A Simple Approach for Local and Global Variable Importance in Nonlinear Regression Models

Emily T. Winn-Nuñez[1,†], Maryclare Griffin[2], and Lorin Crawford[3-5,†]

1 Division of Applied Mathematics, Brown University, Providence, RI, USA
2 Department of Mathematics and Statistics, University of Massachusetts Amherst, Amherst, MA, USA
3 Microsoft Research New England, Cambridge, MA, USA
4 Department of Biostatistics, Brown University, Providence, RI, USA
5 Center for Computational Molecular Biology, Brown University, Providence, RI, USA

† Corresponding E-mails: emily_winn@brown.edu; lcrawford@microsoft.com

Abstract

The ability to interpret machine learning models has become increasingly important as their usage in data science continues to rise. Most current interpretability methods are optimized to work on either (*i*) a global scale, where the goal is to rank features based on their contributions to overall variation in an observed population, or (*ii*) the local level, which aims to detail on how important a feature is to a particular individual in the dataset. In this work, we present the "GlObal And Local Score" (GOALS) operator: a simple *post hoc* approach to simultaneously assess local and global feature variable importance in nonlinear models. Motivated by problems in statistical genetics, we demonstrate our approach using Gaussian process regression where understanding how genetic markers affect trait architecture both among individuals and across populations is of high interest. With detailed simulations and real data analyses, we illustrate the flexible and efficient utility of GOALS over state-of-the-art variable importance strategies.

Introduction

Over the past decade, "interpretability" has become a major focus in statistical and probabilistic machine learning — particularly in the space of biomedicine. While there remains to be a universal definition for what makes a computational method interpretable (e.g., Carvalho et al., 2019; Guidotti et al., 2018; Hall, 2019), it generally refers to a model's "ability to explain or to present in understandable terms to a human" (e.g., Doshi-Velez and Kim, 2017). The simple structure of linear models gives an intrinsic interpretation to their parameters and, as a result, enables them to be used for downstream tasks that extend beyond prediction. For example, in biomedical applications, linear models have been widely implemented to identify differentially expressed genes and enriched signaling pathways in functional genomics (e.g., Love et al., 2014; Nueda et al., 2014; Ritchie et al., 2015; Robinson et al., 2009), characterize complex trait architecture in genome-wide association studies (e.g., Hayeck et al., 2015; Heckerman et al., 2016; Jiang et al., 2019; Kang et al., 2008, 2010; Korte et al., 2012; Lippert et al., 2011; Loh et al., 2018; Price et al., 2010; Runcie and Crawford, 2019; Zeng and Zhou, 2017; Zhou and Stephens, 2012), and estimate the underlying generative model of gene networks (e.g., Karlebach and Shamir, 2008; Ma et al., 2018; Manno et al., 2018). Part of the utility of these approaches is their ability to provide statistical significance measures such as P-values, posterior inclusion probabilities (PIPs), or Bayes factors — all of which lend a notion of statistical evidence about how important each feature is in explaining an outcome variable. Unfortunately, linear models can be inappropriate or infeasible in practice. The strict additive assumptions underlying linear regression can be a hinderance in many supervised learning tasks where the

variation of a measured response is dominated by nonlinear interactions. As data collection technologies continue to advance, even the most powerful linear models have struggled scale to high dimensions due to both inefficient model fitting procedures (Runcie and Crawford, 2019; Runcie et al., 2021) and increasingly large combinatorial feature spaces when searching over both additive and non-additive effects (e.g., Crawford et al., 2017; Stamp et al., 2022).

Machine learning methods can overcome limitations of linear regression by accommodating nonlinear relationships between features (e.g., through activation units in neural networks or via nonparametric covariance functions in Gaussian processes) and implement scalable training algorithms. However, many machine learning methods are also known to be "black box" since they are not inherently transparent about how parameters are learned in making decisions and predicting outcomes (e.g., DeGrave et al., 2021; Rudin, 2019, 2022). Classically, there are two strategies to achieving interpretability of machine learning methods. The first solution attempts to achieve intrinsic interpretability by limiting the architecture of machine learning methods to simple structures (Ai and Narayanan. R, 2021). As an example, in the biomedical sciences, a recent trend has been to develop customized neural network architectures that are inspired by biological systems (e.g., Bourgeais et al., 2021, 2022; Demetci et al., 2021; Elmarakeby et al., 2021). Rather having fully connected, potentially over-parameterized architectures, these frameworks have partially connected architectures that are based on biological annotations in the literature or derived from other functional relationships that have been identified through experimental validation. Each neural network node then has an intrinsic interpretation because they encode some biological unit (e.g., signaling pathways, protein motif, or gene regulatory network) and each neural network weight connecting nodes represent known relationships between the corresponding entities. A key aspect of this approach is that it depends on reliable domain knowledge to generate these architectures. When this level of information is not available, as is the case for many practical scientific problems, implementing this strategy can be extremely challenging.

The second strategy to gain interpretability uses *post hoc* or *auxiliary* methods to assess the importance of features after a model has been trained. A wide range of such approaches have been proposed in the literature. Although many of these techniques share theoretical connections (Lundberg and Lee, 2016), they generally can be separated into two categories. The first group of methods are "salience methods" (also commonly known as "saliency maps"; Simonyan et al., 2014) which, in their simplest form, provide variable importance by calculating the gradient of a model loss function with respect to each input feature for a class of interest. Kindermans et al. (2019) showed that these types of attribution approaches can be highly unreliable in the presence of simple noise structures. In this paper, we will focus on a second class of explainable methods that produce "sensitivity scores" which quantify variable importance by measuring the amount predictive accuracy that is lost when a particular feature is perturbed. Common examples in this second class of methods include information criterion (Gelman et al., 2014), distributional centrality measures (Crawford et al., 2019; Paananen et al., 2019, 2021; Piironen and Vehtari, 2016, 2017; Woo et al., 2015), Shapley Additive Explanations (SHAP) (Chen et al., 2022; Lundberg and Lee, 2017), and knockoffs (Candès et al., 2018; Sesia et al., 2020, 2021). Each of these methods have been shown to have their advantages, but one limitation they all have in common is that they mainly focus on addressing either (*i*) global interpretability where the goal is to rank/select input features based on their contributions to overall variation in an observed population, or (*ii*) local interpretability which aims to detail how important a feature is to any particular individual in the dataset. In many biomedical applications, it would be ideal to have a measure that leads to conclusions on both scales, simultaneously. For example, in statistical genetics, it is important to understand how a genetic variant contributes to the architecture of a complex trait — but, for the purpose of precision medicine, it is also important to understand how that variant might have disproportionate effects on individuals coming from different subpopulations (e.g., Martin et al., 2019).

In this work, we present the "GlObal And Local Score" (GOALS) operator: a simple approach that builds off of the distributional centrality literature to provide a measure that assesses both local and

global variable importance for features, simultaneously. Our method is entirely general with respect to the modeling approach taken. The only requirements are that we have access to the fitted model and the ability to generate out-of-sample predictions. As an illustration of our approach, we focus on using Gaussian process regression. However, also note that this variable importance approach immediately applies to other methodologies such as neural networks (Richard and Lippmann, 1991). We assess our proposed approach in the context of association mapping (i.e., inference on significant variants or loci) in statistical genetics as a way to highlight data science applications that (*i*) contain outcomes that are driven by many covarying and interacting predictors (e.g., epistasis or gene-by-gene interactions; Crawford et al., 2017) and (*ii*) can contain diverse subsets of populations where the importance of features may not be uniform across all individuals in the data. The remainder of the paper is organized as follows. First, we briefly detail the distributional centrality framework for achieving interpretability in nonlinear regression. Here, we review Gaussian processes, motivate the need for an effect size (regression coefficient) analog for input features, and define the concept of relative centrality which can be used to perform variable importance. In the next section, we derive the GOALS operator and detail its ability to make local and global interpretations for features. Lastly, we show the utility of our methodology with extensive simulations and a real data analysis of complex traits assayed in a heterogenous stock of mice from Wellcome Trust Centre for Human Genetics (Valdar et al., 2006a,b).

Overview: Distributional Centrality for Nonlinear Models

In this work, we will follow positions taken by previous studies and assume that an interpretable statistical method is made up of three key components: (*i*) a motivating probabilistic model, (*ii*) a notion of an effect size (or regression coefficient) for each genetic variant and (*iii*) a statistical metric that determines marker significance according to a well-defined null hypothesis (Crawford et al., 2019). The third component is commonly defined by the task of achieving either global or local interpretability. From a genetics perspective, the main objective of global interpretability is to rank each genetic variant based on its contributions to overall phenotypic variation in an observed population. In contrast, local interpretability aims to provide an explanation on how important a variant is to any particular individual in the data set. The purpose of this section is to review background which allows us to formulate all three of these key components within the context of Bayesian Gaussian process regression for continuous quantitative traits; however, note that extending this theoretical framework to both other nonlinear and nonparametric methodologies (e.g., neural networks; Ish-Horowicz et al., 2019); as well as to categorical phenotypes (e.g., binary traits in case-control studies) (e.g., Zhang et al., 2011) is straightforward. In terms of global interpretability, we will introduce the concept of an effect size analog and describe how distributional centrality measures can be used to perform *post hoc* variable prioritization (also sometimes referred to as performing "variable importance" in certain areas of the literature). We then comment on the landscape of existing approaches to assess local interpretability within these same methods and discuss some the need for unifying these concepts for GWA studies.

Weight-Space Gaussian Process Regression

Consider data from a GWA study with N individuals. We have an N-dimensional vector of quantitative traits $\mathbf{y}$ and an $N \times J$ genotype matrix $\mathbf{X}$ with J denoting the number of single nucleotide polymorphisms (SNPs) encoded as $\{0, 1, 2\}$ copies of a reference allele at each locus. To build intuition, we begin by considering a standard linear model for phenotypes such that

$$\mathbf{y} = \boldsymbol{f} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{f} = \mathbf{X}\boldsymbol{\beta}, \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \tau^2\mathbf{I}) \tag{1}$$

where the function to be estimated $\boldsymbol{f}$ is assumed to be a linear combination of SNPs in $\mathbf{X}$ and their respective effects denoted by the J-dimensional vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_J)$ additive coefficients, $\boldsymbol{\varepsilon}$ is a

normally distributed error term with mean zero and scaled variance term τ^2, and $\mathbf{I}$ denotes an $N \times N$ identity matrix. For convenience, we will assume that the trait of interest has been mean-centered and standardized. In this setting, the phenotypic variance for the trait $\mathbb{V}[\mathbf{y}] = 1$, the proportion of variance explained by genetics $\mathbb{V}[\mathbf{X}\boldsymbol{\beta}] = h^2$ is referred to as the narrow-sense heritability, and $\tau^2 = (1 - h^2)$ makes up the remaining variance (Bulik-Sullivan et al., 2015).

It has been well documented in the literature that the genetic architecture of many complex traits can be driven by nonlinear phenomena such as dominance and epistasis (i.e., gene-by-gene interactions Mackay, 2014; Phillips, 2008). In these cases, the assumption in Eq. (1) that phenotypic variation can be fully explained by additive genetic effects is restrictive. One way to overcome this limitation is to conduct model inference within a high-dimensional function space. In this work, we take a nonparametric approach and conduct inference in reproducing kernel Hilbert space (RKHS) by specifying a Gaussian process (GP) prior

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k_\theta(\mathbf{x}, \mathbf{x}')) \tag{2}$$

where $f(\bullet)$ is defined by its mean function $m(\bullet)$ (which we will consider to be fixed at zero) and positive definite covariance function $k(\bullet, \bullet)$. In practice, we assume that our model is only evaluated on the N individuals in our data. When conditioning on these finite observations (or finite set of locations), the GP prior in Eq. (2) becomes a multivariate normal distribution (Kolmogorov and Rozanov, 1960; Rasmussen and Williams, 2006) and we can write the following "weight-space" nonlinear model for complex traits

$$\mathbf{y} = \boldsymbol{f} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{f} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}), \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \tau^2\mathbf{I}). \tag{3}$$

Here, $\boldsymbol{f} = [f(\mathbf{x}_1), \ldots, f(\mathbf{x}_N)]$ is an N-dimensional normally distributed random variable with mean vector $\mathbf{0}$, and the covariance matrix $\mathbf{K}$ is computed with each element given by $k_{ii'} = k(\mathbf{x}_i, \mathbf{x}_{i'})$ where $\mathbf{x}_i$ and $\mathbf{x}_{i'}$ denote the genotypes of the i-th and i'-th individual, respectively. Many covariance functions have been shown to implicitly account for higher-order interactions between features, which often lead to more accurate characterization of complex data types (Cotter et al., 2011; Demetci et al., 2021; Murdoch et al., 2019; Tsang et al., 2018a,b; Wahba, 1990). Without loss of generality, we will use a radial basis function such that $k_{ii'} = \exp\{-\theta\|\mathbf{x}_i - \mathbf{x}_{i'}\|^2\}$ where the bandwidth parameter θ is set using the "median criterion" approach to maintain numerical stability and avoid additional computational costs (Chaudhuri et al., 2017).

Altogether, there are a few important points to note between the linear model in Eq. (1) and the "weight-space" GP regression model in Eq. (3). First, the GP regression model can be seen as a generalization of a linear mixed model which uses a nonlinear covariance $\mathbf{K}$ to account for genetic interactions instead of the usual (additive) gram matrix $\mathbf{X}\mathbf{X}^\intercal/J$ (e.g., Lippert et al., 2011; Zhou and Stephens, 2012). Second, under the standardized phenotype assumption, the proportion of variation explained by function in the GP model $\mathbb{V}[\boldsymbol{f}] = H^2$ is referred to as the broad-sense heritability of a trait since it accounts for both additive and non-additive genetic effects, and the scaled variance term is now $\tau^2 = (1 - H^2)$. Lastly, like linear regression, the GP model can also be easily extended to accommodate other fixed effects (e.g., age, sex or genotype principal components) (de los Campos et al., 2009; Shi et al., 2012) as well as be adapted to account for interactions between variants and non-genetic risk factors (e.g., environment) (Cuevas et al., 2017; Weissbrod et al., 2016). We will not explicitly consider these fixed effects here and, instead, will leave those explorations to the reader.

Effect Size Analogs and Relative Centrality Measures

In this section, we assume access to some trained Bayesian model with the ability to fully characterize or draw samples from its posterior predictive distribution. A central goal in GWA studies is to jointly infer the true global effect and statistical significance of each SNP given both genotypic and phenotypic measurements for an observed population. One classic strategy to estimating regression coefficients in

Eq. (1) is to use least squares where the response variable is projected onto the column space of the data $\widehat{\boldsymbol{\beta}} := \text{Proj}(\mathbf{X}, \mathbf{y}) = \mathbf{X}^{\dagger}\mathbf{y}$ with $\mathbf{X}^{\dagger}$ denoting some generalized inverse since $J > N$ in many GWA applications. We refer to the vector $\widehat{\boldsymbol{\beta}} = [\widehat{\beta}_1, \ldots, \widehat{\beta}_J]$ as the (additive) effect size for each SNP in the data set.

The effect size analog was developed with the intention of being the nonparametric version of a regression coefficient for each input feature of a nonlinear model (Crawford et al., 2018). In general, this leverages the idea that, $\mathbb{E}[\mathbf{y} \,|\, \mathbf{X}] = \boldsymbol{f}$ when conditioning on N finite observations in Eq. (3). Thus, similar to the linear regression case, the effect size analog can be defined by projecting the smooth nonlinear function onto the column space of the data. While there are many projections one can use (e.g., Kowal, 2021; Woody et al., 2021), we will consider the following least squares-like projection where

$$\widetilde{\boldsymbol{\beta}} := \text{Proj}(\mathbf{X}, \boldsymbol{f}) = \mathbf{X}^{\dagger}\boldsymbol{f}. \tag{4}$$

This is a simple way of understanding the genetic relationships that the model has learned. Under the linear projection in Eq. (4), the effect size analogs have the usual interpretation. For example, while holding everything else constant, increasing the j-th feature by 1 will increase $\boldsymbol{f}$ by $\widetilde{\beta}_j$ (Crawford et al., 2018). Importantly, because of the closed-form projection, drawing samples from the posterior distribution of $\boldsymbol{f}$ can be deterministically transformed to samples from the implied posterior distribution of the effect size analogs.

Similar to regression coefficients in linear models, the effect size analog is not enough on its own to determine variable importance. Indeed, there are many ways to achieve global interpretability based on the magnitude of effect size estimates (e.g., Barbieri and Berger, 2004; Hoti and Sillanpää, 2006; Stephens and Balding, 2009), but many of these approaches rely on arbitrary thresholding and fail to theoretically test a null hypothesis. One analogy to traditional Bayesian hypothesis testing for nonparametric regression methods is a *post hoc* approach for association mapping via a series of "distributional centrality measures" using Kullback–Leibler divergence (KLD) (e.g., Alaa and van der Schaar, 2017; Goutis and Robert, 1998; Piironen and Vehtari, 2016, 2017; Smith et al., 2006; Tan et al., 2017; Woo et al., 2015). Assume that we have a collection samples from the implied posterior distribution of the effect size analog. We can summarize the importance of the j-th SNP in our data by taking the KLD between (i) the conditional distribution $p(\widetilde{\boldsymbol{\beta}}_{-j} \,|\, \widetilde{\beta}_j = 0)$ with the effect of that SNP being set to zero and (ii) the marginal distribution $p(\widetilde{\boldsymbol{\beta}}_{-j})$ with the effect of that SNP having been marginalized over. This is defined by solving the following

$$\text{KLD}(j) := \text{KL}\left[p(\widetilde{\boldsymbol{\beta}}_{-j}) \,\|\, p(\widetilde{\boldsymbol{\beta}}_{-j} \,|\, \widetilde{\beta}_j = 0)\right] = \int_{\widetilde{\boldsymbol{\beta}}_{-j}} \log\left(\frac{p(\widetilde{\boldsymbol{\beta}}_{-j})}{p(\widetilde{\boldsymbol{\beta}}_{-j} \,|\, \widetilde{\beta}_j = 0)}\right) p(\widetilde{\boldsymbol{\beta}}_{-j}) \,\mathrm{d}\widetilde{\boldsymbol{\beta}}_{-j}. \tag{5}$$

for each $j = 1, \ldots, J$ variants in the data. We can normalize each of these quantities to obtain a final global association metric

$$\text{RATE}(j) = \text{KLD}(j) / \sum \text{KLD}(l). \tag{6}$$

The above metric is referred to as the "RelATive cEntrality" measure or RATE (Crawford et al., 2019). There are two main takeaways that are important about this metric. First, the KLD(j) value is non-negative, and it equals zero if and only if removing the effect of a given SNP has no impact on explaining the genetic architecture of a trait (i.e., the posterior distribution of $\widetilde{\boldsymbol{\beta}}_{-j}$ is independent of $\widetilde{\beta}_j$). Second, the RATE measure is bounded on the unit interval [0, 1] with the natural interpretation of providing relative evidence of association for each SNP (where values close to 1 suggest greater importance). From a classical hypothesis testing point-of-view, the null under RATE measure assumes that each SNP contributes equally to the phenotypic variance, while the alternative assumes proposes that some SNPs drive the broad-sense heritability of a trait much more than others. Formally, this can stated as

$$H_0 : \text{RATE}(j) = 1/J \quad \text{vs.} \quad H_A : \text{RATE}(j) > 1/J \tag{7}$$

where $1/J$ represents the level that all SNPs in the data have the same relative variable importance.

Limitations of the Current Distributional Centrality Framework

There are several notable issues with effect size analog and RATE framework, particularly for GWA applications. First, calculating both the effect size analog and the KLD in turn for each SNP is computationally expensive even with low-rank matrix approximations (Crawford et al., 2019). Both of these operations involve taking inverses of matrices on the order of J. As the number of variants J grows, these calculations become infeasible. Second, the the significance threshold $1/J \to 0$ as $J \to \infty$, which effectively means that all variants will be considered important for high-dimensional settings. Third, while this framework summarizes the global association for each SNP within the observed population, it lacks the ability to locally explain how important each variant is to each individual in the data. This limits its potential impact, particularly within the context of precision medicine where the goal is to provide individualized patient care. Finally, the least squares projection for the effect size analog in Eq. (4) will only estimate nonlinear effects that are correlated with the linear effects of each SNP (Kowal, 2021; Woody et al., 2021). To see this, define a matrix $\mathbf{Z}$ whose elements are just each column of $\mathbf{X}$ squared. Theoretically, we could define quadratic effects by taking the residuals from the regression of $\boldsymbol{f}$ on $\mathbf{X}$ and regressing them onto $\mathbf{Z}$ in the following way

$$\boldsymbol{\gamma} = (\mathbf{Z}^\intercal \mathbf{Z})^\dagger \mathbf{Z}^\intercal \left(\mathbf{I} - (\mathbf{X}^\intercal \mathbf{X})^\dagger \mathbf{X}^\intercal\right) \boldsymbol{f}.$$

Here, implementing the RATE measure on these new effect sizes $\boldsymbol{\gamma}$ would yield global importance on the quadratic functions of each SNP in the data. However, note that $\boldsymbol{\gamma}$ vanishes if we combine $\widetilde{\boldsymbol{\beta}} + \boldsymbol{\gamma}$ via linear projections onto $\mathbf{X}$. Therefore, if we wanted to study all linear and quadratic effects together, we would instead need to consider a nonlinear projection such as $\widetilde{\boldsymbol{\beta}}^2 + \boldsymbol{\gamma}^2$. The projection operator in Eq. (4) will sometimes miss nonlinear relationships because it only ends up evaluating the part of the nonlinear function $\boldsymbol{f}$ that is linearly associated with each SNP. Each of these issues serve as motivation to develop an alternative and more unified framework for nonlinear models.

Global and Local Score Operators in Nonlinear Models

We now present a simple alternative to achieve interpretability in nonlinear regression models. We will refer to this new summary as the "GlObal And Local Score" (GOALS) operator with the aim to simultaneously identify SNPs that are significantly associated with genetic architecture for a population as well as explain marginal variant effects on an individual level. Again, let $\boldsymbol{f}$ be a nonlinearly estimated function from a weight-space Gaussian process estimated similar to Eq. (3) and consider the scenario where we want to investigate the importance of the j-th feature in explaining what that function has learned from the data. To do so, we can define an N-dimensional vector $\mathbf{g}^{(j)} = \mathbb{E}[\mathbf{y} \mid \mathbf{X} + \boldsymbol{\Xi}^{(j)}]$ where $\boldsymbol{\Xi}^{(j)}$ is an $N \times J$ matrix of all zeros except for the j-th column which we set to be a vector of some positive constant ξ. If we think about the interpretation of a regression coefficient in a linear model as detailing the expected change in the mean response given a ξ-unit increase in the corresponding covariate (holding all else constant), then a natural quantity to understand the importance of each variable is

$$\boldsymbol{\delta}^{(j)} = \boldsymbol{f} - \mathbf{g}^{(j)}. \tag{8}$$

Here, each element of the N-dimensional vector $\boldsymbol{\delta}^{(j)} = (\delta_1^{(j)}, \ldots, \delta_N^{(j)})$ explains the importance of the j-th variable for Gaussian process fit with respect to each sample. The sample average $\bar{\delta}^{(j)} = \sum_i \delta_i^{(j)}/N$ can then be interpreted as a global effect size for the j-th variable within the observed population. Intuitively, in the context of statistical genetics, $\boldsymbol{\delta}^{(j)}$ will be concentrated around zero if the j-th SNP generally has

no effect on the phenotypic variation for a trait. This yields the following natural formulation of a null hypothesis for statistical inference and testing

$$H_0 : \boldsymbol{\delta}^{(j)} = \mathbf{0} \quad \text{vs.} \quad H_A : \boldsymbol{\delta}^{(j)} \neq \mathbf{0} \tag{9}$$

where significantly associated variables have sample means with magnitudes that largely deviate from zero. Since we are assessing a "shift" in function space, each $\boldsymbol{\delta}^{(j)}$ takes into account both additive and nonlinear effects for each variable. Note that the GOALS operator can be flexibly implemented by applying the factor $\boldsymbol{\Xi}^{(j)}$ with any constant and even partitioning the data into subsets for which different values of the constant ξ are used.

Motivated by the popularity of the standard linear model provided in Equation (1) and use of the least-squares like projection in Equation (4), we consider the constant $\xi = 1$. Thus, each $\delta_i^{(j)}$ refers to the expected change in the corresponding phenotype y_i associated with increasing the j-th feature by 1 for individual i. The GOALS operator generalizes the linear model effect size (regression coefficient); when a linear model of the form (1) is assumed, elements of $\boldsymbol{\delta}^{(j)}$ and the average $\bar{\delta}^{(j)}$ will be equal to β_j. This choice of ξ is also well aligned with GWA applications, where SNPs are encoded as $\{0, 1, 2\}$ and increasing a genotype by 1 has a natural interpretation. That said, other values here could also be useful, especially if we assume that genotypes are centered and scaled or want to account for the upper bound of 2 on SNP encodings— we address some of these alternative choices in the Discussion.

Closed-Form Solution and Distributional Properties of GOALS

In this subsection, we describe some of the probabilistic qualities underlying the GOALS operator. Although this approach can be applied to any probabilistic model, we will continue to work under Gaussian process regression as detailed in Eq. (3). To begin, notice that $\boldsymbol{f}$ and each $\mathbf{g}^{(j)}$ are dependent because they are derived from the same set of genotype and phenotype data $\mathbf{X}$ and $\mathbf{y}$, respectively. The joint distribution between the N-dimensional vectors $\mathbf{y}$, $\boldsymbol{f}$, and $\{\mathbf{g}^{(j)}\}_{j=1}^{J}$ can be specified via the following multivariate normal distribution

$$\begin{bmatrix} \mathbf{y} \\ \boldsymbol{f} \\ \mathbf{g}^{(1)} \\ \vdots \\ \mathbf{g}^{(J)} \end{bmatrix} \sim \mathcal{N} \left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{A} & \mathbf{K} & \mathbf{B}^{(1)} & \cdots & \mathbf{B}^{(J)} \\ \mathbf{K} & \mathbf{K} & \mathbf{B}^{(1)} & \cdots & \mathbf{B}^{(J)} \\ \mathbf{B}^{(1)} & \mathbf{B}^{(1)} & \mathbf{C}^{(1)} & \cdots & \mathbf{D}^{(1,J)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{B}^{(J)} & \mathbf{B}^{(J)} & \mathbf{D}^{(J,1)} & \cdots & \mathbf{C}^{(J)} \end{bmatrix} \right) \tag{10}$$

where $\mathbf{A} = \mathbf{K} + \sigma^2\mathbf{I}$ is the marginal variance of the response vector $\mathbf{y}$; $\mathbf{K}$ is the variance of $\boldsymbol{f}$ using the original genotype matrix $\mathbf{X}$ (as in previous notation); $\mathbf{B}^{(j)}$ is the covariance between $\boldsymbol{f}$ and $\mathbf{g}^{(j)}$ using the original matrix $\mathbf{X}$ and the perturbed matrix $\mathbf{X} + \boldsymbol{\Xi}^{(j)}$; $\mathbf{C}^{(j)}$ is the variance of $\mathbf{g}^{(j)}$ using the perturbed matrix $\mathbf{X} + \boldsymbol{\Xi}^{(j)}$; and $\mathbf{D}^{(j,l)}$ is the covariance between $\mathbf{g}^{(j)}$ and $\mathbf{g}^{(l)}$ having perturbed the j-th and l-th feature, respectively. We can simplify the above by utilizing the fact that, when the variance function is the radial basis function, $\mathbf{C}^{(j)} = \mathbf{K}$ for all j. Using this, we can then derive the following joint distribution between $\boldsymbol{f}$ and $\{\mathbf{g}^{(j)}\}_{j=1}^{J}$, conditioned on the data

$$\begin{bmatrix} \boldsymbol{f} \\ \mathbf{g}^{(1)} \\ \vdots \\ \mathbf{g}^{(J)} \end{bmatrix} \Bigg| \, \mathbf{y} \sim \mathcal{N} \left(\begin{bmatrix} \mathbf{K}\mathbf{A}^{-1}\mathbf{y} \\ \mathbf{B}^{(1)}\mathbf{A}^{-1}\mathbf{y} \\ \vdots \\ \mathbf{B}^{(J)}\mathbf{A}^{-1}\mathbf{y} \end{bmatrix}, \begin{bmatrix} \mathbf{K} - \mathbf{K}\mathbf{A}^{-1}\mathbf{K} & \mathbf{B}^{(1)} - \mathbf{K}\mathbf{A}^{-1}\mathbf{B}^{(1)} & \cdots & \mathbf{B}^{(J)} - \mathbf{K}\mathbf{A}^{-1}\mathbf{B}^{(J)} \\ \mathbf{B}^{(1)} - \mathbf{B}^{(1)}\mathbf{A}^{-1}\mathbf{K} & \mathbf{K} - \mathbf{B}^{(1)}\mathbf{A}^{-1}\mathbf{B}^{(1)} & \cdots & \mathbf{D}^{(1,J)} - \mathbf{B}^{(1)}\mathbf{A}^{-1}\mathbf{B}^{(J)} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{B}^{(J)} - \mathbf{B}^{(J)}\mathbf{A}^{-1}\mathbf{K} & \mathbf{D}^{(1,J)} - \mathbf{B}^{(J)}\mathbf{A}^{-1}\mathbf{B}^{(1)} & \cdots & \mathbf{K} - \mathbf{B}^{(J)}\mathbf{A}^{-1}\mathbf{B}^{(J)} \end{bmatrix} \right).$$

Lastly, we can write joint distribution for the GOALS operator $\boldsymbol{\delta}^{(j)} = \boldsymbol{f} - \mathbf{g}^{(j)}$ in Eq. (9) as the following

$$\begin{bmatrix} \boldsymbol{\delta}^{(1)} \\ \vdots \\ \boldsymbol{\delta}^{(J)} \end{bmatrix} \Bigg| \, \mathbf{y} \sim \mathcal{N} \left(\begin{bmatrix} \left(\mathbf{K} - \mathbf{B}^{(1)}\right) \mathbf{A}^{-1}\mathbf{y} \\ \vdots \\ \left(\mathbf{K} - \mathbf{B}^{(J)}\right) \mathbf{A}^{-1}\mathbf{y} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}^{(1)} & \cdots & \boldsymbol{\Sigma}^{(1,J)} \\ \vdots & \ddots & \vdots \\ \boldsymbol{\Sigma}^{(J,1)} & \cdots & \boldsymbol{\Sigma}^{(J)} \end{bmatrix} \right) \tag{11}$$

where

$$\begin{aligned} \boldsymbol{\Sigma}^{(j)} &= \mathbf{K}\mathbf{A}^{-1}\mathbf{K} - \mathbf{B}^{(j)}\mathbf{A}^{-1}\mathbf{B}^{(j)} - \left[\mathbf{B}^{(j)} - \mathbf{B}^{(j)}\mathbf{A}^{-1}\mathbf{K} + \mathbf{B}^{(j)} - \mathbf{B}\mathbf{A}^{-1}\mathbf{B}^{(j)}\right] \\ \boldsymbol{\Sigma}^{(j,l)} &= \mathbf{K} - \mathbf{K}\mathbf{A}^{-1}\mathbf{K} + \mathbf{D}^{(j,l)} - \mathbf{B}^{(j)}\mathbf{A}^{-1}\mathbf{B}^{(l)} - \left[\mathbf{B}^{(j)} - \mathbf{B}^{(j)}\mathbf{A}^{-1}\mathbf{K} + \mathbf{B}^{(l)} - \mathbf{K}\mathbf{A}^{-1}\mathbf{B}^{(l)}\right]. \end{aligned}$$

Theoretically, this results in a joint conditional distribution from which to sample from the posterior distribution of each $\boldsymbol{\delta}^{(j)}$ and obtain local interpretability. As previously mentioned, to investigate the global interpretability of each variant, one can use the sample mean across the local explanations for all observations where $\bar{\delta}^{(j)} = \mathbf{1}^\intercal\boldsymbol{\delta}^{(j)}/N$ with $\mathbf{1}$ being an N-dimensional vector of ones and

$$\begin{bmatrix} \bar{\delta}^{(1)} \\ \vdots \\ \bar{\delta}^{(J)} \end{bmatrix} \Bigg| \, \mathbf{y} \sim \mathcal{N} \left(\begin{bmatrix} \mathbf{1}^\intercal \left(\mathbf{K} - \mathbf{B}^{(1)}\right) \mathbf{A}^{-1}\mathbf{y}/N \\ \vdots \\ \mathbf{1}^\intercal \left(\mathbf{K} - \mathbf{B}^{(J)}\right) \mathbf{A}^{-1}\mathbf{y}/N \end{bmatrix}, \begin{bmatrix} \mathbf{1}^\intercal\boldsymbol{\Sigma}^{(1)}\mathbf{1}/N^2 & \cdots & \mathbf{1}^\intercal\boldsymbol{\Sigma}^{(1,J)}\mathbf{1}/N^2 \\ \vdots & \ddots & \vdots \\ \mathbf{1}^\intercal\boldsymbol{\Sigma}^{(J,1)}\mathbf{1}/N^2 & \cdots & \mathbf{1}^\intercal\boldsymbol{\Sigma}^{(J)}\mathbf{1}/N^2 \end{bmatrix} \right). \tag{12}$$

Therefore, to simulate from the posterior distribution of the sample means, one simply needs to compute the following closed form equations for the first and second moments

$$\begin{aligned} \mathbb{E}\left[\bar{\delta}^{(j)}\right] &= \mathbf{1}^\intercal \left(\mathbf{K} - \mathbf{B}^{(j)}\right) \mathbf{A}^{-1}\mathbf{y}/N \\ \mathbb{V}\left[\bar{\delta}^{(j)}\right] &= \left(\lambda + \alpha_{jj} - 2\psi_j\right)/N^2 \\ \mathbb{V}\left[\bar{\delta}^{(j)}, \bar{\delta}^{(l)}\right] &= \left(\lambda + \alpha_{jl} - \psi_j - \psi_l\right)/N^2 \end{aligned} \tag{13}$$

where $\lambda = \mathbf{1}^\intercal\mathbf{K}\mathbf{1} - \mathbf{1}^\intercal\mathbf{K}\mathbf{A}^{-1}\mathbf{K}\mathbf{1}$; $\alpha_{jl} = \mathbf{1}^\intercal\mathbf{D}^{(j,l)}\mathbf{1} - \mathbf{1}^\intercal\mathbf{B}^{(j)}\mathbf{A}^{-1}\mathbf{B}^{(l)}\mathbf{1}$; and $\psi_j = \mathbf{1}^\intercal\mathbf{B}^{(j)}\mathbf{1} - \mathbf{1}^\intercal\mathbf{K}\mathbf{A}^{-1}\mathbf{B}^{(j)}\mathbf{1}$, respectively.

In current GWA applications, where datasets can include hundreds of thousands of individuals genotyped at millions of markers, it is often desirable to use a more straightforward and scalable computation than sampling estimates of the GOALS measure from the full joint distribution in Eq. (11). To that end, in this work, we will consider the posterior mean in Eq. (11) as estimates of local importance and then take the sample means of these values to get a measurement of global importance. More specifically, these two respective values are taken as the following

$$\widehat{\boldsymbol{\delta}}^{(j)} = \left(\mathbf{K} - \mathbf{B}^{(j)}\right) \mathbf{A}^{-1}\mathbf{y}, \qquad \bar{\delta}^{(j)} = \sum_i \widehat{\delta}_i^{(j)}/N. \tag{14}$$

Scalable Computation

In practice, we can make use of a few additional matrix algebra properties to efficiently compute estimates from the otherwise computationally intensive distribution outlined in Eq. (11). Here, we will continue to assume that the Gaussian process regression uses a radial basis covariance function. First, it is important to note that the only matrix that needs to be recomputed for each SNP j is the matrix $\mathbf{B}^{(j)}$ which measures the covariance between the original $\mathbf{X}$ and the perturbed $\mathbf{X} + \boldsymbol{\Xi}^{(j)}$. When using the radial basis function, this matrix can be derived for the j-th SNP by making the following rank one updates

$$b_{ii'}^{(j)} = k\left(\mathbf{x}_i, \mathbf{x}_{i'} + \boldsymbol{\xi}^{(j)}\right) = \exp\left\{-\theta \left\|\mathbf{x}_i - \left(\mathbf{x}_{i'} + \boldsymbol{\xi}^{(j)}\right)\right\|^2\right\}$$

$$
\begin{aligned}
&= \exp\left\{-\theta\left[\|\mathbf{x}_i - \mathbf{x}_{i'}\|^2 - 2\left(\mathbf{x}_i - \mathbf{x}_{i'}\right)^\intercal \boldsymbol{\xi}^{(j)} + \left\|\boldsymbol{\xi}^{(j)}\right\|^2\right]\right\} \\
&= \exp\left\{-\theta\|\mathbf{x}_i - \mathbf{x}_{i'}\|^2\right\}\exp\{-\theta\left[\xi^2 - 2\xi\left(\mathbf{x}_i - \mathbf{x}_{i'}\right)^\intercal\right]\} \\
&= k(\mathbf{x}_i, \mathbf{x}_{i'})\exp\{-\theta\left[\xi^2 - 2\xi\left(\mathbf{x}_i - \mathbf{x}_{i'}\right)^\intercal\right]\}
\end{aligned}
$$

where, similar to previous notation, $\mathbf{x}_i$ and $\mathbf{x}_{i'}$ are the i-th and i'-th rows of the genotype matrix $\mathbf{X}$, and $\boldsymbol{\xi}^{(j)}$ is a row of the matrix $\boldsymbol{\Xi}^{(j)}$ where the j-th element is set to some positive constant ξ. We can restate the above in matrix notation as

$$
\mathbf{B}^{(j)} = \mathbf{K} \circ \exp\left\{-\theta\left[\xi^2 - 2\xi\left(\mathbf{x}_{\bullet j}\mathbf{1}^\intercal - \mathbf{1}\mathbf{x}_{\bullet j}^\intercal\right)\right]\right\} \tag{15}
$$

where $\mathbf{x}_{\bullet j}$ is the j-th column in the matrix $\mathbf{X}$ and $\circ$ denotes element-wise multiplication. The main summary is that the computation of each $\mathbf{B}^{(j)}$ only relies on linear operations after the initial computation of the radial basis covariance matrix $\mathbf{K}$. These steps extend to other shift invariant covariance functions (e.g., Laplacian and Cauchy) and a similar rank one update procedure can also be shown for the linear gram matrix (see Supplementary Material).

Theoretical Connection to Shapley Additive Explanations

The GOALS operator measures local importance by quantifying the change in function space that occurs when the j-th feature of interest is shifted by some nonzero factor. There is a theoretical connection between this strategy and Shapley Additive Explanations (SHAP) (Lundberg and Lee, 2017) which is a widely used *post hoc* local interpretability metric in the machine learning literature (e.g., Chen et al., 2022). Briefly, Shapley values assign feature importance weights based on game theoretic principles (Roth, 1988; Shapley, 1951) by essentially determining a payoff for all players when each player might have contributed more or less than the others when attempting to achieve the desired outcome. In genetics applications, this is done by considering all possible subsets of genotypes that do not include the j-th genotype $\mathcal{S} \subseteq \mathcal{J}\backslash\{j\}$ and then comparing their performance to the performance of a model trained on the same subset as well as the j-th genotype $\mathcal{S} \cup \{j\}$. This weighted average can be represented as the following formula

$$
\phi_j = \sum_{\mathcal{S}\subseteq\mathcal{J}\backslash\{j\}}\left[\frac{|\mathcal{S}|!(J - |\mathcal{S}| - 1)!}{J!}\right]\left(\boldsymbol{f}_{\mathcal{S}\cup\{j\}} - \boldsymbol{f}_{\mathcal{S}}\right) \tag{16}
$$

where $|\mathcal{S}|$ is number of genotypes in subset $\mathcal{S}$ and $|\mathcal{J}| = J$ is the total number of genotypes in the data. Keeping our notation consistent with previous sections, we say that $\boldsymbol{f}_{\mathcal{S}\cup\{j\}}$ and $\boldsymbol{f}_{\mathcal{S}}$ are the GP regression model fits with and without the j-th genotype added to the genotype subset $\mathcal{S}$, respectively.

Rather than removing a given genotype from each subset and calculating model differences, GOALS perturbs each genotype and calculates the corresponding difference in model fit. However, we can relate SHAP to goals by considering the special case of a single individual $N = 1$. In this case, $g^{(j)} = \mathbb{E}[y\,|\,\mathbf{x} + \boldsymbol{\xi}^{(j)}]$ where $\boldsymbol{\xi}^{(j)}$ is an $1 \times J$ vector of all zeros except for the j-th element which we set to be a vector of some positive constant ξ. Note that we can represent the "shifting" vector $\boldsymbol{\xi}^{(j)}$ as the following

$$
\boldsymbol{\xi}^{(j)} = \xi\left[\mathbb{1}\{j = 1\} \quad \cdots \quad \mathbb{1}\{j = J\}\right] \tag{17}
$$

where $\mathbb{1}\{\bullet\}$ denotes an indicator function which is returns one for the j-th column and 0 otherwise. From this view, we can say that $\mathcal{J}'$ is the set of J indicator random variables which make up elements of $\boldsymbol{\xi}^{(j)}$. We can also therefore rewrite the GOALS operator as

$$
\delta^{(j)} = f - g^{(j)} = f_{\mathcal{J}} - f_{\mathcal{J}\cup\mathcal{J}'}. \tag{18}
$$

If we set $\xi = -x_j$, the GOALS operator behaves similarly to a SHAP value, as $\mathbf{g}^{(j)}$ represents the model fit where the j-th covariate is set to zero. In this case, GOALS could be seen as an approximation to SHAP where GOALS only considers the single subset of $J-1$ genotypes, excluding the j-th genotype, whereas SHAP which considers all every possible subsets of features that do not include the j-th genotype.

Lastly, it is worth noting that there are scenarios where we would expect GOALS and SHAP to provide different local interpretability rankings for the j-th covariate. The factorial in the SHAP weight computation in Eq. (16) favors both the smallest and largest subsets of $\mathcal{J}$ and penalizes subsets $\mathcal{S}$ of the size $|\mathcal{S}| \approx J/2$. In the GWA setting, this means that if the j-th SNP has an effect on the trait of interest via marginal effects, then both GOALS and SHAP are likely to give that SNP a high ranking. If the j-th SNP is only influential on a phenotype through a moderate number of interactions (i.e., within sets of size $J/2$), then GOALS may rank that SNP higher relative to other SNPs than SHAP will. However, on the other hand, if the j-th SNP is influential on trait architecture through pairwise interactions with nearly all other variants genome-wide, then SHAP may provide a higher relative rank for that SNP than GOALS. Furthermore, SHAP may rank SNPS that are highly correlated with each other lower than GOALS, because the difference in model fits $\boldsymbol{f}_{\mathcal{S}\cup\{j\}} - \boldsymbol{f}_{\mathcal{S}}$ may be small when genotype j is highly correlated with genotypes in $\mathcal{S}$. We show that these expectations are supported empirically in the next section.

Results

We now illustrate the benefits of our simple approach for global and local interpretability in extensive simulations and real data analyses. First, we conduct a proof-of-concept simulation study to help the reader build a stronger intuition for how GOALS prioritizes influential variables on both a local and global scale, simultaneously. To provide concrete points of reference, we will also show how the Shapley Additive Explanations (SHAP) (Lundberg and Lee, 2017) approach assigns feature importance weights locally and we will demonstrate how the distributional centrality framework using the effect size analog with RATE performs global interpretability (Crawford et al., 2018, 2019). We also show that GOALS is much more scalable than both methods as both the number of observations and genetic markers increase. For the second analysis in this section, we implement a more realistic simulation scheme to assess how GOALS performs association mapping compared to various *post hoc* variable importance, Bayesian shrinkage, and regularization modeling techniques. Lastly, we apply the GOALS operator to six quantitative traits assayed in a heterogenous stock of mice from Wellcome Trust Centre for Human Genetics (Valdar et al., 2006a,b).

Simulation Studies

The general design of the following simulation studies is commonly used to explore the power of statistical methods for association mapping across different genetic architectures underlying complex traits. Here, we assume that we have some $N \times J$ genotype matrix $\mathbf{X}$ which is used to generate N-dimensional real-valued phenotypes $\mathbf{y}$ based on a combination of linear (additive), interaction (epistatic), and cryptic population stratification effects. We will assume that each synthetic trait has been standardized such that $\mathbb{V}[\mathbf{y}] = 1$. We will also assume that all observed genetic effects explain a fixed proportion of this variance (commonly referred to as the broad-sense heritability of the trait H^2). To explicitly generate data, we select a subset of causal SNPs $\mathcal{J}$ from the genotypic matrix and then use the following linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\boldsymbol{\vartheta} + \mathbf{Z}\boldsymbol{\omega} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \tau^2\mathbf{I}). \tag{19}$$

where $\mathbf{W}$ holds all pairwise interactions between the causal SNPs (i.e., the Hadamard or element-wise product between genotypic vectors of SNPs within $\mathcal{J}$) and $\mathbf{Z}$ contains covariates representing additional population structure in the data. In this study, we will consider $\mathbf{Z}$ to be the top ten principal components

(PCs) from the genotype matrix. The additive and interaction effect sizes for the causal variants are assumed to follow normal distributions; while the effects for non-causal variable set to be zero. Using a parameter ρ, we scale both the additive and pairwise genetic effects so that collectively they explain a fixed proportion of the genetic variance. Namely, we assume that the additive effects detail $\mathbb{V}[\mathbf{X}\boldsymbol{\beta}] = \rho H^2$ (i.e., narrow-sense heritability of a trait h^2) and that the interaction effects make up the remaining $\mathbb{V}[\mathbf{W}\boldsymbol{\vartheta}] = (1-\rho)H^2$. In the simulations shown in this section, we will consider $\rho \in \{0.5, 1\}$, where the former assumes that additive and interaction effects contribute equally to the total phenotypic variance, and the latter assumes only additive effects contribute. Once we obtain the final effect sizes for all causal variants, we draw normally distributed random variables for the population structure coefficients and the random error term. In this scenario, one can think of the combined effect of $\mathbf{Z}\boldsymbol{\omega}$ as structured noise. To this end, we ensure that $\mathbb{V}[\mathbf{Z}\boldsymbol{\omega}] + \mathbb{V}[\boldsymbol{\varepsilon}] = (1 - H^2)$. In all simulations, genotype PCs are not included in any of the assessed model fitting procedures, and no other preprocessing normalizations were carried out to account for the added population structure.

Proof-of-Concept Simulations: Low-Dimensional Analysis. In this subsection, we provide a low-dimensional proof-of-concept simulation study. To accomplish this, we generate a synthetic genotype matrix $\mathbf{X}$ with $N = 2000$ individuals and $J = 25$ SNPs with allele frequencies randomly drawn from a uniform distribution such that we only consider common variants with a minor allele cutoff above 5%. In these simulations, we generate synthetic traits using Eq. (19) by fixing the broad-sense heritability $H^2 = 0.6$ and omitting population structure effects by setting $\boldsymbol{\omega} = \mathbf{0}$. Here, we assume some subset of the variants $\mathcal{J} = \{8, 9, 10, 23, 24, 25\}$ to be causal. We then consider five different simulation scenarios:

- **Scenario I (Additive and Interaction Effects):** The subset $\{23, 24, 25\} \subseteq \mathcal{J}$ are causal SNPs, where all three have additive effects and variants #23 and #24 interact with #25, respectively.
- **Scenario II (Additive and Interactions Effects from Different SNP-sets):** All SNPs in the set $\mathcal{J}$ are causal. SNPs #8-10 only have interaction effects and SNPs #23-25 only have additive effects. Specifically, variants #8 and #9 each interact with #10, separately.
- **Scenario III (Overlapping Additive and Interaction Effects):** All SNPs in the set $\mathcal{J}$ are causal. SNPs #23-25 each have additive effects; while, variant #8 interacts with #10 and #9 interacts with #25, respectively.
- **Scenario IV (Interaction Effects Only):** All SNPs in the set $\mathcal{J}$ are causal only through interaction effects. SNPs #8 and #9 each interact with #10, separately; while, SNPs #23 and #24 each interact with #25, separately.
- **Scenario V (Noise Only):** None of the SNPs in the data have an association with the trait. Represents the case when assumptions of the null model are met.

We want to point out that, while this is indeed a small proof-of-concept study, each of these cases highlight settings that we might experience with real data. For each scenario, we fit a standard GP regression model similar to Eq. (3) under a zero mean prior and a radial basis covariance function.

Figure 1 contains the global variable importance results for GOALS and RATE on Scenarios I-VI for 100 simulated replicates. Here, we perform RATE on a GP model using effect size analogs computed with the linear projection as in Eqs. (4)-(6), while the GOALS operator is calculated on the GP model as in Eq. (14). In Figure 1, the known causal SNPs for each scenario are colored in blue. To compare the null hypotheses for the two approaches, we also display red dashed lines that are drawn at the level of relative equivalence (i.e., $1/J$) for RATE and at zero for GOALS, respectively. Overall, we see that both methods perform similarly in identifying causal SNPs that have both additive and interaction effects on the trait of interest (Figure 1A). However, GOALS proves to be a better discriminator between causal and non-causal SNPs than RATE when interaction effects occur in isolation (i.e., SNPs are involved in an

interactions without necessarily having an additive effect). Importantly, GOALS exhibits a more robust control of the false negative rate in exchange for a slight increase in false discovery for these scenarios (see how the RATE and GOALS operators relate to the null threshold lines in Figures 1B-D). This result highlights the potential limitation of the linear projection that RATE uses to compute the effect size analog and demonstrates its potential to miss associations that stem from nonlinear interactions (especially when SNPs only have non-additive effects such as markers #8 and #9). Lastly, GOALS is better calibrated when traits are generated from complete noise (i.e., when there are no true associations between genotype and phenotype) (Figure 1E). This is due to the fact that the GOALS operator assesses the global importance variables based on their individual contribution to the model fit. While the concept of relative centrality is intuitive, achieving a completely uniform distribution of RATE values at $1/J$ under the null model will rarely happen in practice (especially in applications where spurious associations between correlated input variables and the modeled response can occur). In other words, due to the stochastic nature of data, one variable will always appear relatively more important than another which can lead to ill-informed analyses during downstream tasks under the RATE framework.

Another major contribution of GOALS is that it also provides local explanations of how variables affect model fit for each individual in the data. For example, in GWA applications, this can yield key insight in the event that a genetic variant is biomarker for only a specific subset a population. To demonstrate the utility of GOALS in this case, we consider a sixth simulation scenario where

- **Scenario VI (Population Specific Effects):** The subset $\{22, 23, 24, 25\} \subseteq \mathcal{J}$ are causal SNPs. SNPs #23-25 have additive and interaction effects that are associated with all individuals; while, SNP #22 has an additive effect for only half of the population.

Figure 2 shows the distribution of the local individual-level GOALS operator for SNPs #8, #22, and #25 in this split scenario. As a baseline, we also show results from running a local analysis with SHAP. For clarity, SNP#8 is a non-causal SNP in this scenario. There are a few key takeaways in this empirical illustration. First, when a SNP has a no effect on the trait of interest, the distribution of the local scores for both GOALS and SHAP are centered at 0. Conversely, SNPs with nonzero effects on the phenotype have GOALS and SHAP operators with magnitudes that are centered distinctly away from the origin. One difference here is that the GOALS values tend to have the same sign, while the SHAP metric can be positive or negative. In the case where a SNP has an effect on trait architecture for only a subset of the observed population, the local distribution of the SHAP and GOALS operators will be multimodal allowing for individualized summaries of marker effects on specific observations. In Figure 2, this characteristic is more distinct with the GOALS operator where there is clearer separation in values for SNP #22 in individuals where it has a nonzero effect.

Method Comparisons: High-Dimensional Global Variable Importance. We now assess the power of GOALS and its ability to effectively prioritize causal variables in high-dimensional data settings. In this analysis, we use data from the Wellcome Trust Case Control Consortium (WTCCC) 1 study which initially consisted of 2,938 shared controls with 458,868 SNPs after following the quality control procedures of previous studies (The Wellcome Trust Case Control Consortium, 2007). Missing genotypes were imputed by using the BIMBAM software (`http://www.haplotype.org/bimbam.html`; Servin and Stephens, 2007). In these simulations, we use all polymorphic SNPs with minor allele frequencies (MAFs) above 1% on chromosome 22 to generate continuous phenotypes. Exclusively considering this group of individuals and SNPs resulted in a final dataset consisting of $N = 2{,}938$ samples and $J = 5{,}747$ markers.

During each simulation run, we set the broad-sense heritability to be $H^2 = 0.3$ and consider two choices for the parameter $\rho \in \{0.5, 1\}$. Here, we randomly choose a set of 30 causal SNPs which we break up into two groups. The first group contains a small set of 5 variants, while the second group contains a larger set of 25 variants. All causal markers have additive effects and, when applicable (i.e., where $\rho = 0.5$), there are intergroup pairwise interaction effects (e.g., SNPs in the first group interact with

SNPs in the second group, but never with each other). We also consider simulations with and without population stratification effects by allowing the top ten genotype principal components (PCs) to make up to 10% of the overall variation in the synthetic traits. In total, this resulted in four scenarios based on different parameter combinations: (*i*) $\rho = 1$ and $\mathbb{V}[\mathbf{Z}\boldsymbol{\omega}] = 0$; (*ii*) $\rho = 1$ and $\mathbb{V}[\mathbf{Z}\boldsymbol{\omega}] = 0.1$; (*iii*) $\rho = 0.5$ and $\mathbb{V}[\mathbf{Z}\boldsymbol{\omega}] = 0$; and (*iv*) $\rho = 0.5$ and $\mathbb{V}[\mathbf{Z}\boldsymbol{\omega}] = 0.1$. In other words, scenarios I and II consider traits with additive effects only; while, scenarios III and IV consider traits with both additive and interaction effects. Additionally, scenarios II and IV have the additional complexity of having nonzero population stratification effects contribute to the phenotypic variation which is not observed in scenarios I and III.

We compare the global power of the GOALS measure to a list of association mapping modeling techniques. Specifically, these methods include: (a) the *post hoc* framework of estimating effect size analogs for the input features of a GP regression model and determining their importance using distributional centrality via RATE (Crawford et al., 2019); (b) a genome-wide scan with a single-SNP univariate linear model that is typically used in GWA applications (SCANONE) (Yandell et al., 2007); (c) L1-regularized "least absolute shrinkage and selection operator" (LASSO) regression (Tibshirani, 1996); and (d) the combined regularization utilized by the Elastic Net (Zou and Hastie, 2005). Note that SCANONE produces P-values, and the LASSO and the Elastic Net give magnitudes of regression coefficients. The regularization approaches were fit by first learning tuning parameter values via 10-fold cross validation. Indeed, the SHAP value framework can also be used for *post hoc* assessment of global interpretability by taking the average of local scores across observations for each feature in the data. However, because the SHAP approach considers all possible subsets of genotypes when determining variable importance, it does not scale well to high-dimensional settings. For this reason, we do not consider it for comparison in this simulation study.

Each method is evaluated based on its ability to effectively prioritize the causal SNPs in 100 different simulated datasets. The criteria we use compares the false positive rate (FPR) with the rate at which true variants are identified by each model (TPR). A depiction of these results can be found in Figure 3, where the upper limit of the FPR on the x-axis has been truncated at 0.2. This is further quantified by assessing the entire area under the curve (AUC) in the legend for further comparison. Overall method performance varies depending on the two factors: (a) the presence of interaction effects, and (b) additional structure due to population stratification. For example, most methods perform best in the first simulation scenario where the broad-sense heritability of the synthetic traits is made up of only additive effects (e.g., Figure 3A). This power generally decreases in the presence of structured noise (e.g., Figure 3B) or when pairwise interactions between causal SNPs have nonzero effects on the phenotype (e.g., Figure 3C and 3D). GOALS outperforms LASSO and Elastic Net consistently in every scenario and performs competitively with SCANONE and RATE in every scenario. More specifically, GOALS is a top performer in scenarios with additional population stratification effects at lower false positive rates. In contrast, both RATE and SCANONE lose power when structured noise is added to the simulations. While RATE performs generally well in each of these scenarios, the algorithm often takes much longer than GOALS to run as the number features increases. For a dataset with $J = 500$ features and $N = 1000$ samples, RATE has an average runtime of 60 seconds on computing cluster with 30 nodes whereas GOALS takes only a second to complete. We argue that these simulations highlight GOALS as a reliable option for interpretability given its consistent performance across a wide range of scenarios and its scalability as data sizes increase. GOALS also has the additional benefit of allowing for local variable importance analyses which is something that RATE and SCANONE do not provide.

Global and Local Association Mapping in Heterogenous Stock of Mice

In this section, we apply GOALS to individual-level genetic data from a heterogenous stock of mice collected by the Wellcome Trust Centre of Human Genetics (`http://mtweb.cs.ucl.ac.uk/mus/www/mouse/index.shtml`) (Valdar et al., 2006a,b). The genotypes from this study were downloaded directly using the `BGLR-R` package (Perez and de los Campos, 2014). This study contains $N = 1{,}814$ heterogenous

stock of mice from 85 families (all descending from eight inbred progenitor strains) and 131 quantitative traits that are classified into 6 broad categories including behavior, diabetes, asthma, immunology, haematology, and biochemistry. Phenotypic measurements for these mice can be found freely available online to download (details can be found at `http://mtweb.cs.ucl.ac.uk/mus/www/mouse/HS/index.shtml`). In this study, we focus on three of these complex traits: body weight, percentage of CD8+ cells, and high-density lipoprotein (HDL) content. Each of these phenotypes were previously corrected for sex, age, body weight, season, and year (Valdar et al., 2006a,b). For individuals with missing genotypes, we imputed values by the mean genotype of that SNP in their corresponding family. Only polymorphic SNPs with minor allele frequency above 5% were kept for the analyses. This left a total of $J = 10{,}227$ autosomal SNPs that were available for all mice.

We chose to analyze this particular dataset for a few reasons. The first reason is that RATE has been previously applied to these same three traits to perform nonlinear *post hoc* variable importance (Crawford et al., 2019) — thus, it provides a methodological baseline for the performance of GOALS on the global level. The second reason is that common environmental effects caused by the mice sharing the same cage have been shown to have nonzero contribution to the overall variance observed in these traits (Crawford et al., 2018). Therefore, it means that one might expect to observe varying local SNP effects between mice assigned to different cages. Lastly, the mice in this study are known to be genetically related and the measured have varying levels of broad-sense heritability with nonzero contributions from both additive and non-additive genetic effects (Chen et al., 2012; Valdar et al., 2006b). As result, this dataset represents a realistic mixture of the simulation scenarios we detailed in the previous sections.

For each trait, we fit a GP regression model. Figures 4, S1, and S2 display the variant-level mapping results after assessing *post hoc* variable importance using GOALS and RATE in HDL content, body weight, and the percentage of CD8+ cells, respectively. In these plots, notable SNPs are annotated and color coded according to their nearest mapped gene(s) as cited by the Mouse Genome Informatics database (`http://www.informatics.jax.org/`) (Bult et al., 2019). We also provide summary tables which lists the corresponding GOALS and RATE values for all SNPs (see Tables S1-S3 in the Supplementary Material). In general, GOALS identified genetic signal across more unique genes and chromosomal regions than RATE in all three traits that we studied. This was most apparent in HDL where the top 10 highest ranked SNPs by RATE were all located within two genes on chromosome 1; but, the top 10 highest ranked SNPs by GOALS included 13 relevant genes across five different chromosomes (see Figure 4). In addition to moderate signal on chromosome 1, GOALS also found signal on chromosomes 3, 11, 12,15, and 17. Importantly, many of the candidate SNPs selected by GOALS (and their respective genes) have been previously discovered by past publications as having some functional relationship with HDL content. For example, *Tcq12*, *Hyplip2*, and *Ddiab41* have all been shown to associated with fat, cholesterol, and metabolism (Bult et al., 2019; Gu et al., 1999; Lawson et al., 2011; Moen et al., 2007; Östergren et al., 2015; Valdar et al., 2006b).

There was notable overlap in the findings for RATE and GOALS in the analysis of body weight and the percentage of CD8+ cells. Both methods identified strong signal on the X chromosome for body weight, a genomic region that was also validated by Valdar et al. (2006b) in the original study (see Figure S1). Here, GOALS and RATE detected several adiposity-related genes, including *Obq6* (Chen et al., 2012; Taylor et al., 1999) and *Dbts2* (Cheverud et al., 2004). For the percentage of CD8+ cells, both methods identified many genes on chromosome 17 which are known to greatly determine the ratio of T-cells (Yalcin et al., 2010), and some have been suggested to modulate cell adhesion and motility in the immune system (Kim et al., 2006). Overall, out of the top ten most prioritized variables ranked by GOALS and RATE, there was a 30% overlap for body weight and a 90% overlap for the percentage of CD8+ cells. For the latter, only GOALS prioritized a SNP on Chromosome 4 which harbors genes *Lpq1* involved in pairwise interactions that are associated with lymphocyte percentage (Miller et al., 2020).

Once again, the additional benefit of GOALS is its ability to also perform local variable importance for individual samples. In this particular dataset from the Wellcome Trust Centre of Human Genetics,

shared common environments between mice have been shown to contribute to the phenotypic variation of complex traits (Crawford et al., 2018; Valdar et al., 2006a,b). For example, dietary and immunological phenotypes could depend heavily on the distribution of food and water in each cage. To that end, we assessed the local GOALS metrics for notable SNPs across mice according to the cages in which they were assigned during the study. In Figure 5, we take the two SNPs with the greatest global GOALS value in each trait and plot the local values for the 4 cages with the greatest and least local means. Overall, our proposed measure does indeed seem to capture environmental variation. This is again most apparent in HDL where the top SNPs rs13459070 (chromosome 3) and rs3721166 (chromosome 15) have very different effects on mice in different cages (e.g., Figure 5B and 5E). Altogether, these sets of results would allow practitioners to perform deeper and more nuanced downstream analyses of phenotypic behavior in different populations.

Discussion

In this paper, we proposed the "GlObal And Local Score" (GOALS) operator: a general approach for regression models that assesses variable importance for features at both the local and global levels of data, simultaneously. While this novel *post hoc* interpretability measure can be used for any type of statistical model, we frame GOALS within the context of genome-wide association (GWA) studies where there is an goal to understand how both additive and non-additive sources of genetic variation contribute to disease progression and complex trait architecture. In the main text, we described the probabilistic properties of GOALS assuming we have fit a Gaussian process regression model with a shift-invariant covariance function. We also discussed its similarities to distributional centrality measures (Crawford et al., 2019; Paananen et al., 2019, 2021; Piironen and Vehtari, 2016, 2017; Woo et al., 2015) and Shapley Additive Explanations (SHAP) (Chen et al., 2022; Lundberg and Lee, 2017) which are commonly used approaches for determining global and local feature importance in nonlinear statistical methods, respectively. Through extensive simulations, we showed that our new measure can be used for feature selection and gives comparable state-of-the-art performance even in the presence of population structure (Figures 1-3). The added benefit of GOALS is its ability to also understand how features affect individual samples on the local level and its computational efficiency to reach conclusions with much improved runtime as the dimensions of data increase. In applications to a real GWA dataset from the Wellcome Trust Centre of Human Genetics (Valdar et al., 2006a,b), we showed that GOALS has the ability to identify a greater number of trait-relevant genomic loci in a heterogenous stock of mice that have also been detected in many previous publications (Figures 3 and S1-S2 and Tables S1-S3). The first key part of this analysis was that the ability of GOALS to be incorporate non-additive information allowed it to find genetic signal that were missed by other *post hoc* approaches based on linear approximations. The second main takeaway from the real data GWA analysis is that the GOALS providing local interpretability enables downstream analyses to investigate how and why specific biomarkers are enriched for specific subsets of a population (Figure 5). In this study, we saw how SNPs associated with high density lipoprotein (HDL) content had varying local effects among mice assigned to different cages — potentially, as a result of differing environments such as access to food and water. Ultimately, we hope that GOALS will encourage the continued development of probabilistic machine learning methods that can analyze complex data at the local and global levels.

The current implementation of the GOALS framework offers many directions for future development. First, while GOALS is conceptually applicable to scientific domains outside of statistical genetics, further work is needed to fully assess the best way implement the approach elsewhere. Currently, when assessing variable importance, we fix the perturbation parameter for each feature to be $\xi = 1$. This is motivated by our application to GWA studies where the variables measured are often single nucleotide polymorphisms (SNPs) encoded as $\{0, 1, 2\}$ copies of a reference allele at each locus. Part of our future work will be to explore how to choose ξ or allow for the j-th column of $\boldsymbol{\Xi}^{(j)}$ to take on multiple values such that

GOALS is more widely generalizable to other applications. Additionally, for demonstrative purposes, we allowed our chosen value of ξ to take values of $\tilde{\mathbf{X}}^{(j)}$ outside of the SNP domain space. While we believe this did not impact the overall results, more exploration is needed for values of ξ to respect the domain space and increase interpretability. Secondly, while GOALS provides a measure of general association for nonlinear methods, it cannot be used to directly identify the component (i.e., linear vs. nonlinear) that drives individual variable importance. Thus, despite being able to detect variants that are associated to a response in a nonlinear fashion, GOALS is unable to directly identify the detailed orders of interaction effects. This same limitation also exists with distributional centrality measure such as RATE. A key part of our future work is to continue learning how to disentangle this information (e.g., very similar to the goals of Kowal, 2021; Woody et al., 2021). We also note that we did not explicitly conduct experiments to assess the effect of collinearity on variable selection performance (Gelman and Hill, 2006; Wold et al., 1984), but we did conduct both simulations and data analyses using real genotypic data with various levels of linkage disequilibrium (LD) (i.e., correlation structure) between variants. Therefore, we did implicitly take collinearity into account. In our framework, as long as the nonlinear regression model is capturing the appropriate relationships between the data and the response, GOALS will be able to rank features effectively.

As a third extension, GOALS does not enforce any sparsity or shrinkage when performing variable importance. Thus, while it has a natural null hypothesis (i.e., Eq. (9)), it does not naturally produce a significance threshold for variable selection. Common examples in biomedicine include a Bonferroni-corrected threshold (Gordon et al., 2007) or selection based on a median probability model (Barbieri and Berger, 2004). A natural solution could be to permute the phenotypic labels and refit the model a number of times to choose a GOALS-specific family-wise error rate (FWER) (e.g., Hoti and Sillanpää, 2006; Stephens and Balding, 2009); however, this can be computationally intensive. One alternative could be to sample a collection of $\boldsymbol{\delta}^{(j)}$ from the posterior distribution as specified in Eq. (11) and select significant variables based on a metric like a local false sign rate (Stephens, 2016). Lastly, univariate variable importance methods have been shown to be underpowered in settings where there are many causal variables with small effects. In many GWA applications, recent methods have utilized prior knowledge to test groups of SNPs at a time for enrichment with a particular trait (Carbonetto and Stephens, 2013; Cheng et al., 2020; de Leeuw et al., 2015; Demetci et al., 2021; Ish-Horowicz et al., 2019; Lamparter et al., 2016; Liu et al., 2010; Nakka et al., 2016; Sun et al., 2019; Wu et al., 2010; Zhu and Stephens, 2018). This same group hypothesis extension can also be extended to the GOALS framework by simply perturbing multiple variables at a time.

Software Details

Code for implementing the "GlObal And Local Score" (GOALS) operator is freely available at `https://github.com/lcrawlab/GOALS`, and is written in a combination of `R` and `C++` commands.

Acknowledgements

This research was conducted using computational resources and services at the Center for Computation and Visualization (CCV), Brown University. E.T. Winn-Nuñez was supported by the National Science Foundation Graduate Research Program under Grant No. 1644760. This research was supported by a David & Lucile Packard Fellowship for Science and Engineering awarded to L. Crawford. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of any of the funders.

Author Contributions

All authors conceived the study and developed the methods. MG and LC supervised the project and provided resources. ETWN and LC developed the software. ETWN performed the analyses. All authors wrote and revised the manuscript.

Competing Interests

The authors declare no competing interests.

Figures

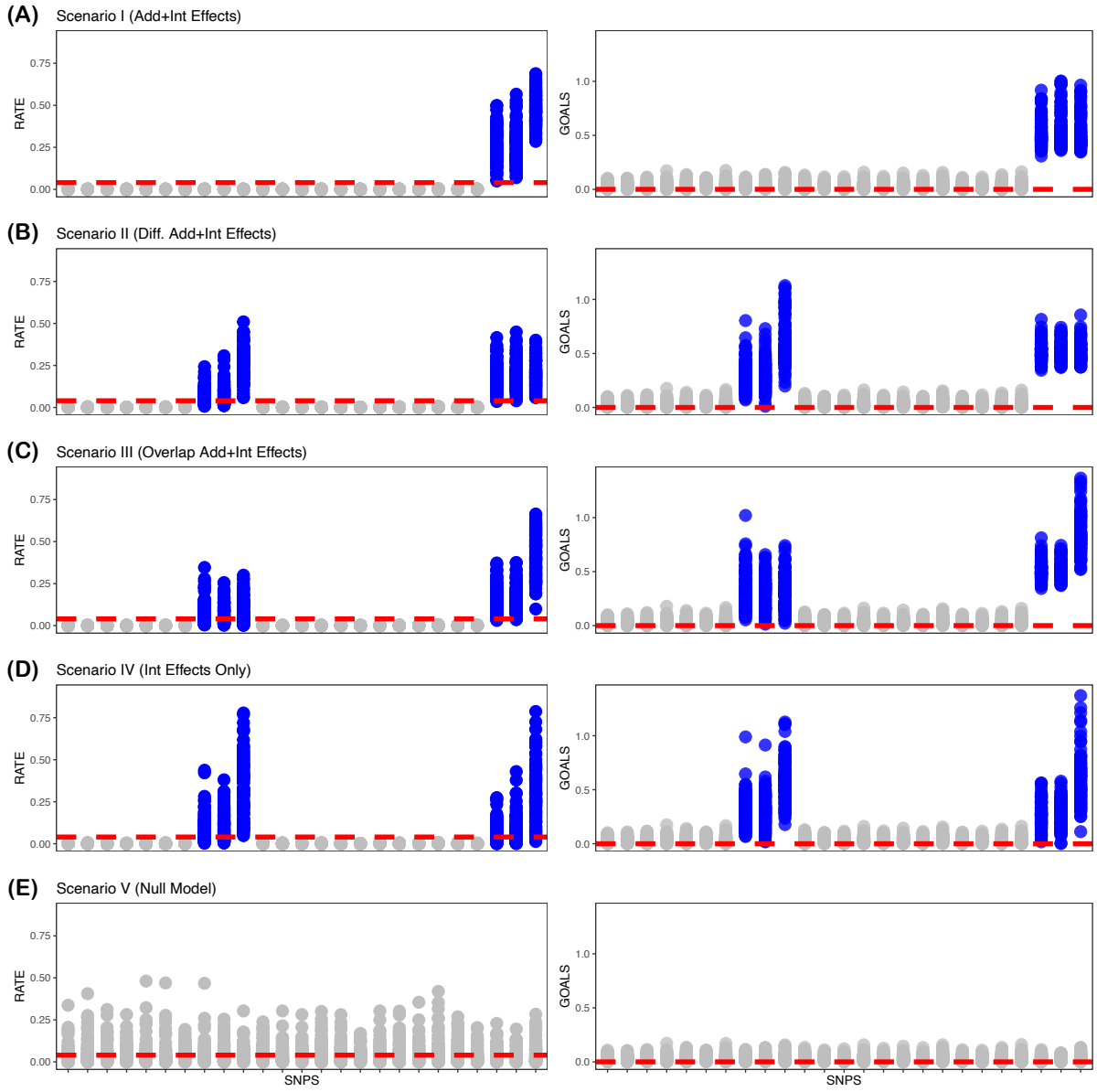

Figure 1. Proof-of-concept simulations to demonstrate how GOALS and RATE globally prioritize associated variants with varying degrees of additive and interaction effects. These simple simulations assume that synthetic traits have broad-sense heritability $H^2 = 0.6$ with interaction effects contributing to $(1 - \rho) = 0\%$ to 50% of the phenotypic variation. SNPs highlighted in blue have nonzero effects within each of the five different scenarios. To compare the null hypotheses for the two approaches, we also display red dashed lines that are drawn at the level of relative equivalence (i.e., $1/J$) for RATE (left column) and at zero for GOALS (right column), respectively. Note that the scales of the y-axes are different because RATE is theoretically bounded on the unit interval [0, 1]. Here, the main takeaway is that, because the GOALS operator measures variable importance in function space, it is more robust to identifying variants whose associations are driven primarily by interaction effects. All results shown in this figure are based on 100 replicates.

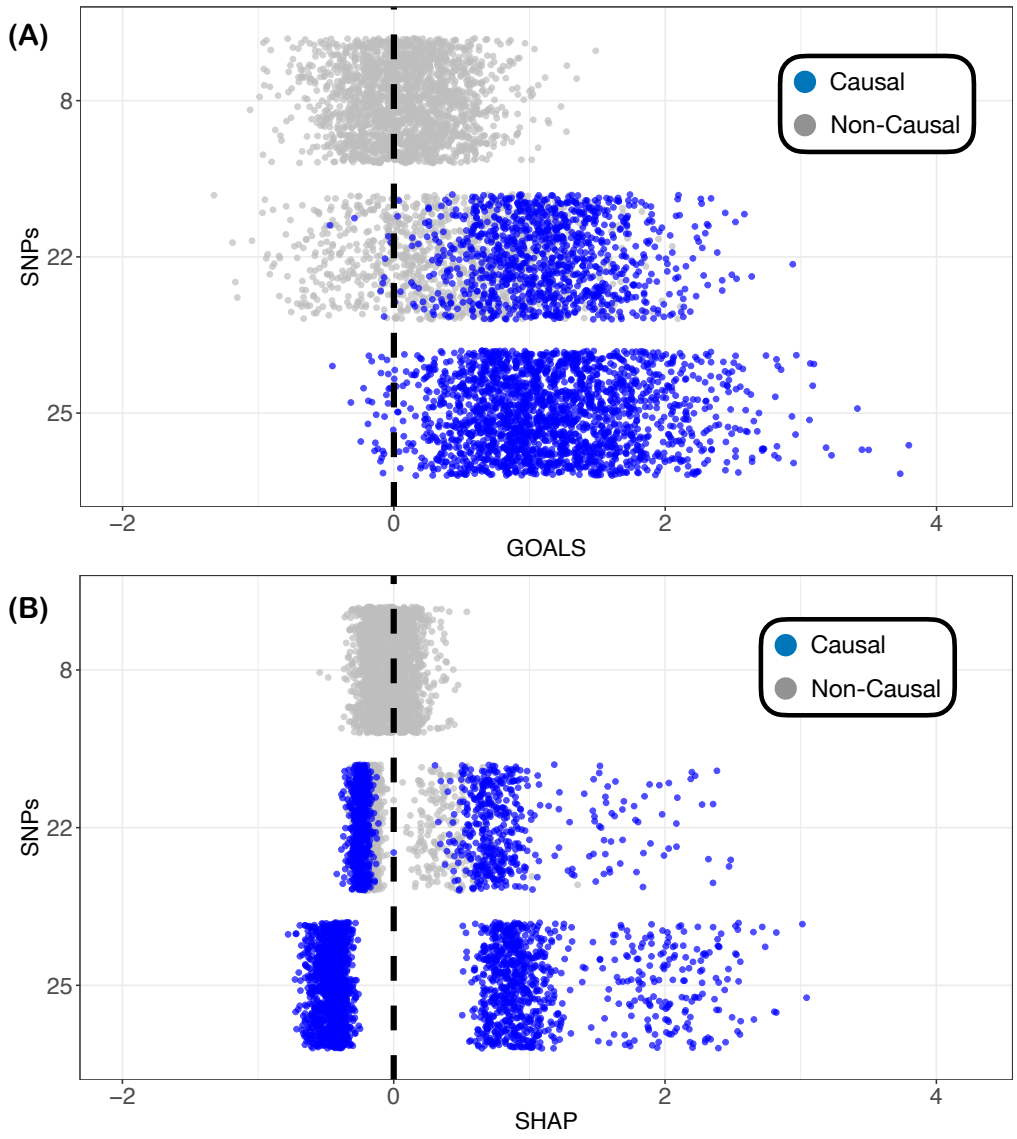

Figure 2. Proof-of-concept simulations to demonstrate how GOALS and SHAP locally prioritize associated variants that have varying level of effects on specific subsets of the population. These simple simulations assume that synthetic traits have broad-sense heritability $H^2 = 0.6$ with interaction effects contributing to $(1 - \rho) = 50\%$ of the phenotypic variation. Here, each point is an individual. We highlight the local variable importance metrics for three specific SNPs according **(A)** GOALS and **(B)** Shapley Additive Explanations (SHAP). In this simulation study, SNP #8 is null feature and does not contribute to the phenotypic variation; SNP #25 has additive and interaction effects that are associated with all individuals; and SNP #22 has an additive effect for only half of the population. A point is blue if the corresponding labeled SNP has a nonzero effect for that individual. The main takeaway of this analysis is that, in the case where a SNP has an effect on trait architecture for only a subset of the observed population, the local distribution of the SHAP and GOALS operators will be multimodal allowing for individualized summaries of marker effects on specific observations. The black dashed line is drawn at zero to represent a threshold where a SNP has no effect for a given sample.

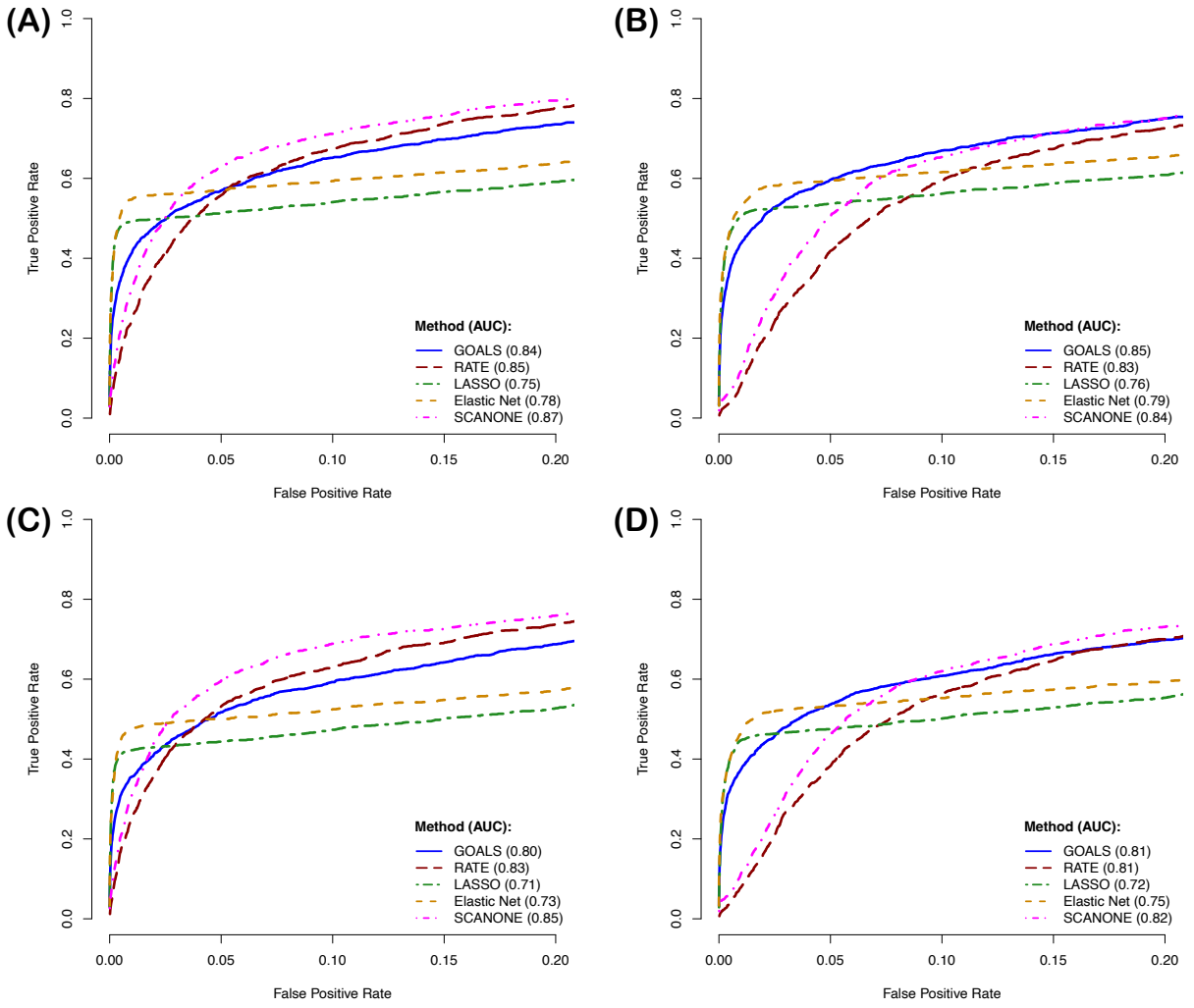

Figure 3. Receiving operating characteristic (ROC) curves comparing the performance of GOALS against other global variable importance approaches in simulations. Here, synthetic traits are simulated with broad-sense heritability $H^2 = 0.3$ with only additive effects in panels **(A)** and **(B)**, and a combination of additive and pairwise interaction effects in panels **(C)** and **(D)**. This is controlled by a free parameter $\rho = \{0.5, 1\}$ which was used to determine the proportion of phenotypic variance that is contributed by additivity. The traits simulated in panels **(B)** and **(D)** also have the additional complexity of having population stratification effects. Competing approaches include: Gaussian process regression with GOALS (blue), Gaussian process regression with RATE (red), LASSO regularization (green), the Elastic Net (yellow), and the SCANONE method (pink). Note that the upper limit of the x-axis (i.e., false positive rate) has been truncated at 0.20. All results are based on 100 simulated replicates.

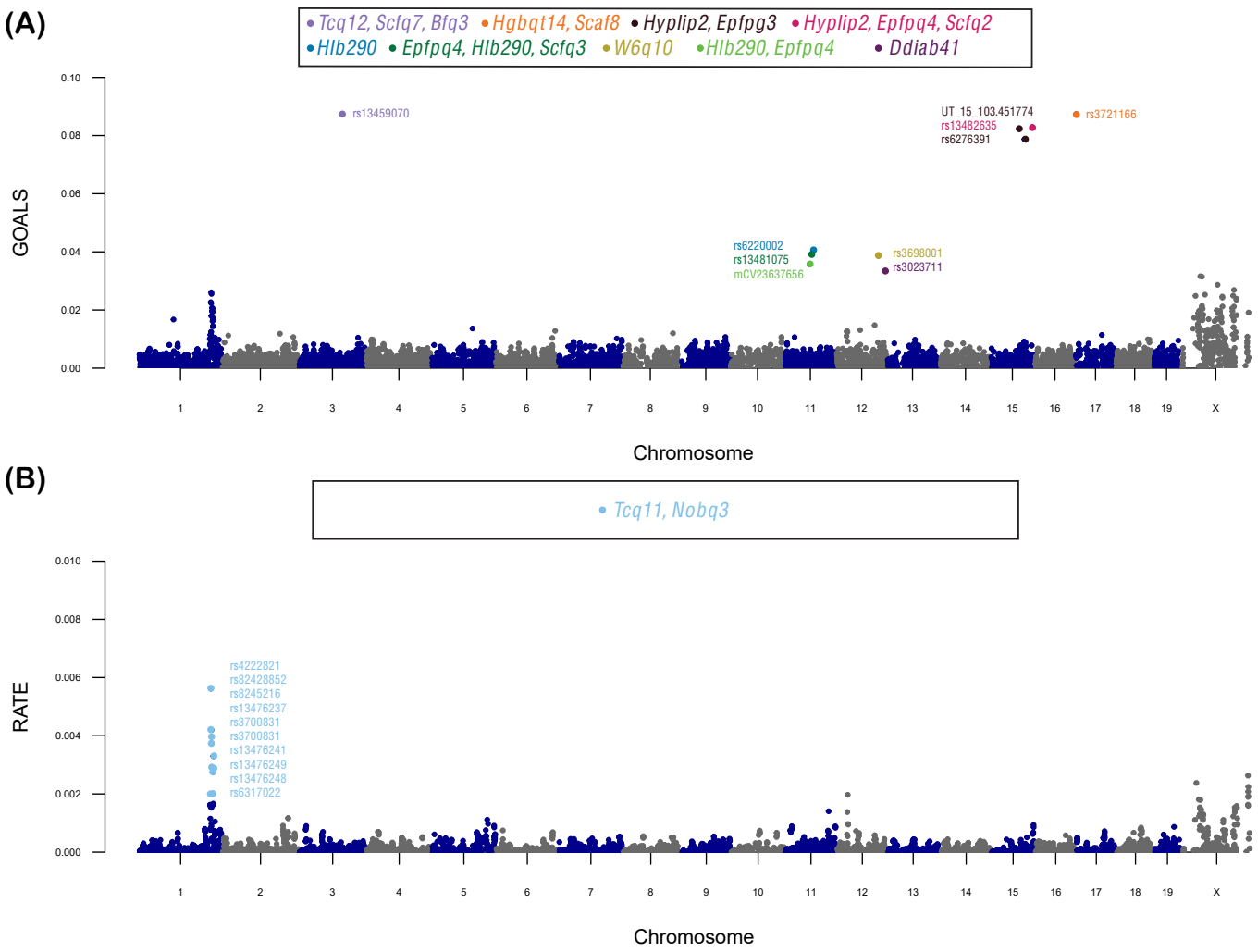

Figure 4. Manhattan plot of variant-level association mapping results for high-density lipoprotein (HDL) content in the heterogenous stock of mice dataset from the Wellcome Trust Centre of Human Genetics (Valdar et al., 2006a,b). Panel **(A)** depicts the global GOALS measure of quality-control-positive SNPs plotted against their genomic positions after running a Bayesian Gaussian process (GP) regression on the quantitative trait. As a direct comparison, in panel **(B)**, we also include results after implementing RATE on the same fitted GP model. In this figure, chromosomes are shown in alternating colors for clarity. The top 10 highest ranked SNPs by GOALS and RATE, respectively, are labeled and color coded based on their nearest mapped gene(s) as cited by the Mouse Genome Informatics database (`http://www.informatics.jax.org/`) (Bult et al., 2019). These annotated genes are listed in the legends of each panel. A complete list of the GOALS and RATE values for all SNPs can be found in Table S1.

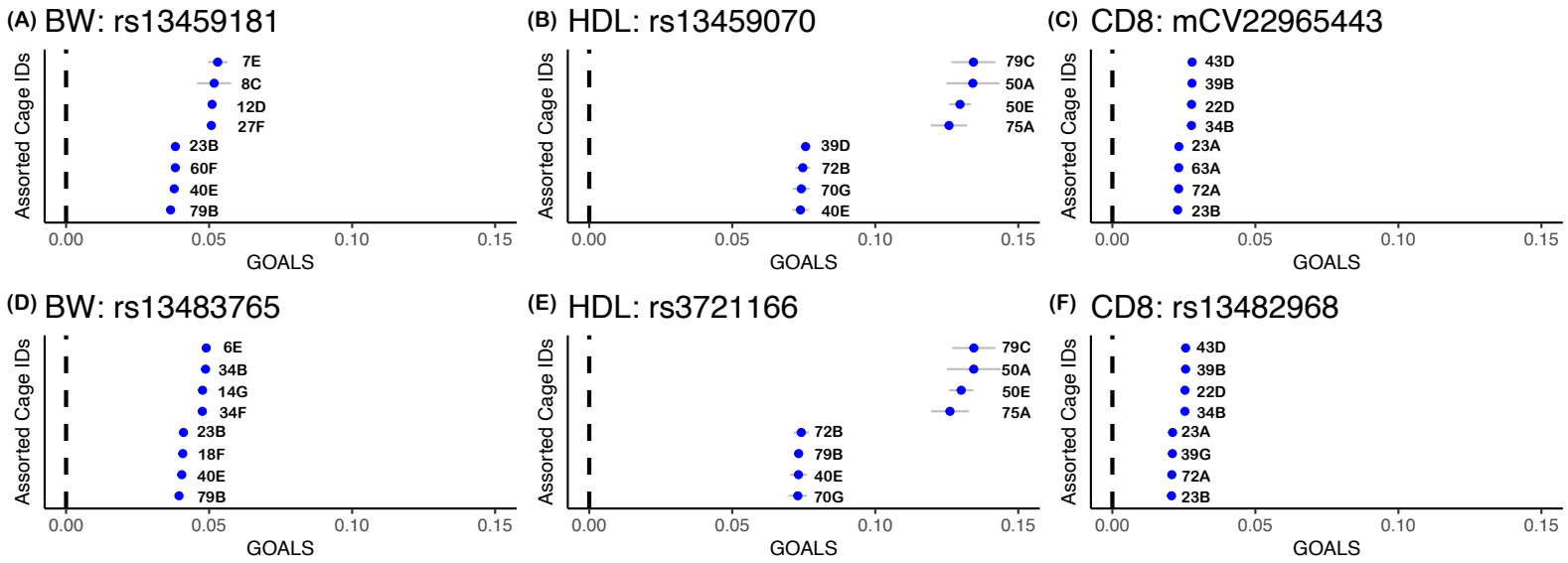

Figure 5. Plot of local variable importance according to GOALS as a function of cage for notable genetic variants in the analysis of the heterogenous stock of mice dataset from the Wellcome Trust Centre of Human Genetics (Valdar et al., 2006a,b) Here, we show how SNPs have varying levels of importance for individual mice depending on the cage they were assigned to in the study. The traits analyzed here include **(A, D)** body weight (BW); **(B, E)** high-density lipoprotein (HDL) content; and **(C, F)** percentage of CD8+ cells. The blue points are mean local GOALS value for each cage and the grey lines show the total distribution. In this plot, we take the two SNPs with the greatest global GOALS value in each trait and plot the local values for the 4 cages with the greatest and least local means. The black dashed line is drawn at zero to represent a threshold where a SNP has no effect for a given mouse.

References

Q. Ai and L. Narayanan. R. Model-agnostic vs. model-intrinsic interpretability for explainable product search. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 5–15, 2021.

A. M. Alaa and M. van der Schaar. Bayesian nonparametric causal inference: information rates and learning algorithms. *arXiv*, page 1712.08914, 2017.

M. M. Barbieri and J. O. Berger. Optimal predictive model selection. *Ann Statist*, 32(3):870–897, 2004. doi: 10.1214/009053604000000238. URL `http://projecteuclid.org/euclid.aos/1085408489`.

V. Bourgeais, F. Zehraoui, M. B. Hamdoune, and B. Hanczar. Deep GONet: self-explainable deep neural network based on gene ontology for phenotype prediction from gene expression data. *BMC Bioinformatics*, 22(S10), May 2021. doi: 10.1186/s12859-021-04370-7. URL `https://doi.org/10.1186/s12859-021-04370-7`.

V. Bourgeais, F. Zehraoui, and B. Hanczar. Graphgonet: a self-explaining neural network encapsulating the gene ontology graph for phenotype prediction on gene expression. *Bioinformatics*, 38(9):2504–2511, 2022.

B. K. Bulik-Sullivan, P.-R. Loh, H. K. Finucane, S. Ripke, J. Yang, N. Patterson, M. J. Daly, A. L. Price, B. M. Neale, and S. W. G. of the Psychiatric Genomics Consortium. Ld score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics*, 47(3):291–295, 2015. doi: 10.1038/ng.3211. URL `https://doi.org/10.1038/ng.3211`.

C. J. Bult, J. A. Blake, C. L. Smith, J. A. Kadin, J. E. Richardson, the Mouse Genome Database Group, A. Anagnostopoulos, R. Asabor, R. M. Baldarelli, J. S. Beal, S. M. Bello, O. Blodgett, N. E. Butler, K. R. Christie, L. E. Corbani, J. Creelman, M. E. Dolan, H. J. Drabkin, S. L. Giannatto, P. Hale, D. P. Hill, M. Law, A. Mendoza, M. McAndrews, D. Miers, H. Motenko, L. Ni, H. Onda, M. Perry, J. M. Recla, B. Richards-Smith, D. Sitnikov, M. Tomczuk, G. Tonorio, L. Wilming, and Y. Zhu. Mouse Genome Database (MGD) 2019. *Nucleic Acids Research*, 47(D1):D801–D806, Jan. 2019. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gky1056. URL `https://academic.oup.com/nar/article/47/D1/D801/5165331`.

E. Candès, Y. Fan, L. Janson, and J. Lv. Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3):551–577, June 2018. ISSN 1369-7412, 1467-9868. doi: 10.1111/rssb.12265. URL `https://onlinelibrary.wiley.com/doi/10.1111/rssb.12265`.

P. Carbonetto and M. Stephens. 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 Genet*, 9(10):e1003770, 2013. URL `https://doi.org/10.1371/journal.pgen.1003770`.

D. V. Carvalho, E. M. Pereira, and J. S. Cardoso. Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8):832, 2019.

A. Chaudhuri, D. Kakde, C. Sadek, L. Gonzalez, and S. Kong. The mean and median criterion for automatic kernel bandwidth selection for support vector data description. *arXiv*, page 1708.05106, 2017.

H. Chen, S. M. Lundberg, and S.-I. Lee. Explaining a series of models by propagating shapley values. *Nature communications*, 13(1):1–15, 2022.

X. Chen, R. McClusky, J. Chen, S. W. Beaven, P. Tontonoz, A. P. Arnold, K. Reue, and A. Attie. The number of x chromosomes causes sex differences in adiposity in mice. *PLoS Genetics*, 8, May 2012. ISSN 1553-7404. doi: 10.1371/journal.pgen.1002709. URL https://dx.plos.org/10.1371/journal.pgen.1002709.

W. Cheng, S. Ramachandran, and L. Crawford. Estimation of non-null snp effect size distributions enables the detection of enriched genes underlying complex traits. *PLoS Genet*, 16(6):1–48, 06 2020. doi: 10.1371/journal.pgen.1008855. URL https://doi.org/10.1371/journal.pgen.1008855.

J. M. Cheverud, T. H. Ehrich, T. Hrbek, J. P. Kenney, L. S. Pletscher, and C. F. Semenkovich. Quantitative trait loci for obesity- and diabetes-related traits and their dietary responses to high-fat feeding in lgxsm recombinant inbred mouse strains. *Diabetes*, 53(12):3328–3336, Dec 2004. ISSN 0012-1797 (Print); 0012-1797 (Linking). doi: 10.2337/diabetes.53.12.3328.

A. Cotter, J. Keshet, and N. Srebro. Explicit approximations of the Gaussian kernel. *arXiv*, page 1109.4603, 2011.

L. Crawford, P. Zeng, S. Mukherjee, and X. Zhou. Detecting epistasis with the marginal epistasis test in genetic mapping studies of quantitative traits. *PLoS Genet*, 13(7):e1006869, July 2017. ISSN 1553-7404. doi: 10.1371/journal.pgen.1006869. URL https://dx.plos.org/10.1371/journal.pgen.1006869.

L. Crawford, K. C. Wood, X. Zhou, and S. Mukherjee. Bayesian Approximate Kernel Regression With Variable Selection. *Journal of the American Statistical Association*, 113 (524):1710–1721, 2018. doi: 10.1080/01621459.2017.1361830.

L. Crawford, S. R. Flaxman, D. E. Runcie, and M. West. Variable prioritization in nonlinear black box methods: A genetic association case study. *The Annals of Applied Statistics*, 13(2):958–989, June 2019. ISSN 1932-6157. doi: 10.1214/18-AOAS1222. URL https://projecteuclid.org/euclid.aoas/1560758434.

J. Cuevas, J. Crossa, O. A. Montesinos-López, J. Burgueño, P. Pérez-Rodríguez, and G. de los Campos. Bayesian genomic prediction with genotype × environment interaction kernel models. *G3 (Bethesda)*, 7(1):41–53, 2017. doi: 10.1534/g3.116.035584. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5217122/.

C. A. de Leeuw, J. M. Mooij, T. Heskes, and D. Posthuma. Magma: generalized gene-set analysis of gwas data. *PLoS Comput Biol*, 11(4):e1004219–, 2015. URL https://doi.org/10.1371/journal.pcbi.1004219.

G. de los Campos, H. Naya, D. Gianola, J. Crossa, A. Legarra, E. Manfredi, K. Weigel, and J. Cotes. Predicting quantitative traits with regression models for dense molecular markers and pedigree. *Genetics*, 182(1):375–385, 2009. URL http://www.genetics.org/content/182/1/375.abstract.

A. J. DeGrave, J. D. Janizek, and S.-I. Lee. Ai for radiographic covid-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3(7):610–619, 2021.

P. Demetci, W. Cheng, G. Darnell, X. Zhou, S. Ramachandran, and L. Crawford. Multi-scale inference of genetic trait architecture using biologically annotated neural networks. *PLoS Genetics*, 17(8):e1009754, 2021.

F. Doshi-Velez and B. Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

H. A. Elmarakeby, J. Hwang, R. Arafeh, J. Crowdis, S. Gang, D. Liu, S. H. AlDubayan, K. Salari, S. Kregel, C. Richter, et al. Biologically informed deep neural network for prostate cancer discovery. *Nature*, 598(7880):348–352, 2021.

A. Gelman and J. Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Analytical Methods for Social Research. Cambridge University Press, 2006. doi: 10.1017/CBO9780511790942.

A. Gelman, J. Hwang, and A. Vehtari. Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, 24(6):997–1016, Nov. 2014. ISSN 0960-3174, 1573-1375. doi: 10.1007/s11222-013-9416-2. URL `http://link.springer.com/10.1007/s11222-013-9416-2`.

A. Gordon, G. Glazko, X. Qiu, and A. Yakovlev. Control of the mean number of false discoveries, bonferroni and stability of multiple testing. *The Annals of Applied Statistics*, 1(1):179–190, 6 2007. doi: 10.1214/07-AOAS102. URL `https://doi.org/10.1214/07-AOAS102`.

C. Goutis and C. P. Robert. Model choice in generalised linear models: a Bayesian approach via Kullback-Leibler projections. *Biometrika*, 85(1):29–37, 1998.

L. Gu, M. W. Johnson, and A. J. Lusis. Quantitative trait locus analysis of plasma lipoprotein levels in an autoimmune mouse model : interactions between lipoprotein metabolism, autoimmune disease, and atherogenesis. *Arterioscler Thromb Vasc Biol*, 19(2):442–453, Feb 1999. ISSN 1079-5642 (Print); 1079-5642 (Linking). doi: 10.1161/01.atv.19.2.442.

R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi. A survey of methods for explaining black box models. *ACM Computing Surveys (CSUR)*, 51(5):93, 2018.

P. Hall. Guidelines for responsible and human-centered use of explainable machine learning. *arXiv preprint arXiv:1906.03533*, 2019.

T. J. Hayeck, N. A. Zaitlen, P.-R. Loh, B. Vilhjalmsson, S. Pollack, A. Gusev, J. Yang, G.-B. Chen, M. E. Goddard, P. M. Visscher, N. Patterson, and A. L. Price. Mixed model with correction for case-control ascertainment increases association power. *The American Journal of Human Genetics*, 96(5):720–730, May 2015. doi: 10.1016/j.ajhg.2015.03.004. URL `https://doi.org/10.1016/j.ajhg.2015.03.004`.

D. Heckerman, D. Gurdasani, C. Kadie, C. Pomilla, T. Carstensen, H. Martin, K. Ekoru, R. N. Nsubuga, G. Ssenyomo, A. Kamali, P. Kaleebu, C. Widmer, and M. S. Sandhu. Linear mixed model for heritability estimation that explicitly addresses environmental variation. *Proceedings of the National Academy of Sciences*, 113(27):7377–7382, July 2016. doi: 10.1073/pnas.1510497113. URL `https://doi.org/10.1073/pnas.1510497113`.

F. Hoti and M. J. Sillanpää. Bayesian mapping of genotype × expression interactions in quantitative and qualitative traits. *Heredity*, 97(1):4–18, May 2006. doi: 10.1038/sj.hdy.6800817. URL `https://doi.org/10.1038/sj.hdy.6800817`.

J. Ish-Horowicz, D. Udwin, S. Flaxman, S. Filippi, and L. Crawford. Interpreting deep neural networks through variable importance. *arXiv preprint arXiv:1901.09839*, 2019.

L. Jiang, Z. Zheng, T. Qi, K. E. Kemper, N. R. Wray, P. M. Visscher, and J. Yang. A resource-efficient tool for mixed model association analysis of large-scale data. *Nature Genetics*, 51(12):1749–1755, Nov. 2019. doi: 10.1038/s41588-019-0530-8. URL `https://doi.org/10.1038/s41588-019-0530-8`.

H. M. Kang, N. A. Zaitlen, C. M. Wade, A. Kirby, D. Heckerman, M. J. Daly, and E. Eskin. Efficient control of population structure in model organism association mapping. *Genetics*, 178(3):1709–1723, Mar. 2008. doi: 10.1534/genetics.107.080101. URL `https://doi.org/10.1534/genetics.107.080101`.

H. M. Kang, J. H. Sul, S. K. Service, N. A. Zaitlen, S. yee Kong, N. B. Freimer, C. Sabatti, and E. Eskin. Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, 42(4):348–354, Mar. 2010. doi: 10.1038/ng.548. URL https://doi.org/10.1038/ng.548.

G. Karlebach and R. Shamir. Modelling and analysis of gene regulatory networks. *Nature reviews Molecular cell biology*, 9(10):770–780, 2008.

S. V. Kim, W. Z. Mehal, X. Dong, V. Heinrich, M. Pypaert, I. Mellman, M. Dembo, M. S. Mooseker, D. Wu, and R. A. Flavell. Modulation of cell adhesion and motility in the immune system by myo1f. *Science*, 314(5796):136–139, Oct 2006. ISSN 1095-9203 (Electronic); 0036-8075 (Linking). doi: 10.1126/science.1131920.

P.-J. Kindermans, S. Hooker, J. Adebayo, M. Alber, K. T. Schütt, S. Dähne, D. Erhan, and B. Kim. *The (Un)reliability of Saliency Methods*, pages 267–280. Springer International Publishing, Cham, 2019. ISBN 978-3-030-28954-6. doi: 10.1007/978-3-030-28954-6_14. URL https://doi.org/10.1007/978-3-030-28954-6_14.

A. N. Kolmogorov and Y. A. Rozanov. On strong mixing conditions for stationary Gaussian processes. *Theory Probab Its Appl*, 5(2):204–208, 1960.

A. Korte, B. J. Vilhjálmsson, V. Segura, A. Platt, Q. Long, and M. Nordborg. A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nature Genetics*, 44 (9):1066–1071, Aug. 2012. doi: 10.1038/ng.2376. URL https://doi.org/10.1038/ng.2376.

D. R. Kowal. Fast, Optimal, and Targeted Predictions Using Parameterized Decision Analysis. *Journal of the American Statistical Association*, pages 1–12, Apr. 2021. ISSN 0162-1459, 1537-274X. doi: 10.1080/01621459.2021.1891926. URL https://www.tandfonline.com/doi/full/10.1080/01621459.2021.1891926.

D. Lamparter, D. Marbach, R. Rueedi, Z. Kutalik, and S. Bergmann. Fast and rigorous computation of gene and pathway scores from snp-based summary statistics. *PLoS Comput Biol*, 12(1):e1004714, 2016. URL https://doi.org/10.1371/journal.pcbi.1004714.

H. A. Lawson, A. Lee, G. L. Fawcett, B. Wang, L. S. Pletscher, T. J. Maxwell, T. H. Ehrich, J. P. Kenney-Hunt, J. B. Wolf, C. F. Semenkovich, and J. M. Cheverud. The importance of context to the genetic architecture of diabetes-related traits is revealed in a genome-wide scan of a lg/j × sm/j murine model. *Mamm Genome*, 22(3-4):197–208, 2011. ISSN 1432-1777 (Electronic); 0938-8990 (Print); 0938-8990 (Linking). doi: 10.1007/s00335-010-9313-3.

C. Lippert, J. Listgarten, Y. Liu, C. M. Kadie, R. I. Davidson, and D. Heckerman. FaST linear mixed models for genome-wide association studies. *Nature Methods*, 8(10):833–835, Sept. 2011. doi: 10.1038/nmeth.1681. URL https://doi.org/10.1038/nmeth.1681.

J. Z. Liu, A. F. Mcrae, D. R. Nyholt, S. E. Medland, N. R. Wray, K. M. Brown, N. K. Hayward, G. W. Montgomery, P. M. Visscher, N. G. Martin, et al. A versatile gene-based test for genome-wide association studies. *American Journal of Human Genetics*, 87(1):139–145, 2010.

P.-R. Loh, G. Kichaev, S. Gazal, A. P. Schoech, and A. L. Price. Mixed-model association for biobank-scale datasets. *Nature Genetics*, 50(7):906–908, June 2018. doi: 10.1038/s41588-018-0144-6. URL https://doi.org/10.1038/s41588-018-0144-6.

M. I. Love, W. Huber, and S. Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15(12):1–21, 2014.

S. Lundberg and S.-I. Lee. A Unified Approach to Interpreting Model Predictions. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 4768–4777, 2017. ISSN 9781510860964. URL https://dl.acm.org/doi/10.5555/3295222.3295230.

S. M. Lundberg and S. Lee. An unexpected unity among methods for interpreting model predictions. *CoRR*, abs/1611.07478, 2016. URL http://arxiv.org/abs/1611.07478.

J. Ma, M. K. Yu, S. Fong, K. Ono, E. Sage, B. Demchak, R. Sharan, and T. Ideker. Using deep learning to model the hierarchical structure and function of a cell. *Nature Methods*, 15(4):290–298, Mar. 2018. doi: 10.1038/nmeth.4627. URL https://doi.org/10.1038/nmeth.4627.

T. F. C. Mackay. Epistasis and quantitative traits: using model organisms to study gene-gene interactions. *Nat Rev Genet*, 15(1):22–33, 2014. URL http://dx.doi.org/10.1038/nrg3627.

G. L. Manno, R. Soldatov, A. Zeisel, E. Braun, H. Hochgerner, V. Petukhov, K. Lidschreiber, M. E. Kastriti, P. Lönnerberg, A. Furlan, J. Fan, L. E. Borm, Z. Liu, D. van Bruggen, J. Guo, X. He, R. Barker, E. Sundström, G. Castelo-Branco, P. Cramer, I. Adameyko, S. Linnarsson, and P. V. Kharchenko. RNA velocity of single cells. *Nature*, 560(7719):494–498, Aug. 2018. doi: 10.1038/s41586-018-0414-6. URL https://doi.org/10.1038/s41586-018-0414-6.

A. R. Martin, M. Kanai, Y. Kamatani, Y. Okada, B. M. Neale, and M. J. Daly. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature Genetics*, 51(4):584–591, Apr. 2019. ISSN 1061-4036, 1546-1718. doi: 10.1038/s41588-019-0379-x. URL http://www.nature.com/articles/s41588-019-0379-x.

A. K. Miller, A. Chen, J. Bartlett, L. Wang, S. M. Williams, and D. A. Buchner. A novel mapping strategy utilizing mouse chromosome substitution strains identifies multiple epistatic interactions that regulate complex traits. *G3 (Bethesda)*, 10(12):4553–4563, Dec 2020. ISSN 2160-1836 (Electronic); 2160-1836 (Linking). doi: 10.1534/g3.120.401824.

C. J. A. Moen, A. P. Tholens, P. J. Voshol, W. de Haan, L. M. Havekes, P. Gargalovic, A. J. Lusis, K. W. van Dyk, R. R. Frants, M. H. Hofker, and P. C. N. Rensen. The hyplip2 locus causes hypertriglyceridemia by decreased clearance of triglycerides. *J Lipid Res*, 48(10):2182–2192, Oct 2007. ISSN 0022-2275 (Print); 0022-2275 (Linking). doi: 10.1194/jlr.M700009-JLR200.

W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44):22071–22080, Oct. 2019. doi: 10.1073/pnas.1900654116. URL https://doi.org/10.1073/pnas.1900654116.

P. Nakka, B. J. Raphael, and S. Ramachandran. Gene and network analysis of common variants reveals novel associations in multiple complex diseases. *Genetics*, 204(2):783–798, 2016.

M. J. Nueda, S. Tarazona, and A. Conesa. Next masigpro: updating masigpro bioconductor package for rna-seq time series. *Bioinformatics*, 30(18):2598–2602, 2014.

C. Östergren, J. Shim, J. V. Larsen, L. B. Nielsen, and J. F. Bentzon. Genetic analysis of ligation-induced neointima formation in an f2 intercross of c57bl/6 and fvb/n inbred mouse strains. *PLoS One*, 10(4):e0121899, 2015. ISSN 1932-6203 (Electronic); 1932-6203 (Linking). doi: 10.1371/journal.pone.0121899.

T. Paananen, J. Piironen, M. R. Andersen, and A. Vehtari. Variable selection for gaussian processes via sensitivity analysis of the posterior predictive distribution. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1743–1752. PMLR, 2019.

T. Paananen, M. R. Andersen, and A. Vehtari. Uncertainty-aware sensitivity analysis using rényi divergences. In *Uncertainty in Artificial Intelligence*, pages 1185–1194. PMLR, 2021.

P. Perez and G. de los Campos. Genome-wide regression and prediction with the bglr statistical package. *Genetics*, 198(2):483–495, 2014.

P. C. Phillips. Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems. *Nat Rev Genet*, 9(11):855–867, 2008. doi: 10.1038/nrg2452. URL `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2689140/`.

J. Piironen and A. Vehtari. Projection predictive model selection for Gaussian processes. *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6, Sept. 2016. doi: 10.1109/MLSP.2016.7738829. URL `http://arxiv.org/abs/1510.04813`. arXiv: 1510.04813.

J. Piironen and A. Vehtari. Comparison of Bayesian predictive methods for model selection. *Statistics and Computing*, 27(3):711–735, May 2017. ISSN 0960-3174, 1573-1375. doi: 10.1007/s11222-016-9649-y. URL `http://link.springer.com/10.1007/s11222-016-9649-y`.

A. L. Price, N. A. Zaitlen, D. Reich, and N. Patterson. New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics*, 11(7):459–463, June 2010. doi: 10.1038/nrg2813. URL `https://doi.org/10.1038/nrg2813`.

C. E. Rasmussen and C. K. I. Williams. *Gaussian processes for machine learning*. Adaptive computation and machine learning. MIT Press, Cambridge, Mass, 2006. ISBN 978-0-262-18253-9. OCLC: ocm61285753.

M. D. Richard and R. P. Lippmann. Neural network classifiers estimate bayesian a posteriori probabilities. *Neural Comput*, 3(4):461–483, Winter 1991. ISSN 1530-888X (Electronic); 0899-7667 (Linking). doi: 10.1162/neco.1991.3.4.461.

M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47–e47, 01 2015. ISSN 0305-1048. doi: 10.1093/nar/gkv007. URL `https://doi.org/10.1093/nar/gkv007`.

M. D. Robinson, D. J. McCarthy, and G. K. Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 11 2009. ISSN 1367-4803. doi: 10.1093/bioinformatics/btp616. URL `https://doi.org/10.1093/bioinformatics/btp616`.

A. E. Roth. *The Shapley value: essays in honor of Lloyd S. Shapley*. Cambridge University Press, 1988.

C. Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

C. Rudin. Why black box machine learning should be avoided for high-stakes decisions, in brief. *Nature Reviews Methods Primers*, 2(1):81, 2022. doi: 10.1038/s43586-022-00172-0. URL `https://doi.org/10.1038/s43586-022-00172-0`.

D. E. Runcie and L. Crawford. Fast and flexible linear mixed models for genome-wide genetics. *PLOS Genetics*, 15(2):e1007978, Feb. 2019. doi: 10.1371/journal.pgen.1007978. URL `https://doi.org/10.1371/journal.pgen.1007978`.

D. E. Runcie, J. Qu, H. Cheng, and L. Crawford. MegaLMM: Mega-scale linear mixed models for genomic predictions with thousands of traits. *Genome Biology*, 22(1), July 2021. doi: 10.1186/s13059-021-02416-w. URL `https://doi.org/10.1186/s13059-021-02416-w`.

B. Servin and M. Stephens. Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS Genet*, 3(7):e114–, 2007. URL https://doi.org/10.1371/journal.pgen.0030114.

M. Sesia, E. Katsevich, S. Bates, E. Candès, and C. Sabatti. Multi-resolution localization of causal variants across the genome. *Nature Communications*, 11(1):1093, 2020. doi: 10.1038/s41467-020-14791-2. URL https://doi.org/10.1038/s41467-020-14791-2.

M. Sesia, S. Bates, E. Candès, J. Marchini, and C. Sabatti. False discovery rate control in genome-wide association studies with population structure. *Proceedings of the National Academy of Sciences*, 118 (40):e2105841118, 2021. doi: 10.1073/pnas.2105841118. URL https://www.pnas.org/doi/abs/10.1073/pnas.2105841118.

L. S. Shapley. *Notes on the N-person Game–I: Characteristic-point Solutions of the Four-person Game*. Rand Corporation, 1951.

J. Q. Shi, B. Wang, E. J. Will, and R. M. West. Mixed-effects gaussian process functional regression models with application to dose–response curve prediction. *Stat Med*, 31(26):3165–3177, 2012. doi: 10.1002/sim.4502. URL https://doi.org/10.1002/sim.4502.

K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *In Workshop at International Conference on Learning Representations*. Citeseer, 2014.

A. Smith, P. A. Naik, and C.-L. Tsai. Markov-switching model selection using Kullback–Leibler divergence. *J Econom*, 134(2):553–577, 2006.

J. Stamp, A. DenAdel, D. Weinreich, and L. Crawford. Leveraging the genetic correlation between traits improves the detection of epistasis in genome-wide association studies. *bioRxiv*, 2022. doi: 10.1101/2022.11.30.518547. URL https://www.biorxiv.org/content/early/2022/12/01/2022.11.30.518547.

M. Stephens. False discovery rates: a new deal. *Biostatistics*, page kxw041, Oct. 2016. ISSN 1465-4644, 1468-4357. doi: 10.1093/biostatistics/kxw041. URL https://academic.oup.com/biostatistics/article-lookup/doi/10.1093/biostatistics/kxw041.

M. Stephens and D. J. Balding. Bayesian statistical methods for genetic association studies. *Nature Reviews Genetics*, 10(10):681–690, Oct. 2009. doi: 10.1038/nrg2615. URL https://doi.org/10.1038/nrg2615.

R. Sun, S. Hui, G. D. Bader, X. Lin, and P. Kraft. Powerful gene set analysis in gwas with the generalized berk-jones statistic. *PLoS Genet*, 15(3):e1007530, 2019. URL https://doi.org/10.1371/journal.pgen.1007530.

S. Tan, R. Caruana, G. Hooker, and Y. Lou. Detecting bias in black-box models using transparent model distillation. *arXiv*, page 1710.06169, 2017.

B. A. Taylor, L. M. Tarantino, and S. J. Phillips. Gender-influenced obesity QTLs identified in a cross involving the KK type II diabetes-prone mouse strain. *Mammalian Genome*, 10(10):963–968, Oct. 1999. ISSN 0938-8990, 1432-1777. doi: 10.1007/s003359901141. URL http://link.springer.com/10.1007/s003359901141.

The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678, 2007. URL http://dx.doi.org/10.1038/nature05911.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996. ISSN 00359246. URL http://www.jstor.org/stable/2346178.

M. Tsang, D. Cheng, and Y. Liu. Detecting statistical interactions from neural network weights. In *International Conference on Learning Representations*, 2018a. URL https://openreview.net/forum?id=ByOfBggRZ.

M. Tsang, H. Liu, S. Purushotham, P. Murali, and Y. Liu. Neural interaction transparency (nit): Disentangling learned interactions for improved interpretability. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018b. URL https://proceedings.neurips.cc/paper/2018/file/74378afe5e8b20910cf1f939e57f0480-Paper.pdf.

W. Valdar, J. Flint, and R. Mott. Simulating the collaborative cross: power of quantitative trait loci detection and mapping resolution in large sets of recombinant inbred strains of mice. *Genetics*, 172(3): 1783–1797, 2006a. ISSN 0016-6731. doi: 10.1534/genetics.104.039313.

W. Valdar, L. C. Solberg, D. Gauguier, S. Burnett, P. Klenerman, W. O. Cookson, M. S. Taylor, J. N. P. Rawlins, R. Mott, and J. Flint. Genome-wide genetic association of complex traits in heterogeneous stock mice. *Nature Genetics*, 38(8):879–887, 2006b. ISSN 1546-1718. doi: 10.1038/ng1840. URL https://www.nature.com/articles/ng1840. Number: 8 Publisher: Nature Publishing Group.

G. Wahba. *Splines models for observational data*, volume 59 of *Series in Applied Mathematics*. SIAM, Philadelphia, PA, 1990.

O. Weissbrod, D. Geiger, and S. Rosset. Multikernel linear mixed models for complex phenotype prediction. *Genome Res*, 26(7):969–979, 2016. URL http://genome.cshlp.org/content/26/7/969.abstract.

S. Wold, A. Ruhe, H. Wold, and W. J. Dunn, III. The collinearity problem in linear regression. the partial least squares (pls) approach to generalized inverses. *SIAM Journal on Scientific and Statistical Computing*, 5(3):735–743, 1984. doi: 10.1137/0905052. URL https://doi.org/10.1137/0905052.

J. H. Woo, Y. Shimoni, W. S. Yang, P. Subramaniam, A. Iyer, P. Nicoletti, M. R. Martínez, G. López, M. Mattioli, R. Realubit, et al. Elucidating compound mechanism of action by network perturbation analysis. *Cell*, 162(2):441–451, 2015.

S. Woody, C. M. Carvalho, and J. S. Murray. Model Interpretation Through Lower-Dimensional Posterior Summarization. *Journal of Computational and Graphical Statistics*, 30(1):144–161, Jan. 2021. ISSN 1061-8600, 1537-2715. doi: 10.1080/10618600.2020.1796684. URL https://www.tandfonline.com/doi/full/10.1080/10618600.2020.1796684.

M. C. Wu, P. Kraft, M. P. Epstein, D. M. Taylor, S. J. Chanock, D. J. Hunter, and X. Lin. Powerful SNP-set analysis for case-control genome-wide association studies. *American Journal of Human Genetics*, 86(6):929–942, 2010.

B. Yalcin, J. Nicod, A. Bhomra, S. Davidson, J. Cleak, L. Farinelli, M. Østerås, A. Whitley, W. Yuan, X. Gan, M. Goodson, P. Klenerman, A. Satpathy, D. Mathis, C. Benoist, D. J. Adams, R. Mott, and J. Flint. Commercially Available Outbred Mice for Genome-Wide Association Studies. *PLoS Genetics*, 6(9):e1001085, Sept. 2010. ISSN 1553-7404. doi: 10.1371/journal.pgen.1001085. URL https://dx.plos.org/10.1371/journal.pgen.1001085.

B. S. Yandell, T. Mehta, S. Banerjee, D. Shriner, R. Venkataraman, J. Y. Moon, W. W. Neely, H. Wu, R. von Smith, and N. Yi. R/qtlbim: QTL with Bayesian interval mapping in experimental crosses. *Bioinformatics*, 23(5):641–643, 2007. doi: 10.1093/bioinformatics/btm011. URL `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4995770/`.

P. Zeng and X. Zhou. Non-parametric genetic prediction of complex traits with latent dirichlet process regression models. *Nature Communications*, 8(1), Sept. 2017. doi: 10.1038/s41467-017-00470-2. URL `https://doi.org/10.1038/s41467-017-00470-2`.

Z. Zhang, G. Dai, and M. I. Jordan. Bayesian generalized kernel mixed models. *J Mach Learn Res*, 12: 111–139, 2011.

X. Zhou and M. Stephens. Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics*, 44(7):821–824, June 2012. doi: 10.1038/ng.2310. URL `https://doi.org/10.1038/ng.2310`.

X. Zhu and M. Stephens. Large-scale genome-wide enrichment analyses identify new trait-associated genes and pathways across 31 human phenotypes. *Nature Communications*, 9(1):4361, 2018.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320, 2005.

Supplementary Text

Scalable Computation for GOALS in Linear Regression

In this section, we show that the "GlObal And Local Score" (GOALS) operator can also be efficiently computed in a linear regression framework. As was done in the main text, consider data from a GWA study with N individuals. We have an N-dimensional vector of quantitative traits $\mathbf{y}$ and an $N \times J$ genotype matrix $\mathbf{X}$ with J denoting the number of single nucleotide polymorphisms (SNPs) encoded as $\{0, 1, 2\}$ copies of a reference allele at each locus. Next, consider a standard linear model

$$\mathbf{y} = \boldsymbol{f} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{f} = \mathbf{X}\boldsymbol{\beta}, \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \tau^2\mathbf{I}) \tag{S1}$$

where the function to be estimated $\boldsymbol{f}$ is assumed to be a linear combination of SNPs in $\mathbf{X}$ and their respective effects denoted by the J-dimensional vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_J)$ additive coefficients, $\boldsymbol{\varepsilon}$ is a normally distributed error term with mean zero and scaled variance term τ^2, and $\mathbf{I}$ denotes an $N \times N$ identity matrix. For convenience, we will assume that the trait of interest has been mean-centered and standardized.

The key identity in this section is that we can equivalently represent the regression in Eq. (S1) as a Gaussian process model with a linear gram kernel where the covariance matrix is written as $\mathbf{K} = \mathbf{X}\mathbf{X}^\intercal$. As in the main text, we can define an N-dimensional vector $\mathbf{g}^{(j)} = \mathbb{E}[\mathbf{y} \,|\, \mathbf{X} + \boldsymbol{\Xi}^{(j)}]$ where $\boldsymbol{\Xi}^{(j)}$ is an $N \times J$ matrix of all zeros except for the j-th column which we set to be a vector of some positive constant ξ. Once again, if we think about the interpretation of a regression coefficient in a linear model as detailing the expected change in the mean response given a ξ-unit increase in the corresponding covariate (holding all else constant), then a natural quantity to understand the importance of each variable is to examine $\boldsymbol{\delta}^{(j)} = \boldsymbol{f} - \mathbf{g}^{(j)}$. We showed that the posterior mean of $\boldsymbol{\delta}^{(j)}$ takes on the form

$$\mathbb{E}\left[\boldsymbol{\delta}^{(j)} \,|\, \mathbf{y}\right] = \left(\mathbf{K} - \mathbf{B}^{(j)}\right)\mathbf{A}^{-1}\mathbf{y} \tag{S2}$$

where, in addition to previous notation, $\mathbf{A} = \mathbf{K} + \sigma^2\mathbf{I}$ is the marginal variance of the response vector $\mathbf{y}$; $\mathbf{K} = \mathbf{X}\mathbf{X}^\intercal$ is the variance of $\boldsymbol{f}$ using the original genotype matrix $\mathbf{X}$; and $\mathbf{B}^{(j)}$ is the covariance between $\boldsymbol{f}$ and $\mathbf{g}^{(j)}$ using the original matrix $\mathbf{X}$ and the perturbed matrix $\mathbf{X} + \boldsymbol{\Xi}^{(j)}$. Since we are working within the context of linear regression, the covariance between $\boldsymbol{f}$ and $\mathbf{g}^{(j)}$ simplifies to the following

$$\mathbf{B}^{(j)} = k(\mathbf{X}, \mathbf{X} + \boldsymbol{\Xi}^{(j)}) = \mathbf{X}(\mathbf{X} + \boldsymbol{\Xi}^{(j)})^\intercal = \mathbf{K} + \mathbf{X}\boldsymbol{\Xi}^{(j)\intercal}. \tag{S3}$$

Note that, because $\boldsymbol{\Xi}^{(j)\intercal}$ is a matrix of all zeros except for the j-th column, we can use Eq. (S3) to simplify Eq. (S2) as the following

$$\mathbb{E}\left[\boldsymbol{\delta}^{(j)} \,|\, \mathbf{y}\right] = -\mathbf{X}\boldsymbol{\Xi}^{(j)\intercal}\mathbf{A}^{-1}\mathbf{y} = \xi\mathbf{x}_{\bullet j}\mathbf{1}^\intercal\mathbf{A}^{-1}\mathbf{y} \tag{S4}$$

where $\mathbf{1}$ is an N-dimensional vector of ones and $\mathbf{x}_{\bullet j}$ is the j-th column in the design matrix $\mathbf{X}$. The main summary is that the computation of Eq. (S2) only relies on linear operations after an initial pre-computation of the term $\mathbf{1}^\intercal\mathbf{A}^{-1}\mathbf{y}$ which can be sped up using matrix decompositions.

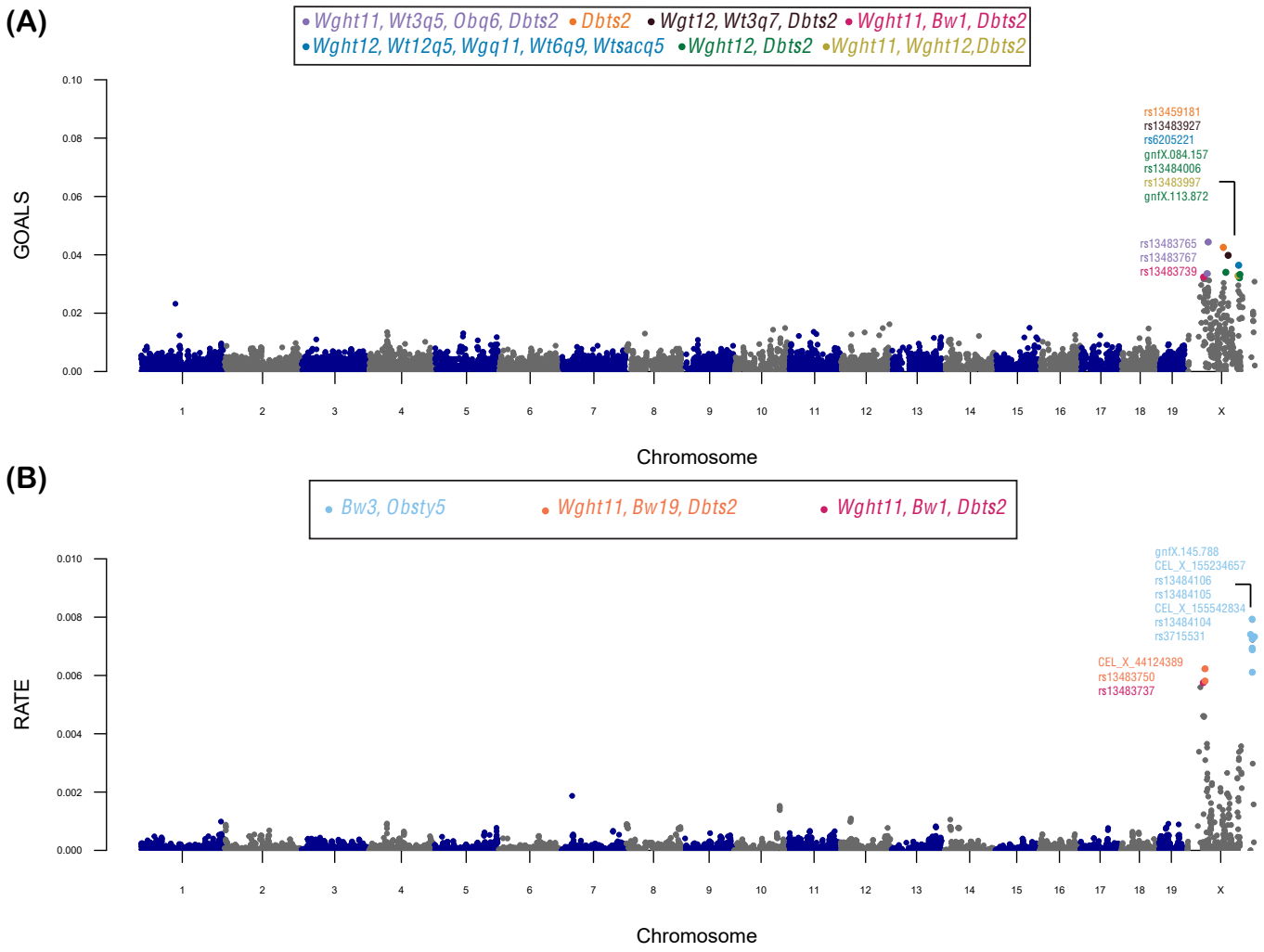

Figure S1. Manhattan plot of variant-level association mapping results for body weight in the heterogenous stock of mice dataset from the Wellcome Trust Centre of Human Genetics (Valdar et al., 2006a,b). Panel **(A)** depicts the global GOALS measure of quality-control-positive SNPs plotted against their genomic positions after running a Bayesian Gaussian process (GP) regression on the quantitative trait. As a direct comparison, in panel **(B)**, we also include results after implementing RATE on the same fitted GP model. In this figure, chromosomes are shown in alternating colors for clarity. The top 10 highest ranked SNPs by GOALS and RATE, respectively, are labeled and color coded based on their nearest mapped gene(s) as cited by the Mouse Genome Informatics database (`http://www.informatics.jax.org/`) (Bult et al., 2019). These annotated genes are listed in the legends of each panel. A complete list of the GOALS and RATE values for all SNPs can be found in Table S2.

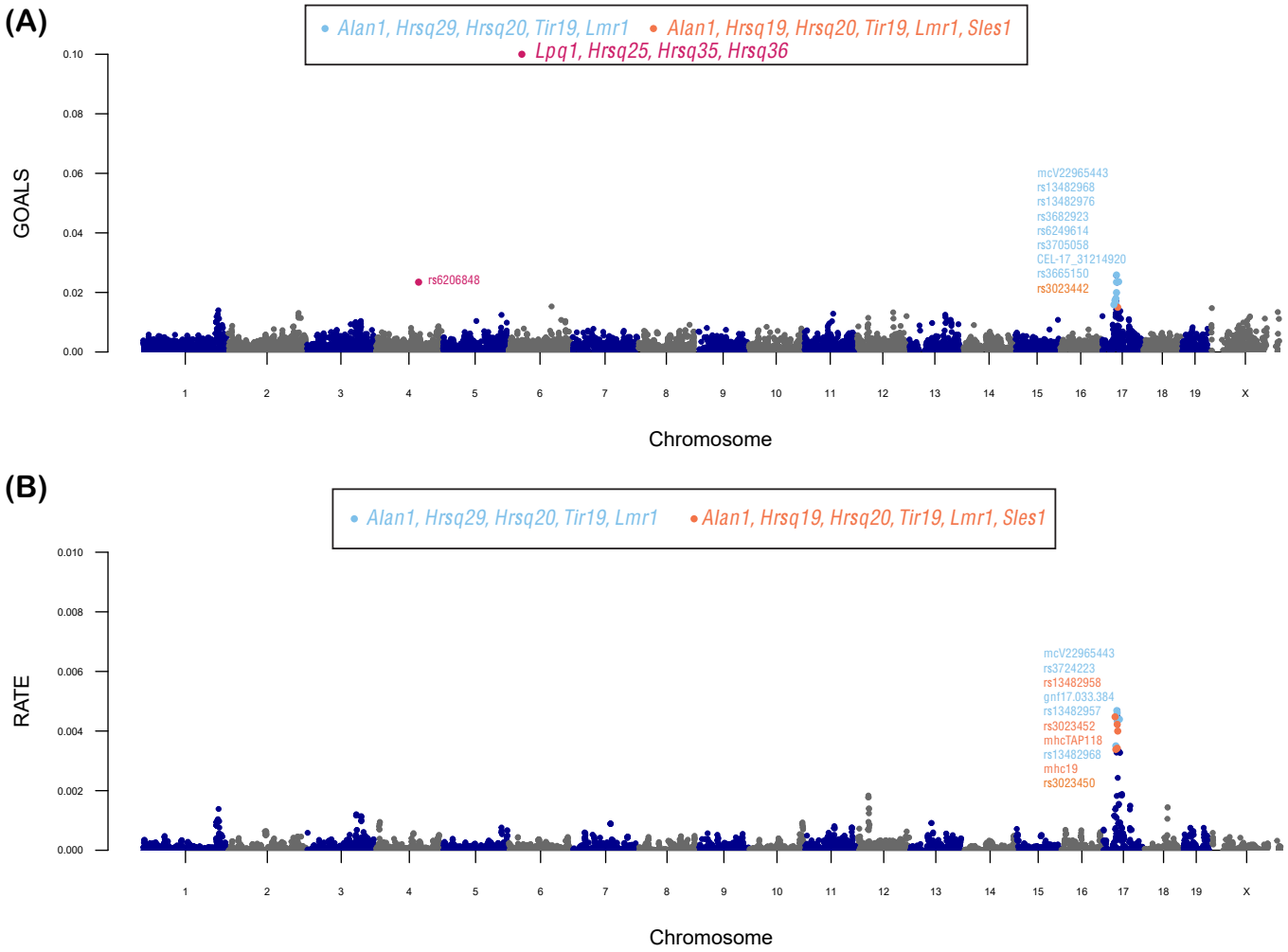

Figure S2. Manhattan plot of variant-level association mapping results for the percentage of CD8+ cell in the heterogenous stock of mice dataset from the Wellcome Trust Centre of Human Genetics (Valdar et al., 2006a,b). Panel **(A)** depicts the global GOALS measure of quality-control-positive SNPs plotted against their genomic positions after running a Bayesian Gaussian process (GP) regression on the quantitative trait. As a direct comparison, in panel **(B)**, we also include results after implementing RATE on the same fitted GP model. In this figure, chromosomes are shown in alternating colors for clarity. The top 10 highest ranked SNPs by GOALS and RATE, respectively, are labeled and color coded based on their nearest mapped gene(s) as cited by the Mouse Genome Informatics database (`http://www.informatics.jax.org/`) (Bult et al., 2019). These annotated genes are listed in the legends of each panel. A complete list of the GOALS and RATE values for all SNPs can be found in Table S3.

Supplementary Tables

Table S1. Genome-wide results for all SNPs in the heterogenous stock of mice dataset while analyzing high-density lipoprotein (HDL). Listed are the RATE and GOALS values for each SNP as computed via Gaussian Processes and the effect size analog. Also listed are the chromosome location and physical position (bp) for each SNP. (XLSX)

Table S2. Genome-wide results for all SNPs in the heterogenous stock of mice dataset while analyzing body weight. Listed are the RATE and GOALS values for each SNP as computed via Gaussian Processes and the effect size analog. Also listed are the chromosome location and physical position (bp) for each SNP. (XLSX)

Table S3. Genome-wide results for all SNPs in the heterogenous stock of mice dataset while analyzing percentage of CD8+ cells. Listed are the RATE and GOALS values for each SNP as computed via Gaussian Processes and the effect size analog. Also listed are the chromosome location and physical position (bp) for each SNP. (XLSX)