Fast and Accurate Bayesian Polygenic Risk Modeling with Variational Inference

Shadi Zabad
McGill University
Simon Gravel
McGill University https://orcid.org/0000-0002-9183-964X
Yue Li (✉ yueli@cs.mcgill.ca)
McGill University https://orcid.org/0000-0003-3844-4865

Article

Keywords:

Posted Date: May 16th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1643278/v1

Fast and Accurate Bayesian Polygenic Risk Modeling with Variational Inference

Shadi Zabad[1], Simon Gravel[2,*] and Yue Li[1,*]

[1]School of Computer Science, McGill University
[2]Department of Human Genetics, McGill University
[*]Correspondence to simon.gravel@mcgill.ca, yueli@cs.mcgill.ca

Abstract

The recent proliferation of large scale genome-wide association studies (GWASs) has motivated the development of statistical methods for phenotype prediction using single nucleotide polymorphism (SNP) array data. These polygenic risk score (PRS) methods formulate the task of polygenic prediction in terms of a multiple linear regression framework, where the goal is to infer the joint effect sizes of all genetic variants on the trait. Among the subset of PRS methods that operate on GWAS summary statistics, sparse Bayesian methods have shown competitive predictive ability. However, existing Bayesian approaches employ Markov Chain Monte Carlo (MCMC) algorithms for posterior inference, which are computationally inefficient and do not scale favorably with the number of SNPs included in the analysis. Here, we introduce Variational Inference of Polygenic Risk Scores (`VIPRS`), a Bayesian summary statistics-based PRS method that utilizes Variational Inference (VI) techniques to efficiently approximate the posterior distribution for the effect sizes. Our experiments with genome-wide simulations and real phenotypes from the UK Biobank (UKB) dataset demonstrated that variational approximations to the posterior are competitively accurate and highly efficient. When compared to state-of-the-art PRS methods, `VIPRS` consistently achieves the best or second best predictive accuracy in our analyses of 18 simulation configurations as well as 12 real phenotypes measured among the UKB participants of "White British" background. This performance advantage was higher among individuals from other ethnic groups, with an increase in R^2 of up to 1.7-fold among participants of Nigerian ancestry for Low-Density Lipoprotein (LDL) cholesterol. Furthermore, given its computational efficiency, we applied `VIPRS` to a dataset of up to 10 million genetic markers, an order of magnitude greater than the standard HapMap3 subset used to train existing PRS methods. Modeling this expanded set of variants conferred modest improvements in prediction accuracy for a number of highly polygenic traits, such as standing height.

1 Introduction

In recent years, with the rapid growth of large-scale biobank data with comprehensive genotyping and phenotyping efforts [1–3], there has been growing interest in developing statistical methods to quantify an individual's disease risk from their genotype data [4–8]. At the same time, these rich biobank data sources have powered many recent analyses of complex traits and diseases, revealing highly polygenic architectures [9–11] with a wide range of effect sizes across different genomic categories [12–14]. Linear models are an important framework for complex trait analysis which allow for the estimation of the additive genetic component of a phenotype, also known as a polygenic score (PGS) or polygenic risk score (PRS) in clinical contexts [5, 15]. Even though many examples of genetic interactions have been documented, such additive effects capture much of the genetic variation underlying human complex traits [16, 17]. Recent work has highlighted the clinical relevance of polygenic scores for some diseases and health conditions [18, 19], especially in applications related to disease risk stratification [20–22] and personalized medicine [23].

Estimating polygenic scores from genome-wide association study (GWAS) data has been an active area of research, with numerous methods recently developed [4, 6, 24–30]. Standard PRS methods formulate the problem of polygenic risk estimation in terms of a multiple linear regression framework, where the goal is to infer the joint effect sizes of all genetic variants on the trait. The most common class of genetic variation considered in these analyses are single nucleotide polymorphisms (SNPs), which are either measured by modern genotyping arrays or statistically imputed using reference haplotypes [31, 32].

Genotyping arrays combined with imputation can accurately capture the genotype of an individual at millions of genetic markers. When paired with modern GWAS sample sizes routinely exceeding hundreds of thousands of individuals, high dimensional data of this scale present several computational challenges. Furthermore, most individual-level GWAS data sources are protected for privacy concerns [33]. These two factors motivated the development of a number of PRS methods that estimate polygenic risk scores based on GWAS summary statistics alone [4, 6, 25–27, 29, 30], which are the marginal test statistic per SNP.

Within this class of summary statistics-based methods, Bayesian PRS models enable a principled way to incorporate prior knowledge as probability distributions over the genetic causal architecture of complex traits. In addition to providing meaningful estimates of parameter uncertainties [34], Bayesian approaches have shown competitive predictive ability, exceeding the predictive performance of heuristic or penalized estimators in many settings [6, 26, 29, 30]. However, a major limitation of existing Bayesian methods is that their scalability is hampered by slow and inefficient inference techniques. While heuristic methods such as Clumping-and-Thresholding (C+T) are routinely applied on millions of SNPs, Bayesian approaches are generally restricted to a subset of approximately one million genetic markers. One of the main reasons for this limitation stems from computational considerations: Most Bayesian PRS methods employ Markov Chain Monte Carlo (MCMC) algorithms to approximate the posterior for the effect sizes [4, 6, 26, 29]. MCMC algorithms are known to be asymptotically accurate but often slow to converge [35,36]. In practice, to obtain accurate posterior estimates, the MCMC chains need to be run for hundreds or thousands of iterations [4,6]. This challenge can be partially remedied with the help of efficient software implementation and enhanced linear algebra routines, which recently enabled scaling up a popular Bayesian PRS method to 2.8 million SNPs [6]. While this is an important advance, these variants still constitute a small fraction of the genetic variation that can be assayed by modern whole genome-sequencing technologies [3,37].

An alternative scheme for approximating the posterior density for the effect sizes is Variational Inference (VI), a fast and deterministic class of algorithms that recast the problem of posterior inference in the form of an optimization problem [35,36,38,39]. Variational methods have seen a surge of interest in the machine learning literature in recent years due to significant advances in stochastic optimization techniques [40,41]. Methods that utilize variational inference have been explored in a wide variety of statistical genetics applications, specifically in the context of linear mixed models (LMMs) [42], association mapping [43,44], fine-mapping [45], and enrichment analysis [46,47], among others. More recently, a theoretical study examined the properties of certain variational approximations to PRS [48].

In this work, we present VIPRS, a Bayesian summary statistics-based PRS method that utilizes variational inference to approximate the posterior for the effect sizes. We conduct a comprehensive set of experiments using simulated and real traits to assess the predictive ability of VIPRS in comparison with the some of the most popular Bayesian and non-Bayesian PRS methods. Overall, we show that VIPRS is a scalable and flexible method that enjoys the speed and efficiency of heuristic approaches such as Clumping-and-Thresholding (C+T), while rivaling state of the art Bayesian methods in terms of its predictive performance. We demonstrate the flexibility of the method by testing its predictions with different families of priors on the effect size, paired with four distinct strategies for tuning the hyperparameters of the model. To illustrate its scalability, we evaluate the predictive accuracy of VIPRS with approximately 10 million SNPs, almost an order of magnitude greater than what most existing PRS methods can handle. This enables us to examine the potential for phenotype prediction of a number highly polygenic traits at an unprecedented scale.

2 Results

2.1 Variational inference of polygenic risk scores (VIPRS)

Given a random sample of individuals from a general population with paired genotype and phenotype data, represented by a genotype matrix $\mathbf{X}_{N\times M}$ and a corresponding phenotype vector $\mathbf{y}_{N\times 1}$, we model the dependence of the phenotype on the genotype via the standard linear model: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. Here, $\boldsymbol{\beta}_{M\times 1}$ is a random and unknown vector of effect sizes for each of the M variants included in the model and $\boldsymbol{\epsilon}_{N\times 1}$ is a vector of residuals. In this setup, the Bayesian formulation of polygenic risk modeling is concerned with inferring the posterior distribution for the effect sizes of the genetic variants on the trait,

$$p(\boldsymbol{\beta} \mid \mathbf{X}, \mathbf{y}, \boldsymbol{\theta}) = \frac{p(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\theta}) p(\boldsymbol{\beta} \mid \boldsymbol{\theta})}{\int p(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\theta}) p(\boldsymbol{\beta} \mid \boldsymbol{\theta}) d\boldsymbol{\beta}} \tag{1}$$

where $\boldsymbol{\theta}$ is a composite term for the hyperparameters of the model, such as the residual variance σ_ϵ^2. The posterior distribution for the effect sizes is proportional to the likelihood $p(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\theta})$ and the prior $p(\boldsymbol{\beta} \mid \boldsymbol{\theta})$, with the latter encapsulating prior beliefs about the genetic architecture of the trait. For continuous traits, assuming that the samples are unrelated and ancestrally homogeneous, the likelihood is commonly modelled with an isotropic Gaussian distribution, $p(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{y}; \mathbf{X}\boldsymbol{\beta}, \sigma_\epsilon^2 \mathbf{I})$.

As for the prior, a common assumption that has proved powerful in predictive settings [4, 29] is to model the effect size of each variant with a two-component mixture distribution: $p(\beta_j) = \pi \mathcal{N}(\beta_j; 0, \sigma_{\boldsymbol{\beta}}^2) + (1 - \pi)\delta_0$. Under this prior distribution, with probability π the effect size is drawn from a Gaussian centered at zero with the scale proportional to σ_β, and with probability $1 - \pi$ from a Dirac delta density centered at zero. The first component captures the contribution of causal variants whereas the second component models the role of variants with null effects. This is sometimes known as the spike-and-slab prior and is used to enforce sparsity on the effect size estimates [49–51]. However, a major difficulty in this context is that this flexible family of priors results in an analytically intractable posterior density, due to the integral in the model evidence $\int p(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\theta}) p(\boldsymbol{\beta} \mid \boldsymbol{\theta}) d\boldsymbol{\beta}$. This requires the use of approximate posterior inference schemes.

Most existing methods for Bayesian PRS estimation are based on Markov Chain Monte Carlo (MCMC) inference algorithms, such as Gibbs sampling [4, 6, 26, 29]. Despite their many successful applications, MCMC methods can be computationally inefficient for performing inference over high-dimensional genome-wide data containing millions of SNPs. Additionally, with MCMC algorithms, it is difficult to assess model convergence [35, 36]. As an alternative, in this study, we present Variational Inference of PRS (VIPRS). We approximate the posterior distribution of the effect sizes by a proposed parametric density, $q(\boldsymbol{\beta}, \mathbf{s})$, and optimize its parameters to minimize a global measure of the divergence from the true posterior [38, 39, 44, 45]. For efficient inference, we use a paired mean-field distribution family [45, 52] that factorizes the joint density of the effect sizes into the product of the individual densities for the effect size of each variant,

$$q(\boldsymbol{\beta}, \mathbf{s}) = \prod_{j=1}^{M} q(\beta_j, s_j).$$

Here, s_j is the posterior Bernoulli variable indicating whether variant j is causal for the trait of interest. Under this posterior variational family, the SNP effect sizes are assumed to be independent, a restrictive assumption that nonetheless works well in practice. In the **Methods** section and accompanying **Supplementary Information**, we show that this formulation results in a fast coordinate-ascent optimization procedure that, under some assumptions, can be expressed solely in terms of GWAS summary statistics. Inference under this model, however, requires setting the hyperparameters $\boldsymbol{\theta} = \{\sigma_\epsilon^2, \sigma_\beta^2, \pi\}$, for which we explored four different strategies.

In the basic formulation of the VIPRS model, we use a Variational Expectation-Maximization (VEM) framework, also known as empirical Bayes. In the E-Step, we update the parameters of the variational density $q(\boldsymbol{\beta}, \mathrm{s})$; in the M-Step, we update the hyperparameters using their maximum likelihood estimates [45, 53]. Additionally, we experimented with grid search (GS) (VIPRS-GS) [54], Bayesian optimization (BO) (VIPRS-BO) [55], and Bayesian Model Averaging (BMA) (VIPRS-BMA) [45] as means to set or integrate out some or all the hyperparameters of the model (**Methods**).

2.2 Genome-wide simulation results

To examine the predictive performance of the VIPRS model compared to existing PRS methods, we simulated quantitative and case-control traits with varying genetic architectures and heritability values. To align our simulations with the real trait analyses in terms of cohort size and composition, we used genotype data for a subset of $\approx 340,000$ unrelated White British individuals from the UK Biobank (**Methods**) and a HapMap3 subset of ≈ 1.1 million genotyped and imputed SNPs. The simulations followed the generative model outlined in the **Methods** section, with the effect size of each variant drawn from the spike-and-slab mixture density and residuals for each individual sampled from an isotropic Gaussian density. For binary traits, we simulated case-control status following the Liability Threshold model [56], with the prevalence set to 15%.

The genetic architectures we simulated varied from sparse to highly polygenic, with the proportion of SNPs contributing to the variation in the trait ranging along a pre-specified grid $\pi \in \{10^{-4}, 10^{-3}, 10^{-2}\}$. Along with controlling for polygenicity, we also varied the proportion of additive genetic variance captured by all causal SNPs, $h^2_{SNP} \in \{0.1, 0.3, 0.5\}$, such that the simulated traits range from the mildly to the highly heritable. For each unique configuration, we simulated 10 independent phenotypes, for a total of 90 traits for each class (binary and continuous). Once the traits were simulated for all individuals in the dataset, we randomly split the sample into 70% training, 15% validation, and 15% testing, with the training set used to generate GWAS summary statistics.

Next, we fit the VIPRS model to the summary statistics from the training data, along with five other commonly-used PRS methods: SBayesR [6], LDPred2 [29], PRScs [26], Lassosum [25], and PRSice2 [27]. The first three methods use the Bayesian framework outlined above for approximate posterior inference, all employing a Gibbs sampling algorithm. They are mainly distinguished by the families of prior density they assign to the effect sizes, among many other algorithmic choices. After fitting each method on the summary statistics from the training data, we used the effect size estimates to generate polygenic scores for individuals in the held-out test set and evaluated their predictive performance. For quantitative traits, we computed the incremental Prediction R^2 for each model, while for binary traits we show the area under the precision-recall curve (AUPRC), a preferable metric in the presence of class imbalance [57]. In addition to the 5 external PRS models, we also examined the predictive performance of the standard VIPRS model trained with the standard Variational EM framework (**Methods**) as well as a version of the VIPRS model, dubbed VIPRS-GS, in which we perform grid search and tune the hyperparameters based on predictive performance on a held-out validation set.

The predictive performance results for this simulation study are summarized in **Fig.** 1, which shows that VIPRS outperforms or is on-par with state-of-the-art PRS methods on almost every genetic architecture tested. In particular, our analyses indicate that VIPRS provides the most benefit for more sparse architectures and highly heritable traits (leftmost panels in **Fig.** 1(a) and 1(b)). Notably, in this particular setting, VIPRS is able to capture most of the additive genetic variance (as measured by the R^2 metric, which is upper-bounded by the heritability), while other Bayesian and non-Bayesian methods often lag behind. For highly polygenic traits with proportion of causal variants set to 1% ($\pi = 0.01$), all models conferred lower predictive accuracy, relative the heritability values that we simulated with. This is because, under our simplified simulation scenario, the larger the number of causal variants, the smaller the effect size per SNP. Consequently, this makes it more difficult for PRS method to pick up the true causal signals, at least given the training sample sizes available. Nonetheless, the VIPRS models conferred higher predictive performance relative to competing methods in this setting as well. This pattern holds for both quantitative (**Fig.** 1(a)) as well as binary case-control phenotypes (**Fig.** 1(b)). This improvement in predictive accuracy comes also with improved computational efficiency, with the run-time of the standard VIPRS model rivaling heuristic and deterministic methods, such as Lassosum and PRSice2 (**Fig.** 3).

One possible explanation for the performance improvement of VIPRS over competing Bayesian models could be the fast model convergence enabled by our variational inference algorithm in contrast to the stochastic MCMC approaches implemented by the other methods. Importantly, our simulation setup followed the same spike-and-slab model generative process assumed by VIPRS as well as other Bayesian methods, such as LDPred2 [4,29]. Therefore, it is important to evaluate these methods on real phenotypes as shown next.

2.3 Application on real phenotypes in the UK Biobank

Given its competitive performance on simulated traits, we next sought to assess the relative predictive ability of the VIPRS model on real phenotypes measured for a subset of $\approx 340,000$ unrelated White British individuals in the UK Biobank. This focus on a large sub-cohort of relatively uniform ancestry helps us achieve sufficient power while reducing confounding due to population structure. A downside of this approach is that it is expected to yield PRS estimates that perform more poorly for individuals of other ancestries [58,59], a limitation that we examine in more detail in the next section.

For this analysis, we extracted and processed phenotypic measurements for 9 quantitative traits and 3 binary traits that are commonly used to benchmark PRS methods (**Table** 1). The traits considered have varying (inferred) genetic architectures and SNP heritabilities. To make full use of the data, we followed a 5-fold cross-validation study design, where in each iteration, 80% of the samples with trait measurements were used to generate the GWAS summary statistics and training the PRS models and 20% were used as an independent test set.

Our experiments show that, across a variety of different phenotypes, VIPRS is competitive with commonly-used Bayesian PRS methods (**Fig.** 2). Within the category of Bayesian PRS methods, the predictive performance of VIPRS is especially distinguished for anthropometric and blood lipid traits (**Fig.** 2(a)). For instance, when compared to the LDPred2 model, which imposes the same spike-and-slab prior on the effect sizes, VIPRS shows an average of 4.6% improvement in prediction R^2 on continuous traits. However, in many cases the standard VIPRS model lags behind the SBayesR [6] and Lassosum [25] models (**Fig.** 2). In addition to the difference in posterior approximation strategy (variational inference versus Gibbs sampling), the SBayesR model differs from the VIPRS model in three other important respects: (1) The prior on the effect size, (2) The estimator for the Linkage Disequilibrium (LD) between variants, and finally (3) the approach for estimating the hyperparameters of the model. We sought to understand the effect of each of these modeling choices on the predictive performance of our model.

To address the first point, we derived and implemented a version of VIPRS called VIPRSMix , where we replaced the spike-and-slab prior on the effect sizes with a Gaussian mixture prior with four mixture components (**Supplementary Information**) [6, 44]. Our experiments show that the more expressive mixture prior improves the performance of the standard VIPRS model on some traits, especially highly heritable and polygenic traits such as standing height and HDL (**Supplementary Fig.** S4), with on average of 2.4% increase in prediction R^2 on continuous traits. However, the improvement is not consistent across all traits and using this prior does not fully bridge the gap between VIPRS and SBayesR.

Secondly, we assessed the impact of the LD estimator on the predictive performance of the VIPRS model by re-fitting the model with three commonly-used estimators for LD: windowed [29, 60], shrinkage [6, 61], and block [26, 62] (**Methods**). Our experiments indicate that, on both simulated and real traits, the windowed estimator for LD, combined with sufficiently large sample sizes, results in the most robust predictive performance (**Supplementary Fig.** S1, S2). Though the block and shrinkage estimators tend to be more robust if the sample size of the LD reference panel is small (**Supplementary Fig.** S1, S2).

Finally, and most importantly, the standard VIPRS model differs from the SBayesR model in terms of its hyperparameter estimation strategy. SBayesR follows a fully Bayesian approach for learning the hyperparameters of the model, assigning them priors and inferring their posterior distributions [6]. By contrast, VIPRS follows a Variational EM algorithm where in the M-Step we set the hyperparameters to their maximum-likelihood estimates [44, 45]. This latter strategy is known to be prone to overfitting or entrapment in local maxima [44, 45, 53, 63, 64]. As an alternative to the VEM framework, we tested three other strategies for tuning the hyperparameters of the model, including grid search [54], Bayesian optimization [55], and Bayesian model averaging [45] (see **Methods**, **Supplementary Fig.** S5,S6). In this context, similar to the Lassosum and LDPred2 methods, we found that by setting some of the hyperparameters of the model via grid search with an independent validation set, VIPRS-GS provides a powerful remedy in most settings (**Fig.** 2, **Supplementary Fig.** S5,S6), resulting in a balanced trade-off between computational speed and predictive accuracy (**Fig.** 2 and 3). Indeed, our results show that the

VIPRS-GS model conferred the highest or second highest predictive performance on all traits tested (**Fig.** 2), consistently exceeding the performance of the VEM-based VIPRS. At the same time, the main drawback of the grid search approach is that, despite the parallel software implementation, it results in a significant slowdown compared to the VEM approach (**Fig.** 3). In terms of predictive performance, the advantage of the grid search is most prominent for highly heritable traits, such as standing height and HDL (**Fig.** 2(a)). For the other traits, SBayesR is on-par or only marginally better. This indicates that the gap in predictive performance between SBayesR and the standard VIPRS model is mostly due to differences in hyperparameter estimation strategy.

2.4 PRS validation in minority populations in the UK Biobank

When trained on GWAS data from a single source population, transferability of PRS estimates across populations is limited [58, 59], with the degradation in predictive accuracy increasing with the increase in allele frequency differentiation (Fst) between populations [59]. At the same time, recent studies of cross-population genetic correlations have demonstrated strong correlations in the genetic architectures of complex traits between various ancestry groups [65, 66]. These correlations imply that PRS models that perform better in the source population will also tend to perform more favorably when applied to the target populations.

To assess this, we extracted genotype and phenotype data for individuals who self-identified as Italian ($N = 6177$), Indian ($N = 6011$), Chinese ($N = 1769$), and Nigerian ($N = 3825$). The self-reported ethnic backgrounds were further validated based on the Principal Components (PC) of the genetic relationship matrix (GRM) [59] (**Methods**). Using the effect size estimates derived from training the PRS models on summary statistics from the White British cohort across the five training folds, we computed a PRS for each individual in the target population. Given the real phenotype measurements for these individuals, we evaluated the predictive performance using relative incremental prediction R^2, where the R^2 in the target population was divided by the R^2 of the best performing model on the test set in the White British cohort.

Our results confirm that for most of the traits analyzed, the models with the best predictive performance on the source population (White British) tend to transfer better to the target populations (**Fig.** 4). Furthermore, consistent with other analyses in this space [58, 59], the drop in predictive accuracy generally tracks with the Euclidean distance between the White British and the target populations in PC space. Interestingly, deviations from this general pattern were observed for LDL and birth weight, which may be due to gene-by-environment interactions [66]. For LDL specifically, we observed strong differentiation in transferability between PRS methods, with VIPRS and VIPRS-GS attaining upwards of 1.5 times the prediction accuracy of the next competing PRS method in individuals of Nigerian and Chinese ancestry (**Fig.** 4). This likely indicates that the VIPRS model correctly inferred the effect sizes of causal variants that are shared across ancestries but were missed by other methods.

2.5 Scaling up VIPRS to 10 million variants

In recent years, with the advent of biobank-scale whole-genome sequencing efforts [3, 37] and improved variant imputation pipelines [31], there has been increasing interest in understanding the extent to which larger and larger sets of genetic variants enable us to better capture the genetic diversity underlying complex traits [67]. This is especially important in light of recent results that showed that a substantial proportion of the missing heritability is due to imperfect tagging of rare causal variants by common SNPs [67]. In this context, the main advantage of VIPRS over competing Bayesian methods is its speed and scalability (**Fig.** 3), thus we wanted to understand the extent to which Bayesian PRS methods could benefit from modeling an expanded set of SNPs. Here, for the first time, we report on the predictive performance of a Bayesian PRS method with approximately 9.6 million measured or imputed genetic variants, between four times to an order of magnitude greater than what previous Bayesian PRS methods were able to handle [4, 6, 26, 29]. This includes all bi-allelic variants with minor allele frequency greater than 0.1% and minor allele count greater than 5 in the White British cohort in the UK Biobank.

Following the same 5-fold cross-validation study design described previously, we observed that modeling an expanded set of SNPs results in improved predictive performance for highly heritable and polygenic traits, such as standing height and HDL, in comparison with the best performing PRS model using a subset of 1.1 million HapMap3 SNPs (**Fig.** 2), which is consistent with previous studies [6] (**Fig.** 5 and **Supplementary Fig.** S4). Concretely, for standing height and HDL in particular, including almost an order of magnitude more variants resulted, on average, in 6.3% and 3.7% relative improvement in prediction R^2, respectively. Importantly, the performance boost is even more pronounced when using a mixture prior with four components (VIPRSMix-10m), indicating that the heterogeneous genetic architecture underlying those complex traits are better modeled by a more flexible prior (**Fig.** 5). When using a mixture prior with the expanded set of variants (VIPRSMix-10m), we see an average of 13.4% and 11.5% relative improvement in prediction R^2 on standing height and HDL, respectively, compared to the standard VIPRS model with the HapMap3 SNPs (**Fig.** 5 and **Supplementary Fig.** S4). However, the improvement is not consistent across all traits. Similar to previous studies [6], we saw that including more variants sometimes led to modest drop in predictive accuracy, perhaps due to increased noise in the PRS estimate. Presumably, imputation errors for rare variants could potentially degrade the performance of the model in this setting. Therefore, we believe that this analysis presents a lower-bound on what could be achieved with more accurate and complete whole genome sequencing data [3, 37]. Finally, we highlight that even after including an order-of-magnitude more variants, the predictive performances of the VIPRS-10m and VIPRSMix-10m models often lag behind the grid search model using the HapMap3 SNPs alone VIPRS-GS (**Fig.** 5 and **Supplementary Fig.** S4). This indicates that the predictive accuracy of the model can potentially be significantly improved with better and more scalable hyperparameter inference strategies (**Discussion**).

2.6 PRS analysis with external GWAS summary statistics

A common use case in the inference of polygenic scores involves settings where the GWAS summary statistics and the LD reference panel are estimated from two different cohorts [8,33]. In other cases, the GWAS summary statistics may be derived from a meta-analysis that combines data from a number of different studies. These settings may present potential mismatches and heterogeneities between of LD reference panel and GWAS summary statistics and are thus challenging to model, often leading to substantial loss in predictive power [29,30,68,69].

To systematically assess the robustness of VIPRS to potential heterogeneities and mismatches between the GWAS cohort and the LD reference panel, we conducted an analysis where we downloaded a number of publicly available GWAS summary statistics (PASS) for some of the traits analyzed previously, including: standing height [70], HDL and LDL [71], BMI [72], Type 2 diabetes [73], and Rheumatoid arthritis [74] (**Table** 1, see **Data Availability**). Most of these studies were conducted in individuals of general European ancestry, some in the form of meta-analysis. Therefore, we would expect some degree of differences between the LD reference panels derived exclusively from the White British cohort in the UK Biobank and the in-sample LD from these GWAS cohorts (which are not available).

We fit VIPRS as well as other PRS methods to the external GWAS summary statistics, providing the same 5-fold validation and testing cohorts within the UK Biobank for the purposes of hyperparameter tuning and evaluation as in the previous analysis. Our results indicate that, for some of the studies analyzed, the standard VIPRS model can be sensitive to mismatches between the GWAS cohort and LD reference panel (**Fig.** 6), though not to the same extent as SBayesR, which failed to converge for all of the quantitative traits analyzed, consistent with earlier work in this area [30]. Interestingly, for the VIPRS model, we saw that using the shrinkage estimator for LD proved to be more robust in the presence of mismatches, especially in the case of standing height (**Supplementary Fig.** S2). Moreover, when using a validation set to tune the hyperparameters of the model, VIPRS-GS recovered most of the drop in performance relative to other PRS methods and showed competitive predictive ability (**Fig.** 6). This suggests that the VIPRS model trained to maximize the Evidence Lower BOund (ELBO) (**Methods**) of the external GWAS data may not generalize well to the UK Biobank individuals. Indeed, our experiments show a partial reversal in the correspondence between the ELBO and the validation R^2 (the metric that VIPRS-GS is optimizing) in the analysis of some of the external summary statistics (**Supplementary Fig.** S12), which explains the poor predictive performance of the standard VIPRS model in those settings.

Given this observation, if an independent validation set is not available, we recommend that users of the VIPRS software run principled tests of LD mismatch and heterogeneity, such as the recently published DENTIST method [69] before fitting the model to GWAS data. In the **Supplementary Text**, we also derived a stochastic estimator of the DENTIST test statistic that can be computed efficiently and provided that as a utility function in our software (see **Code Availability**).

3 Discussion

In this paper, we introduced VIPRS , a fast and flexible Bayesian PRS method that approximates the posterior for the effect sizes of genetic variants on the phenotype using variational inference techniques. Our genome-wide simulation analyses using genotype data from the White British cohort in the UK Biobank demonstrated that variational approximations to the posterior are not only computationally efficient, but they provide highly accurate polygenic score estimates across diverse genetic architectures. Indeed, in some simulation scenarios, VIPRS exceeded the predictive performance of competing Bayesian and non-Bayesian methods by large margins. The competitive predictive accuracy of the VIPRS method replicated in our analyses of real quantitative and binary phenotypes measured for the same UKB participants, though the differences between the methods in this setup were more modest. Similar systematic but mostly modest benefits were observed when PRS methods were applied to individuals from ancestries not included in the training dataset, emphasizing the robustness of the approach. For example, the effect size estimates by VIPRS for LDL cholesterol showed a large enough improvement in performance across ancestries to have potential clinical relevance [75] and make a significant dent in the transferability problem for that trait.

As highlighted throughout the text, we found that many implementation and modeling choices can have a substantial impact on the performance of the VIPRS model in analyses with GWAS summary statistics for real measured traits: hyperparameter tuning strategies, LD estimators, and the prior on the effect size all influenced the predictive performance in ways that varied across phenotypes and experimental setups. This is in addition to the well known challenges faced when there are mismatches between the LD reference panel and the GWAS summary statistics. Overall, in all the setups and experimental conditions that we tested, the grid search approach with spike-and-slab prior and windowed estimator of LD reliably outperformed or rivaled all the other variations of the model as well as previously described PRS methods.

One of the main strengths of the VIPRS model is its computational efficiency, which we exploited to test the predictive performance of the model with approximately 10 million SNPs, almost an order of magnitude greater than the standard HapMap3 subset used to train PRS methods [4, 6, 26, 29]. At this unprecedented scale, we showed that modeling an expanded set of variants results in modest improvements in predictive accuracy for highly polygenic traits, such as standing height and HDL. This is consistent with recent whole-genome sequencing analyses which showed that a considerable proportion of rare causal variants are not well tagged by common SNPs [67].

There are a number of reasons that lead us to believe that the performance metrics that we report here are a lower bound on what could be achieved in modeling large-scale SNP array data. First, the vast majority of the variants that we added beyond the HapMap3 subset are rare and statistically imputed. Rare variant imputation is still a challenging problem and existing algorithms are known to have elevated error rates [76, 77]. We expect that these imputation errors can introduce substantial noise into the PRS estimate, and thus result in decreased predictive accuracy, as we observed for a number of the traits that we analyzed. This difficulty can potentially be addressed by using whole-genome sequencing (WGS) data for GWAS, which may soon be enabled by recent large-scale initiatives by the UKB [37] and TOPMed [3]. Second, residual confounding due to population structure may affect effect size esimation for rare variants [67, 78, 79]. In our GWAS pipeline, we corrected for population stratification using only the top 10 Principal Components (PCs) of the genetic relationship matrix (GRM), which may not adequately capture the more recent demographic history reflected by rare variants [79, 80]. This residual confounding effect may be addressed by increasing the number PCs used in the GWAS analysis [67] or utilizing more genealogically-informed estimates of the GRM [81]. Thirdly, we were only successful in scaling the Variational EM version of the model to the expanded set of variants and, as discussed elsewhere in the text, this VEM-based `VIPRS` can be sub-optimal in some cases. In principle, this can be remedied by using the grid search version of the model, which is demonstrably more accurate in many settings. However, scaling `VIPRS-GS` to this many SNPs proved computationally challenging. Thus, we believe that the predictive ability of the `VIPRS` model with 10 million variants can be further improved with better and more scalable hyperparameter inference techniques, which is an open research direction that we leave for future work.

Despite its competitive predictive ability, we believe that there are a number of modeling choices underlying `VIPRS` that can potentially be improved in future work. Firstly, compared to simulated phenotypes, the generative process for real traits is unknown and likely involves complex and heterogeneous genetic architectures that are not well described by a two-component Gaussian mixture prior. The spike-and-slab prior assumes that all genetic variants have a uniform prior probability of being causal and that the causal SNPs have equal expected contribution to the heritability, which is a simplistic assumption given what is known about the genetic architectures of complex traits [12–14]. This motivated us to explore a more general and flexible Gaussian mixture prior with four mixture components [6]. Our experimental results show that adding mixture components improves accuracy for highly heritable and polygenic traits, such as standing height, but did not systematically improve accuracy for less heritable traits, perhaps because of reduced power to identify the larger number of parameters. Future work using priors informed by functional annotations (e.g. [82, 83]) is a promising avenue to improve accuracy in these cases. Second, our validation analyses in the UKB confirmed that, in general, `VIPRS` as well as other PRS methods do not transfer well across populations or ancestry groups, despite some notable differences between the methods. Recent work has highlighted that transferability in the context of summary statistics-based PRS methods is best achieved when we jointly model the effect sizes of multiple ancestrally homogeneous populations within the same framework [84, 85]. This formulation has proved successful for some Bayesian PRS methods [85] and we believe that a variational approximation to the posterior under such models may prove effective.

Finally, while our results showed that variational approximations to the posterior are a promising alternative to MCMC techniques in predictive settings, it is important to highlight that mean-field variational approaches are known to underestimate the posterior variances and covariances in some cases [39, 86, 87]. In practice, this means that the variational posterior will tend to underestimate the proportion of variance explained (e.g. **Supplementary Fig.**S7 and S8) [44] and may also result in miscalibrated PRS confidence intervals, if such a quantity is sought for some downstream applications [34]. This limitation can be addressed with more expressive variational families [88], such as those derived with variational boosting [89], or alternatively, with help of modern Bayesian inference techniques that combine variational methods and MCMC [90].

4 Methods

4.1 Detailed description of the VIPRS model

We formulate the problem of polygenic prediction in terms of the standard linear model,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2}$$

where $\mathbf{y}$ is an $N \times 1$ vector of phenotypic measurements for N individuals, $\mathbf{X}$ is the $N \times M$ genotype matrix that records the counts of alternative alleles for each individual at each genetic marker, $\boldsymbol{\beta}$ is a vector of effect sizes for each of the M markers, and $\boldsymbol{\epsilon}$ is an $N \times 1$ vector that captures the residual effects on the trait for each individual. For quantitative traits, we assume that the phenotypes follow a Gaussian likelihood, such that $\mathbf{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma_{\epsilon}^2\mathbf{I})$, with σ_{ϵ}^2 being the residual variance. Practically, a central goal of polygenic risk modeling is to arrive at a robust estimate for the effect sizes $\boldsymbol{\beta}$. In the Bayesian framework, this problem is tackled by imposing a prior distribution over the effect sizes and then deriving a solution for the posterior distribution given the data likelihood and the prior,

$$p(\boldsymbol{\beta} \mid \mathbf{X}, \mathbf{y}, \boldsymbol{\theta}) \propto p(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\theta})p(\boldsymbol{\beta}, \boldsymbol{\theta}) \tag{3}$$

where $\boldsymbol{\theta}$ encapsulates all fixed hyperparameters in the model, i.e. parameters that we do not assign a prior. Here, the constant of proportionality is the marginal likelihood or the partition function for the posterior, $\int p(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{s})p(\boldsymbol{\beta}, \boldsymbol{\theta})d\boldsymbol{\beta}$, also known as the model evidence [35,36]. In recent years, considerable work has been devoted to deriving Bayesian PRS models with flexible priors on the effect sizes, such as the continuous shrinkage [26] and mixture priors [6]. In this work, we follow the lead of earlier approaches, e.g. [4,45], and assign a spike-and-slab prior [49,50] on the effect sizes,

$$\beta_j \sim \pi\mathcal{N}(\beta_j; 0, \sigma_{\beta}^2) + (1 - \pi)\delta_0 \tag{4}$$

Here, π is a parameter that denotes the prior probability of a variant being causal, σ_{β}^2 is the prior variance on the effect size of each SNP, and δ_0 is the Dirac delta function. In the simplest formulation of this model, we assume that π and σ_{β}^2 are shared across all SNPs. Thus, π may also be considered as the fraction of variants that are causal for the trait of interest, and the σ_{β}^2 parameter is related to the trait's per-SNP heritability [4,60]. The spike-and-slab prior is a special case of the more general mixture prior:

$$p(\beta_j \mid s_j)p(s_j) = \prod_{k=1}^{K} \mathcal{N}(\beta_j; 0, \sigma_k^2)^{s_{jk}} \prod_{k=1}^{K} \pi_k^{s_{jk}} \tag{5}$$

where s_{jk} is binary indicator for SNP j belonging to the k^{th} mixture component, with π_k and σ_k^2 denoting the the mixing proportion and prior variance for component k, respectively. It is well known that Bayesian linear regression models with a spike-and-slab prior on the effect sizes result in an intractable posterior [45, 49, 50], necessitating the use of approximate posterior inference schemes.

In most of the previous Bayesian PRS formulations, the authors employ a Gibbs sampler, a Markov Chain Monte Carlo (MCMC) technique that relies on conditional conjugacy between the prior and the likelihood, to approximate the posterior distribution of the effect sizes [4,6,26,29]. In this work, we instead leverage a technique known as Variational Inference (VI) [38], which approximates intractable densities by proposing a simple parametric distribution $q(\boldsymbol{\beta}, \mathrm{s})$ and optimizing its parameters to match the true posterior as closely as possible [39]. The closeness between the true posterior and the proposed distribution is measured by the Kullback-Leibler (KL) divergence,

$$\begin{aligned} KL[q||p] &= \mathbb{E}_{q(\boldsymbol{\beta},\mathbf{s})}[\log q(\boldsymbol{\beta}, \mathrm{s})] - \mathbb{E}_{q(\boldsymbol{\beta},\mathbf{s})}[\log p(\boldsymbol{\beta}, \mathrm{s} \mid \mathbf{X}, \mathbf{y}, \boldsymbol{\theta})] && (6) \\ &= \mathbb{E}_{q(\boldsymbol{\beta},\mathbf{s})}[\log q(\boldsymbol{\beta}, \mathrm{s})] - \mathbb{E}_{q(\boldsymbol{\beta},\mathbf{s})}\Big[\log p(\mathbf{y}, \boldsymbol{\beta}, \mathrm{s} \mid \boldsymbol{\theta}, \mathbf{X})\Big] + \log p(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\theta}) && (7) \end{aligned}$$

where $\mathbb{E}_{q(\boldsymbol{\beta},\mathbf{s})}$ is the expectation taken with respect to the proposed distribution [35,36]. However, the KL divergence includes the normalizing constant that made the posterior intractable in the first place. Thus, practitioners typically optimize a surrogate objective known as the Evidence Lower BOund (ELBO) of the log marginal likelihood [35,36,39]:

$$\begin{aligned} \mathcal{L} &= \log p(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\theta}) = \log \int \frac{p(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\theta})p(\boldsymbol{\beta}, \mathrm{s} \mid \boldsymbol{\theta})}{q(\boldsymbol{\beta}, \mathrm{s})} q(\boldsymbol{\beta}, \mathrm{s}) d\boldsymbol{\beta} \\ &\geq \int q(\boldsymbol{\beta}, \mathrm{s}) \log p(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\beta}, \mathrm{s}, \boldsymbol{\theta}) p(\boldsymbol{\beta}, \mathrm{s} \mid \boldsymbol{\theta}) d\boldsymbol{\beta} - \int q(\boldsymbol{\beta}, \mathrm{s}) \log q(\boldsymbol{\beta}, \mathrm{s}) d\boldsymbol{\beta} \\ &= \mathbb{E}_{q(\boldsymbol{\beta},\mathbf{s})}\Big[\log p(\mathbf{y}, \boldsymbol{\beta}, \mathrm{s} \mid \boldsymbol{\theta}, \mathbf{X})\Big] - \mathbb{E}_{q(\boldsymbol{\beta},\mathbf{s})}\Big[\log q(\boldsymbol{\beta}, \mathrm{s})\Big] \equiv ELBO && (8) \end{aligned}$$

Here, the first term in (8) is the expectation of the log joint likelihood of the phenotypes and the effect sizes and the second term corresponds to the negative entropy of the variational distribution. The ELBO (8) and the KL-Divergence in (7) add up to the marginal likelihood: $ELBO + KL[q||p] = \mathcal{L}$. Therefore, maximizing ELBO is equivalent to minimizing the KL-Divergence [35, 36, 39].

The choice of approximating variational distribution $q(\boldsymbol{\beta}, \mathrm{s})$ is a central component in this setting. For simplicity and computational efficiency, we make use of the paired mean-field assumption [35, 36, 52], whereby the density factorizes across the input coordinates, and model the effect size at each locus with a two component Gaussian mixture density [45, 52, 64],

$$q(\boldsymbol{\beta}, \mathbf{s}) = \prod_{j}^{M} q(\beta_j, s_j) = \prod_{j}^{M} \mathcal{N}(\beta_j; \mu_j, \sigma_j^2) Bern(s_j; \gamma_j) \tag{9}$$

Here, $\mu_j, \sigma_j^2, \gamma_j$ are the variational parameters defined for each variant in the dataset and $Bern(s_j; \gamma_j) = \gamma_j^{s_j}(1-\gamma_j)^{1-s_j}$ denotes a Bernoulli distribution with probability γ_j for SNP j. Therefore, the Bernoulli indicator in the proposed distribution approximates the posterior probability that the variant is causal for the trait of interest and the Gaussian component approximates the posterior for the effect size [45]. In the **Supplementary Information**, we provide detailed derivations which show that, under certain assumptions, this variational family leads to the following closed form updates that only depend on GWAS summary statistics and the SNP-by-SNP correlation or Linkage Disequilibrium (LD) matrix, which can be derived from an appropriately-matched reference panel:

$$\sigma_j^2 = \frac{\sigma_\epsilon^2}{N + \sigma_\epsilon^2/\sigma_\beta^2} \tag{10}$$

$$\mu_j = \frac{N\sigma_j^2}{\sigma_\epsilon^2}\Big(\hat{\beta}_j - \sum_{k \neq j} \gamma_k \mu_k R_{jk}\Big) \tag{11}$$

$$\gamma_j = Sigmoid\Big(\log(\frac{\pi}{1-\pi}) + \frac{1}{2} log(\sigma_j^2/\sigma_\beta^2) + \frac{1}{2\sigma_j^2}\mu_j^2\Big) \tag{12}$$

Here, $\hat{\beta}_j = \mathbf{y}^\top \mathbf{x}_j / N$ is the standardized marginal GWAS effect size and $\boldsymbol{R}$ is the LD matrix. This formulation enables us to employ a fast coordinate ascent algorithm to approximate the posterior distributions of the effect sizes.

In addition, to perform inference given some of the unknown fixed parameters, i.e. $\theta = (\pi, \sigma_\beta^2, \sigma_\epsilon^2)$, the basic formulation of the VIPRS model uses the Variational Expectation-Maximization (VEM) algorithm [44, 45, 64], where, in an alternating fashion, in the E-Step we update the variational parameters given the hyperparameters and in the M-Step we update the hyperparameters of the model. In both the E- and M-steps, we update the free parameters of the model to maximize our objective, the ELBO. In **Supplementary Information**, we show that updating the hyperparameters to maximize the ELBO also results in closed-form solutions for those parameters.

Despite the conceptual simplicity of the model described above, fitting such a model to biobank-scale data presents several computational challenges. For instance, the closed-form update equations for some of the variational parameters involve terms that relate to the LD between the focal variant and all other variants in the genome. This can be computationally prohibitive to compute for millions of variants and for hundreds of EM iterations. To overcome this, we follow the lead of other summary-statistics-based PRS methods and use a banded or shrunk LD matrix [4, 6, 26, 29], which results in substantial improvements in speed without substantially degrading predictive performance.

4.2 Hyperparameter tuning strategies

The standard VIPRS model employs a Variational EM framework to infer the hyperparameters $\theta = (\pi, \sigma^2_\beta, \sigma^2_\epsilon)$, where in the M-step, we update each hyperparameter to maximize the surrogate objective, i.e. the ELBO [45, 64]. This strategy works well in many settings, but it is prone to entrapment in local optima [64], which may degrade overall predictive performance of the model. In this work, we explored three alternative strategies for tuning the hyperparameters of the model.

`VIPRS-GS` In the first strategy, we performed grid search (GS) over the hyperparameters of the model, selecting the values that result in the best predictive performance on a held-out validation set [25, 27, 29]. The grid search was performed specifically over the proportion of causal variants π, with the remaining parameters updated according to their maximum likelihood estimates. The grid for π ranged from $\frac{1}{M}$ to $\frac{M-1}{M}$ with 30 equidistant values on a $\log_{10}$ scale, where M is the number of variants included in the model.

`VIPRS-BO` To search over the hyperparameter space without the constraint of a predefined and discrete grid, we experimented with a second hyperparameter tuning technique known as Bayesian Optimization (BO) [55]. In Bayesian Optimization, we assume that there is an underlying unknown function $f(\theta)$ that takes the hyperparameters as input and outputs a certain score that we wish to optimize, such as the training ELBO or the validation R^2. This unknown function is modeled with a Gaussian Process (GP) prior, which allows us to explore the parameter space efficiently while accounting for uncertainty in a principled manner. The other component in this framework is the acquisition function, a heuristic that maps from the GP posterior to information about the most promising regions in hyperparameter space [55, 91]. In our experiments, we used the `scikit-optimize` python package to perform this optimization, with `gp_hedge` as the default acquisition function. The optimizer was allowed to sequentially evaluate up to 20 points in a bounded $1D$ space for the hyperparameter π.

`VIPRS-BMA` In the third strategy, we used a Bayesian Model Averaging (BMA) framework, where we use importance sampling to integrate out some of the hyperparameters of the model, as outlined in [45]. The main idea here is that instead of fixing π to a particular value, we fit the `VIPRS` model along a grid of π values, as in `VIPRS-GS` , and then take a weighted average of the effect size estimates for each SNP based on each model's ELBO [44, 45].

Similar to previous work in this area, we note that these three strategies can be deployed in conjunction with the VEM framework [44, 45, 92], where some of the hyperparameters are updated using their maximum likelihood estimates while the remaining parameters are optimized via the user's strategy of choice. This is important in practice, because with the three hyperparameters of the model, an exhaustive search will require searching over a three-dimensional grid, which can be computationally expensive. Therefore, in our experiments and analyses, for all of the three strategies that we explored, we only implemented a search over the fraction of causal variants π and estimated the other two hyperparameters using the closed-form updates in the M-Step.

4.3 Data preprocessing

To assess the performance of VIPRS on a biobank-scale dataset, we made use of the UK Biobank (UKB), a large database of genomic and phenotypic measurements from 488,377 participants from the United Kingdom [1]. In its latest release, the UKB database has genotype information from 488,377 individuals, from which, after applying standard quality control procedures, we retained data for a total of 337,205 samples. Briefly, the sample quality controls involved selecting unrelated individuals with White British ancestry, defined by the UKB based on self-reported ethnic background as well as Principal Components Analysis (PCA) of the GRM, and who were also included in the PCA and phasing procedures outlined in [1]. We restricted our main analyses to the White British cohort in order to maximize power to detect causal effects while reducing confounding. In addition this, we filtered data for individuals with detected sex chromosome aneuploidy, excess relatedness, or missing genotype rate exceeding 5% from this analysis.

The genetic variants or SNPs included in the study were selected based on a number quality control filters, applied at various stages in the analysis. For the base dataset, we excluded variants with duplicate rsIDs, ambiguous strand, imputation quality score $<$ 0.3, Hardy-Weinberg Equilibrium p-value $< 10^{-10}$, or genotype missingness rate $>$ 0.05. We also removed multi-allelic variants as well as SNPs in long-range LD regions, as specified in Supplementary Table 13 in Bycroft et al. 2018 [1]. In the GWAS analyses or LD matrix construction, we further filtered variants with minor allele count (MAC) < 5 or minor allele frequency (MAF) < 0.1%. This resulted in a total of 9,590,026 bi-allelic variants that were used in the expanded SNP set analyses. Finally, following standard practice in PRS methodologies [6, 29], for the base analyses with the VIPRS model, we restricted to the set of variants in the HapMap3 reference panel [32], resulting in a total of 1,093,308 SNPs. Most of these quality control procedures were carried out with the genetic analysis software tool plink2 [93].

4.4 Construction and efficient representation of LD matrices

An important quantity in the model is the Linkage Disequilibrium (LD) or SNP-by-SNP correlation matrix $\boldsymbol{R}$. The matrix, or its columns, show up mainly in the update equations for the variational parameters μ_j of each SNP j (11), the estimate of the residual variance σ_ϵ^2, as well as in the objective function (i.e., ELBO) (**Supplementary Information**). Our software supports a number of different LD matrix estimators, including sample, block, shrinkage, and windowed estimators.

Sample estimator In the sample estimator, we estimate the sample Pearson correlation coefficient between all SNPs on the same chromosome, which results in a dense matrix. For larger chromosomes and SNP sets, it is impractical to load dense matrices of this scale to memory. To handle data at that scale, we use compressed and chunked on-disk storage with `Zarr` arrays in `python` for fast, multi-threaded read and write access. Then, as we iterate through SNPs in the E-Step, we load the matrix into memory one-chunk at a time, thus allowing us to train `VIPRS` with extremely large LD matrices. In **Supplementary Information**, we describe a procedure that allows us to load the LD matrix only once per iteration, resulting in improved speed and efficiency.

Block estimator In the block LD estimator, we only estimate the sample LD between SNPs that are within the same LD block, as defined by, e.g. LDetect [62]. This is similar to what is done in the `Lassosum` and `PRScs` frameworks [25, 26].

Shrinkage estimator In the shrinkage estimator, we shrink and threshold the entries of the sample LD matrix according to procedure outlined by [6, 61] and implemented in the `gctb` software. Briefly, for the shrinkage estimator, we shrink each element of the LD matrix by a quantity proportional to the distance between pairs of variants j and k in along the chromosome: $\hat{R}_{jk} = R_{jk} \cdot e^{c \cdot d(j,k)}$. In this context, $d(j,k)$ is the distance in centi Morgans between variants j and k and the constant c is related to sample size used to infer the genetic map as well as effective population size [6, 61].

Windowed estimator For the windowed LD estimator, we only consider the correlation between a focal variant with variants that are at most 3 centi Morgan away from it along the chromosome [29, 60]. This estimator results in compact and banded LD matrices that can easily fit in memory on modern compute nodes.

To construct LD matrices for the main analyses of this paper, we selected a random subset of 50,000 individuals from the White British cohort described above. Within that group of individuals, we filtered SNPs with minor allele count (MAC) < 5 or minor allele frequency (MAF) < 0.1%, and again restricted to variants in the HapMap3 reference panel. For the analyses with the expanded set of variants, we only removed the HapMap3 filter. Unless explicitly stated otherwise, the analyses with the `VIPRS` model employed the windowed estimator for LD, with the distnace cutoff set to 3cM. The matrices are stored in compressed `Zarr` array format and are publicly available for download (see **Data Availability**).

4.5 Simulation study

To assess the predictive performance of `VIPRS` on large-scale datasets and for varying genetic architectures, we conducted a genome-wide simulation study using the pre-processed genotype data from the UK Biobank cohort. We simulated quantitative phenotypes according to the generative model outlined in Equations 2 and 4, with 9 different configurations corresponding to three settings of the proportion of causal variants, $\pi = \{10^{-4}, 10^{-3}, 10^{-2}\}$, and three settings of the heritability, $h_g^2 = \{0.1, 0.3, 0.5\}$. For each configuration, we generated 10 traits, for a total of 90 simulated traits. To simulate binary traits, we followed the same procedure, but used the liability threshold model [56] to obtain case-control status, with prevalence set to 15%.

After we generated simulated phenotypes for all individuals in the study ($N = 337,205$), we excluded the 50,000 samples used to generate the LD matrices and randomly split the remaining samples into 70% training ($N = 201,043$), 15% validation, and 15% testing ($N = 43,081$ each). The genotype and simulated phenotype data of the training samples were then used to generate GWAS summary statistics using `plink2` [93].

4.6 Application to real traits from the UKB

To assess the predictive performance of `VIPRS` on real phenotypes, we extracted phenotypic measurements for 9 quantitative and 3 case-control traits for the UKB cohort described previously. The quantitative phenotypes included log-transformed Waist Circumference (WC), log-transformed Hip Circumference (HC), Standing Height (HEIGHT), Birth Weight (BW), log-transformed Body Mass Index (BMI), log-transformed High-Density Lipoprotein (HDL), Low-Density Lipoprotein (LDL), Forced Vital Capacity (FVC), and Forced Expiratory Volume in the first second (FEV1). For each trait, we excluded samples with outlier or extreme values for the trait. For the remaining samples, within each sex separately, we corrected for age and the top 10 Principal Components (PCs) of the genetic relationship matrix (GRM), and then applied a Rank-based Inverse Normal Transform (RINT) on the residuals [94]. To assess the predictive performance on held-out test sets, we performed 5-fold cross-validation. For each split, the training data was further split into 90% training and 10% validation to facilitate running PRS methods that require a validation set to tune their hyperparameters.

The case-control phenotypes included in the analysis are Asthma (prevalence 12.7%), Type 2 Diabetes (T2D) (prevalence 2.3%), and Rheumatoid arthritis (RA) (prevalence 1.7%). To assess the predictive performance on held-out test sets, we performed stratified 5-fold cross validation, followed by splitting the training data into 90% training and 10% validation in a stratified manner to keep the prevalence approximately the same for all subsets of the data.

The phenotypes and associated sample sizes in the UK Biobank are listed in Table 1. The detailed scripts with the extraction and transformation procedure for each phenotype is included in the public repository associated with this publication (**Code Availability**). The 5-fold cross validation procedure was performed using the `scikit-learn` package in `python` [95].

4.7 Validation in minority populations in the UKB

To validate the relative predictive ability of `VIPRS` in individuals of different backgrounds, we used the approach of Prive et al. [59] to identify subgroups of relatively uniform ancestry and ethnicity. Using self-reported ethnic background as well as PCA medoids from [59], we extracted genotype data for individuals of Italian ($N = 6177$), Indian ($N = 6011$), Chinese ($N = 1769$), and Nigerian ($N = 3825$) ancestry. In genetic analyses, those ancestries groups show various levels of allele frequency differentiation (Fst) when compared to the White British cohort [59]. These individuals were selected after applying the same quality control filters as before. Mainly, we retained individuals who were used in the PCA and phasing procedures and filtered samples with detected sex chromosome aneuploidy, excess relatedness, or missing genotype rate exceeding 5% from this analysis.

For each individual in those target populations, we extracted phenotype data for the traits analyzed previously (**Table**1). Then, we used effect size estimates derived from the 5-fold analyses on the White British cohort to generate polygenic scores for individuals in those minority populations. Given these polygenic score estimates, we computed the relative prediction R^2 as the incremental R^2 in the target population divided by the R^2 of the best performing PRS model on the test set in the White British cohort. This metric is designed to highlight the transferability of PRS estimates across different population and ancestry groups.

4.8 PRS method comparison

To compare the predictive performance of VIPRS to state-of-the-art models for polygenic risk prediction with summary statistics, we included a diverse collection of methods with different assumptions and implementations, including stochastic Bayesian methods (`SBayesR` [6], `PRScs` [26], and `LDPred2` [29]), a penalized regression method (`Lassosum` [25]), as well as a C+T method (`PRSice2` [27]).

For each method, we provided the GWAS summary statistics for the simulated and real phenotypes and ran the model with default or recommended settings. For models that require a validation set to tune the hyperparameters, we provided a held-out validation subset of the data. Once the models converge, we output the effect size estimates and then generated polygenic scores for the samples in the test set. Given these polygenic scores, the models were then evaluated for the quality of their predictions.

For quantitative traits, we reported the incremental prediction R^2, defined as the R^2 of a linear model with the PRS and covariates (age, sex, and top 10 PCs) minus the R^2 obtained from a linear model with the covariates alone. For case-control phenotypes, we reported the area under the Precision-Recall curve (AUPRC) between the polygenic score and the binary phenotype. In addition to these predictive metrics, we also compared the run-time of the different methods to gauge their scalability and computational efficiency.

The detailed specification of priors, grid values, and other hyperparameters for each PRS model is shown in the repository accompanying this manuscript (see **Code Availability**).

4.9 Software implementation

The data structures and inference algorithms for the `VIPRS` model are implemented in two `python` packages that are open source and publicly available on GitHub (see **Code Availability**). The first software package, `magenpy`, implements scripts and routines for computing LD matrices and transforming them to `Zarr` array format, simulating complex traits from genotype data, and harmonizing multiple genetic data sources, such as GWAS summary statistics, LD reference panels, functional annotations, etc. The second software package, `viprs`, implements the optimized variational inference algorithms to obtain posterior estimates for the effect sizes. For optimal speed and efficiency, the coordinate ascent routine is written in `cython`, a `python` extension that allows for compiled `python` code by transforming it into the `C` programming language [96]. Both packages follow object-oriented design principles to allow for streamlined user extensions and experimentation by experienced developers. We also provide runner scripts that allow users to perform inference using command-line interfaces.

Code Availability

- The code to process GWAS summary statistics, construct LD matrices, and provide necessary data structures for the `VIPRS` model is publicly available on GitHub: `https://github.com/shz9/magenpy`
- Code to run the `VIPRS` model and perform posterior inference is available via GitHub: `https://github.com/shz9/viprs`
- Scripts to replicate all the analyses and generate the figures for this manuscript are available via GitHub: `https://github.com/shz9/viprs-paper`

Data Availability

- The external GWAS summary statistics were downloaded from the LD-Score regression repository: `https://alkesgroup.broadinstitute.org/LDSCORE/all_sumstats/`
- Pre-computed LD matrices in `Zarr` format are available for download via Zenodo: `https://doi.org/10.5281/zenodo.6529229`

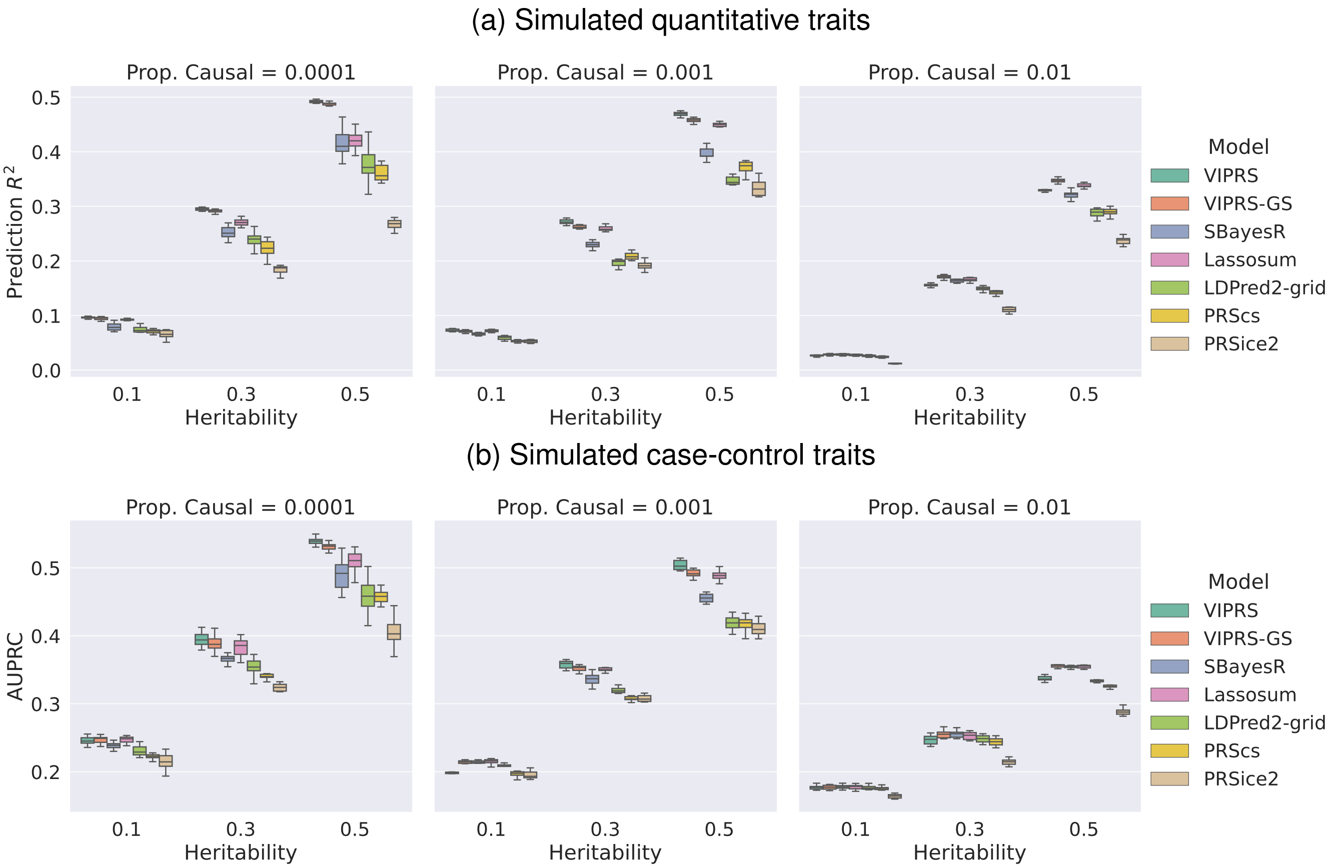
Acknowledgements

We thank members of the Li and Gravel labs for useful feedback and discussions on this project. We are particularly grateful to Doruk Cakmakci and Goodarz Koli Farhood for testing various aspects of the software. We also thank Hans Markus Munter, Stephen Sawcer, and Mark Lathrop for help with data access and Celia Greenwood and Mathieu Blanchette for feedback on the project at various stages. The UK Biobank analyses in this work were conducted under Application 4408. Y.L. is supported by Natural Sciences and Engineering Research Council (NSERC) Discovery Grant (RGPIN-2019-0621), Fonds de recherche Nature et technologies (FRQNT) New Career (NC-268592), and Canada First Research Excellence Fund Healthy Brains for Healthy Life (HBHL) initiative New Investigator start-up award (G249591).

Author contributions

S. G. and Y. L. jointly supervised this work. Y.L. conceived the study. Y.L. and S.Z. developed the methodology. S.Z. implemented the computational software and performed all the experiments. S.Z., Y.L., and S.G. analyzed the results. S.Z. wrote the initial manuscript. All of the authors wrote the final version.

5 Figures



(a) Simulated quantitative traits

(b) Simulated case-control traits

Figure 1: Predictive performance of summary statistics-based PRS methods on simulated **(a)** quantitative and **(b)** case-control phenotypes (15% prevalence). The phenotypes were simulated using real genotype data from the White British cohort in the UK Biobank ($N = 337,225$), leveraging a subset of 1.1 million HapMap3 variants. The simulation scenarios encompass a total of 9 genetic architectures, spanning 3 values for polygenicity (proportion of causal variants) and 3 values for SNP heritability. For each configuration, we simulated 10 independent phenotypes. Each panel shows results for phenotypes simulated with the pre-specified polygenicity and each column within a panel shows performance metrics for phenotypes simulated with a pre-specified SNP heritability. The performance metrics are **(a)** incremental prediction R^2 for quantitative traits and **(b)** area under the Precision Recall curve (AUPRC) for binary traits. The boxplot for each method and simulation configuration shows the quartiles of predictive scores for the 10 simulated phenotypes. The PRS methods shown are our proposed `VIPRS` and `VIPRS-GS` (using grid search to tune model hyperparameters) as well as 5 other baseline models: `SBayesR`, `Lassosum`, `LDPred2` (grid), `PRScs`, and `PRSice2` (C+T).

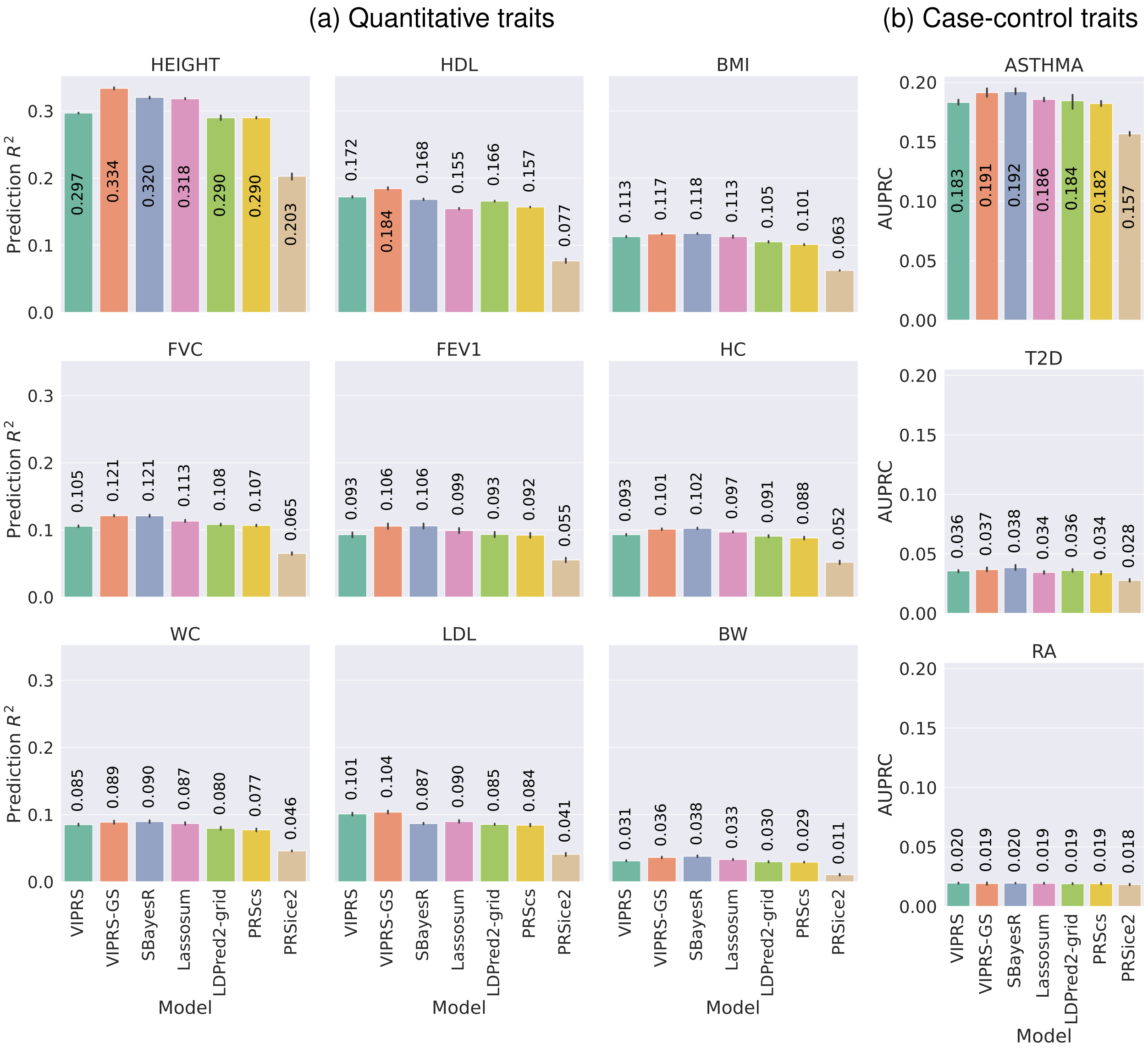


Figure 2: Predictive performance of summary statistics-based PRS methods on real **(a)** quantitative and **(b)** case-control phenotypes in the UK Biobank. The measured phenotypes were pre-processed and analyzed in a 5-fold cross-validation study design and the prediction metrics show the performance of each PRS method in predicting the phenotype in a held-out test set. Each panel shows the predictive performance, in terms of **(a)** incremental R^2 and **(b)** area under the Precision Recall curve (AUPRC), of various PRS methods when applied to a given phenotype. The bars show the mean of the prediction metrics across the 5 folds and the black lines show the corresponding standard errors. The quantitative phenotypes analyzed are standing height (HEIGHT), high-density lipoprotein (HDL), body mass index (BMI), forced vital capacity (FVC), forced expiratory volume in 1 second (FEV1), hip circumference (HC), waist circumference (WC), low-density lipoprotein (LDL) and birth weight (BW). The binary phenotypes analyzed are asthma (ASTHMA), type 2 diabetes (T2D) and Rheumatoid arthritis (RA). The PRS methods shown are our proposed `VIPRS` and `VIPRS-GS` (using grid search to tune model hyperparameters) as well as 5 other baseline models: `SBayesR`, `Lassosum`, `LDPred2` (grid), `PRScs`, and `PRSice2` (C+T).

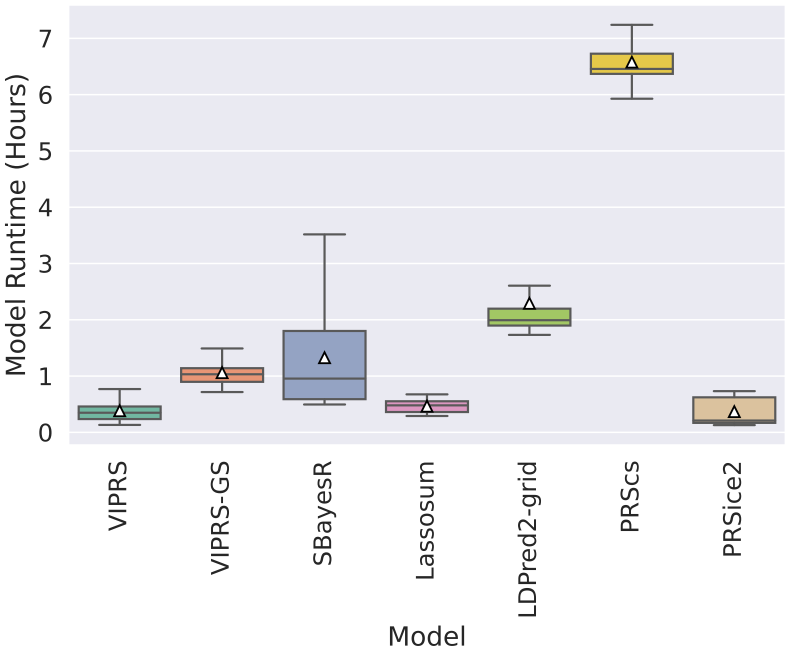


Figure 3: The total runtime (in hours) of the summary statistics-based PRS methods included in the study. The boxplot for each method shows the quartiles of the runtime from a total of 240 independent experiments (180 simulated traits plus 60 real experiments, comprising the 12 real phenotypes analyzed multiplied by the 5 training folds). The white triangles indicates the mean runtime for each method. The PRS methods shown are our proposed `VIPRS` and `VIPRS-GS` (using grid search to tune model hyperparameters) as well as 5 other baseline models: `SBayesR`, `Lassosum`, `LDPred2` (grid), `PRScs`, and `PRSice2` (C+T).

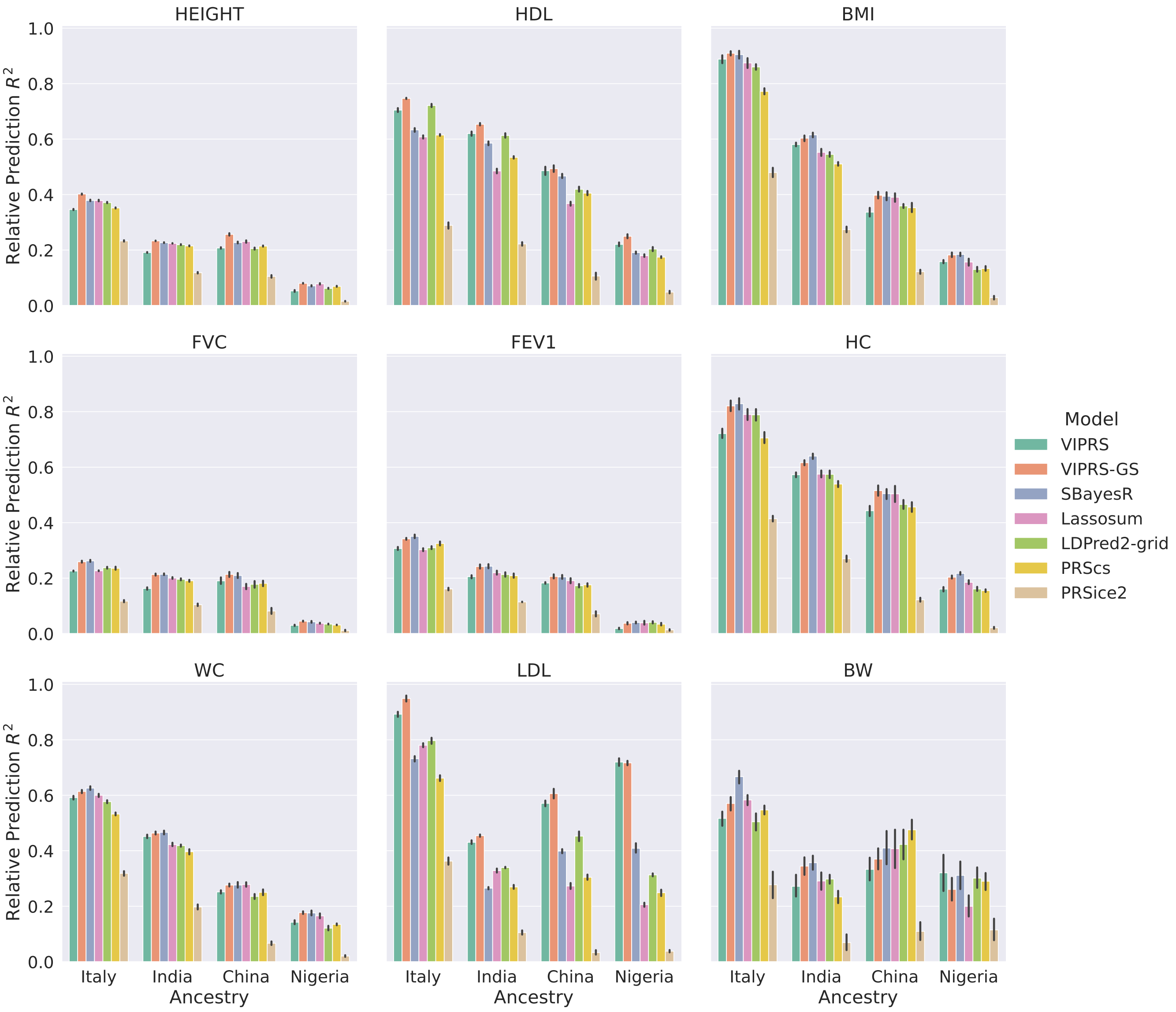


Figure 4: Relative predictive performance of summary statistics-based PRS methods on real quantitative phenotypes in minority populations in the UK Biobank. The PRS models were trained on summary statistics from the White British cohort in the UK Biobank using a 5-fold cross validation design. Then, the effect size estimates from the five training folds were used to perform predictions in individuals of Italian, Indian, Chinese, and Nigerian ancestry. Each panel shows the incremental prediction R^2 in a given ancestry group relative to the prediction R^2 of the best performing model on the White British cohort. The bars show the mean of the relative prediction metric across the 5 training folds and the black lines show the corresponding standard errors. The quantitative phenotypes analyzed are standing height (HEIGHT), high-density lipoprotein (HDL), body mass index (BMI), forced vital capacity (FVC), forced expiratory volume in 1 second (FEV1), hip circumference (HC), waist circumference (WC), low-density lipoprotein (LDL) and birth weight (BW). The PRS methods shown are our proposed `VIPRS` and `VIPRS-GS` (using grid search to tune model hyperparameters) as well as 5 other baseline models: `SBayesR`, `Lassosum`, `LDPred2` (grid), `PRScs`, and `PRSice2` (C+T).

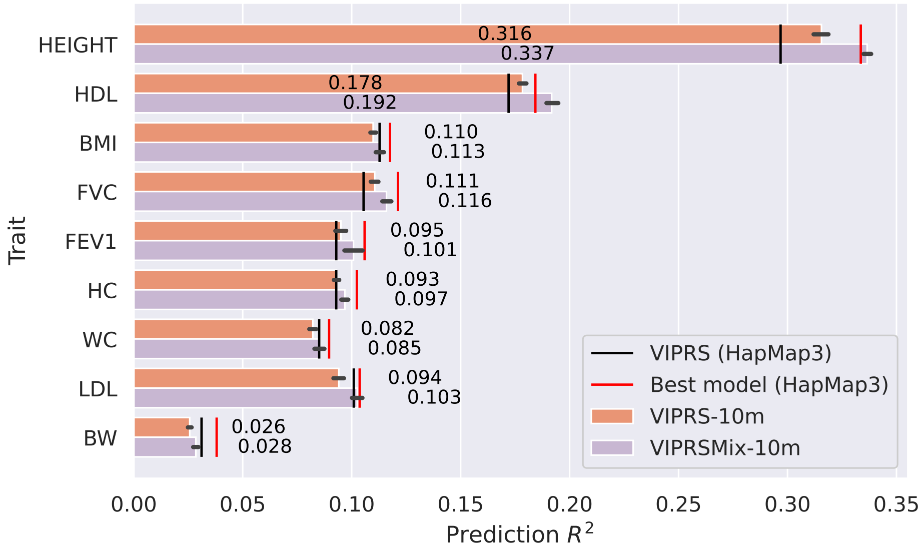


Figure 5: Predictive performance of the VIPRS method on real quantitative traits in the UK Biobank using up to 10 million genotyped and imputed SNPs. Each horizontal bar shows the predictive performance, in terms of incremental Prediction R^2, of two versions of the `VIPRS` model, each incorporating GWAS summary statistics data from roughly 9.6 million SNPs: `VIPRS-10m` (orange) is `VIPRS` that uses VEM to update all of the hyperparameters, whereas `VIPRSMix-10m` (purple) is an extended version of the model employing a 4-component sparse mixture prior on the effect sizes. The thick black horizontal lines show the standard errors across the 5-folds from the 5-fold cross-validation scheme. For reference, the figure shows the mean Prediction R^2 for the standard `VIPRS` model with the ≈ 1.1 million HapMap3 SNPs only (black vertical lines) as well as the best performing model using the HapMap3 SNPs only (red vertical lines). The quantitative phenotypes analyzed are standing height (HEIGHT), high-density lipoprotein (HDL), body mass index (BMI), forced vital capacity (FVC), forced expiratory volume in 1 second (FEV1), hip circumference (HC), waist circumference (WC), low-density lipoprotein (LDL) and birth weight (BW).

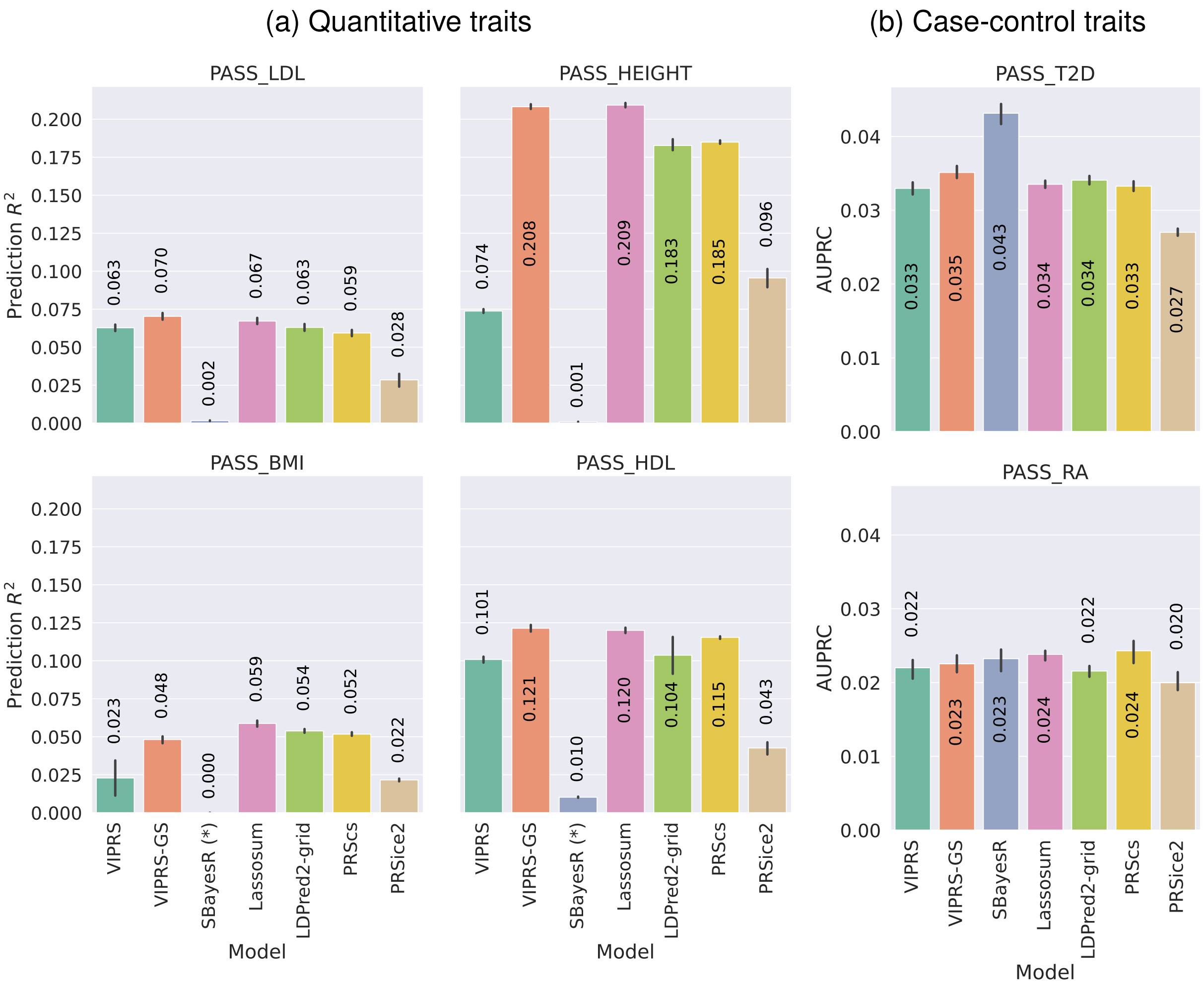


Figure 6: Predictive performance of summary statistics-based PRS methods on real **(a)** quantitative and **(b)** case-control phenotypes using external, publicly available GWAS summary statistics (PASS). Each panel shows the predictive performance, in terms of **(a)** incremental R^2 and **(b)** area under the Precision Recall curve (AUPRC), of various PRS methods when applied to an independent test cohort in the UK Biobank. The bars show the mean and standard error of the prediction metrics across the 5 folds and the black lines show the corresponding standard errors. The quantitative phenotypes analyzed are standing height (PASS_HEIGHT), high-density lipoprotein (PASS_HDL), low-density lipoprotein (PASS_LDL). The binary phenotypes analyzed are type 2 diabetes (PASS_T2D) and Rheumatoid arthritis (PASS_RA). The PRS methods shown are our proposed `VIPRS` and `VIPRS-GS` (using grid search to tune model hyperparameters) as well as 5 other baseline models: `SBayesR`, `Lassosum`, `LDPred2` (grid), `PRScs`, and `PRSice2` (C+T). The asterisk (*) next to the `SBayesR` method in panel **(a)** is to indicate that it did not converge on those traits.

6 Tables

Phenotype	Description	GWAS Source	GWAS sample size	Validation sample size	Test sample size
HEIGHT	Standing height	UKB	242,213	26,913	67,282
HDL	high-density lipoprotein	UKB	211,856	23,540	58,849
BMI	Body mass index	UKB	241,959	26,885	67,211
FVC	Forced Vital Capacity	UKB	221,249	24,584	61,459
FEV1	Forced Expiratory Volume in 1 second	UKB	221,265	24,586	61,463
HC	Hip circumference	UKB	242,311	26,924	67,309
WC	Waist circumference	UKB	242,340	26,927	67,317
LDL	low-density lipoprotein	UKB	230,995	25,667	64,166
BW	Birth weight	UKB	138,300	15,367	38,417
ASTHMA	Asthma	UKB	229,031	25,448	63,620
T2D	Type 2 diabetes	UKB	235,937	26,216	65,538
RA	Rheumatoid arthritis	UKB	186,239	20,694	51,734
PASS_HEIGHT	Standing height	[70]	131,547	26,913	67,282
PASS_HDL	High-density lipoprotein	[71]	97,749	23,540	58,849
PASS_BMI	Body mass index	[72]	122,033	26,885	67,211
PASS_LDL	Low-density lipoprotein	[71]	93,354	25,667	64,166
PASS_T2D	Type 2 diabetes	[73]	60,786	26,216	65,538
PASS_RA	Rheumatoid arthritis	[74]	37,681	20,694	51,734

Table 1: The list of real phenotypes and GWAS data sources analyzed in this study. With each phenotype code, we provide the full name and description, the GWAS data source (UKB or external), as well as the sample sizes for the training, validation, and test sets. The sample sizes for each subset may vary slightly across the 5 folds. We prepended some of the phenotype codes with PASS to indicate that the GWAS summary statistics are external to the UK Biobank and are publicly available. For analyses with the external GWAS summary statistics, the validation and test sets come from the UK Biobank.

References

[1] Bycroft, C. *et al.* The uk biobank resource with deep phenotyping and genomic data. *Nature* **562** (2018).

[2] Kanai, M. *et al.* Genetic analysis of quantitative traits in the japanese population links cell types to complex human diseases. *Nature Genetics* **50** (2018).

[3] Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the nhlbi topmed program. *Nature* **590** (2021).

[4] Vilhjálmsson, B. J. *et al.* Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *American Journal of Human Genetics* **97** (2015).

[5] Torkamani, A., Wineinger, N. E. & Topol, E. J. The personal and clinical utility of polygenic risk scores. *Nature Reviews Genetics* **19** (2018).

[6] Lloyd-Jones, L. R. *et al.* Improved polygenic prediction by bayesian multiple regression on summary statistics. *Nature Communications* **10** (2019).

[7] Lewis, C. M. & Vassos, E. Polygenic risk scores: From research tools to clinical instruments. *Genome Medicine* **12** (2020).

[8] Choi, S. W., Mak, T. S. H. & O'Reilly, P. F. Tutorial: a guide to performing polygenic risk score analyses. *Nature Protocols* **15** (2020).

[9] O'Connor, L. J. *et al.* Extreme polygenicity of complex traits is explained by negative selection. *American Journal of Human Genetics* **105** (2019).

[10] Zeng, J. *et al.* Widespread signatures of natural selection across human complex traits and functional genomic categories. *Nature Communications* **12** (2021).

[11] Johnson, R. *et al.* Estimation of regional polygenicity from gwas provides insights into the genetic architecture of complex traits. *PLoS Computational Biology* **17** (2021).

[12] Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature Genetics* **47** (2015).

[13] Gazal, S. *et al.* Linkage disequilibrium–dependent architecture of human complex traits shows action of negative selection. *Nature Genetics* **49**, 1421–1427 (2017). URL `https://doi.org/10.1038/ng.3954`.

[14] Speed, D., Holmes, J. & Balding, D. J. Evaluating and improving heritability models using summary statistics. *Nature Genetics* **52** (2020).

[15] Chatterjee, N., Shi, J. & García-Closas, M. Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nature Reviews Genetics* **17** (2016).

[16] Hivert, V. *et al.* Estimation of non-additive genetic variance in human complex traits from a large sample of unrelated individuals. *American Journal of Human Genetics* **108** (2021).

[17] Palmer, D. S. *et al.* Analysis of genetic dominance in the uk biobank. *bioRxiv* (2022). URL `https://www.biorxiv.org/content/early/2022/01/14/2021.08.15.456387`. `https://www.biorxiv.org/content/early/2022/01/14/2021.08.15.456387.full.pdf`.

[18] Lambert, S. A., Abraham, G. & Inouye, M. Towards clinical utility of polygenic risk scores. *Human Molecular Genetics* **28** (2019).

[19] Hao, L. *et al.* Development of a clinical polygenic risk score assay and reporting workflow. *Nature Medicine* (2022). URL `https://doi.org/10.1038/s41591-022-01767-6`.

[20] Khera, A. V. *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature Genetics* **50** (2018).

[21] Dai, J. *et al.* Identification of risk loci and a polygenic risk score for lung cancer: a large-scale prospective cohort study in chinese populations. *The Lancet Respiratory Medicine* **7** (2019).

[22] Sugrue, L. P. & Desikan, R. S. What are polygenic scores and why are they important? *JAMA - Journal of the American Medical Association* **321** (2019).

[23] Natarajan, P. *et al.* Polygenic risk score identifies subgroup with higher burden of atherosclerosis and greater relative benefit from statin therapy in the primary prevention setting. *Circulation* **135** (2017).

[24] Moser, G. *et al.* Simultaneous discovery, estimation and prediction analysis of complex traits using a bayesian mixture model. *PLoS Genetics* **11** (2015).

[25] Mak, T. S. H., Porsch, R. M., Choi, S. W., Zhou, X. & Sham, P. C. Polygenic scores via penalized regression on summary statistics. *Genetic Epidemiology* **41** (2017).

[26] Ge, T., Chen, C. Y., Ni, Y., Feng, Y. C. A. & Smoller, J. W. Polygenic prediction via bayesian regression and continuous shrinkage priors. *Nature Communications* **10** (2019).

[27] Choi, S. W. & O'Reilly, P. F. Prsice-2: Polygenic risk score software for biobank-scale data. *GigaScience* **8** (2019).

[28] Qian, J. *et al.* A fast and scalable framework for large-scale and ultrahigh-dimensional sparse regression with application to the uk biobank. *PLOS Genetics* **16**, 1–30 (2020). URL `https://doi.org/10.1371/journal.pgen.1009141`.

[29] Privé, F., Arbel, J. & Vilhjálmsson, B. J. Ldpred2: Better, faster, stronger. *Bioinformatics* **36** (2020).

[30] Zhou, G. & Zhao, H. A fast and robust bayesian nonparametric method for prediction of complex traits using summary statistics. *PLoS Genetics* **17** (2021).

[31] Mitt, M. *et al.* Improved imputation accuracy of rare and low-frequency variants using population-specific high-coverage wgs-based imputation reference panel. *European Journal of Human Genetics* **25** (2017).

[32] Altshuler, D. M. *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* **467** (2010).

[33] Pasaniuc, B. & Price, A. L. Dissecting the genetics of complex traits using summary association statistics. *Nature Reviews Genetics* **18** (2017).

[34] Ding, Y. *et al.* Large uncertainty in individual polygenic risk score estimation impacts prs-based risk stratification. *Nature Genetics* **54**, 30–39 (2022). URL `https://doi.org/10.1038/s41588-021-00961-5`.

[35] Bishop, C. M. Bishop - pattern recognition and machine learning - springer 2006 **58** (2014).

[36] Murphy, K. P. *Probabilistic Machine Learning: An Introduction* (2012).

[37] Halldorsson, B. V. *et al.* The sequences of 150,119 genomes in the uk biobank. *bioRxiv* (2022). URL `https://www.biorxiv.org/content/early/2022/03/01/2021.11.16.468246`. `https://www.biorxiv.org/content/early/2022/03/01/2021.11.16.468246.full.pdf`.

[38] Jordan, M. I., Ghahramani, Z., Jaakkola, T. S. & Saul, L. K. An introduction to variational methods for graphical models. *Machine learning* **37**, 183–233 (1999).

[39] Blei, D. M., Kucukelbir, A. & McAuliffe, J. D. Variational inference: A review for statisticians. *Journal of the American Statistical Association* **112** (2017).

[40] Hoffman, M. D., Blei, D. M., Wang, C. & Paisley, J. Stochastic variational inference. *Journal of Machine Learning Research* **14** (2013).

[41] Kingma, D. P. & Welling, M. Auto-encoding variational bayes (2014).

[42] Loh, P. R. *et al.* Efficient bayesian mixed-model analysis increases association power in large cohorts. *Nature Genetics* **47** (2015).

[43] Logsdon, B. A., Hoffman, G. E. & Mezey, J. G. A variational bayes algorithm for fast and accurate multiple locus genome-wide association analysis. *BMC Bioinformatics* **11** (2010).

[44] Demetci, P. *et al.* Multi-scale inference of genetic trait architecture using biologically annotated neural networks. *bioRxiv* (2020).

[45] Carbonetto, P. & Stephens, M. Scalable variational inference for bayesian variable selection in regression, and its accuracy in genetic association studies. *Bayesian Analysis* **7** (2012).

[46] Carbonetto, P. & Stephens, M. Integrated enrichment analysis of variants and pathways in genome-wide association studies indicates central role for il-2 signaling genes in type 1 diabetes, and cytokine signaling genes in crohn's disease. *PLoS Genetics* **9** (2013).

[47] Zhu, X. & Stephens, M. Large-scale genome-wide enrichment analyses identify new trait-associated genes and pathways across 31 human phenotypes. *Nature Communications* **9** (2018).

[48] Spence, J. Flexible mean field variational inference using mixtures of non-overlapping exponential families. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. & Lin, H. (eds.) *Advances in Neural Information Processing Systems*, vol. 33, 19642–19654 (Curran Associates, Inc., 2020). URL https://proceedings.neurips.cc/paper/2020/file/e3a54649aeec04cf1c13907bc6c5c8aa-Paper.pdf.

[49] Mitchell, T. J. & Beauchamp, J. J. Bayesian variable selection in linear regression. *Journal of the American Statistical Association* **83** (1988).

[50] George, E. I. & McCulloch, R. E. Approaches for bayesian variable selection. *Statistica Sinica* **7** (1997).

[51] Ishwaran, H. & Rao, J. S. Spike and slab variable selection: frequentist and bayesian strategies. *The Annals of Statistics* **33**, 730–773 (2005).

[52] Titsias, M. K. & Lázaro-Gredilla, M. Spike and slab variational inference for multi-task and multiple kernel learning (2011).

[53] Tzikas, D. G., Likas, A. C. & Galatsanos, N. P. The variational approximation for bayesian inference. *IEEE Signal Processing Magazine* **25** (2009).

[54] James, G., Witten, D., Hastie, T. & Tibshirani, R. *Introduction to Statistical Learning with Applications in R*, vol. 11 (2019).

[55] Snoek, J., Larochelle, H. & Adams, R. P. Practical bayesian optimization of machine learning algorithms. vol. 4 (2012).

[56] Falconer, D. S. The inheritance of liability to diseases with variable age of onset, with particular reference to diabetes mellitus. *Annals of Human Genetics* **31** (1967).

[57] Fernández, A. *et al. Learning from Imbalanced Data Sets* (2018).

[58] Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature Genetics* **51**, 584–591 (2019). URL https://doi.org/10.1038/s41588-019-0379-x.

[59] Privé, F. *et al.* Portability of 245 polygenic scores when derived from the uk biobank and applied to 9 ancestry groups from the same cohort. *The American Journal of Human Genetics* **109**, 12–23 (2022). URL https://www.sciencedirect.com/science/article/pii/S0002929721004201.

[60] Bulik-Sullivan, B. *et al.* Ld score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics* **47** (2015).

[61] Wen, X. & Stephens, M. Using linear predictors to impute allele frequencies from summary or pooled genotype data. *Annals of Applied Statistics* **4** (2010).

[62] Berisa, T. & Pickrell, J. K. Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics* **32** (2016).

[63] Khan, M. E., Bouchard, G., Marlin, B. M. & Murphy, K. P. Variational bounds for mixed-data factor analysis (2010).

[64] Ročková, V. & George, E. I. Emvs: The em approach to bayesian variable selection. *Journal of the American Statistical Association* **109** (2014).

[65] Galinsky, K. J. *et al.* Estimating cross-population genetic correlations of causal effect sizes. *Genetic epidemiology* **43**, 180–188 (2019). URL https://pubmed.ncbi.nlm.nih.gov/30474154.

[66] Shi, H. *et al.* Population-specific causal disease effect sizes in functionally important regions impacted by selection. *Nature Communications* **12**, 1098 (2021). URL https://doi.org/10.1038/s41467-021-21286-1.

[67] Wainschtein, P. *et al.* Assessing the contribution of rare variants to complex trait heritability from whole-genome sequence data. *Nature Genetics* **54**, 263–273 (2022). URL https://doi.org/10.1038/s41588-021-00997-7.

[68] Privé, F., Arbel, J., Aschard, H. & Vilhjálmsson, B. J. Identifying and correcting for misspecifications in gwas summary statistics and polygenic scores. *bioRxiv* (2022). URL https://www.biorxiv.org/content/early/2022/04/13/2021.03.29.437510. https://www.biorxiv.org/content/early/2022/04/13/2021.03.29.437510.full.pdf.

[69] Chen, W. *et al.* Improved analyses of gwas summary statistics by reducing data heterogeneity and errors. *Nature Communications* **12**, 7117 (2021). URL https://doi.org/10.1038/s41467-021-27438-7.

[70] Allen, H. L. *et al.* Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467** (2010).

[71] Teslovich, T. M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466** (2010).

[72] Speliotes, E. K. *et al.* Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nature Genetics* **42** (2010).

[73] Morris, A. P. *et al.* Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature Genetics* **44** (2012).

[74] Okada, Y. *et al.* Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506** (2014).

[75] Wu, H. *et al.* Polygenic risk score for low-density lipoprotein cholesterol is associated with risk of ischemic heart disease and enriches for individuals with familial hypercholesterolemia. *Circulation: Genomic and Precision Medicine* **14**, e003106 (2021).

[76] Hoffmann, T. J. & Witte, J. S. Strategies for imputing and analyzing rare variants in association studies. *Trends in genetics : TIG* **31**, 556–563 (2015). URL `https://pubmed.ncbi.nlm.nih.gov/26450338`.

[77] Shi, S. *et al.* Comprehensive assessment of genotype imputation performance. *Human Heredity* **83**, 107–116 (2018). URL `https://www.karger.com/DOI/10.1159/000489758`.

[78] O'Connor, T. D. *et al.* Fine-scale patterns of population stratification confound rare variant association tests. *PloS one* **8**, e65834–e65834 (2013). URL `https://pubmed.ncbi.nlm.nih.gov/23861739`.

[79] Zaidi, A. A. & Mathieson, I. Demographic history mediates the effect of stratification on polygenic scores. *eLife* **9**, e61548 (2020). URL `https://doi.org/10.7554/eLife.61548`.

[80] Mathieson, I. & McVean, G. Differential confounding of rare and common variants in spatially structured populations. *Nature Genetics* **44**, 243–246 (2012). URL `https://doi.org/10.1038/ng.1074`.

[81] Fan, C., Mancuso, N. & Chiang, C. W. K. A genealogical estimate of genetic relationships. *The American Journal of Human Genetics* URL `https://doi.org/10.1016/j.ajhg.2022.03.016`.

[82] Márquez-Luna, C. *et al.* Incorporating functional priors improves polygenic prediction accuracy in uk biobank and 23andme data sets. *Nature Communications* **12**, 6052 (2021). URL `https://doi.org/10.1038/s41467-021-25171-9`.

[83] Zhang, Q., Privé, F., Vilhjálmsson, B. & Speed, D. Improved genetic prediction of complex traits from individual-level data or summary statistics. *Nature Communications* **12**, 4192 (2021). URL `https://doi.org/10.1038/s41467-021-24485-y`.

[84] Cai, M. *et al.* A unified framework for cross-population trait prediction by leveraging the genetic correlation of polygenic traits. *The American Journal of Human Genetics* **108**, 632–655 (2021). URL `https://doi.org/10.1016/j.ajhg.2021.03.002`.

[85] Ruan, Y. *et al.* Improving polygenic prediction in ancestrally diverse populations. *medRxiv* (2021). URL `https://www.medrxiv.org/content/early/2021/08/24/2020.12.27.20248738`. `https://www.medrxiv.org/content/early/2021/08/24/2020.12.27.20248738.full.pdf`.

[86] Turner, R. E. & Sahani, M. *Two problems with variational expectation maximisation for time series models*, 104–124 (Cambridge University Press, 2011).

[87] Giordano, R., Broderick, T. & Jordan, M. I. Covariances, robustness, and variational bayes (2017). URL `https://arxiv.org/abs/1709.02536`.

[88] Zhang, C., Butepage, J., Kjellstrom, H. & Mandt, S. Advances in variational inference. *IEEE Transactions on Pattern Analysis Machine Intelligence* **41**, 2008–2026 (2019).

[89] Miller, A. C., Foti, N. J. & Adams, R. P. Variational boosting: Iteratively refining posterior approximations. In Precup, D. & Teh, Y. W. (eds.) *Proceedings of the 34th International Conference on Machine Learning*, vol. 70 of *Proceedings of Machine Learning Research*, 2420–2429 (PMLR, 2017). URL `https://proceedings.mlr.press/v70/miller17a.html`.

[90] Salimans, T., Kingma, D. P. & Welling, M. Markov chain monte carlo and variational inference: Bridging the gap (2014). URL `https://arxiv.org/abs/1410.6460`.

[91] Agnihotri, A. & Batra, N. Exploring bayesian optimization. *Distill* **5** (2020).

[92] Carbonetto, P., Zhou, X. & Stephens, M. varbvs: Fast variable selection for large-scale regression (2017).

[93] Chang, C. C. *et al.* Second-generation plink: Rising to the challenge of larger and richer datasets. *GigaScience* **4** (2015).

[94] McCaw, Z. R., Lane, J. M., Saxena, R., Redline, S. & Lin, X. Operating characteristics of the rank-based inverse normal transformation for quantitative trait analysis in genome-wide association studies. *Biometrics* **76**, 1262–1272 (2020).

[95] Pedregosa, F. *et al.* Scikit-learn: Machine learning in python. *the Journal of machine Learning research* **12**, 2825–2830 (2011).

[96] Behnel, S. *et al.* Cython: The best of both worlds. *Computing in Science & Engineering* **13**, 31–39 (2011).

[97] Rockova, V. & George, E. Negotiating multicollinearity with spike-and-slab priors. *Metron* **72**, 217–229 (2014).

[98] Virtanen, P. *et al.* SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* **17**, 261–272 (2020).

[99] Wang, J., Clark, S. C., Liu, E. & Frazier, P. I. Parallel bayesian global optimization of expensive functions. *Operations Research* **68** (2020).

[100] Yang, J., Zeng, J., Goddard, M. E., Wray, N. R. & Visscher, P. M. Concepts, estimation and interpretation of snp-based heritability (2017).

Supplementary Information

Shadi Zabad^{1}, Simon Gravel^{2,*}, Yue Li^{1,*}

^{1}School of Computer Science, McGill University, ^{2}Department of Human Genetics, McGill University

†Equal contribution *Correspondence: simon.gravel@mcgill.ca, yueli@cs.mcgill.ca

S1 VIPRS model derivation and implementation details

S1.1 Model formulation

We model the dependence of a quantitative phenotype y on the genotype matrix $\mathbf{X}$ via the standard linear model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{1}$$

where $\mathbf{y}$ is a vector of phenotypic measurements for N individuals, $\mathbf{X}_{N\times M}$ is a matrix of allelic counts for each individual and genetic marker, $\boldsymbol{\beta}$ is a vector of effect sizes, and $\boldsymbol{\epsilon}$ is a vector of residual effects for each individual. We assume that the phenotype conditioned on the effect sizes and genotypes follows an isotropic Gaussian likelihood $\mathbf{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma_{\epsilon}^{2}\boldsymbol{I})$. In Bayesian inference, we treat all unknown parameters as latent random variables and assign them a prior distribution. We assume that the effect sizes $\boldsymbol{\beta}$ have, as a prior, a factorized Gaussian mixture distribution with K components [6, 44],

$$\begin{aligned} p(\boldsymbol{\beta}, \mathbf{s}) &= \prod_{j=1}^{M} p(\beta_j|s_j)p(s_j) = \prod_{j=1}^{M}\Big(\prod_{k=1}^{K}\mathcal{N}(\beta_j; 0, \sigma_k^2)^{s_{jk}}\Big)\Big(\prod_{k=1}^{K}\pi_k^{s_{jk}}\Big) \\ p(\beta_j) &= \sum_{k=1}^{K}\pi_k\,\mathcal{N}(\beta_j; 0, \sigma_k^2) \end{aligned} \tag{2}$$

where s_{jk} is a latent random variable that takes value 1 if SNP j belongs to component k in the mixture, and 0 otherwise. The parameters π_k and σ_k^2 are the mixing proportions and the prior variance for component k, respectively. To allow for sparsity in the model structure, we assume that one of the components, e.g. the last component, is a degenerate Gaussian density with zero variance (i.e., the Dirac delta density δ_0). In this case, when $K = 2$, the above mixture prior reduces to the standard spike-and-slab prior [49–51], where the spike corresponds to the null component and the slab density corresponds to the non-null component. Given this model structure, the goal of Bayesian inference is to estimate the posterior for the effect sizes,

$$
\begin{aligned}
p(\boldsymbol{\beta}|\mathbf{X},\mathbf{y},\boldsymbol{\theta}) &= p(\mathbf{y}|\mathbf{X},\boldsymbol{\beta},\sigma_\epsilon^2)p(\boldsymbol{\beta}\mid\boldsymbol{\sigma}^2,\boldsymbol{\pi})/Z &(3)\\
&\propto p(\mathbf{y}|\mathbf{X},\boldsymbol{\beta},\sigma_\epsilon^2)p(\boldsymbol{\beta}\mid\boldsymbol{\sigma}^2,\boldsymbol{\pi}) &(4)
\end{aligned}
$$

where $\boldsymbol{\theta} = \{\sigma_\epsilon^2, \boldsymbol{\sigma}^2, \boldsymbol{\pi}\}$ is a collection of hyperparameters for the model and $Z = \int p(\mathbf{y}|\mathbf{X},\boldsymbol{\beta},\sigma_\epsilon^2)p(\boldsymbol{\beta}\mid\boldsymbol{\sigma}^2,\boldsymbol{\pi})d\boldsymbol{\beta}$ is the marginal likelihood or the partition function. Therefore, the posterior is proportional to the product of the likelihood and the prior.

S1.2 Variational inference

The mixture prior on $\boldsymbol{\beta}$ results in intractable posterior because the analytical closed-form of the marginal likelihood is not available. Thus, we need to devise a scheme for approximate posterior inference. In this work, we use variational inference, a deterministic approximate inference scheme that involves specifying a parametric distribution family $q(\boldsymbol{\beta}, \mathrm{s})$ and optimizing its parameters to match the true posterior as closely as possible [35, 36, 39]. We use the same family of distributions as the prior, which assumes that the effect size at each SNP follows a mixture of K Gaussians (with the last Gaussian as a spike with zero variance) [44]. For tractability and computational efficiency, we make use of the paired mean-field assumption [52], which factorizes the joint distribution of the effect sizes and causal indicators into a product of independent densities for each SNP,

$$
q(\boldsymbol{\beta},\mathrm{s}) = \prod_{j=1}^{M} q(\beta_j, s_j) = \prod_{j=1}^{M} q(\beta_j \mid s_j)q(s_j) = \prod_{j=1}^{M}\Big(\prod_{k=1}^{K}\mathcal{N}(\mu_{jk},\sigma_{jk}^2)^{s_{jk}}\Big)\Big(\prod_{k=1}^{K}\gamma_{jk}^{s_{jk}}\Big) \tag{5}
$$

where $\mu_{jk}, \sigma_{jk}^2, \gamma_{jk}$ for all j and k are the variational parameters that require tuning. Given this posterior distribution family, for notational convenience, we define the first and second moments for the effect size of SNP j,

$$
\begin{aligned}
\eta_j &\equiv \mathbb{E}_q[\beta_j] = \sum_k \gamma_{jk}\mu_{jk}\\
\zeta_j &\equiv \mathbb{E}_q[\beta_j^2] = \sum_k \gamma_{jk}(\mu_{jk}^2 + \sigma_{jk}^2)\\
Var_q(\beta_j) &= \zeta_j - \eta_j^2 = \sum_k \gamma_{jk}(\mu_{jk}^2 + \sigma_{jk}^2) - \Big(\sum_k \gamma_{jk}\mu_{jk}\Big)^2.
\end{aligned} \tag{6}
$$

where $\mathbb{E}_q$ and Var_q denote the expectation and variance taken with respect to the proposed distribution q, which abbreviates $q(\boldsymbol{\beta}, \mathrm{s})$. Furthermore, η_j and ζ_j denote the posterior mean for the effect size and squared effect size, respectively.

Given proposed distribution family q, the variational inference scheme proceeds by tuning its parameters to approximate the posterior. The measure of discrepancy between the proposed distribution and the true posterior that we optimize in this work is the Kullback-Leibler (KL-)

Divergence [39],

$$
\begin{aligned}
KL(q||p) &= \mathbb{E}_q\left[\log\frac{q(\boldsymbol{\beta},\mathbf{s})}{p(\boldsymbol{\beta},\mathbf{s}\mid\mathbf{X},\mathbf{y},\boldsymbol{\theta})}\right] = \mathbb{E}_q\left[\log q(\boldsymbol{\beta},\mathbf{s})\right] - \mathbb{E}_q\left[\log p(\boldsymbol{\beta},\mathbf{s}\mid\mathbf{X},\mathbf{y},\boldsymbol{\theta})\right] \\
&= \mathbb{E}_q\left[\log q(\boldsymbol{\beta},\mathbf{s})\right] - \mathbb{E}_q\left[\log p(\mathbf{y},\boldsymbol{\beta},\mathbf{s}\mid\mathbf{X},\boldsymbol{\theta})\right] + \log p(\mathbf{y}\mid\mathbf{X},\boldsymbol{\theta})
\end{aligned} \tag{7}
$$

The first term in (7) is the negative entropy of the proposed distribution, whereas the second term is the expectation of the complete log-likelihood with respect to proposed density [39]. The last term is the log marginal likelihood or the model evidence, the normalizing factor that makes the inference intractable. In this setting, the model evidence is a constant with respect to the variational distribution and this fact motivates the use of a surrogate objective, known as the Evidence Lower BOund (ELBO) of the log marginal likelihood [35, 36, 39],

$$
ELBO = \mathbb{E}_q\left[\log p(\mathbf{y},\boldsymbol{\beta},\mathbf{s}\mid\mathbf{X},\boldsymbol{\theta})\right] - \mathbb{E}_q\left[\log q(\boldsymbol{\beta},\mathbf{s})\right] \tag{8}
$$

The ELBO combines the expectation of the complete log-likelihood and the entropy of the variational distribution, thus balancing approximation quality against model complexity [35]. Given the model formulation outlined above, the ELBO takes the following form [44, 45],

$$
\begin{aligned}
ELBO = &-\frac{N}{2}\log(2\pi\sigma_\epsilon^2) - \frac{1}{2\sigma_\epsilon^2}\left(\|\mathbf{y}-\mathbf{X}\boldsymbol{\eta}\|_2^2 + \sum_{j=1}^{M} X_j^\top X_j Var_q(\beta_j)\right) \\
&-\sum_{j=1}^{M}\sum_{k=1}^{K}\gamma_{jk}\log(\frac{\gamma_{jk}}{\pi_k}) + \sum_{j=1}^{M}\sum_{k=1}^{K-1}\frac{\gamma_{jk}}{2}\left[1+\log(\frac{\sigma_{jk}^2}{\sigma_k^2}) - \frac{\mu_{jk}^2+\sigma_{jk}^2}{\sigma_k^2}\right].
\end{aligned} \tag{9}
$$

where $\|\mathbf{y}-\mathbf{X}\boldsymbol{\eta}\|_2^2 = (\mathbf{y}-\mathbf{X}\boldsymbol{\eta})^\top(\mathbf{y}-\mathbf{X}\boldsymbol{\eta})$. Given the ELBO as the objective function to maximize, we can derive a coordinate ascent Variational EM (VEM) algorithm to learn the variational parameters, as well as the fixed hyperparameter $\theta = \{\pi_1,\ldots,\pi_K,\sigma_1^2,\ldots,\sigma_K^2,\sigma_\epsilon^2\}$ [44, 45, 92]. As detailed below, we update the variational parameters in the E-Step and the hyperparameters in the M-Step, both with the aim of maximizing the ELBO. This process is repeated until convergence.

S1.3 Variational EM algorithm

The E-Step In the E-Step, we update the variational parameters for each SNP j and mixture component k in an iterative fashion, using closed-form updates that can be obtained by taking the partial derivatives of the ELBO with respect to target parameters and solving for the roots

[44,45]:

$$\begin{aligned}
\sigma_{jk}^2 &= \frac{\sigma_\epsilon^2}{X_j^\top X_j + \sigma_\epsilon^2/\sigma_k^2} \\
\mu_{jk} &= \frac{\sigma_{jk}^2}{\sigma_\epsilon^2}\left(\mathbf{y}^\top X_j - \sum_{m\neq j} X_j^\top X_m \eta_m\right) \\
u_{jk} &= \log(\pi_k) + \frac{1}{2}\log(\frac{\sigma_{jk}^2}{\sigma_k^2}) + \frac{1}{2\sigma_{jk}^2}\mu_{jk}^2 \\
\gamma_{jk} &= \frac{\exp(u_{jk})}{\sum_{k=1}^K \exp(u_{jk})} = Softmax(\mathbf{u}_j)_k
\end{aligned} \tag{10}$$

The $Softmax$ function applied to the intermediate vector $\mathbf{u}_j$ arises from the need to obtain properly normalized posterior densities [35], such that the mixture probabilities for each SNP j sum to 1. For the null component (i.e., the spike component), by taking the limit as the prior variance σ_k^2 approaches zero ($\lim_{\sigma_k^2 \to 0}$) over the update equations above, we see that, as expected, $\sigma_{jk}^2 = \mu_{jk} = 0$ and $u_{jk} = \log(\pi_k)$.

The M-Step In the M-Step, we update the hyperparameters of the model θ to maximize the ELBO as the approximate log marginal likelihood [45]. Similar to the setup in the E-Step, this is done by taking the partial derivatives of the ELBO w.r.t. each hyperparameter θ_i and solving to obtain:

$$\begin{aligned}
\pi_k &= \frac{1}{M}\sum_{j=1}^M \gamma_{jk} \\
\sigma_k^2 &= \frac{\sum_{j=1}^M \gamma_{jk}(\mu_{jk}^2 + \sigma_{jk}^2)}{\sum_{j=1}^M \gamma_{jk}} \\
\sigma_\epsilon^2 &= \frac{1}{N}\Big(\|\mathbf{y} - \mathbf{X}\boldsymbol{\eta}\|_2^2 + \sum_{j=1}^M X_j^\top X_j Var_q(\beta_j)\Big)
\end{aligned} \tag{11}$$

S1.4 Expressing the model in terms of GWAS summary statistics

The model formulation above and the resultant closed-form update equations require access to individual-level GWAS data, which can potentially limit its scalability and applicability in many practical settings. An important contribution of this work, therefore, was to re-frame the various components of the model such that it only requires GWAS summary statistics for training [33].

Concretely, by examining the closed-form updates in Eq. (10) and Eq. (11) as well as the ELBO definition in Eq. (8), we see that the individual-level data factors in the updates for the variational parameters σ_{jk}^2, μ_{jk}, the residual variance σ_ϵ^2, as well as the ELBO. Under the assumption that both the phenotype vector $\mathbf{y}$ and the genotype matrix $\mathbf{X}$ have been standardized column-wise to have zero mean and unit variance, the following equivalences hold:

$$
\begin{aligned}
X_j^\top X_j &= N \\
X_j^\top X_m &= NR_{jm} \\
\mathbf{y}^\top X_j &= N\hat{\beta}_j \\
\mathbf{y}^\top \mathbf{y} &= N
\end{aligned}
$$

where N is the GWAS sample size, $\mathbf{R}$ is the $M \times M$ Linkage Disequilibrium (LD) matrix that records the pairwise Pearson Correlation coefficient between all pairs of SNPs j and m, and $\hat{\beta}_j$ is the standardized marginal GWAS effect size. In this setup, the matrix $\mathbf{R}$ can be derived from an appropriately-matched reference panel [33]. In many GWAS applications and software implementations (e.g. `plink2` [93]), the genotype matrix is not standardized. To deal with this in practice, we use the pseudo-correlation estimate from Mak et al. (2017) [25] to obtain standardized effect sizes,

$$
\hat{\beta}_j = \frac{z_j}{\sqrt{N - 1 + z_j^2}}
$$

where z_j is the marginal GWAS z-score of SNP j. With this in hand, we can write the update equations for the variational parameters in terms of GWAS summary statistics:

$$
\begin{aligned}
\sigma_{jk}^2 &= \frac{\sigma_\epsilon^2}{N + \frac{\sigma_\epsilon^2}{\sigma_k^2}} \\
\mu_{jk} &= \frac{N\sigma_{jk}^2}{\sigma_\epsilon^2}\Big(\hat{\beta}_j - \sum_{m \neq j} R_{jm}\eta_m\Big)
\end{aligned}
$$

Similarly, in the M-Step, the update for the residual variance, σ_ϵ^2, can be written purely in terms of GWAS summary statistics as well:

$$
\begin{aligned}
\sigma_\epsilon^2 &= \frac{1}{N}\Bigg[\mathbf{y}^\top\mathbf{y} - 2\mathbf{y}^\top\mathbf{X}\boldsymbol{\eta} + \boldsymbol{\eta}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\eta} + \sum_{j=1}^{M} X_j^\top X_j Var_q(\beta_j)\Bigg] \\
&= 1 - 2\sum_{j=1}^{M}\hat{\beta}_j\eta_j + 2\sum_{j=1}^{M}\sum_{m<j} R_{jm}\eta_j\eta_m + \sum_{j=1}^{M}\zeta_j
\end{aligned}
$$

Finally, and using the same expansion for the squared loss as above, we can re-write the objective, the ELBO in Eq. (8), in terms of GWAS summary statistics only,

$$
\begin{aligned}
ELBO = &-\frac{N}{2}\log(2\pi\sigma_\epsilon^2) - \frac{N}{2\sigma_\epsilon^2}\left(1 - 2\sum_{j=1}^{M}\hat{\beta}_j\eta_j + 2\sum_{j=1}^{M}\sum_{m<j}R_{jm}\eta_j\eta_m + \sum_{j=1}^{M}\zeta_j\right) \\
&- \sum_{j=1}^{M}\sum_{k=1}^{K}\gamma_{jk}\log(\frac{\gamma_{jk}}{\pi_k}) + \sum_{j=1}^{M}\sum_{k=1}^{K-1}\frac{\gamma_{jk}}{2}\Big[1 + \log(\frac{\sigma_{jk}^2}{\sigma_k^2}) - \frac{\mu_{jk}^2 + \sigma_{jk}^2}{\sigma_k^2}\Big].
\end{aligned}
$$

These new update equations form the backbone of the proposed VIPRS model and the new summary statistics-based EM framework is summarized in Algorithm 1.

S1.5 Implementation details

S1.5.1 Initialization

The Variational EM algorithm for approximate posterior inference is prone to entrapment in local optima, especially in regimes with severe multi-collinearity as is the case with genotype data [45, 64, 97]. Thus, finding an approximate solution that is close to the global optimum

Algorithm 1: Variational Expectation-Maximization (VEM) algorithm for Bayesian PRS inference with GWAS summary statistics

```
Data: Marginal GWAS summary statistics for $M$ SNPs: Standardized $\hat{\boldsymbol{\beta}}$ or $z$-scores;
GWAS Sample size $N$;
Appropriately-matched LD matrix $\boldsymbol{R}$
Result: Posterior estimates for effect sizes by the variational parameter $\boldsymbol{\eta}$
initializeModelParameters();
while not converged do
    /* E-Step:  Update variational parameters */;
    for $j = 1;\ j <= M;\ j = j + 1$ do
        for $k = 1;\ k < K;\ k = k + 1$ do
            $\sigma^2_{jk} \leftarrow \frac{\sigma^2_\epsilon}{N + \sigma^2_\epsilon / \sigma^2_k}$;
            $\mu_{jk} \leftarrow \frac{N\sigma^2_{jk}}{\sigma^2_\epsilon}\left(\hat{\beta}_j - \sum_{m \neq j} R_{jm}\eta_m\right)$;
            $u_{jk} \leftarrow \log(\pi_k) + \frac{1}{2}\log(\sigma^2_{jk}/\sigma^2_k) + \frac{1}{2\sigma^2_{jk}}\mu^2_{jk}$;
        end
        $\boldsymbol{\gamma_j} \leftarrow Softmax(\boldsymbol{u_j})$;
    end
    /* M-Step:  Update hyperparameters */;
    for $k = 1;\ k < K;\ k = k + 1$ do
        $\pi_k \leftarrow \frac{1}{M}\sum_j \gamma_{jk}$;
        $\sigma^2_k \leftarrow \sum_j \gamma_{jk}(\mu^2_{jk} + \sigma^2_{jk}) / \sum_j \gamma_{jk}$;
    end
    $\sigma^2_\epsilon \leftarrow 1 - 2\sum_{j=1}^{M} \hat{\beta}_j\eta_j + 2\sum_{j=1}^{M}\sum_{m>j} R_{jm}\eta_j\eta_m + \sum_{j=1}^{M}\zeta_j$;
end
/* Obtain the final posterior estimates */;
for $j = 1;\ j <= M;\ j = j + 1$ do
    /* The posterior mean for the effect size*/;
    $\eta_j \leftarrow \sum_{k=1}^{K-1}\gamma_{jk}\mu_{jk}$;
    /* The posterior inclusion probability (PIP)*/;
    PIP($j$) $\leftarrow \sum_{k=1}^{K-1}\gamma_{jk}$;
end
```

requires a well-informed and data-driven parameter initialization strategy. This is particularly important for the hyperparameters of the model, $\theta = \{\pi_1, \dots, \pi_K, \sigma_1^2, \dots, \sigma_K^2, \sigma_\epsilon^2\}$, since they are fixed in the generative process outlined above and affect the update equations of all SNPs. For the hyperparameters of the model, we employ the following strategy for initialization:

- For the **mixing proportions** π_k, we draw the overall proportion of causal variants, $\pi = \sum_{k=1}^{K-1} \pi_k$, from a uniform between $[0.005, 0.1]$, informed by empirical estimates for a large number of complex traits [10,11]. Given this random draw, we set the proportion of variants that belong to the null component to $\pi_K = 1 - \pi$. The non-null mixture proportions are initialized randomly using a Dirichlet distribution with order $K - 1$ and concentration parameters set to $\alpha_k = 1$ for all $k \in \{1, \dots, K-1\}$. Then, the sampled probabilities are multiplied by π to ensure that the overall mixing proportions sum to 1.

- For the **residual variance** σ_ϵ^2, we leverage a fast and simple estimator for the SNP heritability or proportion of variance explained by the SNPs included in the model. For the simple estimator of heritability, we use the formula from [4],

$$\hat{h}_g^2 = \frac{(\overline{\chi}^2 - 1)M}{\bar{l}N}$$

  where $\bar{l}$ is the average LD score, $\overline{\chi}^2$ is the average marginal Chi-Squared statistic, and N is the GWAS sample size, and M is the number of SNPs included in the model. Given this estimate of the heritability, we initialize the residual variance as $\sigma_\epsilon^2 = 1 - \hat{h}_g^2$.

- For the **prior variances on the effect size** σ_k^2, we follow the SBayesR model specification where the different Gaussian components are on different scales, e.g. $\boldsymbol{\sigma}^2 = \{d_1\sigma_\beta^2, d_2\sigma_\beta^2, \dots, d_{K-1}\sigma_\beta^2\}$ [6]. In our model initialization, we set multiplier to be $d_k = 10^{-(k-1)}$, so that the first component has variance σ_β^2, the second component has variance $0.1\sigma_\beta^2$, the third component has variance $0.01\sigma_\beta^2$, etc. The shared parameter σ_β^2 is initialized to be,

$$\sigma_\beta^2 = \frac{\hat{h}_g^2}{M \sum_{k=1}^{K-1} d_k \pi_k},$$

  where $\hat{h}_g^2$ is the naïve SNP heritability estimate, M is the number of SNPs included in the model, d_k's are the multipliers, and π_k's are the mixing proportions described earlier. In the case of a spike-and-slab prior where $K = 2$, it is easy to see that this simply reduces to the common assumption that $\sigma_\beta^2 = \frac{h_g^2}{\pi M}$, where π is the overall proportion of causal variants.

Once the hyperparameters have been initialized, we initialize the variational parameters $\gamma_{jk}, \mu_{jk}, \sigma_{jk}^2$ for each SNPs $j \in \{1, \dots, M\}$ and each mixture component $k \in \{1, \dots, K\}$. The posterior means for all the mixture components, $\mu_{jk} \forall k \in \{1, \dots, K\}$, are simply initialized to be zero. The posterior mixing proportions γ_{jk}'s, on the other hand, are initialized according to the corresponding priors, i.e. $\gamma_{jk} = \pi_k$ for all j and k. Finally, the posterior variances σ_{jk}^2's are initialized according to Equation 10, taking the initial values for the prior residual variance σ_ϵ^2 and prior variance σ_k^2 as input.

S1.5.2 The *Alpha* prior on the effect size

In studies of the genetic architectures of complex traits, it has been observed that there is a non-linear dependence between the minor allele frequency (MAF) of a genetic variant and its effect size [10, 13, 14]. This MAF-dependent architecture of complex traits is usually modelled with what is known as the "alpha" prior, where the effect size of a SNP is assumed to be drawn from a density where the variance is related to its frequency in the population, e.g.,

$$Var(\beta_j) \propto \Big(p_j(1-p_j)\Big)^{1+\alpha}.$$

Here, p_j is the minor allele frequency of SNP j and α is a free parameter that determines the level of dependence between the effect size and the allele frequency. We sought to examine whether incorporating such a prior into the `VIPRS` framework meaningfully improves prediction accuracy. To this end, we derived the `VIPRSAlpha` model, which merges the spike-and-slab prior with the alpha prior, resulting in a mixture density that takes the following form:

$$\beta_j = \pi \mathcal{N}\Big(\beta_j; 0, \big(p_j(1-p_j)\big)^{1+\alpha}\sigma_\beta^2\Big) + (1-\pi)\delta_0.$$

This prior differs from the spike-and-slab in the sense that the per-SNP heritability of causal variants can vary depending on their allele frequency and on the parameter α. However, it is more general and in fact reduces to the standard spike-and-slab prior when $\alpha = -1$.

Given a certain α value, it is straightforward to show that the update equations in the E-Step (Equation (10)) are not affected by this change, except that we need to replace all instances of σ_β^2 with $\big(p_j(1-p_j)\big)^{1+\alpha}\sigma_\beta^2$ for each SNP j. By the same token, this also applies to the closed-form updates for proportion of causal variants π and the residual variance σ_ϵ^2 in the M-Step (Equation (11)). However, for the prior variance on the effect size σ_β^2, the update equation becomes:

$$\sigma_\beta^2 = \frac{\sum_{j=1}^{M} \gamma_{jk}(\mu_j^2 + \sigma_j^2)\big(p_j(1-p_j)\big)^{-(1+\alpha)}}{\sum_{j=1}^{M} \gamma_j}$$

Given this formulation, however, we still need a way to set the unknown parameter α. In our software implementation, we allow the user to either set a fixed value for alpha (e.g. $\alpha = -0.25$) or infer its value in the M-Step by optimizing with respect to the ELBO. In the latter case, due to the non-linearity, we use the `L-BFGS-B` optimizer in `scipy` [98] to tune the value of α, where the objective function that we minimize is the negative ELBO plus a quadratic penalty term that is equivalent to the standard normal prior imposed on α by [10]: $\mathcal{L} = -ELBO + \alpha^2$. In our experiments, (**Supplementary Fig.** S3 and S4), we optimized the value of α in the M-Step.

S1.5.3 Hyperparameter tuning strategies

Hyperparameter tuning is important in the Variational EM framework outlined above. Previous work has shown that the EM framework for this model is sensitive to parameter initialization and, therefore, may get stuck in local optima, especially in the presence of strong multicollinearity [45, 64, 97]. This difficulty is compounded by the fact that the model uses point estimates for these hyperparameters, obtained by maximizing the surrogate objective, the ELBO. These maximum likelihood estimates of the hyperparameters do not capture the true uncertainty around them and may lead the model to overfit to the training data [63].

As a remedy, we explored three other strategies for setting or tuning the hyperparameters of the model. These strategies can be deployed in conjunction with the EM framework, or in a hybrid manner. For instance, the mixing proportions π_k can be optimized via cross-validation while the remaining parameters are updated as before in the M-Step [44, 45, 92]. Some of these strategies require access to a held-out validation dataset where we can directly measure the predictive performance associated with each hyperparameter setting.

The three hyperparameter tuning strategies are:

Grid Search (GS): Grid search is a popular strategy for tuning model hyperparameters that has been shown to result in improved predictive performance for a variety of PRS methods [25–27, 29]. This strategy involves two steps that the user can control: **(1)** Grid construction and **(2)** Specifying model selection metric.

For grid construction, the user can choose which hyperparameters to optimize via cross-validation, including the overall proportion of causal variants π, the residual variance σ_ϵ^2, and the prior variance on the effect size σ_β^2. For each hyperparameter, the user specifies the number of values to test as well as the scale of the grid points. For instance, for the proportion of causal variants π, we test 30 equidistant points from $\frac{1}{M}$ to $\frac{M-1}{M}$ on a $\log_{10}$ scale. If the user provides an estimate for any of the hyperparameters, the software can generate a “localized” grid. For example, given a heritability estimate $\hat{h}_g^2$, the grid for the residual variance σ_ϵ^2 can be multiples of the value provided, e.g. $\sigma_\epsilon^2 \in \{0.5(1-\hat{h}_g^2), 0.75(1-\hat{h}_g^2), (1-\hat{h}_g^2), 1.25(1-\hat{h}_g^2), 1.5(1-\hat{h}_g^2)\}$ [29]. Once the grid has been specified, the `VIPRS-GS` model is trained, in parallel, for each hyperparameter combination. The hyperparameters that are not in the grid are updated using the maximum-likelihood estimates derived in Eq. (11). In our experiments, we found that performing grid search over the overall proportion of causal variants π with 30 grid points results in a good trade-off between predictive performance and computational efficiency.

In order to select the best hyperparameter combination, we experimented with two model selection criteria. The first metric is the prediction R^2 (or (AUPRC) for case-control phenotypes) on a held-out validation dataset [25, 27, 29]. This approach is generally robust and less prone to over-fitting [54], though it requires that the user has access to an independent validation set. As alternative, we also experimented with using a second metric for model selection: the training ELBO. The ELBO is a lower bound on the marginal log-likelihood and it balances model fit against model complexity [35, 36]. Given that the standard EM algorithm can get stuck in local optima, exploring many modes of the model using grid search can, in principle, get us closer to the global optimum.

Bayesian Model Averaging (BMA): A different approach for dealing with the unknown hyperparameters, proposed by [45], is to integrate them out using importance sampling. Concretely, similar to the grid search setup, the idea consists in fitting `VIPRS` with a grid of reasonable values for the hyperparameters. However, instead of selecting the model with the best ELBO or best predictive performance on a validation set, we average the posterior parameter estimates using the importance weights $w(\theta^{(g)})$ [45]:

$$\eta_j = \frac{\sum_{g=1}^{G} \eta_j^{(g)} w(\theta^{(g)})}{\sum_{g=1}^{G} w(\theta^{(g)})}$$

where $\eta_j = \mu_j^{(g)} \gamma_j^{(g)}$ and $\mu_j^{(g)}$ and $\gamma_j^{(g)}$ are the expected causal effect and causal indicator, respectively, given the hyperparameters $\theta^{(g)}$. We followed earlier works and take the estimate for the ELBO at convergence as a proxy for the weights $w(\theta^{(g)}))$ [44, 45, 92].

Bayesian Optimization (BO): In Bayesian Optimization, we assume that there is an underlying black-box function $f(\theta)$ that takes the hyperparameters as input and outputs a certain score that we wish to optimize, such as the training ELBO or the validation R^2 [55]. This unknown function is modeled with a Gaussian Process (GP) prior, which allows for exploring the hyperparameter space efficiently while accounting for uncertainty in a principled manner. The other component in this framework is the acquisition function, a heuristic that maps from the GP posterior to information about the most promising regions in hyperparameter space [55, 91]. In our experiments, we used the `scikit-optimize` python package to perform this optimization, with `gp_hedge` as the default acquisition function. By default, the optimizer is allowed to sequentially evaluate up to 20 points in the space of hyperparameters.

In our main experiments, we used Bayesian Optimization to search over the hyperparameter π, while updating the remaining remaining hyperparameters using their maximum likelihood estimates, as before [44, 92]. The sequential nature of standard Bayesian optimization procedures makes it slow in practice, compared to e.g. grid search where we can explore different models in parallel. This could potentially be addressed in future work by using parallel implementations of Bayesian optimization [99].

S1.5.4 Efficient access and multiplication with the LD matrix

A major bottleneck in the coordinate ascent procedure outlined above is the storage and multiplication with the Linkage-Disequilibrium (LD) matrix $\boldsymbol{R}$ or its columns. In our software implementation, the matrix $\boldsymbol{R}$, whether banded, shrunk or dense, is stored on disk in a compressed and chunked array format known as `Zarr` arrays in `python`. During model fitting, if the per-chromosome matrix does not fit in memory, the software is designed to access it chunk by chunk in order to update the model parameters while managing limited memory resources.

While the `Zarr` library provides fast multi-threaded read and write access to the chunked arrays, it still incurs a heavy computational burden compared to arrays that are available in memory. Therefore, in order to get optimal speed and efficiency, ideally we need to minimize read access for the LD matrix when stored on disk. In a naïve implementation of the coordinate ascent procedure outlined above, we would need to access the LD matrix at least 3 times: **(1)** To update the variational parameters μ_{jk}'s, **(2)** to update the residual variance σ_{ϵ}^2, and **(3)** to compute the ELBO, our surrogate objective.

Here we will outline a procedure similar to the "right-hand" updating scheme of SBayesR [6] that keeps track of a per-SNP term, denoted by q_j, which makes it possible to access the LD matrix only once per iteration. We define the $q-$factor for SNP j as:

$$q_j = \sum_{k \neq j} \eta_k R_{jk}$$

In words, q_j is the sum of the posterior mean for the effect size of all neighboring variants, weighted by their Pearson correlation coefficient with the focal SNP j. In our implementation, we update the q_j's during the E-Step. In particular, as we cycle through the SNPs, we compute a partial update to q_j by including only SNPs whose posterior parameters have been updated already:

$$q_j \leftarrow \sum_{k<j} \eta_k R_{jk}.$$

Here, $\leftarrow$ denotes the assignment operator. And then for all SNPs indexed by $k < j$, we update their corresponding q_k factors by adding the weighted posterior mean for SNP j,

$$q_k \leftarrow q_k + \eta_j R_{jk}.$$

The benefit of keeping track of this term is that we can then use it when updating the residual variance σ_{ϵ}^2 or computing the ELBO without unnecessarily accessing the LD matrix again. The update for residual variance σ_{ϵ}^2 becomes:

$$\sigma_{\epsilon}^2 = 1 - 2\sum_{j=1}^{M} \hat{\beta}_j \eta_j + \sum_{j=1}^{M} \eta_j q_j + \sum_{j=1}^{M} \zeta_j.$$

Similarly, the ELBO can be expressed in terms of the q-factor as:

$$\begin{aligned} ELBO = & -\frac{N}{2}\log(2\pi\sigma_{\epsilon}^2) - \frac{N}{2\sigma_{\epsilon}^2}\left(1 - 2\sum_{j=1}^{M} \hat{\beta}_j \eta_j + \sum_{j=1}^{M} \eta_j q_j + \sum_{j=1}^{M} \zeta_j\right) \\ & - \sum_{j=1}^{M}\sum_{k=1}^{K} \gamma_{jk} \log(\frac{\gamma_{jk}}{\pi_k}) + \sum_{j=1}^{M}\sum_{k=1}^{K-1} \frac{\gamma_{jk}}{2}\Big[1 + \log(\frac{\sigma_{jk}^2}{\sigma_k^2}) - \frac{\mu_{jk}^2 + \sigma_{jk}^2}{\sigma_k^2}\Big]. \end{aligned}$$

S1.5.5 Heritability estimation

An important quantity to help diagnose model fit and quality of posterior approximation is the narrow-sense heritability, or proportion of additive variance explained (PVE) by the SNPs included in the model [100]. In the context of the linear model outlined in Equation 2, SNP heritability is defined as:

$$h^2_{SNP} = \frac{Var(\mathbf{X}\boldsymbol{\beta})}{Var(\mathbf{y})} = \frac{Var(\mathbf{X}\boldsymbol{\beta})}{Var(\mathbf{X}\boldsymbol{\beta}) + \sigma^2_\epsilon}$$

where $Var(\mathbf{X}\boldsymbol{\beta})$ denotes the genotypic variance and $Var(\mathbf{y})$ is the phenotypic variance, which we assume to be a linear sum of the genotypic and residual variances. Therefore, given this formulation, estimating SNP-heritability is equivalent to the task of estimating the genotypic and residual variances under the posterior distribution [6, 44].

In our model, the residual variance σ^2_ϵ is a point estimate that we update in the M-Step of the Variational EM algorithm and that point estimate is used in the denominator above to compute an estimate of SNP-heritability. As for the genotypic variance, assuming that both $\mathbf{X}$ and $\boldsymbol{\beta}$ are random, we can write:

$$\begin{aligned} Var(\mathbf{X}\boldsymbol{\beta}) &= \mathbb{E}[(\mathbf{X}\boldsymbol{\beta})^2] - \mathbb{E}[\mathbf{X}\boldsymbol{\beta}]^2 \\ &= \mathbb{E}_q\left[\frac{1}{N}\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}\right] = \sum_j \sum_{k\neq j} \eta_j\eta_k R_{jk} + \sum_j \zeta_j \end{aligned}$$

where the expectation is taken with respect to all the random components of the model. The second equality follows from our assumption that the genotype matrix is standardized to have mean zero and unit variance. The above estimator for the genotypic variance under the variational posterior can be simplified and expressed in terms of the $q-$factor described previously,

$$Var(\mathbf{X}\boldsymbol{\beta}) = \sum_j \eta_j q_j + \zeta_j.$$

Previous work has noted that the Variational approximation to the posterior tends to systematically underestimate SNP heritability or PVE [44, 45]. In our experiments, we saw that the estimated heritability is accurate for sparse genetic architectures, though it tends to be systematically underestimated for highly polygenic traits (**Supplementary Fig.** S7).

S1.6 A heuristic test of mismatch between GWAS summary statistics and LD reference panel

Summary statistics-based PRS methods can be sensitive to heterogeneities between GWAS summary statistics and the LD reference panel [8, 29, 68]. For some Bayesian methods, this mismatch can result in unpredictable behavior, with the posterior mean for the effect sizes exploding in magnitude and the estimated SNP-heritability exceeding 1 in some circumstances[1]. We observed that these inference difficulties can arise in the context of the VIPRS model, with the model objective, the ELBO, rapidly diverging after a few iterations. Thus, we sought ways to detect and eliminate strong inconsistencies between the GWAS cohort and the LD reference panel.

In general, if the requisite data is available, we encourage users to run principled and comprehensive tests of heterogeneity, such as the DENTIST method of Chen et al. (2021) [69]. However, in case this is not feasible, due to e.g. using precomputed LD matrices without direct access to the individual-level data of the reference panel, we designed a fast, heuristic and noisy estimator of the DENTIST p-value ($\boldsymbol{P}_{\text{DENTIST}}$) that leverages only pre-computed summary statistics. As explained below, our experiments indicate that this heuristic approach can help inference robustness by identifying problematic SNPs.

The heuristic test of mismatch works as follows: For every variant i included in the analysis, we randomly sample a fixed number of neighbors G with replacement. The neighbors in this case are defined by the pre-computed LD matrix. For instance, for the windowed estimator of LD, the neighbors are only SNPs that are within a pre-specified window around the focal SNP. Then for every sampled neighbor j, we compute the T_d statistic as [69],

$$T_d(i,j) = \frac{(z_i - z_j R_{ij})^2}{1 - R_{ij}^2}.$$

The final T_d statistic for variant i, then, is the average across the SNP's sampled neighbors $\mathcal{S}(i)$,

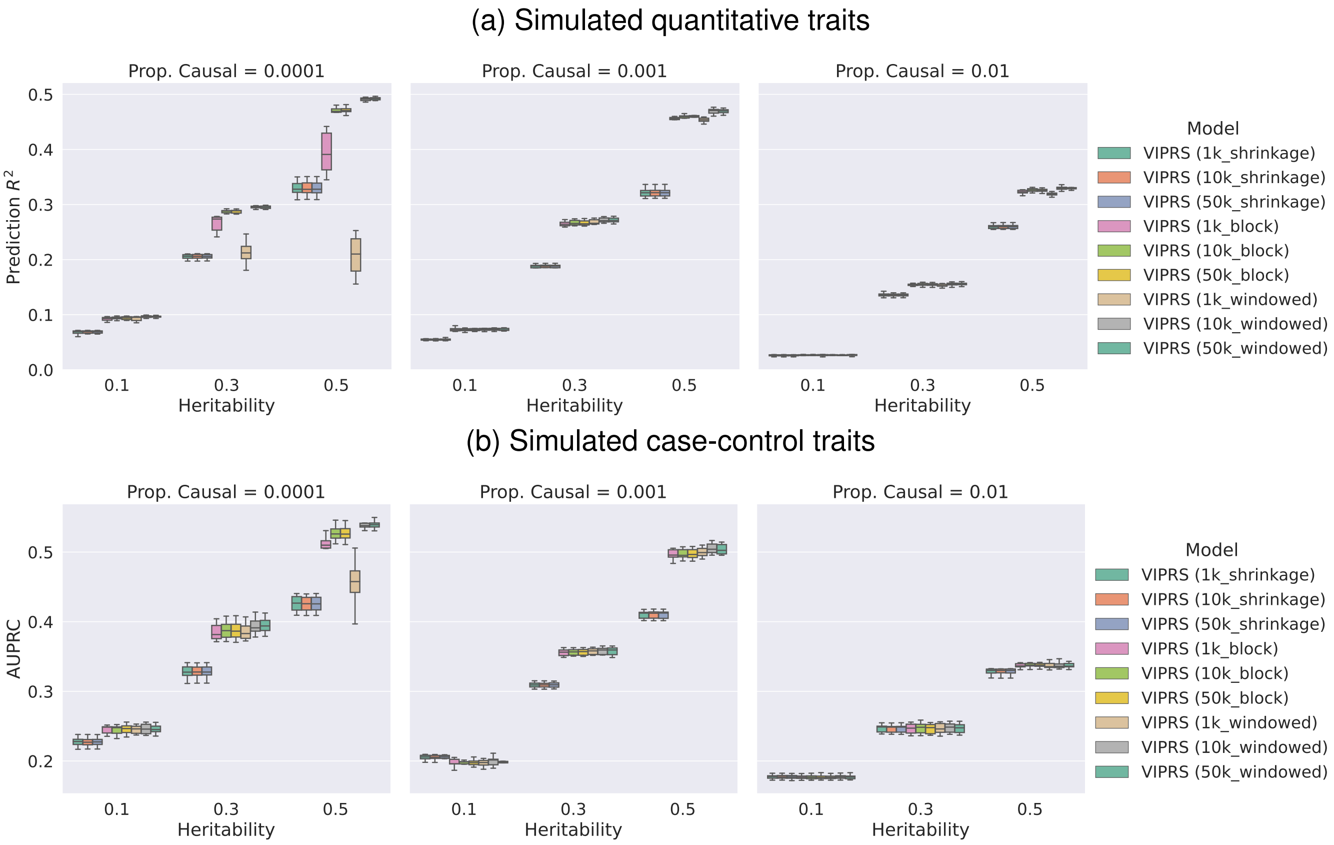
$$T_d(i) = \frac{1}{|\mathcal{S}(i)|} \sum_{j \in \mathcal{S}(i)} T_d(i,j)$$

Given these noisy estimates of the T_d statistic, we then compute a p-value, assuming that it remains roughly Chi-squared distributed with 1 degree-of-freedom [69]. If the p-value is less than a Bonferroni threshold of $\frac{0.05}{M}$, where M is the number of variants included in the model, then those variants are filtered out. In practice, we found that a single round of filtering sometimes does not eliminate all inconsistencies. Thus, in our software pipeline, we follow the authors of the DENTIST method and apply the filters iteratively until model convergence is achieved or we reach a predefined maximum number of iterations.

[1] This behavior has been partially documented in the context of the SbayesR model. See “Achieving SBayesR convergence for Locke et al. 2015 - BMI GWAS and Wood et al. 2014 – Height GWAS.”.

S2 Supplementary Figures

S2.1 Performance of VIPRS with different LD estimators and sample sizes for the LD reference panel



(a) Simulated quantitative traits

(b) Simulated case-control traits

Figure S1: Predictive performance of the VIPRS model on simulated phenotypes using different estimators and sample sizes for the LD reference panel. The phenotypes were simulated using real genotype data from the White British cohort in the UK Biobank ($N = 337,225$), leveraging a subset of 1.1 million HapMap3 variants. The simulation scenarios encompass a total of 9 genetic architectures, spanning 3 values for polygenicity (proportion of causal variants) and 3 values for SNP heritability. For each configuration, we simulated 10 independent phenotypes. Each panel shows results for phenotypes simulated with the pre-specified polygenicity and each column within a panel shows performance metrics for phenotypes simulated with a pre-specified SNP heritability. The performance metrics are **(a)** incremental prediction R^2 for quantitative traits and **(b)** area under the Precision Recall curve (AUPRC) for binary traits. The boxplot for each method and simulation configuration shows the quartiles of predictive scores for the 10 simulated phenotypes. The parenthesis in each method name indicate the sample size of the LD reference panel (1000, 10000, 50000) as well as the LD estimator (shrinkage, block, windowed).

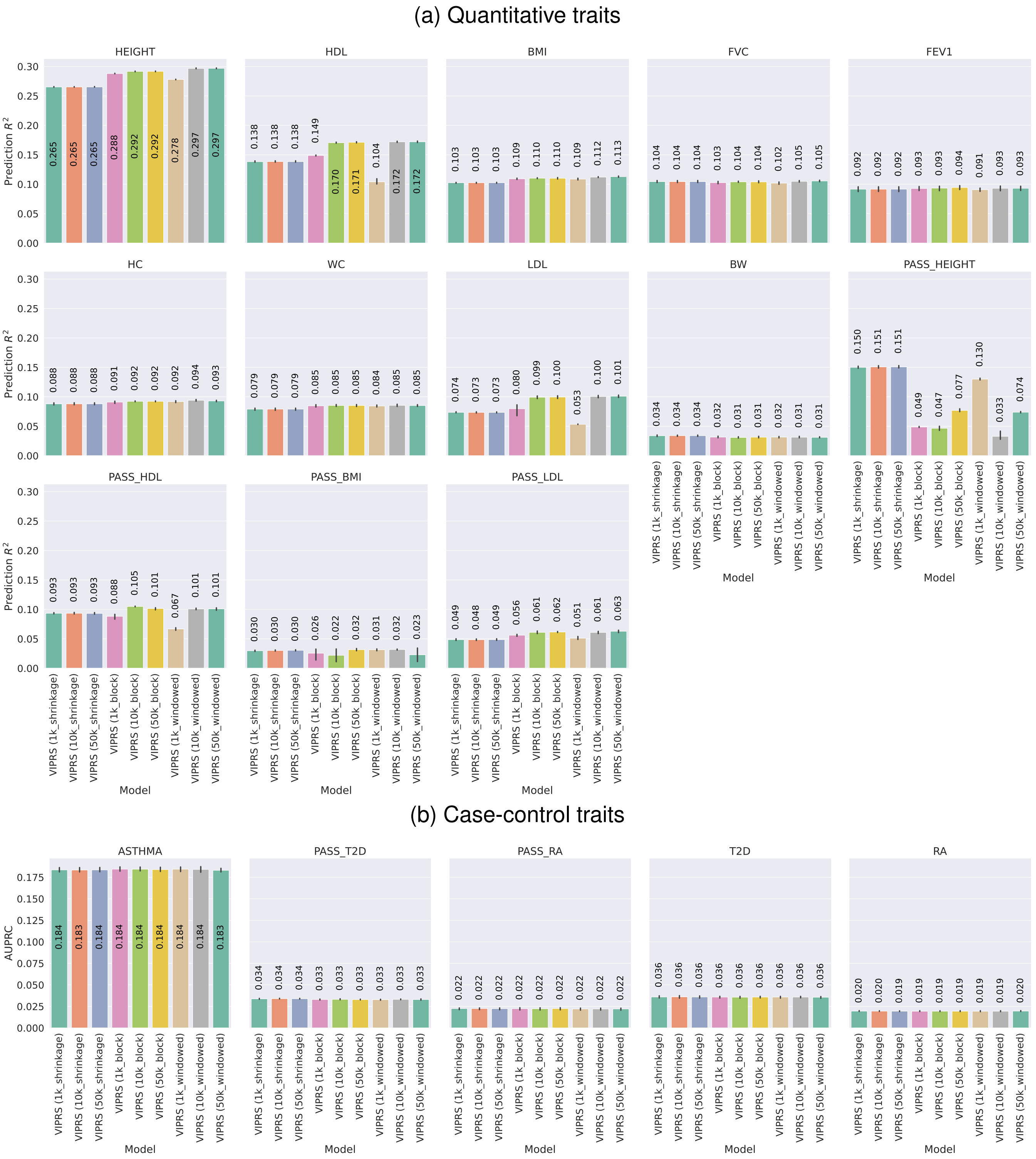
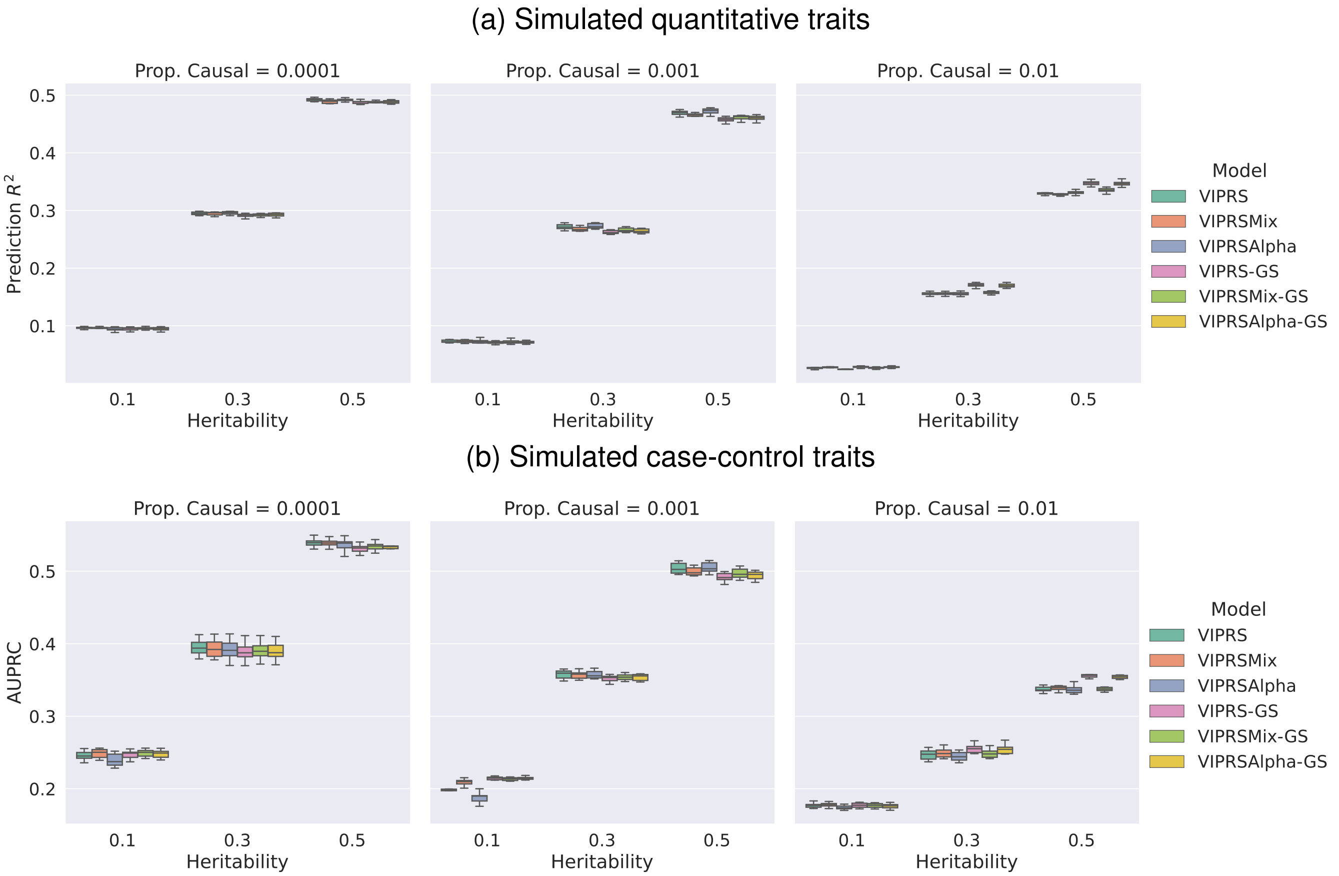


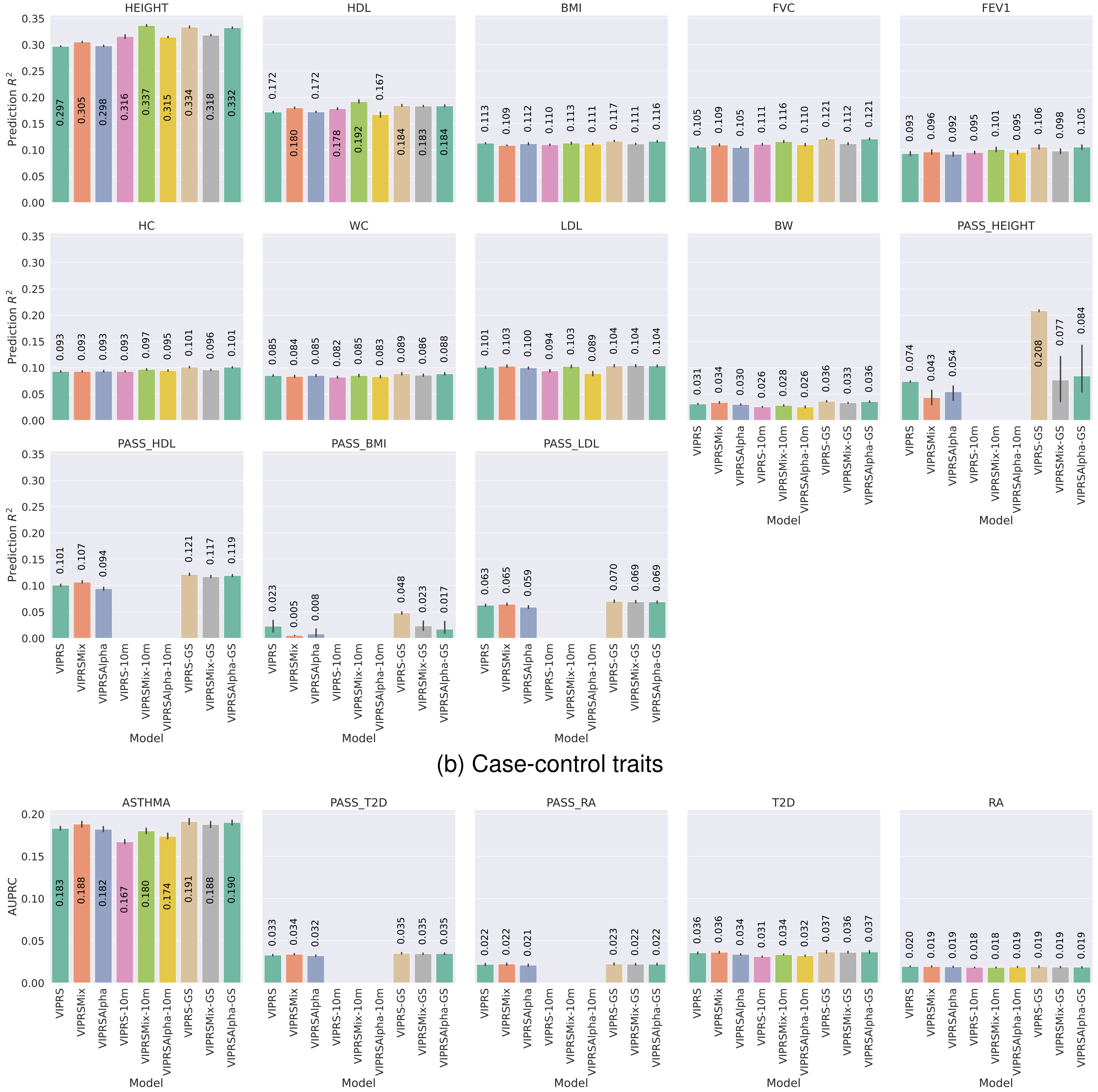
Figure S2: Predictive performance of the `VIPRS` model on real **(a)** quantitative and **(b)** binary phenotypes in the UK Biobank using different estimators and sample sizes for the LD reference panel. Each panel shows the predictive performance, in terms of **(a)** incremental R^2 and **(b)** area under the Precision Recall curve (AUPRC), of each PRS model when applied to a given phenotype. The bars show the mean of the prediction metrics across the 5 folds and the black lines show the corresponding standard errors. The phenotypes analyzed are described in detail in Table 1.The parenthesis in each method name indicate the sample size of the LD reference panel $(1000, 10000, 50000)$ as well as the LD estimator (shrinkage, block, windowed).

S2.2 Performance of the VIPRS model with different families of priors on the effect size


(a) Simulated quantitative traits

(b) Simulated case-control traits

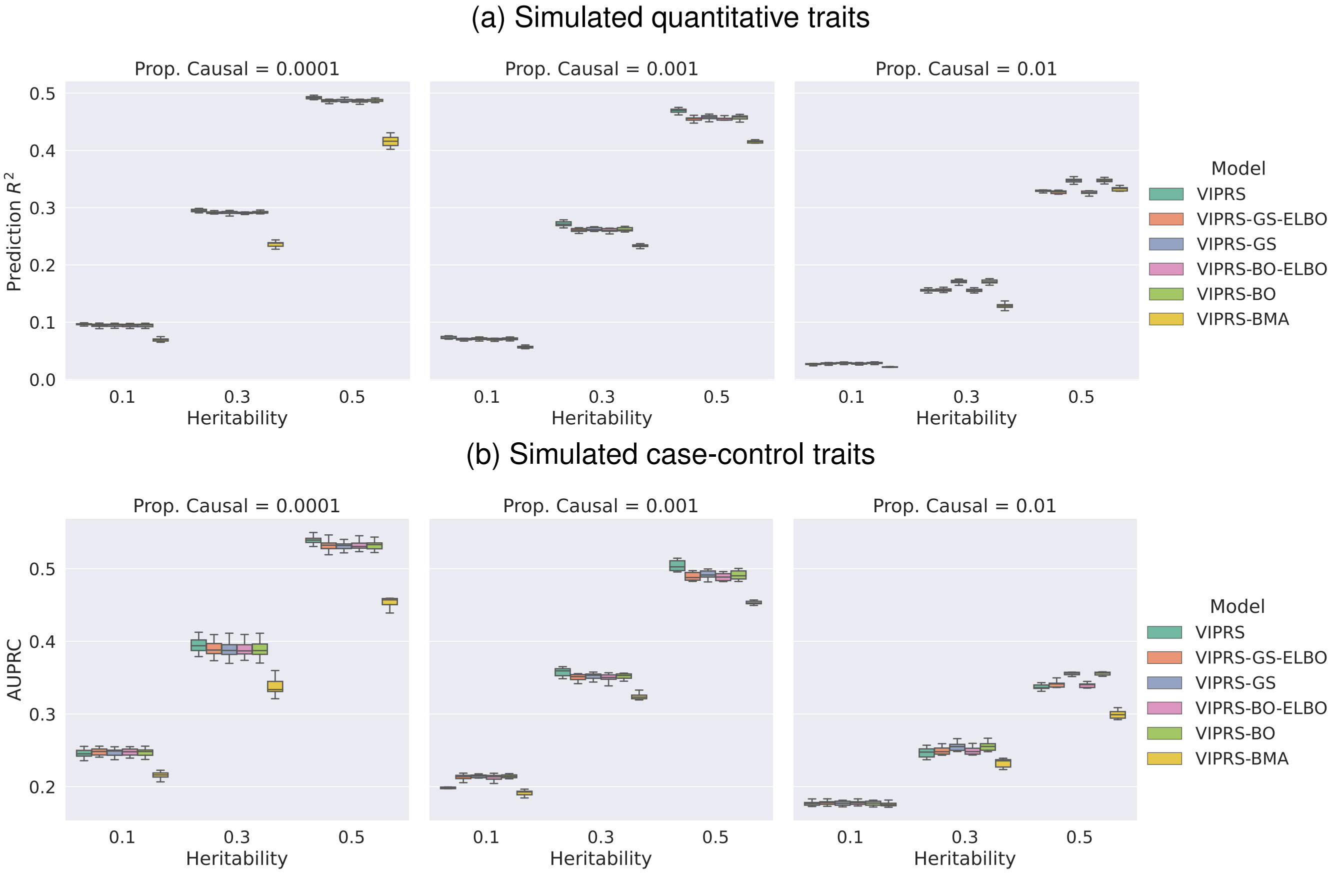
Figure S3: Predictive performance of the VIPRS model on simulated phenotypes using different families of priors on the effect size. The phenotypes were simulated using real genotype data from the White British cohort in the UK Biobank ($N = 337,225$), leveraging a subset of 1.1 million HapMap3 variants. The simulation scenarios encompass a total of 9 genetic architectures, spanning 3 values for polygenicity (proportion of causal variants) and 3 values for SNP heritability. For each configuration, we simulated 10 independent phenotypes. Each panel shows results for phenotypes simulated with the pre-specified polygenicity and each column within a panel shows performance metrics for phenotypes simulated with a pre-specified SNP heritability. The performance metrics are **(a)** incremental prediction R^2 for quantitative traits and **(b)** area under the Precision Recall curve (AUPRC) for binary traits. The boxplot for each method and simulation configuration shows the quartiles of predictive scores for the 10 simulated phenotypes. The methods tested are VIPRS (spike-and-slab prior), VIPRSMix (sparse Gaussian mixture prior with four components), and VIPRSAlpha (spike-and-slab + Alpha prior). In addition, we show the performance of the grid-search version of the model with the aforementioned three priors (VIPRS-GS, VIPRSMix-GS, VIPRSAlpha-GS), where the search is over the proportion of causal variants π.


(a) Quantitative traits

(b) Case-control traits

Figure S4: Predictive performance of the `VIPRS` model on real **(a)** quantitative and **(b)** binary phenotypes in the UK Biobank using different families of priors on the effect size. Each panel shows the predictive performance, in terms of **(a)** incremental R^2 and **(b)** area under the Precision Recall curve (AUPRC), of each PRS model when applied to a given phenotype. The bars show the mean of the prediction metrics across the 5 folds and the black lines show the corresponding standard errors. The phenotypes analyzed are described in detail in Table 1.The methods tested are `VIPRS` (spike-and-slab prior), `VIPRSMix` (sparse Gaussian mixture prior with four components), and `VIPRSAlpha` (spike-and-slab + Alpha prior). These methods were evaluate with the HapMap3 SNPs as well as the 9.6 million expanded set of variants. For the latter case, we added the `10m` suffix to the method name. In addition, we show the performance of the grid-search version of the model with the aforementioned priors (`VIPRS-GS`, `VIPRSMix-GS`, `VIPRSAlpha-GS`), where the search is over the proportion of causal variants π. **NOTE**: We did not run the methods with 9.6 million variants on external summary statistics since they only contained marginal test statistics for a limited set of SNPs.

S2.3 Performance of the `VIPRS` model with different hyperparameter tuning strategies


(a) Simulated quantitative traits

(b) Simulated case-control traits

Figure S5: Predictive performance of the `VIPRS` model on simulated phenotypes using different hyperparameter tuning strategies. The phenotypes were simulated using real genotype data from the White British cohort in the UK Biobank ($N = 337,225$), leveraging a subset of 1.1 million HapMap3 variants. The simulation scenarios encompass a total of 9 genetic architectures, spanning 3 values for polygenicity (proportion of causal variants) and 3 values for SNP heritability. For each configuration, we simulated 10 independent phenotypes. Each panel shows results for phenotypes simulated with the pre-specified polygenicity and each column within a panel shows performance metrics for phenotypes simulated with a pre-specified SNP heritability. The performance metrics are **(a)** incremental prediction R^2 for quantitative traits and **(b)** area under the Precision Recall curve (AUPRC) for binary traits. The boxplot for each method and simulation configuration shows the quartiles of predictive scores for the 10 simulated phenotypes. The hyperparameter tuning strategies tested are `VIPRS` (Variational EM), `VIPRS-GS-ELBO` (grid search with the ELBO as criterion), `VIPRS-GS` (grid search with accuracy on the validation set as criterion), `VIPRS-BO-ELBO` (Bayesian optimization with the ELBO as criterion), `VIPRS-BO` (Bayesian optimization with accuracy on the validation set as criterion), and `VIPRS-BMA` (Bayesian Model Averaging). In these experiments, the hyperparameter search is over the proportion of causal variants π.

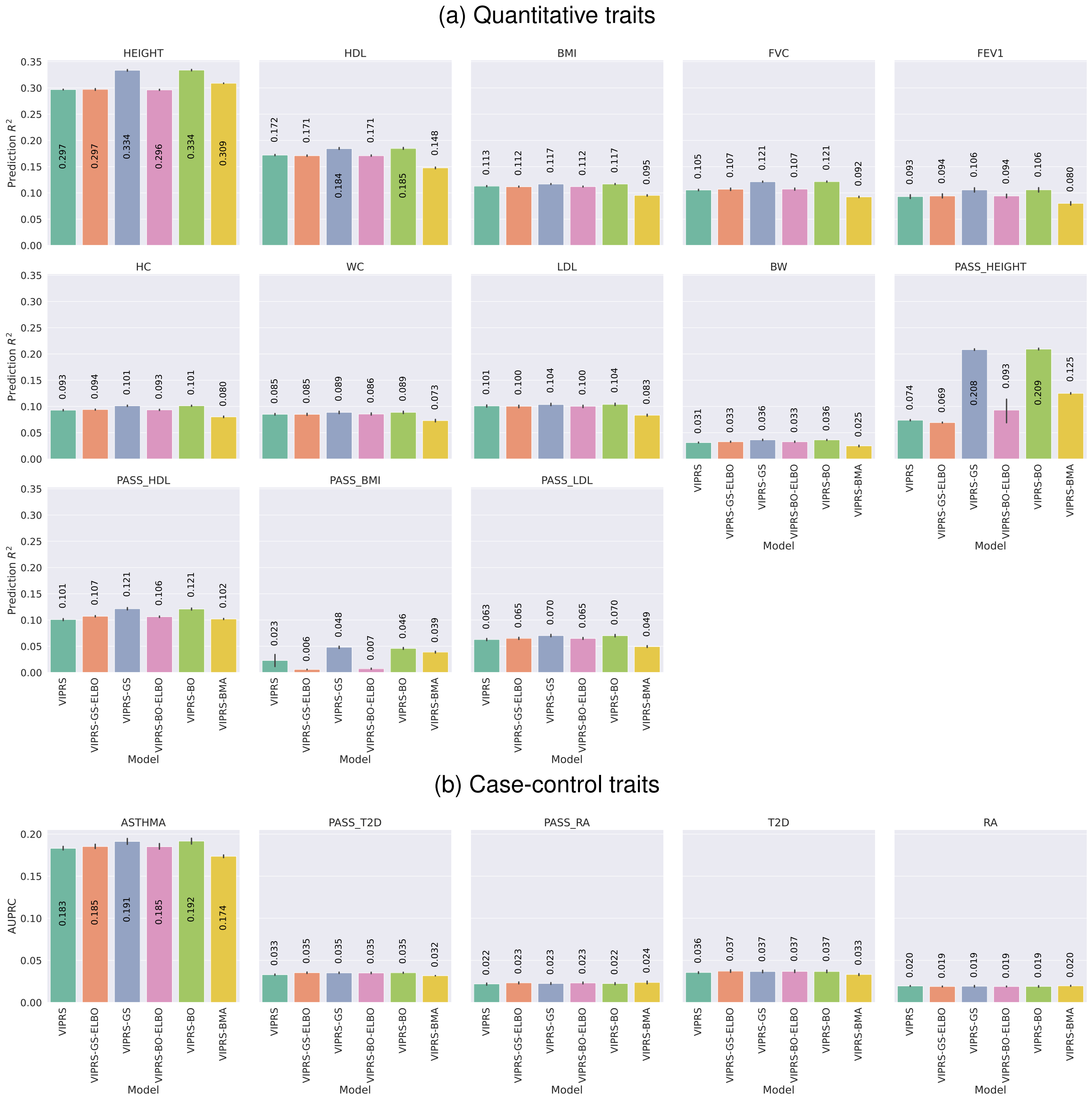
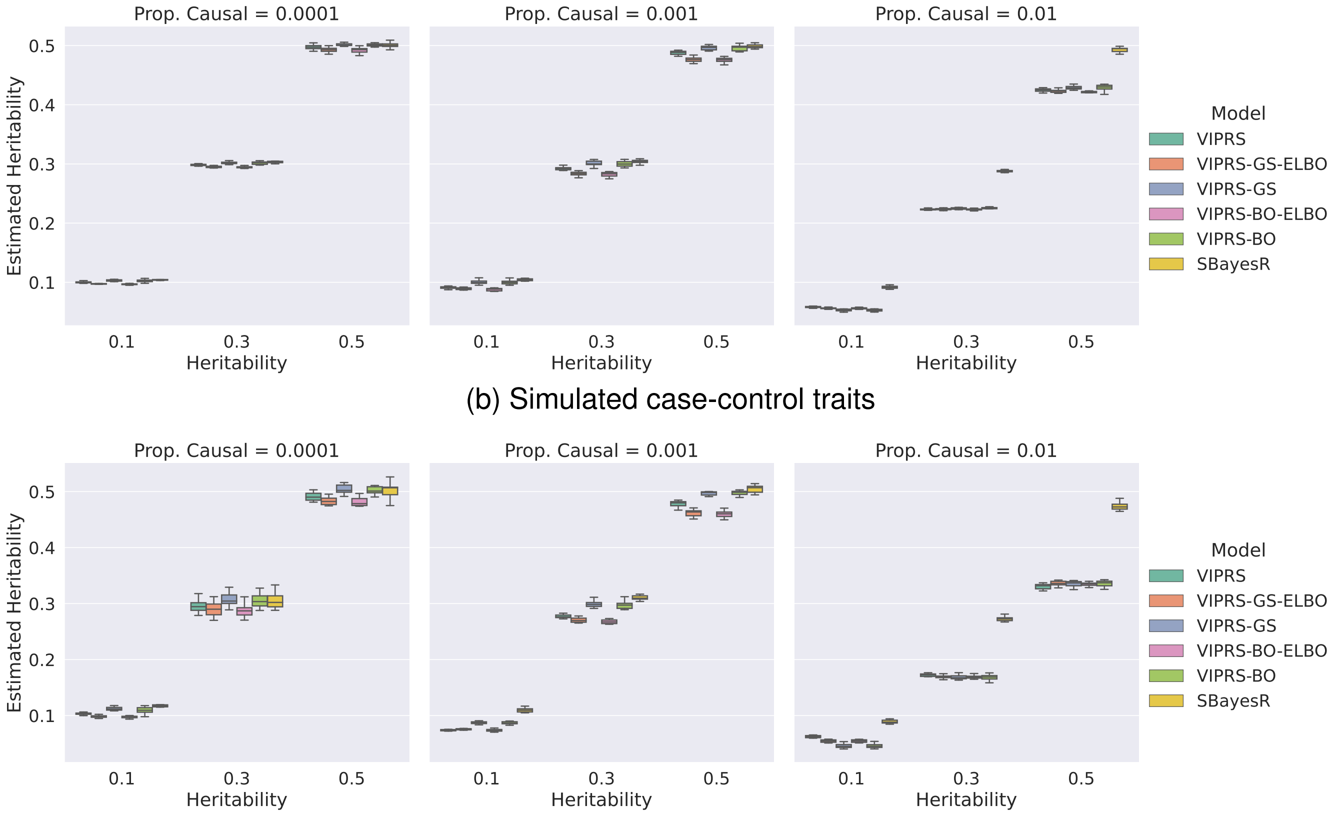


Figure S6: Predictive performance of the `VIPRS` model on real **(a)** quantitative and **(b)** binary phenotypes in the UK Biobank using different hyperparameter tuning strategies. Each panel shows the predictive performance, in terms of **(a)** incremental R^2 and **(b)** area under the Precision Recall curve (AUPRC), of each PRS model when applied to a given phenotype. The bars show the mean of the prediction metrics across the 5 folds and the black lines show the corresponding standard errors. The phenotypes analyzed are described in detail in Table 1.The hyperparameter tuning strategies tested are `VIPRS` (Variational EM), `VIPRS-GS-ELBO` (grid search with the ELBO as criterion), `VIPRS-GS` (grid search with accuracy on the validation set as criterion), `VIPRS-BO-ELBO` (Bayesian optimization with the ELBO as criterion), `VIPRS-BO` (Bayesian optimization with accuracy on the validation set as criterion), and `VIPRS-BMA` (Bayesian Model Averaging). In these experiments, the hyperparameter search is over the proportion of causal variants π.

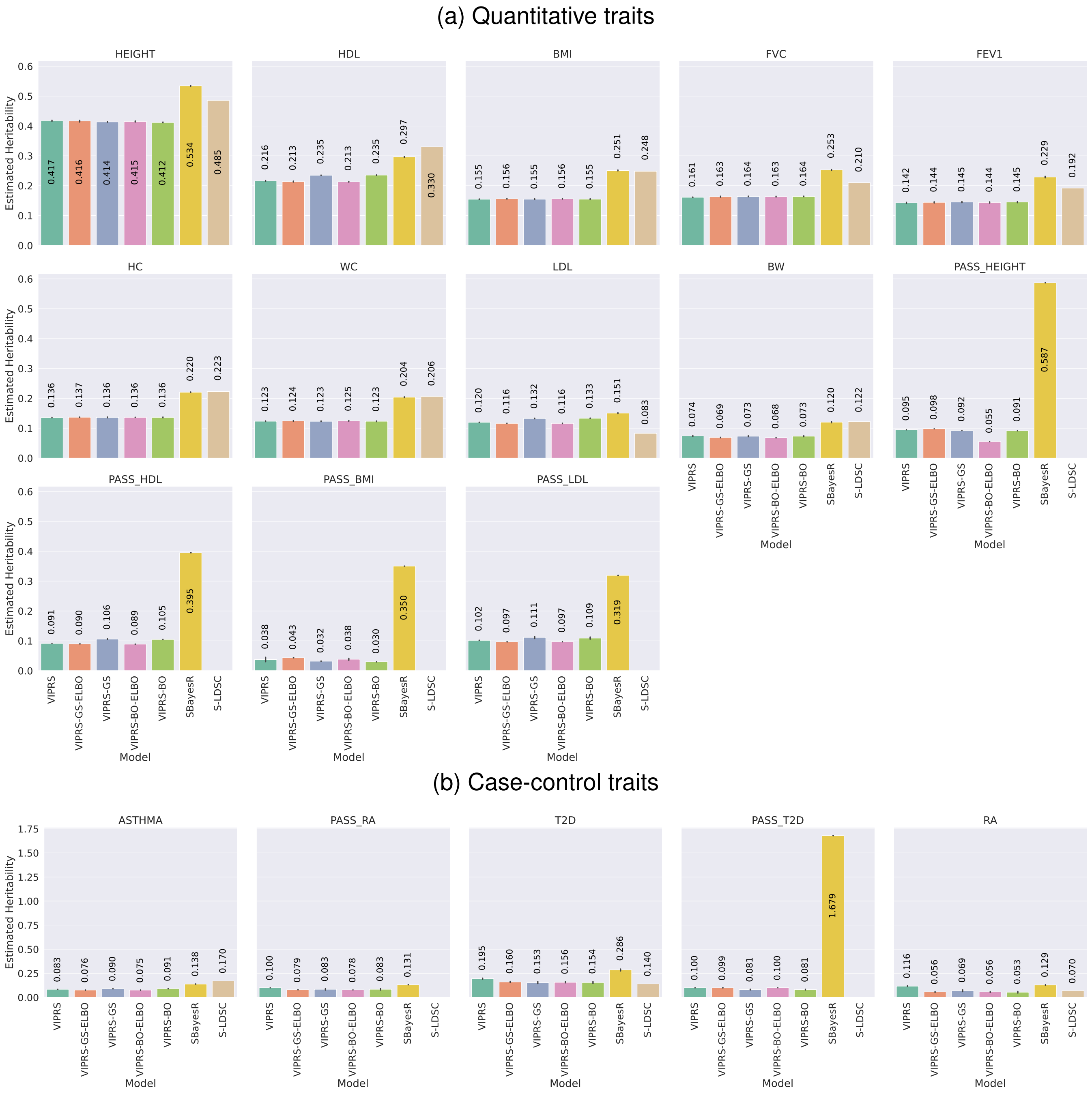
S2.4 Hyperparameter estimation: SNP heritability and polygenicity



(a) Simulated quantitative traits

(b) Simulated case-control traits

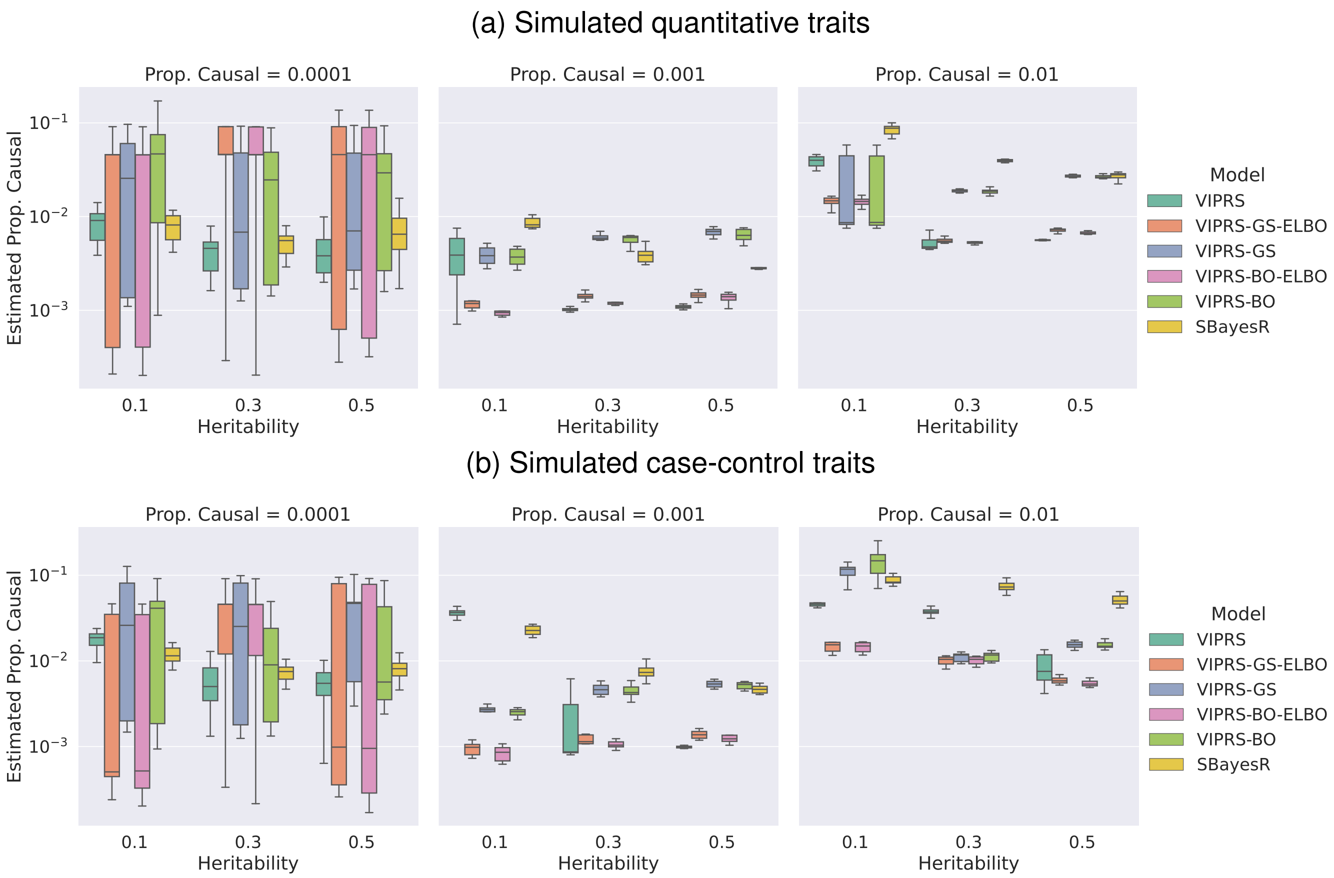
Figure S7: Estimates of SNP heritability or proportion of variance explained (PVE) by summary statistics-based PRS methods for simulated **(a)** quantitative and **(b)** case-control traits. The phenotypes were simulated using real genotype data from the White British cohort in the UK Biobank ($N = 337,225$), leveraging a subset of 1.1 million HapMap3 variants. The simulation scenarios encompass a total of 9 genetic architectures, spanning 3 values for polygenicity (proportion of causal variants) and 3 values for SNP heritability. For each configuration, we simulated 10 independent phenotypes. Each panel shows results for phenotypes simulated with the pre-specified polygenicity and each column within a panel shows performance metrics for phenotypes simulated with a pre-specified SNP heritability. The boxplot for each method and simulation configuration shows the quartiles of heritability estimates for the 10 simulated phenotypes. The methods tested are VIPRS (Variational EM), VIPRS-GS-ELBO (grid search with the ELBO as criterion), VIPRS-GS (grid search with accuracy on the validation set as criterion), VIPRS-BO-ELBO (Bayesian optimization with the ELBO as criterion), VIPRS-BO (Bayesian optimization with accuracy on the validation set as criterion), and SBayesR.



(a) Quantitative traits

(b) Case-control traits

Figure S8: Estimates of SNP heritability or proportion of variance explained (PVE) by summary statistics-based PRS methods for real **(a)** quantitative and **(b)** case-control traits in the UK Biobank. The bars show the mean estimate of SNP heritability across the 5 folds and the black lines show the corresponding standard errors. The phenotypes analyzed are described in detail in Table 1. The methods tested are `VIPRS` (Variational EM), `VIPRS-GS-ELBO` (grid search with the ELBO as criterion), `VIPRS-GS` (grid search with accuracy on the validation set as criterion), `VIPRS-BO-ELBO` (Bayesian optimization with the ELBO as criterion), `VIPRS-BO` (Bayesian optimization with accuracy on the validation set as criterion), `SBayesR`, and, for reference, `S-LDSC` (estimates are taken from Ben Neale lab's Heritability Browser: `https://nealelab.github.io/UKBB_ldsc/h2_browser.html`).



(a) Simulated quantitative traits

(b) Simulated case-control traits

Figure S9: Estimates of polygenicity or proportion of causal variants by summary statistics-based PRS methods for simulated **(a)** quantitative and **(b)** case-control traits. The phenotypes were simulated using real genotype data from the White British cohort in the UK Biobank ($N = 337,225$), leveraging a subset of 1.1 million HapMap3 variants. The simulation scenarios encompass a total of 9 genetic architectures, spanning 3 values for polygenicity (proportion of causal variants) and 3 values for SNP heritability. For each configuration, we simulated 10 independent phenotypes. Each panel shows results for phenotypes simulated with the pre-specified polygenicity and each column within a panel shows performance metrics for phenotypes simulated with a pre-specified SNP heritability. The boxplot for each method and simulation configuration shows the quartiles of polygenicity estimates for the 10 simulated phenotypes. The methods tested are `VIPRS` (Variational EM), `VIPRS-GS-ELBO` (grid search with the ELBO as criterion), `VIPRS-GS` (grid search with accuracy on the validation set as criterion), `VIPRS-BO-ELBO` (Bayesian optimization with the ELBO as criterion), `VIPRS-BO` (Bayesian optimization with accuracy on the validation set as criterion), and `SBayesR`.

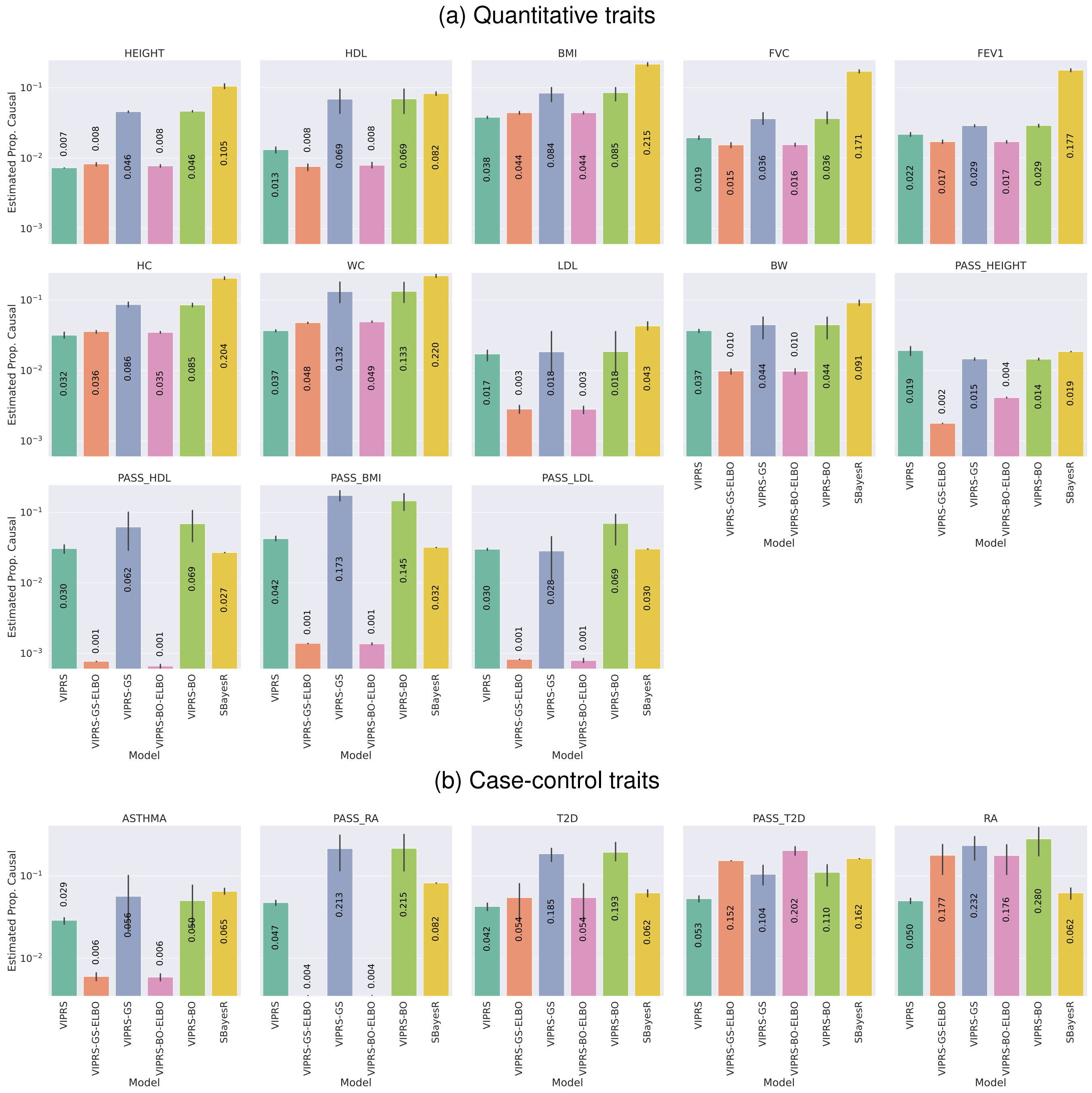


Figure S10: Estimates of polygenicity or proportion of causal variants by summary statistics-based PRS methods for real **(a)** quantitative and **(b)** case-control traits in the UK Biobank. The bars show the mean estimate of polygenicity across the 5 folds and the black lines show the corresponding standard errors. The phenotypes analyzed are described in detail in Table 1. The methods tested are VIPRS (Variational EM), VIPRS-GS-ELBO (grid search with the ELBO as criterion), VIPRS-GS (grid search with accuracy on the validation set as criterion), VIPRS-BO-ELBO (Bayesian optimization with the ELBO as criterion), VIPRS-BO (Bayesian optimization with accuracy on the validation set as criterion), and SBayesR.

S2.5 Correlation between polygenic score estimates of different PRS methods

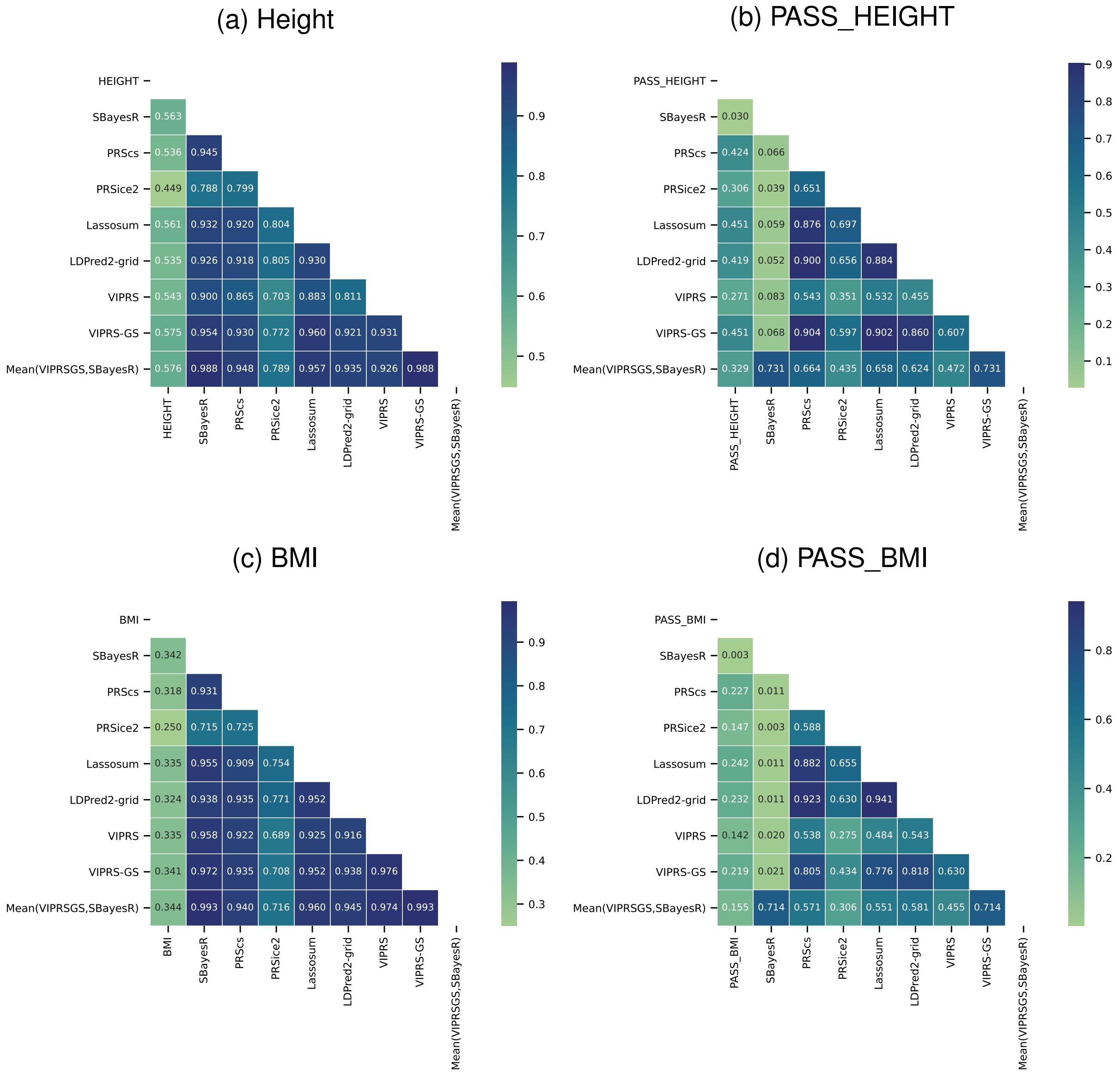


Figure S11: Pearson correlation coefficient between PRS estimates for the UK Biobank individuals derived from different summary statistics-based PRS methods for **(a-b)** standing height (HEIGHT) and **(c-d)** body mass index (BMI) . Each panel shows a heatmap for a different phenotype and/or GWAS summary statistics source. Panels **(a-b)** show the correlations between PRS estimates for standing height where the summary statistics were derived from **(a)** training samples within the UK Biobank and **(b)** Allen et al. (2010) [70]. Panels **(c-d)** show the correlations between PRS estimates for BMI where the summary statistics were derived from **(c)** training samples within the UK Biobank and **(d)** Speliotes et al. 2010 [72]. Each cell shows the Pearson correlation coefficient between a pair of PRS methods. In addition to the phenotype itself, we show the PRS correlations between SBayesR, PRScs, PRSice2, Lassosum, LDPred2-grid, VIPRS, VIPRS-GS, and Mean(VIPRSGS,SBayesR) (the averaged PRS estimate between SBayesR and VIPRS-GS).

S2.6 Relationship between the ELBO and validation R^2 in analyses with real quantitative traits

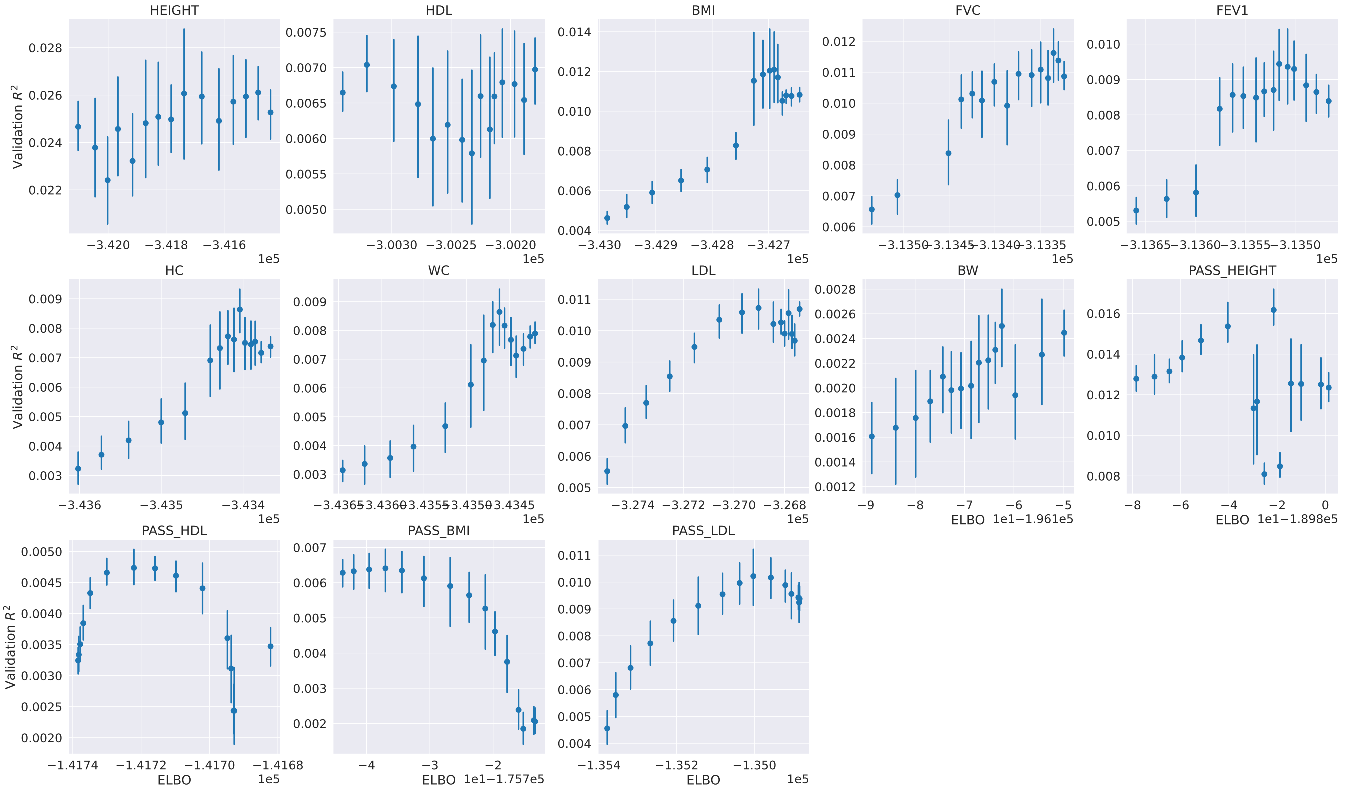


Figure S12: The correspondence between the Evidence Lower BOund (ELBO) of the `VIPRS-GS` models at convergence and the prediction R^2 on the held-out validation set in the UK Biobank. When we train the `VIPRS-GS` version of the model, we are in effect training 30 separate `VIPRS` models with different π values. At convergence, we store the maximum value for the ELBO as well as the prediction R^2 on the validation set. Here, we compare the ELBO value at convergence to the corresponding validation R^2. The phenotypes analyzed are described in detail in Table 1. For readability, for each phenotype and/or GWAS data source (panel), the values for the ELBO on the x-axis were grouped into 15 bins. The dots show the mean validation R^2 within that bin and across the 5 training folds and the vertical lines show the corresponding standard errors. Since we perform model fit on each chromosome separately, this figure shows the metrics from the model training with chromosome 2.