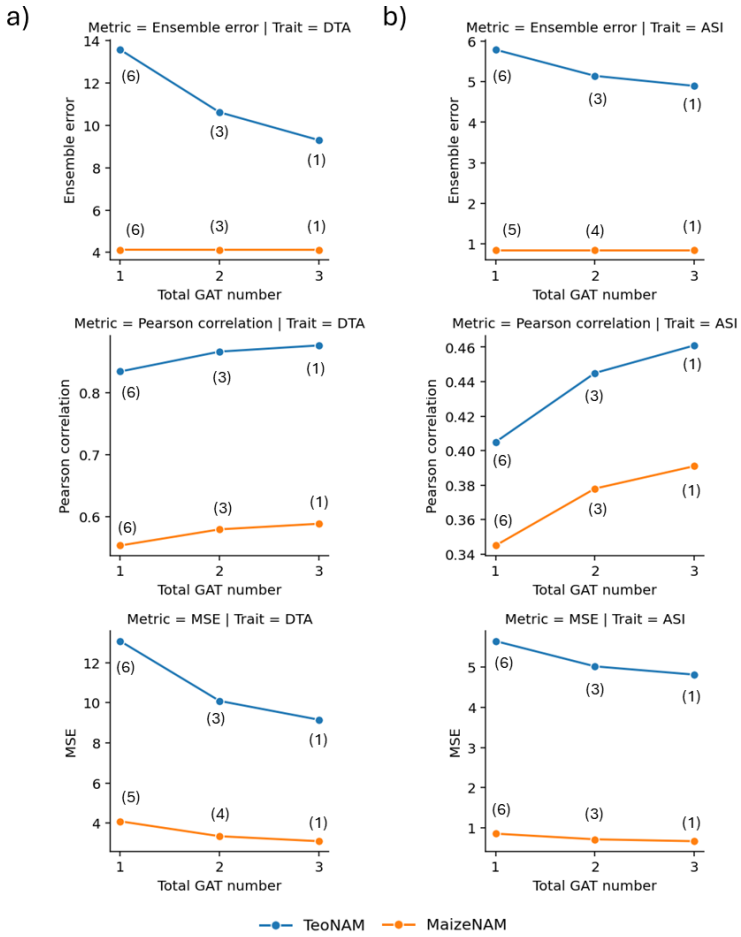

Figure 4: The metric transition of the naïve ensemble-average models with differences in the total number of included GAT models for a) the days to anthesis (DTA) and b) anthesis to silking interval (ASI) traits in both TeoNAM and MaizeNAM datasets. The mean value of ensemble error from the Diversity Prediction Theorem, Pearson correlation and mean squared error (MSE) were used as metrics. The total number of naïve ensemble-average models was 7, considering all the possible combinations of the GAT models that include single GAT models. Number 1, 2 and 3 of the x-axis represent the total number of the GAT models included in the naïve ensemble-average model. The values in brackets for each dot represent the average rank of each naïve ensemble-average model in each metric. Smaller rank values indicate a high preference in the ranking.

Table 2: Spearman correlation between the ranking of the naïve ensemble-average models in the ensemble error (first term) of the Diversity Prediction Theorem and the prediction performance metrics (Pearson correlation and mean squared error (MSE)) for the days to anthesis (DTA) and anthesis to silking interval (ASI) traits in the TeoNAM and MaizeNAM datasets

Trait	Metric	Spearman correlation	
		TeoNAM	MaizeNAM
DTA	Pearson correlation	1.000	0.929
	MSE	1.000	0.643
ASI	Pearson correlation	0.929	0.429
	MSE	0.929	0.929

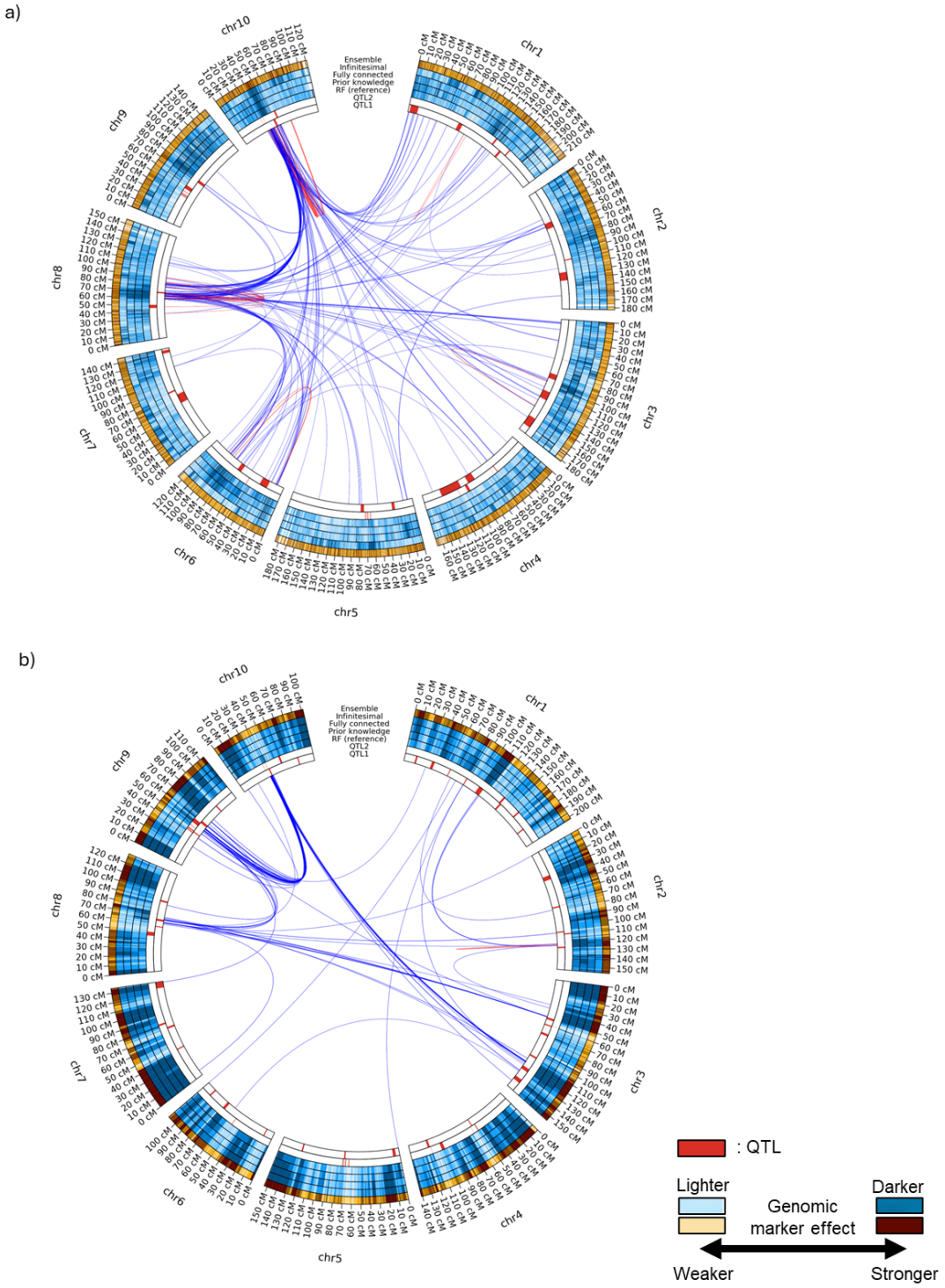

Figure 5: Circos plots for the days to anthesis (DTA) trait for the a) TeoNAM and b) MaizeNAM datasets. The innermost ring (QTR1) represents QTL regions in red identified by Chen et al. (2019) for the TeoNAM dataset and Buckler et al. (2009) for the MaizeNAM dataset, respectively. The second innermost ring (QTL2) represents the QTL regions identified by Wisser et al. (2019). The subsequent four blue rings represent the genomic marker effects from random forest as reference (RF (reference)) used to extract prior knowledge, the prior knowledge model, the fully connected model and the infinitesimal model. The outermost ring in orange shows the estimated genomic marker effects of the naïve ensemble-average model. The intensity of the colours shows the strength of the genomic marker effects represented in ten levels based on the quantiles. The colour becomes darker as the genomic marker effects become higher. Links between two genomic marker regions represent genomic marker-by-marker interactions detected by RF using the Shapley scores. The genomic marker pairs with the Shapley scores included in the top 0.01% were selected for both datasets. Thicker lines show stronger marker-by-marker interactions. The red and blue lines represent genomic marker-by-marker interactions within and between chromosomes, respectively.

Supplementary Material

Table S1: The number of genomic markers (SNPs) at the original and after linkage disequilibrium (LD) filtering and records from recombinant inbred lines (RILs) in the TeoNAM and MaizeNAM datasets. Each subpopulation contains a different total number of SNPs in the TeoNAM dataset. The total number of SNPs after LD filtering depends on the combination of records included in the training set in each prediction scenario. The total number of records varies depending on the subpopulation for both datasets.

Dataset	SNPs		Total records
	Original	After LD filtering	
TeoNAM	11,396 - 16,110	268 - 342	438 - 616
MaizeNAM	1,106	170 - 251	126 - 196

Table S2: The evaluation of the naïve ensemble-average models with differences in the total number of included GAT models for the days to anthesis (DTA) and anthesis to silking interval (ASI) traits in both TeoNAM and MaizeNAM datasets, measured in the ensemble error from the Diversity Prediction Theorem, Pearson correlation and mean squared error (MSE). The total number of naïve ensemble-average models was 7, considering all the possible combinations of the GAT models that include single GAT models. Number 1, 2 and 3 of the columns represent the number of the GAT models included in the naïve ensemble-average model. The sign "±" indicates standard error. The values in brackets represent the average rank and the smaller rank value indicates a high preference in the ranking.

Metric	Dataset	DTA			ASI		
		1	2	3	1	2	3
Ensemble error	TeoNAM	13.578±0.013 (6)	10.620±0.010 (3)	9.304±0.017 (1)	5.789±0.006 (6)	5.144±0.006 (3)	4.897±0.010 (1)
	MaizeNAM	4.125±0.002 (6)	4.122±0.002 (3)	4.121±0.002 (1)	0.858±0.008 (5)	0.858±0.006 (4)	0.858±0.010 (1)
Pearson correlation	TeoNAM	0.834±0.000 (6)	0.866±0.000 (3)	0.876±0.000 (1)	0.405±0.034 (6)	0.445±0.022 (3)	0.461±0.032 (1)
	MaizeNAM	0.554±0.001 (6)	0.580±0.001 (3)	0.589±0.002 (1)	0.345±0.023 (6)	0.378±0.016 (3)	0.391±0.007 (1)
MSE	TeoNAM	13.066±0.001 (6)	10.087±0.001 (3)	9.151±0.001 (1)	5.650±0.001 (6)	5.021±0.001 (3)	4.815±0.001 (1)
	MaizeNAM	4.092±0.001 (5)	3.350±0.001 (4)	3.102±0.002 (1)	0.853±0.007 (6)	0.712±0.004 (3)	0.665±0.006 (1)