Knowledge-driven approaches to improve genomic prediction in plants

Muhammad Farooq

Knowledge-driven approaches to improve genomic prediction in plants

Muhammad Farooq

Propositions

1. Prioritisation of DNA polymorphisms must be routine practice for genomic prediction.
   (this thesis)

2. The curse of dimensionality due to big data can be dealt with by more data.
   (this thesis)

3. The journey from the first to the final version of source code is analogous to the structure of the tree of life.

4. Although diversity is the spice of life, it is the spice of pain and suffering for data scientists.

5. Science begins with curiosity and ends in business.

6. Developing countries must progress by scientific collaboration instead of by reverse engineering.

Propositions belonging to the thesis, entitled

Knowledge-driven approaches to improve genomic prediction in plants

Muhammad Farooq
Wageningen, 20 September 2023

Knowledge-driven approaches to improve genomic prediction in plants

Muhammad Farooq

Thesis committee

Promotors
Prof. Dr Dick de Ridder
Professor of Bioinformatics
Wageningen University & Research

Prof. Dr Shahid Mansoor
Professor of Genomics
National Institute for Biotechnology and Genetic Engineering (NIBGE), Faisalabad, Pakistan

Co-promotors
Dr Aalt-Jan van Dijk
Associate Professor, Bioinformatics
Wageningen University & Research

Dr Harm Nijveen
Researcher, Bioinformatics
Wageningen University & Research

Other members
Prof. Dr R.F. Veerkamp, Wageningen University & Research
Prof. Dr M. Suarez Diez, Wageningen University & Research
Dr C.A. Maliepaard, Wageningen University & Research
Dr K. van Hulzen, Genetwister Technologies B.V. Wageningen

This research was conducted under the auspices of the Graduate School Experimental Plant Sciences (EPS)

Knowledge-driven approaches to improve genomic prediction in plants

Muhammad Farooq

Thesis

submitted in fulfilment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus,
Prof. Dr A.P.J. Mol,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Wednesday 20 September 2023
at 11 a.m. in the Omnia Auditorium.

Muhammad Farooq
Knowledge-driven approaches to improve genomic prediction in plants,
171 pages.

PhD thesis, Wageningen University, Wageningen, The Netherlands (2023)
With references, with summary in English

ISBN: 978-94-6447-794-8
DOI: https://doi.org/10.18174/634641

Contents

CHAPTER

1

Introduction

1.1 Evolution of plant breeding: from observations to a data-driven approach

Plant breeding is the science of developing improved plant cultivars with desired traits mainly related to food security and consumer requirements. The so-called green revolution in the mid-20th century proposed a scientific approach to increase food production through genetic improvement and better management practices (Khush 2001). However, the human population is expected to increase up to ~9 billion by 2050, with an annual growth rate of ~1.3%. Global food production needs a ~100% increase of the current rate to maintain the food-population balance (Royal Society of London 2009, Godfray, Beddington et al. 2010). In this century, more complex challenges, including climate change and shifting consumer preferences exacerbate the urgency to produce improved varieties adapted to these requirements at a much higher pace. With recent stagnating global crop yield growth, the relatively straightforward plant breeding approaches developed in the past need to be upgraded to improve more complex and economically valuable traits at the desired speed (Pingali 2012).

The breeding process generally consists of two phases: induction or recruitment of genetic variation for the trait(s) of interest, followed by (recurrent) selection of genotypes with favourable trait values. Genetic variation is induced in plant populations through different approaches, including crossing, mutagenesis etc. Alternatively, a large reservoir of diverse germplasm might be readily available in the natural genetic pool, from which one can directly recruit parental germplasms for selection. Subsequently, the “crossing - evaluation - selection” cycles are repeated until the breeding objectives are met. In any case, the selection process is a key step for developing better genotypes (NEI 1960), hence minor improvements in selection methodology might significantly increase the efficiency of the breeding process.

Historically, since the origin of agriculture until the first hybridisation experiment by Kölreuter in the 1760s (Roberts 1929), plant breeding was carried out by selection through observation. This resulted in in-field cultivation of wild variants and domestication of crops into landraces, and subsequently in mass selection of cultivars. With the discovery of the laws of heredity in the 19th-20th century, breeding turned to controlled mating, which led to the development of targeted approaches like pedigree, ideotype, population and hybrid breeding (Breseghello and Coelho 2013). The pivotal concept was that the observed phenotypic variation could be explained by genetic variation at specific chromosomal segments (Morgan 1911, Acquaah 2012).

With technological advancements in high-throughput DNA sequencing, low-cost genotyping of molecular markers, especially single nucleotide polymorphisms (SNPs), became feasible. This led to the era of modern plant breeding (Lamichhane and Thapa 2022), relying on the identification of genomic variants (markers) at the DNA nucleotide level associated with the observed phenotypic variation. For quantitative phenotypes, the strength of associations of a set of markers can be determined using statistical models,

resulting in a quantitative trait locus (QTL) map. A QTL is an individual genomic locus that is significantly associated with the trait of interest. The power with which QTLs can be detected depends on the linkage disequilibrium (LD) between markers and the QTLs. Therefore, associations are often measured in mapping populations derived from a few parental lines, that have limited recombination events. The genetic variation at a QTL can then be traced through the breeding process to select for the desired trait, referred to as marker-assisted selection (MAS). A limitation to this approach is the modest genetic diversity caused by using only a few parental lines in the mapping population (Scott, Ladejobi et al. 2020). Moreover, the size of QTL regions is often too large, resulting into imprecise localisation of genetic variations. To explore an increased volume of genetic variations at higher resolution, large genetically diverse germplasms, e.g. natural populations or diversity panels, are used with dense SNP genotyping in genome-wide association studies (GWAS), though their detection power is limited by the smaller LD blocks in these populations compared to the mapping populations.

A practical limitation to both QTL mapping and GWAS is the mandatory QTL identification step prior to selection. This implies that the QTLs need to explain large genetic effects to be able to be detected easily by the statistical model. On the other hand, small effects, which are characteristic for the complex traits, are usually hard to identify, especially when testing fairly large numbers of markers (p) with small numbers of observations (n). The non-additive allele effects (epistasis, dominance) add another layer of complexity for identifying all underlying genetic variance. This comes with the problem known as “missing heritability”, which implies that the sum of variances explained by all of the identified QTLs is less than the total genetic variance for a phenotype (Brachi, Morris et al. 2011, Makowsky, Pajewski et al. 2011). This might be due to limited statistical power to detect small effects in $n << p$ situations, allowing identification of only a subset of QTLs. Moreover, common present confounding factors like population structure and incorrect statistical assumptions for the genetic architecture might also limit QTLs identification (Ehsani, Janss et al. 2016). As a result, QTL/GWAS strategies are not very effective for breeding of complex polygenic traits, governed by many small effect QTLs, along with allelic interactions (Wray, Yang et al. 2013).

1.1.1 Genomic selection: promises and challenges

Instead of first testing associating subsets of SNPs to a phenotype, genome-wide SNPs can be used to estimate their cumulative effect for the phenotypes, omitting the QTL identification step. The use of dense genotyping makes it likely that at least one SNP would be in LD with each QTL, and the use of all SNPs together without regard for significance of their individual association enables small effects to be taken into account. This approach was first introduced by Meuwissen, Hayes et al. (2001) to predict total additive genomic values, called breeding values, using a shrinkage-based regression model. These breeding values can be used as pseudo-phenotypes for selection in the

breeding process, possibly much earlier than the actual phenotype is available – a process known as Genomic Selection (GS) (*Figure 1.1*) (Tessema, Liu et al. 2020, Krishnappa, Savadi et al. 2021).

In the breeding process, progress is usually quantified in terms of genetic gain, i.e. the annual improvement in breeding values. The goal is to increase the rate of genetic improvement over the "crossing - evaluation - selection" cycles, while upholding the genetic diversity, quantified in terms of genetic gain. The expected genetic gain, i.e. the change in average breeding value per unit time, can be calculated from the response to selection in one selection cycle using the parameters from the breeder's equation (Lush 1937, Falconer and Mackay 1996, Walsh 2004, Xu, Li et al. 2017):

$$\Delta G = \frac{ir\sigma_a}{L} \qquad (1.1)$$

Here, ΔG is the annual rate of genetic gain, *r* is the selection response accuracy, measured as correlation between true and genomic estimated breeding values (gEBVs) in GS, *i* is the selection intensity i.e. the mean deviation of selected genotypes in units of phenotypic standard deviation, σ_a is the standard deviation of breeding values, and *L* is duration of the breeding cycle measured in years. The GS framework can increase the genetic gain by improving components of equation (1.1): increasing scale and precision of genotyping and phenotyping to improve selection intensity, viz. intensive genotyping / phenotyping, results in a larger likelihood of selecting the best genotypes as the next parents; improving estimation precision of parameters , i.e. *r* and σ_a, due to the use of SNP markers instead of phenotypic measurements; and reducing the generation interval due to early selection based on breeding values instead of realized phenotypes.

Central to the GS framework is the precise estimation of genetic parameters i.e. σ_a^2, h^2 etc, using a supervised prediction model, called genomic prediction (GP). Major factors affecting the performance of GP include total number and quality of measured traits, genotyping density and allele frequencies, train/test populations design, i.e. sizes, structure and relatedness, genetic architecture of the trait and algorithmic characteristics of the model (Berro, Lado et al. 2019). A thorough understanding of the impact of these factors on performance of different GP methods is critical to yield successful GS.

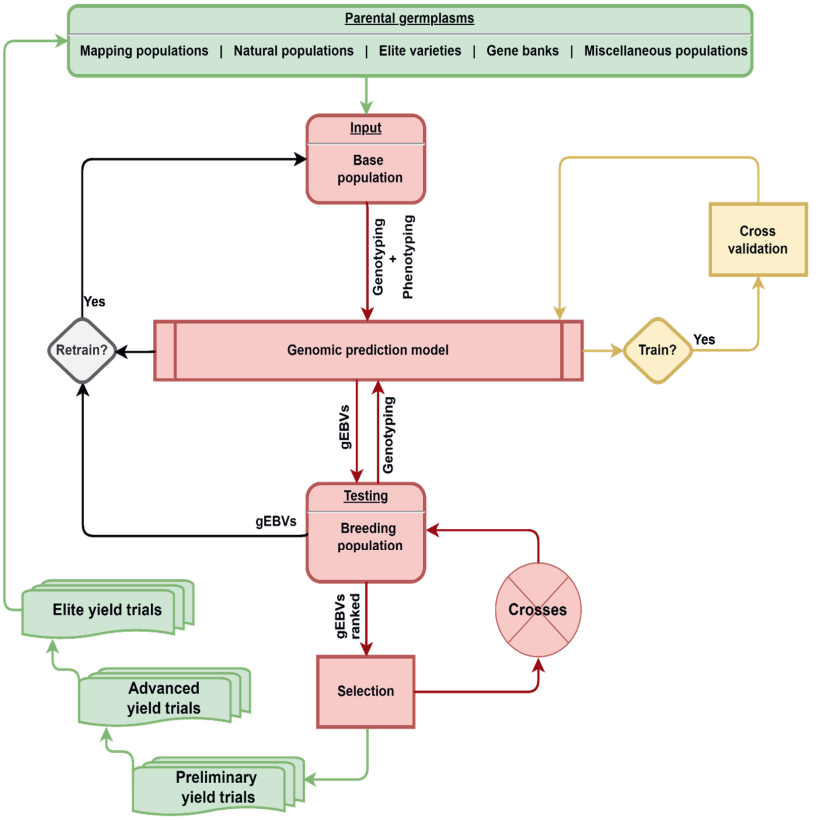


Figure 1.1: Genomic Selection (GS) based plant breeding workflow.

A typical workflow for a genomic selection-based plant breeding program. The training population can be taken from any of the parental germplasms and can be updated based on a breeding population at any breeding cycle, as well as, from preliminary yield trials, advanced yield trials or elite yield trials. The genomic prediction (GP) model is optimised using cross-validation (CV) of the training population and prediction accuracy is calculated based on the correlation between true phenotype and the predicted genomic estimated breeding value (gEBV) on the validation population. The breeding cycle consists of repeated selection, crossing and genotyping of the breeding populations, where selection is based on gEBV provided by the GP model.

1.2 Genomic prediction: past developments and current trends

Statistical models have been the first choice for many GP applications. These include linear mixed effect models (LMMs), where SNP effects are assumed to be random, with different assumptions about their effects distributions. Other linear regression methods employing shrinkage and treating SNP effects as fixed have also been used. These include ridge, lasso, elastic net regression and their extensions (Ogutu, Schulz-Streeck et al. 2012). The solutions to the LMMs are usually found as Best Linear Unbiased Estimates (BLUEs) and Best Linear Unbiased Predictions (BLUPs) (Henderson 1975,

Henderson 1984, Henderson 1985) for the fixed and random effects, respectively. The genomic BLUP (GBLUP) is a popular method that estimates breeding values using population kinship inferred from SNPs and is extensively applied due to its simplifying assumptions, decent performance and low time complexity (Hayes, Bowman et al. 2009). Mainly, SNP effects are modelled as additively acting allele effects and assumed to follow a normal distribution with equal variance; therefore, only one variance parameter needs to be estimated. However, multiple extensions have been proposed to accommodate unequal variances or weighted SNP effects (Zhang, Liu et al. 2010, Zhang, Ding et al. 2011, Tiezzi and Maltecca 2015). A Bayesian framework can also be used to accommodate both equal and unequal SNP variances, using parametric priors. Bayesian methods with unequal SNP variances have shown superior performance over GBLUP in many studies (Wolc, Arango et al. 2016), and many Bayesian model variants have been proposed, that differ in terms of their priors specifications. These include BayesA, BayesB, BayesCπ, BayesDπ and BayesR etc (Meuwissen, Hayes et al. 2001, Habier, Fernando et al. 2011, Gianola 2013).

In a rather generalised notation, a univariate LMM for genomic prediction (GP) can be written as:

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Z_d u_d} + \mathbf{Z_g u_g} + \boldsymbol{\varepsilon} \qquad (1.2)$$

Here, $\mathbf{y}$ is a vector of phenotypic measurements; $\boldsymbol{\beta}$ is a vector of fixed effects, including an estimate of the overall population mean; $\mathbf{u_d}, \mathbf{u_g}, \boldsymbol{\varepsilon}$ are vectors of non-genetic, genetic and residual effects, respectively; and $\mathbf{X}, \mathbf{Z_d}$ and $\mathbf{Z_g}$ are their design matrices. SNP effects can then be assumed to have a uniformly distributed or a flat prior and estimated as fixed effect, as in the case of ridge or lasso regression (see 1.4.1); or considered to be distributed normally and predicted using a frequentist statistical paradigm, e.g. a Restricted Maximum Likelihood Estimator (REML), or in a Bayesian framework e.g. by a Gibbs sampler.

Instead of parametric methods, semi-parametric methods like the Gaussian mixture model and Reproducible Kernel Hilbert Space (RKHS) regression can been used to account for more complex relations between SNPs and phenotypes (De los Campos, Gianola et al. 2010). Moreover, non-parametric machine learning (ML) approaches have gained popularity due to their ability to learn from the data. These include support vector machines (SVMs), bagged or boosted decision trees and their ensembles, e.g. random forest (RF) or extreme gradient boosting (XGBoost), artificial neural networks (ANNs) etc. SVMs, RFs and XGBoost have become popular due to their competitive performance, while not having many hyperparameters (Ogutu, Piepho et al. 2011, Ghafouri-Kesbi, Rahimi-Mianji et al. 2017). More recently, deep learning (DL) is widely tested because of its inherent potential to learn complex genotype-phenotype relations using dense layers of neurons, and previous successes in regulatory genomics (Eraslan, Avsec et al. 2019, Montesinos-López, Martín-Vallejo et al. 2019). The effectiveness of DL for GP is still an

open question because, at the current scale of SNP data, model complexity can become very high, with millions of estimable parameters and prone to overfitting due to the availability of limited data (i.e. $n << p$) (Azodi, Bolger et al. 2019, Montesinos-López, Montesinos-López et al. 2021). Nevertheless, many studies on GP have demonstrated DL as a potential competitor to both LMM and ML methods. Overall, the methodological evolution has resulted in a considerable number of choices in the modelling space and selecting the best method in a particular GS setting remains an open research question.

1.3 Challenges for genomic prediction performance

Generally, a GP model is developed on a training (reference) population for which both genotypes and phenotypes are known, and applied to a test (breeding) population for which only the genotypes are known, to predict gEBVs or phenotypes. Prediction accuracy is then often measured as the correlation between observed and predicted phenotype values or gEBVs of the training data. In practice, a preliminary model version is tested on a held-out subset of the training data for parameter optimisation using, for example, repeated random sampling or cross-validation schemes. The aim is to develop a model which is not overfitted on the training data, robust to outliers, able to deal with confounding factors (like population structure, low allele frequencies or LD patterns, etc.), and with precise parameter estimates with high prediction accuracy on the test data. Here, I discuss two main challenges and their possible remedies, that will be central to this thesis.

1.3.1 High dimensionality

High dimensionality is a common characteristic of GP datasets, because of the desire to capture all QTL effects from their linked SNPs. So dense genotyping is often employed, using whole-genome sequencing (WGS) or high density SNP genotyping arrays, resulting in millions (M) of SNP genotypes. For instance, a dataset of 3M imputed SNPs was recently published for the model plant *Arabidopsis thaliana* (Arouisse, Korte et al. 2020), and ~29M SNPs were reported in a study with 3,010 Asian rice genomes (Wang, Mauleon et al. 2018) using WGS. On the other hand, phenotypic observations are usually available for only a few hundred plants, sometimes including replicates with the same genotype. Despite significant progress in high-throughput phenotyping (HTP) during the last decade, phenotyping plants in comparable numbers to those of the SNPs is still not practical. This results in a situation often termed “the curse of dimensionality” (Manthena, Jarquín et al. 2022), where the number of observations n is far smaller than the number of variables p ($n << p$). As a consequence, the models become very complex with millions of estimable parameters, are prone to overfit on the limited training data and do not generalise well on the test data.

1.3.2 Genetic architecture

Complex traits are highly polygenic and governed by both additive genetic and non-additive (dominance, epistasis) effects (Varona, Legarra et al. 2018). If GS is applied to breeding populations with large heterozygosity, significant dominance effects can be realized (Ishimori, Hattori et al. 2020). On the other hand, in conventional LMM modelling breeding values by definition are considered to be additive (Meuwissen, Hayes et al. 2001). This necessitates incorporation of non-additive effects to increase the prediction accuracy. Moreover, the total genetic variance underlying complex phenotypes is differentially enriched across multiple genomic regions, represented by the heritability models (Speed, Holmes et al. 2020), that can differ between related traits. Collectively, the influence of genetic architectures on prediction accuracy could be very strong, if not carefully considered during GP modelling (Daetwyler, Pong-Wong et al. 2010).

1.4 Improvement strategies

1.4.1 Mitigating high-dimensionality

To deal with the $n << p$ problem, a constraint can be applied to shrink the SNP effects, so that many of the small effects are shrunk towards zero. Ridge (Arthur and Robert 1970) and lasso regression (Tibshirani 1996) introduced the ℓ_2 and ℓ_1 norm of SNP effects as a penalty term in the loss function, respectively. With LMMs, the shrinkage-based estimation of genetic effects is conducted by considering SNP effects as random (Meuwissen, Hayes et al. 2001). The level of shrinkage is estimated from the genetic and residual variances of the data, rather than tuning it as a hyperparameter, as in the case of ridge and lasso. Employing parameter shrinkage, also called model regularisation or simplification, has become standard practice in dealing with the high-dimensional data, although an optimal sparse solution is not guaranteed.

Other approaches select a subset of SNPs based on their potential importance, or remove highly correlated SNPs (Li, Zhang et al. 2018, Yin, Zhang et al. 2020, Selga, Koc et al. 2021). For instance, SNPs can be ranked based on their importance as inferred from the strength of phenotypic associations in linear models (Torstensson 2017), or by any algorithmic importance score, e.g. Gini impurities in Random Forest (Wang, Aggarwal et al. 2017). A downside of this approach is that the subsequent GP model performance completely relies on the algorithmic limitations of the feature selection method. Moreover, due to often imprecise parameter estimates, optimal SNP subset selection is usually challenging. An approach to remove highly correlated SNPs uses binning of contiguous regions on the genome and selects only one SNP out of a bin (Du, Wei et al. 2018), but an optimal selection of bin size is not straightforward. Dimensionality reduction can also be achieved by transformation of the genotype matrix into a low-dimensional representation (Manthena, Jarquín et al. 2022). For instance, singular value decomposition of the genotype matrix can be employed to perform GP using a smaller

number of principal components as features instead of using all SNPs (Pintus, Gaspa et al. 2012, Du, Wei et al. 2018, Odegard, Indahl et al. 2018).

Some of the ML methods already discussed can deal with the high-dimensional data by design. For instance, the Random Forest (Breiman 2001), with random sub-sampling of both training data and features, is not much affected by an increase in dimensionality and weighted subsampling has been shown to increase robustness towards high-dimensionality even further (Nguyen, Zhao et al. 2015). Still, the trade-off between pruning of the feature space, simplification of the parameter space and generalisation is often a difficult job, which requires further improvements in GP methodologies.

1.4.2 Incorporating genetic architectural features in the model

Complex traits might be highly polygenic with both additive and non-additive modes of allele actions, e.g. dominance and epistasis. LMMs (both BLUP and Bayesian) have been heuristically equipped to accommodate non-additive effects. For instance, for GBLUP the random genetic effects component in equation (1.2) has been extended with an estimated population kinship / covariance matrix incorporating dominance and interaction effects. This can be achieved by estimation of the genomic kinship matrix by modelling the relationship between a pair of genotypes using some non-linear kernel function between SNPs, that can intrinsically capture non-linear relations. For instance, in case of RKHS, a Gaussian kernel is used to estimate genomic kinships between a pair of genotypes (De los Campos, Gianola et al. 2010). The dominance effects can be explicitly incorporated into allele coding of the genotype matrix as deviations from the additive effects (Alves, da Costa et al. 2020), and similarly epistasis can be modelled as a mathematical formulation of the interactions between additive and dominance effects (Jiang and Reif 2015, Vieira, Dos Santos et al. 2017). Moreover, altering the distribution of SNP effects from normal to a heavier tailed student's *t*-distributions as a marginal prior (e.g. in BayesA and BayesB) or a distributions mixture with multivariate *t*-distributions prior (e.g. in BayesCπ, BayesDπ and BayesR) aided accommodating different trait architectures; resulted into different variants of Bayesian methods (Habier, Fernando et al. 2011, Gianola 2013, Mollandin, Rau et al. 2020). Alternatively, nonlinear ML / DL methods can consider nonlinear relationships between SNPs by design. For instance, the hierarchical tree structure of the decision trees can intrinsically accommodate regulatory dependences between SNPs, such that the SNP explaining larger effect, due to its regulatory role, comes at top of others in the tree (Schmalohr, Grossbach et al. 2018). To this end, these methods have sought considerable attention, but a continuous development is needed to accommodate various aspects of nonlinear genetic effects (Yoshida and Koike 2011, Wright, Ziegler et al. 2016, Nguyen and Le 2018, Orlenko and Moore 2021).

Incorporating SNP- or region-specific genomic variances is yet another approach (Zhang, Liu et al. 2010) to incorporate genetic architecture. For instance, BayesA, BayesB and BayesDπ methods can account for locus-specific variances, whereas, BayesR can have

a Dirichlet prior mixtures for each locus (Mollandin, Rau et al. 2020). Similarly, extensions of GBLUP with region-specific variances have been proposed. For example, MultiBLUP partitions the total genomic variance into multiple components, based on groups of SNPs whose associations are first tested collectively, and the groups with significant associations are retained in the model as separate random effects (Speed and Balding 2014).

In a nutshell, LMMs are strongly influenced by statistical prior assumptions. Alternatively, data-driven non-parametric or machine learning approaches received popularity (van Dijk, Kootstra et al. 2021, Yan and Wang 2023), because they can learn from the data itself. This, in turn, requires a large amount of data for a good model fit, which is scarcely available for GP. In this connection, the statistical tricks employed for the conventional methodology (e.g. regularisation), together with new approaches can, therefore, be tested for superior performance.

1.4.3 Incorporating knowledge in the model

The knowledge is any piece of information that appears useful by increasing prediction performance of genomic prediction. Moreover, I refer to the knowledge as any additional source of information, with its own representation, other than genotypes and the target phenotype, on either SNPs (predictors), individual plants / genotyped accessions or the target phenotype. The information on SNPs can be taken from the pre-existent information on genes linked to the input SNPs, called prior biological knowledge; whereas, measurements of endophenotypes (e.g. transcriptomics, proteomics, metabolomics etc), multiple traits and environment can be used as additional information on individuals or phenotypes, respectively.

For instance, in the stated above 'MultiBLUP' approach, SNP groups can be inferred from a priori known groups of gene functions. This allows prioritisation of groups of SNPs based on prior biological knowledge; and the model is expected to converge more quickly, generalise well over the unseen data, because publicly available information could be generated from multiple experiments, estimate larger genetic variance and possibly increase the predictive ability (Edwards, Sorensen et al. 2016, MacLeod, Bowman et al. 2016, Sarup, Jensen et al. 2016, Fang, Sahana et al. 2017, Rohde, Demontis et al. 2017).

1.4.3.1 Sources of prior biological knowledge

Before diving deeper into the potential utilisation of prior biological knowledge, it is important to know what sources of knowledge are readily available.

1) Sequence properties

DNA sequences contain specific signatures that can be used to predict their biological functions. Such patterns can be informative because they might differently associate to the phenotype. Genic, intergenic, coding / non-coding, upstream / downstream regions

have different nucleotide preferences, motifs and roles in controlling gene expression. For instance, GC content varies across the genomes as well as within and across genes (Glémin, Clément et al. 2014). The sequence patterns and accessibility status of the transcription factor binding sites in the promotor regions explain the gene regulation mechanism underlying the observable phenotypes. The locations and patterns of these DNA sequences provide contextual and functional information that can be attributed to the SNPs within / near these sequences. For instance, SNPs in coding regions might alter the translated protein completely. A regulatory SNP within promotors, enhancers, silencers or upstream and downstream of protein coding genes is more influential than a SNP within intron of a gene.

Other than SNPs, larger sequence polymorphism blocks, for instance, segmental duplications, deletions, insertions or copy number variations might also be inferred from these sequence properties (Wijfjes, Smit et al. 2019) and can potentially be associated with plant traits (Dolatabadian, Patel et al. 2017). In summary, sequence properties can help group SNPs with related characteristics for use in GP modelling.

2) Gene function information

Genes have specific annotated roles, which are commonly represented using an ontology / controlled vocabulary of annotation terms, known as Gene Ontology (GO) (Ashburner, Ball et al. 2000). The GO is a hierarchical graph-based resource, organising annotation terms at various levels of abstractions from three different aspects: biological processes, molecular functions and cellular components. For model organisms like *Arabidopsis thaliana*, several genes are annotated with GO terms derived from experimental analysis (Ashburner, Ball et al. 2000, Rhee, Beavis et al. 2003). Sequence homology and other methods of computational inference are often used to assign GO terms to unannotated genes of other organisms (Gene Ontology 2021). Besides GO annotations, some other publicly available databases contain functional annotations. For instance, the GENES database of Kyoto Encyclopaedia of Genes and Genomes (KEGG) (Kanehisa and Goto 2000, Kanehisa, Furumichi et al. 2022) contains a collection of gene catalogues with up-to-date annotations of gene functions, curated manually from published literature. Functional groups of genes can also be inferred from gene expression data, where clusters of co-expressed genes can be computationally predicted and genes in a cluster are expected to be involved in a similar function (Serin, Nijveen et al. 2016, Ma, Zhao et al. 2018).

3) Protein-protein interaction (PPI) information

Proteins act together in many different ways to perform a certain biological function. These interactions can be predicted based on protein sequence and structure properties (Bryant, Pozzati et al. 2022) and validated using biochemical assays. Over the past few years, computational approaches have achieved great success in predicting protein-protein interactions (PPI) to complement and guide costly and labour-intensive

experimental studies. Resultantly, several databases are dedicated to hosting such interactions, keeping record of their prediction / validation evidence (Kanehisa and Goto 2000, Mering, Huynen et al. 2003, Mostafavi, Ray et al. 2008).

PPI information can be used to infer protein functions based on the assumption that interacting proteins could be involved in a similar function (guilt by association). Thus, PPI provides a rich resource in the form of a (evidence-weighted) network of protein interactions and their functional linkages. For instance, the STRING database (Szklarczyk, Gable et al. 2021) contains PPI interactions for ~14,000 species, including ~16 million predicted functional associations between *Arabidopsis thaliana* proteins, out of which ~3.6 million are experimentally validated. Similarly, the KEGG PATHWAY database (Kanehisa, Furumichi et al. 2022) contains manually curated high-level gene pathways, whereas AraCyc (Rhee, Zhang et al. 2006) contains computationally derived biochemical pathways of compounds, intermediates, cofactors, reactions, genes, proteins, and protein subcellular locations. Information compiled for model plants may be transferred to non-model plants, and multiple sources can be combined thorough computational approaches. For instance, AraNet v2 (Lee, Yang et al. 2015) integrates 19 diverse data types to yield a comprehensive and reliable gene functional interaction network for *Arabidopsis thaliana*.

1.4.3.2 Strategies to incorporate prior biological knowledge in the models

In the context of the available sources of prior knowledge, two approaches are relevant and used to develop new improved methodologies in this thesis:

1) Differential regularisation / prioritisation

Uniform regularisation for all SNPs does not correspond to the genetic layout of many complex traits and may cancel out some small but useful effects. A better alternative could be to employ weighted regularisation based on heritability models of the traits. Such weights could be based on prior biological knowledge, i.e. certain SNPs or genomic segments could be differentially regularised based on their relation to the trait or the underlying biology. In principle, any information on similarity in function and interaction of biological entities linked to SNPs, i.e. genes, transcripts, protein and metabolites etc., can help group these SNPs.

A straight-forward implementation for differential prioritisation using GBLUP is to test each functionally related group of SNPs separately, by partitioning the total genetic effect in equation (1.2) into two components as proposed in (Edwards, Thomsen et al. 2015, Edwards, Sorensen et al. 2016). The approach, called Genomic Feature BLUP (GFBLUP), can predict breeding values using group specific genomic relationships. The authors tested a number of pathway / GO terms, one by one, on whether the related SNPs could explain more genetic variance than the remaining SNPs. Instead of a single pathway / GO term at a time, it has also been proposed to partition the total genetic effect

over multiple SNP groups (Speed and Balding 2014). Similarly, knowledge-based groups can define classes of SNPs in the Bayesian framework; for instance, BayesRC uses group-specific Dirichlet priors (MacLeod, Bowman et al. 2016). A limitation to this approach is that SNP groups cannot overlap, to avoid over-estimation of SNP effects using multiple random components, whereas genes can be annotated with multiple pathway / GO terms. A solution to these issues could be to incorporate SNP partitioning into non-parametric ML / DL methods, which will be explored further in this thesis.

2) Knowledge-guided networks

GP models can capitalise on interaction information, potentially describing epistasis relationships, to design novel methods. In particular, various DL architectures have been proposed to do so. Such knowledge-primed models derive connections between nodes in the neural network from available biological network data (Kang, Ding et al. 2017, Ma, Yu et al. 2018, Snow, Noghabi et al. 2019, Fortelny and Bock 2020, van Hilten, Kushner et al. 2020, Bourgeais, Zehraoui et al. 2021). In addition to potential benefits for prediction performance, these approaches can also be suitable for developing explainable models (Azodi, Tang et al. 2020, Mieth, Rozier et al. 2020, Novakovsky, Dexter et al. 2022).

Despite many successful demonstrations of GP models using prior knowledge, practical use is still limited as performance varies across traits, populations and species. Moreover, aggregating different knowledge sources and taking their validity into account is challenging and requires further research.

1.4.4 Exploiting information from additional data types

Over the past few years, alongside developments in high-throughput genotyping of genome-wide polymorphisms (genomics), phenotyping of multiple traits per plant (phenomics) has also undergone remarkable progress. Current HTP systems can take multiple measurements per trait, at a broad range of growth and development stages and in carefully controlled environments (Sandhu, Mihalyov et al. 2021). On the other hand, in-field monitoring sensors (fixed or moveable) can provide a wealth of spectral imagery or health parameters, scaling up phenomics for GP modelling (Ecarnot, Compan et al. 2013). The spectra at particular electromagnetic wavelengths contains indirect information about plant physiological or agronomical traits, derived from biochemical responses at the molecular level, through reflectance, absorption or transmittance of light energy (Sandhu, Mihalyov et al. 2021). Commonly used spectral reflectance indices (SRI) are derived from visible and near-infrared imaging, reflecting plant vegetative growth, pigmentation, photosynthesis efficiency and hydration status. Alternatively, broad-spectrum spectroscopy, e.g. hyperspectral imaging, has also shown to have great potential (Rincent, Charpentier et al. 2018, Zhu, Leiser et al. 2021, Zhu, Maurer et al. 2021, Robert, Auzanneau et al. 2022). On the other hand, advances in high-throughput molecular characterisation have made molecular profiling of downstream biological strata

such as transcriptomics, proteomics and metabolomics etc, recordable as cellular manifestations of the observable phenotypes (Hu, Campbell et al. 2021).

These new data types come with new challenges as well as opportunities for GP. Large numbers of heterogeneous measurements become available for the same plants (or sometimes different, but genetically related plants). The resulting datasets may have different modalities, dimensionality, data types and structures. Missing measurements and genotypes makes imputation and use in GP even more challenging (Flores, Claborne et al. 2023). Efficient integration of these heterogeneous datasets into GP has, therefore, become an important topic in modern data-driven plant breeding (Subramanian, Verma et al. 2020). Each dataset can be used as a standalone information resource for predicting phenotypes. For instance, using only the multi- / hyperspectral data, often called phenomics prediction (Edlich-Muth, Muraya et al. 2016, Rincent, Charpentier et al. 2018, Adak, Murray et al. 2021, Brault, Lazerges et al. 2021, Robert, Auzanneau et al. 2022), has been reported to potentially be as effective as genomic prediction. In principle, any -omics dataset can serve for GP, where in many cases metabolomics measurements could be considered to be closest to the observable phenotypes, as metabolism is influenced by other levels in the cell (the genome, transcriptome and proteome). However, metabolites generally have a higher turn-over rate than other biomolecules, such as the genome, which is generally stable throughout the life span of an organism (Michel, Wagner et al. 2021). Contrasting evidences of superiority of metabolic prediction (MP) individually over GP has been reported in different studies on maize and barley (Riedelsheimer, Czedik-Eysenberg et al. 2012, Schrag, Westhues et al. 2018, Gemmer, Richter et al. 2020). On the other hand, transcriptomic prediction (TP) has been reported to be as good as GP or even better (Frisch, Thiemann et al. 2010, Guo, Magwire et al. 2016). Combining these downstream -omics with GP has shown to be a promising approach than their individual use for prediction (Schrag, Westhues et al. 2018).

Augmenting these data types with genomic information, i.e. introducing them as an extra source of information for the model, can be advantageous because it can exploit correlated information from multiple datasets. In this thesis, I however focus on using phenomics data by itself, which is increasingly common in plant breeding because of the growing availability of cost-effective HTP hardware and software. The conventional GP methodology based on univariate LMMs can be extended to multi-variate / multi-trait GP (Bhatta, Gutierrez et al. 2020). Alternatively, non-parametric ML / DL can be used, to avoid parametric assumptions (Sandhu, Patil et al. 2021). Nevertheless, LMMs may still be relevant because using multiple traits drastically increases the number of parameters to estimate, and hence model complexity.

Incorporating genetic architectural characteristics, prior information of the underlying genetics or information from additional data types, shifts GP modelling towards a knowledge-driven approach. This comes with its own challenges related to strategies to

include knowledge, assess credibility and generalisation, benchmarking and standardisation.

1.5 Outline and contribution of this thesis

This thesis explores knowledge-driven genomic prediction, both in application and through development of novel methods. Two approaches are explored: 1) using prior biological knowledge, 2) using information from correlated traits. The use of prior knowledge was tested in LMMs for differential prioritisation of markers, and in the development of a novel deep learning framework with both differential prioritisation and sparsity induction. **Chapter 2** explores the methodological space for GP to identify applications in which LMMs, ensemble machine learning methods, i.e. Random Forests (RFs) and extreme gradient boosting machines (XGBoost), perform best. **Chapter 3** applies differential prioritisation based on various gene ontology (GO) terms and co-expressed gene clusters in LMMs to predict complex traits, like photosynthetic light use efficiency (φ_{PSII}) and projected leaf area (PLA) in *Arabidopsis thaliana*. **Chapter 4** presents a novel Multilayer Perceptron (MLP) framework, called PRIORNET, that can prioritise markers based on a provided list of genes / GO / pathways, and uses protein-protein interactions to implement a sparse network architecture. **Chapter 5** explores the potential of using photosynthesis-related traits as component / secondary traits for improving genomic prediction of PLA, using two-trait and multi-trait modelling. Moreover, it considers using multiple measurements to observe the impact of light dynamics along the growth trajectory.

CHAPTER

2

Genomic prediction in plants: opportunities for ensemble machine learning based approaches

Muhammad Farooq, Aalt D.J. van Dijk, Harm Nijveen,
Shahid Mansoor and Dick de Ridder

This chapter was published as:

Farooq, Muhammad, et al. "Genomic prediction in plants: opportunities for ensemble machine learning based approaches." F1000Research 11.802 (2022): 802.

Abstract

Many studies have demonstrated the utility of machine learning (ML) methods for genomic prediction (GP) of various plant traits, but it is still largely unclear if and when ML should be chosen over conventionally used, often simpler parametric methods,. Predictive performance of GP models might depend on a plethora of factors including sample size, number of markers, population structure and genetic architecture. Here, we investigate which problem and dataset characteristics are related to good performance of ML methods for genomic prediction. We compare the predictive performance of two frequently used ensemble ML methods (Random Forest and Extreme Gradient Boosting) with parametric methods including genomic best linear unbiased prediction (GBLUP), reproducing kernel Hilbert space regression (RKHS), BayesA and BayesB. To explore problem characteristics, we use simulated and real plant traits under different genetic complexity levels determined by the number of Quantitative Trait Loci (QTLs), heritability (h^2 and h_e^2), population structure and linkage disequilibrium between causal nucleotides and other SNPs. Decision tree based ensemble ML methods are a reasonable choice for phenotypes with allelic interactions and are comparable to Bayesian methods for additive phenotypes in the case of large effect Quantitative Trait Nucleotides (QTNs). Furthermore, we find that ML methods are susceptible to confounding due to population structure but less sensitive to low linkage disequilibrium than linear parametric methods. Overall, this provides insights into the usefulness of ML in GP as well as guidelines for practitioners.

Keywords:

ANN: artificial neural network

BLUPs: Best Linear Unbiased Predictions

GBLUP: Genomic Best Linear Unbiased Prediction

GP: Genomic Prediction

MLP: Multilayer Perceptron

QTL: Quantitative Trait Locus

QTN: Quantitative Trait Nucleotide

RF: Random Forest

RKHS: Reproducing Kernel Hilbert Spacing

SNP: single nucleotide polymorphism

SVM: Support Vector Machine

SVR: Support Vector Regression

XGBoost: Extreme Gradient Boosting

2.1 Introduction

The phenotype of an individual is based on its genetic makeup, the environment and the interplay between them. In plant and animal breeding, the genomic prediction (GP) model, using a genome-wide set of markers, is an integral component of the genomic selection-based approach (Meuwissen, Hayes et al. 2001). A GP model is constructed on a reference population for which both genotypes and corresponding phenotypes are known, mostly employing a cross-validation strategy, and applied to related populations with only genotypes known. The total genomic value, estimated from the GP model, is used as a pseudo-phenotype to select the best parents for the next generation(s). In general, phenotypes differ from each other in terms of their genetic complexity, ranging from simple/monogenic to complex/polygenic. These differences impact the potential performance of GP. Complex traits are predominantly governed by a combination of additive and non-additive (e.g. dominant / recessive, epistatic etc.) allele effects, which makes GP challenging for these traits (Moore, Amos et al. 2015). The genetic architecture of complex traits is characterized by moderate to large numbers of Quantitative Trait Loci (QTLs) with small to medium effect sizes and no or few large effect QTLs (Korte and Farlow 2013). Moreover, the ratio of additive to non-additive genetic variance may differ even for closely related traits. Besides the actual genetic variance level, its distribution over the genome is also a determinant of the trait architecture (Speed and Balding 2019). Next to genetic architecture, population structure plays a role as well (*Figure 2.1*): prediction accuracies are influenced by inconsistent relatedness among samples due to ancestral allele frequency imbalance among sub-populations (population structure) or cryptic structures, e.g. familial relationships; linkage disequilibrium (LD) structure, due to inbreeding or selection pressure; varying relatedness between training and test populations, e.g. over the course of a breeding cycle; and sizes of reference and effective populations (Zhao, Chen et al. 2012).

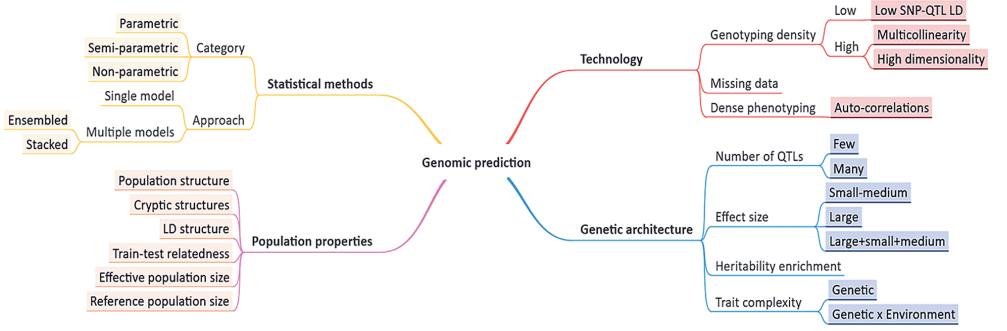


Figure 2.1: Genomic prediction characteristics.

Factors affecting genomic prediction performance, often measured as correlation between true phenotype values and those predicted by a model.

Technological advances and statistical frameworks used bring new challenges (*Figure 2.1*). Genotyping and/or phenotyping technologies can now generate millions of markers and thousands of phenotypic measurements, e.g. in time series, increasing the dimensionality of the prediction problem. For example, using a high-density SNP array (or imputing SNPs based on a low-density array) increases the likelihood of getting many markers in LD with the true QTL (high SNP-QTL LD). It can increase total explained variance (Ogawa, Matsuda et al. 2016), but may induce multicollinearity among SNPs. Consequently, SNP selection prior to predictive modelling has been reported to provide superior performance compared to simply using a dense marker set (Veerkamp, Bouwman et al. 2016). In contrast, low-density genotyping can miss important SNPs in LD with, or weakly linked to, the QTLs, leading to inferior prediction performance (de Los Campos, Sorensen et al. 2019).

Statistical genetics approaches have traditionally focused on formulating phenotype prediction as a parametric regression of one or more phenotypes on genomic markers, treating non-genetic effects as fixed or random in a linear equation. The resulting GP models are biologically interpretable but might yield poor performance for complex phenotypes, as linear regression fails to capture the more complex relations (Pérez-Rodríguez, Gianola et al. 2012). This approach also requires proper translation of prior knowledge on the genetics underlying phenotypes into parametric distributions. Although statistical distributions can help describe genetic architecture, devising a specific distribution for each phenotype is impractical. Therefore, many variations of linear regression were proposed by relaxing statistical assumptions; the main differences lie in their estimation framework and prior assumptions on the random effects (for an overview, see 'Models'). Alternatively, machine learning (ML) offers a more general set of non-parametric methods that can model phenotypes as (non) linear combinations of genotypes. Moreover, these methods can jointly model the problem, e.g. strong learners can be stacked (Sapkota, Boatwright et al. 2020) or weak learners can be combined in an ensemble. Examples include Support Vector Machines (SVMs), (ensembles of) decision trees and artificial neural networks (ANNs). No statistical assumptions are required in advance; therefore, these methods should be able to pick up more complex genetic signals that are missed by linear models. The downside is the large amount of data required for learning these models from the data.

The performance of ML methods in GP problems has previously been compared using simulated and real phenotypes. Some were found to perform better under non-additive allelic activity (Howard, Carriquiry et al. 2014, Abdollahi-Arpanahi, Gianola et al. 2020); however, a clear link between simulated and real phenotypes is often missing, or only a specific breeding population structure is considered. For example, Barbosa, da Silva et al. (2021) compared the performance of ML and statistical methods in a simulated F_2 population of 1,000 individuals and 2,010 SNPs using 26 simulated phenotypes. They varied the heritability and number of QTLs and included dominant and epistatic effects.

They observed that ML methods performed better at low QTL numbers and hypothesized that a reason for this is that with fewer controlling genes, epistatic interactions are more important. But it is still unclear if this is a general conclusion towards a population with different characteristics e.g. natural populations. Moreover, there are conflicting reports on performance of ML (Howard, Carriquiry et al. 2014, Grinberg, Orhobor et al. 2020). For example, ANNs have been reported to perform worse in some applications and are comparable to competing methods in others (Bellot, de Los Campos et al. 2018, Abdollahi-Arpanahi, Gianola et al. 2020). Ensemble decision tree methods, combining the output of a large number of simple predictors, have proven better for some traits but not for others (Ogutu, Piepho et al. 2011, Ghafouri-Kesbi, Rahimi-Mianji et al. 2017, Azodi, Bolger et al. 2019). Gradient boosting showed improved performances for many real traits (Li, Zhang et al. 2018, Yan, Xu et al. 2021) but was inferior to random forests on simulated datasets (Ogutu, Piepho et al. 2011). Furthermore, the impact of population structure and low SNP-QTL LD on the performance of ML methods is still unclear.

In this paper, we investigate which GP characteristics (genetic architecture, population properties and genotype/phenotype measurement technology) a priori point to a better performance for either traditional statistical approaches or ML-based methods. We compare GP performance of two ensemble methods, Random Forests (RF) and Extreme Gradient Boosting (XGBoost), to that of linear mixed models, GBLUP, BayesA, BayesB and RKHS regression with averaged multi-Gaussian kernels. We focus on typical applications in plant breeding to explore various GP characteristics, including the ratio of the total number of markers to the number of samples (p/n), genetic complexity, QTN effect sizes and distributions, additive vs. epistatic heritabilities, sparse vs. dense genotyping and population structure.

2.2 Methods

2.2.1 Data

Simulations:

In a first experiment, artificial genotypes were simulated, in combination with associated phenotype values. Genotype data was simulated for a diploid population with a minor allele frequency (MAF) of 0.4, using a binomial distribution, where each allele was the outcome of a binomial trial. The genotype dataset was coded as {0=AA, 1=Aa, 2=aa}. We fixed MAF for all SNPs, in order not to incorporate the impact of allele frequencies because MAF of a QTL can impact its heritability estimation and ultimately prediction accuracy of the GP model. Moreover, in this way we observed equal and reasonably statistical power for each SNP during allele effects estimation. To explore GP characteristics (*Figure 2.1*), different levels of genetic complexity and dimensionality, defined as the ratio of total number of SNPs to the sample size ($c = p/n$), were simulated. For the high dimensionality scenarios, sample size was fixed at n = 500, because

reference populations of this size are feasible for genotyping and phenotyping in genomic selection studies. Using values of c = {2, 10, 20, 40, 120}, the number of SNPs varied up to p = 60,000 (120*500). Similarly, for the low dimensionality scenarios, the number of SNPs was fixed at p = 500 and sample size was varied up to n = 3,000 to arrive at c = {1, 1/2, 1/4, 1/6}. Subsequently, Quantitative Trait Nucleotides (QTNs) were randomly selected from these simulated SNP sets to generate phenotypes. We selected either 5, 50, 100, p/2 or p QTNs, corresponding to a range of low to high genetic complexity, coupled with a narrow-sense heritability ranging from 0.1 to 0.7. A phenotype with a high number of QTNs and low heritability is more complex than one with few QTNs and higher heritability.

Phenotype datasets were generated using the simplePHENOTYPES v1.3.0 R package (Fernandes and Lipka 2020). Additive polygenic phenotypes were simulated using additive modes of allele effects, as follows:

$$\mathbf{y} = \beta_1\mathbf{QTN}_1 + \beta_2\mathbf{QTN}_2 + \beta_3\mathbf{QTN}_3 + \cdots + \beta_k\mathbf{QTN}_\mathrm{k} + \boldsymbol{\varepsilon} \qquad (2.1)$$

Here, β_i describes the effect size of the i^{th} QTN, where $\mathbf{QTN}_i$ is a vector containing the allele dosages for the i^{th} QTN for all samples. The residuals ($\boldsymbol{\varepsilon}$) were sampled from a normal distribution $N(0, \sqrt{(1- h^2)})$. Three different approaches were used to sample the effect sizes: (i) The narrow-sense heritability (h^2) determines the variance of effect sizes distribution: one effect β is randomly sampled from $N(0, \sqrt{h^2})$ and equally divided among all of the q QTNs, such that each QTN is assigned an effect size of $\beta_i = \beta/q$, referred to as 'simulations with equal/uniform effects' in the text. This allows, smaller effect sizes to be generated by increasing the number of QTNs, thereby simulating increasing genetic complexity by lowering effect sizes. (ii) To further explore genetic complexity, we used equation (2.1) to generate another set of phenotypes where the first QTN is assigned a larger effect than others. For this, we chose an effect two standard deviations away from the mean of the effect sizes distribution (large effect) and the rest small to medium were allowed to be sampled up to one standard deviation way from the mean from $N(0, \sqrt{h^2})$. (ii) As a third case, we sampled all QTN effects randomly from the effect sizes distribution.

For non-additive phenotypes, broad-sense heritability was set at most to 0.8, so the distribution of residuals is $N(0,\sqrt{0.2})$. We considered only epistasis, ignoring other factors such as dominance. Adding an additional term for epistasis to equation (2.1) results in:

$$\mathbf{y} = \beta_1\mathbf{QTN}_1 + \beta_2\mathbf{QTN}_2 + \beta_3\mathbf{QTN}_3 + \cdots + \beta_k\mathbf{QTN}_\mathrm{k} + \beta_e(\mathbf{QTN}_{e1} * \mathbf{QTN}_{e2}) + \boldsymbol{\varepsilon} \quad (2.2)$$

The epistatic heritability (h_e^2) was set analogous to the additive heritability (h^2), such that $H^2 = h^2 + h_e^2$. The *additive* × *additive* epistasis model was used, with only a single pairwise interaction. The epistatic effect β_e was sampled from $N(0,\sqrt{h_e^2})$ and attributed to a single interacting pair of markers ($e1$, $e2$) such that $\beta_e = \beta_{e1} \times \beta_{e2}$. We sampled this interacting pair from the set of additive QTNs; therefore, each interacting marker will always have some main effect. As for additive phenotypes, we also created epistatic

phenotypes with one large effect QTN. The total number of settings (scenarios considered in *Table 2.1*) for the simulated GP characteristics was 135 per phenotype class, i.e. additive and epistatic. For each class, phenotypes were simulated with and without a large effect QTN. Thus, in total 810 (135 × 2 × 3) simulated phenotypic scenarios were generated, each having five independent phenotypic traits. These will be referred to as 'simdata' in the text.

Table 2.1: Simulation scenarios permutations.

					Additive phenotypes $H^2 = h^2$	**Epistatic phenotypes** $H^2 = h^2 + h_e^2 = 0.8$
#Markers (*p*) / #Samples (*n*)	**Ratio (*c=p/n*)**	**#QTNs (*q*)**	**#Markers (*p*) / #QTNs (*q*)**	**Ratio (*d=p/q*)**	h^2	$h^2 + h_e^2$
500/3k	0.17	5, 50, 100, 250, 500	500/5, 500/50, 500/100, 500/250, 500/500	100, 10, 5, 2, 1	0.1, 0.4, 0.7	0.7+0.1, 0.4+0.4, 0.1+0.7
500/2k	0.25	5, 50, 100, 250, 500	500/5, 500/50, 500/100, 500/250, 500/500	100, 10, 5, 2, 1	0.1, 0.4, 0.7	0.7+0.1, 0.4+0.4, 0.1+0.7
500/1k	0.50	5, 50, 100, 250, 500	500/5, 500/50, 500/100, 500/250, 500/500	100, 10, 5, 2, 1	0.1, 0.4, 0.7	0.7+0.1, 0.4+0.4, 0.1+0.7
500/500	1	5, 50, 100, 250, 500	500/5, 500/50, 500/100, 500/250, 500/500	100, 10, 5, 2, 1	0.1, 0.4, 0.7	0.7+0.1, 0.4+0.4, 0.1+0.7
1k/500	2	5, 50, 100, 500, 1k	1k/5, 1k/50, 1k/100, 1k/250, 1k/1k	200, 20, 10, 2, 1	0.1, 0.4, 0.7	0.7+0.1, 0.4+0.4, 0.1+0.7
5k/500	10	5, 50, 100, 2.5k, 5k	5k/5, 5k/50, 5k/100, 5k/250, 5k/5k	1k, 100, 50, 2, 1	0.1, 0.4, 0.7	0.7+0.1, 0.4+0.4, 0.1+0.7
10k/500	20	5, 50, 100, 5k, 10k	10k/5, 10k/50, 10k/100, 10k/250, 10k/10k	2k, 200, 100, 2, 1	0.1, 0.4, 0.7	0.7+0.1, 0.4+0.4, 0.1+0.7
20k/500	40	5, 50, 100, 10k, 20k	20k/5, 20k/50, 20k/100, 20k/250, 20k/20k	4k, 400, 200, 2, 1	0.1, 0.4, 0.7	0.7+0.1, 0.4+0.4, 0.1+0.7
60k/500	120	5, 50, 100, 30k, 60k	60k/5, 60k/50, 60k/100, 60k/250, 60k/60k	12k, 1.2k, 600, 2, 1	0.1, 0.4, 0.7	0.7+0.1, 0.4+0.4, 0.1+0.7

* Note: 1k=1000.

Real datasets

To compare trends observed in simulations with outcomes obtained with real traits, publicly available wheat genotype and phenotype data were taken from Norman, Taylor et al. (2017). This includes 13 traits: biomass, glaucousness, grain protein, grain yield, greenness, growth habit, leaf loss, leaf width, Normalised Difference Vegetative Index (NDVI), physiological yellows, plant height, test weight (TW) and thousand kernel weight (TKW). This particular dataset was chosen as it contains a fairly large number of genotypes (n = 10,375) each genotyped for p = 17,181 SNPs. The impact of population structure, training set size, marker density and its interaction with population structure was assessed in a study by the same authors (Norman, Taylor et al. 2018) and GBLUP prediction accuracies were reported to saturate when training set size was greater than 8,000. We used the same settings, with five-fold cross-validation repeated for five times (training set size 8,300, validation set size 2,075).

The data was generated from a small-plot field experiment for pre-screening of germplasm containing some genotypes that are sown in multiple plots, thus containing spatial heterogeneity with correlation between closely located plots and imbalance in the number of phenotypes per genotype. Soil elevation and salinity, spatial coordinates and virtual blocks (made available on request by the authors) were taken as covariates:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{Z_g g} + \boldsymbol{\varepsilon} \qquad (2.3)$$

Here, $\mathbf{X}$ is the $n \times 4$ design matrix for the fixed effects and overall mean, $\mathbf{b}$ is a 4×1 vector of fixed effects, i.e. soil salinity and elevation; $\mathbf{Z}$ is an $n \times 3$ design matrix for non-genetic random effects $\mathbf{u}$, i.e. range, row and block; $\mathbf{Z_g}$ is the $n \times k$ design matrix for genotypes $\mathbf{g}$ for a maximum of k replicates, and $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of residuals. The Best Linear Unbiased Estimates (BLUEs) of genotypes were used for GP; in this way, we take care of the experimental design factors. Note that equation (2.3) does not contain any SNP information, instead only genotype accessions are used to obtain their adjusted phenotypes.

2.2.2 Population structure analysis

To analyse the influence of population structure on the performance of different GP methods, we used a population of the *Arabidopsis thaliana* RegMap panel (Horton, Hancock et al. 2012) with known structure, containing 1,307 accessions including regional samples (Extended data, *Figure S6*). Additive phenotypes were simulated using narrow-sense heritabilities h^2 = 0.1, 0.4 and 0.7, with equal effect QTNs. The genotypes, available from the *Arabidopsis* 250k SNP array, were further pruned for LD and minor allele frequency (MAF > 5%) using PLINK v1.9 (Purcell, Neale et al. 2007). LD pruning was carried out using a window size of 500 markers, stride of 50 and pairwise r^2 threshold of 0.1, using the '--indep-pairwise' command. This implies that a set of markers in the 500-marker window with squared pairwise correlation greater than 0.1 is greedily pruned from

the window until no such pairs remain. This dataset will be referred to as 'STRUCT-simdata' in the text.

The effect of population structure was also assessed on real data: a genotype dataset of 300 out of the 1,307 RegMap accessions, phenotyped for the sodium accumulation trait with a strongly associated gene (Baxter, Brazelton et al. 2010). This should resemble one of our simulation scenarios, i.e. high heritability (e.g. h^2 = 0.7) with few QTNs (e.g. 5) of large effect. This dataset will be referred to as 'STRUCT-realdata' in the text.

To correct for population structure, we used principal components corresponding to the top ten highest eigenvalues as fixed effects in the models for GBLUP, RKHS regression, BayesA and BayesB (Hoffman 2013). Principal component analysis (PCA) was performed on the allele dosage matrix using the 'prcomp' method in R, with centering and scaling. For random forest and XGBoost, we used these top principal components as additional features in the models.

2.2.3 Analysis of SNP-QTN linkage disequilibrium (LD)

To explore the impact of varying LD between SNP markers and actual QTNs on the performance of GP methods, we used two other datasets: one with real genotypes and simulated phenotypes, the other with real genotypes and real traits.

For the first dataset, we selected a natural population with minimal structure, balanced LD, genotyped at roughly equal genomic spacing and mostly inbred lines: the 360 accessions in the core set of the *Arabidopsis thaliana* HapMap population (Baxter, Brazelton et al. 2010). Genotype data of 344 out of the 360 core accessions was obtained from Farooq, van Dijk et al. (2020) containing 207,981 SNPs. The phenotypes were simulated using one of the scenarios in the Section 'Simulations'. The total number of SNPs was kept close to the number of samples and genetic complexity was kept low, to study the impact of SNP-QTN LD only. To this end, we simulated additive phenotypes with h^2 = 0.7 and 5 QTNs with equal effects. Linkage disequilibrium between SNPs was calculated as squared pairwise Pearson correlation coefficient (r^2) using PLINK v1.9 (Purcell, Neale et al. 2007). Input sets of 500 SNPs were selected randomly from pairs with either low LD ($r^2 \leq 0.5$) or high LD ($r^2 > 0.9$); these two sets were used to train two prediction models using each GP method: one model was trained on the QTNs that were used to generate the phenotype, another on QTN-linked SNPs (closest on the genome) instead of the QTNs themselves, from the low or high LD SNPs pool. To avoid spurious correlations between SNPs in both models, non-QTN-linked SNPs were sampled from a different chromosome. We restricted the sampling of QTNs and the QTN-linked SNPs to chromosome 1, whereas the remaining non-QTN SNPs were sampled from chromosome 2. We refer to this dataset as 'LD-simdata' in the text.

For the second dataset, we used three soybean traits (HT: height, YLD: yield and R8: time to R8 developmental stage) phenotyped for the SoyNam population (Xavier, Muir et

al. 2016). This dataset contains recombinant inbred lines (RILs) derived from 40 biparental populations and the set of markers have been extensively selected for the above traits. Moreover, high dimensionality is not an issue as the dataset contains 5,014 samples and 4,235 SNPs. We refer to this dataset as 'LD-soy' in the text. A complete list of datasets used in this study has been provided in *Table 2.2* and achieved into public repositories.

Table 2.2: List of datasets.

ID	Description	*n*	*p*
simdata	Simulated dataset used to explore GP characteristics of trait genetic complexity, population properties and dimensionality.	See Methods section 2.2.1 for details.	
Wheat	Real wheat dataset from Norman, Taylor et al. (2017) containing 13 traits of varying genetic complexity. These traits are referred to by abbreviations: BM: Biomass, PH: Plant Height, NDVI: Normalised Difference Vegetative Index, LL: Leaf Loss, LW: Leaf Width, GY: Grain Yield, GL: Glaucousness, GP: Grain Protein, Y: Physiological Yellows, TW: Test Weight of grains, TKW: Thousand Kernel Weight, GH: Growth Habit, GR: Greenness	10,375	17,181
STRUCT-simdata	Real structured RegMap panel genotype data of *Arabidopsis thaliana* with simulated phenotypes data used to analyse the effect of population structure	1,307	15,662
STRUCT-realdata	A subset of the real *Arabidopsis thaliana* structured RegMap panel genotype data with real phenotype data of the sodium accumulation trait used to analyse the effect of population structure	300	169,881
LD-simdata	An unstructured set accessions from the core set of the *Arabidopsis thaliana* HapMap population with known genotype data and simulated phenotype data to study the impact of LD	344	48,343
LD-soy	Real soybean dataset of with real phenotypes (R8, HT: height and YLD: yield) for studying the impact of low SNP-QTN LD (Xavier, Muir et al. 2016)	5,014	4,235

2.2.4 Models

A wide range of statistical models have been proposed for GP. Most widely applied are Linear Mixed Models (LMMs), which use whole-genome regression to tackle multicollinearity and high-dimensionality with shrinkage during parameter estimation, employing either a frequentist approach, e.g. restricted maximum likelihood (REML), or Bayesian theory (Xavier 2019). Below, we briefly describe the GP methods used in our experiments. For (semi) parametric methods, we used BGLR v1.1.0 with default settings of hyperparameters (Pérez and de Los Campos 2014); for Random Forests, the ranger R package v0.14.1 (Wright and Ziegler 2017); and for XGBoost, h2o4gpu v0.3.3 (Tang, Gill et al. 2021).

a) Parametric models

GBLUP

The genomic best linear unbiased prediction (GBLUP) method uses a Gaussian prior with equal variance for all markers and a covariance matrix between individuals, called the genomic relationship matrix (GRM), calculated using identity by state (IBS) distances between markers for each pair of samples (Calus, Bouwman et al. 2016). SNP effects are modelled as random effects that follow a normal distribution with zero mean and common variance, and are estimated by solving the mixed model equation:

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{g} + \boldsymbol{\varepsilon} \qquad (2.4)$$

Here, $\mathbf{g}$ is an $n \times 1$ vector of the total genomic value of an individual, captured by all genomic markers; $\boldsymbol{\mu}$ is the overall population mean; and $\boldsymbol{\varepsilon}$ is an n-vector of residuals. The genomic values $\mathbf{g}$ and residuals were assumed to be independent and normally distributed as $\mathbf{g} \sim N(0, \mathbf{G}\,\sigma_g^2)$, $\boldsymbol{\varepsilon} \sim N(0, \mathbf{I}\,\sigma_\varepsilon^2)$. Here $\mathbf{G}$ is the GRM, calculated using the rrBLUP v4.6.1 package (Jacquin, Cao et al. 2016) in R, providing variance-covariance structure for genotypes and $\mathbf{I}$ is the identity matrix. Due to the small number of estimable parameters, GBLUP is computationally fast but the assumption of normality only holds when most effects are close to zero and only a few are larger. The limitation of this approach is that it captures only linear relationships between individuals and assumption of equal variance for all marker effects may not be truly valid for many traits.

Bayesian methods

Several Bayesian methods with slight variations in their prior distributions have been proposed to model different genetic architectures (Gianola 2013) e.g. BayesA, using a scaled t-distribution; Bayesian LASSO or BL (Park and Casella 2008), using a double-exponential; BayesCπ (Habier, Fernando et al. 2011) and BayesBπ (Meuwissen, Hayes et al. 2001), both utilising two-component mixture priors with point mass at zero and either

a Gaussian or scaled *t*-distribution, respectively. To control the proportion of zero effect markers, the hyperparameter 'π' was set equal to 0.5, resulting in a weakly informative prior. For simplicity, we refer to BayesBπ as BayesB in the text. The model in equation (2.5) was solved for posterior means in both BayesA and BayesB with the only difference in priors of β_j:

$$\mathbf{y} = \boldsymbol{\mu} + \sum_j^J \mathbf{x}_j \beta_j + \boldsymbol{\varepsilon} \qquad (2.5)$$

Here, $\boldsymbol{\mu}$ is the intercept, $\mathbf{x}_j$ is an *n*-vector of allele dosages for each SNP and β_j is the effect of SNP *j* out of a total of *J* SNPs.

b) Semi-parametric models

Reproducing Kernel Hilbert Spaces (RKHS) regression is a general semiparametric method that models pairwise distances between samples by a Gaussian kernel and can therefore better capture nonlinear relationships than GBLUP. In fact, GBLUP is a special case of RKHS regression, with a linear kernel (De los Campos, Gianola et al. 2010, Jiang and Reif 2015). We used RKHS regression as a representative semi-parametric model, because it not only employs prior assumptions for random components in LMM equation (2.6), but also learns hyperparameters from the data itself:

$$\mathbf{y} = \boldsymbol{\mu} + \sum_{l=1}^{3} \mathbf{g}_l + \boldsymbol{\varepsilon} \qquad (2.6)$$

In contrast to the GBLUP model in equation (2.4), the RKHS regression model has three random genetic components $\mathbf{g} = \sum_{l=1}^{3} \mathbf{g}_l$, such that $\mathbf{g}_l \sim N(0, \mathbf{K}_l \sigma^2_{\mathrm{gl}})$; where $\mathbf{K}_l$ is the kernel evaluated for the l^{th} component using l^{th} bandwidth (b_l), as described below. This kernel matrix $\mathbf{K}$ is used as genomic relationship matrix, where $\boldsymbol{K} = \{k(\boldsymbol{x_i}, \boldsymbol{x_j})\}$ is an *n* × *n* matrix of Gaussian kernels applied to the average squared-Euclidean distance between genotypes:

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-b\left(\sum_{k=1}^{p}(x_{ik} - x_{jk})^2\right)/p\right) \qquad (2.7)$$

The kernel $k(\boldsymbol{x_i}, \boldsymbol{x_j})$ is a covariance function that maps genetic distances between pairs of individuals $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$ onto a positive real value. The hyperparameter *b*, called the bandwidth, controls the rate at which this covariance function drops with increasing distance between pairs of genotypes. Tuning this parameter for range of values between 0 and 1 could be computationally inefficient. So, instead of tuning *b*, we used a kernel averaging method (De los Campos, Gianola et al. 2010), such that multiple kernels, corresponding to possible bandwidth values b_l = {0.2, 0.5, 0.8}, were averaged.

c) Ensemble machine learning models

Random Forest

The Random Forest (RF) regressor uses an ensemble of decision trees (DTs) that are each grown using bootstrapping (random sampling with replacement of samples), and a random subset of SNPs. The test sample prediction is made by averaging all unpruned DTs as;

$$\hat{f}_{RF}^{D}(x) = \frac{1}{D}\sum_{k=1}^{D}\tau(\boldsymbol{x}, \psi_k) \qquad (2.8)$$

Here $\boldsymbol{x}$ is the test sample genotype using an RF τ with *D* decision trees, for which ψ_k is the k^{th} tree. An RF has a number of hyperparameters that need to be tuned, for which we used grid search using the caret v6.0.92 R package (Kuhn, Wing et al. 2020). We used 500 trees in the forest for all analyses and tuned 'mtry' and 'nodesize' hyperparameters to control tree shapes. The total number of SNPs randomly selected at each tree node, i.e. mtry, was selected from { *p*/3, *p*/4, *p*/5, *p*/6} and the minimum size of terminal nodes below which no split can be tried, i.e. nodesize, was selected from {0.01, 0.05, 0.1, 0.2, 0.3} times the number of training samples in each cross-validation fold.

Extreme Gradient Boosting (XGBoost)

We used XGBoost, a specific implementation of the Gradient Boosting (GB) method. Similar to the Random Forest, Gradient Boosting is an ensemble method, using weak learners such as DTs. The main difference is that an RF aggregates independent DTs trained on random subsets of data (bagging), whereas GB grows iteratively (boosting) by selecting samples in the subsequent DTs based on sample weights obtained in previous DTs, related to how well samples are predicted already by these previous DTs.

Hyperparameters were tuned using a grid search through five-fold cross-validation on each training data fold. We searched over max_depth = {2, 3, 4, 50, 100, 500}, colsample_bytree = {0.1, 0.2, 0.3, 0.5, 0.7, 0.9} and subsample = {0.7, 0.8, 0.9}.

2.2.5 Performance evaluation

Model performance was evaluated based on prediction accuracy, which was measured as the Pearson correlation coefficient (*r*) between observed phenotypic values and predicted genomic values of the test population. For each model, five repeats of five-fold cross-validation were performed, so in total 25 values of *r* were used to compare performances. Statistical comparison between different models was performed by comparing prediction accuracies of each pair of models as a whole, i.e. on all values of *p/n* together using Wilcoxon rank-sum test.

2.2.6 Assessment of trait non-additivity

To link GP performance in simulation scenarios with performance on real data, an assessment of the nature of real traits (i.e. additive or epistatic) was used. To obtain a proxy for additivity of the trait, we assumed that if a trait has a higher proportion of additive variance compared to other traits, estimated with the same model, it will be more additive. To verify this on our simulated dataset scenarios (*Table 2.1*) for epistatic phenotypes, we used the linear mixed model:

$$\mathbf{y} = \boldsymbol{\mu} + (\mathbf{g}_a + \mathbf{g}_e) + \boldsymbol{\varepsilon} \qquad (2.9)$$

Here $\mathbf{g}_a$ defines a set of additive genotype effects such that $\mathbf{g}_a \sim N(0, \sigma_a^2\, \mathbf{G})$, where $\mathbf{G}$ is the genomic relationship matrix (GRM) calculated as described by VanRaden (2008). Moreover, $\mathbf{g}_e \sim N(0, \sigma_e^2\, \mathbf{E})$ is a vector of epistatic genetic effect and $\boldsymbol{\varepsilon}$ is a vector of residuals. Here, $\mathbf{E}$ is the GRM ($\mathbf{G}$ o $\mathbf{G}$). The ratio of additive genetic variance to the epistatic genetic variance (σ_a^2/σ_e^2) was calculated for both the simulated dataset and real wheat traits to assess their relative non-additivity. We tested our assumption on simulated phenotypes (Extended data, *Figure S1*), showing simulated amounts of non-additive heritability to indeed be negatively related to empirical additive heritability.

2.3 Results

2.3.1 ML outperforms traditional methods for GP

Previously, numerous GP methods were tested for different traits of varying genetic architectures using low or high density marker sets, but it is still unclear for which (class of) GP problems applying machine learning (ML) can be beneficial (Pérez-Rodríguez, Gianola et al. 2012). To investigate the role of underlying characteristics (*Figure 2.1*), we generated an extensive set of simulated genotype-phenotype data (simdata: see 2.2.1). This data was analysed using the linear parametric methods GBLUP, BayesA and BayesB; the nonlinear semi-parametric regression method RKHS, using a Gaussian multi-kernel evaluated as average squared-Euclidean distance between genotypes (De los Campos, Gianola et al. 2010); and popular nonlinear ML methods, i.e. support vector regressor (SVR), random forest regressor (RF), extreme gradient boosting (XGBoost) regression trees and a fully-connected feed forward artificial neural network i.e. Multilayer Perceptron (MLP). The simulations covered a variety of trait scenarios (from simple to more complex), as shown in *Table 2.1*. Simple oligogenic traits correspond to simulation scenarios with larger heritabilities, additive allele effects and small numbers of QTNs; complex traits can have both additive and non-additive allele effects (only epistatic here) with small heritabilities and large numbers of QTNs. For additive phenotypes, narrow-sense heritability was set equal to broad-sense heritability and for the epistatic phenotypes, the sum of narrow-sense and epistatic heritability was set equal to the broad-

sense heritability. The extent of phenotypic additivity in both simulations and real datasets was calculated using the ratio of additive genetic variance to the epistatic genetic variance (σ_a^2/σ_e^2) using equation (9). In the results presented below, SVR and MLP were excluded because their performances were significantly lower than the tree-based ensemble ML methods (i.e. RF and XGBoost) on a subset of our simulation scenarios (Extended data, Appendix I). Moreover, the applicability of neural networks/deep learning for GP in the feature space is still limited due to their high tendency toward overfitting under high-dimensionality until they are properly regularized or feature selection is employed (Azodi, Bolger et al. 2019, Montesinos-López, Martín-Vallejo et al. 2019, Montesinos-López, Montesinos-López et al. 2021).

2.3.2 ML methods perform well for simple traits

Many non-mendelian plant traits are fairly simple, where only one or a few QTLs explain a large proportion of phenotypic variance, called oligogenic traits. If these QTLs are identified by the GP model, prediction performance can be pretty high. In our simulations (*Table 2.1*), this scenario is investigated using additive phenotypes with narrow-sense heritability (h^2) equal to 0.7 and a total number of QTNs equal to 5. We then alternatively attribute equal effects to all QTNs, assign a larger effect to the first QTN in equation (2.1) compared to other the QTNs, or sample the QTN effects from a Gaussian distribution (see Section 'Simulations').

The results in *Figure 2.2A*, *Figure 2.2B* and *Figure 2.2C* illustrate that the performance of Bayesian methods and ML was significantly better (p value < 0.01; Extended data, Table S1) than that of genomic relationship-based methods (GBLUP, RKHS). The performance of ML methods was slightly poorer than that of Bayesian methods when all QTNs effects were equal (*Figure 2.2A*) or sampled from a Gaussian distribution (*Figure 2.2C*) but comparable when one of them had a larger effect size (*Figure 2.2B*). Therefore, although not outperforming the other methods, ensemble ML methods seem to be reasonable choices for simple traits.

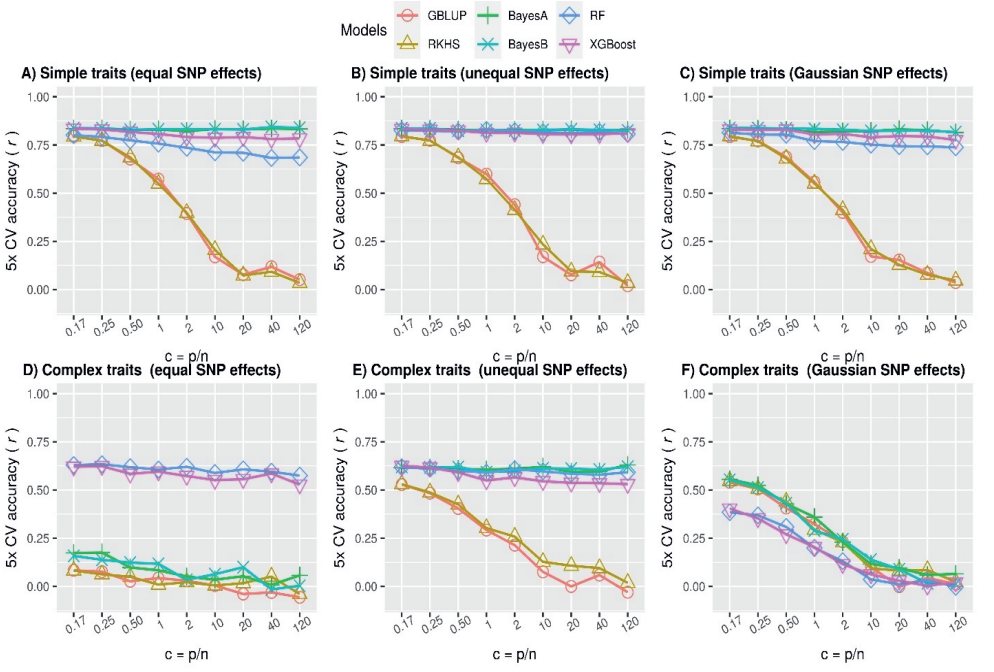


Figure 2.2: Comparison of prediction performances using simulated simple and complex phenotypes.

Performance of parametric (GBLUP), semi-parametric regression (RKHS), parametric Bayesian (BayesA, BayesB) and nonparametric ML (RF and XGBoost) methods as average accuracy over *5*-fold cross-validation of test data. Here accuracy is defined as Pearson correlation coefficient between true and predicted values. Each panel is a subset of the simulated scenarios in 'simdata' for a particular heritability and #QTNs. The ratio of the number of markers to the number of samples (*c* = *p*/*n*) increases from left to right in each subplot. A) Simple traits, simulated as polygenic phenotypes with only additive effects such that #QTNs is equal to 5 and h^2 is 0.7, using equation (2.1), with all QTNs having equal effects. The largest standard error of mean for all values of *c* for each of the model was 0.023, 0.018, 0.007, 0.008, 0.018 and 0.009 for GBLUP, RKHS, BayesA, BayesB, RF and XGBoost respectively; B) similar to A, except one of the QTN had a large effect than others. The largest standard error of mean for all values of *c* for each of the model was 0.022, 0.022, 0.006, 0.007, 0.006 and 0.008 for GBLUP, RKHS, BayesA, BayesB, RF and XGBoost respectively; C) similar to A and B, except QTN effects were sampled from a Gaussian distribution; D) Complex traits, simulated as polygenic phenotypes with both additive and epistatic effects such that #QTNs equal to *p*/2 and h^2 is equal to 0.4, using equation (2.2), such that all QTNs had equal additive effects. Two of the QTNs were attributed to the epistatic effect such that Broad-sense heritability was set to 0.8 ($H^2 = h^2 + h_e^2 = 0.8$). The largest standard error of mean for all values of *c* for each of the models was 0.03; E) similar to D, except one of the QTN had a large effect than others; F) similar to D and E, except QTN effects were sampled from a Gaussian distribution (see methods).

2.3.3 ML methods outperform parametric methods for complex traits

Complex polygenic traits may contain a large effect QTL along with many small to medium effect QTLs (Goddard, Kemper et al. 2016). Despite assuming perfect LD between SNPs and their corresponding QTLs, their detection remains challenging through conventional univariate regression models that are followed by strict multiple testing corrections. Moreover, shrinkage of random effects towards zero in multivariate regression models restricts them from growing too large. Thus, many true small effects may be ignored in the analysis. SNPs may also have non-additive effects, which could cause a large amount of variance to remain unexplained and narrow-sense heritabilities to be low, when modelled by their additive action only.

This genetic complexity was simulated by increasing the number of QTNs, decreasing the narrow-sense heritability and keeping overall effect sizes equal, thereby letting the effect sizes per QTN become proportionally smaller. The QTNs were randomly chosen from the simulated SNPs pool by setting *k* equal to half of the total number of SNPs (*p*/2) in equation (2.2), keeping equal effect sizes for all QTNs and h^2 equal to 0.4. Moreover, similar to simple traits, the other two scenarios, i.e. unequal effect sizes and normally distributed effect sizes, were also simulated. Two QTNs were randomly selected to have a fairly large pairwise interaction effect, corresponding to an epistatic heritability h_e^2 equal to 0.4. The results in *Figure 2.2D* illustrate that ML methods significantly outperformed all methods for complex phenotypes when all of the QTNs had equal effects (*p-value* < 0.01; Extended data, Table S2). Interestingly, when one of the QTN had a larger effect size or was attributed with most of the variance, the Bayesian methods performed on par with ML (*Figure 2.2E*), but when the effect sizes followed a Gaussian distribution (*Figure 2.2F*), ML was outperformed by the other methods. This confirms that parametric methods work well if the effects distribution matches the statistical prior assumptions. In reality, genetic variance may not be attributed to a single Gaussian for other than infinitesimal model, instead it could be decomposed into multiple distributions enriched in multiple chromosomal localisations defined by heritability models (Speed, Holmes et al. 2020). This phenotype complexity is usually unknown and difficult to accurately assess, which provides room for the ML methods.

2.3.4 ML methods are generally suitable for epistatic phenotypes

For complex phenotypes, we observed that ML outperformed LMMs under highly polygenic phenotypes with epistatic effect and equivalent to Bayesian LMMs when at least one QTN had larger effect (*Figure 2.2D* and *E*). To explore further, we investigated a range of additive and non-additive fractions of heritabilities, with or without a large effect QTN and from Gaussian distribution defined in our simulation scenarios (*Table 2.1*).

For additive phenotypes with equal QTN effect sizes, performance of ML methods was poorer than that of Bayesian methods under all scenarios; with an increase in genetic complexity (lowering h^2 and increasing the number of QTNs), performance dropped below

that of GBLUP and RKHS as well (Extended data, *Figure S2A*). Therefore, ML methods are not beneficial for this setting. For epistatic phenotypes however, ML outperformed all methods including the Bayesian methods for all scenarios (Extended data, *Figure S2B*), with random forests generalizing the best. ML methods are thus best suited for epistatic traits and do not necessarily need large main effects to be present. Note that although RKHS regression has been reported to better capture epistatic relationships between markers (Jiang and Reif 2015), it did not perform well in our simulations; perhaps it needs more careful tuning of the bandwidth of the Gaussian distributions, rather than using multi-kernel averaging or require matching prior allele effects distributions (see Discussion, 'Tree-based ensemble ML methods are a reasonable choice for GP').

For the phenotypes explained by a large effect QTN and many small effect QTNs (Extended data: *Figures S3A* and *S3B*), Bayesian methods perform comparable to ML methods for both additive and epistatic phenotypes under all simulation scenarios, although RF gave slightly better performance for epistatic phenotypes with large epistatic heritability (for $h_e^2 = 0.7$) and dimensionality (*p/n* > 2). This could be because the large effect QTN explains most of the additive variance and is easily picked by Bayes and ML methods, but RF has the added advantage of picking up the nonlinear signal, when main effects got smaller with the increase in number of QTNs. XGBoost gave relatively poor performance, especially at smaller heritabilities (0.1 and 0.4) and larger *p*/*n* ratios, while GBLUP and RKHS regression performance was consistently poor in all scenarios.

For both additive and epistatic phenotypes (Extended data: *Figures S4a*, and *S4b*), the ensemble ML methods were still superior over BLUPs and comparable to Bayes when effect sizes were sampled from a Gaussian distribution for a small number of QTNs (e.g. *q* = 5, h^2= 0.7, h_e^2 = 0.1), but the advantage diminishes when *q* increases and approaches the infinitesimal model i.e. *q*=*p.*

In conclusion, our simulation results indicate that ML works well when a fair proportion of broad-sense heritability is contributed by allele interaction effects or a few large effect QTNs.

2.3.5 ML performance is robust to high-dimensional GP

Genomic prediction is usually employed on a genome-wide set of markers to yield total genomic value, but the training population size is limited, i.e. a high dimensional problem. This results into more statistical power to detect QTLs with many SNPs in LD but comes with obscured genetic variance when added together. Consequently, it leads to an overestimation of allelic variances or genomic relationships, overfitting on training samples and reduced performance on unseen data. To investigate the susceptibility of different GP methods for this issue, we analysed how prediction accuracy varied depending on the ratio of markers vs samples (c = *p*/*n* > 1).

In general, the results with different simulation settings of 'simdata' for additive phenotypes show that performance is negatively related to an increase in dimensionality when main effects got smaller due to decreasing heritability or increasing total number of QTNs (Extended data: *Figure S2A*, *Figure S3A* and *Figure S4A*). This implies that for simple traits having one or few large effect QTNs (*Figure 2.2A* to *C*), performance degradation is not a severe issue for Bayesian and ML methods but it can still be a potential problem for genetic distance-based methods i.e. GBLUP and RKHS., presumably because of increased uninformative markers in calculating the genetic kinships. For the epistatic phenotypes, high dimensionality still doesn't affect ML until we have sufficiently large main effects (*Figures 2.2A* and *2B*; Extended data: *Figure S2B*, *Figure S3B* and *Figure S4B*). Here, for the case when main effects were sampled from a Gaussian distribution, increasing polygenicity is analogous to having many small main effects; so, despite having epistatic effects, performance goes down for all methods. In the nutshell, this shows that the conclusions drawn in Section 'ML methods perform well for simple traits' and Section 'ML methods outperform parametric methods for complex traits' holds under high-dimensionality.

2.3.6 Case study in wheat

To see whether our simulation results hold on real traits, we used a dataset of 13 wheat traits (Norman, Taylor et al. 2017) for a fairly large number of samples (10,375 lines) and 17,181 markers ($c \approx 1.6$). These markers have been selected by strict screening criteria, therefore, many of them could be informative. Earlier, insights of the genetic complexity for some of these traits have been reported (Norman, Taylor et al. 2017, Norman, Taylor et al. 2018). For example, glaucousness was reported to be a simple trait, but grain yield to be more complex (Norman, Taylor et al. 2018). The results in *Figure 2.3* clearly indicate that five-fold cross-validated prediction accuracies (r) were higher for both ML methods when the fraction of additive variance was small (i.e. traits were fairly non-additive) and slightly lower or comparable to both Bayesian and GBLUP/RKHS regression methods otherwise. This is in line to what we observed in our simulations: for simple traits (*Figure 2.2A* and *B*) ML performance was either comparable to Bayesian or slightly poorer, but for complex traits it was consistently better (*Figure 2.2C*). For example, leaf width, glaucousness, growth habit, leaf loss, plant height, test weight and thousand kernel weight traits had greater than 80% of their genetic variance explained only by additive variance components and performance of ML relative to Bayesian methods and GBLUP/RKHS regression was either at par or lower than that. On the other hand, biomass, grain protein, grain yield, yellowness and in particular NDVI had smaller fractions of additive variance and, relative to the other methods, ML performed better. Hence, results on this experimental dataset match with the findings in our simulations that ML is best suited for the prediction of more complex traits and a potential candidate for simple traits as well.

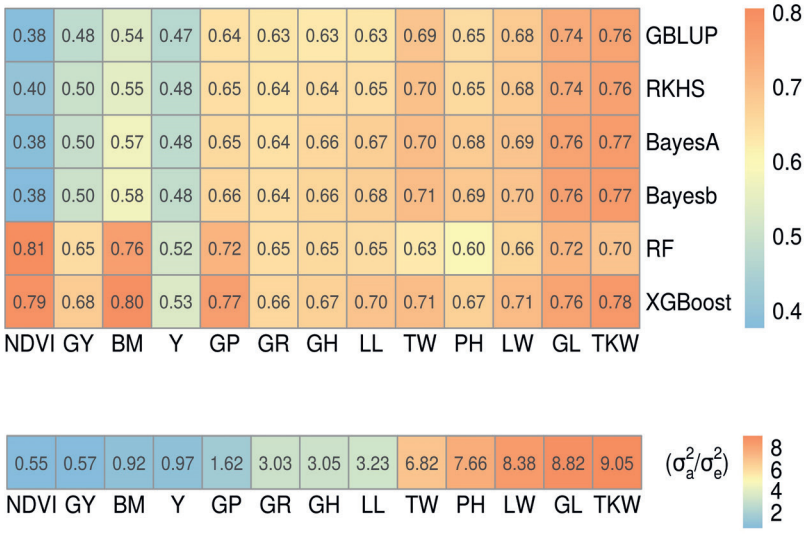


Figure 2.3: Prediction accuracies of wheat traits.

Top: prediction accuracies for GP models on wheat traits, reported as the mean Pearson correlation coefficient (*r*) of 5-fold cross-validation. Trait abbreviations are given in *Table 2.2*. Bottom: fraction of additive to epistatic genetic variance calculated using equation (2.9) for each trait. Traits were sorted in ascending order of the fraction σ_a^2/σ_e^2 (left to right).

2.3.7 ML methods are sensitive to population structure

Population structure (PS) is a well-known confounding factor that results in decreased diversity in training populations (Norman, Taylor et al. 2018) and unrealistic inflated parameter estimates, e.g. for (co) variances of random effects in LMMs (Visscher, Hemani et al. 2014). Parametric and nonparametric ML methods, based on their modelling assumptions and approaches, may be differently sensitive to PS. To assess the impact of population structure on ML methods, we used real genotype data with a known population structure and combined it with both simulated (STRUCT-simdata) and real phenotypes (STRUCT-realdata). Only additive phenotypes were simulated, with varying complexity and dimensionality scenarios, as described earlier in Section 'Simulations'. The STRUCT-simdata contains all 1,307 *Arabidopsis* RegMap accessions (Horton, Hancock et al. 2012). To exclude the impact of multicollinearity among SNPs, only uncorrelated markers were retained after pruning with pairwise squared correlation coefficient ($r^2 < 0.1$, see Section 'Population structure analysis'), leaving 15,662 SNPs, but keeping the population structure intact (Extended data, *Figure S7*). This results in a ratio $c = p/n$ of approximately 12 (15,662/1,307), a setting comparable to the simulation results presented in *Figure 2.2A.*

Correction for PS was carried out by including the top ten principal components corresponding to the largest eigenvalues as fixed effects into the mixed model equations or as additional features for ML methods. For the simulated phenotypes (Extended data, *Figure S6*), average pairwise difference of test accuracies before and after correcting for PS was slightly higher for ML methods (RF: 0.03 and XGBoost: 0.04) than for LMMs (GBLUP: 0.01, RKHS: 0.01, BayesA: 0.01 and BayesB: 0.00). Moreover, the correction resulted into relatively elevated accuracies for the scenarios with larger number of QTNs or low heritabilities. This illustrates that with smaller #QTNs and larger heritabilities (h^2= 0.7, #QTNs = 5), effect sizes per QTN were larger; therefore, confounding due to PS was less of a concern. With the decrease in effect sizes per QTN (increase in #QTNs and decrease in h^2), correction became more important for reliable predictions. From this, we can argue that confounding due to PS should be generally corrected for, but particularly for complex phenotypes having low heritability and large numbers of QTNs with small-medium effect sizes.

To further explore this behaviour, we used real phenotypes of the sodium accumulation trait in *Arabidopsis thaliana* (STRUCT-realdata) using a subset of the same genotypes dataset. Here, we expected to have at least one large effect QTN for this trait, because *AtHKT1;1* locus, encoding a known sodium (Na^+) transporter, has been reported to be a major factor controlling natural variation in leaf Na^+ accumulation capacity (Baxter, Brazelton et al. 2010). Similar to the outcomes on 'STRUCT-simdata', correction for PS increased prediction accuracies of all methods on test data; whereas, GBLUP showed the lowest average difference ($\Delta\mu$ = 0.03) in performance before and after correction (*Figure 4*). In contrast to 'STRUCT-simdata', XGBoost had the largest average difference ($\Delta\mu$ = 0.1) but for RF the difference was comparable to LMMs ($\Delta\mu$ = 0.05). From the above outcomes, we conclude that ML methods, like other GP methods, are sensitive to confounding due to PS and correcting for this can further improve performance for complex phenotypes. However, it is still unclear to which extent or for which GP problem characteristics different methods are more advantageous or more sensitive to PS.

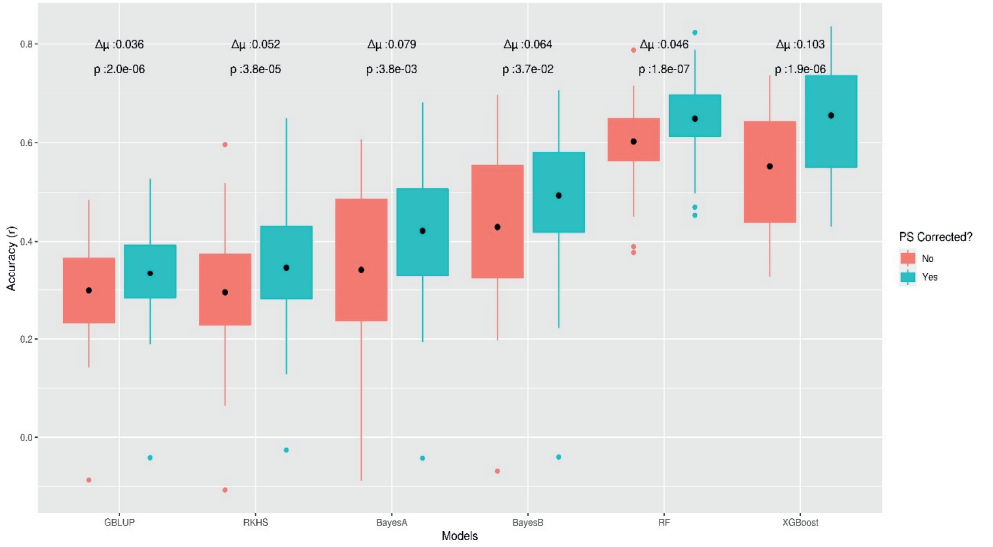


Figure 2.4: Effect of correction for population structure for the sodium accumulation trait in Arabidopsis thaliana.

Boxplots present Pearson correlation coefficients (*r*) found in *5*-fold cross-validation, on test data from 'STRUCT-realdata'. Here Δμ is the average difference between pairwise predictions before and after correction and for each model, the nonparametric Wilcoxon rank sum test was used to assess statistical significance.

2.3.8 ML methods can tackle low SNP-QTN LD

The utility of GP in genomic selection is based on the assumption that there are ample markers within a densely genotyped set of markers which are in LD with the QTLs (Meuwissen, Hayes et al. 2001). The actual QTNs are generally unknown, but SNPs in LD can be used to (partially) capture their effect, depending on the actual correlation and allele frequencies. Therefore, it is worthwhile to investigate the impact of SNP-QTN correlation levels on GP performance (Uemoto, Sasaki et al. 2015). We used two settings, one with real genotypes and simulated phenotypes (LD-simdata), a second with real genotypes and real traits (LD-soy).

In simulations, GP model performance is evaluated based on the difference in prediction accuracies between a model trained on the actual QTNs and a model trained on SNPs in LD (QTN-linked SNPs). Our results show that when SNPs are highly correlated to QTNs (which is likely the case for densely genotyped markers set and $r^2 > 0.9$), all methods perform equally well and the SNP-based model predictions are very close to those of the actual QTN based models (Extended data, *Figure S8*). On the other hand, for low LD between SNPs and QTNs, there was in general a difference between median prediction

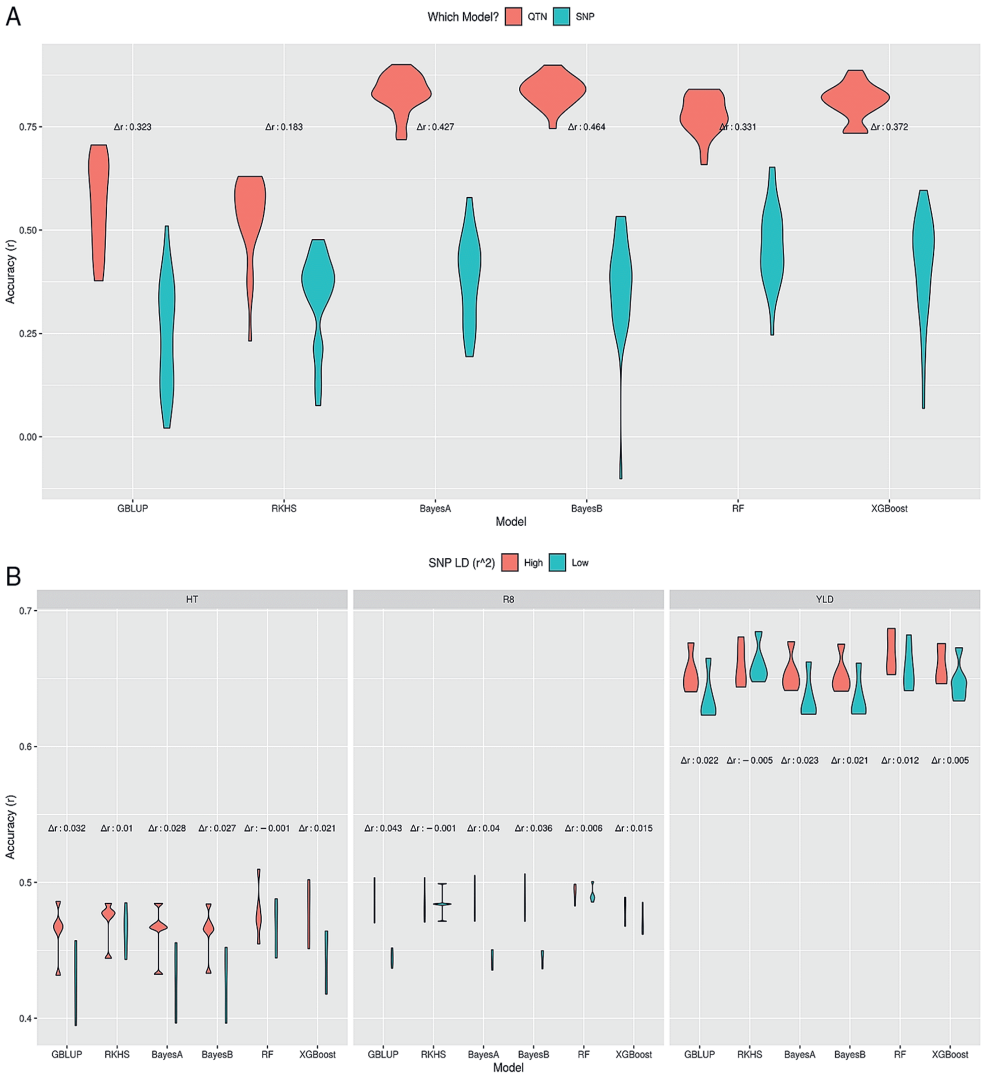


Figure 2.5: Effect of SNP-QTN LD on prediction accuracy.

Prediction accuracy of different GP methods on simulated (A) and real soybean (B) datasets for high and low LD between SNPs and actual QTNs. The difference in median accuracies between these scenarios is indicated as Δ*r*. A) LD-sim data, low SNP-QTN LD ($r^2 \leq 0.5$). B) LD-soy data, low ($r^2 \leq 0.5$) SNP-QTN LD vs. all SNPs (high LD).

accuracies (Δ*r*) of the QTN and SNP-based models (*Figure 2.5A*). This difference varied between methods, from 0.18 for RKHS regression to 0.43–0.46 for the Bayesian methods, with GBLUP and ML methods between these (0.32–0.37). The relative robustness of particularly the Random Forest model in these circumstances compared to the Bayesian methods, in combination with its good performance in many simulations,

supports its usefulness for GP.As a real genotype and phenotype dataset, we used three Soybean traits, i.e. height, time to R8 developmental stage and yield (LD-soy). The motivation was to choose a low-dimensional real dataset with highly correlated SNPs to understand the impact of SNP-QTL LD only. The complete set of markers (4,235 SNPs) had many correlated SNPs, such that only 261 were left with low LD ($r^2 \leq 0.5$). Here, in contrast to LD-simdata, where we knew the QTNs in advance, we assumed that many SNPs could be linked to QTNs, because ~94% of all markers had $r^2 > 0.5$. So, we compared two models: one with all markers (the benchmark model), and one with low LD ($r^2 \leq 0.5$). A similar pattern was observed, as shown in *Figure 2.5B*, i.e. RKHS regression, RF and XGBoost were most robust against low SNP-QTN LD, with negligible differences between median accuracies, where GBLUP and the Bayes methods had higher differences. Moreover, the prediction accuracies were similar to previously reported values for these traits (Azodi, Bolger et al. 2019).

In conclusion, GP methods that model SNP-QTN or SNP-SNP relation as a nonlinear function (RKHS, RF, XGBoost) were more stable under low SNP-QTN LD compared to other methods (GBLUP, BayesA, BayesB). Moreover, RF seems to couple good prediction performance with reliability under low SNP-QTN LD.

2.4 Discussion

2.4.1 There is room for ML in genomic prediction

Genomic prediction has long been the realm of parametric methods, but recently nonlinear supervised ML methods have become increasingly popular. Yet literature is unclear on the characteristics of GP problems that warrant application of ML methods. This study fills this gap and concludes that nonlinear tree-based ensemble ML methods, especially Random Forests, can be a safe choice along with traditional methods for simple as well as complex polygenic traits where epistatic allele interaction effects are present. We simulated different scenarios mimicking the reality at a broader level e.g. the case of simple oligogenic traits (*Figure* 2.2, panel A, B & C), where they outperformed BLUPs but not Bayesian LMMs. A similar trend can be observed in real data of Sodium accumulation trait (*Figure* 2.*4*), where we studied the impact of LD. On the other hand, for complex traits scenarios (*Figure* 2.*2*, panel D, E, F), Random Forests either outperformed when again a large effect was present (panel D &E) or were inferior to other methods, when all effects followed the Gaussian distribution (panel F). The latter (panel F) is prevalently observed for many complex traits, where accuracies are roughly comparable for all methods. Moreover, ML methods are robust to high dimensionality, although further improvements, e.g. statistical or prior knowledge driven regularization, may improve performance. ML methods are particularly useful compared to the frequently used GBLUP and RKHS regression given their higher performance. While Bayesian methods often perform on par with ML models, this is mainly when there are large effect QTNs and/or additive phenotypes. Moreover, Bayesian methods are prone to overfitting

in case of small sample sizes ($p/n > 1$), which is less of an issue with ML, especially with RF (Extended data: *Figures S9A* and *S9B*).

2.4.2 Tree-based ensemble ML methods are a reasonable choice for GP

A wide range of parametric, semi-parametric and nonparametric methods can be used for GP, but it is impractical to test all for a particular application. The choice for a suitable method strongly depends on the GP problem characteristics, described in *Figure 2.1*. While GP methodology can be compared using various model evaluation metrics (BIC, AIC, log likelihoods), we focused on their utility from a breeder's perspective, so we compared only their prediction accuracies. We found that GP methods based on modelling the distance between genotypes using covariance structure(s), inferred from genomic markers (GBLUP and RKHS), were generally inferior to Bayesian and ML methods and less robust to high-dimensional problems likely because all of the p SNPs were used always to calculate the kinship matrices, whereas, either 5, 50, 100, $p/2$ or exactly p SNPs were chosen as QTNs. When q is fairly less than p, makes the kinship matrix too noisy due to the large number of markers that are unrelated to the phenotype but are used in the calculation of the GRM. Hence, we expect equal accuracies for increasing number of QTNs (q), keeping the other factors (p, n and h^2) fixed. *Figure S5* (Extended data) clearly illustrates that these methods indeed have constant prediction accuracies with increasing q values, while the accuracies of the other methods drop due to decreasing effect sizes. This further explains that their performance can be improved by removing unrelated markers from the GRM, for instance using biological knowledge about markers (Zhang, Liu et al. 2010, Farooq, van Dijk et al. 2020).

The parametric LMM equations can be solved using a Bayesian framework. Bayesian methods define prior SNP effects distributions to model different genetic architectures. Instead of a single distribution for all marker effects (e.g. BRR), it could be defined for each individual marker (e.g. BayesA). Mixture distributions have also been proposed (e.g. BayesC, BayesB). From the Bayesian alphabet, we used BayesA and BayesB as representatives because the first scenario, i.e. a single distribution for all markers, has been covered by GBLUP. Our results illustrate that these methods outperform GBLUP and RKHS regression when large effect QTNs are present, for both additive and epistatic phenotypes. On the other hand, tree-based ensemble ML methods had either comparable performance to Bayesian methods (for simple traits) or superior performance (for complex traits). Capitalising on the results from Appendix-I (Extended data) that these ML methods had better performances than other ML methods (SVR and MLP), we can argue that these tree-based ML methods are a reasonable choice to conduct GP.

2.4.3 Population structure analysis

Population structure can affect GP performance. Our results show that without correcting for population structure, test accuracies were lower than after correction for all methods. However, ML seems to be slightly more sensitive because the average difference

between each pairwise test data accuracies was higher than other methods in the simulated data.

Confounding due to population structure can also be due to the frequently employed random cross-validation strategy for predictive modelling (Norman, Taylor et al. 2018). In random cross-validation, the reference population is randomly divided into subsets, one of which is iteratively selected for testing while the remaining subsets are used to train the model. While samples are all part of a test set once, under population structure some subpopulations may be over or under-represented in the training set. As a result, the model may get overfitted. A solution could be to use stratified sampling instead. On the other hand, parameter estimation may get misguided by within subgroup allele frequency differences rather than the overall true phenotype associated variance.

The impact of population structure can be dealt with in many ways. Conventionally, principal components of the SNP dosages or genomic relationship matrix are introduced as fixed effects in the mixed model equations (Guo, Tucker et al. 2014, Bermingham, Pong-Wong et al. 2015, Zhang, Wang et al. 2017). Alternatively, phenotypes and genotypes can be adjusted by the axis of variations before predictive modelling (Zhao, Chen et al. 2012). Nevertheless, some residual structure often remains in the datasets, so it is important to check sensitivity of GP models to this confounding factor. Since ML methods (RF and XGBoost) do not employ any statistical prior and learn the association patterns from the data itself, they may be more sensitive to structure, as we found in our simulation results. But this is not clearly evident from the real phenotypes, so we cannot generalize this conclusion from our simulations.

2.4.4 Effect of SNP-QTN linkage disequilibrium

Despite technological improvements, low density SNP panels are usually cost-effective for routine genomic selection. Increasing marker density does not necessarily increase prediction accuracy, since accuracy is not a linear function of SNP density only (Technow, Riedelsheimer et al. 2012, Wang, Yu et al. 2017, Zhang, Wang et al. 2017). Instead, many GP problem characteristics (*Figure 2.1*) jointly affect performance. However, using low density SNP panels can negatively affect prediction performance, since relevant SNPs in LD with the QTLs can either be completely missing or SNPs only in low LD may be present. As a result, allele frequencies between SNPs and QTNs can be quite different, resulting in incorrect predictions (Uemoto, Sasaki et al. 2015). Despite this, low SNP density can still be sufficient for populations with larger LD blocks, e.g. F_2 populations, where QTL detection power is highest and in this case, we shouldn't expect much improvement by increasing marker density. But it becomes an important consideration when LD starts to decay and population relatedness decreases in the subsequent crosses of the breeding cycle. In this context, our study addresses the question of whether certain GP methods, especially ML, are more sensitive to low SNP-QTL LD. The results using

both simulated and real traits indicate that SNP-QTL LD could also be an important determinant of suitable GP methodological choice and that ML is robust against low LD.

A weak SNP-QTL correlation implies that the SNP is a weak predictor of phenotype and there is an imperfect match between the genotypic distribution and the actual underlying genetic distribution of the phenotype. When using penalized regressions, this can result in different shrinkage for the SNP than that required by the actual QTN, thereby leading to a low genetic variance attribution to that SNP. Therefore, we may expect better prediction by nonparametric ML methods, as they may better learn weak genetic signals and are more robust to low SNP-QTL LD problems. On the other hand, the semiparametric RKHS regression method, which measures genetic similarity between individuals by a nonlinear Gaussian kernel of SNP markers, also performed better than GBLUP and Bayesian methods under low SNP-QTN LD. The reason could be that under low SNP-QTN LD, true pair-wise genetic covariance estimation would be less accurate due to losing many important markers and considering all of them equal contributors towards total genetic covariance. In case of RKHS regression, a Gaussian distribution defines a SNP's probable contribution towards total genetic covariance, which becomes more realistic in this scenario because fewer important SNPs are left than in the high SNP-QTN LD case. The Bayesian methods (BayesA and BayesB) had the largest decrease in test performance under low SNP-QTN LD compared to high SNP-QTN LD. This could be due to the application of penalties on individual marker effects, which shrinks the weak SNP-QTN associations towards zero for each SNP.

2.4.5 ML outperformed parametric methods for predicting complex wheat traits

Bread wheat breeding has huge impact on worldwide food security and socio-economic development (Tessema, Liu et al. 2020). Therefore, minor improvements in GP methodology leading to overall genetic gain can have high impact. In this study, we used a large (10,375 lines) Australian germplasm panel, genotyped with a high quality custom Axiom^{TM} Affymetrix SNP array and phenotyped for multiple traits with varying complexity levels (Norman, Taylor et al. 2017). The authors showed that genomic selection was superior to marker-assisted selection (MAS) by employing GBLUP with two random genetic components (referred to as full-model in their text). Our results clearly indicate that ML can perform well for complex bread wheat traits, e.g. grain yield, yellows, greenness, biomass and NDVI. However, for NDVI, the larger difference between LMMs and ML could be due to low phenotypic variance and heritability for this trait in this dataset. All of these traits except grain yield can be measured using high-throughput automated phenotyping (Rabab, Breen et al. 2021). This is an interesting finding since, with the rapid advances in low cost high-throughput phenotyping systems, attention is shifting towards measuring component traits, e.g. vegetative indices, rather than final yields. ML methods can predict these traits more accurately, as evident from our analysis.

2.5 Conclusions and outlook

Based on simulated and real data, we conclude that tree-based ensemble ML methods can be useful for GP for both simple and complex traits. Moreover, these methods can work for both low- and high-density genotyped populations and a competitive choice for practical plant breeding. In practice, which method works best depends on the particular problem determined by genetic architecture of the trait, population size and structure and data dimensionality. Between bagged (Random Forests) or boosted (XGBoost) decision tree ensemble methods, Random Forest seems to be a good first choice for GP given their generalization performance. Furthermore, population structure should properly be corrected to obtain stable performance. It would be interesting to investigate to what extent these ML methods can benefit from statistical or prior knowledge-based regularization techniques.

2.6 Data availability

2.6.1 Underlying data

All datasets analysed during the current study are already published and publicly available and references to their authors or repositories have been mentioned in the text.

2.6.2 Extended data

The Extended data has been deposited to Figshare, and available as: Extended data for 'Genomic prediction in plants: opportunities for ensemble machine learning based approaches'.

This project contains the following Extended data:

i. Supplementary figures: Farooq, Muhammad (2022): Supplementary figure V2. figshare. Figure. https://doi.org/10.6084/m9.figshare.21705944.v1
   - *Figure S1*. Assessment of phenotypic class (additive or epistatic).
   - *Figure S2*. Comparison of test data prediction performance using simulated phenotypes with equal effects QTNs.
   - *Figure S3*. Comparison of test data prediction performance using simulated phenotypes with unequal effects QTNs.
   - *Figure S4*. Comparison of test data prediction performance using simulated phenotypes with QTN effects sampled from Gaussian distribution.
   - *Figure S5*. Effect of increasing number of QTNs to the total number of SNPs ratio on prediction performances using simulated additive phenotypes phenotypes.
   - *Figure S6*. Effect of population structure correction on GP model accuracies.
   - *Figure S7*. Principal Component Analysis (PCA) of *Arabidopsis thaliana* RegMap 1,307 accessions using uncorrelated set of markers.

- *Figure S8*. Effect of high SNP-QTN LD (r^2>0.9) on prediction accuracy.
- *Figure S9*. Comparison of training data prediction performances using simulated phenotypes with one large effect QTN.
- *Figure S10*. Comparison of prediction performances of parametric, semi-parametric and ML methods using simulated phenotypes without a large effect QTN for epistatic phenotypes.

ii. Supplementary tables: http://www.doi.org/10.6084/m9.figshare.19918729
- *Table S1*. Simple Traits
- *Table S2*. Complex Traits

iii. Appendix 1: Selection of machine learning (ML) candidates for genomic prediction. http://www.doi.org/10.6084/m9.figshare.19919023

2.6.3 Software availability

Source code available from: https://git.wur.nl/faroo002/pub2

Archived at the time of publication: https://doi.org/10.5281/zenodo.6734259

License: GPL version 3

CHAPTER

3

Prior biological knowledge improves genomic prediction of growth-related traits in *Arabidopsis thaliana*

Muhammad Farooq, Aalt D.J. van Dijk, Harm Nijveen, Mark G.M. Aarts, Willem Kruijer, Thu-Phuong Nguyen, Shahid Mansoor and Dick de Ridder

This chapter was published as:

Farooq, Muhammad, et al. "Prior biological knowledge improves genomic prediction of growth-related traits in Arabidopsis thaliana." Frontiers in Genetics 11 (2020): 609117.

Abstract

Prediction of growth-related complex traits is highly important for crop breeding. Photosynthesis efficiency and biomass are direct indicators of overall plant performance and therefore even minor improvements in these traits can result in significant breeding gains. Crop breeding for complex traits has been revolutionized by technological developments in genomics and phenomics. Capitalizing on the growing availability of genomics data, genome-wide marker-based prediction models allow for efficient selection of the best parents for the next generation without the need for phenotypic information. Until now such models mostly predict the phenotype directly from the genotype and fail to make use of relevant biological knowledge. It is an open question to what extent the use of such biological knowledge is beneficial for improving genomic prediction accuracy and reliability.

In this study, we explored the use of publicly available biological information for genomic prediction of photosynthetic light use efficiency (Φ_{PSII}) and projected leaf area (PLA) in *Arabidopsis thaliana*. To explore the use of various types of knowledge, we mapped genomic polymorphisms to Gene Ontology (GO) terms and transcriptomics-based gene clusters, and applied these in a Genomic Feature Best Linear Unbiased Predictor (GFBLUP) model, which is an extension to the traditional Genomic BLUP (GBLUP) benchmark.

Our results suggest that incorporation of prior biological knowledge can improve genomic prediction accuracy for both Φ_{PSII} and PLA. The improvement achieved depends on the trait, type of knowledge and trait heritability. Moreover, transcriptomics offers complementary evidence to the Gene Ontology for improvement when used to define functional groups of genes. In conclusion, prior knowledge about trait-specific groups of genes can be directly translated into improved genomic prediction.

3.1 Introduction

Due to breakthroughs in DNA sequencing technology over the past decade, high-throughput genotyping is now a routine practice in plant breeding (Rimbert, Darrier et al. 2018). Phenotyping is undergoing a similar revolution: large phenomics facilities are being developed that can rapidly score large germplasm collections of plants in a range of different environments (Flood, Kruijer et al. 2016, Crain, Mondal et al. 2018). These technological developments have made it possible to acquire datasets describing genotypes and phenotypes for large numbers of individuals at an extended temporal scale. Despite recent advances in phenomics it is still more expensive and laborious than genotyping. To make the most use of phenomics datasets, Genomic Selection (GS) based breeding programs aim to predict unobserved phenotypes of individuals based on genotypes alone. This has the twofold benefit of reducing breeding costs and speeding up breeding programs as plants can be genotyped in the seedling stage and selected accordingly, thus negating the need to grow large populations to maturity and scoring them all to obtain breeding values based on phenotypes. GS usually models the unobserved phenotypes as additive effects of all genetic markers (total additive genomic value or breeding value) in the test population using a genomic prediction (GP) model. This GP model is based on a reference population which has both been genotyped and phenotyped for the trait(s) of interest (Meuwissen, Hayes et al. 2001). The performance of GP depends on many factors, including genetic architecture, reference population size and structure and heritability (Karaman, Cheng et al. 2016). However, GP accuracy, usually defined as the correlation (Pearson's *r*) between observed phenotypes and predicted breeding values, is generally lower for complex traits than for simpler ones (Morgante 2018). This is because such traits are affected by many loci with small to moderate effects, along with non-additive genetic (dominance, epistasis) and genotype-by-environment (GxE) interactions (Falconer and Mackay 1996). Incorporating epistasis into GP models has been reported to improve performance in selfing plant species but may not work for outcrossing species; therefore, additive GP models are still the primary choice (Jiang and Reif 2015).

In GP models, each individual's genetic or breeding value is modelled as the sum of additive marker effects. Despite advancements in phenomics, phenotyping data is still usually only available for a few traits of several hundreds of individuals (n), compared to millions of genetic markers (p). GP models tackle this curse of dimensionality ($p > n$) by regularization (Meuwissen, Hayes et al. 2001). When marker effects are fixed, this comes in the form of a penalty term added to the log-likelihood, as in LASSO or ridge regression. More frequently, marker effects are considered random, and regularization is achieved through prior distributions on the marker effects. The variance in these priors is directly related to the heritability, and can be estimated either using REML, or a fully Bayesian approach. In the classical GBLUP-approach, a single normal distribution with equal

variance is assumed for all marker effects (VanRaden 2008). More recently, mixture distributions have been considered (Moser, Lee et al. 2015). The prior could e.g. be a mixture of Gaussian distributions with large and small variances, and a point mass at zero, allowing a marker to have respectively large or small effects, or no effect at all (MacLeod, Bowman et al. 2016). Moreover, restrictions on the shape of the probability distribution, usually Gaussian, can be relaxed (e.g. *t*-distribution) to accommodate genetic architectures having a larger number of high to moderate effect sizes (Gianola 2013) or another suitable distribution can be exploited instead. In spite of these refinements, it is usually impossible to find the true causal variants when p > n, which may lead to suboptimal prediction. Therefore, several authors suggested that *a priori* available biological knowledge may be incorporated in GP models, prioritizing likely causal markers, and ultimately improving prediction accuracy (Edwards, Sorensen et al. 2016, Ehsani, Janss et al. 2016, Wang, Zhou et al. 2018).

Two types of biological knowledge have been considered in the literature: first, knowledge on biological properties of genes and their associated markers and second, knowledge in the form of secondary phenotypes. The latter typically concerns -omics data, and is modeled using additional relatedness matrices (Guo, Magwire et al. 2016, Morgante 2018, Azodi, Pardo et al. 2020) or penalized selection indices (Lopez-Cruz, Olson et al. 2020). Although such -omics data can in principle be generated for the GP reference population, the use of more general publicly available information is often more feasible and cost-effective. We therefore focus on biological properties of genes and markers, such as Gene Ontology (GO) and post-GWAS QTL information. The GO provides a structured resource of functional classes of gene products based on orthology, represented into three biological domains, i.e. molecular function, cellular component and biological process (Ashburner, Ball et al. 2000). Similar functional groupings can be achieved from transcriptomic experiments based on the assumption that functionally related genes are expressed together. These clusters of co-expressed genes may be enriched in multiple GO terms or pathways. Such information can be incorporated by allowing the GP model to put more weight on either certain individual markers (Legarra and Ducrocq 2012, MacLeod, Bowman et al. 2016) or groups of markers (Edwards, Sorensen et al. 2016). Various modelling approaches have been proposed to enable use of such data (Zhang, Liu et al. 2010, Speed and Balding 2014, Edwards, Sorensen et al. 2016, Ehsani, Janss et al. 2016, Guo, Magwire et al. 2016, Fragomeni, Lourenco et al. 2017). Here we use the Genomic Feature Best Linear Unbiased Predictor (GFBLUP) approach proposed by Edwards et al., 2016. GFBLUP extends GBLUP by partitioning the total genomic variance into two sub-components to weigh different genomic regions differently. This allows incorporating prior biological knowledge about groups of variants by treating each region as a separate random genetic effect with different variance. Subsequently, researchers applied this approach to various traits (Sarup, Jensen et al. 2016, Fang, Sahana et al. 2017, Rohde, Demontis et al. 2017, Gebreyesus, Bovenhuis

et al. 2019). While prior biological knowledge has thus been used to improve GP accuracy, the question remains what type of knowledge is most useful and how much the genetic architecture impacts the potential for improvement of particular traits.

In this study, we investigate improvement in GP performance using two sources of publicly available biological knowledge, i.e. Gene Ontology (GO) and clusters of co-expressed genes (COEX). This information was incorporated using the GFBLUP modelling approach, grouping markers in genes according to either their predicted function or co-expression respectively. As complex traits of study, we focused on photosynthetic light use efficiency of photosystem II (Φ_{PSII}) and projected leaf area (PLA) in *Arabidopsis thaliana*. Both of these traits are related, in the sense that the Φ_{PSII} directly illustrates the photosynthetic light use efficiency and can capture the most immediate physiological and regulatory response to varying irradiance levels (van Rooijen, Aarts et al. 2015), whereas growth in PLA is the net outcome of unit leaf photosynthetic capacity over time (Weraduwage, Chen et al. 2015, Liu, Peacock et al. 2020).

3.2 Materials and Methods

3.2.1 Datasets

3.2.1.1 Genotype data

Genotype data of the 360 natural accessions in the core set of the *Arabidopsis thaliana* HapMap population, representing its global diversity, was obtained using Affymetrix 250k SNP array (Zhang and Borevitz 2009, Baxter, Brazelton et al. 2010). The HapMap accessions were chosen as most accessions are more or less equally interrelated, so modelling is not heavily affected by population structure. Phenotypes of 344 accessions were available, so 16 accessions were removed from the analysis (CS76104, CS76112, CS76254, CS76257, CS76121, CS28051, CS28108, CS28808, CS28631, CS76086, CS76138, CS76212, CS76196, CS76110, CS76117, CS76118). Genotype data were subjected to quality control and all genotypes with a missing call in any accession were removed. Only 510 (0.24%) markers had minor allele frequency (MAF) <0.01 and 14,824 (6.9%) had MAF <0.05 (*Figure S12*). To incorporate the effects of rare alleles along with common alleles in the GP model, the MAF filtering threshold was set at 0.01. Of subsequent markers in a window of 50bp with a Pearson correlation coefficient (*r*) greater than 0.999, one was removed, using PLINKv1.9 (Purcell, Neale et al. 2007). In total, 214,051 SNPs passed quality filtering, 213,541 remained after MAF filtering and 207,981 SNPs were available after LD pruning for the analyses. The resulting minimal distance between SNPs was found to be ~550bp.

3.2.1.2 Phenotype data

The light use efficiency of Photosystem II electron transport (Φ_{PSII}) dataset was obtained from van Rooijen et al., 2017, who measured it using chlorophyll fluorescence via NIR

imaging at 790 nm. In this dataset, Φ_{PSII} was recorded three times a day; under 100 µmol m^{-2} s^{-1} (low light) for two days and for four continuous days after induction of high light stress at 550 µmol m^{-2} s^{-1} to study the photosynthetic acclimatory response. We measured PLA every 3 hours starting from the afternoon of day 22 after sowing until early morning of day 29 using the *'Phenovator'* high-throughput automated phenotyping system (Flood, Kruijer et al. 2016), which results in total of 54 timepoints for this trait (*Table S8*). Technical mis-match errors between the imaging system and the coordination of image analysis software were identified for some replicates at some time points for a small number of genotypes, but these were not found to influence overall results and the data was thus retained. Data of timepoints on day 22 was excluded from the analyses due to their relatively low coefficient of variation.

The *Phenovator* system has been designed to screen Arabidopsis plants for photosynthesis and growth on a larger temporal scale in a carefully controlled environment with minimal noise. The plants are grown over a table, spatially arranged into sowing blocks, imaged using a moveable monochrome camera recording 12 plants per image, and processed using an image processing software (available on demand from the authors). The system design allows spatial uniformity and temporal reproducibility by minimizing the design parameter variances. Therefore, we expected low variances of interactions between genotype and the design parameters; whereas, within image position and sowing position could have larger main effects and thus could be corrected for. Phenotypic values were taken as the average of one to four replicates of Best Linear Unbiased Estimators (BLUE) using the linear mixed model adjusted for experimental design factors (*Table S9*) that were described in Flood et al., 2016. For this experiment, the important design factors are spatial row (*x*) and column (*y*) coordinate, the image position and the sowing block. Thus, the BLUE for phenotypic mean is calculated based on this equation, implemented in R with the *lmer* function (supplemental R script) using the lme4 package (Bates, Sarkar et al. 2007):

$$\mathbf{Y} = \mathbf{Genotype} + \mathbf{x} + \mathbf{y} + \mathbf{image\_position} + \mathbf{sowing\_block} + \mathbf{error} \quad (3.1)$$

where 'Genotype' is used as fixed effect and the other factors are defined as random effects.

Both traits, at all measurement times, showed approximately normal distributions (*Figure S13 & 14*). The distributions are leptokurtic and left skewed for both traits (except for a few measurements for PLA on day 14 and day 15). The coefficients of variation under low light conditions for Φ_{PSII} ranged from 1.95-2.30% and 2.92-7.58% under high light and 18.73-27.04% for PLA (*Table S1*). Correlation between subsequent measurement times was high ($r > 0.9$) for both traits, except between measurements under low versus high light conditions of Φ_{PSII}; therefore, these were analyzed separately.

3.2.1.3 Biological priors

Co-expressed gene groups were obtained from the Arabidopsis expression compendium by Movahedi (2011). GO data was retrieved using the R package 'org.At.tair.db' (Carlson 2019) and genes were annotated using 'GO.db' (Carlson 2019) irrespective of evidence codes. The set of genes in GO terms were up-propagated along the GO tree, such that each GO group in our analysis comprised of a set of all those genes attributed to itself or to all of its child terms. The up-propagated sets of genes were retrieved using the 'GO2ALLTAIRS' method in the 'org.At.tair.db' package. Markers in genes linked to a specific GO term or COEX cluster were used in the analyses.

Moreover, a set of 7,242 photosynthesis related genes was manually compiled (*Table S6*) using four publicly available sources: KEGG (Kanehisa 2001) pathways related to photosynthesis (i.e. ath00195, ath00197, ath00710); the Arabidopsis pathway database AraCyc for four photosynthesis pathways (i.e. Calvin cycle, photorespiration, oxygenic, light reaction); genes annotated with GO terms directly related to photosynthesis machinery; and all 51 priority genes selected for GWAS of photosynthesis acclamatory response identified by for this HapMap population.

3.2.2 Statistical analysis

3.2.2.1 Linear mixed models

The Linear Mixed Model (LMM) with one random genomic component was used as baseline. This model (3.2), known as Genomic Best Linear Unbiased Prediction (GBLUP) (Habier, Fernando et al. 2007, VanRaden 2008) was used to predict marker effects, calculate genomic heritability (h^2_{GBLUP}) and the total additive genomic values, which is the sum of all marker effects:

$$\tilde{\mathbf{y}} = \boldsymbol{\mu} + \mathbf{g} + \boldsymbol{\varepsilon} \qquad (3.2)$$

Here, $\tilde{\boldsymbol{y}}$ is an nx1 vector of adjusted phenotypes as described in section 5.1.2, μ is the overall mean, $\mathbf{g}$ is an nx1 vector of genomic values captured by all genomic markers such that $\mathbf{g} = \hat{\boldsymbol{g}}$ and $\boldsymbol{\varepsilon}$ is an n-vector of residuals. The random genomic values $\mathbf{g}$ and residuals were assumed to be independent, normally distributed as $\mathbf{g} \sim N(0, \mathbf{G}\sigma^2_g)$, $\boldsymbol{\varepsilon} \sim N(0, \mathbf{I}\sigma^2_e)$. Here $\mathbf{G}$ is the genomic relationship matrix (GRM), providing variance-covariance structure of genotypes calculated from all genomic markers and $\mathbf{I}$ is the identity matrix.

Accordingly, for each GO and COEX gene groups, another linear mixed model similar to GBLUP but with two random genomic components (3.3), known as Genomic Feature Best Linear Unbiased Predictor (GFBLUP) (Edwards, Sorensen et al. 2016) was applied:

$$\tilde{\mathbf{y}} = \boldsymbol{\mu} + \mathbf{f} + \mathbf{r} + \boldsymbol{\varepsilon} \qquad (3.3)$$

This model differs from GBLUP in that the total estimated genomic value ($\hat{\mathbf{g}} = \hat{\mathbf{f}} + \hat{\mathbf{r}}$) is partitioned into genomic value captured by markers in a GO/COEX group ($\hat{\mathbf{f}}$) and by the

remaining markers ($\mathbf{r}^{\wedge}$), such that $\mathbf{f} \sim N(0, \mathbf{G_f}\sigma_f^2)$, $\mathbf{r} \sim N(0, \mathbf{G_r}\sigma_r^2)$ and $\boldsymbol{\varepsilon} \sim N(0, \mathbf{I}\sigma_\varepsilon^2)$. For both GBLUP and GFBLUP, total genomic value $\hat{\mathbf{g}}$ of the test population was predicted conditional on observed phenotypes of the training population, using the approach mentioned by Edwards et al., 2016. The genomic relationship matrix $\mathbf{G}$ in the GBLUP model was constructed based on all genomic markers such that $\mathbf{G} = \mathbf{WW}'/m$, where $\mathbf{W}$ is an $n \times m$ genotype matrix (n genotypes and m markers), centered and scaled such that its i^{th} column $\mathbf{w_i} = (\mathbf{z_i} - 2p_i)/\sqrt{2p_i(1-p_i)}$, where $\mathbf{z_i}$ is the i^{th} column vector of $\mathbf{Z}$ having minor allele counts (0, 1 or 2) as entries and p_i is the MAF of the i^{th} marker. In our case, all genotypic locations were homozygous, so genotypes are coded as 0 or 2. For the GFBLUP model, the genomic relationship matrix $\mathbf{G_f}$ for each GO or COEX group was calculated from the markers linked to that group; $\mathbf{G_r}$ was constructed from the remaining markers.

The MultiBLUP model (3.4) was constructed according to the Adaptive MultiBLUP strategy proposed by (Speed and Balding 2014). Briefly, the total genome was divided into adjacent but 50% overlapping regions of 10kb. The genomic markers within these regions were tested as a group to estimate their association with the phenotype ($p < 10^{-5}$) and adjacent regions were merged if $p_{\text{Bonferroni}} < 0.05$. Subsequently, separate covariance matrices $\mathbf{K}_1$, $\mathbf{K}_2$, ..., $\mathbf{K}_M$ were constructed for each region (M regions in total) based on its markers and genomic values $\mathbf{g}_1$, $\mathbf{g}_2$, ..., $\mathbf{g}_M$ were estimated. The GRM based on all markers (equivalent to GBLUP) was used if no region was found significant. The total genomic value is $\hat{\mathbf{g}} = \sum_{m=1}^{M} \hat{\mathbf{g}}_m$ with i.i.d. $\mathbf{g}_m \sim N(0, \mathbf{K}_m\sigma_m^2)$ and $\boldsymbol{\varepsilon} \sim N(0, \mathbf{I}\sigma_e^2)$:

$$\tilde{\mathbf{y}} = \boldsymbol{\mu} + \sum_{m=1}^{M} \mathbf{g}_m + \boldsymbol{\varepsilon} \qquad (3.4)$$

Variance components in all of these LMMs were estimated using the average information restricted maximum-likelihood (REML) procedure (Johnson and Thompson 1995) implemented in the *greml* method of the R package *qgg* (Rohde, Fourie Sørensen et al. 2020) for GBLUP/GFBLUP, using a maximum of 100 iterations at a tolerance level of 10^{-5}; and LDAK v5.1 (http://dougspeed.com/) for MultiBLUP.

Total additive genomic value was predicted using 8-fold cross-validation. This involved training the model using 301 (78%) genotypes and using the remaining 43 for testing in each fold. The exact same accessions were used for both GBLUP and GFBLUP during each split to enable a fair comparison. Prediction accuracy of models was defined as Pearson correlation (r) between observed phenotypic values and predicted genomic values of the test population in each fold. The procedure was repeated 10 times, thus modelled predictive ability distributions consisted of 80 correlations or fewer if variances were over- or underestimated as described earlier by simulation studies (Kruijer, Boer et al. 2015). For comparison between models, the median of these correlations was used, and significance of the difference was tested using the non-parametric Wilcoxon–Mann–Whitney test for assessing significant differences in median accuracy between GBLUP

and GFBLUP. Subsequently, *p*-values were adjusted for multiple-testing correction by calculating False Discovery Rate (FDR) based on total number of GO/COEX groups multiplied by total number of time points (Edwards, Sorensen et al. 2016). For Φ_{PSII} we also analyzed results without FDR adjustment, which are referred as "informative" as opposed to "significant" throughout the text.

3.2.2.2 Model performance evaluation

GFBLUP models were compared to the benchmark GBLUP based on their goodness of fit, predictive ability and estimated genomic parameters. Using the likelihood ratio test (LRT) we tested the null-hypothesis $\sigma_f^2 = 0$. LRT *p*-values were based on the asymptotic distribution of the LRT-statistic, which is a mixture of a point mass at 0 and a χ^2-distribution with 1 degree of freedom (d.o.f.) (Edwards, Thomsen et al. 2015). The significantly improved GFBLUP models ($p_{LRT} < 0.05$) having predictive abilities greater than the benchmark GBLUP (i.e. *p-value* of Wilcoxon-Mann-Whitney tests < 0.05) were filtered for subsequent analysis. Genomic parameters were calculated from variance estimates of both models to analyze only models passing the abovementioned filtering criteria. This includes total genomic heritability explained ($h^2_{GBLUP} = \sigma_g^2/(\sigma_g^2 + \sigma_e^2)$) and proportion of genomic heritability explained by an individual GO/COEX group in GFBLUP models ($h_f^2 = \sigma_f^2/(\sigma_f^2 + \sigma_r^2 + \sigma_e^2)$). In order to check if we obtained a higher number of *PSGENES* in GO/COEX groups than expected by chance, we used the chi-square test with 1 d.o.f. to compare the observed versus expected frequencies of *PSGENES* in these groups.

3.2.3 Semantic clustering of GO terms

Informative GO terms were clustered based on their semantic similarity using the *Revigo* (Supek, Bošnjak et al. 2011) web server with *'SimRel'* semantic similarity metric equal to 0.7. The resulting GO clusters were plotted using a Multidimensional Scaling (MDS) plot in R, where maximum %gain in accuracy by each GO term was used to color the bubbles. GO terms enriched in COEX groups were found using the PANTHER classification system (Mi, Muruganujan et al. 2019). Fisher's exact test was used for calculating enrichment *p*-values followed by multiple testing correction using the FDR, reporting enrichment at $p < 0.05$. These enriched GO terms were sorted in order of their GO hierarchical tree such that a child term was below its parent; thus, the most specific GO terms are the child GO terms in the bottom of that tree, were used for subsequent analysis.

3.3 Results

3.3.1 Genomic prediction of complex growth related traits

Previously, van Rooijen et al. (2017) conducted a GWAS on *A. thaliana* photosynthesis. In particular, they measured the light use efficiency of photosystem II electron transport

(Φ_{PSII}) for 344 accessions of the Arabidopsis HapMap population, switching from low light (100 µmol m^{-2} s^{-1}) to high light (550 µmol m^{-2} s^{-1}) irradiance at the onset of day 25. In total, they took 6 measurements before and 12 after applying light stress to identify potential QTLs during acclimation to high light. As we intend to use this population to explore the utility of biological knowledge in genomic prediction, we combined projected leaf area (PLA), another indicator of plant growth, with Φ_{PSII}. We first assessed whether GP works with reasonable performance for these complex traits. For this purpose, a classical Genomic Best Linear Unbiased Prediction (GBLUP) model was constructed to assess how well the infinitesimal modelling assumptions fit and to calculate markers-based heritability. In this model (3.2), all marker effects are treated as arising from a single normal distribution $N(0, \mathbf{G}\sigma_g^2)$ having one random genetic component, to regress each individual phenotype measurement over all markers simultaneously. At low light (LL) levels, mean prediction accuracy for Φ_{PSII} is lower (Pearson's *r* between predicted and observed phenotypic values ranging from 0.16±0.02 to 0.22±0.01) than at high light (HL, Pearson's *r* ranging from 0.40±0.01 to 0.48±0.01), as shown in ***Figure 3.1A***. Prediction accuracy for PLA (***Figure 3.1B)*** ranges from 0.06±0.01 to 0.17±0.01 and rises with the increase in plant size and simultaneously decreases with increase in phenotypic coefficient of variation. Genomic heritability (h^2_{GBLUP}) for Φ_{PSII} ranged from 0.08-0.13 under LL and 0.56-0.87 under HL, and 0.05-0.17 for PLA (***Figure S1***). Differences in prediction accuracy for Φ_{PSII} between LL and HL are in line with differences in genomic heritability, in accordance with the observation that genomic prediction accuracy is generally positively correlated with heritabilities (Hayes, Bowman et al. 2009). Moreover, for ~1.2% of the GBLUP models for PLA, h^2_{GBLUP} was zero because of undetermined genomic variance, whereas for Φ_{PSII} ~7% of genomic variances were estimated to be 100% ($h^2_{GBLUP} = 1$), which is clearly an over-estimation (***Figure S2)***. As reported by Kruijer et al. 2015, it was expected (based on 5000 simulated traits) that ~10-15% of GBLUP

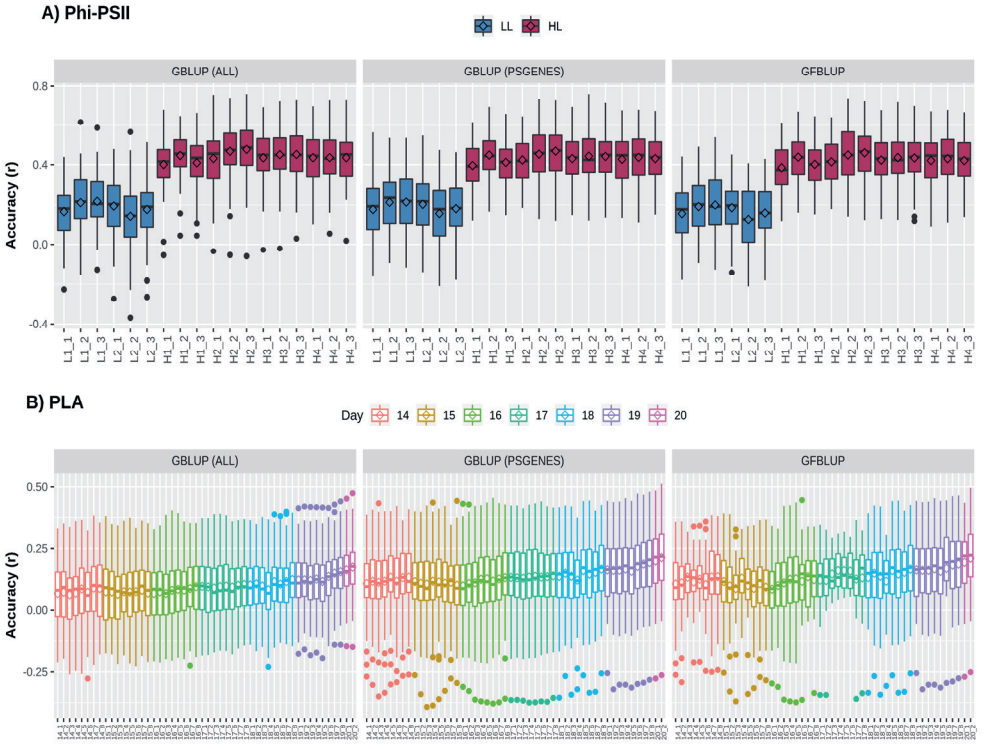


Figure 3.1: Genomic prediction accuracy for photosynthetic light efficiency and projected leaf area using high level prior biological knowledge (PSGENES).

Cross-validation based assessment of prediction accuracy as Pearson's *r* between true and predicted values using three models: (i) GBLUP using all genomic markers (ALL), (ii) GBLUP using only pre-selected photosynthesis related genic markers (*PSGENES*) and (iii) GFBLUP using *PSGENES* as a genomic feature in two genomic components. (A) Accuracy for Φ_{PSII} for two days (6 time points) under low light irradiance levels (LL) and four days (12 time points) under high light irradiance (HL). (B) Accuracy for PLA measured 8 times per day from day 14 after sowing to day 20, where day 20 has only two measurements.

models could have variance components that cannot be estimated for this population, so we discarded these models from our analysis.

An extension of GBLUP is MultiBLUP (Speed and Balding 2014), using multiple random genetic components in the model (3.4), thus allowing differential weighting of groups of genomic markers, each having a separate kinship matrix derived from that group. We applied MultiBLUP using adjacent overlapping chromosomal partitions of 10kb (yielding best performance when testing window sizes of 1 to 100kb) to check if multiple kinship matrices or genomic variance decomposition improve prediction. The results (*Figure S3*) indicate that performance was close to that of GBLUP and could not be improved further. This could be because most models ended up with only one background kinship matrix

during cross-validation and many of these genomic regions did not meet the significance threshold ($p_{Bonferroni} < 0.05$) during association testing. In summary, these results show that predictive performance for these complex traits is low and there may be room for improvement by incorporating prior biological knowledge, decomposing the total genomic variance into biologically relevant subsets.

3.3.2 High-level biological knowledge does not necessarily improve genomic prediction

The next question is whether predictive performance can be improved by using only markers residing within genes that are known to be linked to the traits of interest. The idea comes from previous studies, in which a subset of markers was associated to biological relevant genes and achieved a genomic value similar to the total genomic value achieved when using all SNPs (VanRaden, Tooker et al. 2017, Li, Zhang et al. 2018). Here, we selected 7,242 photosynthesis related genes, referred to as *PSGENES* in the text, from public repositories (see M&M) and constructed a GBLUP model based only on these. The Genomic Relationship Matrix (GRM) was constructed from all markers within the ORFs of *PSGENES*, leaving ~17% of the total genotyped markers after filtering. Interestingly, the models performed equally well (*Figure 3.1*) as the GBLUP based on all markers for both traits, with a slight improvement in predictive ability for PLA (max. ~6% increase in accuracy). Subsequently, to assess whether this pre-selected subset of markers can improve results if they are weighted differently than the rest of markers, we constructed another model using the GFBLUP modelling approach (Edwards, Sorensen et al. 2016) (3.3) having two genomic components. In this model, the markers within *PSGENES* were treated as one genomic component and the remaining markers as a second genomic component. Again, this model showed similar predictive performance as GBLUP, with some reduction in variability for PLA, but could not improve the accuracy further (*Figure 3.1*). From this, we conclude that prior biological knowledge-based selection of functionally relevant genes is potentially useful, but an optimal grouping may be important to improve GP further.

3.3.3 More fine-grained biological knowledge is useful for improving genomic prediction

To assess whether prior information from publicly available resources can help improve GP performance, we tested grouping of genes based on Gene Ontology (GO) terms and previously reported clusters of co-expressed genes (COEX) of *Arabidopsis thaliana* in multiple tissues and developmental stages (Movahedi, Van de Peer et al. 2011). Each of the three GO sub-ontologies, Biological Process (BP), Molecular Function (MF) and Cellular Component (CC), was used. The corresponding groups of markers in a GO or COEX group, called a genomic feature (GF), were used in GFBLUP (3.3) using a separate model for each feature with two genomic components, i.e. one with markers from the GF and the other with the remaining markers (rGF). The predictive performance was

compared to that of the GBLUP benchmark using all markers with identical sets of 8-fold cross-validation test populations. Each group of markers based on GO or COEX was treated as a separate random effect in its respective GFBLUP model for which its contribution to the total genomic variance was calculated (see M&M). For each GF, the effects of all corresponding markers were assumed to follow a normal distribution with equal variance, but different from the remaining markers; that is, the markers in the GF are differentially weighted and prioritized from the rest.

In total, 7,297 GO terms and 12,419 disjoint COEX gene groups were linked to at least one marker. The total number of genes ranged between 1 and 24,998 for the GO features and between 1 and 3,384 for the COEX groups (*Figure S4, Table S4*); the number of markers ranged between 0 and 109,723 for the GO features and 4 and 19,621 for the COEX groups. Due to the hierarchical GO structure, the 95th percentile of the total number of genes within GO features was lower (496) as compared to COEX (2,466). Note that both GO and COEX groups may overlap, i.e. a gene can be in multiple functionally related GO/COEX groups. In the following results, the improvement in genomic prediction has been quantified in terms of percent gain in accuracy compared to the GBLUP benchmark, GFBLUP model's goodness of fit measured using likelihood ratio test (LR), and genomic heritability (h^2_{GBLUP}) and proportion of genomic heritability explained by a genomic feature (h^2_f).

3.3.3.1 GO informed prediction

7,297 GO terms were tested with repeated 8-fold cross-validation at multiple measurements of a trait, leading to a total of ~10 million GFBLUP model accuracies for Φ_{PSII} and ~29 million for PLA (*Figure S5*). The models for which variance was apparently over-estimated ($h^2_f > 0.99$) or undetermined ($h^2_f < 0.01$) were not considered for subsequent analysis. This was the case for ~50% of the models for both traits, indicating that only selected biological groups are potentially helpful.

We initially analyzed the highest gain in prediction performance obtained by any GO term at any time point. For Φ_{PSII}, 'salicylic acid biosynthesis' (BP) provided the highest increase in accuracy (~60%), for Φ_{PSII} measurements under low light on the second day (*Figure 3.2, Table S2A*). For the GO sub-ontologies CC and MF, 'organelle outer membrane' and 'phosphatase activity' respectively yielded highest gains in these categories under low light (~43% and 37% respectively; *Table S2A*). None of the GO terms yielded a significant improvement after high light stress; however, some GO terms, e.g. 'protein containing complex' yielded an increase in accuracy higher than the benchmark but not passing our model evaluation criteria wholly (*Figure 3.3*). For PLA, the largest improvement (~197%) was obtained by the biological process 'monocarboxylic acid biosynthesis' (*Figure 3.2*, *Table S2B*). The best performing MF and CC terms for PLA were 'exopeptidase activity' and 'chloroplast part' (~185% and ~178% respectively; *Figure 3.3, Table S2B*).

Interestingly, these best CC terms for both traits are directly related to photosynthesis, which lends credibility to the usefulness of the GO terms to capture relevant prior biological knowledge.

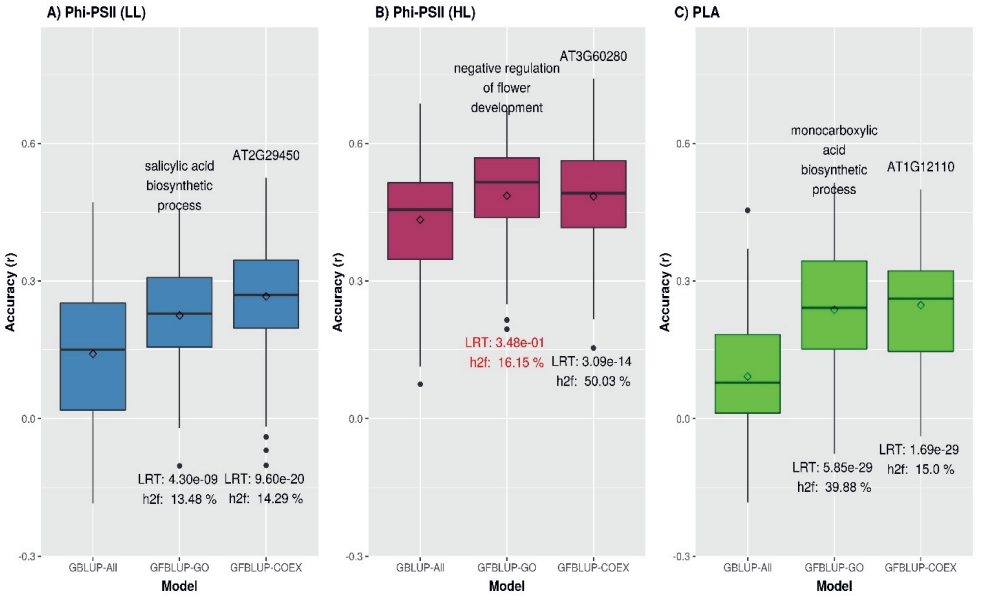


Figure 3.2: Biological priors can help improve genomic prediction accuracy for growth related traits.

Prediction accuracy of the best overall GFBLUP models using Gene Ontology (GO) and co-expression (COEX) gene groups, compared to the GBLUP benchmark (without prior biological knowledge). Since the GBLUP model was evaluated for each measurement of both traits, so GBLUP here is shown for the corresponding time point where improvement by GFBLUP was observed. Accuracy was calculated as Pearson's *r* between true vs. predicted values and significant different was evaluated using likelihood ratio test (LRT) at 5% significance threshold. Moreover, proportion of total genetic variance explained by only the group of SNPs in GO or COEX is mentioned as h_f^2. The GBLUP-ALL model uses all markers in GBLUP; GFBLUP-GO and GFBLUP-COEX models use the top GO terms and COEX (see text for details). (A) Accuracy for Φ_{PSII} under low light irradiance levels (LL). (B) Accuracy for Φ_{PSII} under high light irradiance (HL). Here, despite showing some improvement, the GFBLUP-GO model did not pass all of our model evaluation criteria (see Model Performance Evaluation). (C) Accuracy for PLA.

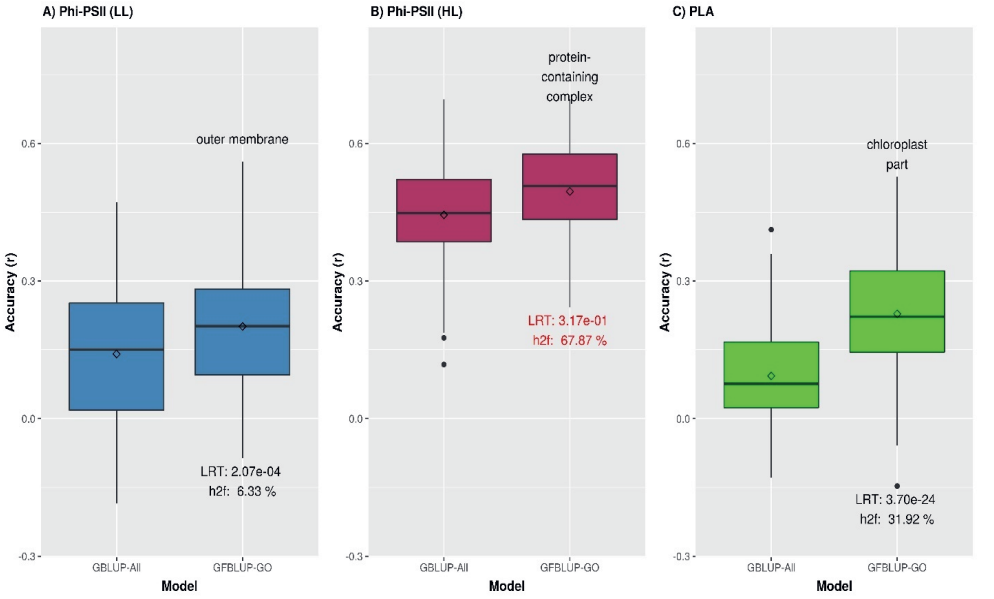


Figure 3.3: Biological priors based on top GO cellular components improving genomic prediction accuracy for growth related traits.

Prediction accuracy of the best GFBLUP models using Gene Ontology (GO) cellular components gene groups, compared to the GBLUP benchmark (without prior biological knowledge). The accuracy of benchmark model may differ between corresponding *Figures 3.2* and *3.3*, because it is calculated from the corresponding time point, where improvement by GFBLUP was realized. Accuracy was calculated as Pearson's *r* between true vs. predicted values and significant different was evaluated using likelihood ratio test (LRT) at 5% significance threshold. Moreover, proportion of total genetic variance explained by only the group of SNPs in GO or COEX is mentioned as h_f^2. The GBLUP-ALL model uses all markers in GBLUP; GFBLUP-GO models use the top GO cellular component terms mentioned in the text above. The text in the bottom of boxplots shows the likelihood ratio test *p*-value (LRT) and proportion of genomic heritability explained (h_f^2) by corresponding GO model. (A) Accuracy for Φ_{PSII} under low light irradiance levels (LL). (B) Accuracy for Φ_{PSII} under high light irradiance (HL). Similar to *Figure 3.2*, the GFBLUP-GO model did not pass all of our model evaluation criteria (see Model Performance Evaluation), though showing some improvement. (C) Accuracy for PLA.

In total, 43 GO terms (BP:34, CC:6, MF:3) were potentially informative (i.e. Wilcoxon–Mann–Whitney test *p*-values <0.05, without multiple testing correction), showing a tendency to improve Φ_{PSII} traits and yielding a significant increase in GFBLUP model accuracy (*Figure S6A, Figure S7, Table S2A*) compared to GBLUP. The overall gain in accuracy for these informative GO features ranged between 23% and 60%. The GO terms' hierarchical redundancy was removed using GO trimming (Jantzen, Sutherland et al. 2011) and the remaining 40 informative terms fell broadly into six biological clusters (*Figure 3.4, Figure S9*): i) hormonal regulation; ii) cellular development; iii) transport; iv) metabolism; v) catabolism and vi) macromolecular complex assembly, organization and

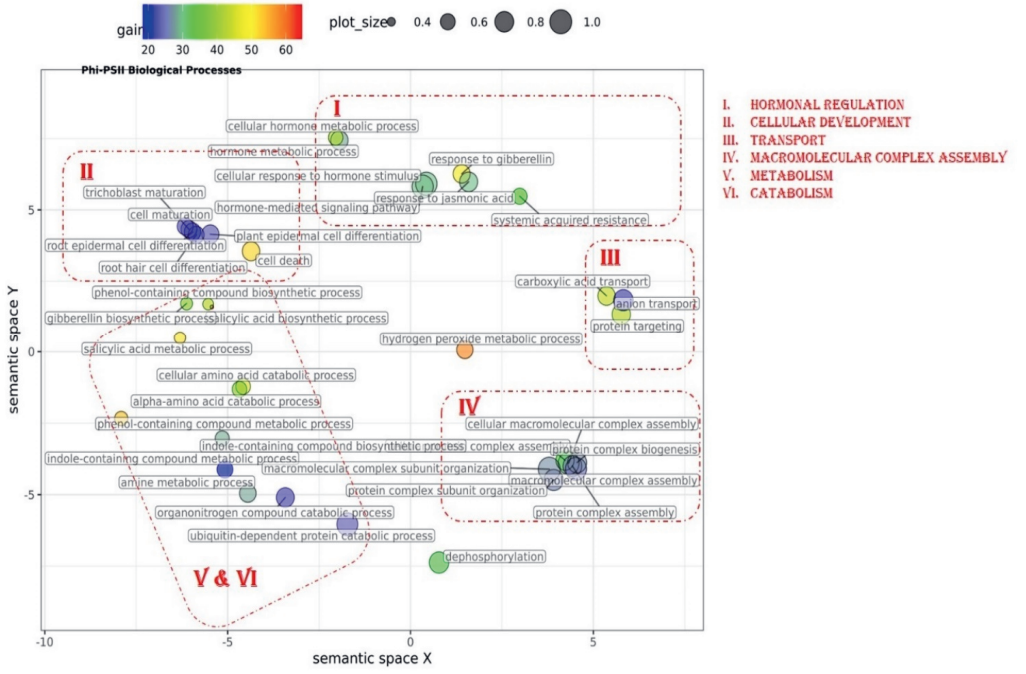


Figure 3.4: Semantic clustering of GO informed prediction for Φ_{PSII}.

Multidimensional scaling (MDS) plot of the representative subset (i.e., terms remaining after the redundancy reduction) of biological process GO terms informative for Φ_{PSII}. Semantically similar GO terms are clustered based on the "*SimRel"* semantic similarity measure using *Revigo*. Dot size is proportional to the number of genes annotated with the GO term, such that more general GO terms have larger circles. The *x* and *y* coordinates indicate relative cluster distances in 2 dimensions. The %gain of a particular GO term is indicated by the circle colour.

biogenesis. The cellular component terms were semantically clustered into organellar membranes and photosynthesis machinery sub-compartments, whereas molecular function terms were related to transmembrane transport and phosphatase activities.

For PLA, 52 GO terms (BP:41, CC:6, MF:5) resulted in significant improvement (p_{FDR}< 0.05) in predictive ability (*Figure 3.6*, *Figure S6C, Table S2B*) and the gain in accuracy ranged between 104% and 197%. After removal of hierarchical redundancy, semantic grouping of the remaining 45 GO terms showed that they involved a number of growth and developmental processes. Biological process GO terms fell into ~8 clusters (*Figure 3.5, Figure S10*) related to development, defence response, stress response, cell cycle regulation, metabolism, molecular biosynthesis, cellular component organization and transport. The molecular function terms were clustered into two groups including exopeptidase and methyltransferase activities. The cellular component terms included the photosynthesis machinery (i.e. chloroplast) and endoplasmic reticulum. Comparison

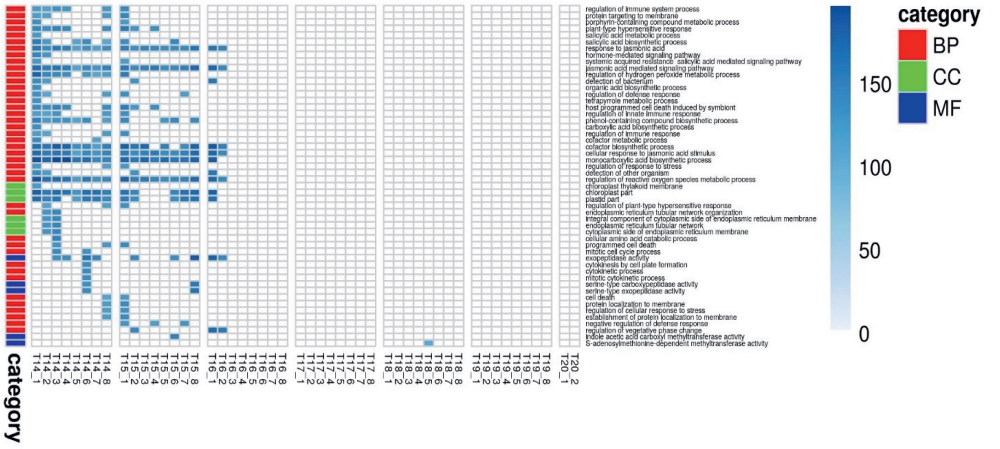


Figure 3.5: Improvement in genomic predictive performance using GO for PLA.

All GO terms that significantly improve GFBLUP models for PLA with %gain in accuracy (*r*) over GBLUP. Each GO term has a separate model for individual measurements indicated as T{day}_{Number of measurement}. The colour bar identifies the GO terms as Biological Process (BP), Cellular Component (CC) and Molecular Function (MF).

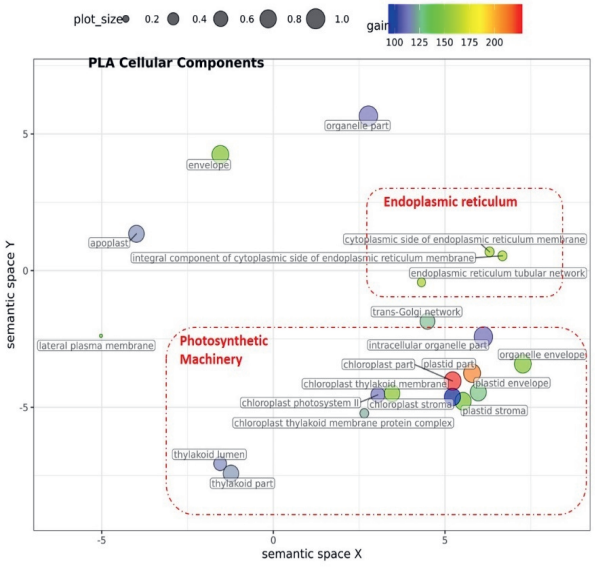


Figure 3.6: Semantic clustering of GO informed prediction for PLA.

Multidimensional scaling (MDS) plot of the representative subset (i.e., terms remaining after the redundancy reduction) of cellular component GO terms informative for PLA. Semantically similar GO terms are clustered based on the "*SimRel*" semantic similarity measure using *Revigo* (Supek, Bošnjak et al. 2011). Dot size is proportional to the number of genes annotated with the GO term, such that more general GO terms have larger bubbles. The *x* and *y* coordinates indicate relative virtual cluster distances in 2 dimensions. The %gain of a particular GO term is indicated by the bubble colour.

of average accuracy over multiple folds of GO models (*Figure S6A,C*) indicate that many models performed better than GBLUP. Some of these passed our significance threshold (see model evaluation criteria, M&M) at a particular trait measurement but appeared to improve prediction performance for other measurement points as well.

The maximum number of genes annotated with the informative GO terms for Φ_{PSII} and significant GO terms for PLA were 1,358 and 1,245 respectively. These GO terms appeared at multiple levels of the GO hierarchical structures, including parent and child terms closely related to photosynthesis and growth (*Table 3.1*). Moreover, many genes were common with the pre-selected photosynthesis related *PSGENES*: 42% and 58% for Φ_{PSII} and PLA respectively, significantly more than what expected by chance ($p_{\chi^2_{df:1}} < 0.05$). Total genomic heritability (h^2_{GBLUP}) was negatively correlated with predictive gain ($r_{\Phi PSII}$=-0.77, r_{PLA}=-0.5). The genomic heritability explained individually (h^2_f) by the informative GO terms ranged between 6% and 31% for Φ_{PSII} and between 3% and 43% for PLA (*Table S2A,B*). Interestingly, the markers associated with these GO terms constituted only 0.1-3.3% of the total markers for Φ_{PSII} and 0.005-2.8% for PLA. This indicates that to improve predictive ability, genomic variance can be decomposed based on biologically meaningful sets of genes scattered over the genome rather than lie in adjacent regions such as in the MultiBLUP analysis above. Moreover, h^2_f is positively correlated with GO gene group size ($r_{\Phi_{PSII}}$=0.87, r_{PLA} =0.77) as well as with the likelihood ratio ($r_{\Phi_{PSII}}$ =0.60, r_{PLA} =0.65) of both trait models, indicating that incorporating meaningful prior subsets into the GFBLUP model improves goodness of fit.

From this we infer that GO-based prior knowledge can improve GP performance. The improvement is most prominent for traits with low heritability, where some of the GO terms appeared more frequently for PLA than Φ_{PSII} at multiple measurement times.

3.3.3.2 COEX informed prediction

Similar to genomic features based on GO, we made subsets of markers based on COEX clusters by selecting the markers within the ORFs of genes which were part of a given COEX cluster. Similar to GO based models, COEX models with zero and with 100% variance explained were discarded (*Figure S5*). In general, more COEX models pass our model evaluation threshold (*Figure S6B,D*) and they have a higher likelihood ratio than GO based models. This could be due to the genic overlap between groups and the enrichment of multiple related GO terms within a group.

For Φ_{PSII} we found 172 informative COEX gene groups potentially improving predictive ability, one of which was statistically significant (p < 0.05) after correcting for multiple testing using FDR (*Figure S6B, Figure S8*). 355 COEX groups significantly improved predictive ability for PLA (*Figure 3.7, Figure S6D*, *Table S3A,B*). The gain in accuracy

was higher for PLA (80% to 243%) than for Φ_{PSII} (7% to 89%) and was negatively correlated with genomic heritability ($r_{\Phi_{PSII}}$=-0.86, r_{PLA}=-0.56), like for GO informed prediction. This improvement was attributed to a maximum of only ~5% of the total genomic markers in all groups. Interpretation of COEX gene groups is not as straightforward as of GO terms, which by nature carry an informative name. Interestingly, ~90% of genes were common in the COEX groups for both traits, possibly due to the relatedness of the traits. To attach biological meaning to these groups we performed GO enrichment analysis on all groups together. We found 113 BP, 29 MF and 24 CC most specific GO terms enriched in these clusters. The top 10 GO terms with highest fold enrichment include photosynthesis machinery, i.e. chloroplast stroma (GO:0009570), chloroplast envelope (GO:0009941) cellular components; ATPase activity coupled with transmembrane ion transport (GO:0015662); and glucose metabolic process *(Figure S11, Table S5*). These results indicate that trait-specific co-expressed gene functional groups can also help improve prediction performance and that these groups capture biologically relevant functions.

Table 3.1: Known trait-specific GO terms improving genomic prediction performance for both traits.
The proportion of explained genomic heritability (h_f^2) by a GO term, likelihood ratio (LR) between GFBLUP and GBLUP models, Wilcoxon–Mann–Whitney test p-value, total number of genes and markers, %gain in accuracy (r), correlation between genomic relationship matrices based on GO term markers ($\mathbf{G}_f$) and remaining markers ($\mathbf{G}_r$) and total genomic heritability (h^2_{GBLUP}), for different trait specific GO terms that are common to both GO and COEX based analyses. For GO terms, the type is indicated – molecular function (MF), biological process (BP) and cellular component (CC).

Φ_{PSII}										
GO ID	Ontology	Type	h^2_f	LR	pvalue (unadj)	#gene	#marker	%gain	Cor(G_f,G_r)	h^2_{GBLUP}
GO: 0009543	chloroplast thylakoid lumen	CC	0.07	10.5	$1.48x10^{-2}$	71	218	33	0.59	0.09
GO: 0031968	organelle outer membrane	CC	0.06	12.5	$4.3x10^{-3}$	72	345	40	0.61	0.08
GO:0044429	mitochondrial part	CC	0.14	47.1	$2.3x10^{-3}$	298	1069	38	0.81	0.09
GO:0005740	mitochondrial envelope	CC	0.13	8.43	$2.7x10^{-2}$	255	914	25	0.79	0.12
PLA										
GO ID	Ontology	Type	h^2_f	LR	pvalue (adj)	#gene	#marker	%gain	Cor(G_f,G_r)	h^2_{GBLUP}
GO:0044434	Chloroplast part	CC	0.32	101	$5.26x10^{-5}$	1211	5658	178	0.94	0.07
GO:0009535	chloroplast thylakoid membrane	CC	0.14	10	$4.9x10^{-2}$	322	1139	121	0.81	0.07
GO:0000911	cytokinesis by cell plate formation	BP	0.15	34	$9.6\ x10^{-3}$	204	1465	134	0.81	0.07
GO:0010090	trichome morphogenesis	BP	0.04	30	$8.3\ x10^{-4}$	31	65	154	0.4	0.06
GO:0010321	regulation of vegetative phase change	BP	0.14	18	$4.9\ x10^{-3}$	425	1512	106	0.84	0.07
GO:0048366	leaf development	BP	0.10	48	$1.96\ x10^{-5}$	99	487	187	0.62	0.06
GO:0090698	post-embryonic plant morphogenesis	BP	0.04	7	$8.3\ x10^{-7}$	4	11	207	0.2	0.06

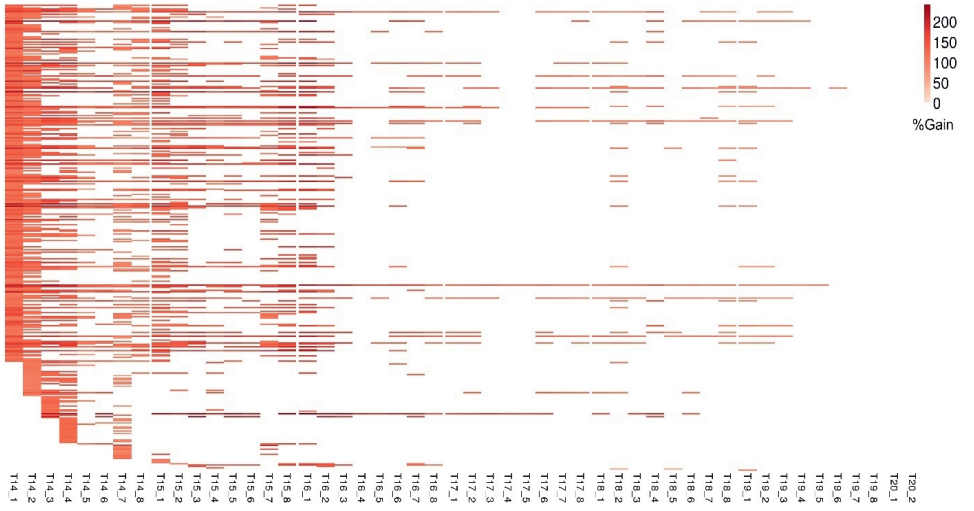


Figure 3.7: Improvement in genomic prediction performance using co-expressed gene clusters for PLA.

All COEX clusters that significantly improve GFBLUP models for PLA with %gain in accuracy (*r*) over GBLUP. Each COEX cluster has a separate model for individual measurements indicated as T{day}_{Number of measurement}. The clusters are ordered according to "cluster_sr_no" column in Supplementary Table 3B.

Similar to GO informed prediction, ~34% of COEX genes were common to the pre-selected photosynthesis related genes (*PSGENES*) for both traits, but here this is close to what we expect by chance. This indicates that, even though the COEX groups contain only a limited subset of all genes, they are not biased towards photosynthesis genes. The gain in predictive ability and explained genomic heritability (h_f^2) for Φ_{PSII} by the top COEX gene group was higher (89% resp. 14%) than those for the top GO feature (60% resp. 13%). Similarly, for PLA the top COEX gene group achieved a higher accuracy gain (242%) than the top GO group (197%), as shown in *Figure 3.2.* Notwithstanding these differences, we observed that many genes were common between GO and COEX based prediction for both traits (21% and 19% of all models passing the evaluation criteria for Φ_{PSII} and PLA resp.). These common genes in COEX based prediction were mainly enriched for many fundamental photosynthesis and growth related GO terms (*Table S7A,B*), e.g. light harvesting in photosystem I and photosynthetic electron transport in photosystem II (BP), chloroplast (CC) and ATP binding (MF).

The largest informative COEX groups for Φ_{PSII} and for PLA only differ slightly in sizes (3,176 and 2,840 genes respectively), but on average, COEX groups were larger than the GO groups for both traits. The 95th percentile of genomic heritability explained individually

by the COEX groups (h_f^2) was 70% for Φ_{PSII} and 39% for PLA, indicating that some Φ_{PSII} models could be over-estimated. Analogous to GO, h_f^2 was positively correlated with COEX gene group sizes ($r_{\Phi_{PSII}}$=0.88, r_{PLA}=0.40) and likelihood ratio ($r_{\Phi_{PSII}}$=0.27, r_{PLA}=0.22), indicating that incorporating meaningful prior subsets into the COEX model improved goodness of fit.

Together, our results illustrate that both of the meaningfully specific GO terms and more general COEX groups of genes with interrelated functions may improve GP predictive performance.

3.4 Discussion

3.4.1 Predicting photosynthesis

In this work, we aimed at improving GP performance by exploiting publicly available biological knowledge to group genes in three different ways: using our knowledge about the trait, using the Gene Ontology and using co-expression. Instead of developing new methodology, we focused on using existing BLUP methods, widely used in animal and plant breeding, to explore new sources of biological prior knowledge, e.g. clusters of co-expressed genes. The GFBLUP methodology was initially proposed for *Drosophila melanogaster* using Gene Ontology data as biological prior knowledge (Edwards, Sorensen et al. 2016). We also investigated to what extent different traits benefit from and the use of prior knowledge. Our results support a strong influence of different trait genetic architectures, since performance improvement was more evident for leaf area phenotypes than for Φ_{PSII}.

The approach can be generally applied to complex traits, but here we focused on photosynthesis and plant size. Besides serving as a case study, photosynthesis is also interesting in its own right, for two reasons. First, the genetic architecture of photosynthesis, though well-studied over the previous decades, is still poorly described in the quantitative genetic context (van Rooijen, Kruijer et al. 2017). Secondly, it is an important target for improvement in crop breeding (Long, Marshall-Colon et al. 2015). Modest improvements in photosynthesis efficiency by engineering photorespiratory pathways have demonstrated enormous yield gains (Kromdijk, Głowacka et al. 2016, South, Cavanagh et al. 2019). The yield model of Monteith (Monteith 1977) suggests that increased light use efficiency of photosystem II holds great potential to meet global food challenges by increasing the conversion efficiency of intercepted irradiance into biomass (ε_c) (van Bezouw, Keurentjes et al. 2019). Another determinant of plant growth rate is leaf area growth, involving precise regulation of photosynthesis machinery and growth hormones such as auxin (Zhang, Hu et al. 2017). Leaf area measurements from fluorescence based non-destructive optical phenotyping systems, can be efficiently used

to screen plants at different growth stages with varying levels of photosynthetic rates (Weraduwage, Chen et al. 2015). Therefore, improved GP models for these traits could have impact in future crop breeding.

Following Edwards et al., 2016, we studied accuracy on internal test sets within the HapMap population. Further work is needed for data-driven selection of the most relevant terms for prediction on external test sets. For example, a possible strategy may be to select the feature with highest genomic variance explained, or with lowest p-value in the LRT we described. Our results indicate that biological priors driven GP models can be used to rank groups of genes potentially associated to the trait of interest along with improving prediction performance. The GWAS conducted on the same HapMap population for photosynthetic light use efficiency of photosystem II identified that the *A. thaliana* 'Yellow Seedling 1' gene is involved in photosynthesis acclimation response (van Rooijen, Kruijer et al. 2017). This *YS1* gene is annotated with GO Cellular Component terms chloroplast, intracellular membrane-bounded organelle and mitochondrion and GO Biological Process terms thylakoid membrane organization and photosystem II assembly. Our results using GO and COEX GP (*Table 3.1*) clearly demonstrate that these GO terms were most prevalent to improve the prediction and explain a large amount of genomic heritability. This indicates that genomic prediction and GWAS support each other as potentially useful tools for forward genetics.

The gain of predictive accuracy of the GP models compared to the base-model is trait-specific and negatively correlates with genomic heritability, which is promising for breeding at low h^2. This inverse relation may be due to the fact that we deal with highly polygenic, complex traits: many physiological and regulatory biological processes are involved in Φ_{PSII} under high light stress, e.g. PSII repair, ROX etc. Our models, testing groups of genes individually, may not be able to improve performance for such cases. Another potential explanation lies in the ability of GFBLUP to capture small genetic variance at low h^2 in a separate random component, potentially including known causal genes, which is not possible in GBLUP.

3.4.2 Exploiting biological knowledge to improve genomic prediction

With recent technological advances in both field and controlled environment high-throughput phenotyping systems, phenotypes can be measured at unprecedented scales. Phenotypes can vary in space and time due to genetics and environment alone, genotype-by-environment (GxE) interactions as well as stochastic and development effects. Component variances due to these factors can be calculated by precise modelling. If multiple measurements are available, GP models can be developed on individual measurements, treated as individual phenotypes, or on derived parameters, e.g. growth curves. We found that at each measurement timepoint, at least some GO (in

particular cellular component terms) or COEX group could help to improve performance, and some were more frequent (*Figure 3.4, Figure S7*). For example, for Φ_{PSII} no single GO or COEX gene group was capable of improving GP accuracy for all time points (either LL or HL separately), but a number of gene groups were able to improve PLA at multiple measurements (although not always meeting the threshold for significance). Phenotyping at an extended scale and GP modelling thus provides an opportunity to obtain biological insights. As an alternative to modelling at each timepoint separately, a whole time series or growth curve can be used instead. We did not pursue this here, as time series data is not generally available in most practical scenarios and we were interested to learn whether performance improvement was specific to growth stages and conditions e.g. models for Φ_{PSII} behaved differently under low and high light conditions.

Here, we mainly investigated two approaches to incorporate publicly available trait-specific biological information into GP, i.e. pre-selecting a list of genes and selecting sets or groups of genes based on predicted functional (i.e. GO) or expression (COEX) information. The approach using predicted functional information proved to be more useful in this context, but more approaches and sources of information can also be incorporated with a focus on prioritizing biologically related genomic regions. Moreover, knowledge from multiple heterogeneous sources can be combined to further pinpoint potential QTLs, termed as poly-omics GP models (Wheeler, Aquino-Michaels et al. 2014, Uzunangelov, Wong et al. 2020). These information sources may include (i) predicted variants effects, (ii) gene functions e.g. GO, COEX, (iii) networks of gene-gene and protein-protein interactions, stored in public resources like STRING (Mering, Huynen et al. 2003), GeneMANIA (Warde-Farley, Donaldson et al. 2010); (iv) pathways, in which genes are grouped e.g. KEGG (Kanehisa and Goto 2000); (v) previously generated GWAS and QTL results which indicate involvement of particular regions for specific traits e.g. AraGWAS (Togninalli, Seren et al. 2020) , AraQTL (Nijveen, Ligterink et al. 2017), (vi) known connections to phenotypes and (vii) endophenotypes, usually measured using -omics data at different stages of genetic information flow towards phenotypes. The reliability of these sources of information is an important factor for credible analysis. Information describing the (un)certainty of annotations is generally available in the form of a score (e.g. for gene functions based on GO evidence scores or reliability scores generated by a prediction method). It remains an open question how to incorporate such scores in the process of using the biological knowledge for GP.

Our first approach, pre-selecting a gene list, seems to be naive but can be useful as a baseline for comparison with more complex statistical procedures. The group based approach is usually based on gene function, but this heavily depends on computational prediction, as for most of the genes in plants and animals, no experimental function annotation is available (Radivojac, Clark et al. 2013). Function prediction is often based

on sequence similarity, which works well for predicting molecular functions but less so for biological processes. Using expression compendia based on multiple experiments poses an interesting alternative, since genes with similar expression patterns are more likely functionally related, hence more likely involved in the same biological process(es) (Kourmpetis, van Dijk et al. 2011). Alternatives are to define phenotype associated genomic regions based on differential gene expression levels (Fang, Sahana et al. 2017) or metabolite levels and metabolic fluxes (Tong, Küken et al. 2020), or to construct haplotypes in genic regions based on their ontology information (Gao, Teng et al. 2018). The GP requiring genomics inferred relationship matrices (GRM), e.g. GBLUP and its variants, can make use of information derived from these sources to construct a population variance-covariance structure (Zhang, Liu et al. 2010, Zhang, Ding et al. 2011, Fragomeni, Lourenco et al. 2017). A simple approach is to include multiple random effects for each knowledge source yielding its own variance-covariance structure for the population under study, in the mixed model equations (Guo, Magwire et al. 2016). One way to combine multiple omics datasets is to prepare a Composite Relationship Matrix (CRM) as a linear combination of Genomic Relationship Matrices (GRMs), Expression Relationship Matrices (XRMs), Metabolome Relationship Matrices (MRMs), MicroRNA Relationship Matrices (miRMs) etc. (Wheeler, Aquino-Michaels et al. 2014).

3.4.3 Alternative models for genomic prediction

Linear mixed model (LMM)-based genomic prediction, as used in this work, makes use of raw genotypes and parameter regularization to estimate thousands of SNP marker effects using only a few hundred observations ($p >> n$), employing different prior statistical assumptions on these parameters. This makes the approach fairly simple and interpretable; therefore, biological knowledge can be incorporated straightforwardly by employing these statistical assumptions. But with the increase in the ratio between markers and available phenotypes, serious overfitting problems may be encountered in these models (González-Recio, Rosa et al. 2014), leading to a need to use prior knowledge in regularization. A more general set of statistical learning methods are Machine Learning (ML) methods for prediction and classification, capable of dealing with the dimensionality problem in a more flexible manner. In these methods, phenotypes are regressed on nonlinear functions of genotypes rather than raw genotype values, compromising model interpretability but potentially improving prediction performance. Several studies have reported the use of Support Vector Machines (SVM), Reproducing Kernel Hilbert Spaces Regression (RKHS), Neural Networks (NN), Random Forests (RF) and boosting (De los Campos, Gianola et al. 2010, Ogutu, Piepho et al. 2011) for genomic prediction. Still, low prediction accuracy remains a problem for complex traits. It will be interesting to further explore how biological knowledge can be incorporated into ML

approaches for GP. One way could be to involve a knowledge driven regularization-based approach as demonstrated for disease prediction in human (Deng and Runger 2013).

3.5 Conclusion

The wealth of publicly available transcriptomics and Gene Ontology based prior biological knowledge can be incorporated for genomic prediction of photosynthetic light use efficiency of photosystem II electron transport (Φ_{PSII}) and PLA. Significant improvement in prediction accuracy over the benchmark GBLUP model was obtained for several GO terms and COEX groups. This improvement is trait-specific and negatively correlates with genomic heritability; whereas, for projected leaf area we found more added value than for Φ_{PSII}. Many known photosynthesis-specific GO terms lead to improvements, providing evidence of the potential usefulness of this approach in future breeding practice. We foresee incorporation of heterogeneous prior biological information into machine learning algorithms as an active area of research in future.

3.6 Supplementary material

Supplementary information available at:
https://www.frontiersin.org/articles/10.3389/fgene.2020.609117/full#supplementary-material

Author Contributions

MGMA and PN provided the genotype and phenotype datasets, MF performed the analyses. DDR, ADJvD and HN were involved in designing the analyses and interpreting the results. WK helped with statistical analysis. MF wrote the manuscript with DDR, ADJvD, HN and SM. All authors read the final manuscript.

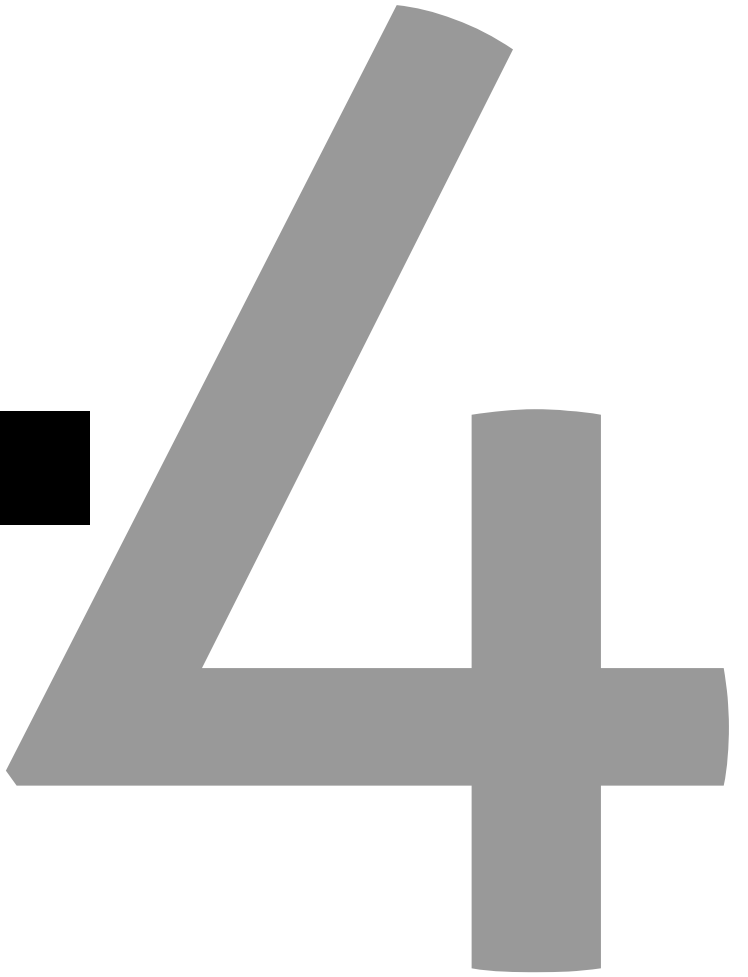
Acknowledgments

The authors are grateful for the support of the WUR sandwich PhD programme, funded by the Wageningen Graduate Schools. We are thankful to Pádraic J Flood of Plant Breeding, Wageningen University & Research for reviewing the manuscript.

Data Availability Statement

All data and scripts have been uploaded to the Wageningen University & Research git server (https://git.wur.nl/faroo002/pub1).

CHAPTER



PRIORNET: a framework for improving multilayer perceptron for genomic prediction using SNP prioritisation and protein-protein interaction information

Muhammad Farooq, Aalt D.J. van Dijk, Harm Nijveen,
Shahid Mansoor and Dick de Ridder

Abstract

Genomic prediction can benefit from the application of deep learning due to its ability to capture nonlinear relationship between genotypes and phenotypes (traits). However, given large numbers of single nucleotide polymorphisms typically used as input features in combination with the limited availability of training samples, standard models such as fully connected Multilayer Perceptrons (MLPs) are too complex. Weight regularization is usually employed to induce sparsity, and constraining complexity in a biologically informed way could improve generalization and interpretability. For this purpose, a wealth of trait-specific prior biological knowledge on gene functions and protein interactions can be utilised. The added value of using prior knowledge has been demonstrated for standard genomic prediction approaches, but it is still an open question how to incorporate it into deep learning. Here, we propose PRIORNET, an MLP using prior knowledge prioritisation along with sparsity induction using protein-protein interactions to improve prediction performance. The resulting approach drastically reduces model complexity and improves prediction accuracies compared to a benchmark fully connected MLP, for both simulated and real traits.

4.1 Introduction

Deep learning (DL) is regularly applied in computational biology due to its potential to learn complex nonlinear relationships among biological entities (Gazestani and Lewis 2019, van Dijk, Kootstra et al. 2021). Genomic prediction (GP) involves the prediction of phenotypes based on genotypes. The utility of DL for GP is still under discussion because of the typical small number of samples available, usually few hundreds, compared to millions of single nucleotide polymorphisms (SNPs) measured (Bellot, de Los Campos et al. 2018, Pérez-Enciso and Zingaretti 2019, Montesinos-López, Montesinos-López et al. 2021, Gill, Anderson et al. 2022). Hence, DL performance has been reported to be either comparable to or worse than other methods on simulated and real traits (Abdollahi-Arpanahi, Gianola et al. 2020). Performance is expected to be particularly poor for complex traits, where many SNPs with small to medium effects contribute to the phenotype (Azodi, Bolger et al. 2019).

The Multilayer Perceptron (MLP) class of neural networks has already been widely tested for GP (Montesinos-López, Montesinos-López et al. 2018, Khaki and Wang 2019, Montesinos-López, Martín-Vallejo et al. 2019, Zingaretti, Gezan et al. 2020, Sandhu, Lozada et al. 2021), where each SNP is represented as an individual feature at the input layer. The network is fully connected from input to the output, with all nodes (neurons) in subsequent layers connected by an edge with accompanying weight. For the number of SNPs that are currently routinely measured, the model may require estimation of millions of parameters, making it prone to overfitting due to the limited numbers of samples. A simple solution is to reduce the dimensionality of the feature space by selecting a subset of SNPs using feature selection methods, for instance based on random forest derived importance scores (Wang, Aggarwal et al. 2017) or on SNP effect estimates derived from linear models (Torstensson 2017). However, cascading a feature selector before a fully connected network makes the subsequent prediction by the neural network reliant on the precision of the feature reduction method.

To reduce the number of parameters throughout an MLP, regularization is often applied to induce sparsity, e.g. by including the ℓ_1 norm of edge weights in the loss function to shrink them towards zero (Lemhadri, Ruan et al. 2021). Strong shrinkage may however cancel out minor effects of many potentially causal SNPs with small to medium effects – characteristic for complex traits. Instead of removing some SNPs, or applying strong penalties, biological knowledge could be applied to give more weight to *a priori* known genomic loci with relevance to the specific trait of interest than to other SNPs (the "background"). This assumes that several (if not all) relevant loci are known *a priori* and that the true causal SNPs are located in or near these loci or in linkage disequilibrium (LD) with them (Gazestani and Lewis 2019). Basing network design fully on biological knowledge can yield highly transparent and interpretable models (Ma, Yu et al. 2018, van

Hilten, Kushner et al. 2020), but the (near) exhaustive prior knowledge required is not available for any complex trait. Here, we investigate the use of more limited prior knowledge to improve learning performance.

SNPs may have certain properties due to their nature, impact, genomic context and allele frequencies, based on which they may affect gene expression and ultimately the observable phenotypes. Therefore, it makes sense to group them based on their class, if known *a priori*. Such *a priori* knowledge could include SNP effect types (e.g. synonymous / nonsynonymous) and location in the genome (e.g. exonic / intronic / intergenic variants) etc. Moreover, genes have functional roles in the cell as indicated by their annotations (e.g. Gene Ontology), transcription patterns (e.g. co-expression) and post-translation behaviour (e.g. protein-protein interactions). The SNPs associated to genes may therefore, also be grouped by functional relation of genes to each other and to the phenotypes. Here, we explore two different types of prior biological knowledge relevant towards this end. First, knowledge of relevance of genes, variants or genetic loci to traits, for example derived from genome-wide association studies (GWAS), QTL mapping, gene expression studies or functional annotations etc. This information is deposited in public repositories and can be retrieved for traits. Based on this prior knowledge, SNP groups that are putatively functionally related to the trait of interest are prioritised (i.e. regularised less strongly) than others. The second type of knowledge involves relations between SNPs based on interactions of the genes they are located in. Protein-protein interaction (PPI) information, either experimentally measured or computationally predicted (Szklarczyk, Gable et al. 2019), are used to group sets of SNPs for sparse connections.

Taking both types of knowledge into account, results in sparser, better generalizable models. We test our approach using simulated and real traits with simple and complex genetic architectures. As a simple trait, we explore sodium accumulation in *Arabidopsis thaliana* leaves for which GWAS indicated a single strong effect QTL; and for complex traits, we explore flowering time and seed germination ability in dark.

4.2 Methods

4.2.1 GP Models

Benchmark MLP

A Multilayer Perceptron (MLP) is a fully connected feed-forward artificial neural network, with one or multiple hidden layers between input and output (*Figure S1*). In this study, the input layer contains a separate node for each SNP, while the output consists of a single node for the phenotype to be predicted. Each node of the hidden layers transforms the values from the previous layer with a weighted linear summation followed by an activation function:

$$h_{ij} = f(w_{1(j-1)}x_{1(j-1)} + w_{2(j-1)}x_{2(j-1)} + \cdots + w_{p(j-1)}x_{p(j-1)} + b_{ij}) \quad (4.1)$$

Here h_{ij} is the value of the i^{th} node in the j^{th} hidden layer, $x_{i(j-1)}$ is the value of the i^{th} node in the $(j\text{-}1)^{\text{th}}$ hidden layer, $w_{i(j-1)}$ is its corresponding learned weight and b_{ij} is the learned bias parameter. The function *f* is the activation function, which in our case is optimized (see below), except for the output node where a linear activation function is used. To train the network for prediction of quantitative traits, the mean squared error between network predictions and true values was used as a loss function, in combination with ℓ_2 regularization of the vector of network weights $\mathbf{w}$. The regularization term was optimised using the weight decay hyperparameter λ. This leads to the following loss function between network outputs $\hat{y}$ and target values y:

$$J(\theta) = \frac{1}{2n}\sum_{k=1}^{n} ||y_k - \widehat{y_k}||_2^2 + \lambda\|\mathbf{w}\|_2^2 \quad (4.2)$$

where $\boldsymbol{\theta} = \{\mathbf{w}, \mathbf{b}\}$ is the set of all parameters and *n* is the size of the training dataset. Hyperparameters include numbers of nodes per hidden layer, the activation function for all hidden layers, weight decay, learning rate, dropout, total number of epochs, batch size and the optimization algorithm. In cross-validation, the best hyperparameter combination for each fold of training data was determined using a second (inner) 5-fold cross-validation loop, based on the R^2 score between true and actual output. We employed grid search to optimise activation functions from [ReLU, Sigmoid, Tanh], optimizers from [Adam, SGD, RMSprop], dropout values from the set [0, 0.1, 0.2], learning rates from [0.01, 0.001, 0.0001, 0.00001], weight decay from [0.01, 0.001] and hidden layer sizes from [[1000, 50], [100, 50, 20, 10, 5], [50, 20, 10, 5], [50, 10], [20, 10], [10], [20], [50], [128, 64, 32, 16, 2]] (where the vector length indicates the number of hidden layers and the vector elements the number of nodes in each layer). Based on initial exploratory trials, the maximum number of epochs was set to 500, with early stopping enabled, and the batch size to 50. All models were implemented using the skorch wrapper api v0.12.0 in the PyTorch v1.13 machine learning framework, Python. The models were tested on CPUs and an NVIDIA Tesla T4 x16 PCIe Gen3 GPU. For computational time assessment per model, we used a CPU server with Intel(R) Xeon(R) CPU E5-2670v3 @2.30GHz, utilising a single core and a single thread.

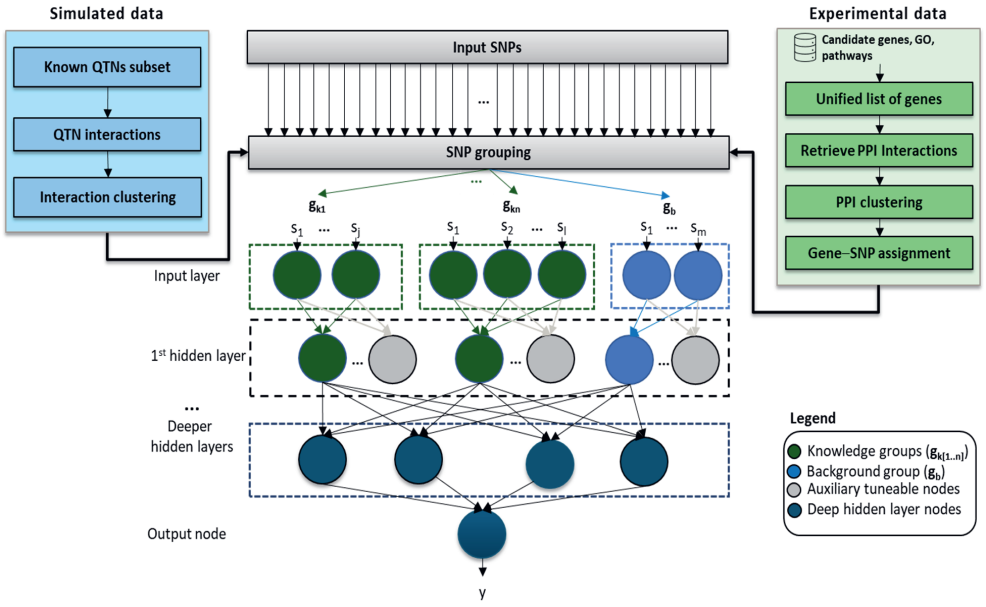


Figure 4.1: PRIORNET architecture.

The network has *p* SNPs at the input layer, one or more hidden layers and one output node. For the experimental data, list of genes, GO terms or pathway IDs available as prior biological knowledge are parsed until each SNP is associated to a group, i.e. either a knowledge group or the background group. On the other hand, for simulated data, SNPs are directly clustered and associated to the knowledge and background groups. These groups are connected to one or more nodes in the 1st hidden layer. The total number of hidden layers, number of nodes per layer, number of nodes per group, activation function and the optimiser are hyperparameters, found using grid search. Green or blue edges between input and 1st hidden layers indicate edges that are minimally required between them, whereas the nodes and edges in grey are auxiliary, tuned by the grid search. Similarly, black coloured edges are mandatory connections in the fully connected deeper hidden layers.

PRIORNET: genomic prediction using SNP prioritisation and protein-protein interaction information

Our proposed network architecture ("PRIORNET") is shown in *Figure 4.1*. The network is constructed as follows:

a) **Input:** knowledge is provided as a list of genes potentially related to the trait of interest, based on e.g. Gene Ontology terms or pathways.
b) **Interaction information retrieval:** a network containing interactions between all gene products (proteins) provided in the input knowledge, is retrieved from the STRING database (see below).

c) **Protein-protein interaction clustering:** possibly overlapping gene clusters that we call “knowledge groups” are found in the protein-protein interaction (PPI) network by identifying highly cohesive sub-networks.
d) **SNP assignment:** SNPs within genes, either with or without flanking regions, are associated to these knowledge groups.

All other SNPs, including those left out from the input list of genes (a) or those for which PPI interactions are unavailable, are merged into one separate “background group”.

e) **Pruning [optional]:** in case of high-dimensionality, each SNP group can be filtered using some feature selection method for dimensionality reduction (see section ‘SNP pruner’).
f) **Prioritisation:** nodes in the input layer are divided into two sets: one for the knowledge groups and the other for the background group. The number of nodes in the 1st hidden layer connected to all knowledge groups and the number of nodes connected to the background group are tuned hyperparameters, influencing the prioritisation of prior knowledge.
g) **Knowledge group node assignment:** the number of nodes for each individual knowledge group in the 1st hidden layer is calculated based on PPI sub-clustering; see equation (4.3) below for details.
h) **Mapping input to the 1st hidden layer:** input SNPs in each knowledge group are connected to their designated set of nodes in the 1st hidden layer; the background group is fully connected to another single set of nodes.
i) **Network:** the 2nd hidden layer and onwards are fully connected.
j) **Output:** the output layer consists of one node, outputting the predicted trait value.

Hyperparameters were tuned as for the benchmark MLP, except that for nodes in the first hidden layer two separate sets of hyperparameters were used, one for the designated knowledge group(s) and the other for the background group. The total number of nodes assigned to all knowledge groups (n_k) and the background group (n_b) was selected from {9, 19, 49, 99, 127, 999} and {1, 10, 20, 50, 100}, respectively. Note that the search space of network architectures for both MLP and PRIORNET is the same, but both can choose different architectures based on hyperparameter tuning. This implies that, for each n_k tested during the hyperparameter optimisation, each individual knowledge group is assigned nodes proportional to its size with respect to all SNPs in all knowledge groups. For instance, a group g with m_g SNPs will be assigned n_g nodes as follows:

$$n_g = n_k \left\lfloor \frac{m_g}{\sum_{g=1}^{G} m_g} \right\rceil \qquad (4.3)$$

Random forest

A random forest (RF) is an ensemble of decision trees, using bagging with random SNP subsampling to grow the trees. We used the Python Scikit-learn module (Pedregosa, Varoquaux et al. 2011) with TPESampler in optuna (Akiba, Sano et al. 2019) for hyperparameter tuning. The numbers of SNPs considered at each tree node were selected from [p/3, p/4, p/5, p/6, $\sqrt{p}$] out of the total number of SNPs p. The number of trees was taken from [100, 200, 300, 400, 500] by optimization and the minimum fraction of samples required for splitting at an internal node was selected from [0.01, 0.05, 0.1, 0.2, 0.3] of the total number of training samples. Hyperparameter optimization of the random forest was performed in the same inner cross-validation loop as described above for the optimization of the neural network hyperparameters.

SNP pruner

PRIORNET incorporates an optional pruning step as an additional way to prevent over-parametrisation in dealing with high dimensional GP problems to. For this, one or more feature selection methods can be used before PRIORNET, in any combination. Here, we utilised a random forest for each group of SNPs, passing all SNPs scored with a positive importance value (Wang, Aggarwal et al. 2017, Li, Raidan et al. 2019) to PRIORNET. SNP importance values were calculated on the training data as the mean decrease in impurity. If a group is left without any relevant SNPs, it is left intact for learning by PRIORNET itself, to minimise reliance on the feature selector. The number of SNPs considered at each tree node was selected from [$n_g/3$, $n_g/4$, $n_g/5$, $n_g/6$, $\sqrt{n_g}$] for each knowledge group and from n_b for the background group, with n_g and n_b as defined earlier. PRIORNET models developed using this approach are referred as GROUPRF-PRIORNET in the text. Instead of this pruning of individual groups, we tested an overall pruning approach as well, i.e. one RF with all SNPs, which we refer to as RF-PRIORNET in the text. The hyperparameters for RF as an overall pruner were chosen from the same search space as described above for the RF as a predictor.

Genomic Best Linear Unbiased Predictor (GBLUP)

GBLUP solves a linear mixed model equation that models the genetic covariance between genotyped accessions in terms of their similarity, estimated using a relationship matrix based on genomic markers (VanRaden 2008):

$$\mathbf{y} = \mathbf{1}_n\mu + \mathbf{g} + \boldsymbol{\varepsilon} \qquad (4.4)$$

Here, $\mathbf{g}$ is an $n \times 1$ vector of the total genomic value of all individuals, captured by all genomic markers; μ is the overall mean; and $\boldsymbol{\varepsilon}$ is an n-vector of residuals. The genomic values and residuals are assumed to be independent and normally distributed as $\mathbf{g} \sim N(0,\mathbf{G}\sigma_g^2)$ and $\boldsymbol{\varepsilon} \sim N(0,\mathbf{I}\sigma_\varepsilon^2)$. Here, $\mathbf{G}$ is the genomic relationship matrix, calculated using

the rrBLUP package (Endelman 2011) in R and **I** is the identity matrix. GBLUP is computationally efficient because it assumes equal variance for all genomic values and is practically equivalent to ridge regression. We used the Bayesian implementation to solve the above equation, implemented in the R package BGLR (Pérez and de Los Campos 2014).

Least Absolute Shrinkage and Selection Operator (LASSO) regression

Lasso is an efficient regularization-based regression method for high dimensional data (Ogutu, Schulz-Streeck et al. 2012), solving the following model using least squares:

$$\mathbf{y} = \mathbf{1}_n\mu + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad (4.5)$$

Here $\boldsymbol{\beta}$ is an $n \times 1$ vector of SNP effects treated as fixed, that are multiplied with the n x p genotype matrix $\mathbf{X}$ containing genotypes (encoding AA:-1, AB:0, BB:1) for bi-allelic SNPs; μ is the overall mean and $\boldsymbol{\varepsilon}$ is an n-vector of residuals $\boldsymbol{\varepsilon} \sim N(0, \mathbf{I}\sigma_\varepsilon^2)$. The above model, with mean-centred variables, is regularized by implementing an ℓ_1 penalty with a weight λ on least squares estimates, optimizing the following function for $\boldsymbol{\beta}$:

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\{\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1\} \qquad (4.6)$$

We used the LassoCV implementation of the Scikit-learn Python module with default settings.

Elastic net regression

The elastic net (Ogutu, Schulz-Streeck et al. 2012) uses both ℓ_1 and ℓ_2 penalties as regularization terms for the model in (6), optimizing

$$\widehat{\boldsymbol{\beta}} = \left(1 + \frac{\lambda_2}{n}\right) \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\{\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_2\|\boldsymbol{\beta}\|_2^2 + \lambda_1\|\boldsymbol{\beta}\|_1\} \qquad (4.7)$$

Bayesian methods

The Bayesian regression methods use the linear equation (4.5) by assuming SNP effects as random and solve the linear mixed effects models with the Bayesian framework to estimate the SNP effects. The model is solved for posterior means of the SNP effects using a Gibbs sampler implemented in the R package BGLR (Pérez and de Los Campos 2014). There is a large set of Bayesian methods with different prior distributions (Gianola 2013); for example, BayesA, uses a scaled *t*-distribution; Bayesian LASSO (Park and Casella 2008) uses a double-exponential; and BayesCπ (Habier, Fernando et al. 2011) and BayesBπ (Meuwissen, Hayes et al. 2001) both utilise two-component mixture priors with point mass at zero and either a Gaussian or scaled *t*-distribution, respectively. Here we refer to BayesBπ as BayesB, as we fixed the hyperparameter π to 0.5 to assume a weak prior.

4.2.2 Model performance assessment

We evaluated model performance on test data using Pearson's correlation coefficient (*r*) between observed phenotypic values and predictions. To estimate prediction accuracy, we repeated the 5-fold cross-validation 100 times. This yielded 500 accuracy measurements for comparing each method using a non-parametric (Wilcoxon signed rank sum) statistical test.

4.2.3 SNP importance attribution

After training PRIORNET, SNP importance values were determined using the integrated gradients method (Sundararajan, Taly et al. 2017) implemented in the Captum Python library (Kokhlikyan, Miglani et al. 2020). This was done per fold in the cross-validation scheme, and average importance scores over all folds are reported.

4.2.4 Data

Simulated data

For a proof-of-concept, we used a public dataset of *n* = 599 historical wheat lines from the CIMMYT global wheat program (Crossa, Campos et al. 2010). The dataset contains contained *p* = 1,279 markers (coded as presence/absence of an allele, with minor allele frequency greater than 5%), genotyped using Diversity Array Technology (DArT). Polygenic complex phenotypes were simulated by assigning 10 randomly selected Quantitative Trait Nucleotides (QTNs) with additive and epistatic allele effects:

$$\mathbf{y} = \sum_{i=1}^{10} \beta_i^{\text{add}} \mathbf{q}_i + \sum_{i=2}^{5} \beta_i^{\text{int}} (\mathbf{q_1} \times \mathbf{q}_i) + \sum_{i=7}^{10} \beta_i^{\text{int}} (\mathbf{q_6} \times \mathbf{q}_i) + \boldsymbol{\varepsilon} \quad (4.8)$$

Here, β_i^{add} describes the additive effect of the *i*th QTN, where $\mathbf{q}_i$ is a vector containing the allele dosages for the *i*th QTN for all samples and β_i^{int} is a pairwise interaction effect of QTN 1 with QTNs 2-5 and of QTN 6 with QTNs 7-10. This results in two distinct sets of interacting SNPs. The additive effects β_i^{add} were sampled from $N(0, \sqrt{h^2})$, where narrow-sense heritability h^2 was set to 0.4. The interaction effects β_i^{int} were sampled from $\Gamma\left(1, \frac{2}{3}\right)$. Total epistatic heritability (h_e^2) , i.e. the sum of the interaction terms, was also set to 0.4, so the broad-sense heritability ($H^2 = h^2 + h_e^2$) was 0.8 and residuals were sampled from ~$N(0,\sqrt{0.2})$.

Experimental data

We used different real polygenic traits with both simple and complex genetic architectures reported in GWAS studies. As a relatively simple trait, we used sodium accumulation measured for a global representative population of *Arabidopsis thaliana* (Baxter, Brazelton et al. 2010). The genotype data contained *n* = 300 of the 1,307 RegMap accessions (Horton, Hancock et al. 2012) with *p* = 169,881 SNPs. Previous analyses by

Baxter, Brazelton et al. (2010) found one strongly associated QTL centred on the sodium transporter gene (*AtHKT1;1*).

For a more complex trait, we used flowering time in *Arabidopsis*, a well-studied trait with many associated genes (Bouché, Lobet et al. 2015). A genotype dataset (p = 183,366 SNPs) and accompanying a set of 107 phenotypic measurements including 23 flowering related traits were available from Atwell, Huang et al. (2010). All the flowering traits were strongly positively correlated and the GWAS analysis found many common genes. We used the "flowering time in green house" trait, due to its large broad-sense heritability and larger sample size (n = 166). As another example of a complex trait, we used "seed germination ability in dark", an early development-related trait from the same study. This comprises 93 samples collected in the absence of light, where the phenotype is the percentage of non-dormant seeds that can germinate during one week of cold exposure at 4^{o}C.

4.2.5 Prior biological knowledge

The PRIORNET architecture can take any list of genes, gene clusters based on Gene Ontology (GO) terms, pathways or other functional annotations related to the trait of interest as prior knowledge, which are eventually all converted to SNP groups. If GO terms are provided, then the genes annotated with these terms are retrieved. For the sodium accumulation trait, we used the "sodium ion transport" term (GO:0006814); for flowering time we used "flower development" (GO:0009908) and for seed germination ability in dark we used "seed germination" (GO:0009845). Genes attributed to these GO terms (or one of its children in the GO hierarchical structure); including all types of annotation evidence (i.e. both based on experimental data and based on computational predictions) were obtained from GO mapping of TAIR identifiers using tairGO2ALLTAIRS method in org.At.tair.db R package (Carlson 2019). For these genes, all predicted functional protein-protein interactions (PPIs) were retrieved from the STRING database (Szklarczyk, Gable et al. 2019) and used to create a network. This network may further contain fairly overlapping subnetworks, so ClusterONE (Nepusz, Yu et al. 2012) was used to find highly cohesive clusters. Finally, all SNPs within 2kb upstream or downstream of the genes in a cluster are collected as a knowledge group; all SNPs not assigned to a knowledge group or not having a PPI interaction end up in a background group.

The genetic variance attributed to SNPs in a knowledge group was quantified as the ratio h_k^2/h^2 for each knowledge group, by the following linear mixed model;

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{k} + \mathbf{b} + \boldsymbol{\varepsilon} \qquad (4.9)$$

The above model splits genetic variance into two random components (Edwards, Sorensen et al. 2016, Farooq, van Dijk et al. 2020), such that the total genomic value ($\hat{\mathbf{g}} = \hat{\mathbf{k}} + \hat{\mathbf{b}}$) is partitioned into the part captured by SNPs in the knowledge group ($\hat{\mathbf{k}}$) and the

part by SNPs in the background group ($\hat{\mathbf{b}}$). The effect sizes were distributed as $k \sim N(0, \mathbf{G}_k \sigma_k^2)$ and $b \sim N(0, \mathbf{G}_b \sigma_b^2)$ and residuals were distributed as $\boldsymbol{\varepsilon} \sim N(0, \mathbf{I}\sigma_{\boldsymbol{\varepsilon}}^2)$. Here, $\mathbf{G}_k = \frac{\mathbf{W}\mathbf{W}'}{m}$ is the additive genomic relationship matrix based on *m* knowledge group SNPs, where $\mathbf{W}$ is a matrix of *n* genotypes by *m* markers, centred and scaled such that its *i*th column $\mathbf{w}_i = \frac{(\mathbf{z_i} - 2p_i)}{\sqrt{2p_i(1-p_i)}}$. Here $\mathbf{z}_i$ is the *i*th column vector of the matrix $\mathbf{Z}$ containing minor allele counts (0, 1, or 2) as entries and p_i is the minor allele frequency of the *i*th marker. The matrix $\mathbf{G}_b$ was calculated in the same way on SNPs in the background group. The above mixed model was implemented using the BGLR (Pérez and de Los Campos 2014) R package.

For simulated data, QTNs were used directly as input instead of genes, such that each QTN represents one gene (*Figure 4.1*). Interactions were defined using equation (4.8) and the genes represented by QTNs were subsequently subjected to ClusterONE for cluster assignments. Since we defined pairwise interactions between QTN 1 to 5 and QTN 6 to 10, ClusterONE always yielded two well-separated clusters. We then utilised either both clusters or one of them as knowledge groups to reflect full or partial knowledge, respectively.

4.3 Results

The Multilayer Perceptron (MLP) is a densely connected neural network architecture in which each node in a hidden layer is connected to all nodes in the previous and next layer (*Figure S1*). Since typical GP datasets contain thousands of genotyped SNPs, an MLP used for GP has many edges between the layers and a corresponding large number of weights to optimize, based on a limited number of samples. While a fully connected first hidden layer may help capture the joint effect of interacting SNPs, removing likely unimportant edges based on prior knowledge may result in a simpler network with equivalent or better performance.

In PRIORNET (*Figure 4.1*), we select genes known *a priori* to be functionally related to the trait at hand, and cluster them by their protein-protein interaction network derived from the STRING database. SNPs in or near genes in such a cluster are referred as a knowledge group, and the corresponding network inputs are connected to separate nodes in the 1st hidden layer, which have no other incoming edges. All other SNPs are combined in a single background group and the corresponding inputs are connected to a different set of nodes. Network tuning then enables us to assign a higher number of nodes to the knowledge groups in total than to the background. This setup allows differential prioritisation of trait-specific knowledge aiming to increase the sparsity, generalization and interpretability of the network.

4.3.1 PRIORNET significantly improves the prediction accuracy of MLP

As a proof of concept, simulated phenotypes were used (see Methods), based on genotype data for which the number of SNPs (*p*) was roughly double the number of samples (*n*). Ten SNPs were randomly selected as causal, from here on called Quantitative Trait Nucleotides (QTNs), with both additive effects (h^2 = 0.4) and non-additive, pairwise interaction effects (h_e^2 = 0.4). To test the maximum predictive ability of PRIORNET, we used specific prior knowledge – i.e. all of the QTNs were provided as prior knowledge and grouped into two knowledge groups, using a QTN clustering step (see Methods). Since both QTN interactions groups had equal numbers of QTNs, they were assigned an equal number of nodes in the first hidden layer (equation (4.3)). Moreover, given the very specific input knowledge we expected hyperparameter optimization to assign the smallest number of nodes (ideally, one) to the background group. *Figure* S2A shows that network tuning indeed found one background node for most of the cross-validation folds.

The prediction accuracy of PRIORNET (*Figure 4.2A*) on test data is significantly higher ($p_{\text{Wilcoxon}} < 0.05$) than that of the benchmark MLP and all other methods. The accuracy of the parametric methods (Lasso, Elastic Net, BayesA and BayesB) is similar and approaches the narrow-sense heritability (h^2 = 0.4) as they can only capture the additive effects of SNPs (see Methods) (Farooq, van Dijk et al. 2022). The accuracy of the Random Forest is higher, presumably because it can capture interaction effects. The prediction accuracy of the benchmark MLP was worse than that of both the linear and ML methods. PRIORNET achieved the highest accuracy, close to the expected value of $r \sim$ 0.89, i.e. the square root of total simulated broad-sense heritability (H^2 = 0.8). This shows that PRIORNET can achieve high accuracy by prioritising the input prior knowledge and can be a useful alternative for genomic prediction, provided highly specific knowledge is used. Moreover, PRIORNET reduces model complexity by splitting the fully connected first hidden layer into subgroups of fully connected components. This reduces the total number of weight parameters up to ~10-fold compared to the benchmark MLP (*Figure S3*).

4.3.2 PRIORNET performance degrades gracefully with incomplete prior knowledge

The used prior knowledge may contain SNPs that do not contribute to the phenotype and thus add noise to the input. To investigate the impact of such noise, we moved randomly selected SNPs from the background group to the knowledge group, keeping the specific knowledge (all QTNs in the knowledge groups) intact. The results in *Figure 4.2B* show that when almost all background group SNPs (10 times the number of QTNs) were added to the knowledge groups, model accuracy as expected dropped to levels similar to those

of the benchmark MLP. For smaller numbers of irrelevant SNPs however, performance loss is accordingly smaller.

In realistic scenarios, biological knowledge about a trait is usually incomplete; hence, the background group may also contain relevant QTNs. To investigate how this affects PRIORNET performance, we assigned one QTN cluster (QTN 1-5) to the knowledge group and included the other QTN cluster in the background group. The resulting "partial knowledge" model indeed performed worse than PRIORNET (*Figure 4.2C*), but still better than the benchmark MLP. In conclusion, PRIORNET performance decreases when the used knowledge is incomplete or diluted, but degradation is graceful.

4.3.3 PRIORNET performance is sensitive to the inclusion of correct interactions

The PRIORNET architecture makes use of prior information at two steps. First, as trait-specific knowledge in the form of functionally relevant groups of genes, which is used to differentially prioritize weight assignment to knowledge and background SNPs. The other is interaction information between genes in the knowledge group, to create subgroups used to induce sparse connections between the input and the 1st hidden layer. This means that the network will have fewer edges between the input and the 1st hidden layer than the fully connected benchmark MLP and may thus be less prone to overfitting.

The protein-protein interactions may contain false positives and negatives, because they are predicted based on computational approaches. To investigate the advantage of incorporating correct interaction information in PRIORNET, we conducted an experiment using simulated data, providing false positive interaction information. Instead of two QTN clusters of five QTNs, as in section 4.2.5 and explained in the Methods, we created five clusters of two QTNs each. In each of these five clusters, both QTNs are from different original clusters (*Figure S4*). The right boxplot in *Figure 4.2C* ("False interactions") show that when false interactions are introduced, prediction accuracy drops a bit below that of the model with only true interactions (red dotted line). Together, these results illustrate that incorporating meaningful interactions (true positives) and pruning unnecessary connections can improve prediction accuracy and yield a simpler model.

4.3.4 PRIORNET works well on experimental data with simple genetic architecture

We next tested PRIORNET on real traits, first with a relatively simple genetic architecture. For this, we used the sodium accumulation trait in *Arabidopsis thaliana*, for which GWAS identified a large effect QTL, explaining large amount of genetic variance, on chromosome 4 centred at the *AtHKT1;1* gene (Baxter, Brazelton et al. 2010). We used the GO term of *AtHKT1;1,* i.e. "sodium ion transport", as prior knowledge, with another 15 genes annotated to it. The PPI network was retrieved for these genes, but no further subgrouping of this network was identified by ClusterONE, so PRIORNET used only a single knowledge group next to the background group. A total of 153 SNPs were located

within these genes, including their 2kb flanking regions. The knowledge is quite specific for the trait, as these SNPs explain a large proportion of the additive genetic variance (h_k^2/h^2: 81%, using equation (4.9)). We therefore expected the network to assign more nodes in the 1st hidden layer to the knowledge group than to the background group. This is this indeed the case, as shown in *Figure S2B*: the knowledge group was assigned [1000, 128, 100, 1000, 1000] and background group was assigned [20, 10, 1, 20, 1] nodes in the five cross-validation folds.

Note that we used SNP pruning before training PRIORNET for all real trait examples, since the input contained far too many SNPs compared to the number of samples ($p >> n$). We explored two ways of utilising a random forest before PRIORNET to prune features: one where the RF was trained on all SNPs (RF-PRIORNET) and a one where an RF was trained independently for each knowledge and background group (GROUPRF-PRIORNET). The latter approach was found to yield higher prediction accuracies (*Table 4.1*) for all real traits, possibly because variable selection in a subset of SNPs is easier. Alternatively, a fully connected benchmark MLP was also tested without and with pruning (MLP and RF-MLP, respectively).

The prediction accuracy of PRIORNET without pruning was higher, on average, than that of the benchmark MLP without pruning (*Table 4.1*), though statistically not significant. But for both GROUPRF-PRIORNET and RF-PRIORNET, it was significantly higher ($p_{wilcoxon} < 0.05$) than that of the benchmark MLP as well as that of the MLP after pruning (RF-MLP). Moreover, accuracy of RF-MLP was lower than that of the benchmark MLP, implying that pruning simplifies networks, but does not necessarily improve performance.

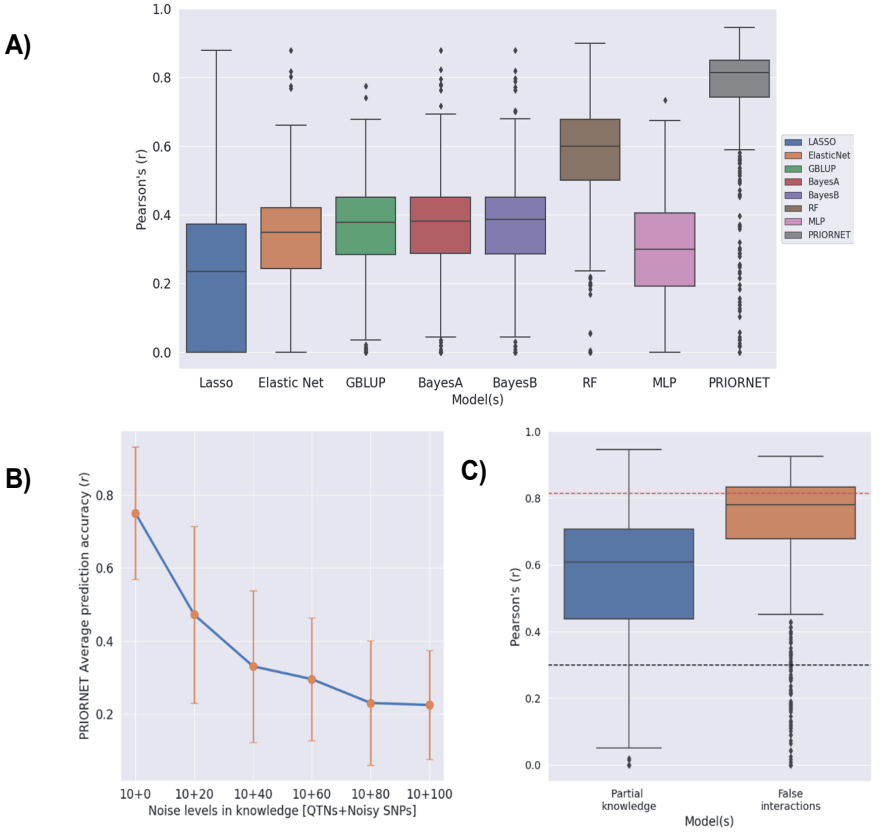


Figure 4.2: PRIORNET performance using simulated data.

A) Prediction accuracies of PRIORNET compared to different parametric (Lasso, Elastic Net, GBLUP, BayesA, BayesB), nonparametric machine learning (random forest, RF) and neural network (MLP) methods are plotted as average test accuracy over 100 times 5-fold cross-validation of simulated data. Here, accuracy is defined as Pearson correlation coefficient (*r*) between true and predicted values.
B) The influence of irrelevant SNPs (noise) in the prior knowledge was evaluated by adding up *n* SNPs from the background group (noise) into knowledge groups, containing *q*=10 QTNs in total. The value of *n* was chosen as {0, 2, 4, 6, 8, 10} times *q* (x-axis). Each dot represents the average prediction accuracy, and the error bars represent the standard deviation of the accuracy distributions.
C) The impact of partial prior knowledge and the advantage of including interaction information on prediction accuracies. Here, the red dotted line (top) is the median accuracy of PRIORNET without noise and when perfect knowledge is provided to the network; the black dotted line (bottom) is the median accuracy of the benchmark MLP. The left boxplot indicates the case when only half of the QTNs, i.e. *q*=5, were provided as knowledge. The right boxplot is the case when SNPs are still grouped into knowledge and background, but instead of mapping true QTN interactions between input and 1st hidden layer, incorrect QTN interactions were used for the knowledge groups (see *Figure S4B*).

4.3.5 Case study using experimental data with a complex genetic architecture

Complex traits are highly polygenic in nature and a large part of their genetic variance may be governed by non-additive allele actions (dominance, epistasis etc.). Moreover, many SNPs can have small effects that are challenging for a prediction model to pick up. We tested two traits related to growth and development: flowering time in the greenhouse and seed germination ability in the dark (Atwell, Huang et al. 2010). For the flowering time trait, we used all 492 genes annotated with the "flower development" GO term as prior knowledge. The associated SNPs explained ~55% of the additive genetic variance (h_k^2/h^2) for this trait. For seed germination, we used the GO term "seed germination", with 182 annotated genes that together explain ~47% of the additive genetic variance. The prediction accuracy of GROUPRF-PRIORNET is significantly higher than that of the benchmark MLP and RF-MLP for both these complex traits (*Table 4.1*), whereas PRIORNET without pruning and RF-PRIORNET tend to perform better than MLP and RF-MLP, though not significantly so.

Table 4.1: Performance of PRIORNET on real traits.
Five-fold cross-validated prediction accuracies (Pearson's correlation coefficient (*r*) between true and predicted values) of PRIORNET and benchmark MLP models for real traits, given as mean±standard deviation. Random forests were used for SNP pruning before both models using either all input SNPs (prefixed: RF-) or for each group (prefixed: GROUPRF-).

(b): Benchmark MLP model without pruning

(*): Significance assessed using Wilcoxon signed rank sum test between a models versus benchmark i.e. MLP^b.

#SNPs / #samples	Trait	Trait nature	Knowledge	#SNPs in knowledge groups	h_K^2/ h^2	MLP^b	RF-MLP	PRIORNET	GROUPRF-PRIORNET	RF-PRIORNET
~170k / 300	Sodium accumulation	Oligogenic	Sodium ion transport GO:0006814	153	81%	0.39±0.09	0.36±0.11	0.47±0.13	0.54±0.07*	0.56±0.12*
~183k / 166	Flowering time	Complex	Flower development GO:0009908	2,302	55%	0.66±0.11	0.64±0.11	0.65±0.12	0.71±0.07*	0.68±0.06
~183k / 93	Seed germination ability in dark	Complex	Seed germination GO:0009845	1,854	47%	0.35±0.21	0.28±0.13	0.45±0.14	0.49±0.12*	0.38±0.11

4.3.6 Genetic insights using PRIORNET

PRIORNET can be considered a semi-transparent architecture, given its explicit use of prior knowledge in connecting the input to the first hidden layer combined with “black-box” hidden and output layers. It can be relevant to inspect whether SNPs in the background group are associated to the trait, due to imperfect knowledge. This calls for *post-hoc* interpretation algorithms, which are increasingly popular for deep learning applications in -omics data analytics. These take a trained model and aim to identify relevant (combinations of) input features and quantify their importance for the model’s performance (Novakovsky, Dexter et al. 2022). Specifically, we used the integrated gradients (IG) method, that defines the attribution of a SNP as the integral of the gradients along the path from a baseline input to the actual input. The baseline is defined as the value of the input for which the network produces a reference value; in our case, zero input was used as baseline.

The results for the sodium accumulation trait (*Figure 4.3A*) show that the phenotype is mostly governed by the SNPs assigned to the input list of genes provided as knowledge (coloured green). All other SNPs received very small or zero importance values. For both complex traits, as expected, SNPs in the knowledge groups received high importance values compared to SNPs in the background group. However, many background SNPs were found to be important as well (*Figure 4.3B, Figure 4.3C*), possibly indicating that the used prior knowledge was incomplete. For instance, in close proximity of the top 3 important SNPs from the background group, three genes related to flowering were found, which were not in our input set based on prior knowledge (*Table S1*). These include genes involved in pollen tube guidance (Zhong, Liu et al. 2019), trichome morphogenesis (Matías-Hernández, Aguilar-Jaramillo et al. 2015) and regulation of sphingolipid biosynthetic process (Michaelson, Napier et al. 2016). Similarly, for the trait germination ability in dark, genes involved in root development (Bareke 2018) and response to salicylic acid (Lee, Kim et al. 2010) were found in the neighbourhood of important SNPs (*Table S2*). The latter is interesting given that salicylic acid is known to promote seed germination (Gao, Liu et al. 2021).

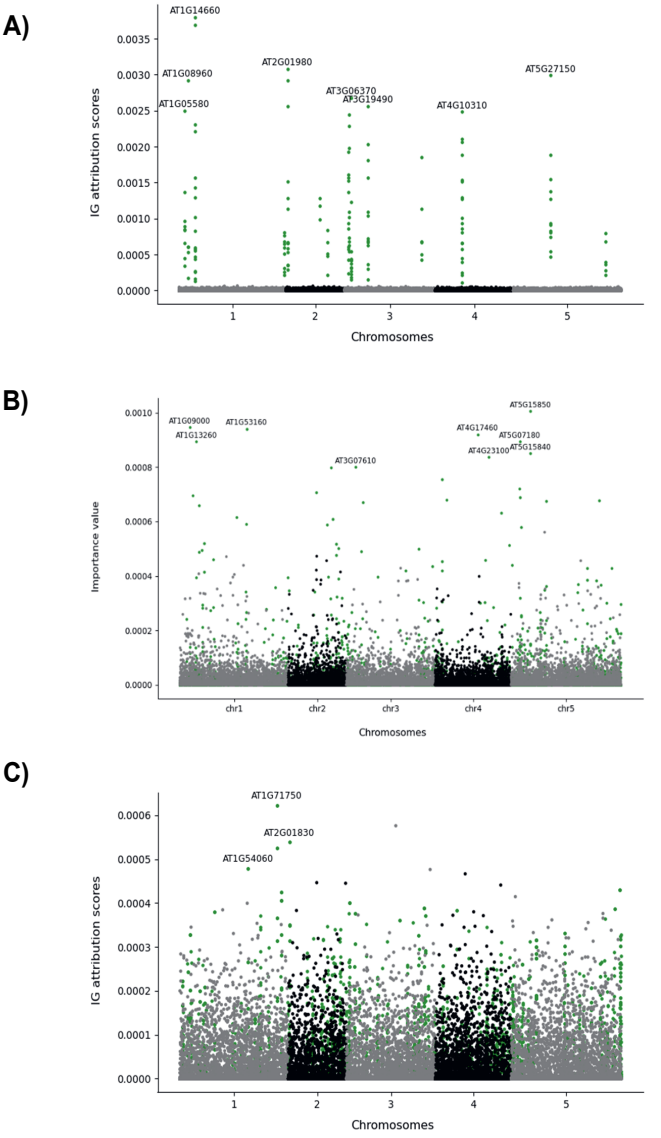


Figure 4.3: SNP importance values attributed by PRIORNET.

SNP importance as integrated gradient-based global attribution scores averaged over 5 folds. The top scoring SNPs from the knowledge group are indicated by associated gene IDs within 2kb genomic distance. If more than one closely located SNP is associated with a gene, then the gene ID is displayed once. The SNPs in the prior knowledge provided to the model are marked in green. On the *x*-axis, the marker positions on the chromosome are indicated from left to right. **A)** Sodium accumulation trait. **B)** Flowering time in greenhouse. **C)** Germination ability in dark.

4.4 Discussion

4.4.1 PRIORNET outperforms standard methods for GP

Our experiments with simulated data indicate that PRIORNET comes close to the maximum achievable performance when complete and accurate knowledge about a phenotype is available. Moreover, compared to both linear parametric methods and random forests, it outperforms all of these standard methods. On the other hand, the results with experimental data showed significant improvement over the standard MLP. In reality, prior knowledge is incomplete and noisy. Nevertheless, biological knowledge is being progressively generated for many complex traits. PRIORNET can make use of such knowledge to obtain improved prediction performance and potentially improve over other ML methods as well. However, the concept of knowledge prioritisation is more general, and other methods can also be tested for improvement with a similar strategy.

4.4.2 Pruning is key for PRIORNET for high dimensional data

The assumption underlying GP is that at least one SNP is in high LD with the QTL(s). Therefore, genotyping is conducted at high density (Meuwissen, Hayes et al. 2001). This may result in closely co-located, multi-collinear SNPs which add little or no information, next to possibly many non-associated SNPs. In contrast, training samples are usually limited. Consequently, input pruning is critically important, especially for complex models like (deep) neural networks (Abdollahi-Arpanahi, Gianola et al. 2020). Among different approaches proposed for dimensionality reduction (Manthena, Jarquín et al. 2022), we used variable importance scores obtained from random forests, where a positive value implies association to the phenotype (Li, Raidan et al. 2019). We compared calculating these on all SNPs (Wang, Aggarwal et al. 2017) (RF-PRIORNET, RF-MLP) or for each SNP group (GROUPRF-PRIORNET). The latter approach was found to perform better for both simple and complex traits, perhaps because it restricts the search space to a more meaningful set of SNPs to begin with. However, in principle PRIORNET can be combined with any other linear or nonlinear feature selection method(s), like lasso or evolutionary algorithms. An important distinction here is whether such methods can already take interactions between SNPs into account (like random forests or evolutionary algorithms) or not (like lasso).

4.4.3 PRIORNET is efficient for differential prioritisation of SNPs

Complex traits are governed by many SNPs with small to medium effects and possible inter- or intra-allelic interactions. Capturing these main and interaction effects in models is challenging, particularly under high dimensionality (many SNPs) and limited numbers of samples. Selecting potentially causal SNPs based on previous biological insights on the trait may thus help. Earlier studies using linear mixed models have demonstrated that such prioritisation based on prior knowledge could improve prediction accuracies in linear

mixed effect models (LMMs) (Edwards, Sorensen et al. 2016, Farooq, van Dijk et al. 2020). However, a limitation of such prioritisation into LMMs is that a SNP can only be in one of the random effects of the linear equation to avoid over-estimation of SNP effects. However, this is not the case for the non-parametric method like PRIORNET, where input SNPs from knowledge and background might be connected, in case a SNP is associated to the genes in both groups. This, in turn, will allow to learn nonlinear relations and more meaningful biologically, where a gene / SNP could contribute in multiple biological processes. Moreover, we find that PRIORNET indeed always allocates a higher number of nodes to SNPs in knowledge groups than to SNPs in the background group.

As an alternative approach, differential prioritisation could also be achieved by group-wise regularisation (Li and Li 2008) in which the network weights in the regularisation term (equation (4.2)) can be split into weights of knowledge and background groups and different regularisation strengths (λ) could be assigned. However, this only applies to weights between the input layer and the 1st hidden layer. PRIORNET imposes different penalisation to the knowledge versus the background group more comprehensively, since assigning more nodes in the 1st hidden layer yields a higher number of edges for the knowledge group in the 2nd hidden layer, and so on. The resulting network in fact can be seen as consisting of two subnetworks, whose weights are optimised simultaneously using a single loss function, but with the knowledge subnetwork being dominant.

PRIORNET induces sparsity by completely removing (hard masking) potentially unneeded connections between the input layer and the 1st hidden layer, based on sub-clustering of the PPI network. Alternatively, a soft masking approach could be adopted, where knowledge of known interactions could be utilised to only regularise the weights (Kang, Ding et al. 2017). This approach is computationally more expensive, as it involves multiplication of a masking function with the network weight matrix at each optimisation step. One disadvantage of our hard masking approach is that one might remove potentially important connections. PRIORNET minimises this risk by retrieving all computationally predicted functional interactions, instead of limiting ourself to only the experimentally verified directly interacting protein interactions (which will be incomplete).

4.5 Conclusion

We developed PRIORNET, a sparse MLP framework employing prior biological knowledge for improved GP performance. Given trait-related genes, PRIORNET can prioritise SNPs by assigning them more nodes than other SNPs, and it makes use of protein-protein interaction data to retain only biologically meaningful edges. The resulting network supports interpretation of connections between the input layer and the 1st hidden layer, but allows to learn further nonlinear relations through any number of subsequent hidden layers.

We have demonstrated the effectiveness of this approach using simulated data and tested on both simple and complex polygenic traits, providing related sets of genes derived from GO terms. We hypothesise that the PRIORNET approach can work for any type of prior knowledge that can be provided in the form of a group of trait-related genes. With increasing availability of reliable trait-specific prior knowledge of various types, the methodology can thus potentially improve prediction performance even further.

4.6 Supplementary material

Supplementary information available at: https://doi.org/10.5281/zenodo.7969553

Supplementary figures:

Figure S1: Multilayer Perceptron (MLP)

The fully connected Multilayer Perceptron model with p SNPs at the input layer, one or more hidden layers and one output node. For purposes of visualization, p is set to 7. The number of hidden layers, number of nodes per layer, activation functions and the optimiser are hyperparameters, found using a grid search. All nodes except the output node have a nonlinear activation function.

Figure S2: Prioritisation ability

The figure illustrates the prioritisation ability of PRIORNET by the total number of nodes on the 1st hidden layer assigned during hyperparameter tuning using equation (3) for each of knowledge and background groups. A) Simulated data: frequency of total nodes assigned to each of the 500 models (i.e. 5 folds cross-validations repeated 100 times); B) Sodium accumulation trait: total nodes assigned to each of the 5 folds of training data, where knowledge group was assigned [1000, 128, 100, 1000, 1000], and background group was assigned [20, 10, 1, 20, 1] nodes.

Figure S3: Comparison of model parametric complexities of MLP and PRIORNET.

The comparison of model parametric complexities in terms of total number of weight parameters between input ─ 1st hidden layers is shows for the simulated data. Different PRIORNET models are shown from left to right, by increasing number of noisy SNPs in the knowledge group. The noisy SNPs are the ones, not a QTN and part of the background group. For instance, PRIORNET [10+0] contains 10 QTNs and no noisy SNP, whereas, PRIORNET [10+20] contains 20 noisy SNPs along with 10 QTNs, and so on.

Figure S4: Simulated QTN clustering schemes

The figure illustrates two QTN clustering schemes used for simulated data. A) QTNs are clustered, using equation (4.8), into two groups, such that QTN 1─5 belong to the first cluster, and QTN 6─10 to the second; B) To illustrate the effect of false positive interaction information, the clustering scheme is modified for the same QTNs in (A), such that one QTN is picked from the 1st cluster and the second from the 2nd cluster.

Supplementary tables:

Table S1: Genes related to flowering residing in close proximity to the important SNPs from the background group for flowering time trait

Table S2: Genes related to seed germination residing in close proximity to the important SNPs from the background group for seed germination ability in dark trait.

CHAPTER

5

Improving genomic prediction of biomass using photosynthesis-related traits in *Arabidopsis thaliana*

Muhammad Farooq, Aalt D.J. van Dijk, Harm Nijveen, Mark G. M. Aarts, Thu-Phuong Nguyen, Raul Wijfjes, Shahid Mansoor and Dick de Ridder

Abstract

Total biomass is an important target trait for plant breeders. Recently, high-throughput non-invasive phenotyping systems have made it possible to monitor plant growth at high spatial and temporal resolutions. Alongside biomass, many other biochemical traits can be measured, including components of photosynthesis, an important determinant of growth. Given the availability of large genomic resources, a key task is to relate trait measurements to genomic variation. Here, we assess whether secondary photosynthesis-related traits can improve performance in genomic prediction of Projected Leaf Area (PLA), a proxy for biomass. We performed two-trait and multi-trait genomic prediction using the Genomic Best Linear Unbiased Predictor (GBLUP), considering individual as well as multiple measurements of the secondary traits for predicting each measurement of PLA. Our results show that, although there is a strong influence of the moment of measurement and light intensity perturbations on prediction accuracies, all photosynthetic traits can improve PLA prediction. We achieved up to a three-fold gain in accuracy over the single-trait model. We conclude that photosynthesis-related traits are genetically related to PLA and that two-trait or multi-trait genomic prediction using photosynthesis traits can improve PLA prediction.

5.1 Introduction

Leaf area determines the light interception capacity of a plant (Weraduwage, Chen et al. 2015). Growth in leaf area can be used as a proxy for plant biomass accumulation during the vegetative phase of development. With the development of high-throughput phenotyping (HTP) systems, leaf area can now be efficiently measured using Near Infrared Imaging (NIR) at different growth stages (Flood, Kruijer et al. 2016). HTP systems thus provide an opportunity to explore the genetic variation underlying biomass accumulation. Total leaf area is often determined by Projected Leaf Area (PLA), i.e. the area of the horizontal plane. PLA has a complex genetic architecture (Farooq, van Dijk et al. 2020), is a determinant of the total photosynthetic capacity (Hu, Lu et al. 2020), and can be measured from the chlorophyll fluorescence. HTP systems are capable of estimating PLA from fluorescence signals along with photosynthetic performance in a single experiment (Flood, Kruijer et al. 2016). But utilising photosynthesis is challenging due to its inherent properties, i.e. sensitivity to light dynamicity, temperature and other environmental variations (van Bezouw, Keurentjes et al. 2019). Given the apparent relationship between PLA and photosynthesis, it is still an open question to what extent genetic variation associated with these traits overlaps and if they can be improved together in plant breeding efforts, or if one can be utilised for improving the other.

Over the past two decades, plant breeding has been modernised by employing new molecular and computational tools, such as Genomic Selection (GS). The GS framework uses whole-genome SNP markers to develop a genomic prediction (GP) model on a training population for which both genotypes and phenotypes are known. This model can then be used to predict the total genomic value of individuals in a test population, for which only genotypic information is available (Meuwissen, Hayes et al. 2001). The GP model is central to GS-based breeding programs and its prediction accuracy determines overall selection accuracy. The accuracy of GP depends on combined effects of population properties, genetic complexity factors and the modelling framework (Farooq, van Dijk et al. 2022). Different strategies, such as incorporating population and trait properties as part of the modelling framework, can be employed to improve GP predictive ability. Using *a priori* information on genes, proteins and their interactions (Edwards, Sorensen et al. 2016, Fang, Sahana et al. 2017) or estimating endophenotypes based on -omics measurements (Lozano, del Carpio et al. 2017, Haile, N'Diaye et al. 2020) has resulted in promising improvements.

In its most basic form, GP is performed as a single-trait approach (ST-GP). However, the capability of HTP systems, in particular, to measure multiple related traits together, inspires modelling two or more traits simultaneously in genomic prediction, so-called multi-trait genomic prediction (MT-GP). The relationship between traits is exploited by including them in the same prediction equation, with the aim to improve prediction performance compared to ST-GP. In practice, when so-called secondary or component

traits are added with the goal of predicting a specific trait of interest, it is preferred that the secondary trait(s) is (are) easy and cheap to measure. For example, multiple spectral reflectance indices, representing plant physiological and biochemical characteristics, can be calculated using spectral radiometry and used as secondary traits (Sandhu, Mihalyov et al. 2021). In addition, secondary traits preferably have higher heritabilities than the primary trait (Velazco, Jordan et al. 2019, Bhatta, Gutierrez et al. 2020, Arouisse, Theeuwen et al. 2021) and share sufficient genetic variation. The level of shared genetic variation, called genetic correlation, indicates the extent to which MT models can improve over ST. For instance, in case of a single secondary trait, the accuracy on a target trait improves when its heritability is lower than the heritability of the secondary trait times the squared genetic correlation (Schulthess, Wang et al. 2016, Arouisse, Theeuwen et al. 2021).

Many studies have assessed the potential of MT-GP, using both simulated and real traits, using different approaches. For instance, secondary traits can be modelled as an additional random effect of their phenotypic values in a univariate mixed linear model (Azodi, Pardo et al. 2020, Arouisse, Theeuwen et al. 2021). Alternatively, a multivariate model can be used to jointly model all traits with a joint distribution accounting for their genetic (co)variance (Bhatta, Gutierrez et al. 2020). For instance, a multi-variate Genomic Best Linear Unbiased Prediction (GBLUP) can be constructed based on a common Genomic Relationship Matrix (GRM) for all traits, with the implicit assumption of equal (co)variance of all loci. However, Karaman, Lund et al. (2018) extended it with heterogeneous SNP (co)variances for multiple traits derived from Bayesian regression based SNP effects of the individual trait. In case of large numbers of secondary traits, penalized selection indices-based BLUP (SI-BLUP) was proposed by Arouisse, Theeuwen et al. (2021). SI-BLUP reduces the number of secondary traits by estimating their penalised regression coefficients, which also has a potential application in cases when secondary traits are not measured for the test population. Atanda, Steffes et al. (2022) proposed a sparse-testing-aided MT-GP framework in cases where sparse phenotyping is used, i.e. when not all individuals in the training and/or test populations have all secondary traits measured. MT-GP methods have shown improved prediction performance over ST-GP in many applications, including predicting baking quality traits in wheat (Lado, Vázquez et al. 2018), harvest index, grain yield, grain number, spike partitioning index and fruiting efficiency in wheat (Shahi, Guo et al. 2022), grain yield and protein content in rye (Schulthess, Wang et al. 2016), agronomic and malting quality traits in barley (Bhatta, Gutierrez et al. 2020) and nutritional traits in pea (Atanda, Steffes et al. 2022). Given the success with which MT-GP can exploit relations between traits, it is interesting to explore whether photosynthesis-related traits can help improve prediction of leaf biomass, which in turn can shed light on the relation between these traits.

In this study, we used five photosynthesis parameters (Kramer, Johnson et al. 2004), including photosystem II electron transport efficiency (Φ_{PSII}), maximum quantum yield (F_v/F_m), non-photochemical quenching (NPQ), yield for other non-light induced energy losses (Φ_{NO}) and the fraction of absorbed light dissipated by NPQ (Φ_{NPQ}) as secondary traits to predict projected leaf area (PLA). We used either a single measurement of these traits, or measurements at multiple timepoints. To investigate practical breeding scenarios, we compared the use of two cross-validation schemes, with secondary traits either measured on both the training and test population or measured only on the training population. Each secondary trait was modelled individually together with each of the target PLA traits (2T-GP), and multiple photosynthesis-related traits were jointly modelled together with the PLA traits (MT-GP). This allowed us to answer three main research questions: 1) can we improve genomic prediction accuracy of PLA using photosynthesis traits?; 2) which of the photosynthesis traits is most suitable to do so?; and 3) can multiple measurements per trait for photosynthesis be advantageous?

5.2 Methods

5.2.1 Plant material and phenotyping setup

A set of 175 accessions from the Dutch *Arabidopsis thaliana* Map (DartMap) diversity panel was used together with the Col-0 and Ely accessions (Wijfjes 2021). Seeds were pre-germinated on filter paper with demi water in petri dishes by stratification treatment for five days. Subsequently, they were germinated at 24°C for 16 hours. The germinated seeds were transferred to 4×4 cm rockwool blocks to grow into plants. From this point onwards, experiment time is indicated as days after sowing (DAS).

The population was grown in a complete randomized block design in the Phenovator II system (PlantScreen Robotics XY system, Photon System Instruments™) embedded in the Netherlands Plant Eco-phenotyping Center (NPEC; www.npec.nl), a prototype of the Phenovator (Flood, Kruijer et al. 2016); as a result, eight blocks equivalent to eight replicates per accession of the DartMap were available. Plants were supplied with standard Hyponex solution for Arabidopsis in a climate chamber where day and night temperatures were 20°C and 18°C, respectively. The photoperiod was 12 hours (light on from 8.00 to 20.00), in which light intensity increased and decreased gradually in the first and last hour, respectively. The light treatment over the course of the experiment is described in *Figure S1* and contains four phases. Constant light treatment of 300 $\mu mol.m^{-2}.s^{-1}$ was given in phase 1 (the first 11 DAS) and phase 3 (17 to 20 DAS). Light fluctuation between 100 and 900 $\mu mol.m^{-2}.s^{-1}$ for five hours a day (from 11.15 to 16.15) was introduced in phase 2 (12 to 16 DAS) and phase 4 (21 to 25 DAS). A fast fluctuation regime (15 min frequency) was applied from 12 to 15 DAS and on 25 DAS; a slower fluctuation regime, with a frequency of 60 min, was applied 16 DAS and from 21 to 24 DAS.

5.2.1.1 Phenotyping, image processing and data collection

PLA and photosynthetic parameters were measured every day from 8 DAS until the end of the experiment at 26 DAS, 19 days in total. Pulse-amplitude modulated fluorescence analysis in dark adapted-state (in the night) results in the fluorescence kinetic components (F_m and F_o) and the deduced maximum quantum yield of PSII (F_v/F_m). Similarly fluorescence analysis in light adapted-state (during the day) provides fluorescence kinetic components (F_{mp} and F_p) and the deduced Photosystem II efficiency (Φ_{PSII}) (Murchie and Lawson 2013) were measured twice a day in the photoperiod starting at 9.45 and 17.45 after plants were allowed to acclimate at constant light of 300 µmol.m^{-2}.s^{-1} for at least 45 min. The maximum quantum yield (F_v/F_m) was measured once a day during the night, starting from 0.00. The plant size (projected leaf area, PLA) at night was also obtained once per day, at the same time as the F_v/F_m measurement. Images were processed with a mask correction of 3,500 for light adapted-measurements and 4,500 for dark adapted- measurements. The F_{mp} and F_p values at each Φ_{PSII} measurement and those of F_o and F_m in the night of the same day were used to calculate the NPQ, Φ_{NO} and Φ_{NPQ} photosynthetic parameters (Kramer, Johnson et al. 2004, Baker 2008). As a result, there are 19×2 timepoints for NPQ, Φ_{PSII}, Φ_{NO} and Φ_{NPQ} and 19 timepoints for F_v/F_m.

5.2.2 Phenotypic data analysis

The Best Linear Unbiased Estimates (BLUE) of each accession were calculated by treating genotypes as fixed effects and all other design factors as random effects:

$$y_{ijkmnl} = \mu + T_i + B_j + X_{jm} + Y_{jn} + G_k + \epsilon_{ijkmnl} \qquad (5.1)$$

Here, y_{ijkmnl} is the phenotypic measurement of the l^{th} plant of genotype k, on the i^{th} growth chamber table tray T, sown at coordinates (m, n) within the j^{th} block B. Here, X_{jm}, X and Y_{jn} represent the spatial position of kth genotype G_k within the block. Accordingly, the residual term of the corresponding plant is denoted as ϵ_{ijkmnl} and G contains the genotype labels as factors, for which BLUEs are estimated. The variances of each of these components were estimated by treating all factors in equation (5.1), including genotypes, as random effects. Subsequently, broad-sense (H^2) and narrow-sense heritabilities (h^2) of all traits were calculated as:

$$H^2 = \frac{\sigma_g^2}{\sigma_g^2+\sigma_t^2+\sigma_b^2+\sigma_{bx}^2+\sigma_{by}^2+\sigma_\varepsilon^2} \qquad (5.2a)$$

$$h^2 = \frac{\sigma_a^2}{\sigma_a^2+\sigma_\varepsilon^2} \qquad (5.2b)$$

Here, σ_g^2 is the total genetic variance among accessions, σ_a^2 is the genetic variance due to additive SNP effects only, σ_t^2 is the variance among trays, $\sigma_b^2, \sigma_{bx}^2, \sigma_{by}^2$ are variations between and within blocks respectively, and σ_ε^2 is the residual variance. Both equations (5.1) and (5.2) were implemented using the *lmer* function from the lme4 v1.1-31 R

package (Bates, Sarkar et al. 2007). One accessions (340) was left out from subsequent analyses due to missing data. Relationships among phenotypes were analysed using Pearson's correlation coefficient (*r*) of their BLUE estimates for all measurements. The dependence between multiple measurements of a trait was analysed by calculating autocorrelations, i.e. the correlation of a timeseries with lagged versions of itself. The autocorrelations were calculated for each sample individually and the mean value of all samples was reported.

5.2.3 Genotypic data analysis

The genotyping data was obtained from chapter 5 of the doctoral dissertation of Wijfjes (2021). In total, 5,972,233 variants (SNPs and INDELs) were retrieved, including variants in organellar genomes. These variants were filtered using PLINK (Purcell, Neale et al. 2007) to keep only biallelic SNPs for the 5 autosomes, with a minor allele frequency cut-off of 5% and squared pairwise correlation r^2 less than 0.9 among 500 adjacent SNPs. This left 302,783 SNPs for GP. The realized relationship or kinship matrix was estimated on these SNPs using GEMMA v 0.98.1 (Zhou and Stephens 2012). Accessions showing high redundancy based on their kinship estimates (1098, 1220, 1509, 2078, 2356, 2395, 2442, 40710) were removed from the analysis. In summary, after removing redundant accessions and the accession with missing data (described earlier), 166 accessions were available for GP.

5.2.4 Genomic prediction models

GP was evaluated using three models, single trait (ST), two-trait (2T) and multi-trait (MT), depending upon the number of secondary traits used. The ST model used no secondary trait, the 2T model used one secondary trait in addition to the primary PLA trait and the MT model used multiple secondary traits. All models were evaluated for each of the 19 daily PLA measurements; the 2T or MT models were either evaluated using one daily measurement of the secondary trait(s) or using multiple daily measurements.

The Genomic Best Linear Unbiased Predictions (GBLUP) for ST and 2T/MT were obtained using Bayesian ridge regression (BRR) with 6,000 burn-in iterations and 1,000 iterations of the Gibbs sampler. We used the *BGLR* R package (Pérez and de Los Campos 2014) for ST and the *MTM* package (de los Campos G 2013) for 2T/MT models.

The ST model is as follows:

$$\boldsymbol{y} = \mathbf{1}_n\mu + \mathbf{Zu} + \boldsymbol{\varepsilon} \qquad (5.3)$$

where **y** is an (n × 1) vector of BLUEs of a single phenotypic measurement of PLA for n genotypes, μ is the overall mean and **Z** is the incidence matrix linking genomic effects **u** to the phenotypes. Here, **u** is an (n × 1) vector, such that u_i is the genomic effect of the i^{th} genotype for the PLA measurement, assumed to follow a normal distribution **u** ~

$N(0,\mathbf{G}\sigma_g^2)$, where **G** is the genomic relationship matrix and σ_g^2 is the additive genetic variance.

The 2T/MT model extends equation (5.3) using a Bayesian multivariate Gaussian model that estimates an unstructured variance-covariance matrix between traits ($\mathbf{G_o}$) and residual matrix (**R**) (Lado, Vázquez et al. 2018, Atanda, Steffes et al. 2022), as follows:

$$\mathbf{Y} = \mathbf{1}_n\boldsymbol{\mu}^{\mathbf{T}} + \mathbf{ZU} + \boldsymbol{\epsilon} \tag{5.4}$$

where **Y** is the ($n \times q$) matrix of BLUEs of q traits, including one PLA measurement and one or more secondary trait measurements, for n accessions; $\boldsymbol{\mu}^{\mathbf{T}}$ is a ($1 \times q$) vector of means for each trait multiplied by the ($n \times 1$) vector of one's $\mathbf{1}_n$; **Z** is the ($q \times q$) incidence matrix linking predicted genomic effects **U** for all traits, where **U** is an ($n \times q$) matrix which follows a multivariate normal distribution $\mathbf{U} \sim N(\mathbf{0},\mathbf{G_o}\otimes\mathbf{G})$ and $\in$ is an ($n \times q$) matrix of residuals such that $\in \sim N(\mathbf{0},\mathbf{R}\otimes\mathbf{I})$. Here, $\otimes$ indicates the element-wise product; **G** is the ($n \times n$) genomic relationship matrix, which is prespecified, and $\mathbf{G_o}$ and **R** are the ($q \times q$) variance-covariance matrices of genetic and residual effects, which are estimated for all traits. The matrices $\mathbf{G_o}$ and **R** are estimated using a Gibbs sampler, assuming $\mathbf{G_o}$ is an unstructured matrix and **R** a diagonal matrix with uninformative scaled inverse Chi-square prior distribution (Schulthess, Wang et al. 2016).

5.2.5 Trait correlation inference

The relationship between PLA and photosynthesis parameters was studied in terms of their phenotypic correlation (r_P). The phenotypic correlation between two traits x and y can be defined as the sum of their coheritability ($c_{x,y}$) and coenvironmentability ($e_{x,y}$), where $r_{Px,y}$, $c_{x,y}$ and $e_{x,y} \in [-1, 1]$, with the constraint that $|c_{x,y}| + |e_{x,y}| \leq 1$ (NEI 1960, Falconer and Mackay 1996, Vasquez-Kool 2019):

$$r_{P_{x,y}} = c_{x,y} + e_{x,y} \tag{5.5}$$

The coheritability $c_{x,y}$ of any two traits is a measure of their genetic correlation, weighted by their heritabilities. On the other hand, coenvironmentability $e_{x,y}$ represents the joint influence of all factors that affect the observable relationship between traits and are not yet accounted for by genetic factors. For instance, the coheritability between a measurement of PLA **t** and a secondary trait **s** over all genotypes in a single measurement 2T model is:

$$c_{t,s} = \sqrt{h_t^2 * h_s^2}\; r_{A_{t,s}} \text{ , with } r_{A_{t,s}} = \frac{\mathrm{cov}(A_t, A_s)}{\sqrt{\mathrm{var}(A_t) \times \mathrm{var}(A_s)}} \tag{5.6}$$

Here, h_t^2 and h_s^2 are narrow-sense heritabilities and $r_{A_{t,s}}$ is the genetic correlation due to shared additive genetic effects between **t** and **s**. We estimated $c_{t,s}$ with the value of $r_{A_{t,s}}$ calculated using the realized unbiased estimation of inter-trait additive genetic variance-covariance matrix ($\mathbf{G_o}$), obtained using equation (5.4) (Sandhu, Mihalyov et al. 2021). The

off-diagonal elements of $\mathbf{G_o}$ are the covariances ($\mathrm{cov}(A_t, A_s)$), the diagonal contains additive genetic variance ($\mathrm{var}(A_t), \mathrm{var}(A_s)$) estimates of both traits.

5.2.6 Model performance assessment and cross-validation schemes

Given the (*n* x *m*) matrices of PLA ($\mathbf{P}=\{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_n\}$) and for each of the secondary trait ($\mathbf{S}=\{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n\}$), prediction accuracy of all models was calculated as Pearson's correlation coefficient (*r*) between the true and predicted values in a 5-fold cross-validation (CV) setting, repeated 10 times for each measurement of PLA. Accuracies were compared using an unpaired *t* test at a 5% *p*-value threshold. We used different cross-validation schemes (CVS) for both ST and 2T/MT, depending upon the availability of secondary traits. In CVS-0 (equation 5.7), accuracy was calculated based on true and predicted values of PLA as mentioned earlier. Conversely, for 2T/MT, two types of CV were used which differed in whether secondary traits were available in the test set. CVS-1 (equation 5.8) included all secondary traits for both training and testing, whereas CVS-2 (equation 5.9) only used secondary traits during training. In case of multiple measurements per secondary trait, all measurements until the day of PLA measurement were used for training (CVS-1, CVS-2) and for testing (CVS-2). For the i^{th} day measurement of PLA ($\mathbf{p}_i$), the response variable of equation (5.4) was divided into training and test sets according to the 5-fold CV. To summarize, for each of the training (trn) and test (tst) fold, the cross-validation schemes are as follows:

CVS-0 (ST model):

$$\mathbf{Y}^{\mathrm{trn}} = \{\mathbf{p}_i^{\mathrm{trn}}\} \tag{5.7a}$$

$$\mathbf{Y}^{\mathrm{tst}} = \{\ \ \} \tag{5.7b}$$

CVS-1 (2T/MT models):

$$\mathbf{Y}^{\mathrm{trn}} = \begin{cases} \{\mathbf{p}_i^{\mathrm{trn}}\} \cup \{\cup_j^N \mathbf{s}_i^{\mathrm{trn,j}}\}, & \text{single measurement analysis} \\ \{\mathbf{p}_i^{\mathrm{trn}}\} \cup \{\cup_j^N \{\cup_{k=1}^i \mathbf{s}_k^{\mathrm{trn,j}}\}\}, & \text{multiple measurement analysis} \end{cases} \tag{5.8a}$$

$$\mathbf{Y}^{\mathrm{tst}} = \begin{cases} \{\cup_j^N \mathbf{s}_i^{\mathrm{tst,j}}\}, & \text{single measurement analysis} \\ \{\cup_j^N \{\cup_{k=1}^i \mathbf{s}_k^{\mathrm{tst,j}}\}\}, & \text{multiple measurement analysis} \end{cases} \tag{5.8b}$$

CVS-2 (2T/MT models):

$$\mathbf{Y}^{\mathrm{trn}} = \begin{cases} \{\mathbf{p}_i^{\mathrm{trn}}\} \cup \{\cup_j^N \mathbf{s}_i^{\mathrm{trn,j}}\}, & \text{single measurement analysis} \\ \{\mathbf{p}_i^{\mathrm{trn}}\} \cup \{\cup_j^N \{\cup_{k=1}^i \mathbf{s}_k^{\mathrm{trn,j}}\}\}, & \text{multiple measurement analysis} \end{cases} \tag{5.9a}$$

$$\mathbf{Y}^{\mathrm{tst}} = \{\ \ \} \tag{5.9b}$$

Here, in CVS-1 and CVS-2, *N* is the number of secondary traits in the model (e.g. for 2T, *N* = 1) and **s** is a measurement for a secondary trait. This implies that $\mathbf{p}_i^{\mathrm{tst}}$ is predicted by the model given one or more secondary trait(s) in the test data (2T/MT) or given only PLA

(ST). Importantly, for the traits measured twice per day, both measurements of that day were added to the set, unless they were treated as separate morning and afternoon traits.

The gain/loss in predictive abilities of 2T/MT models compared to the baseline ST model was calculated as $\triangle d = \left(\left(med(r_{\mathrm{2T/MT}\frac{2T}{MT}}) - med(r_{\mathrm{ST}})\right)/med(r_{\mathrm{ST}})\right)$, where, $med(r_{\mathrm{2T/MT}})$ is the median accuracy over 10 repetitions of 5-fold cross-validation of the 2T or MT model, and similarly $med(r_{\mathrm{ST}})$ is the median accuracy of the ST model. Hence, $\triangle d$ indicates the relative increase or decrease in accuracy compared to the ST model.

5.3 Results

Multi-trait genomic prediction (GP) relies on shared genetic mechanisms between a target trait and one or more secondary traits. If secondary traits are genetically correlated with the target trait and have high heritabilities, they can add information to the univariate GP model. The traditional Genomic Best Linear Unbiased Prediction (GBLUP) models can be extended to incorporate this information by using the genomic (co)variance between traits (equation 5.4). Here, we investigate the relation between PLA and photosynthetic secondary traits by comparing single-trait models and multi-trait models.

5.3.1 Phenotype data description

Projected leaf area (PLA) was used as the target trait together with five photosynthesis-related secondary traits, including non-photochemical quenching (NPQ), maximum potential quantum efficiency of Photosystem II (F_v/F_m) and components of captured light destination, i.e. electron transport efficiency of photosystem II (Φ_{PSII}), ratio of incoming energy lost via non-regulated processes (Φ_{NO}) and the ratio of excited electrons energy that goes towards non-photochemical quenching (Φ_{NPQ}). All traits were measured for 19 days in total, from 8 to 26 days after sowing (DAS). Two light conditions, which are constant light and fluctuating light, were used to observe the sensitivity of the photosynthesis machinery and to mimic a realistic dynamic environment. For this, light intensity was fluctuated during 12-16 and 21-25 DAS, corresponding to days 5-9 and 14-18 of measurements, respectively. All traits except F_v/F_m and PLA were measured twice a day: early morning and late afternoon (see Methods, section 2.1).

The broad-sense heritability (H^2) for PLA was, on average, close to 0.32 until 16 DAS and then started decreasing gradually until slightly below 0.2 (***Figure 5.1)***. The broad-sense heritability of F_v/F_m consistently increased in the experiment, starting at 0.2, climbing up to that of PLA between 10-16 DAS and then increasing further to ~0.4. All other trait heritabilities fluctuated periodically between morning and afternoon measurements and reached their maximum at 22 DAS, ranging between 0.56-0.64. After reaching this maximum, there generally was a slow decrease until the final day of measurements. Heritability of Φ_{PSII} was highest on average for most days. The periodic

behaviour in the heritabilities of Φ_{NPQ} and NPQ was disturbed after the second fluctuating light treatment was applied from 21 DAS.

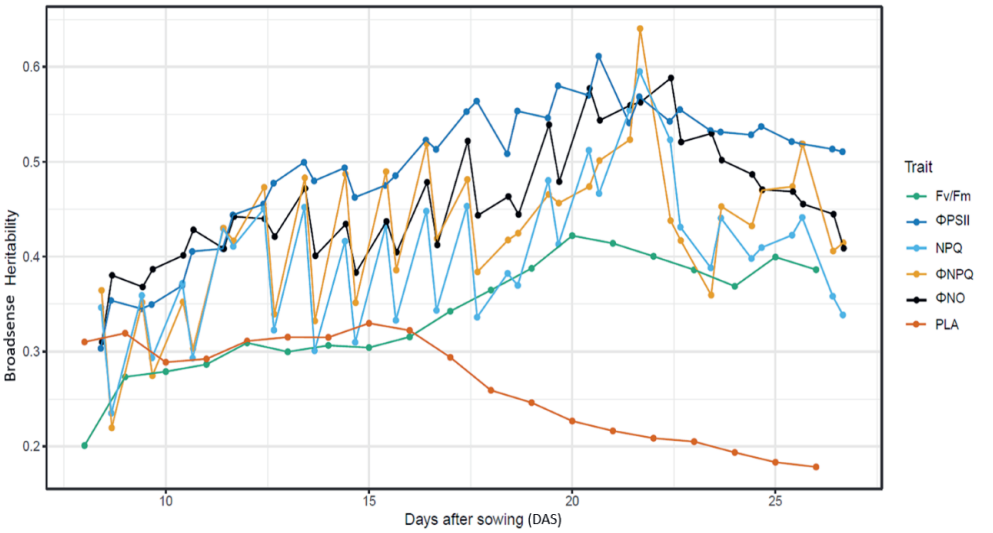


Figure 5.1: Broad-sense heritabilities.

Broad-sense heritabilities (H^2) of PLA and the five photosynthesis traits, measured from 8-26 DAS (19 days). Two measurements were recorded per day (early morning and late afternoon) for all traits, except for PLA and the photosynthesis parameter (F_v/F_m) which have only one measurement per day. Here, F_v/F_m was measured on dark adapted leaves, followed by PLA, and all other traits were measured simultaneously on light adapted leaves, either next morning or afternoon.

We used BLUE corrected phenotypes (BLUEs) in our GP modeling, calculated by adjusting raw measurements for the experimental design using equation (5.1). The BLUEs were correlated within as well as across traits. The correlations between subsequent measurements of a trait, as expected, were higher than with later ones (*Figure S2*). Moreover, as measurements clearly differed between morning and afternoon (*Figure 5.2A*), we analysed them separately in our analysis from here on. All traits, except Φ_{PSII}, were moderately correlated to PLA (*Figure 5.2B*). The Φ_{NO} trait was negatively correlated with all other traits and Φ_{PSII} was negatively correlated with Φ_{NPQ}. Since the phenotypic correlation can be broken down into additive genetic and residual components, its sign and magnitude do not immediately indicate whether the secondary trait can help improve GP of the target trait. Therefore, we evaluated all photosynthesis parameters as secondary traits in the subsequent analysis. Moreover, it would be interesting to know if Φ_{PSII}, if not as a whole, is genetically correlated to PLA for some measurements.

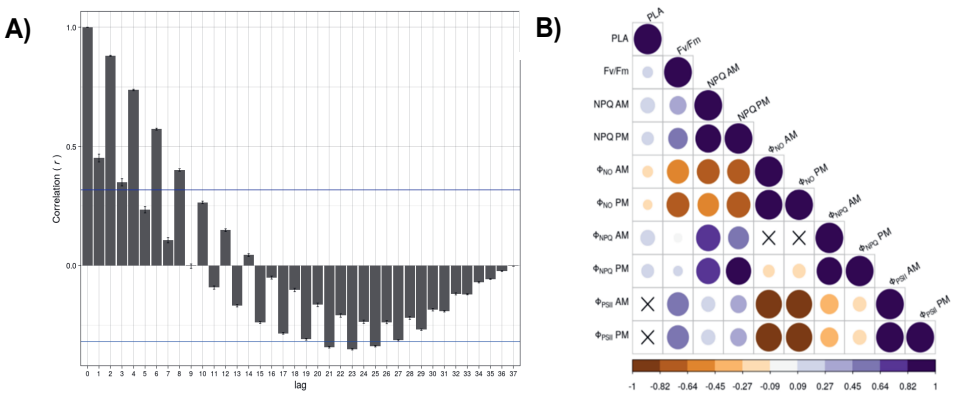


Figure 5.2: Phenotypic relationships.

Phenotypic relationship within and between traits. **A)** Autocorrelation among all (both morning and afternoon) Φ_{NPQ} measurements. Here, the blue lines represent the 95% confidence interval; **B)** Pearson's correlation coefficient (*r*) between all traits, separately for morning (AM) and afternoon (PM) measurements. Significance of the correlation was assessed based on two-sided correlation test, at a threshold of 1%. The crosses indicate non-significant *p*-values, dot sizes and colour both represent correlation.

5.3.2 Photosynthesis improves genomic prediction of PLA using multi-trait modelling

We first set out to investigate the effect of adding secondary trait(s) to the GP model, for each measurement (i.e. each time point). To this end, we used two cross-validation schemes, CVS-1 and CVS-2 (equations 5.8, 5.9). Both include secondary trait(s) in training, but only CVS-1 includes these in the test set as well. As a baseline, we predicted PLA starting from the first day of measurement (8 DAS) until the 19th day (26 DAS) using the ST-GP model and the CVS-0 cross-validation scheme. The multi-trait models then either included each secondary trait one by one, along with PLA at any time of measurement (2T-GP), or included more than one secondary trait (MT-GP).

a) 2T-GP analysis

We first predicted PLA with each of the secondary traits individually, measured on the same day (2T-GP). In general, the results in *Figure 5.3A* illustrate that compared to ST-GP, 2T-GP showed a gain in prediction accuracies (Δd), higher in the early days, and more moderate in later timepoints. This was the case for most secondary traits when they were included in the test set (CVS-1). On the other hand, accuracies did not improve under CVS-2 (*Figure S3A*).

Specifically, for CVS-1 the largest gain (Δd) at any timepoint was observed for NPQ in the morning (2.66×) followed by Φ_{NPQ} (2.28×) on DAS 9 (*Figure 5.3A*). No secondary trait gave consistent improvement in predicting PLA for all 19 measurements. On the other hand, no trait was consistently poor either. The use of Φ_{NPQ}-AM improved performance

most often (17 out of 19, *Figure S5A*), Φ_{NO}-PM the least (10 out of 19). These results illustrate that all photosynthesis parameters have the potential to improve PLA prediction to some extent, but that the timepoint of measurement is an important factor. Based on our findings, F_v/F_m, Φ_{NPQ}-AM and NPQ-AM have the largest potential to improve over the baseline model.

b) MT-GP analysis

Figure 5.2B indicates that the BLUEs of most of the traits are correlated (positively or negatively) to each other and to PLA. Moreover, all these fluorescence parameters intrinsically complement each other (Klughammer and Schreiber 2008). For example, the sum of quantum yields of photochemical energy conversion in photosystem II (Φ_{PSII}) and non-photochemical energy conversions (Φ_{NPQ}, Φ_{NO}) must be unity. Moreover, measuring F_v/F_m and Φ related traits differs only in the measuring protocol for total photosynthetic quantum yield, where the former uses dark adapted samples and the latter an illumination strategy. To learn whether combining multiple traits could further increase prediction accuracy, we tried different combinations of secondary traits along with the corresponding day measurement for PLA. The results using CVS-1 (*Figure 5.3B*) show that the maximum gain was slightly higher (~3.1×) on DAS 9 than for 2T-GP, when all Φ measurements (both morning and afternoon) were included. The use of all traits with either morning or afternoon measurements, or only Φ parameters with morning measurements, improved accuracy for most measurements (17 out of 19, *Figure S5B*), the use of all Φ parameters the least (13×).

Like for 2T-GP, when using CVS-2 (*Figure S3B*), generally MT-GP accuracy is comparable ST-GP, although slight improvements were observed when using the Φ morning and afternoon measurements together (from DAS 14 onwards, max Δd: 0.26× on DAS 21) or when only their morning or afternoon measurements were considered (DAS 20 onwards).

Based on this analysis, we conclude that in general, PLA prediction using multiple photosynthesis parameters in an MT-GP setting could only improve over the 2T-GP model to some extent. The improvement was most prominent when measurements taken in the morning were included in the model. Moreover, the gain varied over the growth trajectory, as for the 2T-GP model.

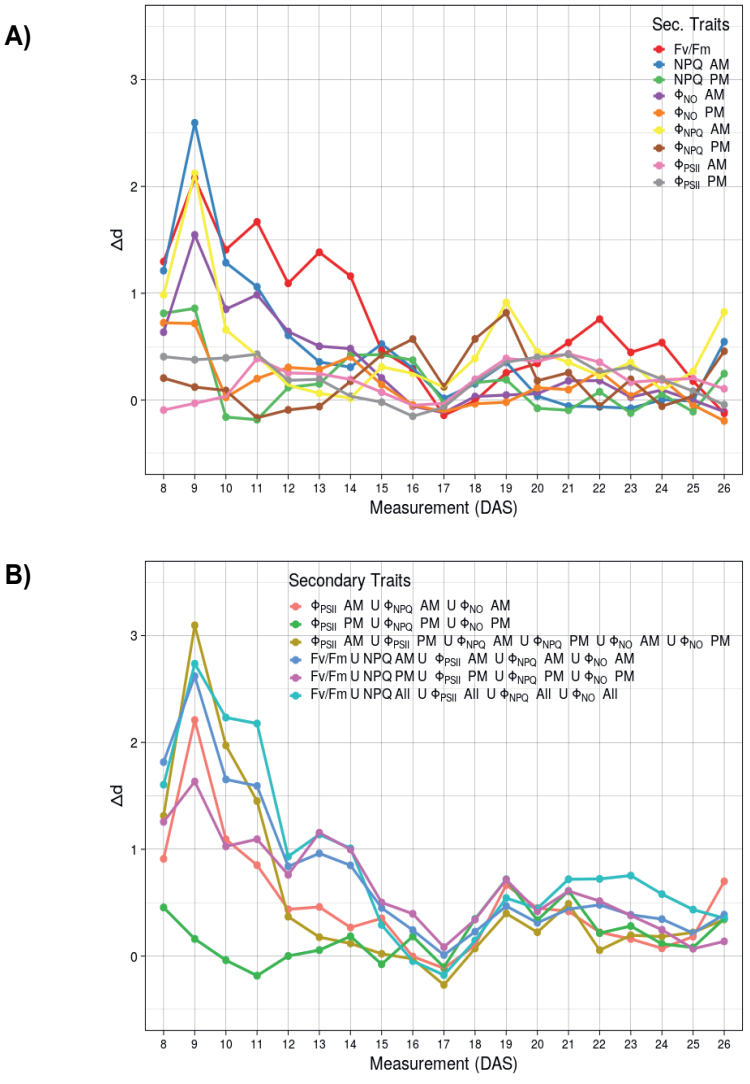


Figure 5.3: Accuracy improvement of 2T/MT-GP models using single measurements-based analysis under CVS-1.

Relative increase or decrease of median prediction accuracies (Δ*d*) of two-trait (2T) or multi-trait (MT) models versus the single-trait baseline. The accuracy for each model was calculated as Pearson's correlation coefficient (*r*) between true and predicted values of PLA, calculated using 10 times repeated 5-folds cross-validation (10×5 = 50 estimates). The secondary traits were available for the test data under the CVS-1 scheme, and each model was constructed using measurements from both target and secondary traits at the same time point. **A)** 2T-GP model, for each secondary trait with each measurement of PLA, where morning (AM) and afternoon (PM) measurements were considered separately. **B)** MT-GP model, for combinations of secondary traits with each measurement of PLA, and morning (AM) and afternoon (PM) were considered separately as well as together (ALL).

5.3.3 Incorporating multiple measurements consistently improves predictions under dynamic light conditions

a) Multiple measurement-based 2T-GP analysis shows improved prediction accuracy

The analysis above, based on single measurements, showed that accuracy improvement of 2T and MT models fluctuates considerably. The gain decreased after the first dynamic light treatment (DAS 12-16) and then increased slightly again (*Figure 5.3*). This could be because photosynthesis is sensitive to environmental factors like light fluctuations as well as growth stage, time of the day and measurement strategy (van Bezouw, Keurentjes et al. 2018). We therefore next investigated whether using multiple past timepoints on the growth trajectory could improve accuracy. Note that while we still use the terms 2T and MT to describe how many different types of traits were included, in this case also for 2T multiple measurements for the secondary trait were included. Next to the distinction between morning and afternoon measurements already in place, we further subdivided measurements according to the constant and fluctuating light treatments (see Methods, section 3.1). This gave a total of 39 trait sets to be tested. Results for the initial measurements (day 1-3) are expected to be similar to those of single measurement-based analyses, as little or no history is taken into account in the early days.

We first assessed the use of all traits individually in the 2T model, using CVS-1 (equation 5.8). The top section of the heatmap in *Figure 5.4A* shows that for all secondary traits accuracy improved, when all measurements (morning and afternoon) were used. On the other hand, morning or afternoon measurements used separately did not improve accuracy for *all* measurements, although NPQ-PM and Φ_{NO}-PM did do so for 17 out of 19 days (*Figure 5.4A*). Like before, highest gains were observed in the early days. When using only constant light measurements, results were quite similar to what was observed using all measurements. Under fluctuating light however, only some traits helped improve accuracy consistently, indicating that light dynamics are critical for photosynthesis-based analysis. The maximum gain (Δd) was observed for NPQ morning measurements (2.63) followed by Φ_{NPQ} (2.58×) on DAS 9 (*Figure 5.4A*), when both morning and afternoon measurements were included. Under constant light, these numbers slightly decreased to 2.52× and 2.48× on DAS 9 respectively. The average gain ($\overline{\triangle d}$) over all measurements ranged from 0.19× for Φ_{NO}-PM & Φ_{PSII}-PM to 1.45× for Φ_{NPQ}. Overall, these results illustrate that including past measurements in the 2T-GP model generally increases accuracy consistently.

The 2T models using CVS-2 again, in general, did not improve accuracy (*Figure S4A*), and instead tended to be worse than ST-GP for all traits. Nevertheless, small improvements up to ~0.10× were observed toward the final days of the experiment when only measurements under constant light were considered. This indicates that, though

multiple measurements could be useful, measuring secondary traits on the test data is important.

b) Multiple measurement-based MT-GP analysis robustly improves prediction accuracy

Above, MT-GP using single measurements improved overall prediction performance over the 2T model. To see whether this holds when multiple measurements are used, we tried different secondary trait sets in the MT model, as shown in *Figure 5.4B*. The results using CVS-1 illustrate that all traits help improved accuracy for all measurements, except for one timepoint for three traits under fluctuating light (*Figure 5.4B*). This is clearly an improvement over single measurement-based MT-GP analysis, where some traits did not consistently improve accuracy. Compared to 2T models with multiple measurements under fluctuating light, MT-GP results were less affected by light dynamicity, although the gain was similar. The maximum gain at any time for any trait was also similar (~2.84×) to single measurement-based MT analysis for the same day and trait (day 2, all Φ morning and afternoon measurements). Compared to multiple measurements-based 2T-GP analysis, the average gain ($\overline{\triangle d}$) over all measurements was slightly higher and ranged from ~1.55× for the traits set with all Φ morning and afternoon measurements to 0.43× for the same trait under fluctuating light. Together, this indicates that both 2T and MT models utilising multiple past measurements, improve over single measurement-based models, but that incorporating multiple traits mainly makes prediction gain over the growth trajectory more even.

Under CVS-2, none of secondary traits improved accuracy for all measurements, except when using all traits with all measurements, yielding a slight increase in accuracy gain (up to ~10%) in the early days of the experiment. In summary, the above results illustrate that multiple measurement-based MT-GP is preferred over 2T-GP as performance is more stable.

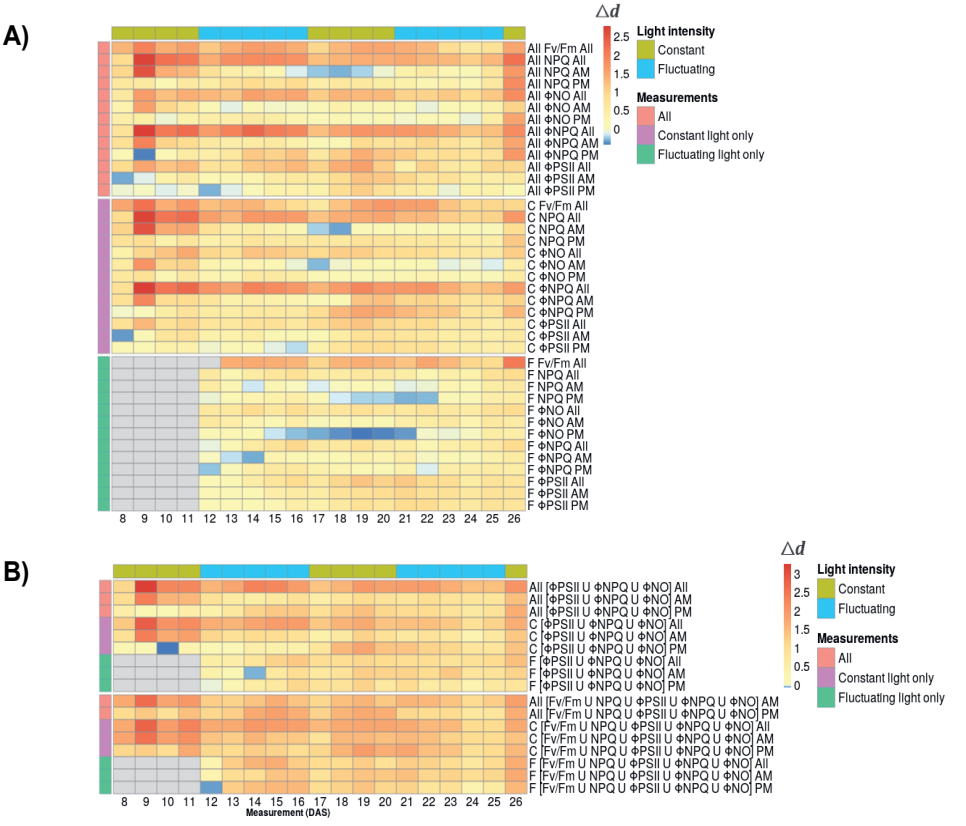


Figure 5.4: Accuracy improvement of 2T/MT-GP models using multiple measurements-based analysis under CVS-1.

Relative increase or decrease ($\triangle d$) in median prediction accuracies of two-trait (2T) or multi-trait (MT) GP models versus the single-trait (ST) baseline. Each model, developed under CVS-1, uses all past measurements until the moment of measurement of PLA. The median accuracy is estimated based on 10 times repeated 5-folds cross-validation (10×5 = 50 estimates). Each cell corresponds to the prediction of a specific PLA measurement (column) employing a certain set of secondary traits (row), indicated as [Light treatment-Name-Subset] on the right. The light treatments are 'All': all measurements, 'C': all measurements only under constant light, 'F': all measurements only under fluctuating light. The trait names indicate the photosynthesis-related traits Φ_{PSII}, F_v/F_m, NPQ, Φ_{NO} and Φ_{NPQ}, followed by 'AM'`: only morning measurements, 'PM': only afternoon measurements or 'All': all measurements. **A)** 2T-GP. **B)** MT-GP. The colours in the top row indicate the light intensity treatment applied for each timepoint, and the leftmost column indicates secondary trait sets, based on the light treatment.

5.3.4 All photosynthesis parameters are genetically correlated to PLA

The key underlying assumption for improved prediction performance of multi-trait GP is that the target and secondary trait(s) are correlated, either positively or negatively (Falconer and Mackay 1996). Analogous to the classical quantitative genetics theory, where phenotypic variance of a trait is divided into genetic and non-genetic components along with their interplay, phenotypic correlation between traits can be partitioned into shared genetic and residual correlations. This shared genetic component, or genetic correlation, describes the degree to which genetic variation is shared between the traits, known as pleiotropy, and could thus serve as an indicator for improvement in MT-GP (Schulthess, Wang et al. 2016). However, a larger genetic correlation may not translate into a larger shared genetic component if heritabilities are small. To this end, Vasquez-Kool (2019) proposed to weigh genetic correlations by the geometric mean of heritabilities, called coheritability (*c*).

To illustrate the relationship between secondary traits and PLA, we evaluated coheritabilities and gains in accuracies in the 2T models for single measurement-based analyses. Note that coheritability is not a static measure, and varies over the growth trajectory (*Figure S6*), ranged between -0.14 < *c* < 0.17. Results (*Figure 5.5*) show that the gain (△*d*) increased when coheritability increased in magnitude, either positive or negative, and not for smaller absolute values. The coheritability is non-zero for most photosynthesis parameters (*Figure 5.5, Figure S6*), indicating that they are genetically correlated with PLA, and can thus, in principle, be used to improve GP in a multi-trait setting. Specifically, Φ_{PSII}, that showed a non-significant overall phenotypic correlation (*Figure 5.2B*), still had non-zero coheritability for many measurements. Moreover, although we had a negative overall phenotypic correlation for Φ_{NO}, coheritabilities were positive towards later measurements on the growth trajectory.

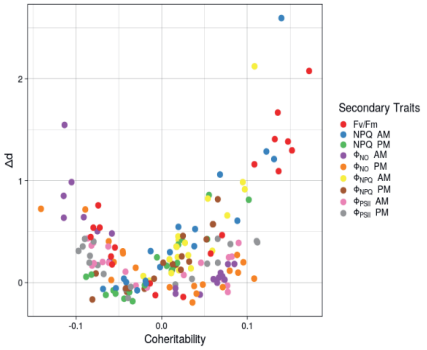


Figure 5.5: Relationship between coheritability and gain in accuracy.

Gains in accuracy (Δ*d*) vs. coheritabilities of each measurement of PLA with each secondary trait, measured on the same day.

5.4 Discussion

5.4.1 Multi-trait modelling allows to improve genome prediction of biomass

Exploiting relations between traits is interesting, because many genes have pleiotropic modes of action and selection drives their co-inheritance and co-improvement. The observable phenotypic correlation between traits can be quantified in terms of different genetic parameters, such as heritability, genetic covariance / correlation and coheritability, etc. Generally, MT-GP aims to exploit genetic correlation by estimating the effect of pleiotropy, along with the effects for all markers. This, in turn, requires a strong effect ─ either positive or negative ─ of the pleiotropic (common genetic) loci. Accordingly, genetic covariance / correlation should be high, provided narrow-sense heritabilities of both target and secondary traits are also high for a given population. However, most often target traits are fairly complex, with low heritabilities; so intrinsically, this encourages the use of secondary traits which are relatively simple and bear high heritabilities. Nevertheless, it is still an open question to what extent a particular complex secondary trait, with heritabilities comparable to the target trait, may be able to improve a target trait.

Plant leaf biomass and leaf area are complex traits, involving a plethora of genes and pathways turning on and off at different growth stages (Habyarimana, De Franceschi et al. 2020). Photosynthesis is another complex trait with well-known genetic pathways, but its application as a breeding target is usually limited because of practical constraints like variation between top and shaded leaves due to canopy architecture and high sensitivity to photoperiod, light intensities and stresses (van Bezouw, Keurentjes et al. 2018). Roughly 3,000 genes are known to be related to photosynthesis in the nuclear genome of *Arabidopsis thaliana* (Richly and Leister 2004). Although the core photosynthetic machinery is quite conserved, substantial polymorphisms with small effects in the nuclear genomes have been observed through mapping and association studies (van Rooijen, Kruijer et al. 2017, Oakley, Savage et al. 2018). Moreover, many QTLs have been identified for photoprotection sensitivity under fluctuating light conditions, through non-photochemical quenching (NPQ) measurements (Theeuwen, Logie et al. 2022) and for light use efficiency of photosystem II (van Rooijen, Aarts et al. 2015) in *Arabidopsis thaliana*. Here we could achieve ~3 folds improvement in accuracy of GP of PLA over the ST benchmark and found clear evidence of shared genetics between these traits. Therefore, even though they are genetically complex, using such secondary traits for improving predictions of another complex trait, like biomass, is appealing.

In addition to the genetic motivations, practical concerns like cost and ease of phenotyping are important factors for utilising photosynthesis as a secondary trait. The advent of high-throughput phenotyping systems both in the field and in growth chambers (Flood, Kruijer et al. 2016) now allows to measure different photosynthesis parameters

on large populations. Measurements are usually based on chlorophyll fluorescence to estimate maximum efficiency (e.g. F_v/F_m), photochemical harvesting efficiency (e.g. Φ_{PSII}) or losses (e.g. NPQ, Φ_{NPQ}, Φ_{NO}). These parameters can be easily determined by HTP systems, in real time or with a small lag with respect to PLA measurements. This provides a clear opportunity to improve complex target traits such as biomass.

5.4.2 Timeseries data is potentially useful for 2T/MT-GP

Our experimental design incorporated light fluctuation both at slow and fast rates (*Figure S1*) over a number of days. As a result, we observed high variation in prediction improvement between subsequent measurements, specifically during single measurement-based analysis (*Figure 5.3*). Specifically, a strong dip was observed on DAS 17 immediately after the fluctuating light treatment until DAS 16, after which gain gradually increased. Interestingly, no such dip was observed for heritabilities (*Figure 5.1*), but a similar trend was observed for coheritabilities (*Figure S6*). This suggests that genetic correlation is the underlying factor for improvement in MT-GP. The results also help explain why most secondary traits helped improve accuracy at early timepoints, despite having comparable heritabilities to PLA (*Figure 5.1*), but not so much at the later timepoints (*Figure 5.3*). The reason could be that the absolute value of coheritability between photosynthesis and PLA is somewhat higher for the earlier measurements than for the later ones (*Figure S6*). This may be because heritability of PLA gradually decreases after DAS 16 (*Figure 5.1*). Nevertheless, getting improved predictions at early stages is useful for practitioners.

Models including multiple past time points (CVS-1) are able to capitalise over the genetic correlations from the preceding measurements, and improve prediction accuracy rather consistently, including for the DAS 17. This indicates that it is useful to phenotype at multiple timepoints for photosynthesis-related traits to tackle dynamic light conditions. Note that in high resolution phenotyping, successive measurements might be correlated. Using large numbers of correlated secondary traits with a small number of samples may inflate the (co)variance estimates of the traits, resulting in imprecise genetic correlation estimates. These can potentially undermine the expected gain in accuracy through MT-GP. These issues can possibly be addressed by dimensionality reduction of the secondary traits (Falconer and Mackay 1996, Arouisse, Theeuwen et al. 2021).

5.4.3 Phenotyping schemes strongly affect proper utility of MT-GP

We considered two cross-validation schemes, i.e. measuring secondary traits on the training population as well as on the test population (CVS-1) or only on the training population (CVS-2). However, other schemes could be applicable, based on different practical scenarios. For instance, secondary measurements can be performed much earlier in the season, on the same plants or different plants of the same genotypes for

which target traits will be measured. This scenario is somewhat similar to our multiple measurement-based analysis under CVS-1, except that in that case secondary measurements were available until the time of PLA measurement. In another scenario, secondary traits may not be measured for a number of plants in the training and test populations, to conserve cost or due to measurement errors. This yields a sparse matrix of measurements –CVS-2 is an extreme case where no measurements are available at all for the test set. These missing values hinder the precise estimation of realized relationships between training and test data; therefore, we may not expect a clear advantage of measuring secondary traits on training data for predicting the test data. In this situation, joint modelling of traits (2T/MT) does not help, as is evident from our results. A possible workaround could be to use a two-step approach: first the genomic GBLUPs of the secondary trait measurements are used to predict the missing values; next, these predictions are used as random effects in a second univariate mixed effect linear model (ST-GP), as an additional explanatory variable, to predict the target trait (Arouisse, Theeuwen et al. 2021). Both scenarios investigated (CVS-1 and CVS-2) suggest that secondary traits are best be measured as extensively as possible, to achieve maximum prediction gain, as in *Figure 5.3* and *Figure 5.4* for CVS-1.

5.5 Conclusion

Photosynthesis-related traits are genetically correlated with PLA and can therefore be used in multi-trait genomic prediction to increase accuracy over using just a single trait. Using multiple measurements can make a model more resilient to fluctuating light conditions. Prediction accuracy improves only if secondary traits are measured on both the training and test population, but not when secondary traits were measured on the training population alone. This means that practical application for now is likely limited.

5.6 Data availability

All data can be made available from the corresponding author on request.

5.7 Supplementary material

Supplementary information available at: https://doi.org/10.5281/zenodo.7971910

Supplementary figures:

Figure S1: Experimental design with light intensity treatments

Measurement plan has been shown along with the light treatments, such that the red arrows indicate PLA and F_v/F_m measurements in the night, grey arrows indicate all other photosynthesis-related trait measurements. The first grey arrow is the measurement in the morning (AM) and the latter is in the afternoon (PM). The light treatment divides the total measurements into two main classes: 1) under constant light (DAS 8 – 11, 17–20 and 26); 2) under fluctuating light (DAS 12 –16 and DAS 21–25). The measurements under fluctuating light can further be categorised into either under fast fluctuating light

treatment (DAS 12–15 and 25) or slower fluctuating light treatment on DAS 16 and 21–24 (see Methods, section 5.2.1).

Figure S2: Phenotypic correlations within traits

Pairwise Pearson's correlation coefficients (r) between all measurements, for each a trait.

Figure S3: Accuracy improvement of 2T/MT-GP models using single measurements-based analysis under CVS-2

Relative increase or decrease of median prediction accuracies (Δd) of two-trait (2T) or multi-trait (MT) models versus the single-trait baseline. The accuracy for each model was calculated as Pearson's correlation coefficient (r) between true and predicted values of PLA, calculated using 10 times repeated 5-folds cross-validation (10 x 5 = 50 estimates). The secondary traits were available for the test data under the CVS-1 scheme, and each model was constructed using measurements from both target and secondary traits at the same time point. A) 2T-GP model, for each secondary trait with each measurement of PLA, where morning (AM) and afternoon (PM) measurements were considered separately. B) MT-GP model, for combinations of secondary traits with each measurement of PLA, and morning (AM) and afternoon (PM) were considered separately as well as together (ALL = {AM ∪ PM}).

Figure S4: Accuracy improvement of 2T/MT-GP models using multiple measurements-based analysis under CVS-2

Relative increase or decrease (△d) in median prediction accuracies of two-trait (2T) or multi-trait (MT) GP models versus the single-trait (ST) baseline. Each model, developed under CVS-1, uses all past measurements until the moment of measurement of PLA. The median accuracy is estimated based on 10 times repeated 5-folds cross-validation (10 x 5 = 50 estimates). Each cell of the heatmap for all PLA measurements (columns) employed a certain set of secondary traits (rows). The secondary traits are labelled as [Light treatment-Name-Subset]. The light treatments are 'All': all measurements, 'C': all measurements only under constant light, 'F': all measurements only under fluctuating light. The trait names indicate the photosynthesis-related traits ΦPSII, Fv/Fm, NPQ, ΦNO and ΦNPQ, followed by 'AM'`: only morning time measurements, 'PM': only afternoon measurement or 'All': all measurements. A) 2T-GP. B) MT-GP. The colours in the top row indicate the light intensity treatment applied for each timepoint, and the leftmost column indicates secondary trait sets, based on the light treatment.

Figure S5: Prediction gain per timepoint for 2T/MT models using CVS-1, under single measurements based analysis

The distribution of gain/loss of prediction accuracy (Δd) for all 19 days. Each dot represents the difference in medians of 10 times repeated 5-fold cross-validation accuracy estimates (10 x 5 = 50 estimates) of 2T/MT and the single-trait (ST) model. A) 2T-GP analysis. B) MT-GP analysis.

Figure S6: Coheritability over the growth trajectory

Coheritability for each timepoint for each secondary trait, using CVS-1 for single measurement-based analysis.

CHAPTER

6

Discussion

In this thesis, I demonstrated increased genomic prediction (GP) performance for Linear Mixed effect Models (LMMs) using prior biological knowledge on gene annotation and gene expression to group and prioritise SNPs. I also showed that using phenomics of correlated traits as additional information increased performance in LMMs. Inspired by the prioritisation approach, I proposed a novel deep learning framework named PRIORNET. PRIORNET is able to prioritise user-supplied knowledge for the trait of interest and accommodate partial, incomplete or noisy input knowledge.

Here, I discuss a number of issues related to knowledge-driven GP approaches, both generally and specific to this thesis. My work has contributed by demonstrating good performance of knowledge-driven approaches in certain circumstances, both by developing novel methods and through application. I will discuss remaining limitations, challenges and promises in practical scenarios below. Moreover, I will outline a number of topics in which I expect methodological innovations to empower future research.

6.1 Challenges for knowledge-driven GP methods

6.1.1 Benchmarking

Despite improvements realised in multiple studies (MacLeod, Bowman et al. 2016, Fang, Sahana et al. 2017, Lozano, del Carpio et al. 2017, Michel, Wagner et al. 2021), knowledge-driven GP is not yet widely used in practice. This could be due to the fact that improvements reported are generally inconsistent. For instance, leveraging transcriptomics data has been reported to only nominally increase prediction accuracies for casava (Lozano, del Carpio et al. 2017) while substantial improvement was observed for wheat (Michel, Wagner et al. 2021). Benchmarking these methods could be one way to address this challenge. In this context, different research questions should be answered, such as;

i. Which GP method should be considered as the benchmark for a given knowledge type and representation?
ii. How much knowledge is minimally required?
iii. What is the quantifiable impact of limiting factors including weak or uncertain knowledge and population properties?
iv. How much missing data in the knowledge is tolerable?
v. How are model metrics (e.g. learning curve, error rate, complexity) affected by differences in the quality and quantity of knowledge?
vi. Do models need to be retrained or adapted to continuously evolving knowledge?

Addressing these challenging research questions will be required to properly adopt knowledge-driven approaches in breeding programs. For instance, related to the first question: given the knowledge is represented as a lists of genes known a priori, the GFBLUP model (Chapter 3) can serve as a benchmark knowledge model. The model can

prioritise one group of SNPs, linked to each group of genes, at a time. The group of genes representing best performance for a trait, can then be used as a benchmark.

The impact of weak or uncertain knowledge can be studied by grouping knowledge based on its quality (Lee, Cohen et al. 2009). For instance, physically verified and computationally predicted PPIs can be grouped separately, that allows the model to learn coefficients based on their quality of evidence. The impact of missing data can be analysed in a comprehensive evaluation of various imputation methods on simulated data.

6.1.2 Knowledge source reliability

An important challenge in the use of publicly available datasets is their reliability. For instance, high-throughput experimental techniques for genome-wide function or interaction screens lead to spurious measurements; therefore, the presence of (considerable numbers of) false negatives / positives should be considered in the GP model (Deng, Sun et al. 2003, Gaudet and Dessimoz 2017). Also, gene ontology (GO) terms and protein-protein interactions (PPI) are largely computationally predicted, or transferred based on homology from model organisms to other organisms. The use of distantly related species for annotation can bias the overall distributions of annotations for a target species. While the quality of computationally inferred gene functions and interactions has increased over the years (Škunca, Altenhoff et al. 2012, Makrodimitris, van Ham et al. 2020), the proportion of experimentally verified annotations is often very small compared to the computationally predicted ones. Moreover, evidence codes represent only what type of experiment or analysis supports an annotation, and do not reflect the varying degrees of precision and confidence of different types of experiments.

These issues are addressed by some researchers. For instance, Wei, Zhang et al. (2020) proposed comparing the ratio of annotated genes for the same GO term across different taxonomic groups, to detect erroneous annotations. Buza, McCarthy et al. (2008) devised GO Annotation Quality (GAQ) scores for assessing quality of an annotation term as a product of its depth (i.e. distance from the root to an annotation term) with expert derived rank of the evidence codes. Lin, Liu et al. (2009) observed a higher likelihood score of a true PPI interaction when more high-throughput experiments reported an interaction, using a Bayesian network analysis approach. Another approach, that I took in Chapter 4 with PRIORNET, is to develop models that can somehow deal with noisy knowledge. I did not filter any input knowledge and left the distinction between useful and useless data to the GP models. I suggest to incorporate likelihood / evidence scores for prior knowledge as regularisation constraints in the loss function. Another option could be to use them for defining the initialisation distribution of network weights. It is important to realise, that most of the work done in this thesis uses the model plant *Arabidopsis*

thaliana, for which highly curated knowledge is available. This is not the case for many other plants, including commercial crops.

6.1.3 Knowledge representation

I used prior knowledge in the form of the functional classes of SNPs based on GO, COEX or a list of related genes (Chapter 3, Chapter 4) and SNPs / genes interactions from STRING (Mering, Huynen et al. 2003, Szklarczyk, Gable et al. 2019) retrieved as pairwise PPIs. In case of PPIs, similar information can be found in different databases, with variable quality, volume, schema and output formats. Examples include GeneMANIA (Mostafavi, Ray et al. 2008), BioGRID (Oughtred, Rust et al. 2021) and PlaPPISite (Yang, Yang et al. 2020) etc. In addition to high-throughput experimental approaches, PPI information can also be extracted computationally from other sources of information (Farooq, Shaukat et al. 2021), such as primary and 3D protein structures (Li, Gong et al. 2018, Bepler and Berger 2021), gene expression profiles (Chin, Chen et al. 2010) and the GO database. For instance, the GO is an ontology graph, organised as a directed acyclic graph (DAG) structure, representing relations between GO terms. Ieremie, Ewing et al. (2022) used a deep learning based 'transformer' architecture to learn functional interactions between genes, based on the semantic similarities between their GO terms in this graph. The variety of information sources, different ways to retrieve / compile interactions and types of interaction information (e.g. pairwise, transitive, higher-order interactions) make it important to represent PPIs with standardised probabilistic confidence scores (Lee, Date et al. 2004), in a generalisable format, that can directly be compared and formulated into regularisation constraints for the GP models (Roychowdhury, Diligenti et al. 2021).

Lists of trait-related genes can likewise be inferred from multiple sources, including co-localisation, co-expression or co-regulation in genomics / transcriptomics or from pathways. Such lists are continuously updated based on literature and different sources might contain varying information. Relying on one or a limited set of repositories will obviously not yield the best possible performance for the knowledge-driven models. Compiling all known information remains an open challenge.

Despite easy availability of knowledge in interpretable textual or markup formats, customised plugins / APIs are often required to pre-process the data for subsequent analysis. Developing a unified representation schema from a comprehensive set of knowledge, mined from a linked set of heterogeneous sources (Rueden, Mayer et al. 2023), could be a good way to address the issues arising from evolving knowledge and variable representations across repositories. For instance, web technologies offer data management platforms for storage, querying, retrieval and sharing of heterogeneous sets of linked biological data objects along with their semantics in the form of knowledge graphs, as semantic webs. For instance, Neo4j (Mondal, Do et al. 2022), AgroLD

(Venkatesan, Tagny Ngompé et al. 2018) and KnetMiner (Hassani-Pak 2017) provide data management platforms supporting the Resource Description Format (RDF) data model, processed with SPARQL as query language, and RDF schema and OWL standards for storing ontologies and vocabularies.

6.1.4 Knowledge integration

An important consideration is where in the GP model knowledge should best be used. In my view, this depends on the type of knowledge and its representation, algorithmic characteristics of the model and choice of the modeller. The literature reports on different approaches to incorporate knowledge, such as;

i. Integrating into the parametric assumptions / Bayesian priors
ii. Utilising knowledge for SNPs pre-selection
iii. Learning with knowledge-derived constraints
iv. Incorporating in model design / architecture
v. Forming an overall hypothesis
vi. Combining with the training data (individuals)
vii. Combining with the phenotype

Each of these approaches comes with specific issues. For instance, parametric linear models (including LMMs) can incorporate *a priori* information in the form of statistical assumptions on the unknown parameters, for example by letting statistical distributions reflect the genetic architecture (Chapter 3) and population characteristics. For this, the Bayesian framework provides an elegant way to specify priors (Gao, Li et al. 2015); which using Bayes equation yields posteriors. However, prior elicitation and its optimal specification remains an open challenge, because it is often difficult to transform domain knowledge of various types and representations into well-defined prior distributions (Mikkola, Martin et al. 2021, Tian, Lewis-Beck et al. 2023).

A second integration approach is to preselect SNPs before using them as model input, using any source of available information. Such information could be inferred from the training data itself, for instance by testing associations of SNPs with the phenotype; or based on prior biological knowledge. However, both approaches come with issues. For instance, SNP selection based on training data may yield poor performance, as accurate estimation of SNP effects with limited training data is challenging. On the other hand, using biological knowledge makes preselection dependent on expert judgement and requires highly curated knowledge, which is scarcely available in the real scenarios. Quantitative trait nucleotides (QTNs) can usually not be accurately located, so all SNPs in close proximity to the potentially causal genes/loci are selected. Nevertheless, preselection reduces the dimensionality and can increase prediction accuracy. In Chapter

4, I preselected SNP based on both model derived importance scores and trait-specific knowledge.

An obvious choice for prior knowledge incorporation is to formulate a constrained learning problem (Rueden, Mayer et al. 2023). This implies that knowledge about SNP grouping and interactions is added as a constraint to the loss function. For instance, separate regularisation terms can be used for SNPs related to the trait at hand and for the remaining ones. Similarly, different regularisation terms can be used for known and unknown interactions (Andel and Masri 2015). For the non-linear tree or network-based models, e.g. decision trees and artificial neural networks, interaction rules can be used to define node connectivity (Diligenti, Roychowdhury et al. 2017). However, biological regulation is complex for the complex traits, making translation of knowledge rules into simple model constraints challenging. For instance, instead of simple pairwise epistatic interactions, many complex traits are governed by tertiary or higher-order interactions (Saha, Perrin et al. 2022, Singhal, Veturi et al. 2023). Likewise, gene expression is regulated by a plethora of transcription factors and non-coding RNAs in complex regulatory networks (Moore, Amos et al. 2015). A way forward could be to base GP models completely on biological networks. In this case, the model design remains fixed (Ma, Yu et al. 2018, Snow, Noghabi et al. 2019, Fortelny and Bock 2020, Bourgeais, Zehraoui et al. 2021) and is interpretable and transparent. This is likely only suitable when highly curated knowledge of a trait is available; a more realistic approach is to let the model reflect the currently known biological network and to provide the means to learn new connections. PRIORNET (Chapter 4) uses such an approach, where certain parts of the network are based on knowledge, but hidden layers can still learn yet unknown SNP interactions.

A useful approach to compensate small training datasets is to exploit additional/secondary information on the training samples or the target trait. For instance, predictors other than SNPs can be scored for the same set of individuals to explain larger genetic variance for the phenotype, when combined with SNPs. Such predictors can include endophenotypes like transcriptomics, proteomics and metabolomics etc., or component / secondary traits (Chapter 5). Integration can use the previously mentioned strategies including constrained learning, design embedding (Huang, Chaudhary et al. 2017, Hu, Xie et al. 2019), or through use as separate covariance sources in linear models (Wheeler, Aquino-Michaels et al. 2014, Deniz Akdemir 2019). However, missing data often arises because not all biomolecules are measured in all samples, e.g. due to their difference in expression over space and time, measurement cost or instrumental sensitivities. While different imputation strategies have been proposed already (Flores, Claborne et al. 2023), this remains an open area for further development.

6.2 The future of plant breeding

Over the last couple of decades, the genetic gain through modern plant breeding mainly increased due to technological advances in high-throughput genotyping, phenotyping and molecular profiling. Progress is expected to continue towards increasingly accurate and cheaper screening of larger populations, measuring (newer) endophenotypes at greater resolutions, recording large numbers of traits and environmental covariates at greater spatiotemporal scales, and large-scale functional studies using high-throughput experimental and computational approaches. This will expand the existing prior knowledge-base. The ability to generate such massive datasets generates opportunities for increased performance of GP models, but also calls for precise tuning of overall breeding programs. This implies that instead of improving only the genetic gain over time, multiple components of breeding programs, including germplasm, phenotyping, data generation and human resources are expected to improve jointly. Increasingly cheaper SNP array-based genotyping allows for larger reference populations, providing more training data to improve genomic prediction performance. However, germplasm resources are not expected to match the number of DNA polymorphisms we can measure. The resulting limited training data situation will thus remain, making high dimensionality (i.e. the $p >> n$ problem) a persistent characteristic of GP.

Despite limited training data, the focus is likely to shift towards data-driven complex machine learning/deep learning models to capitalise on the inherent nonlinear patterns among heterogeneous datasets, for which conventional methods are usually suboptimal. Their intrinsic large training data requirement is likely to be addressed by newer approaches, including the knowledge-driven approaches discussed in this thesis. For instance, transfer learning, i.e. reusing previously trained models, can be useful to capitalise on the earlier learning on a different but similar population. This promises increased robustness, generalisability and accuracy of the resulting model (Weiss, Khoshgoftaar et al. 2016). Active learning is another approach to increase training data by predicting labels for widely available unlabelled data prior to predictive modelling (Yao, Zhu et al. 2016). Reinforcement learning can be used for overall efficient resource allocation in a GS-based breeding program (Moeinizade, Hu et al. 2022). Modulation of the compound effects of different factors in a breeding program can compensate for the limited training data. These newer techniques provide an opportunity for knowledge-driven approaches, such as those discussed in this thesis. Moreover, I foresee further developments in a number of directions which I will discuss below.

6.2.1 New approaches to knowledge retrieval

Retrieval of prior biological knowledge often requires mining literature or databases by experts, because trait-specific knowledge may not be readily accessible. I used a list of

gene ontology (GO) terms and clusters of co-expressed genes (COEX) for the LMM-based analyses of the photosynthetic light use efficiency and projected leaf area traits (Chapter 3) and found some interesting ones only after testing each of the ~7,297 GO terms and ~12,000 COEX clusters, which is computationally expensive. In contrast, PRIORNET (Chapter 4) used a pre-specified list of trait-specific genes / GO terms for the sodium accumulation in leaves, flowering time and seed germination ability in the dark traits. To ease pre-selection of knowledge, an automated approach could be exploited, for which language models could be a reasonable choice. For instance, large language models (LLMs) such as ChatGPT (Open AI, San Francisco, CA, USA) (OpenAI 2022) could be asked to compile a potential list of genes / GO terms / pathways for both of these analyses. LLMs can easily be integrated into a GP model through an API; nevertheless, some human intervention will still be required to parse and validate outcomes because they might be inaccurate, untruthful, or otherwise misleading at times. As an example, we could obtain 10 out of 16 genes, used as knowledge for PRIORNET, for the sodium accumulation, 20 out of 491 for flowering time and 10 out of 182 for seed germination traits along with some other genes using ChatGPT v3.5 (after calling it three times and combining the slightly different lists of genes it produced each time). Generic LLMs are likely to keep improving in the near future, and custom implementations could be developed for large scale data querying.

6.2.2 Genotype data: from SNP arrays to whole genome sequencing

GP is conducted using genome-wide SNPs, based on the assumption that at least one SNP will be found in LD with each QTL. Most commonly low-density (LD) or high-density (HD) SNP arrays are utilised, for which SNPs have been pre-selected by researchers for their possible associations with the traits of interests, based on different association studies. HD arrays are chosen to increase the likelihood of capturing larger numbers of QTLs, for a reasonable cost of genotyping. Nevertheless, the maximum number of SNPs that can be queried using an array is usually far lower than the actual number of SNPs in a population. This may result in missing useful genetic variation, and hence QTLs, and create an intrinsic bias due to their arbitrary selection. An obvious choice is to increase the number of SNPs through imputation or whole genome sequencing (WGS).

WGS data can yield a much higher number of SNPs than an HD SNP array, including the true QTNs (i.e. not relying on LD between measured and true QTNs). Since LD can vary across populations, compared to SNP-based GP, WGS-based models are potentially more accurate over a broader range of populations, with low relatedness between individuals (Ros-Freixedes, Johnsson et al. 2022). Recently, due to the rapidly decreasing cost of DNA sequencing, WGS-based GP has become feasible, but for whole populations, containing hundreds of genotypes, it is still not very cost effective. Therefore, a mixed strategy is often used: some of the germplasm is sequenced at greater depth

than others, while other accessions are genotyped using SNP arrays. Whole-genome genotypes for the entire population can then be imputed.

Since WGS data potentially contains all genetic variation, a considerable increase in GP prediction accuracy is expected. Earlier, a ~30% increase in prediction accuracy persisting over many generations was observed using simulated data (Druet, Macleod et al. 2014); however, in real populations only marginal increases could be realized (MacLeod, Hayes et al. 2014). This is usually because of genotype imputation errors, limited effective population sizes or noise in the real WGS data. In Chapter 2, I observed that with increasing numbers of non-causal SNPs (noise) in the model, prediction accuracies start decreasing for all models, unless strong effect QTNs are present and can be found by the variable selection methods (Bayes, Random Forest, XGBoost). The number of non-causal SNPs will be far higher in WGS than in array data, leading to significantly more parameters to estimate and thus requiring more training data. This implies that to make best out of WGS data, noise needs to be repressed. One possible strategy could be to prune SNPs prior to incorporation into the model, based on their association scores in a preliminary univariate analysis (Brondum, Su et al. 2015, VanRaden, Tooker et al. 2017, Lopez, An et al. 2021). However, the predictions are likely to get overfitted on the training population, because population structure is implicitly accounted for during SNP effects estimation for both preselection and prediction on the same population (Veerkamp, Bouwman et al. 2016). A more generic approach is to preselect or prioritise potentially causal SNPs based on prior biological knowledge, as I did for both LMMs (Chapter 3) and PRIORNET (Chapter 4), or to use predicted functional variant impact as an alternative (Teng, Huang et al. 2020, Benegas, Batra et al. 2022).

DL methods may also prove relevant for WGS-based GP (Alharbi and Rashid 2022). For instance, Convolutional Neural Networks (CNNs) are useful to preserve the local SNP context (i.e. LD between closely linked / physically co-located SNPs) and can reduce dimensionality by pooling (Vaz and Balaji 2021). CNNs are capable of selecting relevant SNPs based on their strength of association, reducing the chip bias inherent in manual curation of SNP array data. However, an identical convolution filter is applied to each genomic locus/region, which is likely not in line with the underlying biology. An alternative is to apply different filters to different genomic regions, proposed as local CNN (LCNN) (Pook, Freudenthal et al. 2020). Although this significantly increases the computational complexity, prior knowledge on region-specific biological annotations can be used to select the number of filters (Hou, Tao et al. 2021).

6.2.3 Improved xAI using prior biological knowledge

GP models are not just for making predictions: they can also help understand the trait associated SNPs / loci by accounting for long-distance SNP interactions (Wolc and

Dekkers 2022). A first choice is to use linear models, as these provide straight-forward interpretability of these associations for humans. Nonlinear ML / DL models are potentially competitive in terms of prediction performance, with a potential to become superior due to increasingly cheaper data generation (Montesinos-López, Martín-Vallejo et al. 2019), but most are considered to be black boxes. Recently, new methods and techniques have been developed to open these up and increase their interpretability, referred to as explainable Artificial Intelligence (xAI) (Azodi, Tang et al. 2020, Molnar 2020). Such techniques either provide a local interpretation of the input features, i.e. for individual samples, or a global interpretability of outcomes based on all input samples (Novakovsky, Dexter et al. 2022). The methods usually take a pretrained model and quantify importance attributions for different model components (e.g. layers, edges, nodes) or only the input SNPs for a particular output. However, these methods have not been evaluated thoroughly for elucidating true SNP associations, and benchmarked on the known information; raising doubts for their practical use (Rieger, Singh et al. 2020). Moreover, given millions of input SNPs, accurate estimation of individual SNP importance is computationally challenging. PRIORNET in Chapter 4 therefore used the integrated gradients (IG) method, which is faster than the more accurate Shapley values (Jethani, Sudarshan et al. 2021, Holzinger, Saranti et al. 2022). The IG method was able to retrieve most of the knowledge SNPs, assigning them high importance. This suggests that prior knowledge can provide a way to validate SNP associations. In turn, knowledge-driven penalisation can be used to deal with high-dimensional SNP data and ignore spurious correlations. Based on this, I anticipate further developments for knowledge-driven xAI methods for SNP data.

6.2.4 New data sources

Given technological advances, data generation for GP is not a bottleneck anymore; instead, the focus has shifted towards processing and integration of data into state-of-the-art methods. New data types representing different physiological and biochemical characteristics of plants are being generated and their potential for improving GP needs to be investigated. In this thesis, along with genomics, I used datasets derived from transcriptomics and phenomics, and observed significant improvements. However, other choices are readily available, for instance field phenomics, enviromics (Crossa, Fritsche-Neto et al. 2021), proteomics (Azimi, Kaufman et al. 2020), metabolomics (Tong, Küken et al. 2020) and fluxomics (Emwas, Szczepski et al. 2022) etc. Here, I briefly discuss two of these promising data types.

i. Field phenomics

The term ‘phenomics’ covers automated high-throughput phenotyping (HTP). Accurate HTP allows to capture phenotypic variation in a cost-effective manner, providing larger populations to help increase GP accuracy in the breeder’s equation (section 1.1.1).

Another advantage is the ability to screen multiple traits at a time and exploit their genetic correlations, as well as correlations between traits and environments. Following the developments in carefully controlled HTP facilities, similar advances are now being realised in field-based phenotyping, testing plants in real environments at lower capital cost (Deery and Jones 2021). The implementation of field phenomics depends on its target measurements objectives, carried out by utilising the appropriate sensing systems and the types of sensors. The sensing systems may include fixed sensor points with greater field views, field buggies, manned / unmanned airborne aircrafts operable at low and altitudes and satellites etc. Moreover, different sensors based on reflectance (e.g. multispectral / hyperspectral / RGB), depth (e.g. LiDAR) and thermal imaging can be incorporated within the sensing system, generating large amounts of phenomics data, ranging from the individual plant level to the whole field at a time. As such measurements can assess or reflect a broad range of plant physiological aspects (e.g. height, biomass, leaf count, tiller counts etc) and biochemical responses (e.g. photochemical reflectance index or fluorescence studies of photosynthesis), the efficacy for crop improvement depends on the type of dataset. In my work, I tested photosynthesis and PLA, measured in a simulated dynamic light environment (Chapter 5). However, the simulated settings can be different from the real environment. For instance, in comparison to our experimental settings, cloud shading frequency, sensor / robot movement noise, wind speed, temperature and plant-background separability can differ between the growth-chamber and field conditions. Additionally, different sensing systems and sensor types can be employed for in-field photosynthetic measurements, resulting into higher trait variability (Herritt, Long et al. 2021). It would be interesting to test our outcomes on data measured in the field conditions. Moreover, I used only traits derived from reflectance at narrow wavelength bands; it would be good to explore a broader range of reflectance spectra, obtained from multispectral / hyperspectral cameras. For example, a recent study demonstrated selection based only on hyperspectral imaging to be competitive to genomic selection (Rincent, Charpentier et al. 2018, Lane, Murray et al. 2020, Zhu, Leiser et al. 2021, Robert, Auzanneau et al. 2022).

Major challenges in integrating phenomics into GP are the increase in dimensionality of the phenotypic space with the same number of training samples, large number of correlated phenotypes and missing data (Crossa, Fritsche-Neto et al. 2021). Montesinos-López, Montesinos-López et al. (2017) proposed using a Bayesian functional regression model considering 250 reflectance bands at discrete wavelengths, along with genomic, pedigree, main effects of genotypes, environments and their interactions. Models including a wavelength x environment interaction were found to predict more accurately across environments. These models were able to handle correlated wavelengths by selecting only a subset of them. Another choice is to use selection indices (SIs), e.g. a linear SI that is simply a weighted sum of the measured phenotypes. The weights are

derived by maximizing correlation between target phenotype and the SI. Lopez-Cruz, Olson et al. (2020) demonstrated higher accuracy when using regularized SIs derived from hyperspectral data than when using standard SIs and vegetation indices. More recently, Arouisse, Theeuwen et al. (2021) proposed to reduce dimensionality of secondary traits instead of using all of these wavelengths, and proposed penalised SIs useful for the case when individual plot level data is available. As discussed in Chapter 5, multi-trait models capitalise over the genetic correlation between multiple phenotypes and this common genetic effect (pleiotropic effect) is augmented with the SNPs based genetic effects for the target traits. However, an underlying assumption is uniform (co)variance between multiple phenotypes. Incorporating phenotype-specific covariance or prior knowledge about genetic architectures of the individual phenotypes would be an interesting research direction.

ii. Enviromics

The drive to account for diverse environments in plant breeding is not new. However, the high cost involved in large-scale environmental trials, a lack of (sufficiently accurate) environmental descriptors, and a lack of statistical methodology to efficiently integrate such descriptors with molecular and phenomics data have been major challenges. Therefore, most conventional GP models are able to explain only part of the phenotypic variation governed by the genetic component, with the remaining variation, including variation due to environment and its interaction with genotypes (GxE), often goes unexplained. Commonly, GxE is accounted for by performing multi-environmental trials, using a representative set of environments. However, recent advances in remote sensing, soil-born sensors, global information systems, historical weather and temperature records can provide cost-effective environment data at much larger scales than before, enabling measuring of large numbers of environment descriptors (envirotyping) and untangling a higher proportion of GxE variance and phenotypic plasticity. Climate change and the corresponding biotic and abiotic stresses are the major driving forces to use these technologies to evaluate large scale environmental data, providing yet another type of -omics data called ‘enviromics’ (Costa-Neto and Fritsche-Neto 2021, Resende, Piepho et al. 2021). Key research questions are, predicting a suitable environment for a given germplasm, predicting the performance of a line not evaluated before in a specific environment, developing robust GP models for predicting in unforeseen environments, identification of best sites for conducting experimental trials etc.

Different statistical methods have explored the potential of multi-environmental covariates, using reaction norms (Jarquin, Crossa et al. 2014, Jarquin, Lemes da Silva et al. 2017, Sukumaran, Jarquin et al. 2018), kernel methods (Cuevas, Crossa et al. 2016, Cuevas, Crossa et al. 2017, Costa-Neto, Fritsche-Neto et al. 2021) and DL methods (Cuevas, Montesinos-López et al. 2019). More recently, large-scale enviromics data has

been incorporated using crop-growth models (Rincent, Malosetti et al. 2019, Robert, Le Gouis et al. 2020) and historical weather data (de Los Campos, Pérez-Rodríguez et al. 2020). Integrating different environmental covariates into networks, referred to as 'enviromics assembly', has been proposed to answer the key enviromics-related questions (stated above) under more growing conditions (Costa-Neto, Crossa et al. 2021, Costa-Neto, Galli et al. 2021).

Despite its importance, throughout this thesis, I have ignored the effect of environment. This was partly because the phenotype data available was generated from a sophisticated 'Phenovator II' HTP system, designed and tested with high experimental repeatability. Also, my focus was to elucidate the impact of knowledge, so models were compared on their relative performance with respect to benchmarks (without knowledge incorporated), instead of their absolute performance. Nevertheless, in the near future I foresee significant attention to integrating enviromics together with genomics and phenomics (Crossa, Fritsche-Neto et al. 2021), and incorporation of prior knowledge in the form of differentially weighted environmental covariates with *a priori* known interactions and expert opinions. For example, interaction between sowing time, temperature and disease outbreak is often known *a priori*; this can be incorporated as an explicit interaction in the model.

6.2.5 Training data expansion

Thus far, I mainly focussed on two approaches to deal with the limited training data problem for GP: using additional information from secondary traits / endophenotypes and using prior biological knowledge. An alternative approach is to augment real training populations by generating synthetic data using generative models (von Werra, Schöngens et al. 2019). This implies that for a training population with SNP genotypes, genetic polymorphism patterns are learnt by the model, and new genotypes are generated based on these patterns for the new samples. Similarly, with recent developments in the synthetic biology, (re)design of eukaryotic genomes is becoming practical (Zhang, Mitchell et al. 2020, Yelmen, Decelle et al. 2021, Hazra, Kim et al. 2022). The synthetic samples can serve to increase the effective population size, maintaining the original population characteristics, e.g. allele frequencies, linkage disequilibrium and population structure etc.; or can help introduce low levels of noise over the characteristic distributions. The former helps improve a model's predictive performance, and the latter is aimed at increasing robustness to unseen data. In this regard, Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) are specialised DL based generative models (Pérez-Enciso and Zingaretti 2019, Montesinos-López, Montesinos-López et al. 2021). GANs and VAEs themselves require large training data and training time and sometimes face non-convergence and instability (Kingma and Welling 2019, Chen 2021). With recent methodological improvements in these models (Jabbar, Li et al.

2021), I anticipate considerable applications of deep generative models coupled with all other GP methods.

I propose to investigate a knowledge-driven approach using synthetic data generation, as a follow-up of my work in this thesis. The strategy is to inform new training sample generation by incorporating anticipated effects of selection on population, genomes and trait values in view of the whole GS-based breeding program (*Figure 1.1*). In GS, after successive selections, genetic relatedness between train and test populations decreases. Since GP is usually developed on the training population and the accuracy depends on the train-test relatedness, the accuracy decreases accordingly (Cericola, Jahoor et al. 2017, Berro, Lado et al. 2019, Merrick, Herr et al. 2022). A solution is to retrain the model after each selection round, by including the previous selection population. Since GS aims to speed up the breeding cycle through early selection using breeding values rather than waiting for the true phenotypes to appear, the model can only be re-trained using the predicted breeding values as phenotypes for the next selection, which could be less accurate. Moreover, the first selection might still affect subsequent selection decisions.

A better choice could be to anticipate alleles throughout successive selections from the base population using forward-in-time genome evolution (Kessner and Novembre 2014) along with incorporating the breeding goals. The simulated selected samples are then included *a priori* in the training set. For instance, given the parental genotypes and the breeding goals (e.g. projected trait values, heritabilities and selection sizes), alleles are simulated and breeding values are estimated guided by a specific recombination frequency. The resulting new allelic combinations can, therefore, increase the effective population size of the training set. In practice, such a concept can be realized by framing the whole breeding program as an optimisation problem, where objective function(s) on the rate of genetic gain or narrow-sense heritability can be optimised. For instance, Aasim, Ayhan et al. (2023) predicted the optimal culture medium for plant regeneration frequency and maximum shoot count by simulating generations of crossovers and mutations, with a cascade of artificial neural networks and genetic algorithms.

In summary, the presented knowledge-driven approach provides a way to unbox the traditionally used black-box approach for genomic prediction. Providing accurate biological context allows the model to predict better and paves the way to provide mechanistic insights. Moreover, this knowledge-driven approach can be used to overcome the inherent limitations of a predominant data-driven approach to future plant breeding. In this thesis, I demonstrated its applicability on both simulated and real datasets, developed innovative methodology and used some knowledge resources. There still is considerable room for further developments by expanding both knowledge sources and the methodological framework, including the areas discussed above. I therefore anticipate an exciting time ahead.

References

Aasim, M., A. Ayhan, R. Katırcı, A. Ş. Acar and S. A. Ali (2023). "Computing artificial neural network and genetic algorithm for the feature optimization of basal salts and cytokinin-auxin for in vitro organogenesis of royal purple (Cotinus coggygria Scop)." Industrial Crops and Products **199**: 116718.

Abdollahi-Arpanahi, R., D. Gianola and F. Peñagaricano (2020). "Deep learning versus parametric and ensemble methods for genomic prediction of complex phenotypes." Genetics Selection Evolution **52**(1): 1_15.

Acquaah, G. (2012). History of Plant Breeding. Principles of Plant Genetics and Breeding**:** 22-39.

Adak, A., S. C. Murray and S. L. Anderson (2021). "Temporal phenomic predictions from unoccupied aerial systems can outperform genomic predictions." bioRxiv: 2021.2010.2006.463310.

Akiba, T., S. Sano, T. Yanase, T. Ohta and M. Koyama (2019). Optuna: A next-generation hyperparameter optimization framework. Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining.

Alharbi, W. S. and M. Rashid (2022). "A review of deep learning applications in human genomics using next-generation sequencing data." Human Genomics **16**(1): 1-20.

Alves, A. A. C., R. M. da Costa, T. Bresolin, G. A. Fernandes Júnior, R. Espigolan, A. M. F. Ribeiro, R. Carvalheiro and L. G. de Albuquerque (2020). "Genome-wide prediction for complex traits under the presence of dominance effects in simulated populations using GBLUP and machine learning methods." Journal of Animal Science **98**(6).

Andel, M. and F. Masri (2015). "Sparse Omics-network Regularization to Increase Interpretability and Performance of SVM-based Predictive Models." IEEE Explore.

Arouisse, B., A. Korte, F. van Eeuwijk and W. Kruijer (2020). "Imputation of 3 million SNPs in the Arabidopsis regional mapping population." The Plant Journal **102**(4): 872-882.

Arouisse, B., T. P. J. M. Theeuwen, F. A. van Eeuwijk and W. Kruijer (2021). "Improving Genomic Prediction Using High-Dimensional Secondary Phenotypes." Frontiers in Genetics **12**.

Arthur, E. H. and W. K. Robert (1970). "Ridge regression: biased estimation for nonorthogonal problems." Technometrics **12**(1): 55.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight and J. T. Eppig (2000). "Gene ontology: tool for the unification of biology." Nature Genetics **25**(1): 25-29.

Atanda, S. A., J. Steffes, Y. Ian, M. A. Al Bari, J.-H. Kim, M. Morales, J. P. Johnson, R. Saludares, H. Worral, L. Piche, A. Ross, M. Grusak, C. Coyne, R. McGee, J. Rao and N. Bandillo (2022). "Multi-trait genomic prediction improves selection accuracy for enhancing seed mineral concentrations in pea." The Plant Genome **n/a**(n/a): e20260.

Atwell, S., Y. S. Huang, B. J. Vilhjálmsson, G. Willems, M. Horton, Y. Li, D. Meng, A. Platt, A. M. Tarone, T. T. Hu, R. Jiang, N. W. Muliyati, X. Zhang, M. A. Amer, I. Baxter, B. Brachi, J. Chory, C. Dean, M. Debieu, J. de Meaux, J. R. Ecker, N. Faure, J. M. Kniskern, J. D. G. Jones, T. Michael, A. Nemri, F. Roux, D. E. Salt, C. Tang, M. Todesco, M. B. Traw, D. Weigel, P. Marjoram, J. O. Borevitz, J. Bergelson and M. Nordborg (2010). "Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines." Nature **465**(7298): 627-631.

Azimi, A., K. L. Kaufman, J. Kim, M. Ali, G. J. Mann and P. Fernandez-Penas (2020). "Proteomics: An emerging approach for the diagnosis and classification of cutaneous squamous cell carcinoma and its precursors." Journal of Dermatological Science **99**(1): 9-16.

Azodi, C. B., E. Bolger, A. McCarren, M. Roantree, G. de Los Campos and S.-H. Shiu (2019). "Benchmarking parametric and Machine Learning models for genomic prediction of complex traits." G3: Genes, Genomes, Genetics **9**(11): 3691_3702.

Azodi, C. B., J. Pardo, R. VanBuren, G. de los Campos and S.-H. Shiu (2020). "Transcriptome-Based Prediction of Complex Traits in Maize." The Plant Cell **32**(1): 139-151.

Azodi, C. B., J. Tang and S.-H. Shiu (2020). "Opening the Black Box: Interpretable Machine Learning for Geneticists." Trends in Genetics **36**(6): 442-455.

Baker, N. R. (2008). "Chlorophyll fluorescence: a probe of photosynthesis in vivo." Annu. Rev. Plant Biol. **59**: 89-113.

Barbosa, I. d. P., M. J. da Silva, W. G. da Costa, I. de Castro Sant'Anna, M. Nascimento and C. D. Cruz (2021). "Genome-enabled prediction through machine learning methods considering different levels of trait complexity." Crop Science **61**(3): 1890_1902.

Bareke, T. (2018). "Biology of seed development and germination physiology." Advances in Plants & Agriculture Research **8**(4): 336-346.

Bates, D., D. Sarkar, M. D. Bates and L. Matrix (2007). "The lme4 package." R package version **2**(1): 74.

Baxter, I., J. N. Brazelton, D. Yu, Y. S. Huang, B. Lahner, E. Yakubova, Y. Li, J. Bergelson, J. O. Borevitz and M. Nordborg (2010). "A coastal cline in sodium accumulation in Arabidopsis thaliana is driven by natural variation of the sodium transporter AtHKT1; 1." PLoS Genetics **6**(11): e1001193.

Bellot, P., G. de Los Campos and M. Pérez-Enciso (2018). "Can deep learning improve genomic prediction of complex human traits?" Genetics **210**(3): 809_819.

Benegas, G., S. S. Batra and Y. S. Song (2022). "DNA language models are powerful zero-shot predictors of non-coding variant effects." bioRxiv: 2022.2008. 2022.504706.

Bepler, T. and B. Berger (2021). "Learning the protein language: Evolution, structure, and function." Cell Syst **12**(6): 654-669.e653.

Bermingham, M. L., R. Pong-Wong, A. Spiliopoulou, C. Hayward, I. Rudan, H. Campbell, A. F. Wright, J. F. Wilson, F. Agakov, P. Navarro and C. S. Haley (2015). "Application of high-dimensional feature selection: evaluation for genomic prediction in man." Sci Rep **5**: 10312.

Berro, I., B. Lado, R. S. Nalin, M. Quincke and L. Gutiérrez (2019). "Training Population Optimization for Genomic Selection." The Plant Genome **12**(3): 190028.

Bhatta, M., L. Gutierrez, L. Cammarota, F. Cardozo, S. Germán, B. Gómez-Guerrero, M. F. Pardo, V. Lanaro, M. Sayas and A. J. Castro (2020). "Multi-trait Genomic Prediction Model Increased the Predictive Ability for Agronomic and Malting Quality Traits in Barley (Hordeum vulgare L.)." G3 Genes|Genomes|Genetics **10**(3): 1113-1124.

Bouché, F., G. Lobet, P. Tocquin and C. Périlleux (2015). "Flowering Interactive Database [FLOR-ID]."

Bourgeais, V., F. Zehraoui, M. Ben Hamdoune and B. Hanczar (2021). "Deep GONet: self-explainable deep neural network based on Gene Ontology for phenotype prediction from gene expression data." BMC Bioinformatics **22**(10): 455.

Brachi, B., G. P. Morris and J. O. Borevitz (2011). "Genome-wide association studies in plants: the missing heritability is in the field." Genome Biol **12**(10): 232.

Brault, C., J. Lazerges, A. Doligez, M. Thomas, M. Ecarnot, P. Roumet, Y. Bertrand, G. Berger, T. Pons and P. François (2021). "Interest of phenomic prediction as an alternative to genomic prediction in grapevine." bioRxiv.

Breiman, L. (2001). "Random Forests." Machine Learning **45**(1): 5-32.

Breseghello, F. and A. S. G. Coelho (2013). "Traditional and Modern Plant Breeding Methods with Examples in Rice (Oryza sativa L.)." Journal of Agricultural and Food Chemistry **61**(35): 8277-8286.

Brondum, R. F., G. Su, L. Janss, G. Sahana, B. Guldbrandtsen, D. Boichard and M. S. Lund (2015). "Quantitative trait loci markers derived from whole genome sequence data increases the reliability of genomic prediction." J Dairy Sci **98**(6): 4107-4116.

Bryant, P., G. Pozzati and A. Elofsson (2022). "Improved prediction of protein-protein interactions using AlphaFold2." Nature Communications **13**(1): 1265.

Buza, T. J., F. M. McCarthy, N. Wang, S. M. Bridges and S. C. Burgess (2008). "Gene Ontology annotation quality analysis in model eukaryotes." Nucleic Acids Research **36**(2): e12-e12.

Calus, M. P., A. C. Bouwman, C. Schrooten and R. F. Veerkamp (2016). "Efficient genomic prediction based on whole-genome sequence data using split-and-merge Bayesian variable selection." Genet Sel Evol **48**(1): 49.

Carlson, M. (2019). "GO.db: A set of annotation maps describing the entire Gene Ontology. R package version 3.10.0.".

Carlson, M. (2019). "org.At.tair.db: Genome wide annotation for Arabidopsis, R package version 3.10.0."

Cericola, F., A. Jahoor, J. Orabi, J. R. Andersen, L. L. Janss and J. Jensen (2017). "Optimizing Training Population Size and Genotyping Strategy for Genomic Prediction Using Association Study Results and Pedigree Information. A Case of Study in Advanced Wheat Breeding Lines." PLoS One **12**(1): e0169606.

Chen, H. (2021). Challenges and corresponding solutions of generative adversarial networks (GANs): a survey study. Journal of Physics: Conference Series, IOP Publishing.

Chin, C.-H., S.-H. Chen, C.-W. Ho, M.-T. Ko and C.-Y. Lin (2010). "A hub-attachment based method to detect functional modules from confidence-scored protein interactions and expression profiles." BMC bioinformatics **11**(1): 1-9.

Costa-Neto, G., J. Crossa and R. Fritsche-Neto (2021). "Enviromic assembly increases accuracy and reduces costs of the genomic prediction for yield plasticity." bioRxiv: 2021.2006.2004.447091.

Costa-Neto, G. and R. Fritsche-Neto (2021). "Enviromics: bridging different sources of data, building one framework." Crop Breeding and Applied Biotechnology **21**.

Costa-Neto, G., R. Fritsche-Neto and J. Crossa (2021). "Nonlinear kernels, dominance, and envirotyping data increase the accuracy of genome-based prediction in multi-environment trials." Heredity **126**(1): 92-106.

Costa-Neto, G., G. Galli, H. F. Carvalho, J. Crossa and R. Fritsche-Neto (2021). "EnvRtype: a software to interplay enviromics and quantitative genomics in agriculture." G3 **11**(4): jkab040.

Crain, J., S. Mondal, J. Rutkoski, R. P. Singh and J. Poland (2018). "Combining high-throughput phenotyping and genomic information to increase prediction and selection accuracy in wheat breeding." The plant genome **11**(1).

Crossa, J., G. d. l. Campos, P. Pérez, D. Gianola, J. Burgueno, J. L. Araus, D. Makumbi, R. P. Singh, S. Dreisigacker and J. Yan (2010). "Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers." Genetics **186**(2): 713-724.

Crossa, J., R. Fritsche-Neto, O. A. Montesinos-Lopez, G. Costa-Neto, S. Dreisigacker, A. Montesinos-Lopez and A. R. Bentley (2021). "The Modern Plant Breeding Triangle: Optimizing the Use of Genomics, Phenomics, and Enviromics Data." Frontiers in Plant Science **12**.

Cuevas, J., J. Crossa, O. A. Montesinos-Lopez, J. Burgueno, P. Perez-Rodriguez and G. de Los Campos (2017). "Bayesian Genomic Prediction with Genotype x Environment Interaction Kernel Models." G3 (Bethesda) **7**(1): 41-53.

Cuevas, J., J. Crossa, V. Soberanis, S. Pérez-Elizalde, P. Pérez-Rodríguez, G. d. l. Campos, O. Montesinos-López and J. Burgueño (2016). "Genomic prediction of genotype× environment interaction kernel regression models." The plant genome **9**(3): plantgenome2016.2003.0024.

Cuevas, J., O. Montesinos-López, P. Juliana, C. Guzmán, P. Pérez-Rodríguez, J. González-Bucio, J. Burgueño, A. Montesinos-López and J. Crossa (2019). "Deep Kernel for Genomic and Near Infrared Predictions in Multi-environment Breeding Trials." G3 Genes|Genomes|Genetics **9**(9): 2913-2924.

Daetwyler, H. D., R. Pong-Wong, B. Villanueva and J. A. Woolliams (2010). "The Impact of Genetic Architecture on Genome-Wide Evaluation Methods." Genetics **185**(3): 1021-1031.

de los Campos G, Grueneberg A. (2013). "MTM: An R-Package for Genetic Multi-Trait Recursive and Factor Analyses Models (version 1.0)."

De los Campos, G., D. Gianola, G. J. Rosa, K. A. Weigel and J. Crossa (2010). "Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods." Genetics Research **92**(4): 295-308.

De los Campos, G., D. Gianola, G. J. M. Rosa, K. A. Weigel and J. Crossa (2010). "Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods." Genet. Res. **92**(4): 295_308.

de Los Campos, G., P. Pérez-Rodríguez, M. Bogard, D. Gouache and J. Crossa (2020). "A data-driven simulation platform to predict cultivars' performances under uncertain weather conditions." Nature communications **11**(1): 4876.

de Los Campos, G., D. A. Sorensen and M. A. Toro (2019). "Imperfect linkage disequilibrium generates phantom epistasis (& perils of big data)." G3: Genes, Genomes, Genetics **9**(5): 1429_1436.

Deery, D. M. and H. G. Jones (2021). "Field phenomics: will it enable crop improvement?" Plant Phenomics **2021**.

Deng, H. and G. Runger (2013). "Gene selection with guided regularized random forest." Pattern Recognition **46**(12): 3483-3489.

Deng, M., F. Sun and T. Chen (2003). "Assessment of the reliability of protein-protein interactions and protein function prediction." Pac Symp Biocomput: 140-151.

Deniz Akdemir, J. I. S. e. (2019). "Adventures in Multi-Omics I: Combining heterogeneous data sets via relationships matrices." bioRxiv.

Diligenti, M., S. Roychowdhury and M. Gori (2017). Integrating Prior Knowledge into Deep Learning. 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA).

Dolatabadian, A., D. A. Patel, D. Edwards and J. Batley (2017). "Copy number variation and disease resistance in plants." Theoretical and Applied Genetics **130**: 2479-2490.

Druet, T., I. M. Macleod and B. J. Hayes (2014). "Toward genomic prediction from whole-genome sequence data: impact of sequencing design on genotype imputation and accuracy of predictions." Heredity (Edinb) **112**(1): 39-47.

Du, C., J. Wei, S. Wang and Z. Jia (2018). "Genomic selection using principal component regression." Heredity (Edinb) **121**(1): 12-23.

Ecarnot, M., F. Compan and P. Roumet (2013). "Assessing leaf nitrogen content and leaf mass per unit area of wheat in the field throughout plant cycle with a portable spectrometer." Field Crops Research **140**: 44-50.

Edlich-Muth, C., M. M. Muraya, T. Altmann and J. Selbig (2016). "Phenomic prediction of maize hybrids." Biosystems **146**: 102-109.

Edwards, S. M., I. F. Sorensen, P. Sarup, T. F. Mackay and P. Sorensen (2016). "Genomic Prediction for Quantitative Traits Is Improved by Mapping Variants to Gene Ontology Categories in Drosophila melanogaster." Genetics **203**(4): 1871-1883.

Edwards, S. M., B. Thomsen, P. Madsen and P. Sørensen (2015). "Partitioning of genomic variance reveals biological pathways associated with udder health and milk production traits in dairy cattle." Genetics Selection Evolution **47**(1): 60.

Ehsani, A., L. Janss, D. Pomp and P. Sørensen (2016). "Decomposing genomic variance using information from GWA, GWE and eQTL analysis." Animal Genetics **47**(2): 165-173.

Emwas, A.-H., K. Szczepski, I. Al-Younis, J. I. Lachowicz and M. Jaremko (2022). "Fluxomics-new metabolomics approaches to monitor metabolic pathways." Frontiers in Pharmacology **13**: 299.

Endelman, J. B. (2011). "Ridge regression and other kernels for genomic selection with R package rrBLUP." The plant genome **4**(3).

Eraslan, G., Ž. Avsec, J. Gagneur and F. J. Theis (2019). "Deep learning: new computational modelling techniques for genomics." Nature Reviews Genetics **20**(7): 389-403.

Falconer, D. and T. Mackay (1996). "Introduction to quantitative genetics. 1996." Harlow, Essex, UK: Longmans Green **3**.

Falconer, D. and T. Mackay (1996). "Introduction to quantitative genetics. Essex." UK: Longman Group.

Fang, L., G. Sahana, P. Ma, G. Su, Y. Yu, S. Zhang, M. S. Lund and P. Sorensen (2017). "Exploring the genetic architecture and improving genomic prediction accuracy for mastitis and milk production traits in dairy cattle by mapping variants to hepatic transcriptomic regions responsive to intra-mammary infection." Genetics Selection Evolution **49**(1): 44.

Fang, L., G. Sahana, P. Ma, G. Su, Y. Yu, S. Zhang, M. S. Lund and P. Sorensen (2017). "Use of biological priors enhances understanding of genetic architecture and genomic prediction of complex traits within and between dairy cattle breeds." BMC Genomics **18**(1): 604.

Fang, L., G. Sahana, P. Ma, G. Su, Y. Yu, S. Zhang, M. S. Lund and P. Sørensen (2017). "Exploring the genetic architecture and improving genomic prediction accuracy for mastitis and milk production traits in dairy cattle by mapping variants to hepatic transcriptomic regions responsive to intra-mammary infection." Genetics Selection Evolution **49**(1): 44.

Farooq, M., A. D. van Dijk, H. Nijveen, M. G. Aarts, W. Kruijer, T.-P. Nguyen, S. Mansoor and D. d. Ridder (2020). "Prior biological knowledge improves genomic prediction of growth-related traits in Arabidopsis thaliana." Frontiers in Genetics **11**: 1810.

Farooq, M., A. D. van Dijk, H. Nijveen, S. Mansoor and D. de Ridder (2022). "Genomic prediction in plants: opportunities for ensemble machine learning based approaches." F1000Research **11**(802): 802.

Farooq, Q. U. A., Z. Shaukat, S. Aiman and C. H. Li (2021). "Protein-protein interactions: Methods, databases, and applications in virus-host study." World J Virol **10**(6): 288-300.

Fernandes, S. B. and A. E. Lipka (2020). "simplePHENOTYPES: SIMulation of Pleiotropic, Linked and Epistatic PHENOTYPES." bioRxiv: 2020.2001.2011.902874.

Flood, P. J., W. Kruijer, S. K. Schnabel, R. van der Schoor, H. Jalink, J. F. H. Snel, J. Harbinson and M. G. M. Aarts (2016). "Phenomics for photosynthesis, growth and reflectance in Arabidopsis thaliana reveals circadian and long-term fluctuations in heritability." Plant Methods **12**(1): 14.

Flores, J. E., D. M. Claborne, Z. D. Weller, B. M. Webb-Robertson, K. M. Waters and L. M. Bramer (2023). "Missing data in multi-omics integration: Recent advances through artificial intelligence." Front Artif Intell **6**: 1098308.

Fortelny, N. and C. Bock (2020). "Knowledge-primed neural networks enable biologically interpretable deep learning on single-cell sequencing data." Genome Biology **21**(1): 190.

Fragomeni, B. O., D. A. L. Lourenco, Y. Masuda, A. Legarra and I. Misztal (2017). "Incorporation of causative quantitative trait nucleotides in single-step GBLUP." Genetics Selection Evolution **49**(1): 59.

Frisch, M., A. Thiemann, J. Fu, T. A. Schrag, S. Scholten and A. E. Melchinger (2010). "Transcriptome-based distance measures for grouping of germplasm and prediction of hybrid performance in maize." Theor Appl Genet **120**(2): 441-450.

Gao, N., J. Li, J. He, G. Xiao, Y. Luo, H. Zhang, Z. Chen and Z. Zhang (2015). "Improving accuracy of genomic prediction by genetic architecture based priors in a Bayesian model." BMC Genet **16**: 120.

Gao, N., J. Teng, S. Ye, X. Yuan, S. Huang, H. Zhang, X. Zhang, J. Li and Z. Zhang (2018). "Genomic Prediction of Complex Phenotypes Using Genic Similarity Based Relatedness Matrix." Frontiers in Genetics **9**: 364.

Gao, W., Y. Liu, J. Huang, Y. Chen, C. Chen, L. Lu, H. Zhao, S. Men and X. Zhang (2021). "MES7 Modulates Seed Germination via Regulating Salicylic Acid Content in Arabidopsis." Plants **10**(5): 903.

Gaudet, P. and C. Dessimoz (2017). Gene Ontology: Pitfalls, Biases, and Remedies. The Gene Ontology Handbook. C. Dessimoz and N. Škunca. New York, NY, Springer New York**:** 189-205.

Gazestani, V. H. and N. E. Lewis (2019). "From genotype to phenotype: augmenting deep learning with networks and systems biology." Current Opinion in Systems Biology **15**: 68-73.

Gebreyesus, G., H. Bovenhuis, M. S. Lund, N. A. Poulsen, D. Sun and B. Buitenhuis (2019). "Reliability of genomic prediction for milk fatty acid composition by using a multi-population reference and incorporating GWAS results." Genetics Selection Evolution **51**(1): 16.

Gemmer, M. R., C. Richter, Y. Jiang, T. Schmutzer, M. L. Raorane, B. Junker, K. Pillen and A. Maurer (2020). "Can metabolic prediction be an alternative to genomic prediction in barley?" PLoS One **15**(6): e0234052.

Gene Ontology, C. (2021). "The Gene Ontology resource: enriching a GOld mine." Nucleic Acids Res **49**(D1): D325-D334.

Ghafouri-Kesbi, F., G. Rahimi-Mianji, M. Honarvar and A. Nejati-Javaremi (2017). "Predictive ability of Random Forests, Boosting, Support Vector Machines and Genomic Best Linear Unbiased Prediction in different scenarios of genomic evaluation." Animal Production Science **57**(2): 229_236.

Gianola, D. (2013). "Priors in whole-genome regression: the bayesian alphabet returns." Genetics **194**(3): 573-596.

Gill, M., R. Anderson, H. Hu, M. Bennamoun, J. Petereit, B. Valliyodan, H. T. Nguyen, J. Batley, P. E. Bayer and D. Edwards (2022). "Machine learning models outperform deep learning models, provide interpretation and facilitate feature selection for soybean trait prediction." BMC plant biology **22**(1): 1-8.

Glémin, S., Y. Clément, J. David and A. Ressayre (2014). "GC content evolution in coding regions of angiosperm genomes: a unifying hypothesis." Trends in Genetics **30**(7): 263-270.

Goddard, M., K. Kemper, I. MacLeod, A. Chamberlain and B. Hayes (2016). "Genetics of complex traits: prediction of phenotype, identification of causal polymorphisms and genetic architecture." Proceedings of the Royal Society B: Biological Sciences **283**(1835): 20160569.

Godfray, H. C., J. R. Beddington, I. R. Crute, L. Haddad, D. Lawrence, J. F. Muir, J. Pretty, S. Robinson, S. M. Thomas and C. Toulmin (2010). "Food security: the challenge of feeding 9 billion people." Science **327**(5967): 812-818.

González-Recio, O., G. J. Rosa and D. Gianola (2014). "Machine learning methods and predictive ability metrics for genome-wide prediction of complex traits." Livestock Science **166**: 217-231.

Grinberg, N. F., O. I. Orhobor and R. D. King (2020). "An evaluation of machine-learning for predicting phenotype: studies in yeast, rice, and wheat." Mach. Learn. **109**(2): 251_277.

Guo, Z., M. M. Magwire, C. J. Basten, Z. Xu and D. Wang (2016). "Evaluation of the utility of gene expression and metabolic information for genomic prediction in maize." Theoretical and Applied Genetics **129**(12): 2413-2427.

Guo, Z., D. M. Tucker, C. J. Basten, H. Gandhi, E. Ersoz, B. Guo, Z. Xu, D. Wang and G. Gay (2014). "The impact of population structure on genomic prediction in stratified populations." Theor Appl Genet **127**(3): 749_762.

Habier, D., R. L. Fernando and J. C. Dekkers (2007). "The impact of genetic relationship information on genome-assisted breeding values." Genetics **177**(4): 2389-2397.

Habier, D., R. L. Fernando, K. Kizilkaya and D. J. Garrick (2011). "Extension of the bayesian alphabet for genomic selection." BMC Bioinformatics **12**: 186.

Habyarimana, E., P. De Franceschi, S. Ercisli, F. S. Baloch and M. Dall'Agata (2020). "Genome-Wide Association Study for Biomass Related Traits in a Panel of Sorghum bicolor and S. bicolor × S. halepense Populations." Frontiers in Plant Science **11**.

Haile, J. K., A. N'Diaye, E. Sari, S. Walkowiak, J. E. Rutkoski, H. R. Kutcher and C. J. Pozniak (2020). "Potential of genomic selection and integrating "omics" data for disease evaluation in wheat." Crop Breeding, Genetics and Genomics **4**(3).

Hassani-Pak, K. (2017). KnetMiner-An integrated data platform for gene mining and biological knowledge discovery, Doctoral dissertation, Bielefeld University, Bielefeld.

Hayes, B. J., P. J. Bowman, A. C. Chamberlain, K. Verbyla and M. E. Goddard (2009). "Accuracy of genomic breeding values in multi-breed dairy cattle populations." Genetics Selection Evolution **41**(1): 1-9.

Hayes, B. J., P. J. Bowman, A. J. Chamberlain and M. E. Goddard (2009). "Invited review: Genomic selection in dairy cattle: progress and challenges." Journal of Dairy Science **92**(2): 433-443.

Hazra, D., M.-R. Kim and Y.-C. Byun (2022). "Generative Adversarial Networks for Creating Synthetic Nucleic Acid Sequences of Cat Genome." International Journal of Molecular Sciences **23**(7): 3701.

Henderson, C. (1984). "Applications of linear models in animal breeding" Applications of linear models in animal breeding. University of Guelph, Guelph, ON, Canada.

Henderson, C. R. (1975). "Best linear unbiased estimation and prediction under a selection model." Biometrics: 423-447.

Henderson, C. R. (1985). "Best Linear Unbiased Prediction Using Relationship Matrices Derived from Selected Base Populations." Journal of Dairy Science **68**(2): 443-448.

Herritt, M. T., J. C. Long, M. D. Roybal, D. C. Moller, T. C. Mockler, D. Pauli and A. L. Thompson (2021). "FLIP: FLuorescence Imaging Pipeline for field-based chlorophyll fluorescence images." SoftwareX **14**: 100685.

Hoffman, G. E. (2013). "Correcting for Population Structure and Kinship Using the Linear Mixed Model: Theory and Extensions." PLOS ONE **8**(10): e75707.

Holzinger, A., A. Saranti, C. Molnar, P. Biecek and W. Samek (2022). Explainable AI Methods - A Brief Overview. xxAI - Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers. A. Holzinger, R. Goebel, R. Fong et al. Cham, Springer International Publishing**:** 13-38.

Horton, M. W., A. M. Hancock, Y. S. Huang, C. Toomajian, S. Atwell, A. Auton, N. W. Muliyati, A. Platt, F. G. Sperone, B. J. Vilhjalmsson, M. Nordborg, J. O. Borevitz and J. Bergelson (2012). "Genome-wide patterns of genetic variation in worldwide Arabidopsis thaliana accessions from the RegMap panel." Nature Genetics **44**(2): 212-216.

Hou, W., X. Tao and D. Xu (2021). "Combining Prior Knowledge With CNN for Weak Scratch Inspection of Optical Components." IEEE Transactions on Instrumentation and Measurement **70**: 1-11.

Howard, R., A. L. Carriquiry and W. D. Beavis (2014). "Parametric and nonparametric statistical methods for genomic selection of traits with additive and epistatic genetic architectures." G3 (Bethesda) **4**(6): 1027_1046.

Hu, H., M. T. Campbell, T. H. Yeats, X. Zheng, D. E. Runcie, G. Covarrubias-Pazaran, C. Broeckling, L. Yao, M. Caffe-Treml, L. Gutiérrez, K. P. Smith, J. Tanaka, O. A. Hoekenga, M. E. Sorrells, M. A. Gore and J.-L. Jannink (2021). "Multi-omics prediction of oat agronomic and seed nutritional traits across environments and in distantly related populations." bioRxiv: 2021.2005.2003.442386.

Hu, W., Z. Lu, F. Meng, X. Li, R. Cong, T. Ren, T. D. Sharkey and J. Lu (2020). "The reduction in leaf area precedes that in photosynthesis under potassium deficiency: the importance of leaf anatomy." New Phytologist **227**(6): 1749-1763.

Hu, X., W. Xie, C. Wu and S. Xu (2019). "A directed learning strategy integrating multiple omic data improves genomic prediction." Plant Biotechnology Journal **17**(10): 2011-2020.

Huang, S., K. Chaudhary and L. X. Garmire (2017). "More Is Better: Recent Progress in Multi-Omics Data Integration Methods." Frontiers in Genetics **8**(84).

Ieremie, I., R. M. Ewing and M. Niranjan (2022). "TransformerGO: predicting protein–protein interactions by modelling the attention between sets of gene ontology terms." Bioinformatics **38**(8): 2269-2277.

Ishimori, M., T. Hattori, K. Yamazaki, H. Takanashi, M. Fujimoto, H. Kajiya-Kanegae, J. Yoneda, T. Tokunaga, T. Fujiwara, N. Tsutsumi and H. Iwata (2020). "Impacts of dominance effects on genomic prediction of sorghum hybrid performance." Breeding Science **70**(5): 605-616.

Jabbar, A., X. Li and B. Omar (2021). "A survey on generative adversarial networks: Variants, applications, and training." ACM Computing Surveys (CSUR) **54**(8): 1-49.

Jacquin, L., T. V. Cao and N. Ahmadi (2016). "A Unified and Comprehensible View of Parametric and Kernel Methods for Genomic Prediction with Application to Rice." Front Genet **7**: 145.

Jantzen, S. G., B. J. G. Sutherland, D. R. Minkley and B. F. Koop (2011). "GO Trimming: Systematically reducing redundancy in large Gene Ontology datasets." BMC Research Notes **4**(1): 267.

Jarquin, D., J. Crossa, X. Lacaze, P. Du Cheyron, J. Daucourt, J. Lorgeou, F. Piraux, L. Guerreiro, P. Perez, M. Calus, J. Burgueno and G. de los Campos (2014). "A reaction norm model for genomic selection using high-dimensional genomic and environmental data." Theor Appl Genet **127**(3): 595-607.

Jarquin, D., C. Lemes da Silva, R. C. Gaynor, J. Poland, A. Fritz, R. Howard, S. Battenfield and J. Crossa (2017). "Increasing Genomic-Enabled Prediction Accuracy by Modeling Genotype x Environment Interactions in Kansas Wheat." Plant Genome **10**(2).

Jethani, N., M. Sudarshan, I. Covert, S.-I. Lee and R. Ranganath (2021) "FastSHAP: Real-Time Shapley Value Estimation." arXiv:2107.07436 DOI: 10.48550/arXiv.2107.07436.

Jiang, Y. and J. C. Reif (2015). "Modeling epistasis in genomic selection." Genetics **201**(2): 759-768.

Jiang, Y. and J. C. Reif (2015). "Modeling Epistasis in Genomic Selection." Genetics **201**(2): 759-768.

Johnson, D. and R. Thompson (1995). "Restricted maximum likelihood estimation of variance components for univariate animal models using sparse matrix techniques and average information." Journal of Dairy Science **78**(2): 449-456.

Kanehisa, M. (2001). "Prediction of higher order functional networks from genomic data." Pharmacogenomics **2**(4): 373-385.

Kanehisa, M., M. Furumichi, Y. Sato, M. Kawashima and M. Ishiguro-Watanabe (2022). "KEGG for taxonomy-based analysis of pathways and genomes." Nucleic Acids Res.

Kanehisa, M. and S. Goto (2000). "KEGG: kyoto encyclopedia of genes and genomes." Nucleic acids Research **28**(1): 27-30.

Kang, T., W. Ding, L. Zhang, D. Ziemek and K. Zarringhalam (2017). "A biological network-based regularized artificial neural network model for robust phenotype prediction from gene expression data." BMC Bioinformatics **18**(1): 565.

Karaman, E., H. Cheng, M. Z. Firat, D. J. Garrick and R. L. Fernando (2016). "An upper bound for accuracy of prediction using GBLUP." PLoS ONE **11**(8).

Karaman, E., M. S. Lund, M. T. Anche, L. Janss and G. Su (2018). "Genomic Prediction Using Multi-trait Weighted GBLUP Accounting for Heterogeneous Variances and Covariances Across the Genome." G3 (Bethesda, Md.) **8**(11): 3549-3558.

Kessner, D. and J. Novembre (2014). "forqs: forward-in-time simulation of recombination, quantitative traits and selection." Bioinformatics **30**(4): 576-577.

Khaki, S. and L. Wang (2019). "Crop Yield Prediction Using Deep Neural Networks." Frontiers in Plant Science **10**(621).

Khush, G. S. (2001). "Green revolution: the way forward." Nat Rev Genet **2**(10): 815-822.

Kingma, D. P. and M. Welling (2019). "An Introduction to Variational Autoencoders." Foundations and Trends® in Machine Learning **12**(4): 307-392.

Klughammer, C. and U. Schreiber (2008). " " PAM application notes **1**(2): 201-247.

Kokhlikyan, N., V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya and S. Yan (2020). "Captum: A unified and generic model interpretability library for pytorch." arXiv preprint arXiv:2009.07896.

Korte, A. and A. Farlow (2013). "The advantages and limitations of trait analysis with GWAS: a review." Plant Methods **9**: 29.

Kourmpetis, Y. A., A. D. van Dijk, R. C. van Ham and C. J. ter Braak (2011). "Genome-wide computational function prediction of Arabidopsis proteins by integration of multiple data sources." Plant Physiology **155**(1): 271-281.

Kramer, D. M., G. Johnson, O. Kiirats and G. E. Edwards (2004). "New Fluorescence Parameters for the Determination of QA Redox State and Excitation Energy Fluxes." Photosynth Res **79**(2): 209.

Krishnappa, G., S. Savadi, B. S. Tyagi, S. K. Singh, H. M. Mamrutha, S. Kumar, C. N. Mishra, H. Khan, K. Gangadhara, G. Uday, G. Singh and G. P. Singh (2021). "Integrated genomic selection for rapid improvement of crops." Genomics **113**(3): 1070-1086.

Kromdijk, J., K. Głowacka, L. Leonelli, S. T. Gabilly, M. Iwai, K. K. Niyogi and S. P. Long (2016). "Improving photosynthesis and crop productivity by accelerating recovery from photoprotection." Science **354**(6314): 857-861.

Kruijer, W., M. P. Boer, M. Malosetti, P. J. Flood, B. Engel, R. Kooke, J. J. Keurentjes and F. A. van Eeuwijk (2015). "Marker-based estimation of heritability in immortal populations." Genetics **199**(2): 379-398.

Kuhn, M., J. Wing, S. Weston, A. Williams, C. Keefer, A. Engelhardt, T. Cooper, Z. Mayer and B. Kenkel (2020). "caret: Classification and Regression Training. R package version 6.0-86." Avaliable at: https://cran. r-project. org/web/packages/caret/caret. pdf (accessed March 20, 2020).

Lado, B., D. Vázquez, M. Quincke, P. Silva, I. Aguilar and L. Gutiérrez (2018). "Resource allocation optimization with multi-trait genomic prediction for bread wheat (Triticum aestivum L.) baking quality." Theor Appl Genet **131**(12): 2719-2731.

Lamichhane, S. and S. Thapa (2022). "Advances from conventional to modern plant breeding methodologies." Plant breeding and biotechnology **10**(1): 1-14.

Lane, H. M., S. C. Murray, O. A. Montesinos-López, A. Montesinos-López, J. Crossa, D. K. Rooney, I. D. Barrero-Farfan, G. N. De La Fuente and C. L. S. Morgan (2020). "Phenomic selection and prediction of

maize grain yield from near-infrared reflectance spectroscopy of kernels." The Plant Phenome Journal **3**(1): e20002.

Lee, A., W. W. Cohen and K. R. Koedinger (2009). A computational model of how learner errors arise from weak prior knowledge. Proceedings of the Annual Conference of the Cognitive Science Society, Austin, TX.

Lee, I., S. V. Date, A. T. Adai and E. M. Marcotte (2004). "A Probabilistic Functional Network of Yeast Genes." Science **306**(5701): 1555-1558.

Lee, S., S. G. Kim and C. M. Park (2010). "Salicylic acid promotes seed germination under high salinity by modulating antioxidant activity in Arabidopsis." New Phytologist **188**(2): 626-637.

Lee, T., S. Yang, E. Kim, Y. Ko, S. Hwang, J. Shin, J. E. Shim, H. Shim, H. Kim, C. Kim and I. Lee (2015). "AraNet v2: an improved database of co-functional gene networks for the study of Arabidopsis thaliana and 27 other nonmodel plant species." Nucleic Acids Res **43**(Database issue): D996-1002.

Legarra, A. and V. Ducrocq (2012). "Computational strategies for national integration of phenotypic, genomic, and pedigree data in a single-step best linear unbiased prediction." Journal of Dairy Science **95**(8): 4629-4645.

Lemhadri, I., F. Ruan, L. Abraham and R. Tibshirani (2021). "LassoNet: A Neural Network with Feature Sparsity." Journal of Machine Learning Research **22**(127): 1-29.

Li, B., N. Zhang, Y. G. Wang, A. W. George, A. Reverter and Y. Li (2018). "Genomic Prediction of Breeding Values Using a Subset of SNPs Identified by Three Machine Learning Methods." Frontiers in Genetics **9**: 237.

Li, C. and H. Li (2008). "Network-constrained regularization and variable selection for analysis of genomic data." Bioinformatics **24**(9): 1175-1182.

Li, H., X.-J. Gong, H. Yu and C. Zhou (2018). "Deep neural network based predictions of protein interactions using primary sequences." Molecules **23**(8): 1923.

Li, Y., F. Raidan, M. N. Sanchez, A. George and A. Reverter (2019). Using random forest to identify snps that decrease accuracy of genomic prediction–behaviour of snps with negative vim values. Proc. Assoc. Advmt. Anim. Breed. Genet.

Lin, X., M. Liu and X.-w. Chen (2009). "Assessing reliability of protein-protein interactions by integrative analysis of data in model organisms." BMC Bioinformatics **10**(4): S5.

Liu, P.-C., W. J. Peacock, L. Wang, R. Furbank, A. Larkum and E. S. Dennis (2020). "Leaf growth in early development is key to biomass heterosis in Arabidopsis." Journal of Experimental Botany **71**(8): 2439-2450.

Long, S. P., A. Marshall-Colon and X.-G. Zhu (2015). "Meeting the global food demand of the future by engineering crop photosynthesis and yield potential." Cell **161**(1): 56-66.

Lopez-Cruz, M., E. Olson, G. Rovere, J. Crossa, S. Dreisigacker, S. Mondal, R. Singh and G. d. l. Campos (2020). "Regularized selection indices for breeding value prediction using hyper-spectral image data." Scientific Reports **10**(1): 8195.

Lopez, B. I. M., N. An, K. Srikanth, S. Lee, J.-D. Oh, D.-H. Shin, W. Park, H.-H. Chai, J.-E. Park and D. Lim (2021). "Genomic prediction based on SNP functional annotation using imputed whole-genome sequence data in Korean Hanwoo cattle." Frontiers in Genetics **11**: 603822.

Lozano, R., D. P. del Carpio, T. Amuge, I. S. Kayondo, A. O. Adebo, M. Ferguson and J.-L. Jannink (2017). "Leveraging Transcriptomics Data for Genomic Prediction Models in Cassava." bioRxiv: 208181.

Lush, J. L. (1937). Animal breeding plans. Ames, Iowa, Collegiate Press, Inc.

Ma, J. Z., M. K. Yu, S. Fong, K. Ono, E. Sage, B. Demchak, R. Sharan and T. Ideker (2018). "Using deep learning to model the hierarchical structure and function of a cell." Nature Methods **15**(4): 290-+.

Ma, X. L., H. S. Zhao, W. Y. Xu, Q. You, H. Y. Yan, Z. M. Gao and Z. Su (2018). "Co-expression Gene Network Analysis and Functional Module Identification in Bamboo Growth and Development." Frontiers in Genetics **9**.

MacLeod, I. M., P. J. Bowman, C. J. Vander Jagt, M. Haile-Mariam, K. E. Kemper, A. J. Chamberlain, C. Schrooten, B. J. Hayes and M. E. Goddard (2016). "Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits." BMC Genomics **17**: 144.

MacLeod, I. M., B. J. Hayes and M. E. Goddard (2014). "The effects of demography and long-term selection on the accuracy of genomic prediction with sequence data." Genetics **198**(4): 1671-1684.

Makowsky, R., N. M. Pajewski, Y. C. Klimentidis, A. I. Vazquez, C. W. Duarte, D. B. Allison and G. de los Campos (2011). "Beyond missing heritability: prediction of complex traits." PLoS Genet. **7**(4): e1002051.

Makrodimitris, S., R. C. H. J. van Ham and M. J. T. Reinders (2020). "Automatic Gene Function Prediction in the 2020's." Genes **11**(11): 1264.

Manthena, V., D. Jarquín, R. K. Varshney, M. Roorkiwal, G. P. Dixit, C. Bharadwaj and R. Howard (2022). "Evaluating dimensionality reduction for genomic prediction." Frontiers in Genetics **13**.

Matías-Hernández, L., A. E. Aguilar-Jaramillo, R. A. Cigliano, W. Sanseverino and S. Pelaz (2015). "Flowering and trichome development share hormonal and transcription factor regulation." Journal of Experimental Botany **67**(5): 1209-1219.

Mering, C. v., M. Huynen, D. Jaeggi, S. Schmidt, P. Bork and B. Snel (2003). "STRING: a database of predicted functional associations between proteins." Nucleic acids Research **31**(1): 258-261.

Merrick, L. F., A. W. Herr, K. S. Sandhu, D. N. Lozada and A. H. Carter (2022). "Optimizing Plant Breeding Programs for Genomic Selection." Agronomy **12**(3): 714.

Meuwissen, T. H., B. J. Hayes and M. E. Goddard (2001). "Prediction of total genetic value using genome-wide dense marker maps." Genetics **157**(4): 1819-1829.

Meuwissen, T. H. E., B. Hayes and M. Goddard (2001). "Prediction of total genetic value using genome-wide dense marker maps." Genetics **157**(4): 1819_1829.

Mi, H., A. Muruganujan, D. Ebert, X. Huang and P. D. Thomas (2019). "PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools." Nucleic acids Research **47**(D1): D419-D426.

Michaelson, L. V., J. A. Napier, D. Molino and J. D. Faure (2016). "Plant sphingolipids: Their importance in cellular organization and adaption." Biochim Biophys Acta **1861**(9 Pt B): 1329-1335.

Michel, S., C. Wagner, T. Nosenko, B. Steiner, M. Samad-Zamini, M. Buerstmayr, K. Mayer and H. Buerstmayr (2021). "Merging Genomics and Transcriptomics for Predicting Fusarium Head Blight Resistance in Wheat." Genes (Basel) **12**(1).

Mieth, B., A. Rozier, J. A. Rodriguez, M. M.-C. Höhne, N. Görnitz and K.-R. Müller (2020). "DeepCOMBI: Explainable artificial intelligence for the analysis and discovery in genome-wide association studies." bioRxiv: 2020.2011.2006.371542.

Mikkola, P., O. A. Martin, S. Chandramouli, M. Hartmann, O. A. Pla, O. Thomas, H. Pesonen, J. Corander, A. Vehtari and S. Kaski (2021). "Prior knowledge elicitation: The past, present, and future." arXiv preprint arXiv:2112.01380.

Moeinizade, S., G. Hu and L. Wang (2022). "A reinforcement Learning approach to resource allocation in genomic selection." Intelligent Systems with Applications **14**: 200076.

Mollandin, F., A. Rau and P. Croiseau (2020). "An evaluation of the interpretability and predictive performance of the BayesR model for genomic prediction." bioRxiv: 2020.2010.2023.351700.

Molnar, C. (2020). Interpretable machine learning, Lulu. com.

Mondal, R., M. D. Do, N. U. Ahmed, D. Walke, D. Micheel, D. Broneske, G. Saake and R. Heyer (2022). "Decision tree learning in Neo4j on homogeneous and unconnected graph nodes from biological and clinical datasets." BMC Medical Informatics and Decision Making **22**(6): 1-13.

Monteith, J. L. (1977). "Climate and the efficiency of crop production in Britain." Philosophical Transactions of the Royal Society of London. B, Biological Sciences **281**(980): 277-294.

Montesinos-López, A., O. A. Montesinos-López, J. Cuevas, W. A. Mata-López, J. Burgueño, S. Mondal, J. Huerta, R. Singh, E. Autrique, L. González-Pérez and J. Crossa (2017). "Genomic Bayesian functional regression models with interactions for predicting wheat grain yield using hyper-spectral image data." Plant Methods **13**(1): 62.

Montesinos-López, A., O. A. Montesinos-López, D. Gianola, J. Crossa and C. M. Hernández-Suárez (2018). "Multi-environment Genomic Prediction of Plant Traits Using Deep Learners With Dense Architecture." G3: Genes|Genomes|Genetics **8**(12): 3813-3828.

Montesinos-López, O. A., J. Martín-Vallejo, J. Crossa, D. Gianola, C. M. Hernández-Suárez, A. Montesinos-López, P. Juliana and R. Singh (2019). "A Benchmarking Between Deep Learning, Support Vector Machine and Bayesian Threshold Best Linear Unbiased Prediction for Predicting Ordinal Traits in Plant Breeding." G3: Genes|Genomes|Genetics **9**(2): 601_618.

Montesinos-López, O. A., A. Montesinos-López, P. Pérez-Rodríguez, J. A. Barrón-López, J. W. R. Martini, S. B. Fajardo-Flores, L. S. Gaytan-Lugo, P. C. Santana-Mancilla and J. Crossa (2021). "A review of deep learning applications for genomic selection." BMC Genomics **22**(1): 19.

Montesinos-López, O. A., A. Montesinos-López, C. M. Hernandez-Suarez, J. A. Barrón-López and J. Crossa (2021). "Deep-learning power and perspectives for genomic selection." The plant genome **14**(3): e20122.

Moore, J. H., R. Amos, J. Kiralis and P. C. Andrews (2015). "Heuristic identification of biological architectures for simulating complex hierarchical genetic interactions." Genet Epidemiol **39**(1): 25_34.

Morgan, T. H. (1911). "An attempt to analyze the constitution of the chromosomes on the basis of sex-limited inheritance in Drosophila." Journal of Experimental Zoology **11**(4): 365-413.

Morgante, F. (2018). Genetic Analysis and Prediction of Complex Traits in Drosophila melanogaster, North Carolina State University.

Moser, G., S. H. Lee, B. J. Hayes, M. E. Goddard, N. R. Wray and P. M. Visscher (2015). "Simultaneous discovery, estimation and prediction analysis of complex traits using a Bayesian mixture model." PLoS Genetics **11**(4): e1004969.

Mostafavi, S., D. Ray, D. Warde-Farley, C. Grouios and Q. Morris (2008). "GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function." Genome biology **9**: 1-15.

Movahedi, S., Y. Van de Peer and K. Vandepoele (2011). "Comparative network analysis reveals that tissue specificity and gene function are important factors influencing the mode of expression evolution in Arabidopsis and rice." Plant physiology **156**(3): 1316-1330.

Murchie, E. H. and T. Lawson (2013). "Chlorophyll fluorescence analysis: a guide to good practice and understanding some new applications." Journal of experimental botany **64**(13): 3983-3998.

NEI, M. (1960). "Studies on the application of biometrical genetics to plant breeding." Plant Breeding **82**(0).

Nepusz, T., H. Yu and A. Paccanaro (2012). "Detecting overlapping protein complexes in protein-protein interaction networks." Nature methods **9**(5): 471-472.

Nguyen, T.-T., H. Zhao, J. Z. Huang, T. T. Nguyen and M. J. Li (2015). A new feature sampling method in random forests for predicting high-dimensional data. Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer.

Nguyen, T. and L. Le (2018). Detection of SNP-SNP Interactions in Genome-wide Association Data Using Random Forests and Association Rules. 2018 12th International Conference on Software, Knowledge, Information Management & Applications (SKIMA).

Nijveen, H., W. Ligterink, J. J. Keurentjes, O. Loudet, J. Long, M. G. Sterken, P. Prins, H. W. Hilhorst, D. De Ridder and J. E. Kammenga (2017). "Ara QTL–workbench and archive for systems genetics in Arabidopsis thaliana." The Plant Journal **89**(6): 1225-1235.

Norman, A., J. Taylor, J. Edwards and H. Kuchel (2018). "Optimising Genomic Selection in Wheat: Effect of Marker Density, Population Size and Population Structure on Prediction Accuracy." G3 (Bethesda).

Norman, A., J. Taylor, E. Tanaka, P. Telfer, J. Edwards, J. P. Martinant and H. Kuchel (2017). "Increased genomic prediction accuracy in wheat breeding using a large Australian panel." Theor Appl Genet **130**(12): 2543_2555.

Novakovsky, G., N. Dexter, M. W. Libbrecht, W. W. Wasserman and S. Mostafavi (2022). "Obtaining genetics insights from deep learning via explainable artificial intelligence." Nature Reviews Genetics.

Oakley, C. G., L. Savage, S. Lotz, G. R. Larson, M. F. Thomashow, D. M. Kramer and D. W. Schemske (2018). "Genetic basis of photosynthetic responses to cold in two locally adapted populations of Arabidopsis thaliana." Journal of Experimental Botany **69**(3): 699-709.

Odegard, J., U. Indahl, I. Stranden and T. H. E. Meuwissen (2018). "Large-scale genomic prediction using singular value decomposition of the genotype matrix." Genet Sel Evol **50**(1): 6.

Ogawa, S., H. Matsuda, Y. Taniguchi, T. Watanabe, Y. Sugimoto and H. Iwaisaki (2016). "Estimation of variance and genomic prediction using genotypes imputed from low-density marker subsets for carcass traits in Japanese black cattle." Anim Sci J **87**(9): 1106-1113.

Ogutu, J. O., H. P. Piepho and T. Schulz-Streeck (2011). "A comparison of random forests, boosting and support vector machines for genomic selection." BMC Proceedings **5 Suppl 3**: S11.

Ogutu, J. O., T. Schulz-Streeck and H. P. Piepho (2012). "Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions." BMC Proc **6 Suppl 2**: S10.

OpenAI, T. (2022). "Chatgpt: Optimizing language models for dialogue." OpenAI.

Orlenko, A. and J. H. Moore (2021). "A comparison of methods for interpreting random forest models of genetic association in the presence of non-additive interactions." BioData mining **14**(1): 1-17.

Oughtred, R., J. Rust, C. Chang, B. J. Breitkreutz, C. Stark, A. Willems, L. Boucher, G. Leung, N. Kolas and F. Zhang (2021). "The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions." Protein Science **30**(1): 187-200.

Park, T. and G. Casella (2008). "The Bayesian Lasso." Journal of the American Statistical Association **103**(482): 681_686.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss and V. Dubourg (2011). "Scikit-learn: Machine learning in Python." the Journal of machine Learning research **12**: 2825-2830.

Pérez-Enciso, M. and L. M. Zingaretti (2019). "A guide on deep learning for complex trait genomic prediction." Genes **10**(7): 553.

Pérez-Rodríguez, P., D. Gianola, J. M. González-Camacho, J. Crossa, Y. Manès and S. Dreisigacker (2012). "Comparison between linear and non-parametric regression models for genome-enabled prediction in wheat." G3: Genes, Genomes, Genetics **2**(12): 1595_1605.

Pérez, P. and G. de Los Campos (2014). "Genome-wide regression and prediction with the BGLR statistical package." Genetics **198**(2): 483_495.

Pingali, P. L. (2012). "Green Revolution: Impacts, limits, and the path ahead." Proceedings of the National Academy of Sciences **109**(31): 12302-12308.

Pintus, M. A., G. Gaspa, E. L. Nicolazzi, D. Vicario, A. Rossoni, P. Ajmone-Marsan, A. Nardone, C. Dimauro and N. P. Macciotta (2012). "Prediction of genomic breeding values for dairy traits in Italian Brown and Simmental bulls using a principal component approach." J Dairy Sci **95**(6): 3390-3400.

Pook, T., J. Freudenthal, A. Korte and H. Simianer (2020). "Using local convolutional neural networks for genomic prediction." bioRxiv.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. De Bakker and M. J. Daly (2007). "PLINK: a tool set for whole-genome association and population-based linkage analyses." The American journal of human genetics **81**(3): 559_575.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. de Bakker, M. J. Daly and P. C. Sham (2007). "PLINK: a tool set for whole-genome association and population-based linkage analyses." Am. J. Hum. Genet. **81**(3): 559-575.

Rabab, S., E. Breen, A. Gebremedhin, F. Shi, P. Badenhorst, Y.-P. P. Chen and H. D. Daetwyler (2021). "A New Method for Extracting Individual Plant Bio-Characteristics from High-Resolution Digital Images." Remote Sensing **13**(6): 1212.

Radivojac, P., W. T. Clark, T. R. Oron, A. M. Schnoes, T. Wittkop, A. Sokolov, K. Graim, C. Funk, K. Verspoor and A. Ben-Hur (2013). "A large-scale evaluation of computational protein function prediction." Nature Methods **10**(3): 221-227.

Resende, R. T., H.-P. Piepho, G. J. M. Rosa, O. B. Silva-Junior, F. F. e Silva, M. D. V. de Resende and D. Grattapaglia (2021). "Enviromics in breeding: applications and perspectives on envirotypic-assisted selection." Theoretical and Applied Genetics **134**(1): 95-112.

Rhee, S. Y., W. Beavis, T. Z. Berardini, G. Chen, D. Dixon, A. Doyle, M. Garcia-Hernandez, E. Huala, G. Lander and M. Montoya (2003). "The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community." Nucleic acids research **31**(1): 224-228.

Rhee, S. Y., P. Zhang, H. Foerster and C. Tissier (2006). AraCyc: Overview of an Arabidopsis Metabolism Database and its Applications for Plant Research. Plant Metabolomics**:** 141-154.

Richly, E. and D. Leister (2004). "An improved prediction of chloroplast proteins reveals diversities and commonalities in the chloroplast proteomes of Arabidopsis and rice." Gene **329**: 11-16.

Riedelsheimer, C., A. Czedik-Eysenberg, C. Grieder, J. Lisec, F. Technow, R. Sulpice, T. Altmann, M. Stitt, L. Willmitzer and A. E. Melchinger (2012). "Genomic and metabolic prediction of complex heterotic traits in hybrid maize." Nat Genet **44**(2): 217-220.

Rieger, L., C. Singh, W. Murdoch and B. Yu (2020). Interpretations are useful: penalizing explanations to align neural networks with prior knowledge. International conference on machine learning, PMLR.

Rimbert, H., B. Darrier, J. Navarro, J. Kitt, F. Choulet, M. Leveugle, J. Duarte, N. Rivière, K. Eversole and J. Le Gouis (2018). "High throughput SNP discovery and genotyping in hexaploid wheat." PloS ONE **13**(1): e0186329.

Rincent, R., J.-P. Charpentier, P. Faivre-Rampant, E. Paux, J. Le Gouis, C. Bastien and V. Segura (2018). "Phenomic Selection Is a Low-Cost and High-Throughput Method Based on Indirect Predictions: Proof of Concept on Wheat and Poplar." G3 Genes|Genomes|Genetics **8**(12): 3961-3972.

Rincent, R., J.-P. Charpentier, P. Faivre-Rampant, E. Paux, J. Le Gouis, C. Bastien and V. Segura (2018). "Phenomic selection: a low-cost and high-throughput alternative to genomic selection." bioRxiv: 302117.

Rincent, R., M. Malosetti, B. Ababaei, G. Touzy, A. Mini, M. Bogard, P. Martre, J. Le Gouis and F. van Eeuwijk (2019). "Using crop growth model stress covariates and AMMI decomposition to better predict genotype-by-environment interactions." Theoretical and Applied Genetics **132**(12): 3399-3411.

Robert, P., J. Auzanneau, E. Goudemand, F.-X. Oury, B. Rolland, E. Heumez, S. Bouchet, J. Le Gouis and R. Rincent (2022). "Phenomic selection in wheat breeding: identification and optimisation of factors influencing prediction accuracy and comparison to genomic selection." Theoretical and Applied Genetics.

Robert, P., J. Le Gouis, B. Consortium and R. Rincent (2020). "Combining crop growth modeling with trait-assisted prediction improved the prediction of genotype by environment interactions." Frontiers in Plant Science **11**: 827.

Roberts, H. F. (1929). "Plant hybridization before Mendel." Plant hybridization before Mendel.

Rohde, P. D., D. Demontis, A. Børglum and P. Sørensen (2017). Improved prediction of genetic predisposition to psychiatric disorders using genomic feature best linear unbiased prediction models. ESHG 2017: European Society of Human Genetics Annual Meeting.

Rohde, P. D., I. Fourie Sørensen and P. Sørensen (2020). "qgg: an R package for large-scale quantitative genetic analyses." Bioinformatics **36**(8): 2614-2615.

Ros-Freixedes, R., M. Johnsson, A. Whalen, C.-Y. Chen, B. D. Valente, W. O. Herring, G. Gorjanc and J. M. Hickey (2022). "Genomic prediction with whole-genome sequence data in intensely selected pig lines." Genetics Selection Evolution **54**(1): 65.

Royal Society of London (2009). Reaping the Benefits: Science and the Sustainable Intensification of Global Agriculture. London.

Roychowdhury, S., M. Diligenti and M. Gori (2021). "Regularizing deep networks with prior knowledge: A constraint-based approach." Knowledge-Based Systems **222**: 106989.

Rueden, L. v., S. Mayer, K. Beckh, B. Georgiev, S. Giesselbach, R. Heese, B. Kirsch, J. Pfrommer, A. Pick, R. Ramamurthy, M. Walczak, J. Garcke, C. Bauckhage and J. Schuecker (2023). "Informed Machine Learning – A Taxonomy and Survey of Integrating Prior Knowledge into Learning Systems." IEEE Transactions on Knowledge and Data Engineering **35**(1): 614-633.

Saha, S., L. Perrin, L. Röder, C. Brun and L. Spinelli (2022). "Epi-MEIF: detecting higher order epistatic interactions for complex traits using mixed effect conditional inference forests." Nucleic Acids Research **50**(19): e114-e114.

Sandhu, K., S. S. Patil, M. Pumphrey and A. Carter (2021). "Multitrait machine- and deep-learning models for genomic selection using spectral information in a wheat breeding program." The Plant Genome **14**(3): e20119.

Sandhu, K. S., D. N. Lozada, Z. Zhang, M. O. Pumphrey and A. H. Carter (2021). "Deep learning for predicting complex traits in spring wheat breeding program." Frontiers in Plant Science **11**: 2084.

Sandhu, K. S., P. D. Mihalyov, M. J. Lewien, M. O. Pumphrey and A. H. Carter (2021). "Combining Genomic and Phenomic Information for Predicting Grain Protein Content and Grain Yield in Spring Wheat." Frontiers in Plant Science **12**.

Sapkota, S., J. L. Boatwright, K. Jordan, R. Boyles and S. Kresovich (2020). "Multi-Trait Regressor Stacking Increased Genomic Prediction Accuracy of Sorghum Grain Composition." Agronomy **10**(9): 1221.

Sarup, P., J. Jensen, T. Ostersen, M. Henryon and P. Sorensen (2016). "Increased prediction accuracy using a genomic feature model including prior information on quantitative trait locus regions in purebred Danish Duroc pigs." BMC Genetics **17**: 11.

Schmalohr, C., J. Grossbach, A. Beyer and M. Clément-Ziza (2018). "Exploiting the Structure of Random Forest for the Detection of Epistatic Interactions. *ISMB/ECCB 2017 Conference*. 2017"

Schrag, T. A., M. Westhues, W. Schipprack, F. Seifert, A. Thiemann, S. Scholten and A. E. Melchinger (2018). "Beyond Genomic Prediction: Combining Different Types of omics Data Can Improve Prediction of Hybrid Performance in Maize." Genetics **208**(4): 1373-1385.

Schulthess, A. W., Y. Wang, T. Miedaner, P. Wilde, J. C. Reif and Y. Zhao (2016). "Multiple-trait- and selection indices-genomic predictions for grain yield and protein content in rye for feeding purposes." Theor Appl Genet **129**(2): 273-287.

Scott, M. F., O. Ladejobi, S. Amer, A. R. Bentley, J. Biernaskie, S. A. Boden, M. Clark, M. Dell'Acqua, L. E. Dixon, C. V. Filippi, N. Fradgley, K. A. Gardner, I. J. Mackay, D. O'Sullivan, L. Percival-Alwyn, M. Roorkiwal, R. K. Singh, M. Thudi, R. K. Varshney, L. Venturini, A. Whan, J. Cockram and R. Mott (2020). "Multi-parent populations in crops: a toolbox integrating genomics and genetic mapping with breeding." Heredity **125**(6): 396-416.

Selga, C., A. Koc, A. Chawade and R. Ortiz (2021). "A Bioinformatics Pipeline to Identify a Subset of SNPs for Genomics-Assisted Potato Breeding." Plants **10**(1): 30.

Serin, E. A. R., H. Nijveen, H. W. M. Hilhorst and W. Ligterink (2016). "Learning from Co-expression Networks: Possibilities and Challenges." Frontiers in Plant Science **7**.

Shahi, D., J. Guo, S. Pradhan, J. Khan, M. Avci, N. Khan, J. McBreen, G. Bai, M. Reynolds, J. Foulkes and M. A. Babar (2022). "Multi-trait genomic prediction using in-season physiological parameters increases prediction accuracy of complex traits in US wheat." BMC Genomics **23**(1): 298.

Singhal, P., Y. Veturi, S. M. Dudek, A. Lucas, A. Frase, K. van Steen, S. J. Schrodi, D. Fasel, C. Weng, R. Pendergrass, D. J. Schaid, I. J. Kullo, O. Dikilitas, P. M. A. Sleiman, H. Hakonarson, J. H. Moore, S. M. Williams, M. D. Ritchie and S. S. Verma (2023). "Evidence of epistasis in regions of long-range linkage disequilibrium across five complex diseases in the UK Biobank and eMERGE datasets." The American Journal of Human Genetics **110**(4): 575-591.

Škunca, N., A. Altenhoff and C. Dessimoz (2012). "Quality of computationally inferred gene ontology annotations." PLoS computational biology **8**(5): e1002533.

Snow, O., H. S. Noghabi, J. Lu, O. Zolotareva, M. Lee and M. Ester (2019). "BDKANN – Biological Domain Knowledge-based Artificial Neural Network for drug response prediction." bioRxiv: 840553.

South, P. F., A. P. Cavanagh, H. W. Liu and D. R. Ort (2019). "Synthetic glycolate metabolism pathways stimulate crop growth and productivity in the field." Science **363**(6422).

Speed, D. and D. J. Balding (2014). "MultiBLUP: improved SNP-based prediction for complex traits." Genome Research **24**(9): 1550-1557.

Speed, D. and D. J. Balding (2019). "SumHer better estimates the SNP heritability of complex traits from summary statistics." Nat Genet **51**(2): 277_284.

Speed, D., J. Holmes and D. J. Balding (2020). "Evaluating and improving heritability models using summary statistics." Nature Genetics **52**(4): 458_462.

Subramanian, I., S. Verma, S. Kumar, A. Jere and K. Anamika (2020). "Multi-omics Data Integration, Interpretation, and Its Application." Bioinformatics and biology insights **14**: 1177932219899051-1177932219899051.

Sukumaran, S., D. Jarquin, J. Crossa and M. Reynolds (2018). "Genomic-enabled Prediction Accuracies Increased by Modeling Genotype x Environment Interaction in Durum Wheat." Plant Genome **11**(2).

Sundararajan, M., A. Taly and Q. Yan (2017). Axiomatic attribution for deep networks. International conference on machine learning, PMLR.

Supek, F., M. Bošnjak, N. Škunca and T. Šmuc (2011). "REVIGO summarizes and visualizes long lists of gene ontology terms." PLoS ONE **6**(7): e21800.

Szklarczyk, D., A. L. Gable, D. Lyon, A. Junge, S. Wyder, J. Huerta-Cepas, M. Simonovic, N. T. Doncheva, J. H. Morris and P. Bork (2019). "STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets." Nucleic acids research **47**(D1): D607-D613.

Szklarczyk, D., A. L. Gable, K. C. Nastou, D. Lyon, R. Kirsch, S. Pyysalo, N. T. Doncheva, M. Legeay, T. Fang, P. Bork, L. J. Jensen and C. von Mering (2021). "The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets." Nucleic Acids Res **49**(D1): D605-D612.

Tang, Y., N. Gill, E. LeDell and V. Ovsyannikov (2021). "h2o4gpu: Interface to 'H2O4GPU'."

Technow, F., C. Riedelsheimer, T. A. Schrag and A. E. Melchinger (2012). "Genomic prediction of hybrid performance in maize with models incorporating dominance and population specific marker effects." Theor. Appl. Genet. **125**(6): 1181_1194.

Teng, J., S. Huang, Z. Chen, N. Gao, S. Ye, S. Diao, X. Ding, X. Yuan, H. Zhang, J. Li and Z. Zhang (2020). "Optimizing genomic prediction model given causal genes in a dairy cattle population." Journal of Dairy Science **103**(11): 10299-10310.

Tessema, B. B., H. Liu, A. C. Sørensen, J. R. Andersen and J. Jensen (2020). "Strategies Using Genomic Selection to Increase Genetic Gain in Breeding Programs for Wheat." Front. Genet. **11**: 578123.

Tessema, B. B., H. Liu, A. C. Sørensen, J. R. Andersen and J. Jensen (2020). "Strategies Using Genomic Selection to Increase Genetic Gain in Breeding Programs for Wheat." Frontiers in Genetics **11**(1538).

Theeuwen, T. P., L. L. Logie, S. Put, H. Bagheri, K. Losinski, J. Drouault, P. J. Flood, C. Hanhart, F. F. Becker and R. Wijfjes (2022). "Plethora of QTLs found in Arabidopsis thaliana reveals complexity of genetic variation for photosynthesis in dynamic light conditions." bioRxiv.

Tian, Q., C. Lewis-Beck, J. B. Niemi and W. Q. Meeker (2023). "Specifying prior distributions in reliability applications." Applied Stochastic Models in Business and Industry **n/a**(n/a): 1-58.

Tibshirani, R. (1996). "Regression shrinkage and selection via the lasso." Journal of the Royal Statistical Society: Series B (Methodological) **58**(1): 267-288.

Tiezzi, F. and C. Maltecca (2015). "Accounting for trait architecture in genomic predictions of US Holstein cattle using a weighted realized relationship matrix." Genetics Selection Evolution **47**(1): 24.

Togninalli, M., Ü. Seren, J. A. Freudenthal, J. G. Monroe, D. Meng, M. Nordborg, D. Weigel, K. Borgwardt, A. Korte and D. G. Grimm (2020). "AraPheno and the AraGWAS Catalog 2020: a major database update including RNA-Seq and knockout mutation data for Arabidopsis thaliana." Nucleic Acids Research **48**(D1): D1063-D1068.

Tong, H., A. Küken and Z. Nikoloski (2020). "Integrating molecular markers into metabolic models improves genomic selection for Arabidopsis growth." Nature Communications **11**(1): 1-9.

Torstensson, E. (2017). "Using LASSO regularization as a feature selection tool."

Uemoto, Y., S. Sasaki, T. Kojima, Y. Sugimoto and T. Watanabe (2015). "Impact of QTL minor allele frequency on genomic evaluation using real genotype data and simulated phenotypes in Japanese Black cattle." BMC Genetics **16**(1): 134.

Uzunangelov, V., C. K. Wong and J. Stuart (2020). "Highly Accurate Cancer Phenotype Prediction with AKLIMATE, a Stacked Kernel Learner Integrating Multimodal Genomic Data and Pathway Knowledge." bioRxiv: 2020.2007.2015.205575.

van Bezouw, R., J. J. B. Keurentjes, J. Harbinson and M. G. M. Aarts (2018). "Converging phenomics and genomics to study natural variation in plant photosynthesis efficiency." Plant J.

van Bezouw, R. F. H. M., J. J. B. Keurentjes, J. Harbinson and M. G. M. Aarts (2019). "Converging phenomics and genomics to study natural variation in plant photosynthetic efficiency." The Plant Journal **97**(1): 112-133.

van Dijk, A. D. J., G. Kootstra, W. Kruijer and D. de Ridder (2021). "Machine learning in plant science and plant breeding." iScience **24**(1): 101890.

van Hilten, A., S. A. Kushner, M. Kayser, M. A. Ikram, H. Adams, C. Klaver, W. J. Niessen and G. V. Roshchupkin (2020). "GenNet framework: interpretable neural networks for phenotype prediction." bioRxiv.

van Rooijen, R., M. G. M. Aarts and J. Harbinson (2015). "Natural Genetic Variation for Acclimation of Photosynthetic Light Use Efficiency to Growth Irradiance in Arabidopsis." Plant Physiology **167**(4): 1412-U1484.

van Rooijen, R., M. G. M. Aarts and J. Harbinson (2015). "Natural Genetic Variation for Acclimation of Photosynthetic Light Use Efficiency to Growth Irradiance in Arabidopsis." Plant Physiology **167**(4): 1412-1429.

van Rooijen, R., W. Kruijer, R. Boesten, F. A. van Eeuwijk, J. Harbinson and M. G. M. Aarts (2017). "Natural variation of YELLOW SEEDLING1 affects photosynthetic acclimation of Arabidopsis thaliana." Nature Communications **8**(1): 1421.

VanRaden, P. M. (2008). "Efficient methods to compute genomic predictions." Journal of Dairy Science **91**(11): 4414_4423.

VanRaden, P. M., M. E. Tooker, J. R. O'Connell, J. B. Cole and D. M. Bickhart (2017). "Selecting sequence variants to improve genomic predictions for dairy cattle." Genetics Selection Evolution **49**(1): 32.

Varona, L., A. Legarra, M. A. Toro and Z. G. Vitezica (2018). "Non-additive Effects in Genomic Selection." Frontiers in Genetics **9**(78).

Vasquez-Kool, J. (2019). "Coheritability and coenvironmentability as concepts for partitioning the phenotypic correlation." bioRxiv: 598623.

Vaz, J. M. and S. Balaji (2021). "Convolutional neural networks (CNNs): Concepts and applications in pharmacogenomics." Molecular Diversity **25**(3): 1569-1584.

Veerkamp, R. F., A. C. Bouwman, C. Schrooten and M. P. Calus (2016). "Genomic prediction using preselected DNA variants from a GWAS with whole-genome sequence data in Holstein-Friesian cattle." Genet Sel Evol **48**(1): 95.

Velazco, J. G., D. R. Jordan, E. S. Mace, C. H. Hunt, M. Malosetti and F. A. van Eeuwijk (2019). "Genomic Prediction of Grain Yield and Drought-Adaptation Capacity in Sorghum Is Enhanced by Multi-Trait Analysis." Frontiers in Plant Science **10**.

Venkatesan, A., G. Tagny Ngompé, N. E. Hassouni, I. Chentli, V. Guignon, C. Jonquet, M. Ruiz and P. Larmande (2018). "Agronomic Linked Data (AgroLD): A knowledge-based system to enable integrative biology in agronomy." PLoS One **13**(11): e0198270.

Vieira, I., J. Dos Santos, L. Pires, B. Lima and M. Balestre (2017). "Assessing non-additive effects in GBLUP model." Genetics and Molecular Research **16**(2).

Visscher, P. M., G. Hemani, A. A. E. Vinkhuyzen, G.-B. Chen, S. H. Lee, N. R. Wray, M. E. Goddard and J. Yang (2014). "Statistical Power to Detect Genetic (Co)Variance of Complex Traits Using SNP Data in Unrelated Samples." PLOS Genetics **10**(4): e1004269.

von Werra, L., M. Schöngens, E. D. Gamsiz Uzun and C. Eickhoff (2019). Generative adversarial networks in precision oncology. Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval.

Walsh, B. (2004). "Population-and quantitative-genetic models of selection limits." Plant breeding reviews **24**(1): 177-226.

Wang, J., Z. Zhou, Z. Zhang, H. Li, D. Liu, Q. Zhang, P. J. Bradbury, E. S. Buckler and Z. Zhang (2018). "Expanding the BLUP alphabet for genomic prediction adaptable to the genetic architectures of complex traits." Heredity (Edinb).

Wang, Q., Y. Yu, J. Yuan, X. Zhang, H. Huang, F. Li and J. Xiang (2017). "Effects of marker density and population structure on the genomic prediction accuracy for growth trait in Pacific white shrimp Litopenaeus vannamei." BMC Genet **18**(1): 45.

Wang, S., C. Aggarwal and H. Liu (2017). Using a random forest to inspire a neural network and improving on it. Proceedings of the 2017 SIAM international conference on data mining, SIAM.

Wang, W., R. Mauleon, Z. Hu, D. Chebotarov, S. Tai, Z. Wu, M. Li, T. Zheng, R. R. Fuentes, F. Zhang, L. Mansueto, D. Copetti, M. Sanciangco, K. C. Palis, J. Xu, C. Sun, B. Fu, H. Zhang, Y. Gao, X. Zhao, F. Shen, X. Cui, H. Yu, Z. Li, M. Chen, J. Detras, Y. Zhou, X. Zhang, Y. Zhao, D. Kudrna, C. Wang, R. Li, B. Jia, J. Lu, X. He, Z. Dong, J. Xu, Y. Li, M. Wang, J. Shi, J. Li, D. Zhang, S. Lee, W. Hu, A. Poliakov, I. Dubchak, V. J. Ulat, F. N. Borja, J. R. Mendoza, J. Ali, J. Li, Q. Gao, Y. Niu, Z. Yue, M. E. B. Naredo, J. Talag, X. Wang, J. Li, X. Fang, Y. Yin, J. C. Glaszmann, J. Zhang, J. Li, R. S. Hamilton, R. A. Wing, J. Ruan, G. Zhang, C. Wei, N. Alexandrov, K. L. McNally, Z. Li and H. Leung (2018). "Genomic variation in 3,010 diverse accessions of Asian cultivated rice." Nature **557**(7703): 43-49.

Warde-Farley, D., S. L. Donaldson, O. Comes, K. Zuberi, R. Badrawi, P. Chao, M. Franz, C. Grouios, F. Kazi and C. T. Lopes (2010). "The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function." Nucleic acids Research **38**(suppl_2): W214-W220.

Wei, X., C. Zhang, P. L. Freddolino and Y. Zhang (2020). "Detecting Gene Ontology misannotations using taxon-specific rate ratio comparisons." Bioinformatics **36**(16): 4383-4388.

Weiss, K., T. M. Khoshgoftaar and D. Wang (2016). "A survey of transfer learning." Journal of Big Data **3**(1): 9.

Weraduwage, S. M., J. Chen, F. C. Anozie, A. Morales, S. E. Weise and T. D. Sharkey (2015). "The relationship between leaf area growth and biomass accumulation in Arabidopsis thaliana." Frontiers in Plant Science **6**(167).

Wheeler, H. E., K. Aquino-Michaels, E. R. Gamazon, V. V. Trubetskoy, M. E. Dolan, R. S. Huang, N. J. Cox and H. K. Im (2014). "Poly-omic prediction of complex traits: OmicKriging." Genetic Epidemiology **38**(5): 402-415.

Wijfjes, R. (2021). "Computational analysis of copy number variation in plants." Doctoral Dissertation, WUR

Wijfjes, R. (2021). "Supplementary Material Chapter 5. Zenodo."

Wijfjes, R. Y., S. Smit and D. de Ridder (2019). "Hecaton: reliably detecting copy number variation in plant genomes using short read sequencing data." BMC Genomics **20**(1): 818.

Wolc, A., J. Arango, P. Settar, J. E. Fulton, N. P. O'Sullivan, J. C. M. Dekkers, R. Fernando and D. J. Garrick (2016). "Mixture models detect large effect QTL better than GBLUP and result in more accurate and persistent predictions." Journal of Animal Science and Biotechnology **7**(1): 7.

Wolc, A. and J. C. M. Dekkers (2022). "Application of Bayesian genomic prediction methods to genome-wide association analyses." Genetics Selection Evolution **54**(1): 31.

Wray, N. R., J. Yang, B. J. Hayes, A. L. Price, M. E. Goddard and P. M. Visscher (2013). "Pitfalls of predicting complex traits from SNPs." Nat Rev Genet **14**(7): 507-515.

Wright, M. N. and A. Ziegler (2017). "ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R." Journal of Statistical Software **77**(1): 1_17.

Wright, M. N., A. Ziegler and I. R. König (2016). "Do little interactions get lost in dark random forests?" BMC Bioinformatics **17**(1): 145.

Xavier, A. (2019). "Efficient Estimation of Marker Effects in Plant Breeding." G3 (Bethesda, Md.) **9**(11): 3855-3866.

Xavier, A., W. M. Muir and K. M. Rainey (2016). "Assessing Predictive Properties of Genome-Wide Selection in Soybeans." G3 **6**(8): 2611_2616.

Xu, Y., P. Li, C. Zou, Y. Lu, C. Xie, X. Zhang, B. M. Prasanna and M. S. Olsen (2017). "Enhancing genetic gain in the era of molecular breeding." Journal of Experimental Botany **68**(11): 2641-2666.

Yan, J. and X. Wang (2023). "Machine learning bridges omics sciences and plant breeding." Trends in Plant Science **28**(2): 199-210.

Yan, J., Y. Xu, Q. Cheng, S. Jiang, Q. Wang, Y. Xiao, C. Ma, J. Yan and X. Wang (2021). "LightGBM: accelerated genomically designed crop breeding through ensemble learning." Genome Biology **22**(1): 271.

Yang, X., S. Yang, H. Qi, T. Wang, H. Li and Z. Zhang (2020). "PlaPPISite: a comprehensive resource for plant protein-protein interaction sites." BMC plant biology **20**(1): 1-11.

Yao, C., X. Zhu and K. A. Weigel (2016). "Semi-supervised learning for genomic prediction of novel traits with small reference populations: an application to residual feed intake in dairy cattle." Genet Sel Evol **48**(1): 84.

Yelmen, B., A. Decelle, L. Ongaro, D. Marnetto, C. Tallec, F. Montinaro, C. Furtlehner, L. Pagani and F. Jay (2021). "Creating artificial human genomes using generative neural networks." PLoS genetics **17**(2): e1009303.

Yin, L., H. Zhang, X. Zhou, X. Yuan, S. Zhao, X. Li and X. Liu (2020). "KAML: improving genomic prediction accuracy of complex traits using machine learning determined parameters." Genome Biology **21**(1): 146.

Yoshida, M. and A. Koike (2011). "SNPInterForest: A new method for detecting epistatic interactions." BMC Bioinformatics **12**(1): 469.

Zhang, A., H. Wang, Y. Beyene, K. Semagn, Y. Liu, S. Cao, Z. Cui, Y. Ruan, J. Burgueno, F. San Vicente, M. Olsen, B. M. Prasanna, J. Crossa, H. Yu and X. Zhang (2017). "Effect of Trait Heritability, Training Population Size and Marker Density on Genomic Prediction Accuracy Estimation in 22 bi-parental Tropical Maize Populations." Front Plant Sci **8**: 1916.

Zhang, M., X. L. Hu, M. Zhu, M. Y. Xu and L. Wang (2017). "Transcription factors NF-YA2 and NF-YA10 regulate leaf growth via auxin signaling in Arabidopsis." Scientific Reports **7**.

Zhang, W., L. A. Mitchell, J. S. Bader and J. D. Boeke (2020). "Synthetic Genomes." Annual Review of Biochemistry **89**(1): 77-101.

Zhang, X. and J. O. Borevitz (2009). "Global analysis of allele-specific expression in Arabidopsis thaliana." Genetics **182**(4): 943-954.

Zhang, Z., X. Ding, J. Liu, D. J. de Koning and Q. Zhang (2011). "Genomic selection for QTL-MAS data using a trait-specific relationship matrix." BMC Proceedings **5 Suppl 3**: S15.

Zhang, Z., J. Liu, X. Ding, P. Bijma, D. J. de Koning and Q. Zhang (2010). "Best linear unbiased prediction of genomic breeding values using a trait-specific marker-derived relationship matrix." PLoS ONE **5**(9).

Zhao, Y., F. Chen, R. Zhai, X. Lin, Z. Wang, L. Su and D. C. Christiani (2012). "Correction for population stratification in random forest analysis." International Journal of Epidemiology **41**(6): 1798_1806.

Zhong, S., M. Liu, Z. Wang, Q. Huang, S. Hou, Y.-C. Xu, Z. Ge, Z. Song, J. Huang and X. Qiu (2019). "Cysteine-rich peptides promote interspecific genetic isolation in Arabidopsis." Science **364**(6443): eaau9564.

Zhou, X. and M. Stephens (2012). "Genome-wide efficient mixed-model analysis for association studies." Nature Genetics **44**(7): 821-824.

Zhu, X., W. L. Leiser, V. Hahn and T. Würschum (2021). "Phenomic selection is competitive with genomic selection for breeding of complex traits." The Plant Phenome Journal **4**(1): e20027.

Zhu, X., H. P. Maurer, M. Jenz, V. Hahn, A. Ruckelshausen, W. L. Leiser and T. Würschum (2021). "The performance of phenomic selection depends on the genetic architecture of the target trait." Theoretical and Applied Genetics.

Zingaretti, L. M., S. A. Gezan, L. F. V. Ferrão, L. F. Osorio, A. Monfort, P. R. Muñoz, V. M. Whitaker and M. Pérez-Enciso (2020). "Exploring deep learning for complex trait genomic prediction in polyploid outcrossing species." Frontiers in plant science **11**: 25.

Summary

Plant breeding is the science to develop new genetically superior cultivars with favourable characters to satisfy human requirements. During the green revolution until 2000, global agricultural production doubled compared to the historical development till that time; but challenges are growing as well. Dealing with the current and anticipated challenges of the rapidly growing human and animal population, climate change and its consequent effects on environment and food security is practically not possible through the current pace for genetic improvements using conventional breeding practices. Improving plant breeding processes is, therefore, dearly required for sustainable growth of the global ecosystem.

Over the past couple of decades, technological advancements in High Throughput Sequencing (HTS) of the genomic DNA has equipped plant breeding with high resolution molecular information, and paved the way forward to mitigate the gap between current versus the required genetic potential of the existing germplasms. With high-resolution genomic markers, breeding process for acquiring desired plant characters becomes much more precise and accurate than in the conventional breeding. Genomic Selection (GS) is one such technique, where genome-wide DNA polymorphisms, often single nucleotide polymorphisms (SNPs), are used to estimate their cumulative effects, as breeding values; and select the best germplasm from a population as future parents. Pivotal to the GS framework is a genomic prediction (GP) model, that capitalises over the accurate SNP information to estimate breeding values and steer the decision making for selection. An improvement in GP can, therefore, be directly translated to the improvement in GS based breeding; and is central to this thesis.

Various factors including genetic architecture, population structure and genotype and phenotype data characteristics affect the prediction performance of GP models. As a result, several methods were proposed along with numerous extensions to account for these characteristic factors. Most commonly, the parametric Linear Mixed effect Models (LMMs) implementing GP as a whole genome regression and many non-parametric machine learning and deep learning methods have been applied. It is important to explore the potential application scenarios for certain methods under different GP problem characteristics (Chapter 2). I found out that a definite conclusion is still hard to draw, but a general guideline can still be made that ensemble ML methods along with Bayesian LMMs (e.g. BayesA and BayesB) are a reasonable choice for the traits predominantly characterised by the presence of large effects. On the other hand, complex polygenic traits governed by many small effects are hard to predict by all methods and prediction performance is generally low for the high-dimensional SNP data. This led me to explore possible solutions to improve GP for complex traits. To this end, this thesis presents novel development and application of improved GP methodologies for complex traits.

A possible choice for improving GP is to incorporate the wealth of prior biological knowledge, available freely from public repositories. However, incorporating such information into GP models still remains an open question. One strategy is to prioritise the genome-wide SNPs based on this information, resulting into different groups of SNPs, which can then be differentially prioritised in the model. At first, I used this strategy using LMMs, where the groups of SNPs were formed using functional information of the genes, in which they belong to (Chapter 3). In this connection, I utilised gene ontology (GO) and coexpressed gene clusters (COEX). The approach increased the prediction accuracy of the commonly used Genomic Best Linear Unbiased Prediction (GBLUP) method with different traits related GO and COEX groups, when tested on growth related traits i.e. photosynthetic light use efficiency of the photosystem II and projected leaf area, in *Arabidopsis thaliana*.

Next, the differential prioritisation approach in LMM was further extended to develop a novel deep learning framework, called PRIORNET, based on the fully connected feed-forward artificial neural network architecture, Multilayer Perceptron (MLP) (Chapter 4). The SNPs were grouped into knowledge and background using trait-related list of genes, GO or pathways etc, known *a priori*. Additionally, knowledge of protein-protein interactions was leveraged to make PRIORNET sparser than MLP. PRIORNET is capable of increasing the prediction accuracy up to the theoretical maximum value of a GP model when highly specific knowledge is provided. It also accommodates for more practical situations when partial or noisy knowledge is provided. Tested on both simulated phenotypes and experimental traits, it outperforms its benchmark MLP.

Another type of knowledge which can be applied to improve GP for complex traits consists of available information on additional traits. I used genetically correlated component traits, referred as secondary traits, measured along the target trait in multi-trait GP (MT-GP) modelling (Chapter 5). To demonstrate the efficacy of MT-GP, different photosynthesis parameters are used as secondary traits to predict biomass as target in *Arabidopsis thaliana*, measured as projected leaf area (PLA). Moreover, I showed that how photosynthesis-related traits can improve PLA predictions up to ~3 folds than predicting it alone, given the incident light is also dynamic.

In conclusion, this thesis aids improving genomic prediction of complex plant traits using knowledge-driven models. While significant improvement was observed in both statistical and machine learning based modelling frameworks, there is considerable room for development using benchmarked knowledge.

Curriculum vitae

Muhammad Farooq was born on July 17, 1983 in Faisalabad, Pakistan. He completed his BSc in Computer Science at Punjab University College of Information Technology (PUCIT), Lahore, in 2006, after which he went on to pursue an MSc in Systems Engineering at Pakistan Institute of Engineering and Applied Sciences (PIEAS), Islamabad, Pakistan. During his MSc, he also worked at National Institute for Lasers and Optronics (NILOP) for his thesis project on optical interference fringe analysis using image processing techniques. After finishing his MSc in 2009, he worked as a software developer at NILOP until December 2013, after which he joined the National Institute for Biotechnology and Genetic Engineering (NIBGE) where he worked as a support bioinformatician until November 2018. He started his PhD in the Bioinformatics group at Wageningen University (WU) in November 2018 and worked under the kind supervision of prof. dr. Dick de Ridder, dr. Aalt-Jan van Dijk, dr. Harm Nijveen and prof. dr. Shahid Mansoor (NIBGE), of which the results are presented in this thesis. Since April 2023, he again works at NIBGE, Pakistan on development and applications of genomic prediction methods in plants and animal breeding.

List of publications (this thesis)

Peer-reviewed

1. **Muhammad Farooq**, Aalt DJ Van Dijk, Harm Nijveen, Mark GM Aarts, Willem Kruijer, Thu-Phuong Nguyen, Shahid Mansoor, and Dick de Ridder. "Prior biological knowledge improves genomic prediction of growth-related traits in Arabidopsis thaliana." *Frontiers in Genetics* 11 (2020): 609117.
2. **Muhammad Farooq**, Aalt DJ van Dijk, Harm Nijveen, Shahid Mansoor, and Dick de Ridder. "Genomic prediction in plants: opportunities for ensemble machine learning based approaches." *F1000Research* 11, no. 802 (2022): 802.

In preparation

3. **Muhammad Farooq**, Aalt DJ van Dijk, Harm Nijveen, Shahid Mansoor, and Dick de Ridder. “PRIORNET: a multilayer perceptron for genomic prediction using SNP prioritisation and protein-protein interaction information”.
4. **Muhammad Farooq**, Aalt D. J. van Dijk, Harm Nijveen, Mark G. M. Aarts, Thu-Phuong Nguyen, Raul Wijfjes, Shahid Mansoor and Dick de Ridder. “Improving genomic prediction of biomass using photosynthesis-related traits in *Arabidopsis thaliana*”.

List of publications (in collaboration with NIBGE)

1. **Muhammad Farooq** ¥, Rubab Zahra Naqvi ¥, Imran Amin, Atiq Ur Rehman, Muhammad Asif, and Shahid Mansoor. "Transcriptome diversity assessment of *Gossypium arboreum* (FDH228) leaves under control, drought and whitefly infestation using PacBio long reads." *Gene* 852 (2023): 147065.
2. Rubab Zahra Naqvi, **Muhammad Farooq**, Syed Ali Asad Naqvi, Hamid Anees Siddiqui, Imran Amin, Muhammad Asif, and Shahid Mansoor. "Big data analytics and advanced technologies for sustainable agriculture." *Handbook of Smart Materials, Technologies, and Devices: Applications of Industry 4.0* (2020): 1-27.
3. Athar Hussain, **Muhammad Farooq**, Rubab Zahra Naqvi, Imran Amin, Khalid Pervaiz, Muhammad Saeed, Muhammad Asif, M. Shahid Mukhtar, and Shahid Mansoor. "Genome-wide identification and classification of resistance genes predicted several decoy domains in Gossypium sp." *Plant gene* 24 (2020): 100250.
4. Sonia Hussain ¥, **Muhammad Farooq** ¥, Hassan Jamil Malik, Imran Amin, Brian E. Scheffler, Jodi A. Scheffler, Shu-Sheng Liu, and Shahid Mansoor. "Whole genome sequencing of Asia II 1 species of whitefly reveals that genes involved in virus transmission and insecticide resistance have genetic variances between Asia II 1 and MEAM1 species." *BMC genomics* 20, no. 1 (2019): 1-13.

¥ Both authors contributed equally

Acknowledgements

Thanks foremost to the Almighty, who blessed me with proper health and wisdom to accomplish this milestone. Then, I would like to thank Dick de Ridder, Aalt-Jan van Dijk and Harm Nijveen for their encouragement, guidance, patience, and for cultivating my interest in doing a PhD in bioinformatics. Your contributions to this thesis and my development as a scientist cannot be overstated. Dick, your ability to keep a sharp mind despite the numerous meetings you endure is impressive and helped me to be more critical of my own work myself. Aalt-Jan, your attention to the details and urge to do things well have certainly rubbed off on me, and were particularly instrumental in finishing all of the work. Harm, your cool-headed and encouraging nature has left a deep impression on my personality. I am glad to have had you all as my supervisors.

Thanks to Mark Aarts for acting as an external advisor and providing datasets for Chapter 3 and 5. I appreciate all of the feedback given on the drafts and during numerous meetings. And to Thu-Phuong Nguyen, for generating high-throughput phenotyping datasets. Also to Raul Wijfjes for providing the genotype datasets for the Chapter 5. All of you contributed to the content and quality of the written output of this thesis, and I am glad that we were able to work closely with each other during the past years.

Thanks to Iris van den Hatert, an MSc student I supervised during the course of this thesis. You taught me a lot about supervision while I worked with you on your project, resulting in a collaboration that eventually became a chapter of this thesis.

Thanks to Marie-José van Iersel and Maria Augustijn of the secretariat for their help in taking care of all things related to administration. And to Harm Nijveen again, Jan van Haarst and Gwen Dawes for keeping all servers up and running. Also thanks to Susan Urbanus for her support from the side of EPS.

Many thanks to all of the people who were part of the Bioinformatics group for putting up with a colleague that literally towered above them for most of the day. You made my time as a PhD enjoyable in too many ways to list here. I am pleased to interact with many of the PhD graduates including the first one. Specifically, I would like to thank Ehsan Motazedi, Siavash Sheikhizadeh Anari, Vittorio Tracanna, Mehmet Akdel, Janani Durairaj, Carlos de Lannoy, Eef Jonkheer, Ronald de Jongh, Barbara Terlouw, Rens Holmer, Kumar Singh, Lotte Pronk, Dirk-Jan van Workum, Hannah Augustijn, Niek de Jonge, Lakhansing Pardeshi, David Meijer, Margi Hartanto, Roven Rommel Fuentes, Sina Majidian, Mohammad Alanjary, Justin van der Hooft, Marnix Medema, Anne Kupczok and Sandra Smit.

Thanks to the reading committee for taking the time to go through my thesis, I look forward to having a discussion with you on the day of the defence.

The research described in this thesis was financially supported by the sandwich PhD programme of Wageningen University & Research (WUR). I would like to thank and acknowledge the support of prof. Mansoor (former Director NIBGE) for allowing me to carry out my research work during the time I was working at NIBGE.

Last, but not the least, remembering my (late) parents, I would like to thank them for their love and support throughout my life. Without them, this day would not have been possible. I would also like to thank my entire family for their warm support, and special thanks to my wife, who took the complete responsibility of the household to spare me for the study.