NTNU
Kunnskap for en bedre verden

Department of Mathematical Sciences

TMA4500 - Industrial Mathematics, Specialization Project

Genomic Prediction in Wild Populations: Exploring the Effect of Model Complexity in Bayesian Principal Component Ridge Regression on Binary Predictions

Author:
Eivor Øsebak-Loe

Supervisor:
Stefanie Muff

December 2025

Abstract

Genomic prediction is a central tool in quantitative genetics for understanding and predicting genetic contributions to traits in natural populations. While many established methods are developed for continuous traits, several biologically important traits, such as dispersal, survival, and recruitment, are recorded as binary outcomes, which complicates prediction and model evaluation. In addition, genomic prediction models require a careful balance between capturing genetic signal and avoiding overfitting. The aim of this project is to investigate how model complexity influences predictive performance in Bayesian Principal Component Ridge Regression (BPCRR) for binary traits.

This project is based on simulated binary phenotypes generated from real SNP data from a wild house sparrow population. Phenotypes were simulated within a generalized linear mixed model framework with varying heritability levels. BPCRR models were fitted using an increasing number of principal components derived from the genotype matrix. Predictive performance was evaluated using three complementary measures: discrimination accuracy on the observed binary scale, measured by the area under the ROC curve (AUC), correlation between true and predicted breeding values on the liability scale, and recovery of additive genetic variance.

The results show that classification performance on the observed scale is largely insensitive to model complexity and is primarily driven by trait heritability. In contrast, prediction accuracy on the liability scale exhibits an optimum at an intermediate number of principal components, with the optimal complexity increasing as heritability increases. Recovery of additive genetic variance improves monotonically with increasing model complexity. Together, these findings demonstrate that the optimal level of model complexity depends on the chosen performance metric, and that improvements in genetic prediction on the liability scale do not necessarily translate into improved classification performance on the observed binary scale.

Contents

List of Figures

List of Tables

1 Introduction

Understanding why individuals differ in their traits is a central goal in evolutionary biology. Traits such as dispersal, number of surviving offspring and recruitment are essential to understand how populations are influenced by natural selection and environmental changes (Lynch and Walsh, 1998). Variation among individuals is caused by a combination of genetic and environmental factors, and *quantitative genetics* provides a framework for describing how these factors influence traits (Falconer and Mackay, 1989). Understanding the genetic factor is necessary to gain insight into a population's long-term chance of survival, and is a central goal of quantitative genetics (Wilson et al., 2010). Classical linear models have been used to model and predict both environmental and genetic effects in both wild and human-managed breeding systems. In managed breeding systems, such as livestock and crop populations, there is already a good understanding of functional relations between genes and expressed traits (Meuwissen et al., 2001; VanRaden, 2008). Generalizing these results to wild populations is not trivial. Wild populations are characterized by fluctuating environments, incomplete and unbalanced data, limited control of mating, life history, and the environment, and often smaller population sizes (Wilson et al., 2010). As a consequence, it is more difficult to determine the genetic contribution to a trait in wild populations, and studying the functional relationships between genes and a trait of interest requires methods that can extract genetic information and remain robust to environmental complexity.

Many methods in quantitative genetics are developed for continuous traits, such as body mass and height. However, many biologically important traits in wild populations are recorded as binary outcomes, which introduces additional challenges when analyzing, compared to continuous traits. A binary observation is recorded as zero or one, and provides little information about the underlying genetic effects. It is therefore common to assume that binary outcomes reflect an underlying continuous process and that the genetic effects influence the binary trait through this process. Prediction is typically carried out using binary regression models, most often formulated as generalized linear mixed models (GLMMs) with a logit or probit link (e.g., Steinsland et al., 2014). The presence of an underlying process introduces an additional layer for assessing accuracy. It is possible to assess the classification prediction into zeros or ones, or the accuracy of the predicted genetic signal on the underlying scale. Concerning binary predictions, these accuracies do not necessarily coincide and may lead to different conclusions about performance. As a consequence, choosing an optimal model is a non-trivial task.

When using statistical models, it is necessary to balance the amount of genetic information included against the risk of overfitting. Both *principal component analysis* (PCA) and different regularization methods are common to control model complexity, and for continuous traits these effects are well documented (Ashraf et al., 2022; Aspheim et al., 2024; Gianola, 2013; Meuwissen et al., 2001). Recent work has shown that Bayesian principal component ridge regression (BPCRR) provides a stable and flexible framework for predictions of continuous traits (Aspheim et al., 2024). This method also allows explicit control over model complexity through the number of included principal components, making BPCRR a natural framework for investigating how this complexity affects predictive performance. While extending BPCRR to binary traits is conceptually straightforward, the presence of the unobserved liability scale introduces additional challenges. In particular, it remains unclear how model complexity influences predictions on the binary scale.

The aim of this project is to quantify how the number of included principal components in BPCRR influences predictive performance for binary traits. The project is motivated by the analysis of real genomic data. However, this project thesis focuses on simulated phenotypes generated using real SNP data, which allows key genetic properties to be controlled and systematically varied. We focus on three complementary aspects of performance: discrimination accuracy on the observed scale, reconstruction of the inherited signal on the liability scale, and recovery of additive genetic variance. By varying both heritability and model complexity, this project provides insight into how different definitions of predictive accuracy interact with model complexity in a binary-trait context. All code necessary for reproducing the results is available in the GitHub repository eivorloe/Prosjektoppgave.

The data for this project are provided by the Department of Biology at NTNU and the Gjærevoll Centre and include phenotypic traits, environmental data, and SNPs. These data were collected from a population of wild house sparrows monitored since 1993, inhabiting an island system in northern Norway (Ranke et al., 2021). In this project, we consider binary trait dispersal.

The project is organized as follows. After introducing relevant background on biological and statistical concepts, we present the methodological framework for extending Bayesian principal component ridge regression to binary outcomes. The results section presents findings of the effects of model complexity on predictive performance across different scales. Finally, we discuss the implications of our findings and highlight directions for future research.

2 Background

2.1 Fundamental Concepts of Quantitative Genetics

2.1.1 Basic Genetic Concepts and Terminology

To understand how genetic information contributes to trait variation, we first introduce the fundamental biological terms used throughout quantitative genetics. All living organisms store hereditary information in deoxyribonucleic acid (DNA), a long molecule arranged into units called *chromosomes*. A *gene* is a segment of the DNA that influences a particular characteristic of an organism (Griffiths, 2015). Each gene occupies a fixed position on a chromosome, known as its *locus*. Genes can occur in different versions, called *alleles*, and alleles of the same gene may lead to different trait expressions. The proportion of individuals in a population carrying an allele is referred to as the *allele frequency*. In organisms that inherit one set of chromosomes from each parent, individuals carry two alleles at every locus. The combination of alleles an individual carries is its *genotype*.

Many traits of organisms vary among individuals, and the observable expression of such a trait is called an individual's *phenotype*. For example, traits such as body mass, eye color, or survival can be observed, and then 1000 grams, blue, and yes are possible phenotypes. Most traits studied in evolutionary biology are *quantitative traits*. In the classical sense, these traits vary continuously within a population rather than forming discrete categories such as "tall/short". Quantitative traits are also *polygenic*, meaning several genes contribute to the realization of the phenotype. The term *genetic architecture* is commonly used to describe this underlying genetic basis, including how many genes contribute to a trait and how their effects are distributed. Most traits are also influenced by environmental factors, such as temperature, nutrition, and parental care, as well as by interactions between the environment and genes. (Falconer and Mackay, 1989).

2.1.2 Phenotypic Variation and Heritability

Phenotypic variation among individuals is a defining feature of populations and provides the basis for evolutionary change and quantitative genetics. To understand how genes and environment contribute to the variation, a simple decomposition is proposed. In the simplest case, the phenotype P of an individual is written as the sum of a genetic component G and an environmental component E,

$$P = G + E\,.$$

Assuming that genetic and environmental components are independent, the total phenotypic variance can be expressed as,

$$V_P = V_G + V_E\,,$$

where V_G is the genetic variance and V_E is the environmental variance (Falconer and Mackay, 1989). The genetic component G can be further decomposed into biologically meaningful parts,

$$G = A + D + I\,,$$

where A represents the *additive genetic effect*, D is the *dominance deviation* (interactions between alleles at the same locus), and I is the *epistatic deviation* (interactions between alleles at different loci) (Lynch and Walsh, 1998). These components sum to the total genetic variance,

$$V_G = V_A + V_D + V_I \,.$$

The additive genetic effect A of an individual is of particular importance because it represents the sum of the average effects of alleles. The value of A, expressed relative to the population mean phenotype, is called an individual's *breeding value* (Wilson et al., 2010). This value describes the expected genetic contribution an individual passes to its offspring, and is the only genetic effect that contributes predictably to evolutionary change in a population. The proportion of phenotypic variance due to additive genetic effects is the *narrow-sense heritability*,

$$h^2 = \frac{V_A}{V_P} \,.$$

This quantity reflects the proportion of variation between individuals in a population that is due to the additive effect of their genotypes, and it is central in both evolutionary biology and applied breeding (Visscher et al., 2012).

2.1.3 Single Nucleotide Polymorphism

To quantify genetic contributions to phenotypic variation, information about genetic differences among individuals is required. With advances in genotyping technology, it is now possible to trace genetic effects directly at the DNA level through *single nucleotide polymorphisms* (SNPs). A SNP is a locus on the DNA where individuals in a population differ by a single nucleotide base (Brookes, 1999). We classify a locus as a SNP if the *minor allele* frequency is at least 1% (Karki et al., 2015). The minor allele is the least frequent allele at a given locus (Conner and Hartl, 2004; Brookes, 1999). In genomic analysis, the combination of alleles at a SNP locus is usually coded as $\{0, 1, 2\}$, representing the number of minor alleles in the genotype.

Most SNPs do not directly alter phenotypes, as they often occur in non-coding regions of the DNA (Visscher et al., 2012). However, they serve as effective genetic markers due to *linkage disequilibrium* (LD), the non-random association of alleles at different loci (Bush and Moore, 2012). Alleles that are physically close on the chromosome tend to be inherited together, and a SNP in strong LD with a causal locus can therefore act as an indirect indicator of its effect (Ashraf et al., 2022). These properties make SNPs highly useful for studying polygenic traits (Syvänen, 2001).

2.2 Animal model and Genomic Prediction for Continuous Traits

2.2.1 From Pedigree-Based to SNP-Based Relatedness

To analyze the biological concepts, quantitative genetics relies on statistical models that describe the genetic and environmental sources of phenotypic variation. The animal model has long been the cornerstone of quantitative genetics. It is a linear mixed model (LMM) that partitions the phenotype of each individual into fixed environmental effects, additive

genetic effects, and residual noise. In its simplest form, for a single phenotype y_i for a single individual i, the animal model can be expressed as

$$y_i = \mu + a_i + \varepsilon_i, \quad i = 1, \ldots, N \ .$$

Here μ is the population mean for the trait, a_i is the additive genetic effect also known as the breeding value, ε_i is an independent and identically distributed residual term, and N is the number of individuals. In LMMs, the residual terms are often assumed to follow a Gaussian distribution with zero mean and unknown variance σ^2. The breeding values $\mathbf{a} = (a_1, a_2, \ldots, a_N)^\top$ are assumed to originate from a multivariate Gaussian distribution, $\mathbf{a} \sim N(\mathbf{0}, V_A\mathbf{G})$, where V_A is the additive genetic variance and $\mathbf{G}$ represents the relatedness between individuals. Several formulations of $\mathbf{G}$ are possible, each defining a different representation of genetic similarity between individuals.

Traditionally, relatedness has been estimated from knowledge of pedigrees. The pedigree-based $\mathbf{G}$ encodes the expected proportion of alleles shared by descent between every pair of individuals in a population. For instance, individuals i and l that are full siblings are estimated to have $\mathbf{G}_{il} = 0.5$, meaning we assume they share half their alleles. However, the pedigree-based method has some limitations. In practice, especially for wild populations, pedigrees are often incomplete or uncertain (Ashraf et al., 2022). Moreover, pedigree-based expected relatedness does not necessarily reflect the actual realized genetic similarity because of random segregation and recombination. These limitations reduce the precision of the estimated additive genetic variance and breeding values.

The introduction of the *genomic relationship matrix* (GRM) was therefore a major advance in quantitative genetics. The expected pedigree-based relatedness estimates are replaced with realized similarity from genetic markers such as SNPs. In this case, $\mathbf{G}_{il} = 0.5$ indicates that individual i and l actually share half of their alleles. A commonly used estimator, proposed by VanRaden (2008), defines the GRM as

$$\mathbf{G} = \frac{\mathbf{Z}\mathbf{Z}^\top}{2\sum_j \rho_j(1-\rho_j)} \ ,$$

where $\mathbf{Z}$ contains the centered genotype markers, and ρ_j is the allele frequency at marker j. The numerator ensures that the GRM behaves similarly to the pedigree matrix (VanRaden, 2008), and using the GRM instead of the pedigree is known as the *genomic animal model*.

The genomic animal model improves the accuracy compared to the traditional animal model, but it also introduces some computational challenges. The matrix scales with the number of individuals, and the cost of inversion grows with its square (Aspheim et al., 2024). These computational demands become limiting for high-dimensional genomic datasets. Moreover, the GRM is derived from the same SNP matrix used in the analysis, such that genetic information enters the model through both the relationship matrix and the marker data.

2.2.2 Marker-Based Regression

A popular alternative to the animal model is the *marker-based regression model*, where the genomic contribution to the phenotype is explicitly represented by the additive effects of individual SNPs. This formulation also serves as the basis for *genomic prediction*, where estimated marker effects are used to predict breeding values for individuals. The

underlying idea is that many traits, such as body mass, wing length, and the binary dispersal trait, are highly polygenic traits (Griffiths, 2015; Steinsland et al., 2014). For a continuous trait measured on N individuals and m SNPs, the marker-based regression model can be written as

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{W}\mathbf{d} + \boldsymbol{\varepsilon} \ , \tag{1}$$

where $\boldsymbol{\mu}$ is the vector for the overall population mean, $\mathbf{X}$ is the matrix of fixed effects like age or sex with regression parameter $\boldsymbol{\beta}$, $\mathbf{Z}$ is the mean-centered SNP matrix with one column per marker and corresponding effect sizes $\mathbf{u}$, $\mathbf{W}$ is the matrix of random effects like island or family with effect sizes $\mathbf{d}$, and $\boldsymbol{\varepsilon}$ is the residual term. The breeding value of individual i is then

$$a_i = \sum_{j=1}^{m} \mathbf{Z}_{ij} u_j \ .$$

In many cases $m \gg N$, making models ill-posed such that the design matrix $\mathbf{Z}$ contains linearly dependent columns. In this case, it is not possible to estimate the effects $\mathbf{u}$ uniquely, as there are more unknown parameters than equations to solve. In practice, this results in overfitting, as the model can fit noise in the training data perfectly and fail to generalize. To obtain meaningful and stable predictions, it is therefore necessary to reduce the number of predictors or constrain their effects through regularization or dimension reduction.

2.2.3 Principal Components for Dimension Reduction

A common strategy to address ill-posedness is to apply some form of dimension reduction on the SNP matrix before fitting the model. Principal component analysis (PCA) provides an intuitive way to summarize genotypic information, by replacing the original SNPs with a smaller set of *principal components* (PCs). Each PC is a weighted sum of the original SNPs, and PCA chooses these weights so that the first component captures the largest overall pattern of variation in the SNP matrix across individuals. Subsequent components capture as much of the remaining variation as possible while being constrained to be orthogonal to the previous ones. The resulting PCs are therefore uncorrelated, and projecting the data onto the first few PCs often preserves most of the variation present in the original data (Abdi and Williams, 2010).

In practice, PCs are obtained through singular value decomposition (SVD) of the SNP-matrix $\mathbf{Z}$ (Aspheim et al., 2024). The SVD of $\mathbf{Z}$ is given by

$$\mathbf{Z} = \mathbf{U}\Lambda\mathbf{V}^\top \ ,$$

where $\mathbf{U}$ has dimensions $m \times m$, Λ is the diagonal $m \times N$ matrix with singular values on the diagonal, and $\mathbf{V}$ is the $N \times N$ matrix of eigenvectors. The singular values in Λ are ordered such that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N$, and each λ_j^2 is proportional to the variance explained by the corresponding PC. For additional details see Abdi and Williams (2010).

The first K PCs are then obtained by projecting the standardized SNPs onto the first K eigenvectors of $\mathbf{Z}^\top\mathbf{Z}$:

$$\mathbf{Z}^\star = \mathbf{Z}\mathbf{V}_{1:K} \ ,$$

where $\mathbf{Z}^\star$ is the projection of the SNP matrix on the K-dimensional PC-space. $\mathbf{V}_{1:K}$ is the first K columns of $\mathbf{V}$. Since the leading PCs capture the largest sources of variation in the

SNP matrix, a relatively small number of components often explains a substantial proportion of the total genotypic variation. As a result, PCA condenses thousands of correlated genetic markers into a smaller set of orthogonal predictors, reducing dimensionality while preserving the core genetic signal needed for modeling.

2.2.4 Shrinkage-Based Bayesian Approaches

Polygenic traits are expressed through small contributions from multiple markers. Applying regularization to the marker effects helps stabilize their estimates and prevents overfitting. A common regularization method is to induce shrinkage with Bayesian methods. Intuitively, shrinkage can be described as weighting prior knowledge against information added by the data. If a model-estimated marker effect is small and uncertain, we trust the prior knowledge more and shrink the effect to zero. On the other hand, effects with estimates further from zero are kept in the model, as we trust the data more. Shrinkage also reduces the variance of the estimated effects by forcing the estimates closer to zero. This reinforces patterns consistent with the data. In other words, the goal is to limit the influence of small and uncertain effects by penalizing model complexity. This results in more reliable predictions and clearer interpretation, especially when the number of markers exceeds the number of individuals.

The simplest Bayesian shrinkage method is ridge regression. This is the Bayesian equivalent of the genomic animal model, and in this case all effects are penalized uniformly (Gianola, 2013). The penalization is applied through the prior assumption on the marker effects, and for ridge regression, we assume a normal prior with identical variance for the effect sizes,

$$u_j|\sigma_u^2 \sim N(0, \sigma_u^2) \ .$$

The variance parameter σ_u^2 is either assumed known or assigned a weakly informative hyper-prior, for example an inverse-gamma distribution (Aspheim et al., 2024). Note that assigning a common prior variance to all marker effects implies independence between the effects. Different prior choices for the marker effects lead to different forms of shrinkage. This flexibility forms the basis of the Bayesian alphabet (Gianola, 2013).

The first model introduced to the alphabet was BayesA (Meuwissen et al., 2001). The normal distribution assumption is retained for the marker effects, but with a variance specific to each marker. Each variance follows a scaled inverse-chi-square prior, which produces heavy tails on the marker effect prior. These heavy tails translate to heavy shrinkage on most effects, while a small number of markers avoid this shrinkage. BayesB modifies the idea from BayesA by allowing different distributions on the marker effects. Each marker has a high prior probability of having exactly zero effect, and a small probability of following a heavy-tailed distribution (Meuwissen et al., 2001). In this scenario, both shrinkage and variable selection are used to exclude SNPs from the model. BayesR generalizes these ideas by allowing a mixture of normal distributions with different variances (Gianola, 2013). In practice, this mixture reflects the assumption that most markers have no effect, while a few markers contribute at different scales, an assumption consistent with findings from Aspheim et al. (2024). BayesR has repeatedly demonstrated competitive predictive accuracy and robustness to the underlying genetic architecture (e.g., Ashraf et al., 2022; Aspheim et al., 2024).

2.2.5 Bayesian Principal Component Ridge Regression (BPCRR)

All members of the Bayesian alphabet model shrinkage directly on SNP effects. An alternative strategy is to combine shrinkage with dimension reduction, which is the idea behind Bayesian principal component ridge regression (BPCRR). In BPCRR, principal components are used as predictors, and Bayesian priors are specified to induce ridge-type shrinkage of the corresponding marker effects. This formulation is motivated by the fact that SNP data rarely satisfy the independence assumption underlying ridge priors due to linkage disequilibrium and population structure (e.g., Aspheim et al., 2024). However, using principal components instead of raw markers resolves this issue, as PCs are uncorrelated by construction while still capturing most of the genetic variation in the data. Applying ridge shrinkage to the PC effects can further stabilize them and prevent overfitting.

The BPCRR model is constructed by first transforming the genotype matrix using principal component analysis. Let $\mathbf{Z}$ denote the mean-centered $N \times m$ SNP matrix with singular value decomposition $\mathbf{Z} = \mathbf{U}\Lambda\mathbf{V}^\top$. The projected marker matrix is then

$$\mathbf{Z}^\star = \mathbf{Z}\mathbf{V}_K \ ,$$

where $\mathbf{V}_K$ is the first K eigenvectors associated with the largest singular values.

Bayesian ridge regression is applied to the PC effects $\mathbf{u}^\star$. Each effect is assigned a normal prior

$$u_j^\star | \sigma^2_{\mathbf{u}^\star} \sim N(0, \sigma^2_{\mathbf{u}^\star}) \ ,$$

which induces shrinkage. This combination of shrinkage and PCs yields a more robust, less prior-sensitive shrinkage than SNP-based ridge models (Aspheim et al., 2024).

A practical advantage of BPCRR, compared to other Bayesian models, is that BPCRR can be fitted within a single linear mixed modeling (LMM) framework using Integrated Nested Laplace Approximation (INLA; Martino and Riebler, 2019), as it allows all model components (fixed, random, and genomic effects) to be estimated simultaneously. In contrast, most implementations of BayesB or BayesR require the use of a two-step method, where fixed and random effects are removed before SNP effects are estimated (Aspheim et al., 2024). Finally, BPCRR has robust predictive performance for different choices of the number of principal components K. Aspheim et al. (2024) shows that the prediction accuracy is high over broad intervals for K, mostly as a consequence of the shrinkage on the PC effects.

2.3 Quantitative Genetics for Binary Traits

2.3.1 Binary Generalized Linear Mixed Model

For binary traits, the response variable represents a probability rather than a continuous outcome. Modeling this probability requires a generalized linear mixed model, which connects a continuous linear predictor to the observed response through a link function. In the binary setting the response variable y_i takes values 0 or 1, and is assumed to follow a Bernoulli distribution with probability of success $p_i = P(y_i = 1)$:

$$y_i \sim \text{Bernoulli}(p_i) \ .$$

The aim is to model how the success probability p_i depends on a set of predictors. This relationship is specified through a linear predictor η_i, which may consist of fixed effects, random effects, or additional genomic components, or a combination. Let $\mathbf{x_i}$ be the vector of fixed effects with effect sizes $\boldsymbol{\beta}$, $\mathbf{w_i}$ be the vector of random effects with effect sizes $\mathbf{d}$, and let $\mathbf{z_i}$ be the genetic effects with effect sizes $\mathbf{u}$, all for individual i. The linear predictor can then be described as,

$$\eta_i = \mathbf{x_i}^\top \boldsymbol{\beta} + \mathbf{z_i}^\top \mathbf{u} + \mathbf{w_i}^\top \mathbf{d} \ .$$

To ensure that η_i maps to a probability p_i, a link function $h(\cdot)$ is specified as part of the model. For the logistic model, the most common link is the logit link,

$$\eta_i = h(p_i) = \log\left(\frac{p_i}{1-p_i}\right) \ .$$

The corresponding inverse link function returns the predicted probability

$$p_i = h^{-1}(\eta_i) = \frac{e^{\eta_i}}{1+e^{\eta_i}} \ . \tag{2}$$

To obtain a binary prediction of y_i, the estimated probability p_i is converted into a class label using a fixed cutoff, typically set to 0.5. The values above are classified as successes, and those below as failures.

$$\hat{y}_i = \begin{cases} 1, & \text{if } p_i \geq 0.5, \\ 0, & \text{if } p_i < 0.5. \end{cases}$$

2.3.2 Variance Components and Heritability on Different Scales

For binary traits, the observed 0/1 outcome does not provide a meaningful scale for describing sources of variation. To enable a quantitative genetic decomposition similar to that used for continuous traits, it is therefore customary to introduce an unobserved continuous latent variable, referred to as the *liability*. On the latent scale, we assume that there exists a fixed threshold that separates the two observed outcomes. Individuals with a linear predictor below this threshold are observed as $y = 0$, whereas individuals with a linear predictor exceeding this threshold are observed as $y = 1$. This representation is commonly referred to as a threshold model. The position of the threshold is determined by the trait's prevalence, which corresponds to the proportion of individuals whose latent liability exceeds the threshold. Low-prevalence traits correspond to thresholds located in the upper tail of the liability distribution, whereas high prevalence implies a lower threshold on the liability scale.

Formulating the model on the latent scale allows decomposition of variation in the same manner as for continuous traits. On the latent scale, genetic and environmental effects act additively, even though their effects on the observed binary outcome are inherently non-linear (Acker and Reid, 2025). In addition to these components, binary models include a third variance term, V_{link}, implied by the link function. This is not a biological variance, but follows from the model structure. For the link function introduced here, $V_{link} = \pi^2/3$ (Steinsland et al., 2014). As a consequence, the latent variance has one constant component that contributes to the total trait variance. Given the variance components on the liability scale, heritability h^2 is defined as the proportion of total liability variation that is due to

additive genetic effects. For binary traits, this is referred to as the latent-scale heritability and is expressed as

$$h^2 = \frac{V_A}{V_A + V_E + V_{link}} \ .$$

This formulation shows that additive genetic variance must be considered alongside both environmental variation and the variance implied by the link function to obtain the total variation on the liability scale.

2.4 Evaluation Metrics for Binary Regression Models

2.4.1 Area Under the ROC Curve (AUC)

Evaluating predictive models requires a clear definition of what is meant by accuracy. While this is relatively straightforward for continuous traits, binary outcomes introduce additional challenges, as different aspects of model performance may be of interest. In particular, prediction can be assessed either by how well a model discriminates between observed 0/1 outcomes, or by how accurately it captures the underlying genetic signal.

We will first consider the discrimination accuracy on the observed scale using the *Area Under the ROC curve* (AUC). The *Receiver Operating Characteristic* (ROC) curve shows the true positive rate against the false positive rate as the classification cutoff on the predicted probabilities ranges from 0 to 1. Therefore, the ROC curve does not depend on a single classification cutoff but rather on how the model ranks individuals based on their predicted probabilities. The AUC summarises the model's ROC curve into a single number by integrating the true positive rate over the false positive rate across all possible classification thresholds. In quantitative genetics the AUC reflects how well the model discriminates individuals with the two observed phenotypes. Importantly, for binary traits, the achievable AUC is constrained by both trait heritability and prevalence. Even with a perfect genetic predictor, these factors impose an upper bound on discrimination performance (Wray et al., 2019). An AUC of 1.0 corresponds to perfect discrimination between the groups, whereas an AUC of 0.5 reflects random guessing. When comparing the predictive performance of different models using AUC, we define the optimal model as the one that maximises AUC.

2.4.2 Accuracy of Predicted Breeding Values

AUC provides a useful measure of discrimination on the observed 0/1 scale. At the same time, predictions rely on the accurate estimation of underlying genetic effects. It is therefore informative to assess the quality of the genetic information on which the model's predictions are based, since good observed-scale discrimination can arise even when the genetic signal is only weakly captured. To assess this aspect of predictive performance, we evaluate the correlation between true and predicted breeding values. This metric complements the AUC, as it evaluates the genetic information contained in the predictions rather than their ability to classify individuals on the observed 0/1 scale. With respect to this correlation, we define the optimal model as the model which maximises the correlation.

The correlation between true and predicted breeding values is subject to a fundamental theoretical constraint. In an additive genetic model, the maximum attainable correlation

between predicted and true breeding values is given by $\sqrt{h^2}$, where h^2 denotes the heritability (Lynch and Walsh, 1998). Although this result is typically stated for continuous traits, the same argument applies to binary outcomes when formulated on the underlying liability scale. Intuitively, this bound reflects that breeding values represent only one component of liability, with the remaining variation arising from environmental and non-additive genetic effects. In practice, attainable accuracy is further constrained by sample size and the genetic architecture. Daetwyler et al. (2008) derive an expression for the expected correlation between true and predicted breeding values with a liability model for binary traits. Let a and $\hat{a}$ be the true and predicted breeding value on the liability scale for any individual. The correlation $r_{a\hat{a}} = corr(a\ , \hat{a})$ is bounded by

$$r_{a\hat{a}} = \sqrt{\frac{\lambda h_o^2}{\lambda h_o^2 + 1}}\ .$$

Here, h_o^2 denotes the heritability on the observed 0/1 scale, and λ represents the amount of training data relative to the genetic complexity of the trait (Daetwyler et al., 2008).

2.4.3 Additive Genetic Variance Estimation

In addition to evaluating individual-level prediction performance, it is also informative to consider how well genomic prediction models recover population-level genetic properties. Additive genetic variance recovery describes how well a model reproduces the amount of additive genetic variation on the latent liability scale and is an important quantity in quantitative genetics. We can compare the true additive genetic variance V_A, with the corresponding model estimated $\hat{V}_A$, by looking at the proportion

$$\frac{\hat{V}_A}{V_A}\ .$$

This ratio quantifies the proportion of variance recovered by the model. A value of one indicates that the model reproduces the true level of additive genetic variance, while values below or above one indicate under- or overestimation, respectively. Although the primary goal of this project is predictive performance, assessing heritability recovery provides a complementary perspective. The 0/1 outcome provides limited information about the underlying genetic architecture. Understanding how model complexity influences access to this underlying architecture provides additional insight into an important aspect of model performance. Together, these evaluation metrics form the basis for the methodological comparisons presented in the next section.

3 Methods

3.1 Data Description

The data motivating this project come from a meta-population of Norwegian house sparrows on an island system off the coast of Helgeland. The meta-population spans 18 islands. These islands are divided into two main habitats: the *inner islands*, closer to the mainland, and the *outer islands*, further from the mainland. Some islands are more isolated, but there is genetic flow (dispersal) between all of them (Ranke et al., 2021), as shown in Figure 1.

The data include a binary dispersal indicator (0/1), 65 244 SNP markers per individual, and additional environmental features. The environmental features are displayed in Table 1. The dataset consists of yearly measurements collected from 1998 to 2019, and contains 4722 unique individuals, of which 1106 dispersed. To limit the scope of this project, the observed phenotypes are not analyzed. The SNP matrix is instead used as the genotypic basis for the simulated traits, to ensure biologically reasonable conditions.

Table 1: Environmental features included in the dataset.

Feature	**Description**
Ringnr	Unique individual identifier
HatchYear	Year the individual hatched
Natal Island	Island where the individual hatched
Adult Island	Island of recruitment
Sex	Sex of the individual
Recruit	Survival to recruitment

3.2 Simulation of Phenotypes

To assess the effect of model complexity on predictive performance, we conducted a simulation study based on SNP data from the Helgeland system. Binary phenotypes were generated within a GLMM framework, with additive genetic effects represented through the marker-based regression model defined in equation (1).

Let $\mathbf{Z}$ denote the centered SNP matrix, and let $\mathbf{u} = (u_1, u_2, \ldots, u_m)^\top$ be the vector of marker effects, where m is the number of SNPs. The vector of breeding values for all individuals is denoted by $\mathbf{a} = (a_1, a_2, \ldots, a_N)^\top$, where N is the number of individuals. To limit the scope of this project, no fixed or additional random effects were included in the model. Under these specifications, the linear predictor in the GLMM coincides with the vector of breeding values, and is defined as

$$\mathbf{a} = \mathbf{Z}\mathbf{u} \, .$$

Each element in $\mathbf{a}$ is given by $a_i = \mathbf{z}_i^\top \mathbf{u}$, and $\mathbf{z}_i^\top$ denotes the centered SNP vector of individual i, which represents the covariates of the model. The marker effects $\mathbf{u}$ were simulated using a BayesR mixture distribution with two components: one component with zero effect and one component with effect sizes from a normal distribution $u_j \sim N(0, \sigma^2)$.

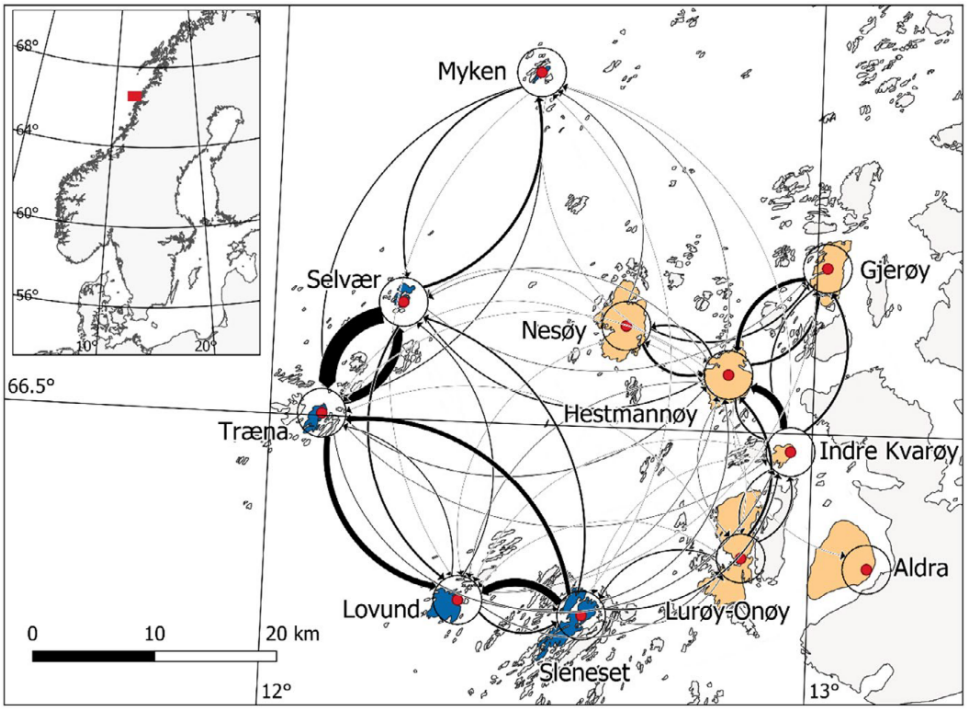

Figure 1: Map of the study area. Inner islands are in yellow, and outer islands are in blue. The arrows display emigration patterns in the period $1993 - 2014$. Increased thickness of the arrows indicates a larger volume of emigration (Ranke et al., 2021).

The proportion of SNPs with zero effect is π_1, and consequently, the distribution of the effects u_j is

$$u_j \sim \pi_1 \cdot 0 + (1 - \pi_1) \cdot N(0, \sigma^2) \quad j = 1, 2, \ldots, m \ ,$$

which yields a sparse yet approximately polygenic genetic architecture.

The vector of breeding values $\mathbf{a}$ was centered to have mean zero and scaled to achieve a target additive genetic variance $\text{Var}(\mathbf{a}) = V_A$. Under a logit link, the residual variance on the liability scale is $V_{link} = \pi^2/3$, and the scaling was done with respect to some desired heritability h^2,

$$V_A = \frac{h^2 \cdot \pi^2/3}{1 - h^2} \ .$$

Centering implies a zero intercept, corresponding to an expected phenotype probability of 0.5. Lastly, phenotypes $\mathbf{y} = (y_1, y_2, \ldots, y_N)^\top$ were sampled from a Bernoulli distribution with probabilities $\mathbf{p} = (p_1, p_2, \ldots, p_N)^\top$ defined by the centered and scaled linear predictor, following the GLMM formulation described in Section 2.3.1.

A summary of the procedure is given in Algorithm 1. Marker effects were simulated from a sparse mixture distribution with a proportion $\pi_1 = 0.95$ of zero effects and variance $\sigma^2 = 10^{-4}$ for non-zero effects. The procedure returns simulated phenotypes together with the corresponding breeding values and marker effects.

Algorithm 1 Simulation of binary phenotypes

1: **Input:** SNP matrix $\mathbf{Z}$, heritability h^2, mixture parameters (π_1, σ^2)
2: Draw marker effects $u_j \sim \pi_1 \cdot 0 + (1-\pi_1)N(0, \sigma^2)$ for $j = 1, \ldots, m$
3: Compute raw breeding values $a_i = \sum_j Z_{ij} u_j$ for $i = 1, \ldots, N$
4: Center $\mathbf{a}$ to mean 0
5: Scale $\mathbf{a}$ such that the additive variance equals $V_A = \frac{h^2 \cdot \pi^2/3}{1-h^2}$
6: Compute individual probabilities $p_i = \text{logit}^{-1}(a_i)$ for $i = 1, \ldots, N$
7: Sample phenotypes $y_i \sim \text{Bernoulli}(p_i)$ for $i = 1, \ldots, N$
8: **Output:** $\{\mathbf{y}, \mathbf{a}, \mathbf{u}, V_A\}$

3.3 Prediction with Bayesian Principal Component Ridge Regression

Using the simulated phenotypes described above, we fitted a Bernoulli GLMM with the BPCRR as the linear predictor. The model specification and estimation procedure are described below.

The phenotypes were modeled using a Bernoulli likelihood with a logit link. For any given individual i, the linear predictor on the liability scale was

$$\text{logit}(p_i) = \eta_i = \alpha + \sum_{k=1}^{K} z_{ik}^{\star} u_k^{\star} \quad i = 1, 2, \ldots, N \ ,$$

where α is the intercept, $z_{ik}^{\star}$ is the k'th PC score for individual i, $u_k^{\star}$ is the corresponding effect, and K is the total number of PCs included in the model. The PCs were computed from the centered SNP matrix using singular value decomposition. All PC effects were assigned the common ridge prior,

$$u_k^{\star} \sim N(0, \sigma_{u^{\star}}^2) \quad k = 1, 2, \ldots, K \ ,$$

with a shared prior variance $\sigma_{u^{\star}}^2$. This prior variance was chosen to ensure that the total additive genetic variance on the liability scale equals the predefined value V_A from the simulation setup. To obtain a prior on the effects that induce the desired level of additive genetic variance, we scale the prior variance according to the empirical variances of the principal components,

$$\sigma_{u^{\star}}^2 = \frac{V_A}{\sum_{k=1}^{K} \text{Var}(\mathbf{z}_k^{\star})} \ ,$$

where $\text{Var}(\mathbf{z}_k^{\star})$ is the variance contribution from the k'th PC $\mathbf{z}_k^{\star}$. This prior is chosen to give ridge-type shrinkage on the PC effects (Aspheim et al., 2024). The intercept α was treated as a fixed effect using INLA's default prior, and posterior means of the PC effects $\hat{u}_k^{\star}$ were used for prediction. Estimated breeding values were obtained as

$$\hat{a}_i = \sum_{k=1}^{K} z_{ik}^{\star} \hat{u}_k^{\star} \quad i = 1, 2, \ldots, N \ ,$$

and predicted probabilities $\hat{p}_i$ were computed by applying the inverse-logit, defined in equation (2), to the linear predictor

$$\hat{p}_i = \frac{e^{\hat{a}_i}}{1 + e^{\hat{a}_i}} \quad i = 1, 2, \ldots, N \ .$$

These probabilities were used to assess predictive performance under different choices of the number of principal components.

3.4 Evaluation of Predictive Performance

Predictive performance was evaluated across simulations and the number of principal components using three measures. Specifically, we assessed the area under the ROC curve (AUC), correlation between true and predicted breeding values on the liability scale, and recovery of additive genetic variance.

Prediction performance on the observed binary scale was evaluated using the AUC. The predicted probability was compared to the simulated phenotype, and the AUC was computed using standard software functions. For comparison, a ceiling AUC was derived by using the true breeding values in the AUC calculations.

To evaluate how well the model recovered the underlying additive genetic signal, prediction accuracy was assessed on the liability scale by comparing true and estimated breeding values. Under the BPCRR model, estimated breeding values $\hat{a}_i$ were obtained from the posterior means of the PC effects $\hat{u}_k^\star$,

$$\hat{a}_i = \sum_{k=1}^{K} z_{ik}^\star \hat{u}_k^\star \ .$$

Accuracy on the liability scale was quantified as the correlation between the true and estimated breeding values,

$$\mathrm{corr}(a_i, \hat{a}_i) \ .$$

Here, a_i denotes the true additive genetic value of individual i generated in the simulation. For comparison, we also calculate the theoretical upper bound for this correlation from the true heritability used in the simulation setup.

Lastly, to evaluate the model's ability to recover the correct additive genetic variance, posterior samples of the individual breeding values were obtained from the BPCRR model. The estimated variance $\hat{V}_A$ is the mean of the posterior distribution of $\mathrm{Var}(\hat{a}_i)$, obtained by sampling from this distribution.

3.5 Experimental Setup

This section describes the experimental design used to evaluate predictive performance and variance recovery across heritability levels and numbers of principal components. Simulations were performed for three heritability scenarios, $h^2 \in \{0.1, 0.2, 0.33\}$, and for each simulated dataset, the BPCRR model was fitted using a grid of $K \in \{100, 200, \ldots, 1500\}$, corresponding to the number of principal components included as genomic predictors.

To quantify variation due to data-splitting and model stochasticity, the model fitting was repeated 20 times. The number of repetitions was chosen to balance estimation stability with computational cost. In each replicate, a new stratified 80/20 train-test split was generated, and the same split was used across all K values to ensure paired comparisons. All predictive performance evaluations were conducted on the test set.

The following implementation details describe how dimensionality reduction, model fitting, and evaluation were carried out in practice. The SNP matrix was centered and scaled prior to PCA to match the simulation design. PCA was computed once on the full SNP matrix, and the resulting PC scores were used consistently across all replicates and values of K. Phenotypes and PCs for test individuals were masked before fitting the BPCRR model, such that only training data contributed to parameter estimation. All models were fitted using Integrated Nested Laplace Approximation (INLA; version 25.06.07), a method for approximate Bayesian inference that applies to latent Gaussian models (Rue et al., 2009). The computational details of INLA are beyond the scope of this project. A summary of the key numerical settings is given in Table 2. Together, these choices define the experimental framework for evaluating predictive performance and variance recovery across heritability levels and numbers of principal components.

Table 2: Numerical settings used in the simulation study and prediction analyses.

Quantity	**Value**
Number of individuals	4 722
Number of SNPs	65 244
Simulated heritabilities	0.1, 0.2, 0.33
Mixture proportion π_1	0.95
Effect variance σ^2	10^{-4}
Train–test split	80/20
Number of replicates	$R = 20$
PC grid	$K \in \{100,\ 200, \ldots, 1500\}$
Number of samples to estimate $\hat{V}_A$	200

4 Results

4.1 Classification Accuracy on the Binary Scale

Overall, the AUC patterns differed between heritabilities. Higher heritability resulted in higher AUC values, but accuracy remained relatively stable across most values of K (Figure 2). Compared with the ceiling AUC, the model recovered approximately $45\% - 55\%$ of the maximum attainable accuracy in each scenario, based on the highest median AUC. In the scenario with $h^2 = 0.1$, the AUC values remained relatively stable across the full range of K, but with some visible increase from 100 to 400 principal components (Figure 2). At $h^2 = 0.2$, a modest increase in the AUC was observed from $K = 100$ to $K = 300$, after which the values declined slightly. For the highest heritability, $h^2 = 0.33$, the AUC values were consistently higher than in the other two settings, and the median changed little as K increased.

The number of principal components resulting in the highest AUC varied notably across replicates for all heritability levels (Table 3). For $h^2 = 0.1$, the optimal value of K was highly variable across replicates, with values ranging from $K = 100$ to $K = 1100$. At $h^2 = 0.2$, the results were more concentrated, and $K = 300$ was selected in 50% of replicates. For $h^2 = 0.33$, the pattern became more dispersed again, although $K = 200$ and 400 appeared most frequently, 30% and 20% respectively. No single value of K consistently maximizes AUC across replicates or heritability scenarios.

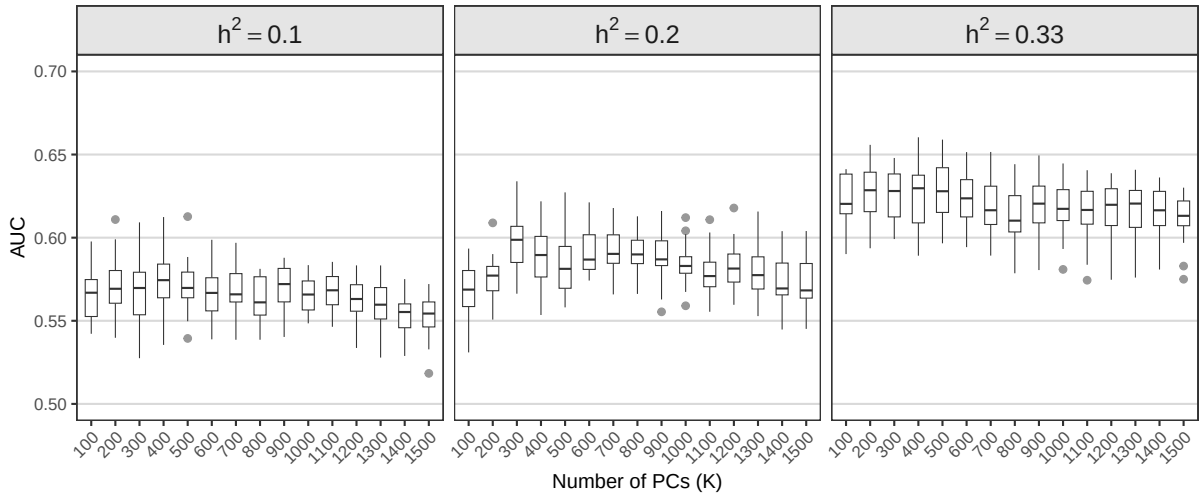

Figure 2: AUC as a function of the number of principal components (K) for three heritability levels. Boxplots show the distribution across 20 independent replicates, with the horizontal line indicating the median AUC. For reference, the ceiling AUC was 0.63 for $h^2 = 0.1$, 0.71 for $h^2 = 0.2$, and 0.78 for $h^2 = 0.33$.

4.2 Prediction Accuracy on the Liability Scale

Across heritabilities, the overall correlation between true and predicted breeding values increased with h^2, but the development of the correlation across the number of principal components K differed between scenarios (Figure 3). For all heritabilities, the correlation

Table 3: Replicate-level selection of the optimal number of principal components (K_{best}) based on AUC. For each heritability level, the table reports the proportion of replicates for which values of K achieved the highest AUC.

h^2	K_{best}	proportion
0.1	100	0.15
0.1	200	0.15
0.1	300	0.05
0.1	400	0.20
0.1	500	0.10
0.1	700	0.10
0.1	900	0.10
0.1	1100	0.15
0.2	200	0.05
0.2	300	0.50
0.2	500	0.05
0.2	600	0.15
0.2	700	0.15
0.2	800	0.05
0.2	1000	0.05
0.33	100	0.10
0.33	200	0.30
0.33	300	0.05
0.33	400	0.20
0.33	500	0.10
0.33	600	0.05
0.33	700	0.05
0.33	900	0.05
0.33	1200	0.05
0.33	1400	0.05

changed more noticeably with K than the AUC, and peaks were visible in each panel of Figure 3. The highest values were observed for $h^2 = 0.33$, followed by $h^2 = 0.2$ and $h^2 = 0.1$. For each heritability scenario, we compared the distribution of correlations at the K value yielding the highest median correlation to the corresponding theoretical upper bound. For low heritability ($h^2 = 0.1$), the entire distribution lay above the theoretical bound of 0.32. At intermediate heritability ($h^2 = 0.2$), the distribution was centered close to the expected bound, with observed values around 0.43 compared to a theoretical maximum of 0.45. In contrast, for high heritability ($h^2 = 0.33$), the entire distribution fell below the theoretical limit of 0.57. For $h^2 = 0.1$, the correlation increased from $K = 100$ to a maximum at $K = 300$, after which it declined steadily across the remaining values of K (Figure 3). At $h^2 = 0.2$, the correlation increased more markedly with increasing K, reaching its highest values around $K = 400$. Beyond this point, the correlation declined as for the lowest heritability. For $h^2 = 0.33$, the correlation values were consistently higher than in the other two scenarios, with a broader peak spanning approximately from $K = 200$ to $K = 700$. After this range, the correlation declined moderately.

The number of principal components yielding the highest correlation varied across replicates for all heritability levels. For $h^2 = 0.1$, the optimal values of K ranged from $K = 100$ to $K = 600$, with 40% replicates selecting $K = 100$ as the optimal choice. At $h^2 = 0.2$, the range of selected values was slightly wider, spanning from $K = 300$ to $K = 1100$, and $K = 400$ was optimal in half of the replicates. For $h^2 = 0.33$, the optimal K values fell between $K = 200$ and $K = 900$, with $K = 200$, 500, and 700 each selected in 25% of the replicates. These results show that the optimal number of components differed both within and between heritability scenarios (Table 4).

Replicate-level results for the AUC and correlation criteria are provided in Appendix A. Across all heritability levels, the optimal number of principal components identified by the AUC and correlation criteria coincided in 14 out of the 60 replicates. In the remaining replicates, the two criteria selected different values of K.

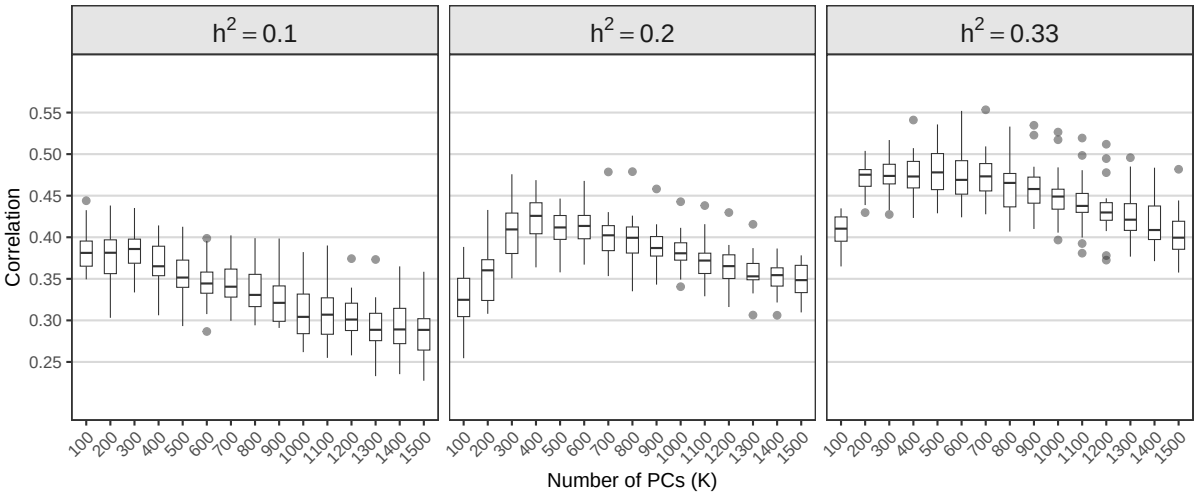

Figure 3: Correlation between true and estimated breeding values as a function of the number of principal components (K) for three heritability levels. Boxplots summarize 20 independent replicates, with the horizontal line indicating the median correlation. For reference, the theoretical attainable correlations were 0.31 for $h^2 = 0.1$, 0.45 for $h^2 = 0.2$, and 0.57 for $h^2 = 0.33$.

4.3 Recovery of Additive Genetic Variance

To assess how model complexity influences the recovery of genetic signal from the binary trait, we examined the proportion of the total additive genetic variance recovered as a function of the number of principal components. Across all heritability scenarios, this proportion increased as the number of principal components K grew, but the rate of this increase differed noticeably between scenarios (Figure 4). When heritability was low, a substantial share of the variance was captured by the first few components, whereas higher heritabilities showed a more gradual rise and did not reach the same level of recovery within the examined range of K. Overall, the curves were steepest for $h^2 = 0.1$ and most gradual for $h^2 = 0.33$. For $h^2 = 0.1$, the model recovered around 60% of the additive genetic variance with the first 100 PCs, and the proportion increased to approximately 90% by $K = 400$. Beyond this point, additional components contributed very little to the recovered variance (Figure 4). At $h^2 = 0.2$, the recovery at $K = 100$ was lower, at

Table 4: Replicate-level selection of the optimal number of principal components (K_{best}) based on correlation. For each heritability level, the table reports the proportion of replicates selecting a given value of K as optimal, based on the highest correlation between predicted and simulated breeding values.

h^2	K_{best}	proportion
0.1	100	0.40
0.1	200	0.20
0.1	300	0.20
0.1	400	0.10
0.1	600	0.10
0.2	300	0.10
0.2	400	0.50
0.2	500	0.05
0.2	600	0.25
0.2	800	0.05
0.2	1100	0.05
0.33	200	0.25
0.33	300	0.10
0.33	400	0.10
0.33	500	0.25
0.33	700	0.25
0.33	900	0.05

about 39%, and the 90% level was reached much later, at around $K = 900$. The overall maximum proportion recovered was slightly lower than in the lowest heritability scenario. For $h^2 = 0.33$, the initial recovery was approximately 31% at $K = 100$, and even with $K = 1500$, the 90% threshold was not reached. The curve increased steeply at low K but continued to rise gradually across the entire range without showing a plateau (Figure 4). In all scenarios, the variability between replicates decreased as K increased, and the estimates became more stable at higher K. With this metric, the optimal number of principal components is $K = 1500$, as the total variance is maximally reconstructed at this value.

When considering the reconstructed absolute value of the additive genetic variance, clear differences between heritability scenarios emerged (Figure 5). For a given number of principal components, higher heritability consistently resulted in greater variance recovery, with curves starting at higher levels and remaining above those for lower heritability across all values of K. As K increases, the recovered heritability values corresponding to different true heritability levels become increasingly separated. The increase between successive values of K is generally larger for higher true heritability levels. We also observe decreasing variability in the estimated heritability as K increases, and a similar leveling-off pattern as in Figure 4 is evident here.

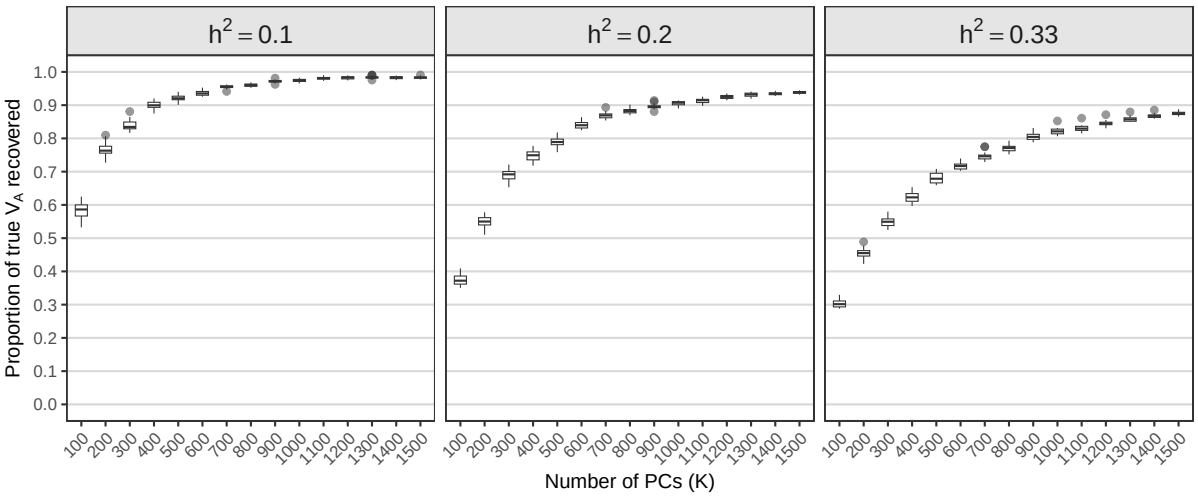

Figure 4: Recovered proportion of additive genetic variance as a function of the number of principal components (K) for three heritability levels. Boxplots show the proportion of the true simulated additive genetic variance (V_A) recovered by BPCRR across 20 independent replicates. Values close to 1 indicate near-complete recovery of the simulated V_A.

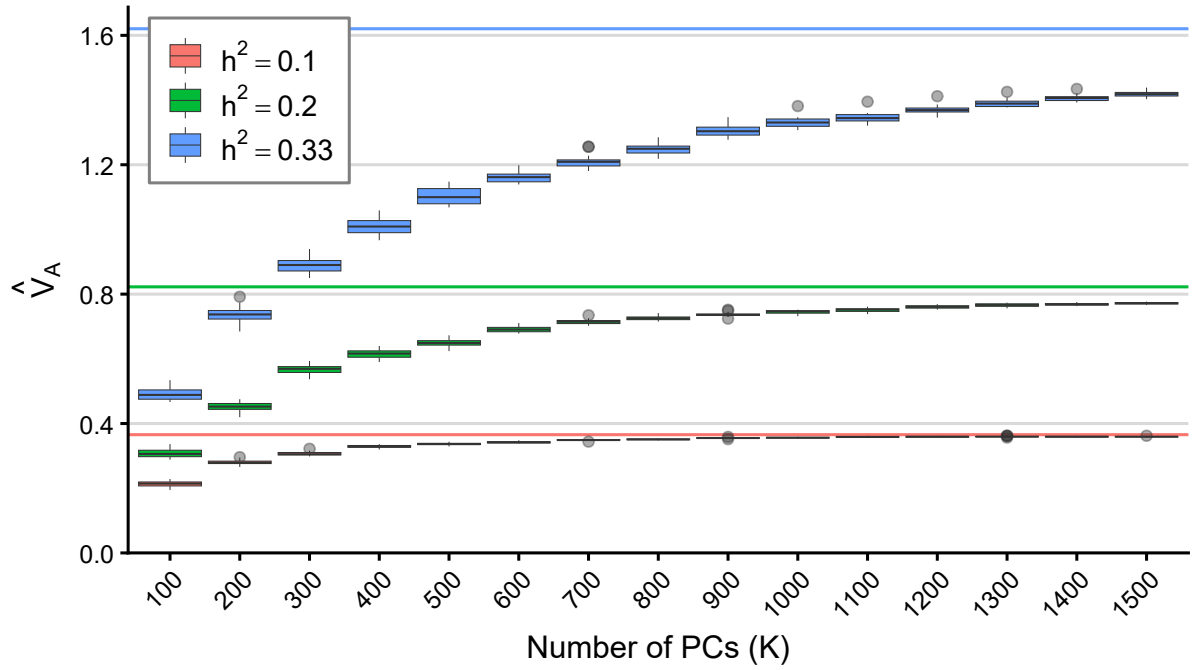

Figure 5: Recovered absolute additive genetic variance on the liability scale as a function of the number of principal components (K). The horizontal lines indicate the true simulated additive genetic variances, and boxplots summarize the distribution across 20 independent replicates.

5 Discussion

In this project, the main goal was to investigate how the complexity of genomic prediction models influences their ability to capture genetic signal and predict binary traits under different levels of heritability. Using Bayesian principal component ridge regression in a binary generalized linear mixed model framework, prediction performance was evaluated across the number of principal components K included, using three complementary metrics: prediction accuracy on the observed phenotypic scale (AUC), correlation on the liability scale, and the amount of genetic variance reconstructed by the model. Across all heritability scenarios, prediction performance on the observed scale was mostly insensitive to model complexity and primarily determined by heritability (Figure 2). The correlation on the liability scale displayed an optimum that shifted towards higher model complexity (K) as heritability increased (Figure 3), and increasing K resulted in improved recovery of additive genetic variance (Figure 4). Together, these results demonstrate that the optimal model complexity depends on the metric and that increasing complexity does not translate into improved predictive performance.

To assess prediction performance on the observed phenotypic scale, we evaluated the area under the ROC curve (AUC). As expected, AUC increases with heritability, reflecting that a larger proportion of phenotypic variation is genetically determined (Figure 2). In contrast, AUC shows only weak sensitivity to the number of principal components, and no clear or consistent optimal value of K emerges across heritability scenarios (Table 3). This pattern arises because AUC evaluates discrimination on the observed binary scale rather than the accuracy of genetic prediction on the liability scale. Even when liability-scale prediction improves with increasing model complexity, phenotypic noise and the non-linear threshold mapping limit how much of this improvement is transferred to the observed scale. At higher values of K, additional principal components may introduce fluctuations in the linear predictor that are too small or too noisy to affect 0/1 discrimination, resulting in a largely flat AUC profile. When interpreted relative to the ceiling AUC, the maximal median AUC corresponds to roughly half of the ceiling AUC. This reflects fundamental constraints on phenotypic prediction: accuracy on the liability scale is itself bounded by heritability (Daetwyler et al., 2008), and only a fraction of that signal is expressed on the observed binary scale. Previous work has shown that gains in liability-scale prediction translate into limited improvements in AUC (Wray et al., 2019), and that substantial variation on the liability scale may appear nearly invisible after thresholding (Acker & Reid, 2025). This raises the question of whether alternative modeling strategies for binary traits, which operate more directly on the liability scale or avoid hard thresholding, could better preserve information from the genetic predictor.

To gain insight into predictive performance on the liability scale, we examine the correlation between true and estimated breeding values. The correlation shows dependence on the number of principal components, with an optimum that shifts toward higher values of K as heritability increases (Figure 3). This pattern was also consistent across replicates (Table 4). In all scenarios, prediction accuracy peaks at early or intermediate values of K and declines when additional components are included. Shrinkage stabilizes estimated effects and limits prediction variance, but this alone is not sufficient to ensure stable correlation when many components are included. The liability-scale correlation measures how well individuals are ranked by their genetic values, and is therefore sensitive to the accumulation of many small, noisy contributions. As a consequence, adding large

numbers of PCs introduces a limited genetic signal with substantial noise. Although each component is strongly shrunk, their combined effect can still reduce accuracy, resulting in a decline in correlation at high K. To place these patterns in a theoretical context, we compare the observed correlations with a theoretical reference point, the square root of heritability. The relationship between the observed correlation and the reference point differs across heritability scenarios (Figure 3). When considering the median correlation at K with maximal median correlation at low heritability, the correlation at the optimal number of components lies above the bound, whereas at intermediate heritability, it closely matches the expected limit. At high heritability, the observed correlation falls systematically below the bound. The different relationships to the theoretical bound across heritability levels reflect how shrinkage-based prediction balances bias and variance differently depending on signal strength. At low heritability, prediction is dominated by noise, and shrinkage substantially improves stability by reducing estimation variance, sometimes yielding correlations above the theoretical bound. As heritability increases, the genetic signal becomes stronger, and shrinkage can dampen the real genetic signal, resulting in systematically lower correlations relative to the bound.

In addition to predictive performance, we examined how model complexity influences the recovery of additive genetic variance. The reconstructed additive genetic variance (V_A), calculated over a grid of PCs, provides a measure of how access to the underlying genetic signal, changes as the number of principal components increases. Expressing the reconstructed additive genetic variance as a proportion of the true additive variance emphasizes differences in how efficiently the genetic signal is recovered across heritability levels, with higher heritabilities showing a slower approach to complete recovery as K increases (Figure 4). The simulation framework involves some stochasticity in the sense that different sets of SNPs may be assigned non-zero effects across heritability scenarios. However, this is not expected to induce systematic differences in the accumulation of genetic variance across principal components. For polygenic traits, the variance explained by additive genetic effects has been shown to scale proportionally with the variance explained by the SNPs (Aspheim et al., 2024), suggesting that the relative recovery patterns observed here primarily reflect differences in total genetic variance rather than properties of signal representation. When the same results are viewed on an absolute scale, the same pattern is observed (Figure 5). In addition, for a fixed number of principal components, higher heritability scenarios consistently exhibit greater variance recovery for the same number of principal components, with curves starting at higher levels and remaining above those for lower heritability. These perspectives are complementary rather than contradictory, capturing differences between relative recovery efficiency and the absolute amount of genetic signal available for recovery. This pattern follows directly from the definition of heritability. Higher heritability implies that a larger proportion of the total phenotypic variance is due to additive genetic effects.

A key insight from this project is the different behavior of prediction performance on the liability and observed scales. While improvements in liability-scale prediction are clear as model complexity increases to an optimal point, these gains are largely masked when performance is evaluated using AUC. In addition, several limitations of the present study should be acknowledged. Liability-scale correlation relies on access to true genetic values and can therefore only be computed in a simulation setting, where it is used as a diagnostic measure rather than a performance metric directly applicable to empirical data. As discussed in the background section, more realistic bounds on prediction accuracy

for binary traits depend on both observed-scale heritability, and the effective amount of training data relative to trait complexity (Daetwyler et al., 2008). From this perspective, the bound used in this project may overstate achievable performance in empirical settings. Moreover, the simulation framework relies on a simplified genetic architecture and a limited number of replicates per heritability level, which may influence the absolute scale of the results. Lastly, heritability was varied through changes in genetic effects rather than environmental variance, which may limit biological interpretation. These factors should be considered when interpreting the quantitative levels of prediction accuracy.

These limitations also point toward natural directions for future work. Running additional simulations and exploring more biologically realistic ways of varying heritability, including changes in environmental variance, could improve robustness and interpretability. Alternative modeling strategies for binary traits, such as approaches that operate more directly on the liability scale, avoid hard thresholding, or even machine learning, may better preserve information from the genetic predictor to the observed 0/1 scale. Finally, applying these methods to empirical data and incorporating non-genetic predictors would help assess their practical relevance.

References

Abdi, H., & Williams, L. J. (2010). Principal component analysis. *WIREs Computational Statistics*, *2*(4), 433–459.

Acker, P., & Reid, J. M. (2025). Full decomposition of phenotypic variance in dichotomous traits: New methods and key implications for behaviour, demography and evolutionary ecology. *Methods in Ecology and Evolution*.

Ashraf, B., Hunter, D. C., Bérénos, C., Ellis, P. A., Johnston, S. E., Pilkington, J. G., Pemberton, J. M., & Slate, J. (2022). Genomic prediction in the wild: A case study in soay sheep. *Molecular Ecology*, *31*(24), 6541–6555.

Aspheim, J. C. H., Aase, K., Bolstad, G. H., Jensen, H., & Muff, S. (2024). Bayesian marker-based principal component ridge regression – a flexible multipurpose framework for quantitative genetics in wild study systems. *bioRxiv*.

Brookes, A. J. (1999). The essence of SNPs. *Gene*, *234*(2), 177–186.

Bush, W. S., & Moore, J. H. (2012). Chapter 11: Genome-wide association studies. *PLOS Computational Biology*, *8*(12), e1002822.

Conner, J. K., & Hartl, D. L. (2004). *A primer of ecological genetics*. Sinauer Associates.

Daetwyler, H. D., Villanueva, B., & Woolliams, J. A. (2008). Accuracy of predicting the genetic risk of disease using a genome-wide approach (M. N. Weedon, Ed.). *PLoS ONE*, *3*(10), e3395.

Falconer, D. S., & Mackay, T. F. (1989). *Introduction to quantitative genetics* (3rd ed.). Longman Group.

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

Griffiths, A. J. F. (2015). *Introduction to genetic analysis*. New York, NY : W.H. Freeman & Company.

Karki, R., Pandya, D., Elston, R. C., & Ferlini, C. (2015). Defining "mutation" and "polymorphism" in the era of personal genomics. *BMC Medical Genomics*, *8*(1), 37.

Lynch, M., & Walsh, B. (1998). *Genetics and analysis of quantitative traits*. Oxford University Press.

Martino, S., & Riebler, A. (2019). Integrated nested laplace approximations (INLA).

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

Ranke, P. S., Araya-Ajoy, Y. G., Ringsby, T. H., Pärn, H., Rønning, B.-E., Jensen, H., Wright, J., & Saether, B.-E. (2021). Spatial structure and dispersal dynamics in a house sparrow metapopulation. *Journal of Animal Ecology*, *90*(12), 2767–2781.

Rue, H., Martino, S., & Chopin, N. (2009). Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *71*(2), 319–392.

Steinsland, I., Larsen, C. T., Roulin, A., & Jensen, H. (2014). Quantitative genetic modeling and inference in the presence of nonignorable missing data: Quantitative genetic nonignorable missing data. *Evolution*, *68*(6), 1735–1747.

Syvänen, A.-C. (2001). Accessing genetic variation: Genotyping single nucleotide polymorphisms. *Nature Reviews Genetics*, *2*(12), 930–942.

VanRaden, P. (2008). Efficient methods to compute genomic predictions. *Journal of Dairy Science*, *91*(11), 4414–4423.

Visscher, P. M., Brown, M. A., McCarthy, M. I., & Yang, J. (2012). Five years of GWAS discovery. *The American Journal of Human Genetics*, *90*(1), 7–24.

Wilson, A. J., Réale, D., Clements, M. N., Morrissey, M. M., Postma, E., Walling, C. A., Kruuk, L. E. B., & Nussey, D. H. (2010). An ecologist's guide to the animal model. *Journal of Animal Ecology*, *79*(1), 13–26.

Wray, N. R., Kemper, K. E., Hayes, B. J., Goddard, M. E., & Visscher, P. M. (2019). Complex trait prediction from genome data: Contrasting EBV in livestock to PRS in humans: Genomic prediction. *Genetics*, *211*(4), 1131–1141.

Appendix

A Replicate-Level Selection of the Number of Principal Components

Table 5: Replicate-level selection of the optimal number of principal components (K_{best}) based on test set AUC. For each heritability level, the table reports the value of K achieving the highest AUC in each replicate, along with the corresponding AUC value.

rep	$\mathbf{h^2 = 0.1}$		$\mathbf{h^2 = 0.2}$		$\mathbf{h^2 = 0.33}$	
	K	AUC	K	AUC	K	AUC
1	400	0.572	300	0.608	200	0.646
2	1100	0.579	300	0.606	400	0.634
3	1100	0.585	600	0.596	400	0.624
4	300	0.580	300	0.599	100	0.621
5	100	0.553	300	0.582	200	0.640
6	700	0.597	700	0.605	400	0.660
7	400	0.564	300	0.592	600	0.628
8	200	0.584	300	0.607	200	0.639
9	700	0.586	1000	0.602	400	0.637
10	100	0.586	700	0.618	900	0.620
11	900	0.578	600	0.581	200	0.626
12	900	0.552	600	0.621	200	0.656
13	500	0.613	500	0.627	1400	0.630
14	400	0.598	200	0.609	100	0.617
15	1100	0.564	300	0.634	500	0.655
16	100	0.598	300	0.607	500	0.646
17	200	0.599	300	0.584	1200	0.628
18	500	0.583	300	0.592	300	0.647
19	400	0.583	800	0.598	200	0.652
20	200	0.611	700	0.597	700	0.652

Table 6: Replicate-level selection of the optimal number of principal components (K_{best}) based on correlation. For each heritability level, the table reports the value of K achieving the highest correlation between predicted and true breeding values in each replicate, along with the corresponding correlation value.

rep	$\mathbf{h^2 = 0.1}$		$\mathbf{h^2 = 0.2}$		$\mathbf{h^2 = 0.33}$	
	K	corr	K	corr	K	corr
1	200	0.416	400	0.453	700	0.477
2	200	0.396	400	0.417	500	0.501
3	600	0.399	400	0.445	700	0.492
4	200	0.400	400	0.425	200	0.476
5	100	0.397	600	0.367	200	0.481
6	300	0.360	500	0.398	400	0.507
7	300	0.394	300	0.442	500	0.517
8	100	0.404	300	0.476	700	0.487
9	400	0.392	1100	0.416	300	0.486
10	100	0.364	400	0.417	500	0.475
11	200	0.438	600	0.412	200	0.462
12	100	0.376	600	0.453	200	0.498
13	400	0.414	400	0.418	300	0.494
14	300	0.413	600	0.399	200	0.477
15	100	0.387	400	0.439	500	0.517
16	100	0.367	400	0.440	900	0.523
17	100	0.444	400	0.393	700	0.444
18	100	0.416	400	0.452	500	0.500
19	600	0.387	800	0.479	400	0.467
20	300	0.417	600	0.432	700	0.553

NTNU

Declaration of AI aids and tools

Have any AI-based aids or tools been used in the creation of this report?

No ✔ Yes

If yes: please specify the aid/tool and area of use below.

Text

✔ **Spell checking.** Are parts of the text checked by: Grammarly, Ginger, Grammarbot, LanguageTool, ProWritingAid, Sapling, Trinka AI or similar tools?

✔ **Text-generation.** Are parts of the text generated by: ChatGPT, GrammarlyGO, Copy.AI, WordAi, WriteSonic, Jasper, Simplified, Rytr or similar tools?

✔ **Writing assistance.** Are one or more of the report's ideas or approach suggested by: ChatGPT, Google Bard, Bing chat, YouChat or similar tools?

If no, sign document. If yes, use of text aids/tools apply to this report -please specify usage here:

Grammarly and ChatGPT was used as support during writing process, including language checks, reprhrasing of text, improving structure and clarity in several sections. ChatGPT was also used to help resolve LaTex formatting issues.

Codes and algorithms

✔ **Programming assistance.** Are parts of the codes/algorithms that i) appear directly in the report or ii)have been used to produce results such as figures, tables or numerical values been generated by: GitHub Copilot, CodeGPT, Google Codey/Studio Bot, Replit Ghostwriter, Amazon CodeWhisperer, GPTEngineer, ChatGPT, Google Bard or other tools.

If yes, use of programming assistance aid/tools apply to this report - please specify usage here:

ChatGPT was used as programming assistance including plotting, debuggig, clean up commenting, understanding error messages, sugestions for improving code, and getting code ready for Markov.

Images and figures

✔ **Image generation.** Are one or more of the reports images/figures generated by: Midjourney, Jasper, WriteSonic, Stability AI, Dall-E or similar tools?

If yes, use of image generator aids/tools apply to this report – please specify usage here:

ChatGPT made code in LaTex for inserting figures and tables.

Other AI aids or tools. Have you used other types of AI aids or tools in the creation of this report?
If yes, please specify usage here:

✔ I am familiar with NTNU's regulations on artificial intelligence. I declare that any use of AI aids or tools are explicitly stated i) directly in the report or ii) in this declaration form.

19.12.2025/Trondheim

Signature/Date/Place