Computational Statistics and Data Analysis 194 (2024) 107914

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

A simple approach for local and global variable importance in nonlinear regression models

Emily T. Winn-Nuñez [a,*], Maryclare Griffin [b], Lorin Crawford [c,d,e,*]

[a] Division of Applied Mathematics, Brown University, Providence, RI, USA
[b] Department of Mathematics and Statistics, University of Massachusetts Amherst, Amherst, MA, USA
[c] Microsoft Research New England, Cambridge, MA, USA
[d] Department of Biostatistics, Brown University, Providence, RI, USA
[e] Center for Computational Molecular Biology, Brown University, Providence, RI, USA

ARTICLE INFO

Keywords:
Interpretability
Gaussian processes
Machine learning
Variable selection

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 data set. In this work, a new operator is proposed called the "GlObal And Local Score" (GOALS): a simple *post hoc* approach to simultaneously assess local and global feature variable importance in nonlinear models. Motivated by problems in biomedicine, the approach is demonstrated using Gaussian process regression where the task of understanding how genetic markers are associated with disease progression both within individuals and across populations is of high interest. Detailed simulations and real data analyses illustrate the flexible and efficient utility of GOALS over state-of-the-art variable importance strategies.

1. Introduction

Over the past decade, "interpretability" has become a major focus in statistical and probabilistic machine learning. 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. Part of the utility of linear models is their ability to provide well-calibrated 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 underpowered and infeasible to implement in practice. The strict additive assumptions underlying linear regression can be a hindrance 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 (e.g., Lin et al., 2022; Lippert et al., 2011; Runcie and

* Corresponding authors.
E-mail addresses: emily_winn@brown.edu (E.T. Winn-Nuñez), lcrawford@microsoft.com (L. Crawford).

https://doi.org/10.1016/j.csda.2023.107914
Received 14 February 2023; Received in revised form 23 October 2023; Accepted 23 December 2023

Available online 22 January 2024
0167-9473/© 2024 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Crawford, 2019; Schulz et al., 2020; Trippe et al., 2021) and increasingly large combinatorial feature spaces when searching over both additive and non-additive effects (e.g., Agrawal et al., 2019; Crawford et al., 2017; Stamp et al., 2023).

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 Ramasamy, 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; Fortelny and Bock, 2020). Rather having fully connected, potentially over-parameterized architectures, these newer frameworks have partially connected architectures that are based on (*i*) annotations in the literature or (*ii*) derived from relationships between features that have been identified through real-world evidence. In the biomedical application example, each neural network node has an intrinsic interpretation because they encode some biological unit (e.g., signaling pathways, protein motif, or gene regulatory network) and each weight connecting nodes represent known relationships between the corresponding units. A key aspect of this partially connected modeling 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 "sensitivity scores" has been proposed in the literature which aim to quantify variable importance by measuring the amount predictive accuracy that is lost when a particular feature is perturbed (Lundberg and Lee, 2016). One group of methods are called "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 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. Other 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 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 data set. In many scientific applications, it would be ideal to have a measure that leads to conclusions on both scales, simultaneously. For example, in human health, it is important to understand how a gene is associated with the general progression of a disease — but, for the purpose of precision medicine, it is also important to understand how certain genes might have disproportionate effects on individuals coming from different subpopulations (e.g., Martin et al., 2019; Smith et al., 2022).

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 a general illustration of our approach, we focus on using Gaussian process regression. However, also note that this variable importance approach immediately applies to other probabilistic methodologies such as neural networks (e.g., see review in Conard et al., 2023). We assess our proposed approach in the context of statistical genetics as a way to highlight data science applications that (*i*) contain outcomes that are driven by many covarying and interacting features (e.g., pairwise interactions between genes; 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 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 heterogeneous stock of mice from Wellcome Trust Centre for Human Genetics (Valdar et al., 2006a,b).

2. 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 feature and (*iii*) a metric that determines the statistical significance of each feature 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. The main objective of global interpretability is to identify features that best explain the variation of an outcome variable within an observed population. In contrast, local interpretability aims to provide an explanation on how important a single feature is to any particular individual in the data set. The purpose of this section is to review background which allows us to demonstrate all three of these key components within the context of Bayesian Gaussian process regression for continuous outcomes; however, note that extending this

theoretical framework to other nonlinear methods (e.g., neural networks; Conard et al., 2023; Ish-Horowicz et al., 2019), as well as to categorical outcomes (e.g., binary class labels 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 various statistical applications.

2.1. Weight-space Gaussian process regression

Consider a data set where $\mathbf{y}$ is a continuous response vector of length N and $\mathbf{X}$ is an $N \times J$ design matrix with N observations and J covariates. To build intuition, we begin by specifying a standard linear regression model to analyze the outcome variable such that

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

where the function to be estimated $\boldsymbol{f}$ is assumed to be a linear combination of the features in $\mathbf{X}$ and their respective effects denoted by the J-dimensional vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_J)$ of additive coefficients, $\boldsymbol{\varepsilon}$ is a normally distributed error term with mean zero and scaled variance term σ^2, and $\mathbf{I}$ denotes an $N \times N$ identity matrix. For convenience, we will assume that the response variable has been centered and standardized to have mean zero and standard deviation equal to one. A nonzero mean and non-unit variance of the response could be incorporated by including and estimating an overall intercept and scale.

It has been well documented that linear models can be underpowered when the variation of the outcome is driven by non-additive effects (e.g., Cheng et al., 2019; Pérez-Cruz et al., 2013; Yoshikawa et al., 2015). For example, in genetics applications, nonlinear models have been shown to outperform linear regression in the presence of gene-by-gene interactions (Jiang et al., 2019; McCaw et al., 2022; Weissbrod et al., 2016; Zhou et al., 2022). In these cases, the assumption in Eq. (1) that the variation in the response $\mathbf{y}$ can be fully explained by additive 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 general nonparametric approach and conduct inference in a reproducing kernel Hilbert space (RKHS) by specifying a Gaussian process (GP) prior over the data such that

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

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 $k_\theta(\bullet, \bullet)$. In practice, we assume that our model is only evaluated on the N observations in our data. When conditioning on these finite samples (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 regression model

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

Here, $\boldsymbol{f} = [f(\mathbf{x}_1), \ldots, f(\mathbf{x}_N)]$ is a normally distributed random variable with mean vector $\mathbf{0}$, and the covariance matrix $\mathbf{K}$ is computed with each element given by $k_{ii'} = k_\theta(\mathbf{x}_i, \mathbf{x}_{i'})$ where $\mathbf{x}_i$ and $\mathbf{x}_{i'}$ denote the features for the i-th and i'-th observation, 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; Crawford et al., 2018; Demetci et al., 2021; Murdoch et al., 2019; Tsang et al., 2018a,b; Wahba, 1990). For the demonstrations in the main text of this paper, we will consider $k_{ii'}$ to be a nonlinear shift-invariant function.

Altogether, the "weight-space" GP regression model in Eq. (3) can be seen as a generalization of the linear model in Eq. (1) which uses a nonlinear covariance $\mathbf{K}$ to account for non-additive interactions between features instead of the usual (additive) gram matrix $\mathbf{X}\mathbf{X}^\intercal/J$ (e.g., Lippert et al., 2011; Zhou and Stephens, 2012). Lastly, the GP model can also be easily extended to accommodate fixed effects that are specific to the observations being studied (e.g., age, socioeconomic status) (de los Campos et al., 2009; Shi et al., 2012). We will not explicitly consider the inclusion of fixed effects here and, instead, will leave those explorations to the reader.

2.2. 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 many statistical applications is to jointly infer the true effect size and statistical significance of each feature that is put into the model. One classic strategy for estimating the regression coefficients in the linear model presented 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 of the design matrix. Under the generative model in Eq. (3) for the response variable $\mathbf{y}$, $\widehat{\boldsymbol{\beta}} = [\widehat{\beta}_1, \ldots, \widehat{\beta}_J]$ is a random variable with elements that can be interpreted as the (additive) effect size for each feature in the data set.

The effect size analog was developed with the intention of being the nonparametric version of a regression coefficient for each feature of a nonlinear model (Crawford et al., 2018). In general, this leverages the idea that $\mathbb{E}[\mathbf{y} \mid \mathbf{X}] = \mathbb{E}[\boldsymbol{f} \mid \mathbf{y}]$ 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 relationships between the features and the response that the nonlinear model has learned. Under the linear projection, the effect size analogs in Eq. (4) 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 feature in our data by taking the KLD between (*i*) the conditional distribution $p(\widetilde{\boldsymbol{\beta}}_{-j} \mid \widetilde{\beta}_j = 0)$ with the effect of that feature being set to zero and (*ii*) the marginal distribution $p(\widetilde{\boldsymbol{\beta}}_{-j})$ with the effect of that feature having been marginalized over. This is defined by solving the following

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

for each $j = 1, \ldots, J$ features 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 feature has no impact on explaining the modeled outcome or response (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 importance for each feature (where values close to 1 suggest greater importance). From a classical hypothesis testing point-of-view, the null under RATE measure assumes that each feature is equally associated with the outcome, while the alternative assumes proposes that some features are much more important than others. Formally, this can be 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 features in the data have the same relative variable importance.

2.3. Limitations of the current distributional centrality framework

There are several notable shortcomings with effect size analog and RATE framework. First, calculating both the effect size analog and the KLD for each feature in turn 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 features J grows, these calculations become infeasible. Second, the significance threshold $1/J \to 0$ as $J \to \infty$, which effectively means that all variables will be considered important for high-dimensional settings. Third, while this framework summarizes the global association for each feature within the observed population, it lacks the ability to locally explain how important variables are to each individual observation in the data. This limits its potential impact, for example, 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 feature (Kowal, 2021; Smith et al., 2023; 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 feature 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 feature. Each of these issues serve as motivation to develop an alternative and more unified framework for nonlinear models.

3. 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 features that are significantly associated with a response variable across a population as well as explain marginal feature effects on an individual level. Again, let $\boldsymbol{f}$ be a function that is estimated from a nonlinear model (e.g., a Gaussian process) 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 define perturbed features $\mathbf{X}+\boldsymbol{\Xi}^{(j)}$, where $\boldsymbol{\Xi}^{(j)}$ is an $N \times J$ matrix with rows $\boldsymbol{\xi}^{(j)}$ equal to all zeros except for the j-th element which we set to be a vector of some positive constant ξ. We then define a random variable $\mathbf{g}^{(j)} = [f(\mathbf{x}_1 + \boldsymbol{\xi}^{(j)}), \dots, f(\mathbf{x}_N + \boldsymbol{\xi}^{(j)})]$. 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 study the difference

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

Here, each element of the length N vector $\boldsymbol{\delta}^{(j)} = (\delta_1^{(j)}, \dots, \delta_N^{(j)})$ reflects the importance of the j-th variable for the model fit with respect to each sample. In other words, it quantifies local variable importance by measuring how much the response changes when a particular feature is perturbed (i.e., similar to the objective of other sensitivity score-based methods). 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, under the null hypothesis of there being no relationship between the j-th covariate and the response, elements of $\boldsymbol{\delta}^{(j)}$ will be concentrated around zero if the j-th covariate generally has no effect on the response variable that is being analyzed. Under the alternative hypothesis of there being some relationship between the j-th covariate and the response, significantly associated variables under the alternative have $\boldsymbol{\delta}^{(j)}$ 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.

3.1. Probabilistic properties of GOALS

Although GOALS can be applied to any probabilistic model for the mean response for arbitrary features, we will demonstrate its properties using a weight-space Gaussian process regression model (e.g., similar to what is 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 data $\mathbf{X}$ and $\mathbf{y}$, respectively. The joint distribution between length N vectors $\mathbf{y}$, $\boldsymbol{f}$, and $\{\mathbf{g}^{(j)}\}_{j=1}^J$ can be specified via the following normal distribution

$$\begin{bmatrix} \mathbf{y} \\ \boldsymbol{f} \\ \mathbf{g}^{(1)} \\ \vdots \\ \mathbf{g}^{(J)} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{A} & \mathbf{K} & \mathbf{B}^{(1)} & \cdots & \mathbf{B}^{(J)} \\ \mathbf{K} & \mathbf{K} & \mathbf{B}^{(1)} & \cdots & \mathbf{B}^{(J)} \\ \left(\mathbf{B}^{(1)}\right)^\intercal & \left(\mathbf{B}^{(1)}\right)^\intercal & \mathbf{C}^{(1)} & \cdots & \mathbf{D}^{(1,J)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \left(\mathbf{B}^{(J)}\right)^\intercal & \left(\mathbf{B}^{(J)}\right)^\intercal & \mathbf{D}^{(J,1)} & \cdots & \mathbf{C}^{(J)} \end{bmatrix} \right), \tag{9}$$

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 design 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 shift-invariant, $\mathbf{C}^{(j)} = \mathbf{K}$ 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

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

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

$$\begin{bmatrix} \boldsymbol{\delta}^{(1)} \\ \vdots \\ \boldsymbol{\delta}^{(J)} \end{bmatrix} \Bigg| \mathbf{y} \sim \mathcal{N}\left(\begin{bmatrix} \left[\mathbf{K}-\left(\mathbf{B}^{(1)}\right)^\intercal\right]\mathbf{A}^{-1}\mathbf{y} \\ \vdots \\ \left[\mathbf{K}-\left(\mathbf{B}^{(J)}\right)^\intercal\right]\mathbf{A}^{-1}\mathbf{y} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}^{(1)} & \cdots & \boldsymbol{\Sigma}^{(1,J)} \\ \vdots & \ddots & \vdots \\ \boldsymbol{\Sigma}^{(J,1)} & \cdots & \boldsymbol{\Sigma}^{(J)} \end{bmatrix} \right), \tag{10}$$

where

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

Theoretically, this results in a joint conditional distribution from which to estimate the posterior distribution of each $\boldsymbol{\delta}^{(j)}$ and obtain local interpretability. However, in many current data science applications, where data sets can include hundreds of thousands of observations that have been collected with millions of features, it is often desirable to use a more scalable computation than sampling

estimates from a full joint distribution. To that end, in this work, we will consider the posterior mean in Eq. (10) 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

$$\widehat{\boldsymbol{\delta}}^{(j)} = \left[\mathbf{K} - \left(\mathbf{B}^{(j)}\right)^{\intercal}\right]\mathbf{A}^{-1}\mathbf{y}, \qquad \bar{\delta}^{(j)} = \sum_{i} \widehat{\delta}_i^{(j)}/N. \tag{11}$$

Derivations of the full joint distribution for the global importance scores $[\bar{\delta}^{(1)}, \ldots, \bar{\delta}^{(J)}]$, as well as an outline of how to extend GOALS to perform variable importance in probabilistic neural networks, can be found in the Supplementary Material.

3.2. 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 Eqs. (10) and (11). For demonstration, we will assume a Gaussian process with a radial basis covariance function $k_{ii'} = \exp\{-\theta\|\mathbf{x}_i - \mathbf{x}_{i'}\|^2\}$ 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). First, it is important to note that the only matrix that needs to be recomputed for each feature 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 feature by making the following rank one updates

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

where, similar to previous notation, $\mathbf{x}_i$ and $\mathbf{x}_{i'}$ are the i-th and i'-th rows of the design 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

$$\mathbf{B}^{(j)} = \mathbf{K} \circ \exp\left\{-\theta\left[\xi^2\mathbf{1}\mathbf{1}^{\intercal} - 2\xi\left(\mathbf{x}_{\bullet j}\mathbf{1}^{\intercal} - \mathbf{1}\mathbf{x}_{\bullet j}^{\intercal}\right)\right]\right\}, \tag{12}$$

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 twofold. First, the magnitude of the GOALS operator computed using Eq. (11) will depend on the value of ξ. This relationship is most easily seen when the covariance function is linear (see the Supplementary Material). The second takeaway 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 (again see Supplementary Material).

3.3. 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 applications, this is done by considering all possible subsets of variables that do not include the j-th feature $\mathcal{S} \subseteq \mathcal{J}\backslash\{j\}$ and then comparing their performance to the performance of a model trained on the same subset as well as the j-th feature $\mathcal{S} \cup \{j\}$. This weighted average can be represented as the following formula

$$\phi_j = \sum_{\mathcal{S}\subseteq\mathcal{J}\backslash\{j\}}\left[\frac{|\mathcal{S}|!(J - |\mathcal{S}| - 1)!}{J!}\right]\left(\boldsymbol{f}_{\mathcal{S}\cup\{j\}} - \boldsymbol{f}_{\mathcal{S}}\right), \tag{13}$$

where $|\mathcal{S}|$ is number of features in the subset $\mathcal{S}$ and $|\mathcal{J}| = J$ is the total number of features 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 feature added to the subset $\mathcal{S}$, respectively.

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

$$\boldsymbol{\xi}^{(j)} = \xi\left[\mathbb{1}\{j = 1\} \quad \cdots \quad \mathbb{1}\{j = J\},\right] \tag{14}$$

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

$$\delta^{(j)} = f - g^{(j)} = f_{\mathcal{J}} - f_{\mathcal{J}\cup\mathcal{J}'}. \tag{15}$$

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 $J-1$ features, excluding the j-th feature, whereas SHAP considers every possible subsets of variables that do not include the j-th feature (which can be computationally intensive for large data sets).

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. (13) favors both the smallest and largest subsets of $\mathcal{J}$ and penalizes subsets S of the size $|S| \approx J/2$. This means that if the j-th feature has an effect on the response via marginal effects, then both GOALS and SHAP are likely to give that feature a high ranking. If the j-th feature is only influential on a response through a moderate number of interactions (i.e., within sets of size $J/2$), then GOALS may rank that feature higher (relative to other features) than SHAP will. However, on the other hand, if the j-th feature is influential through pairwise interactions with nearly all other features in the data set, then SHAP may provide a higher relative rank for that feature than GOALS — although, this scenario is probably least likely to happen in practice. Furthermore, SHAP may rank variables that are highly correlated with each other lower than GOALS — this is because the difference in model fits $\boldsymbol{f}_{S\cup\{j\}} - \boldsymbol{f}_S$ may be small when feature j is highly correlated with features in S. We show that these expectations are supported empirically in the next section.

4. 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 heterogeneous stock of mice from Wellcome Trust Centre for Human Genetics (Valdar et al., 2006a,b).

4.1. Simulation studies

The general design of the following simulation studies has been previously used to explore the power of variable importance methods (Crawford et al., 2018, 2019; Demetci et al., 2021; Smith et al., 2023). Once again, let $\mathbf{X}$ be a design matrix of N observations with J features. To generate synthetic data, we select a subset of causal features from the design matrix and then use the following linear model

$$\mathbf{y} = \sum_{c\in C} \mathbf{x}_c\beta_c + \mathbf{W}\boldsymbol{\tau} + \mathbf{Z}\boldsymbol{\omega} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}), \tag{16}$$

where $\mathbf{y}$ is a synthetic response vector of length N; C represents the set of all randomly selected causal features; $\mathbf{x}_c$ is the c-th causal feature vector with a corresponding nonzero additive effect size β_c; $\mathbf{W}$ is an $N \times M$ dimensional matrix which holds all pairwise interactions between the causal features, with the columns of this matrix assumed to be the Hadamard (element-wise) product between feature vectors of the form $\mathbf{x}_j \circ \mathbf{x}_k$ for the j-th and k-th features; $\boldsymbol{\tau}$ is the M-dimensional vector of interaction effect sizes; $\mathbf{Z}$ contains covariates representing additional population structure between the samples in the data with corresponding effects $\boldsymbol{\omega}$; and $\boldsymbol{\varepsilon}$ is an length N vector of environmental noise. For simplicity, we will consider $\mathbf{Z}$ to be the top ten principal components (PCs) from the design matrix $\mathbf{X}$. In these simulations, we assume that the total variation of the synthetic response variable is $\mathbb{V}[\mathbf{y}] = 1$. We allow the additive and interaction effect sizes to be randomly drawn from standard normal distributions, $\boldsymbol{\beta}_c \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\boldsymbol{\tau} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Next, we scale the additive, pairwise interactions, population structure, and the environmental noise terms so that they collectively explain a fixed proportion of the total variance where

$$\mathbb{V}\left[\sum_{c\in C} \mathbf{x}_c\beta_c\right] = \rho v^2, \qquad \mathbb{V}[\mathbf{W}\boldsymbol{\tau}] = (1-\rho)v^2, \qquad \mathbb{V}[\mathbf{Z}\boldsymbol{\omega}] + \mathbb{V}[\boldsymbol{\varepsilon}] = 1 - v^2. \tag{17}$$

Intuitively, v^2 determines how much variance in the simulated response is due to signal versus noise, while ρ is a mixture parameter which determines how much of the signal is driven by additive versus interaction effects. Below, we will consider studies where $v^2 \in \{0.3, 0.6\}$. We will also assess different cases by setting $\rho \in \{0.5, 1\}$, where the former assumes that additive and interaction effects contribute equally to the total variation in the response, and the latter assumes only additive effects contribute to the signal.

Proof-of-concept simulations: low-dimensional analysis In this subsection, we provide a low-dimensional proof-of-concept simulation study (i.e., $N > J$). To accomplish this, we generate synthetic data $\mathbf{X}$ with $N = 2000$ observations and $J = 25$ covariates where each feature is drawn from a standard normal distribution. In these simulations, we generate synthetic outcome variables using Eq. (16) by fixing the signal-to-noise ratio to be $\upsilon^2 = 0.6$ and omitting population structure effects by setting $\boldsymbol{\omega} = \mathbf{0}$. Here, we assume some subset of the features $C = \{8, 9, 10, 23, 24, 25\}$ to be causal. We then consider five different simulation scenarios:

- **Scenario I (Additive and Interaction Effects):** The subset $\{23, 24, 25\} \subseteq C$ are causal features, where all three have additive effects and features #23 and #24 interact with #25, respectively.
- **Scenario II (Additive and Interactions Effects from Different Groups):** All features in the set C are causal. Features #8-10 only have interaction effects and features #23-25 only have additive effects. Specifically, features #8 and #9 each interact with #10, separately.
- **Scenario III (Overlapping Additive and Interaction Effects):** All features in the set C are causal. Features #23-25 each have additive effects; while, feature #8 interacts with #10 and #9 interacts with #25, respectively.
- **Scenario IV (Interaction Effects Only):** All features in the set C are causal only through interaction effects. Features #8 and #9 each interact with #10, separately; while, features #23 and #24 each interact with #25, separately.
- **Scenario V (Noise Only):** None of the features in the data have an association with the response. 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 in real applications. 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.

Fig. 1 contains the global variable importance results for GOALS using perturbation parameter $\xi = 1$ and RATE on Scenarios I-V 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. (11). In Fig. 1, the known causal features 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. For the alternative simulation Scenarios I-IV, any causal variables with importance scores above the significance thresholds $1/J$ and 0 are considered to be true positives for RATE and GOALS — all other variables with importance scores above these thresholds are false discoveries. In the null simulation Scenario V, all variables should appear below the respective significance thresholds for both methods. Overall, we see that both methods perform similarly in identifying causal features that have both additive and interaction effects on the response (Fig. 1A). However, GOALS proves to be a better discriminator between causal and non-causal features than RATE when interaction effects occur in isolation (i.e., covariates are involved in 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 Fig. 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 features only have non-additive effects such as variables #8 and #9). Lastly, GOALS is better calibrated when data are generated from complete noise (i.e., when there are no true associations between features $\mathbf{X}$ and outcome $\mathbf{y}$) (Fig. 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 features 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 biomedical applications, this can yield key insight in the event that a gene 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 $\{22, 23, 24, 25\} \subseteq C$ are causal features. Features #23-25 have additive and interaction effects that are associated with all individuals; while, feature #22 has an additive effect for only half of the population.

Fig. 2 shows the distribution of the local individual-level GOALS operator for variables #8, #22, and #25 in this split scenario. As a baseline, we also show results from running a local analysis with SHAP. For clarity, variable #8 is a non-causal feature in this scenario. There are a few key takeaways in this empirical illustration. First, when a feature has a no effect on the response, the distribution of the local scores for both GOALS and SHAP are centered at 0. Conversely, features with nonzero effects on the outcome 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 feature has an effect 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 variable importance on specific observations. In Fig. 2, this characteristic is more distinct with the GOALS operator where there is clearer separation in values for variable #22 in samples where it has a nonzero effect.

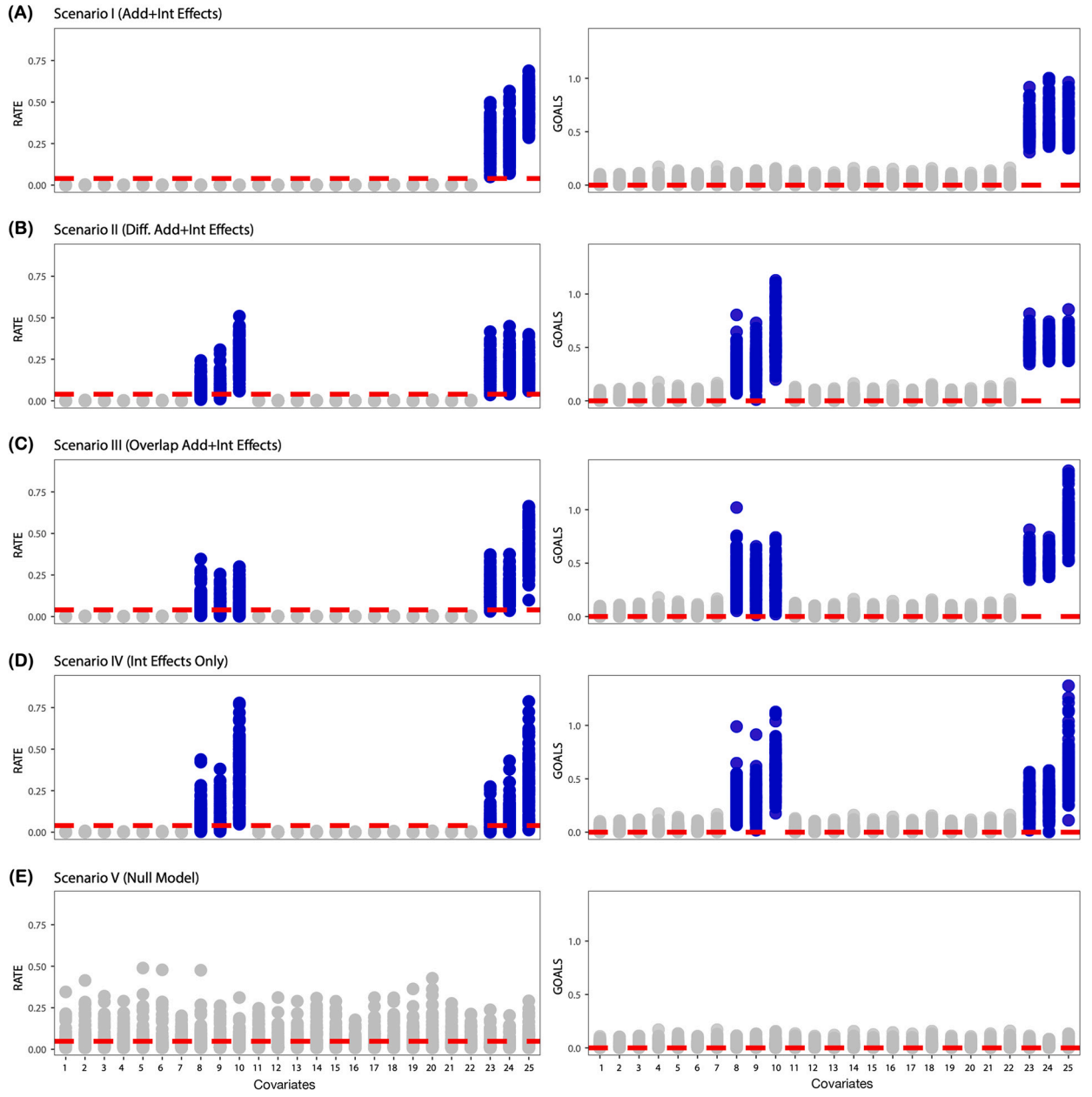

Fig. 1. Proof-of-concept simulations to demonstrate how GOALS and RATE globally prioritize important variables with varying degrees of additive and interaction effects. These simple simulations assume that synthetic responses have a signal-to-noise ratio equal to $\upsilon^2 = 0.6$ with $(1 - \rho) = 0\%$ to 50% of the signal stemming from interaction effects. Points highlighted in blue are covariates that 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 features whose associations are driven primarily by interaction effects. All results shown in this figure are based on 100 replicates. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

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 real data as our design matrix $\mathbf{X}$ to generate a synthetic outcome $\mathbf{y}$. Here, we take data from the Wellcome Trust Case Control Consortium (WTCCC) 1 study which initially consisted of 2,938 samples with 458,868 genetic features. The features in this data are known as single nucleotide polymorphisms (SNPs), each of which are originally encoded as 0, 1, 2 copies of a reference allele at each locus. We follow the same quality control procedures used in previous studies (The Wellcome Trust Case Control Consortium, 2007). Missing data were imputed by using the BIMBAM software (http://www.haplotype.org/bimbam.html; Servin and Stephens, 2007). In these simulations, we use all features

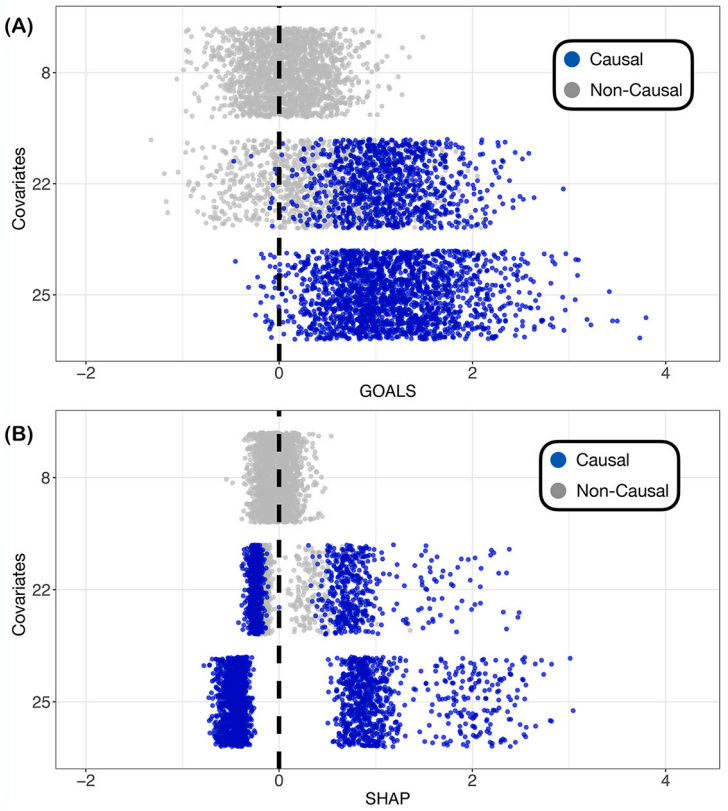

Fig. 2. Proof-of-concept simulations to demonstrate how GOALS and SHAP locally prioritize important variables that have varying level of effects on specific subsets of the population. These simple simulations assume that synthetic responses have a signal-to-noise ratio equal to $\upsilon^2 = 0.6$ with $(1-\rho) = 0\%$ to 50% of the signal stemming from interaction effects. Points highlighted in blue are covariates that have nonzero effects within each of the five different scenarios. Here, each point is an individual. We highlight the local variable importance metrics for three specific features according **(A)** GOALS and **(B)** Shapley Additive Explanations (SHAP). In this simulation study, covariate #8 is null feature and does not contribute to the phenotypic variation; covariate #25 has additive and interaction effects that are associated with all individuals; and covariate #22 has an additive effect for only half of the population. A point is blue if the corresponding labeled covariate has a nonzero effect for that individual. The main takeaway of this analysis is that, in the case where a feature has an effect on the response 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 variable importance on specific observations. The black dashed line is drawn at zero to represent a threshold where a feature has no effect for a given sample.

with minor allele frequencies (MAFs) above 1% on chromosome 22 to generate continuous outcome variables. After preprocessing, all genetic features were centered and scaled to have mean zero and standard deviation equal to one. Exclusively considering this group of individuals and SNPs resulted in a final data set consisting of $N = 2{,}938$ samples and $J = 5{,}747$ features.

During each simulation run, we randomly choose a set of $|C| = 30$ causal SNPs. Next, we set the signal-to-noise ratio $\upsilon^2 = 0.3$ and consider two choices for the contribution stemming from interactions between causal features $\rho \in \{0.5, 1\}$. We also consider simulations with and without population structure effects by allowing the top ten principal components (PCs) from the design matrix $\mathbf{X}$ to make up to 10% of the overall variation in the synthetic outcome variable $\mathbf{y}$. 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 data with only additive effects; while, scenarios III and IV consider data with both additive and interaction effects. Additionally, scenarios II and IV have the additional complexity of having nonzero effects from population structure which is not observed in scenarios I and III.

We compare the global power of the GOALS measure to a list of variable importance techniques. Specifically, these methods include: (a) the *post hoc* framework of estimating effect size analogs for the features used in a GP regression model and determining their importance using distributional centrality via RATE (Crawford et al., 2019); (b) a univariate linear model (SCANONE) (Yandell et al., 2007); (c) L1-regularized "least absolute shrinkage and selection operator" (LASSO) regression (Tibshirani, 1996); (d) the combined regularization utilized by the Elastic Net (Zou and Hastie, 2005); (e) a random forest (RF) (Ishwaran and Lu, 2019) fit with 500 trees; a gradient boosting machine (GBM) (Friedman, 2001) fit with 100 trees; and a Bayesian additive regression tree (BART) (Chipman et al., 2010) fit with 200 trees and 1000 Markov chain Monte Carlo (MCMC) iterations. Note that SCANONE produces *P*-values, and the LASSO and the Elastic Net give magnitudes of regression coefficients. The latter two regularization approaches were fit by first learning tuning parameter values via 10-fold cross validation. Additionally, features in the RF and GBM are ranked by assessing relative influence which is computed by taking the average total decrease in the residual sum of squares after tree splitting on each variable; while, in BART, features are ranked by the average number of times that they are used in decisions for each tree. Indeed, the SHAP value framework can also be used for *post hoc* assessment of global interpretability by taking the average

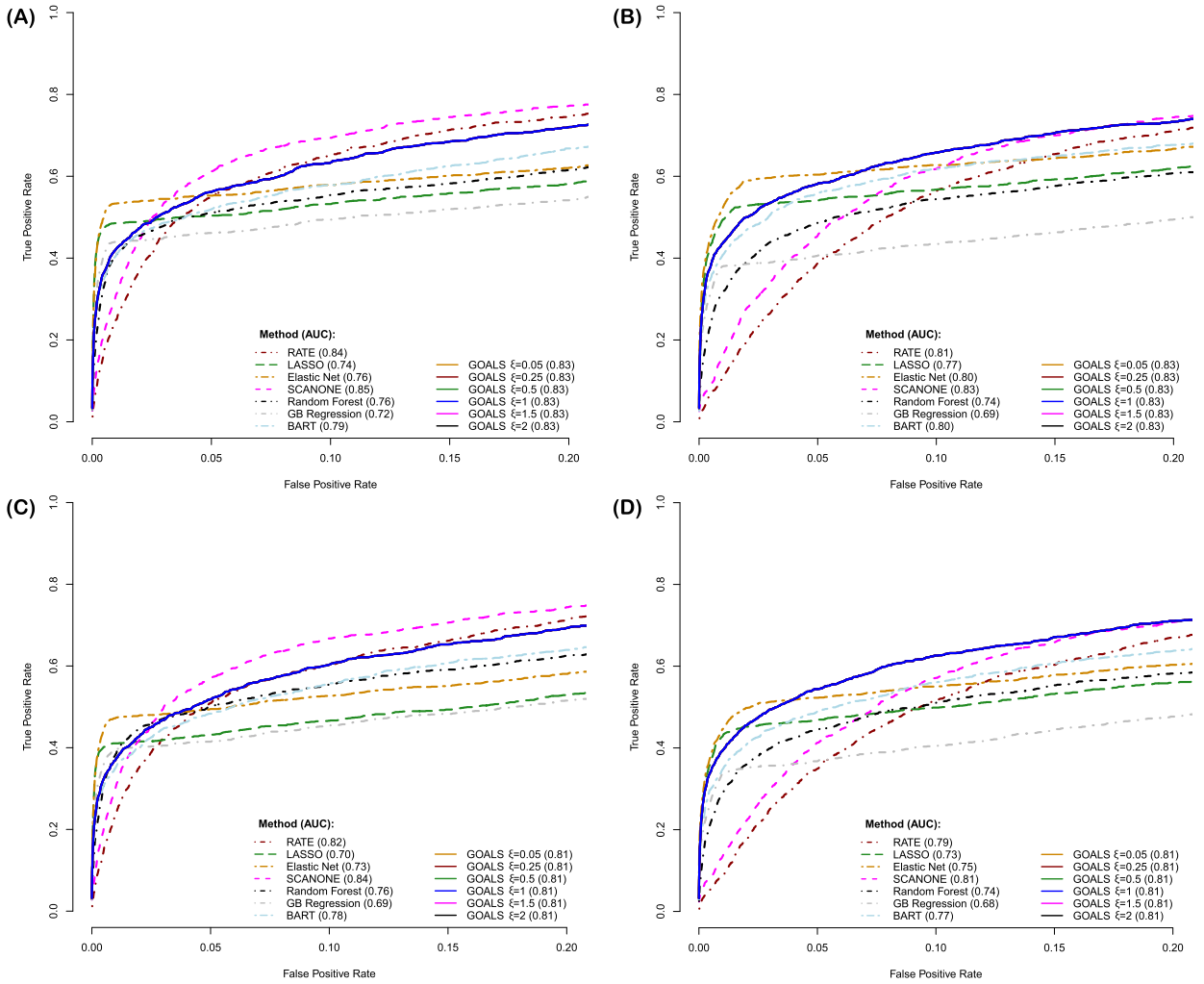

Fig. 3. Receiving operating characteristic (ROC) curves comparing the performance of GOALS against other global variable importance approaches in simulations. Here, synthetic responses are simulated to have a signal-to-noise ratio equal to $v^2 = 0.6$ with only additive effects in panels **(A)** and **(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 signal that is contributed by additivity. The response variables simulated in panels **(B)** and **(D)** also have the additional complexity of having population stratification effects. We show results using Gaussian process regression with GOALS across a wide range of values for the perturbation parameter ξ. Competing approaches include: Gaussian process regression with RATE (red), LASSO regularization (green), the Elastic Net (yellow), the SCANONE method (pink), a random forest (RF) (black), a gradient boosting machine (GBM) (grey), and a Bayesian additive regression tree (BART) (light blue). Methods using GOALS are illustrated as a solid lines, while the competing models are shown as dotted lines. Note that even though the GOALS value may be different depending on the choice of ξ (see Figure S1), the relative importance ranking that it assigns to each feature remains the same. As a result, the performance of GOALS is not sensitive to the choice of ξ in these simulations, so all solid lines fall on top of each other. Lastly, 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.

of local scores across observations for each feature in the data. However, because the SHAP approach considers all possible subsets of features 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 (see next section for its application to a subset of real data).

Each method is evaluated based on its ability to effectively prioritize causal features in 100 different simulated data sets. We consider the 30 variables in the causal set C to be true positives and all other variables to be true negatives. The criteria we use compare the false positive rate (FPR) with the rate at which true causal variables are selected first by each model (TPR). This information is depicted as receiver operating characteristic (ROC) curves in Fig. 3. Specifically, for each method, we rank features from most to least important. Starting with the top ranked variable, we then use a sliding threshold to create a set of "selected" features. During each iteration, we compute the number of true and false positives in the selected set (which we will denote as TP and FP, respectively). We then calculate the TPR and FPR as the following

$$\text{TPR} = \text{TP}/|C|, \qquad \text{FPR} = \text{FP}/(J - |C|), \tag{18}$$

where, again, $|C| = 30$ is the number of causal variables in the simulation and $J - |C|$ then represents the number of true negatives. Fig. 3 illustrates the mean ROC curve across all 100 replicates per simulation scenario, 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 — where a higher AUC points to better model performance.

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 data is generated by causal variables with only additive effects (e.g., Fig. 3A). This power generally decreases in the presence of population structure (e.g., Fig. 3B) or when causal variables are involved in pairwise interactions (e.g., Fig. 3C and 3D). GOALS outperforms LASSO, Elastic Net, RF, GBM, and BART 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 structure at lower false positive rates. 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 data set 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. Lastly, to illustrate the robustness of GOALS, we apply our method while using a range of values for $\xi = \{0.05, 0.25, 0.5, 1, 1.5, 2\}$ and show that the performance of GOALS is relatively robust to the choice made for this parameter in terms of prioritizing the correct causal features. Note however, as mentioned previously when discussing Eqs. (11) and (12), the magnitude of the GOALS operator will be affected by the value of ξ (Figure S1). 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.

4.2. Global and local association mapping in heterogeneous stock of mice

In this section, we apply GOALS to 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 $J = 10{,}227$ genetic features that were available for all mice.

We chose to analyze this particular data set 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 (i.e., signal-to-noise ratios) with nonzero contributions from both additive and non-additive genetic effects (Chen et al., 2012; Valdar et al., 2006b). As result, this data set represents a realistic mixture of the simulation scenarios we detailed in the previous sections.

For each trait, we fit a GP regression model with GOALS and RATE, a random forest (RF) (Ishwaran and Lu, 2019) fit with 500 trees, a gradient boosting machine (GBM) (Friedman, 2001) fit with 100 trees; and a Bayesian additive regression tree (BART) (Chipman et al., 2010) fit with 200 trees and 1000 Markov chain Monte Carlo (MCMC) iterations. Figs. 4 and S2-S6 display the variant-level mapping results via Manhattan plots after assessing variable importance using GOALS with $\xi = 1$, RATE, RF, GBM, and BART in HDL content, body weight, and the percentage of CD8+ cells, respectively. In these plots, larger values mean "enrichment" in a given genomic position. 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 variable importance scores for all SNPs (see Tables S1-S3 in the Supplementary Material). In general, GOALS demonstrated the ability to identify biologically relevant signal in all three traits that were missed by the other competing approaches. This was most apparent in HDL where the top 10 highest ranked SNPs by RATE, RF, GBM, and BART were primarily located within two genes on the first and X chromosomes; but, the top 10 highest ranked SNPs by GOALS included 13 relevant genes across five different chromosomes (see Figs. 4 and S2). In addition to moderate signal on first and X chromosomes, GOALS also found signal on chromosomes 3, 11, 12,15, and 17. We hypothesize that GOALS prioritizes these additional genetic variants because it measures variable importance in function space and, thus, is better positioned to identify features whose associations are driven primarily by non-additive effects. 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 all methods in the analysis of body weight and the percentage of CD8+ cells. For example, each approach identified strong signal on the X chromosome for body weight, a genomic region that was also validated by

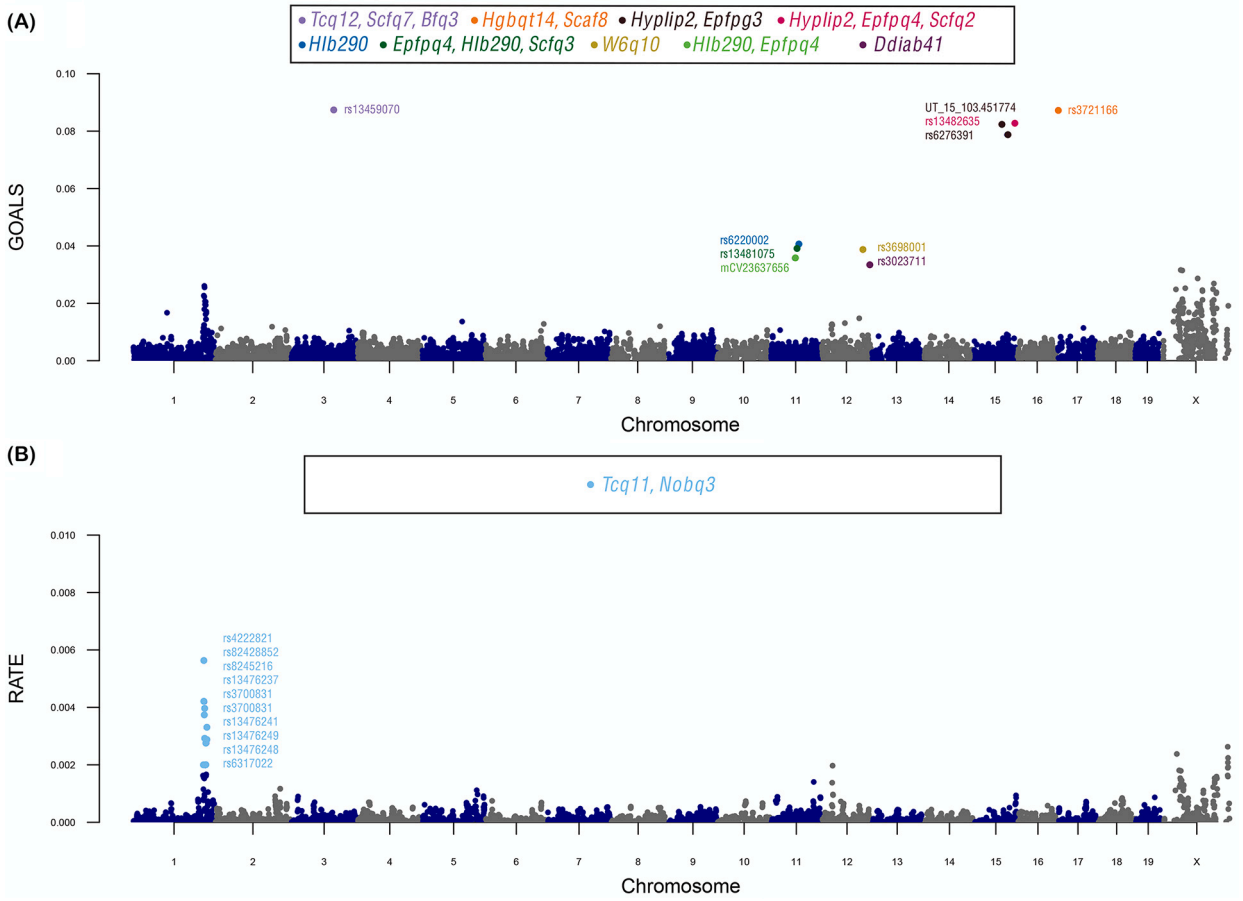

Fig. 4. Manhattan plot of variant-level association mapping results for high-density lipoprotein (HDL) content in the heterogeneous stock of mice data set from the Wellcome Trust Centre of Human Genetics (Valdar et al., 2006a,b). Panel **(A)** depicts the global GOALS measure (with $\xi = 1$) of quality-control-positive 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 comparison of these results to a random forest (RF), a gradient boosting machine (GBM), and a Bayesian additive regression tree (BART) can be found in Figure S2. A complete list of the variable importance values provided by each method for all SNPs can be found in Table S1.

Valdar et al. (2006b) in the original study (see Figures S3 and S4). Here, all methods other than the random forest 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, all 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) (see Figures S5 and S6). Overall, out of the top ten most prioritized variables ranked by *post-hoc* approaches 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 data set 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 Fig. 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. As a direct comparison, we also implement SHAP on the same set of SNPs. Note that due to computational considerations, SHAP is implemented by only considering all possible subsets of features on a given chromosome when computing local variable importance. Specifically, to run SHAP, we limit the data to include 372 SNPs on the X chromosome for body weight (runtime approximately 8 hours), 375 SNPs on chromosome 17 for percentage of CD8+ cells (runtime approximately 8 hours), and 758 SNPs on chromosome 3 for HDL content (runtime approximately 23 hours). GOALS, on the other hand, is implemented on the full genome-wide data set. Overall, our proposed measure does indeed seem to capture environmental variation. This is again most

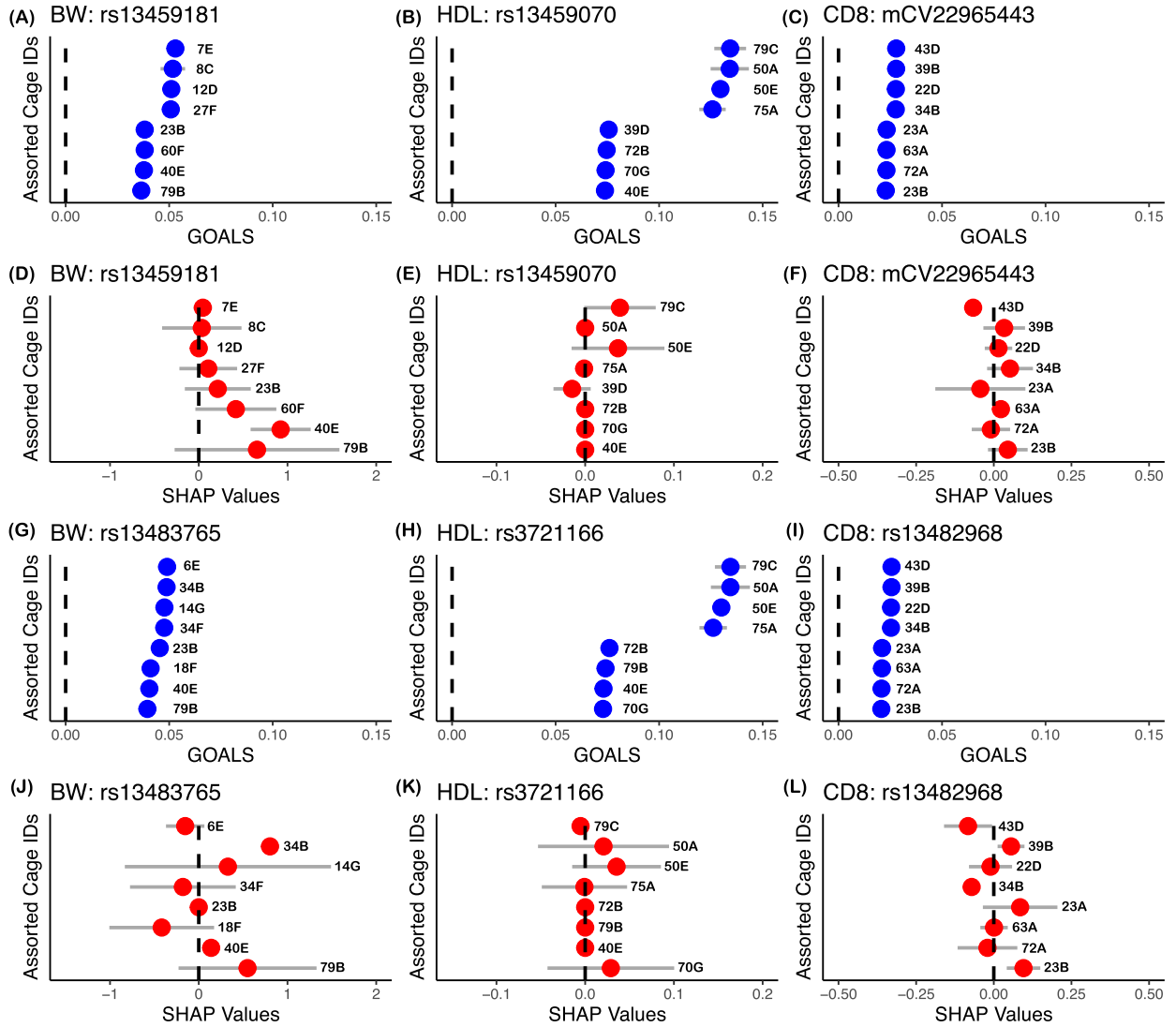

Fig. 5. Plot of local variable importance according to GOALS and SHAP as a function of cage for notable genetic variants in the analysis of the heterogeneous stock of mice data set from the Wellcome Trust Centre of Human Genetics (Valdar et al., 2006a,b). Here, we show how SNPs have varying levels of importance for individual mice depending on the cage they were assigned to in the study. The traits analyzed here include **(A, D, G, J)** body weight (BW); **(B, E, H, K)** high-density lipoprotein (HDL) content; and **(C, F, I, L)** percentage of CD8+ cells. The blue and red points are the mean local GOALS and SHAP values for each cage, respectively, and the grey lines show the total distribution for each score. 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. Note that due to computational considerations, SHAP is implemented by only considering all possible subsets of features on a given chromosome when computing local variable importance. Specifically, to run SHAP, we limit the data to include 372 SNPs on the X chromosome for body weight, 375 SNPs on chromosome 17 for percentage of CD8+ cells, and 758 SNPs on chromosome 3 for HDL content. GOALS is implemented on the full genome-wide data set.

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., Fig. 5 in panels B, E, H, and K). Altogether, these sets of results would allow practitioners to perform deeper and more nuanced downstream analyses of phenotypic behavior in different populations.

5. 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 described the probabilistic properties of GOALS assuming we have fit a Gaussian process regression model with a shift-invariant covariance function. 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 (Figs. 1-3 and S1). 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 data set 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 (Figs. 4 and S2-S6 and Tables S1-S3). The first key part of this analysis showed that GOALS incorporates non-additive information to find genetic signal that were missed by other approaches. The main takeaway from the real data analysis was that GOALS can provide local interpretability which enables downstream analyses to investigate how and why specific biomarkers are enriched for specific subsets of a population (Fig. 5). In this study, we saw how genetic variants 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 provides a measure of general association for nonlinear methods, it cannot be used to directly identify the component (i.e., linear versus nonlinear) that drives individual variable importance. Thus, despite being able to detect variables that are important 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). As another extension, GOALS does not enforce any sparsity or shrinkage when performing variable importance. Thus, while it has a natural null hypothesis, we do not provide a significance threshold for variable selection. Common examples in the statistics literature include a Bonferroni-corrected threshold (Gordon et al., 2007) or selection based on a median probability model (Barbieri and Berger, 2004). One natural solution would be to utilize the posterior variances of $\boldsymbol{\delta}^{(j)}$ and $\tilde{\boldsymbol{\delta}}^{(j)}$ which are derived and provided in the Supplementary Text. Another natural solution could be to permute the response variable 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. (10) 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 applications, particularly in biomedicine, recent methods have utilized prior knowledge to test groups of variables at a time to improve power (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. Software for computing the "RelATive cEntrality" (RATE) measure is carried out in `R` and `Python` code which is freely available at https://github.com/lorinanthony/RATE. The LASSO and elastic net regression models were run using the `glmnet` package in `R` (Friedman et al., 2010), while SCANONE was implemented using the baseline `lm()` function in `R`. The random forest was fit using the `randomForest` package (Liaw and Wiener, 2002), the gradient boosting machine was fit using the `gbm` package (Friedman, 2001), and the Bayesian additive regression tree was fit using the `BART` package (Sparapani et al., 2021) — also all in `R`. Lastly, the SHapley Additive exPlanation approach was implemented using the `shap` package in `Python`.

CRediT authorship contribution statement

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.

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. This study makes use of data generated by the Wellcome Trust Case Control Consortium (WTCCC). A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. Funding for the WTCCC project was provided by the Wellcome Trust under award 076113, 085475, and 090355. 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.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.csda.2023.107914.

References

Agrawal, R., Trippe, B., Huggins, J., Broderick, T., 2019. The kernel interaction trick: fast Bayesian discovery of pairwise interactions in high dimensions. In: Chaudhuri, K., Salakhutdinov, R. (Eds.), Proceedings of the 36th International Conference on Machine Learning. In: Proceedings of Machine Learning Research, vol. 97. PMLR, pp. 141–150. https://proceedings.mlr.press/v97/agrawal19a.html.

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

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

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

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

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

Bult, C.J., Blake, J.A., Smith, C.L., Kadin, J.A., Richardson, J.E., Anagnostopoulos, A., Asabor, R., Baldarelli, R.M., Beal, J.S., Bello, S.M., Blodgett, O., Butler, N.E., Christie, K.R., Corbani, L.E., Creelman, J., Dolan, M.E., Drabkin, H.J., Giannatto, S.L., Hale, P., Hill, D.P., Law, M., Mendoza, A., McAndrews, M., Miers, D., Motenko, H., Ni, L., Onda, H., Perry, M., Recla, J.M., Richards-Smith, B., Sitnikov, D., Tomczuk, M., Tonorio, G., Wilming, L., Zhu, Y., 2019, the Mouse Genome Database Group. Mouse genome database MGD. Nucleic Acids Res. 47 (D1), 801–806. https://doi.org/10.1093/nar/gky1056. ISSN 0305-1048, 1362-4962. https://academic.oup.com/nar/article/47/D1/D801/5165331, 2019.

Candès, E., Fan, Y., Janson, L., Lv, J., 2018. Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection. J. R. Stat. Soc., Ser. B, Stat. Methodol. 80 (3), 551–577. https://doi.org/10.1111/rssb.12265. ISSN 1369-7412, 1467-9868.

Carbonetto, P., Stephens, M., 2013. 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. https://doi.org/10.1371/journal.pgen.1003770.

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

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

Chen, H., Lundberg, S.M., Lee, S.-I., 2022. Explaining a series of models by propagating Shapley values. Nat. Commun. 13 (1), 1–15.

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

Cheng, L., Ramchandran, S., Vatanen, T., Lietzén, N., Lahesmaa, R., Vehtari, A., Lähdesmäki, H., 2019. An additive Gaussian process regression model for interpretable non-parametric analysis of longitudinal data. Nat. Commun. 10 (1), 1798. https://doi.org/10.1038/s41467-019-09785-8. ISSN 2041-1723.

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

Cheverud, J.M., Ehrich, T.H., Hrbek, T., Kenney, J.P., Pletscher, L.S., Semenkovich, C.F., 2004. 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. https://doi.org/10.2337/diabetes.53.12.3328. ISSN 0012-1797 (Print), 0012-1797 (Linking).

Chipman, H.A., George, E.I., McCulloch, R.E., 2010. BART: Bayesian additive regression trees. Ann. Appl. Stat. 4 (1), 266–298. https://doi.org/10.1214/09-AOAS285.

Conard, A.M., DenAdel, A., Crawford, L., 2023. A spectrum of explainable and interpretable machine learning approaches for genomic studies. WIREs: Comput. Stat., e1617. https://doi.org/10.1002/wics.1617.

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

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

Crawford, L., Wood, K.C., Zhou, X., Mukherjee, S., 2018. Bayesian approximate kernel regression with variable selection. J. Am. Stat. Assoc. 113 (524), 1710–1721. https://doi.org/10.1080/01621459.2017.1361830.

Crawford, L., Flaxman, S.R., Runcie, D.E., West, M., 2019. Variable prioritization in nonlinear black box methods: a genetic association case study. Ann. Appl. Stat. 13 (2), 958–989. https://doi.org/10.1214/18-AOAS1222. ISSN 1932-6157. https://projecteuclid.org/euclid.aoas/1560758434.

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

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

DeGrave, A.J., Janizek, J.D., Lee, S.-I., 2021. AI for radiographic COVID-19 detection selects shortcuts over signal. Nat. Mach. Intell. 3 (7), 610–619.

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

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

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

Fortelny, N., Bock, C., 2020. Knowledge-primed neural networks enable biologically interpretable deep learning on single-cell sequencing data. Genome Biol. 21 (1), 190. https://doi.org/10.1186/s13059-020-02100-5.

Friedman, J., Tibshirani, R., Hastie, T., 2010. Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. 33 (1), 1–22. https://doi.org/10.18637/jss.v033.i01.

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat. 29 (5), 1189–1232. https://doi.org/10.1214/aos/1013203451.

Gelman, A., Hwang, J., Vehtari, A., 2014. Understanding predictive information criteria for Bayesian models. Stat. Comput. 24 (6), 997–1016. https://doi.org/10.1007/s11222-013-9416-2. ISSN 0960-3174, 1573-1375.

Gordon, A., Glazko, G., Qiu, X., Yakovlev, A., 2007. Control of the mean number of false discoveries, Bonferroni and stability of multiple testing. Ann. Appl. Stat. 1 (1), 179–190. https://doi.org/10.1214/07-AOAS102.

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

Gu, L., Johnson, M.W., Lusis, A.J., 1999. 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. https://doi.org/10.1161/01.atv.19.2.442. ISSN 1079-5642 (Print). 1079-5642 (Linking).

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D., 2018. A survey of methods for explaining black box models. ACM Comput. Surv. 51 (5), 93.

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

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

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

Ishwaran, H., Lu, M., 2019. Standard errors and confidence intervals for variable importance in random forest regression, classification, and survival. Stat. Med. 38 (4), 558–582. https://doi.org/10.1002/sim.7803.

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

Kim, S.V., Mehal, W.Z., Dong, X., Heinrich, V., Pypaert, M., Mellman, I., Dembo, M., Mooseker, M.S., Wu, D., Flavell, R.A., 2006. Modulation of cell adhesion and motility in the immune system by Myo1f. Science 314 (5796), 136–139. https://doi.org/10.1126/science.1131920. ISSN 1095-9203 (Electronic). 0036-8075 (Linking).

Kindermans, P.-J., Hooker, S., Adebayo, J., Alber, M., Schütt, K.T., Dähne, S., Erhan, D., Kim, B., 2019. The (Un)reliability of Saliency Methods. Springer International Publishing, Cham. ISBN 978-3-030-28954-6, pp. 267–280.

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

Kowal, D.R., 2021. Fast, optimal, and targeted predictions using parameterized decision analysis. J. Am. Stat. Assoc., 1–12. https://doi.org/10.1080/01621459.2021.1891926.

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

Lawson, H.A., Lee, A., Fawcett, G.L., Wang, B., Pletscher, L.S., Maxwell, T.J., Ehrich, T.H., Kenney-Hunt, J.P., Wolf, J.B., Semenkovich, C.F., Cheverud, J.M., 2011. 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. https://doi.org/10.1007/s00335-010-9313-3. ISSN 1432-1777 (Electronic); 0938-8990 (Print); 0938-8990 (Linking).

Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R News 2 (3), 18–22. https://CRAN.R-project.org/doc/Rnews/.

Lin, A., Song, A.H., Bilgic, B., Ba, D., 2022. Covariance-free sparse Bayesian learning. IEEE Trans. Signal Process. 70, 3818–3831. https://doi.org/10.1109/TSP.2022.3186185.

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

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

Lundberg, S., Lee, S.-I., 2017. A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 4768–4777. ISSN 9781510860964. URL https://dl.acm.org/doi/10.5555/3295222.3295230.

Lundberg, S.M., Lee, S., 2016. An unexpected unity among methods for interpreting model predictions. CoRR. arXiv:1611.07478 [abs].

Martin, A.R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B.M., Daly, M.J., 2019. Clinical use of current polygenic risk scores may exacerbate health disparities. Nat. Genet. 51 (4), 584–591. https://doi.org/10.1038/s41588-019-0379-x. ISSN 1061-4036, 1546-1718.

McCaw, Z.R., Colthurst, T., Yun, T., Furlotte, N.A., Carroll, A., Alipanahi, B., McLean, C.Y., Hormozdiari, F., 2022. DeepNull models non-linear covariate effects to improve phenotypic prediction and association power. Nat. Commun. 13 (1), 241. https://doi.org/10.1038/s41467-021-27930-0.

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

Moen, C.J.A., Tholens, A.P., Voshol, P.J., de Haan, W., Havekes, L.M., Gargalovic, P., Lusis, A.J., van Dyk, K.W., Frants, R.R., Hofker, M.H., Rensen, P.C.N., 2007. The Hyplip2 locus causes hypertriglyceridemia by decreased clearance of triglycerides. J. Lipid Res. 48 (10), 2182–2192. https://doi.org/10.1194/jlr.M700009-JLR200. ISSN 0022-2275 (Print). 0022-2275 (Linking).

Murdoch, W.J., Singh, C., Kumbier, K., Abbasi-Asl, R., Yu, B., 2019. Definitions, methods, and applications in interpretable machine learning. Proc. Natl. Acad. Sci. 116 (44), 22071–22080. https://doi.org/10.1073/pnas.1900654116.

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

Östergren, C., Shim, J., Larsen, J.V., Nielsen, L.B., Bentzon, J.F., 2015. 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. https://doi.org/10.1371/journal.pone.0121899. ISSN 1932-6203 (Electronic). 1932-6203 (Linking).

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

Paananen, T., Andersen, M.R., Vehtari, A., 2021. Uncertainty-aware sensitivity analysis using Rényi divergences. In: Uncertainty in Artificial Intelligence. PMLR, pp. 1185–1194.

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

Pérez-Cruz, F., Van Vaerenbergh, S., Murillo-Fuentes, J.J., Lázaro-Gredilla, M., Santamaria, I., 2013. Gaussian processes for nonlinear signal processing: an overview of recent advances. IEEE Signal Process. Mag. 30 (4), 40–50.

Piironen, J., Vehtari, A., 2016. Projection predictive model selection for Gaussian processes. In: 2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP), pp. 1–6. http://arxiv.org/abs/1510.04813. arXiv:1510.04813.

Piironen, J., Vehtari, A., 2017. Comparison of Bayesian predictive methods for model selection. Stat. Comput. 27 (3), 711–735. https://doi.org/10.1007/s11222-016-9649-y. http://link.springer.com/10.1007/s11222-016-9649-y.

Rasmussen, C.E., Williams, C.K.I., 2006. Gaussian Processes for Machine Learning. Adaptive Computation and Machine Learning. MIT Press, Cambridge, Mass. ISBN 978-0-262-18253-9. OCLC: ocm61285753.

Roth, A.E., 1988. The Shapley Value: Essays in Honor of Lloyd S. Shapley. Cambridge University Press.

Rudin, C., 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat. Mach. Intell. 1 (5), 206–215.

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

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

Schulz, M.-A., Yeo, B.T.T., Vogelstein, J.T., Mourao-Miranada, J., Kather, J.N., Kording, K., Richards, B., Bzdok, D., 2020. Different scaling of linear models and deep learning in UKBiobank brain images versus machine-learning datasets. Nat. Commun. 11 (1), 4238. https://doi.org/10.1038/s41467-020-18037-z.

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

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

Sesia, M., Bates, S., Candès, E., Marchini, J., Sabatti, C., 2021. False discovery rate control in genome-wide association studies with population structure. Proc. Natl. Acad. Sci. 118 (40), e2105841118. https://doi.org/10.1073/pnas.2105841118.

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

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

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

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

Smith, S.P., Shahamatdar, S., Cheng, W., Zhang, S., Paik, J., Graff, M., Haiman, C., Matise, T.C., North, K.E., Peters, U., Kenny, E., Gignoux, C., Wojcik, G., Crawford, L., Ramachandran, S., 2022. Enrichment analyses identify shared associations for 25 quantitative traits in over 600,000 individuals from seven diverse ancestries. Am. J. Hum. Genet. 109 (5), 871–884. https://doi.org/10.1016/j.ajhg.2022.03.005. https://www.sciencedirect.com/science/article/pii/S000292972200101X.

Smith, S.P., Darnell, G., Udwin, D., Harpak, A., Ramachandran, S., Crawford, L., 2023. Accounting for statistical non-additive interactions enables the recovery of missing heritability from GWAS summary statistics. https://doi.org/10.1101/2022.07.21.501001. bioRxiv, 2022.07.21.501001. http://biorxiv.org/content/early/2023/06/23/2022.07.21.501001.abstract.

Sparapani, R., Spanbauer, C., McCulloch, R., 2021. Nonparametric machine learning and efficient computation with Bayesian additive regression trees: the BART R package. J. Stat. Softw. 97 (1), 1–66. https://doi.org/10.18637/jss.v097.i01.

Stamp, J., DenAdel, A., Weinreich, D., Crawford, L., 2023. Leveraging the genetic correlation between traits improves the detection of epistasis in genome-wide association studies. G3 Genes Genomes Genet., jkad118. https://doi.org/10.1093/g3journal/jkad118.

Stephens, M., 2016. False discovery rates: a new deal. Biostatistics, kxw041. https://doi.org/10.1093/biostatistics/kxw041.

Stephens, M., Balding, D.J., 2009. Bayesian statistical methods for genetic association studies. Nat. Rev. Genet. 10 (10), 681–690. https://doi.org/10.1038/nrg2615.

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

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

Taylor, B.A., Tarantino, L.M., Phillips, S.J., 1999. Gender-influenced obesity QTLs identified in a cross involving the KK type II diabetes-prone mouse strain. Mamm. Genome 10 (10), 963–968. https://doi.org/10.1007/s003359901141. ISSN 0938-8990, 1432-1777.

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. https://doi.org/10.1038/nature05911.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc., Ser. B, Methodol. 58 (1), 267–288. ISSN 00359246. URL http://www.jstor.org/stable/2346178.

Trippe, B.L., Finucane, H., Broderick, T., 2021. For high-dimensional hierarchical models, consider exchangeability of effects across covariates instead of across datasets. In: Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (Eds.), Advances in Neural Information Processing Systems. https://openreview.net/forum?id=28NikxkK6kJ.

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

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

Valdar, W., Flint, J., Mott, R., 2006a. 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. https://doi.org/10.1534/genetics.104.039313. ISSN 0016-6731.

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

Wahba, G., 1990. Splines Models for Observational Data. Series in Applied Mathematics, vol. 59. SIAM, Philadelphia, PA.

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

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

Woody, S., Carvalho, C.M., Murray, J.S., 2021. Model interpretation through lower-dimensional posterior summarization. J. Comput. Graph. Stat. 30 (1), 144–161. https://doi.org/10.1080/10618600.2020.1796684. ISSN 1061-8600, 1537-2715.

Wu, M.C., Kraft, P., Epstein, M.P., Taylor, D.M., Chanock, S.J., Hunter, D.J., Lin, X., 2010. Powerful SNP-set analysis for case-control genome-wide association studies. Am. J. Hum. Genet. 86 (6), 929–942.

Yalcin, B., Nicod, J., Bhomra, A., Davidson, S., Cleak, J., Farinelli, L., Østerås, M., Whitley, A., Yuan, W., Gan, X., Goodson, M., Klenerman, P., Satpathy, A., Mathis, D., Benoist, C., Adams, D.J., Mott, R., Flint, J., 2010. Commercially available outbred mice for genome-wide association studies. PLoS Genet. 6 (9), e1001085. https://doi.org/10.1371/journal.pgen.1001085. ISSN 1553-7404.

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

Yoshikawa, Y., Iwata, T., Sawada, H., 2015. Non-linear regression for bag-of-words data via Gaussian process latent variable set model. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 29. https://ojs.aaai.org/index.php/AAAI/article/view/9615.

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

Zhou, J., Wong, M.S., Chen, W.-C., Krainer, A.R., Kinney, J.B., McCandlish, D.M., 2022. Higher-order epistasis and phenotypic prediction. Proc. Natl. Acad. Sci. 119 (39), e2204233119. https://doi.org/10.1073/pnas.2204233119.

Zhou, X., Stephens, M., 2012. Genome-wide efficient mixed-model analysis for association studies. Nat. Genet. 44 (7), 821–824. https://doi.org/10.1038/ng.2310.

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

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. R. Stat. Soc., Ser. B, Stat. Methodol. 67 (2), 301–320.