Learning epistatic polygenic phenotypes with Boolean interactions

Merle Behr,[1,12] Karl Kumbier,[2,12] Aldo Cordova-Palomera,[3] Matthew Aguirre,[3,4] Euan Ashley,[5] Atul J. Butte,[6] Rima Arnaout,[6,7] Ben Brown,[1,8] James Priest,[3,9,11]∗ Bin Yu [1,10,11]∗

[1] Department of Statistics, University of California at Berkeley, Berkeley, CA, USA; [2] Department of Pharmaceutical Chemistry, University of California, San Francisco, CA, USA; [3] Department of Pediatrics, Stanford Medicine, Stanford, CA, USA; [4] Department of Biomedical Data Science, Stanford Medicine, Stanford, CA, USA; [5] Division of Cardiovascular Medicine, Stanford Medicine, Stanford, CA, USA; [6] Bakar Computational Health Sciences Institute, University of California, San Francisco, CA, USA; [7] Division of Cardiology, Department of Medicine, University of California, San Francisco, CA, USA; [8] Biosciences Area, Lawrence Berkeley National Laboratory, Berkeley, CA, USA; [9] Current affiliation: Tenaya Therapeutics, San Francisco, CA, USA; [10] Department of Electrical Engineering and Computer Sciences and Center for Computational Biology, University of California at Berkeley, Berkeley, CA, USA; [11] Co-senior authorship [12] These authors contributed equally to this work

∗ Correspondence: jpriest@stanford.edu, binyu@berkeley.edu

Abstract: Detecting epistatic drivers of human phenotypes remains a challenge. Traditional approaches use regression to sequentially test multiplicative interaction terms involving single pairs of genetic variants. For higher-order interactions and genome-wide large-scale data, this strategy is computationally intractable. Moreover, multiplicative terms used in regression modeling may not capture the form of biological interactions. Building on the Predictability, Computability, Stability (PCS) framework, we introduce the epiTree pipeline to extract higher-order interactions from genomic data using tree-based models. The epiTree pipeline first selects a set of variants derived from tissue-specific estimates of gene expression. Next, it uses iterative random forests (iRF) to search training data for candidate Boolean interactions (pairwise and higher-order). We derive significance tests from interactions by simulating Boolean tree-structured null (no epistasis) and alternative (epistasis) distributions on hold-out test data. Finally, our pipeline computes PCS epistasis p-values that evaluate the stability of improvement in prediction accuracy via bootstrap sampling on the test set. We validate the epiTree pipeline using the phenotype of red-hair from the UK Biobank, where several genes are known to demonstrate epistatic interactions. epiTree recovers both previously reported and novel interactions, which represent forms of non-linearities not captured by logistic regression models. Additionally, epiTree suggests interactions between genes such as *PKHD1* and *XPOTP1*, which are unlinked to *MC1R*, as novel candidate interactions associated with the red hair phenotype. Last but not least, we find that individual Boolean or tree-based epistasis models generally provide higher prediction accuracy than classical logistic regression.

Introduction

Epistasis between genetic alleles describes a non-additive relationship between different loci governing a single trait. Despite the fact that epistatic interactions have been widely hypothesized to provide biological insights into the underlying functional mechanism between genotype and phenotype [30, 4], most large-scale studies on polygenic contributions to date have focused on discovering additive effects. Discovering epistatic interactions at work in human biology has been a slow and small-scale process with relatively modest or minimal evidence from large-scale studies [35, 27]. Several definitions of epistasis have co-existed since the term arose more than a century ago [2]. The most commonly accepted statistical form goes back to [12], who defined epistasis as the "deviation from the addition of superimposed effects (...) between different Mendelian factors." However, deviation from additivity does not specify a unique mathematical model for either the null hypothesis of no-epistasis or the alternative of epistasis. First, the notation of additivity depends on the particular scaling of the response (e.g., penetrance for a binary trait) — for example, a multiplicative function becomes additive on a log-scale. This is particularly problematic for standard statistical approaches that model epistasis using a logit transform of penetrance (e.g., in logistic regression). Second, for a fixed scaling there are many ways to write an additive model of individual components — for example taking the inverse normal or other transformation of each feature before running logistic regression. Stated succinctly, Fisher's original definition of epistasis does not specify unique models for the null and the alternative hypotheses. Although this has been noted repeatedly in the literature [36, 9, 28, 29, 31], its consequences for interpretation of statistical results are often not highlighted.

The most common approach to date for statistical modelling of epistasis is via linear (continuous response) or logistic (binary response) regression for two genetic variants (possibly incorporating covariates for population structure and genetic-linkage) [40, 34, 14, 26]. In this setting, the null (i.e. no-epistasis) model has linear additive components for the two involved genes and the alternative (i.e. epistasis) model has an additional multiplicative interaction term. However, complex phenotypes can involve many more than two genes and the functional translation between genes and phenotype can be highly complex. As a result, interactions identified as highly significant by brute-force, pairwise searches with logistic/linear regression can correspond to models that provide a poor fit to the data, particularly when trying to generalize to new/unobserved samples. Moreover, statistical hypothesis testing based on heavily mis-specified null or alternative models are often unstable and can lead to irreproducible results [38, 25, 16, 11].

In addition to the challenges of modeling complex interaction forms, classical approaches face several computational barriers. Brute force searches become computationally intractable for genome-scale data, in particular, beyond pairwise interactions. Models limited to pairwise interactions provide a poor fit for complex traits involving many genes. Moreover, the standard approach of modeling a single interaction at a time may lead to further model misspecification by failing to take the entire genetic background

into account. Finally, analyses typically limit their focus to marginally important variants due to high-dimensional genomic data. As a result, interactions involving variants with weak marginal effects cannot be detected.

In this paper, we propose a new approach, the epiTree pipeline, to detect polygenic, epistatic interactions in genome-scale data, validated by a re-analysis of the red hair phenotype in UK biobank data. The epiTree pipeline divides data into training and test sets (often through a random split when external test set is not available), and then takes the following three steps:

Using the training set, we conduct a two-step biologically inspired dimension reduction for interaction screening. First, we use tissue-specific estimates of gene expression [13] to select a subset of variants for analysis, which improves computational efficiency and stabilizes recovered interactions. We note that other methods could be used for dimension reduction but will not necessarily share the gene-level interpretation. Second, we use iterative random forests (iRF) [1] to search for epistatic interactions of arbitrary size and Boolean forms on a genome-wide scale.

Using the training set, we construct tree-based statistical models for both epistasis and no-epistasis. These models evaluate deviations from additivity – as considered in Fisher's definition of epistasis – directly on the penetrance scale and capture data-driven forms of both individual additive and non-additive interaction components. More specifically, these models make use of classification and regression trees (CART) [6], which can capture flexible forms of non-linearity and Boolean-type behavior. Such a thresholding relationship, as captured by CART functions but not by linear/multiplicative functions, has been empirically observed in many biological processes, in particular, in genetics [22, 17, 21, 19]. Moreover, the interaction behavior captured by a decision tree can be visualized easily and thus, provides direct insights into the directionality of effects. In our data analysis, we generally found that CART models give better prediction accuracy on hold-out test data than linear/multiplicative models as used e.g., in logistic regression.

Based on these tree-based models in 2, we devise an inference or testing procedure, called epiTree test, that builds on the recently proposed predictability, computability, and stability (PCS) framework for veridical data science [43]. Importantly, epiTree test builds on our random sample split (of data into training and test sets), which separates the discovery from the inference stage. In addition, the procedure uses prediction error on the test set to screen models before testing to arrive at (PCS) p-values. Further, epiTree test does not rely on any chi-square distributional approximations, as, e.g., logistic regression, but instead directly evaluates p-values based on bootstrap sub-sampling from the test data. This makes our approach more robust to model-misspecification, which can result in unrealistic small p-values from poor distributional tail approximations [32]. It is worth noting that epiTree test is a stand-alone test that can be used for any given candidate interaction of genes (or variants) relative to a given phenotype.

Finally, we present results using our epiTree pipeline for the phenotype of red hair in the UK biobank.

We rediscover previously reported pairwise interactions surrounding *MC1R* and discover new interactions, for example pairwise interactions that include genes that were not associated previously with red-hair or even hair color more generally (e.g. *UPF3A*, *SIAH2*); pairwise interactions between genes that individually have been associated with hair color but have not been reported as interacting previously; as well as high-order (up to order-4) interactions which other methods cannot typically detect.

Material and Methods

Phenotype and genotype data from the UK Biobank:

As both an illustration of, and positive control for, our approach epiTree, we re-analyzed epistatic interactions for a well-studied, genetically-driven phenotype – that of red hair [26]. Our study is based on self-reported hair color from the UK Biobank (Data field 1747). We consider a set of 337,535 unrelated White British individuals: 15,326 with "Red" hair and 322,209 individuals with "Blonde", "Light Brown", "Dark Brown", "Black", or "Other". For our analysis we focused on a balanced sample, incorporating a random subset of 15,000 red hair individuals with an equally sized random subsample of the controls. To help assess the generalizability of our results to unobserved data, we performed a random stratified sample split into training and test sets with 26K training samples and 4K test samples. We note that, in general, one might prefer non-randomly to randomly sampled test sets, e.g., a data set from an external experiment. Such an external data set was non available to us for this analysis so we use a random sample split here.

We model hair color using $\sim 15,000,000$ common variants imputed to the Haplotype Reference Consortium (HRC) and UK10K reference panels from 800,000 directly genotyped variants, which were obtained by UK Biobank using one of two similar arrays [7]. Details regarding the ascertainment and quality control of these genotypes have been previously described [7, 26]. In brief, genotyped variants were subject to outlier-based filtration on effects due to batch, plate, sex, array, as well as discordance across control replicates. Samples with excess heterozygosity or missingness were excluded from the data release. Imputed variants were further subject to filtration based on Hardy-Weinburg equilibrium, missingness in white British individuals, minor allele frequency ($> 10^{-4}$), and imputation quality.

Two-step procedure for interaction screening with biologically inspired dimension reduction:

We used a two-step procedure for screening interactions to improve computational efficiency and ensure the stability of results. For the first step, we estimated tissue-specific gene expression derived from PrediXcan [13]. The PrediXcan approach includes a set of elastic net models trained to predict tissue-specific gene expression levels on GTEx v7 data [23]. Analysis of gene-expression reduced the number of

genetic features by several orders of magnitude, from $\sim 10^7$ variants to $\sim 10^4$ genes. For the phenotype of red hair we selected the estimates of gene expression derived from skin, which best captures melanocyte biology and pigmentation. Using tissue-specific estimates of gene expression, we searched for non-linear interactions between transcription level and the phenotype of interest using iterative random forests (iRF) [1] (see below). This analysis generated a set of putative gene-level interactions extracted by iRF. In the second step of our analysis, variants within 1MB of the start or end of an interacting gene were selected for identification of variant-level interactions by re-running iRF over a reduced subset of variants for the phenotype of interest. See Figure 1 for an illustration. Our two-step procedure identified variant- and gene-level interactions that capture nonlinear dependencies over arbitrary sets of variants/genes. The ability to model interactions among any set of genes arises directly from our use of iRF, which models the joint contribution of all genes simultaneously. We also explored a complementary one-step approach to directly search for variant-level interactions, by pre-filtering variants based on iRF applied to variant-sub-batches which is discussed in further detail in the Supplementary Information.

iRF non-linear model selection

Given a set of features (either estimated gene expression or individual genetic variants), we applied iRF [1, 18] as a non-linear model selection step to extract candidate interactions. The iRF algorithm fits a series of feature-weighted random forests (RF) [5] to iteratively stabilize feature selection along the decision paths of trees in the forest. After the final iteration, iRF identifies stable candidate interactions by searching for features that frequently co-occur on decision tree paths using random intersection trees (RIT) [33] (see Figure 4 and 5 for the data analysis detailed below and also figures in SI). A key benefit of iRF is that the computational complexity of searching for interactions does not grow exponentially with the size of an interaction. This allowed us to identify high-order (i.e. beyond pairwise) Boolean interactions in a computationally tractable manner, provided they appeared frequently (i.e. were stable) on decision tree paths. We further expanded our set of candidate interactions for PCS epistasis inference (see below) by taking all inter-chromosome pairwise interactions among the top 50 iRF genes (with respect to Gini importance).

Recall that prior to running iRF, we randomly divided the data into training and test sets of 26K and 4K subjects, respectively. Sample splitting allowed us to search for interactions (on training set) and conduct inference (on test set) in distinct datasets, helping to mitigate overfitting. Moreover, the accuracy of iRF on the hold-out test set provides a measure of reproducibility/generalizability for our non-linear model selection, and correspondingly, the identified interactions.

Predictability, Computability, Stability (PCS) inference for epistatic interactions

Classically, the significance of a candidate interaction for a binary phenotype is evaluated through logistic regression (see Supplementary Information for a recap). However, the model imposed by logistic regression, including a rescaling of the phenotype, transformations of gene expression data, and the polynomial interaction term, does not arise from a known biological observation or mechanism. Indeed, additive or multiplicative forms may not capture the stochastic nature of living systems or the threshold effects commonly observed in molecular and cellular phenomena [22, 17, 21, 19]. A mathematically inaccurate representation of the biological mechanism of action between multiple genetic loci may result in unstable and irreproducible results that fit poorly to data.

To address these issues, we evaluated the significance of candidate interactions using a novel PCS epistasis p-value. The PCS framework, recently proposed in [43], is based on the three core principles of data science: Predictability, Computability and Stability (PCS). It unifies and expands ideas from machine learning and statistics by using predictivity as a universal reality (or model) check, appropriate computational strategies for every step of an analysis process including simulating realistic reference/null distributions, and stability analysis to assess reproducibility at every stage of an analysis – from problem formulation/data collection, to modeling including inference, and to post-hoc model interpretations.

We computed PCS epistasis p-values based on models for both non-epistasis (null) and epistasis (alternative). Both models were based on decision trees defined over the set of interacting gene features identified by the iRF. For a given interaction, e.g. between gene A and B, (or a collection of interacting features in the set), the non-epistasis (null) model was defined as an additive combination of decision trees fit to individual features using the training set samples combined with backfitting. That is, the null model (no-epistasis) is of the form

$$H_0 : \; P(y = 1 \mid a, b) = CART_A(a) + CART_B(b),$$

where CART denotes a "Classification And Regression Tree" (see [6]). We stress that the null model H_0 also covers the case where the response $P(y = 1 \mid a, b)$ only depends on one of the genes A or B, or on neither of them, with $CART_A$ and/or $CART_B$ equal to a constant. For higher-order interactions, say of the feature set I with $|I| > 2$, the null model can be defined in a similar way. In this case, lower-order interactions among features $I' \subsetneq I$ must also be taken into account. We address this by defining the null model of no-epistasis based on disjoint partitions of features I into two groups I_1 and I_2, which can be considered as two disconnected graphs (or pathways). Features within each group (pathway) can interact with each other via a decision tree model $CART_{I_i}$, with $i = 1, 2$, but features between the two groups (pathways) cannot. The overall no-epistasis null model for the epiTree p-values then corresponds

to the sum of the two individual trees, such that features between the two groups do not interact with each other. We give details on the construction of p-values for higher-order interactions in the SI.

The epistasis (alternative) model was based on a single decision tree fit on training set samples and all features in the interaction. That is, the alternative model (no-epistasis) is of the form

$$H_1 : \; P(y = 1 \mid a, b) = CART_{AB}(a, b).$$

An example for fitted null and alternative models is shown in Figure 2 for the two genetic features A = *ASIP* and B = *DEF8*, see also the SI figure for the respective response surfaces. A particular advantage of working with decision trees is that they allow for flexible forms of interactions and additive components. As a result, our approach is often more agnostic to the underlying biological mechanism compared with a multiplicative term in logistic regression, as we highlight with several examples in the Results Section (see also our general discussion on this in the Discussion Section). Moreover, by modeling responses directly on the penetrance scale, our approach avoids the need for transforming responses as, e.g., performed in logistic regression. To assess the generalizability of our results, both the epistasis and no-epistasis models were learned on the training set, using the classification and regression tree (CART) algorithm, and evaluated on hold-out test data. We refer to the Supplementary Information for a detailed discussion.

For a given interaction, we compute a PCS p-value using a novel testing framework. Specifically, we evaluate prediction error (via cross-entropy/log-likelihood) of both the epistasis and the no-epistasis models, on hold-out test data. Note that the forms of both models are estimated using the training set only. When the epistasis model has a worse (or equal) prediction error than the no-epistasis model (on the hold-out test data), we report no significant finding by formally setting the PCS p-value equal to one. That is, prediction on the test set is used as a screening as stipulated by PCS inference in [43]. Otherwise, we quantify the improvement in prediction error (difference in cross-entropy/log-likelihood ratio) by calculating a PCS p-value as follows.

We consider B bootstrap-samples (each of size n) of the test data of size n (for the red-hair analysis the test set has size n = 4K). For each bootstrap-sample, we generate a null-perturbation of simulated independent samples of its n-tuple response vector by using the no-epistasis distribution and the feature vectors in the test data set. In fact, for each bootstrap sample, we have a pair of n-vectors of responses, one from the null distribution and one from a bootstrap subsample of the test set responses. For each pair of such n-vector responses, we evaluate whether the improvement in prediction error (difference in cross-entropy/log-likelihood ratio between the no-epistasis and the epistasis model) for the test set is smaller than for the null-perturbation. The PCS p-value is given by the percentage of bootstrap sub-samples where the improvement in prediction error for the test set response is smaller than for the

null-perturbation response. In practice, one should choose B sufficiently large such that the PCS p-value converges. A simple derivation (see Supplementary Information for details), shows that it is easy to approximate the PCS p-value (conditioned on the data) as B tends to infinity, which we have done for the red hair analysis.

We call this new test the epiTree test. As a result, the quality of an interaction is directly related to its prediction accuracy on hold-out test data. A detailed description of underlying assumptions and derivation of the PCS p-value is presented in the Supplementary Information. The results presented next are PCS epistasis p-values for gene level interactions found by iRF. We also present comparisons to the p-values obtained by a standard logistic regression analysis. We note that PCS inference could also be used to assess the significance of interaction terms in a logistic regression model, but we do not consider such an analysis in this paper.

Results

To evaluate how well iRF captures genotype/phenotype relationships for candidate interaction screening, we evaluated prediction accuracy (Predictability in PCS) on the hold-out test set. Details on parameter choice and implementation are given in the Supplementary Information. Figure 3 shows ROC curves for the prediction accuracy of iRF and competitors on the gene level (left) and variant level (right) (all fit using training data): penalized logistic regression with L1 penalty term and random forests (RF). With an area under the ROC curve (AUROC) of 0.93 (95% bootstrap confidence interval was [0.922, 0.936]) on hold-out test data, iRF demonstrates high prediction accuracy and outperforms its closest competitor (penalized logistic regression AUROC = 0.9 and a 95% bootstrap confidence interval of [0.894, 0.913]). (For ranger we obtained AUROC = 0.88 with a 95% bootstrap confidence interval of [0.871, 0.892].)

Figure 4 shows the list of candidate interactions obtained from our gene-level iRF non-linear model selector, while Figure 5 reports results on the variant level (see also figure in the SI). To account for spurious interactions that may arise from linkage disequilibrium [39, 44, 11], we restrict results to interactions among genes that are not all on the same chromosome.

In total the iRF interaction screening found 18 stable inter-chromosomal interactions (satisfying filtering criteria described in the Supplementary Information) of up to order-4. This amounts to a massive dimension reduction step, with respect to all $\sim 10^{16}$ possible order-4 interactions among $\sim 10^4$ genes. Interactions primarily center around genes in the *MC1R* and *ASIP* regions which corroborates previously reported epistatic interaction for the red hair phenotype within a well-described biological pathway, see [26]. We stress that while other works *a priori* restrict their search to interactions involving *MC1R*, our approach does not require such pre-selection.

Interaction terms contribute to an increase in prediction accuracy:

We evaluated the strength and significance of the individual candidate interactions using PCS epistasis p-values. The PCS epistasis p-values are shown in the second column of Figures 4 and 5. Comparisons with the corresponding classical p-values from logistic regression are shown in Supplementary Information. The first column reports test-set prediction errors of the CART models for both, epistasis and non-epistasis. We find that for several interactions, CART tree-based models show a considerable improvement in prediction error relative to logistic regression (e.g. *ASIP*, *TUBB3*; *ASIP*, *VPS9D1*). For the remaining interactions, prediction accuracy is comparable between logistic regression and CART models. Moreover, interactions that are significant with respect to the PCS p-value (< 0.05 with Bonferroni correction) correspond to those for which the epistasis model corresponds to a notable decrease in prediction error relative to the no epistasis model. In contrast, interactions that are significant with respect to the logistic regression p-value show little improvement in prediction accuracy between the epistasis and no epistasis models. This finding suggests that the increased flexibility and thresholding form of the Boolean interactions in CART, which reflect abundance thresholding behavior observed in biomolecular interactions [22, 17, 21, 19], can capture interaction behavior between genes better than a polynomial interaction model.

PCS detects different types of epistasis:

The PCS epistasis p-values and the classical logistic regression p-values can differ substantially, reflecting the fact that each captures a different form of interaction and thus different genotype/phenotype relationships. The PCS p-values, based on CART models, operate directly on the penetrance scale and capture thresholding-type behavior. Logistic regression, on the other hand, considers epistasis on a logit scale, and thus describes a different form of non-additivity, and is restricted to polynomial interaction terms. A key advantage of having CART as the building block for PCS p-values in our epiTree test is that significant interaction behavior is interpretable and easily visualized via the respective decision trees (see Figure 2 for an example). We highlight two examples below.

First, we considered the interaction between *ASIP* and *TUBB3*, which shows the greatest difference between PCS and classical p-values (Figure 4). Figure 6 plots the response surfaces, which indicate $P($ red hair $)$ as a function of interacting features. The top and bottom rows indicate response surfaces for the logistic regression and CART models respectively, while the columns indicate surfaces for the non-epistasis (left) and epistasis (right) models. For comparison, we plot these surfaces along with a smoothed scatter plot of the observed test data. Here, the polynomial interaction term assumed by logistic regression does not capture the highly non-monotone behaviour observed in *TUBB3*. On the other hand, the CART model appears to describe this relationship much better, which is confirmed by the improvement in prediction accuracy relative to logistic regression (see SI figures). In other words, the

PCS p-value detects a non-additive, epistatic relation between the two genes that cannot be captured by the logistic regression model. A similar example for the interaction *ASIP* - *DBNDD1* is shown in the SI.

Note that such non-monotonic behaviour, as it is observed for *TUBB3* in Figure 6, is common among PrediXcan estimated gene expression features. This is due to the fact that estimated expression of a single gene corresponds to a weighted linear combination of discrete variants. When a small number of variants are both highly associated with the response and have large PrediXcan weights for a given gene, the response surface exhibits marked transitions that correspond to different values of the highly weighted variants. This is exactly the case for *TUBB3*, with 25% of its weight mass on the variant rs8048449, which shows strong marginal association with red hair.

Second, we considered the interaction between *DEF8* and *ASIP*, shown in SI. For this interaction, the classical logistic regression p-value is highly significant (with Bonferroni correction), while the PCS epistasis p-value is not. The response surfaces indicate that these models differ most in the region where *DEF8* is small and *ASIP* is large. While the logistic regression model estimates a high probability of red hair in this region, resulting in a significant interaction term (with Bonferroni correction), there is limited data in the region to support how responses behave here. Indeed, all four models (logistic regression non-epistasis, logistic regression epistasis, CART non-epistasis, CART epistasis) achieve nearly identical prediction accuracy (with average cross entropies of 0.624, 0.620, 0.622, 0.630, see Figure 4). In other words, there is no evidence with respect to prediction accuracy that a particular model should be preferred, despite the significance of the classical logistic regression p-value. The SI shows another example for the interaction *ZNF276-RPL36P4* where the PCS p-value is not significant but the logistic regression p-value is. Here, the data clearly shows a non-linear relationship in one of the variables, which cannot be captured by the linear components of the null model for logistic regression. CART can capture such a relation well and outperforms logistic regression with a simple additive relationship. That is, the logistic regression significant p-values appear to be driven by a main effect that is not captured by the linear form of main effect in logistic regression.

We provide a similar discussion for the remaining three pairwise interactions for which either the PCS p-value or the logistic regression p-value was significant (with Bonferroni correction) in the SI.

In summary, we find that CART tree-based models used by our PCS epistasis p-values can capture different types of interaction, and thus lead to novel insights that are not detectable in standard logistic polynomial models. Moreover, the CART decision tree structure provides an intuitive, and accessible visualization of the interaction via the decision tree, which further aids in the interpretation of the results.

Biological findings:

Recapitulation of epistasis between *MC1R* and *ASIP*: For the red hair phenotype our epiTree pipeline recovered several well known epistatic interactions between the *MC1R* and *ASIP* region, both

on the gene and variant levels (Figure 4 and 5). The *MC1R* gene on chromosome 16 is an evolutionarily conserved regulator of pigmentation with a strong association with the red hair phenotype in humans. As a result, other works, studying epistasis related to red-hair, restrict to interactions between variants in the *MC1R* region and other marginally associated regions as a pre-filtering step to reduce the overall search space for interactions, see [26]. We stress that our pipeline does not rely on any such *a priori* knowledge and still recovers the *MC1R* and *ASIP* interaction, making it particularly valuable for phenotypes that have not been studied extensively.

Novel higher order interactions: Our epiTree pipeline recovered several higher order interactions, which is not feasible in most other epistasis pipelines, which typically restrict to pairwise interactions, with a few exceptions where order three interactions are considered, e.g., in [15]. For example, the top order three interaction between *ASIP*, *CDK10*, and *TUBB3* had a significant PCS p-value in 90% of bootstrap replicates (with Bonferroni correction). Moreover, the respective CART interaction model showed a strong increase in prediction accuracy, compared to the additive CART model (no-epistasis), as well as both logistic regression models (epistasis and non-epistasis). We note that *CDK10* and *TUBB3* are both genes on chromosome 16 which are just 200 kb apart. Thus, in this case, we cannot rule out the effect of strong genetic linkage which may result in spurious detection of epistasis.

Suggestions of new genes not associated with red hair previously: In addition to recapitulating well established epistatic interactions between *MC1R* and *ASIP*, our pipeline epiTree also provides new insights for the red hair phenotype. The iRF model selection step identified two genes that were not previously associated with red hair, namely *UPF3A* and *SIAH2*, both in interactions with *MC1R* related genes (*DBNDD1* + *UPF3A* and *CDK10* + *SIAH2*). However, neither the PCS p-values nor the logistic regression p-values were significant for these interactions and also the subsequent iRF variant interaction filtering did not result in any epistasis related to these genes. Consequently, we cannot report those interactions as significant findings. Nonetheless, we do see some further indication for potential association of these genes with red hair. First, in the pairwise gene level search of the top iRF genes, both, *UPF3A* and *SIAH2*, appear in interactions that have significant logistic regression p-value in some of the bootstrap replicates (with Bonferroni correction), see SI figure. We note, however, that the PCS p-value, while providing higher prediction accuracy, was not significant for the same interactions. Second, among all pairwise interactions between the variants that enter the PrediXcan model for *CDK10* and *SIAH2*, respectively, we find that the *SIAH2* variant rs482236 together with another *CDK10* variant shows significant PCS p-value in 20% among bootstrap samples of the test data (with Bonferroni correction), see SI figure. In summary, with the current data at hand, we cannot report any strong statistical evidence for the association of *UPF3A* and *SIAH2* with red hair, but these are suggestive of further investigations with new data.

Evidence of epistasis amongst previously reported genes: For the pairwise inter chromosome gene level analysis of the top 50 iRF genes, we found that an interaction between *PKHD1* (chr 6) and *XPOTP1* (chr 20) had significant PCS p-value among 30% of bootstrap replicates (of the test data) after Bonferroni correction, with the CART model from PCS also showing some improvement in terms of prediction accuracy compared to the logistic regression models. The respective response surface and interaction CART model, shown in SI, suggests a thresholding of gene expression levels with the red hair phenotype. In particular, one observes a decrease in red hair penetrance when gene expression of both, *PKHD1* and *XPOTP1*, are small and an increase when both are large. Otherwise, when one is small and the other is large, one observes an average red hair penetrance of around 50% (recall that we consider a balanced sample in this analysis). This interaction behavior is well described by the tree-based model used in the PCS p-value and results in a significant PCS p-value of $\sim 10^{-3}$. Both, *PKHD1* and *XPOTP1*, have been associated with hair color previously. In particular, in the recent work [26] both genes had reported epistasis with *MC1R*. However, epistasis between the two genes has not been previously described.

PCS epistasis p-values detect diverse interaction forms in simulation studies:

To evaluate the performance of inference methods, we conducted a series of data-based simulations comparing (i) epiTree PCS epistasis p-values (ii) logistic regression p-values (iii) permutation random forest (pRF) p-values [20]. We generated responses using four different models, each corresponding to a different form of interaction, applied to the gene expression data from the red hair case study. For each response generating model, we generated a set of candidate interactions using iRF and computed p-values for each inference method. For three of the four response generating models, epiTree both consistently recovered true interaction(s) (along with subsets of interacting features) and ranked these interactions as most significant across all candidates. In contrast, pRF consistently recovered true interactions in 2/4 response generating models. However, pRF tended to rank any candidate interaction that included the true interaction as highly significant (even if the candidate included inactive features). Finally, logistic regression recovered the true interaction in 2/4 response generating models, though at a lower significance level and with greater variability. For a detailed description of our simulation study and results, see Supplementary Information.

Discussion:

Summary of findings: Here we introduce the epiTree pipeline, building on a novel tree-based method and PCS framework, to detect epistasis. epiTree is grounded in a biologically relevant dimensional-reduction schema derived from tissue-specific gene expression to derive variant-level interactions, and

additionally, may detect interactions of between more than two-loci. Without prior knowledge of the phenotype of red-hair coloration, our approach detects known interactions between *MC1R* locus and nearby genes in the locus related to pigmentation and *ASIP*, as well as novel interactions between *CDK10* and *SIAH2*. Our findings may suggest a previously unrecognized relationship between *SIAH2* and the red hair phenotype. Interestingly, *SIAH2* is known to be associated with breast cancer risk and cellular response to hypoxia – a phenotype that has been associated with melanoma and pigmentation in general [3].

Related work: Previous approaches typically screen variants based on main effects due to computational constraints [8, 10, 15, 26], which may miss interacting genetic variants with weak main effects. The tree-based approach presented here does not limit analysis to variants with a strong main effect, however we note that decision tree-based methods have been proposed before in epistasis, e.g., [8, 15, 37, 41]. In [8] and [15] they are used to obtain marginal feature importance for the individual variant features from which candidate interactions are extracted either via a brute force search or with the help of phasing, in particular, independent of the phenotype under consideration. In contrast, going beyond main effect genes, iRF explicitly exploits the structure of the trees in a forest (at its last iteration) to extract Boolean interactions by interpreting frequently co-occurring features along the paths of trees as interactions. Similar to iRF, [37] and [41] take the joint appearance of features on a decision path as a measure of interaction. [37] interpret every path in a collection of boosted decision trees as a candidate interaction. To ensure the set of reported interactions is a manageable size, the authors restrict their search to interactions among at most d variants (with d = 5 being the default value). In contrast, iRF employs the stability principle [42, 43] – reporting feature combinations that frequently co-occur rather than the entire set of decision paths — to filter the set of candidate interactions with no restriction on size, after soft-dimensionality reduction through iterations of RF. [41], on the other hand, extract a measure of interaction strength from an RF for a pre-selected interaction. Moreover, their approach does not provide a way to extract candidate interactions from the trees as is done in iRF via the RIT algorithm; thus, it also requires a brute force search in practice. We are not aware of any other tree-based method for epistasis that incorporates stability driven bootstrap sub-sampling into its pipeline for screening interactions. More importantly, all these previous works do not develop a significance test or p-value that is based on tree models to evaluate the found interactions.

Limitations and directions for future work: Our application of iRF is based upon biologically-relevant dimensional reduction in the form of estimated tissue-specific expression values derived from GTeX via PrediXcan [13]. While this approach detects both known and previously undescribed epistasis for the red-hair phenotype, it depends on the accuracy of gene expression estimates derived from bulk RNA sequencing on heterogeneous cell types in skin tissue. In this case, we make the assumption that

tissue-specific data is sufficiently representative of melanocyte biology from which the trait of pigmentation and red-hair is derived. This assumption of representativeness is not likely to hold for traits deriving from rare cell types or from tissues and cell-types which are not sampled in GTeX in the first place. Additionally we note that the number of individuals in GTeX is not currently representative of all population strata which serves to limit the application of our approach for detection of epistasis to human phenotypes by excluding diverse genetic populations. Nevertheless, our epiTree test is a stand-alone test for epistasis of any given interaction relative to a phenotype that arises from other knowledge rather than the PrediXcan.

In principle, iRF can detect interactions of arbitrary order. However, the maximal detectable interaction order depends on tree depth—an interaction corresponds to subsets of features on individual decision paths, which scales logarithmically with the number of samples. Another potential limitation of the iRF interaction screening is that it does not explicitly take the genomic location of individual features into account. Even on the gene expression level, features that are close-by on the genome are typically highly correlated. Such features might appear exchangeably in interactions, as observed for the red hair phenotype with *MC1R*, see Figure 4. More generally, it is well known that for random forest based algorithms, correlated features can lead to masking effects, where the effect of a feature might be hidden by another dominating, correlated feature [24]. Finally, as discussed earlier, genetic features that are close-by on the genome might show artificial epistasis that is caused by hidden linkage with unobserved causal variants. In fact, for the red hair phenotype iRF reported various intra-chromosome interactions that we filtered out subsequently. One idea to overcome these limitations, would be to analyze by haplotypes of merged close-by genes into a single hyper-feature. However, two correlated genetic features do not necessarily show the same epistatic behaviour. Thus, there is a general tradeoff. Aggregating features in this way may increase the ability to detect epistasis by concentrating the forest on a single feature. On the other hand, aggregation may lead to loss of specificity that decreases the chance of detecting epistasis.

Although, the CART tree-based model that we propose for the PCS p-value (see Materials and Method and Supplementary Information) can describe significantly more flexible additive and non-additive relationships for the epistasis and non-epistasis model than a single linear or multiplicative term, we also emphasize that not every functional relationship might be well described by a relatively simple CART model. In particular, when a relationship is truly linear or multiplicative, clearly, working with a linear model will be more powerful than a CART approximation. The response surface plots, as shown in e.g., in Figure 6 and discussed in the Results section, are a powerful tool to investigate whether CART or linear models might be more appropriate for the given data at hand. Moreover, we stress that PCS p-values are, in principle, not restricted to CART models, but could be combined with any learning algorithm that appears appropriate for the data. As with any machine learning algorithms there is typ-

ically a tradeoff between interpretability and prediction accuracy. We find the CART tree-based models provide a good balance between both, being sufficiently flexible and very straight-forward to interpret, while capturing the thresholding behavior of interactions among many biomolecules.

The sample splitting paradigm that we follow in our approach holds the advantage of separating the discovery from the inference stage, which prevents overfitting and thus, makes findings more reliable, in general. We note, however, that compared to classical p-values, the need of a separate training set, will generally imply that there is less data available to evaluate the p-values, thus, a decrease in power. However, we note that due to the iRF interaction selection step, in general our approach requires the testing of orders of magnitude fewer hypotheses than any brute-force p-value approach. Thus, multiplicity correction of p-values will be much less of a problem in our pipeline, i.e., less stringent significance thresholds for p-values apply, which can compensate for this effect.

Conclusions: Our new methodology, the epiTree pipeline, represents a novel approach to the identification of Boolean epistatic genetic relationships. epiTree goes well beyond multiplicative interaction, is biologically often more meaningful, and is capable of detecting genetic interactions of order three and higher.

Supplementary Information (SI)

SI includes supplementary text, 20 figures, and 2 tables.

Declaration of Interests

Atul Butte is a co-founder and consultant to Personalis and NuMedii; consultant to Samsung, Mango Tree Corporation, and in the recent past, 10x Genomics, Helix, Pathway Genomics, and Verinata (Illumina); has served on paid advisory panels or boards for Geisinger Health, Regenstrief Institute, Gerson Lehman Group, AlphaSights, Covance, Novartis, Genentech, Merck, and Roche; is a shareholder in Personalis and NuMedii; is a minor shareholder in Apple, Facebook, Alphabet (Google), Microsoft, Amazon, Snap, 10x Genomics, Illumina, CVS, Nuna Health, Assay Depot, Vet24seven, Regeneron, Sanofi, Royalty Pharma, AstraZeneca, Moderna, Biogen, Paraxel, and Sutro, and several other non-health related companies and mutual funds; and has received honoraria and travel reimbursement for invited talks from Johnson and Johnson, Roche, Genentech, Pfizer, Merck, Lilly, Takeda, Varian, Mars, Siemens, Optum, Abbott, Celgene, AstraZeneca, AbbVie, Westat, and many academic institutions, medical or disease specific foundations and associations, and health systems. Atul Butte receives royalty payments through Stanford University, for several patents and other disclosures licensed to NuMedii and Personalis. Atul Butte's research has been funded by NIH, Northrup Grumman (as the prime on an NIH contract), Genentech,

Johnson and Johnson, FDA, Robert Wood Johnson Foundation, Leon Lowenstein Foundation, Intervalien Foundation, Priscilla Chan and Mark Zuckerberg, the Barbara and Gerson Bakar Foundation, and in the recent past, the March of Dimes, Juvenile Diabetes Research Foundation, California Governor's Office of Planning and Research, California Institute for Regenerative Medicine, L'Oreal, and Progenity.

Acknowledgments

This work was supported in part by a Chan Zuckerberg Biohub Intercampus Research Award to EA, AB, RA, BB, JP, and BY. MB was supported by Deutsche Forschungsgemeinschaft (DFG; German Research Foundation) Postdoctoral Fellowship BE 6805/1-1. MB acknowledges partial support from NSF Grant Big Data 60312. BY acknowledges partial support from Army Research Office Grant W911NF1710005 and NSF Grants DMS-1613002 and IIS 1741340. MA is currently supported by the National Library of Medicine under training grant T15 LM 007033. This work was carried out under UK Biobank study number 15860.

Data and Code Availability

The datasets and code generated during this study are available at `https://github.com/merlebehr/epiTree`. The raw genetic and phenotype data are available from UK Biobank.

References

[1] S. Basu, et al. Iterative random forests to discover predictive and stable high-order interactions. *Proceedings of the National Academy of Sciences*, 115(8):1943–1948, 2018.

[2] W. Bateson. *Mendel's Principles of Heredity.* Cambridge Univ. Press, 1909.

[3] B. Bedogni and M. B. Powell. Hypoxia, melanocytes and melanoma - survival and tumor development in the permissive microenvironment of the skin. *Pigment Cell & Melanoma Research*, 22(2):166–174, 2009.

[4] J. T. Bell, et al. Genome-wide association scan allowing for epistasis in type 2 diabetes: 2D GWA scan of type 2 diabetes. *Annals of Human Genetics*, 75(1):10–19, 2011.

[5] L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

[6] L. Breiman, et al. *Classification and Regression Trees.* Chapman and Hall, New York, 1984.

[7] C. Bycroft, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, 2018.

[8] X. Chen, et al. A forest-based approach to identifying gene and gene–gene interactions. *Proceedings of the National Academy of Sciences of the United States of America*, 104(49):19199–19203, 2007.

[9] H. J. Cordell. Epistasis: What it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics*, 11(20):2463–2468, 2002.

[10] H. J. Cordell. Detecting gene–gene interactions that underlie human diseases. *Nature Reviews Genetics*, 10(6):392–404, 2009.

[11] G. de los Campos, D. A. Sorensen, and M. A. Toro. Imperfect linkage disequilibrium generates phantom epistasis (and perils of big data). *G3: Genes, Genomes, Genetics*, 9(5):1429–1436, 2019.

[12] R. A. Fisher. The correlation between relatives on the supposition of mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, 52(2):399–433, 1919.

[13] E. R. Gamazon, et al. A gene-based association method for mapping traits using reference transcriptome data. *Nature Genetics*, 47(9):1091–1098, 2015.

[14] Y. Huang, S. Wuchty, and T. M. Przytycka. eQTL epistasis – challenges and computational approaches. *Frontiers in Genetics*, 4:51, 2013.

[15] R. Jiang, et al. A random forest approach to the detection of epistatic interactions in case-control studies. *BMC Bioinformatics*, 10(Suppl 1):S65, 2009.

[16] K. Kim. Massive false-positive gene–gene interactions by Rothman's additive model. *Annals of the Rheumatic Diseases*, 78(3):437–439, 2019.

[17] O. Kobiler, et al. Quantitative kinetic analysis of the bacteriophage genetic network. *Proceedings of the National Academy of Sciences*, 102(12):4470–4475, 2005.

[18] K. Kumbier, et al. Refining interaction search through signed iterative Random Forests. *bioRxiv:467498*, 2018.

[19] E. Levine and T. Hwa. Small RNAs establish gene expression thresholds. *Current Opinion in Microbiology*, 11(6):574–579, 2008.

[20] J. Li, et al. Detecting gene-gene interactions using a permutation-based random forest method. *BioData mining*, 9(1):1–17, 2016.

[21] J. W. Little. Threshold effects in gene regulation: When some is not enough. *Proceedings of the National Academy of Sciences*, 102(15):5310–5311, 2005.

[22] J. W. Little, D. P. Shepley, and D. W. Wert. Robustness of a gene regulatory circuit. *The EMBO Journal*, 18(15):4299–4307, 1999.

[23] J. Lonsdale, et al. The Genotype-Tissue Expression (GTEx) project. *Nature Genetics*, 45(6):580–585, 2013.

[24] G. Louppe. Understanding random forests: From theory to practice. *arXiv:1407.7502*, 2015.

[25] B. B. McShane, et al. Abandon statistical significance. *The American Statistician*, 73(sup1):235–245, 2019.

[26] M. D. Morgan, et al. Genome-wide study of hair colour in UK Biobank explains most of the SNP heritability. *Nature Communications*, 9:5271, 2018.

[27] A. Nag, M. I. McCarthy, and A. Mahajan. Large-scale analyses provide no evidence for gene-gene interactions influencing type 2 diabetes risk. *Diabetes*, 69(11):2518–2522, 2020.

[28] B. V. North, D. Curtis, and P. C. Sham. Application of logistic regression to case-control association studies involving two causative loci. *Human Heredity*, 59(2):79–87, 2005.

[29] P. C. Phillips. Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems. *Nature reviews. Genetics*, 9(11):855–867, 2008.

[30] M. D. Ritchie. Finding the epistasis needles in the genome-wide haystack. In *Epistasis. Methods in Molecular Biology (Methods and Protocols)*, volume 1253. Humana Press, New York, 2015.

[31] Z. R. Sailer and M. J. Harms. Detecting high-order epistasis in nonlinear genotype-phenotype maps. *Genetics*, 205(3):1079–1088, 2017.

[32] S. Santosh Bangalore, J. Wang, and D. B. Allison. How accurate are the extremely small -values used in genomic research: An evaluation of numerical libraries. *Computational Statistics & Data Analysis*, 53(7):2446–2452, 2009.

[33] R. D. Shah and N. Meinshausen. Random intersection trees. *The Journal of Machine Learning Research*, 15(1):629–654, 2014.

[34] M. Ueki and H. J. Cordell. Improved statistics for genome-wide interaction analysis. *PLOS Genetics*, 8(4):e1002625, 2012.

[35] K. Van Steen and J. H. Moore. How to increase our belief in discovered statistical interactions via large-scale association studies? *Human Genetics*, 138(4):293–305, 2019.

[36] M. J. Wade, et al. Alternative definitions of epistasis: Dependence and interaction. *Trends in Ecology & Evolution*, 16(9):498–504, 2001.

[37] X. Wan, et al. BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies. *The American Journal of Human Genetics*, 87(3):325–340, 2010.

[38] R. L. Wasserstein and N. A. Lazar. The ASA statement on p-values: Context, process, and purpose. *The American Statistician*, 70(2):129–133, 2016.

[39] A. R. Wood, et al. Another explanation for apparent epistasis. *Nature*, 514(7520):E3–E5, 2014.

[40] X. Wu, et al. A novel statistic for genome-wide interaction analysis. *PLoS Genetics*, 6(9):e1001131, 2010.

[41] M. Yoshida and A. Koike. SNPInterForest: A new method for detecting epistatic interactions. *BMC Bioinformatics*, 12:469, 2011.

[42] B. Yu. Stability. *Bernoulli*, 19(4):1484–1500, 2013.

[43] B. Yu and K. Kumbier. Veridical data science. *Proceedings of the National Academy of Sciences*, 117(8):3920–3929, 2020.

[44] Y. Zan, S. K. G. Forsberg, and Ö. Carlborg. On the relationship between high-order linkage disequilibrium and epistasis. *G3: Genes, Genomes, Genetics*, 8(8):2817–2824, 2018.

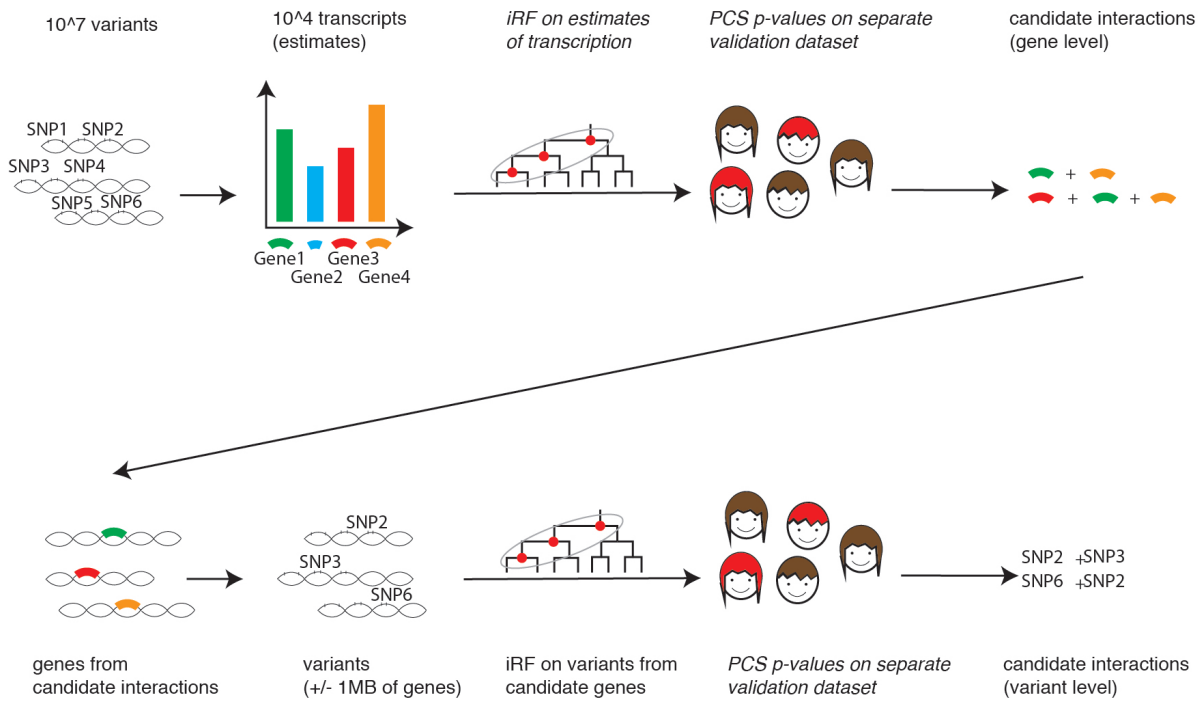

Figure 1: Illustration of the two-step procedure, from gene to variant level, for extraction of candidate interactions. First row from left to right: from approx. 10^7 variants tissue specific transcripts for approx. 10^4 genes are imputed using the software PrediXcan. Then the epiTree pipeline is applied to extract interactions for the gene expression features. Second row from left to right: for all genes that appear in interactions from the first step, variants within 1MB of the start or end of an interacting gene are extracted. Then the epiTree pipeline is applied to extract interactions for the variant features.

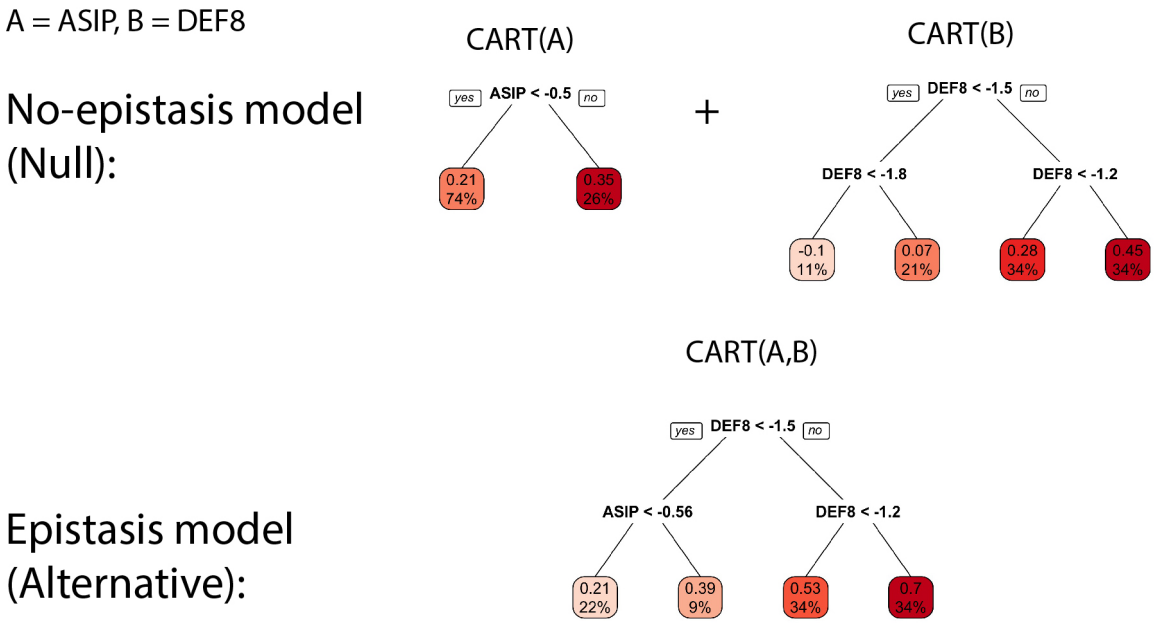

Figure 2: Example for CART based fitted null model (no-epistasis) shown in top row and alternative model (epistasis) shown in bottom row, for gene expression features A = *ASIP* and B = *DEF8*. The decimal digits at the tip nodes correspond to the predicted probability of red hair. The percentage at the tip nodes corresponds to the percentage of training observations falling into this tip node.

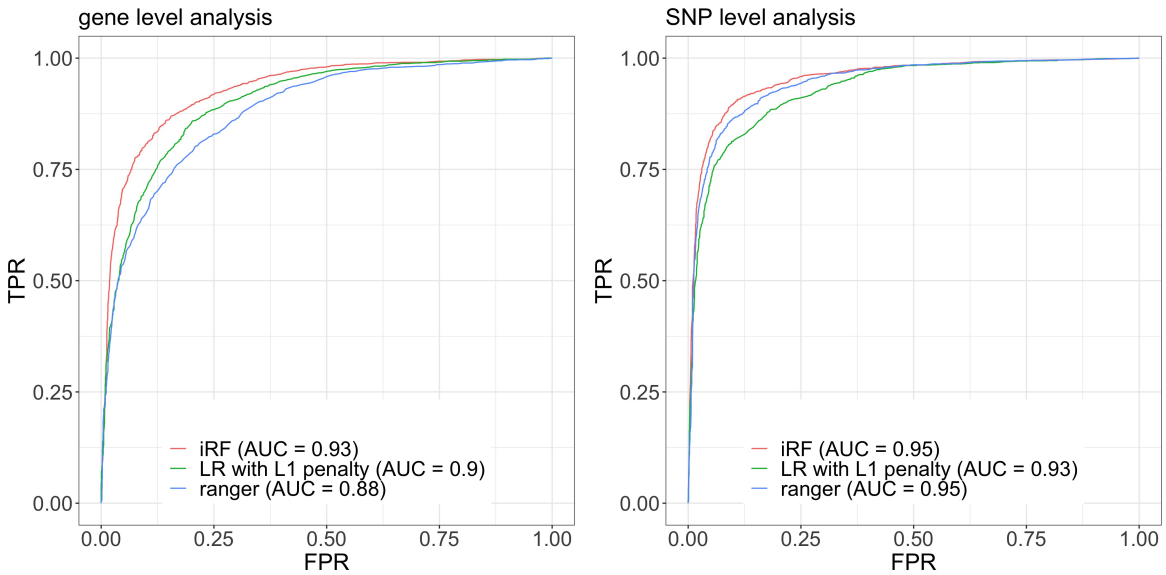

Figure 3: ROC curves of iRF prediction model and competitors on hold out test data. "LR with L1 penalty" stands for a logistic regression model with an additive L1 penalty term on the parameter vector, i.e., a lasso type estimator. The lambda tuning parameter was selected via cross validation using the cv.glmnet R function from the glmnet R package. The "ranger" competitor corresponds to the random forest implementation of the R package ranger with default parameters. Left: gene expression features. Right: Variant features

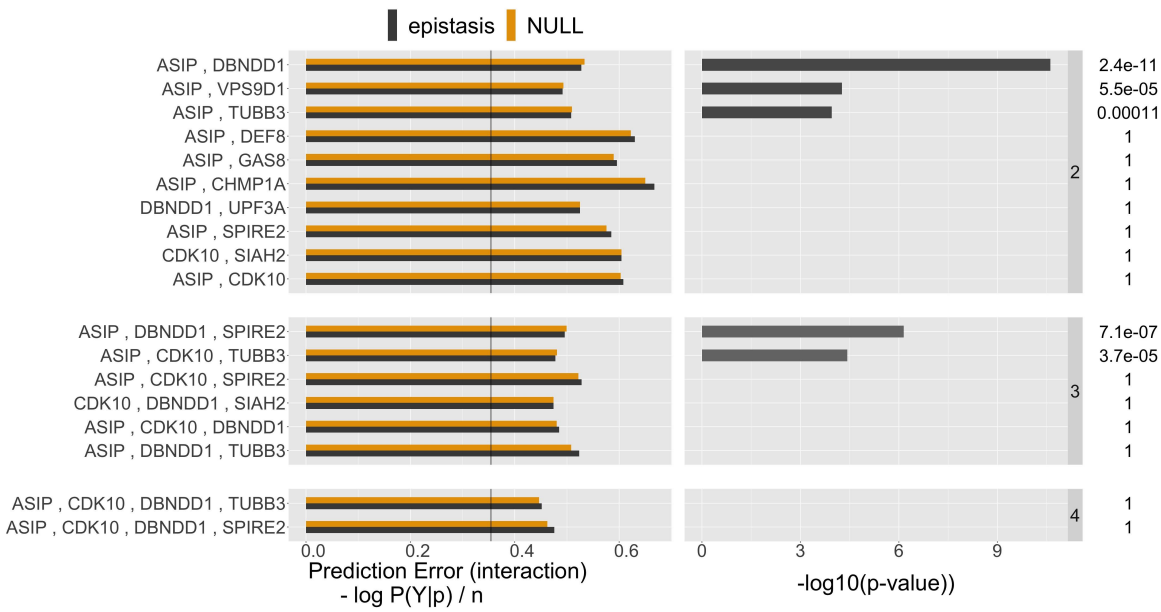

Figure 4: List of stable gene level interactions found by iRF (stability score > 0.5). The fist column shows the prediction error (cross-entropy) on the test data of the learned CART models for both, no-epistasis (NULL, orange) and epistasis (alternative, gray). The second column shows the PCS p-value on a -log10 scale and the numeric value is shown on the very right, up to two significant digits. The black vertical line in the first column shows the prediction error achieved by iRF using all the gene features simultaneously.

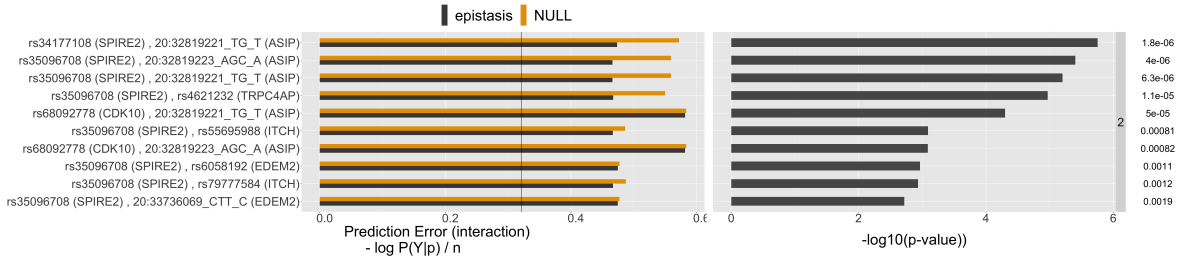

Figure 5: Same as Figure 4 for the top 10 order two variant level interactions.

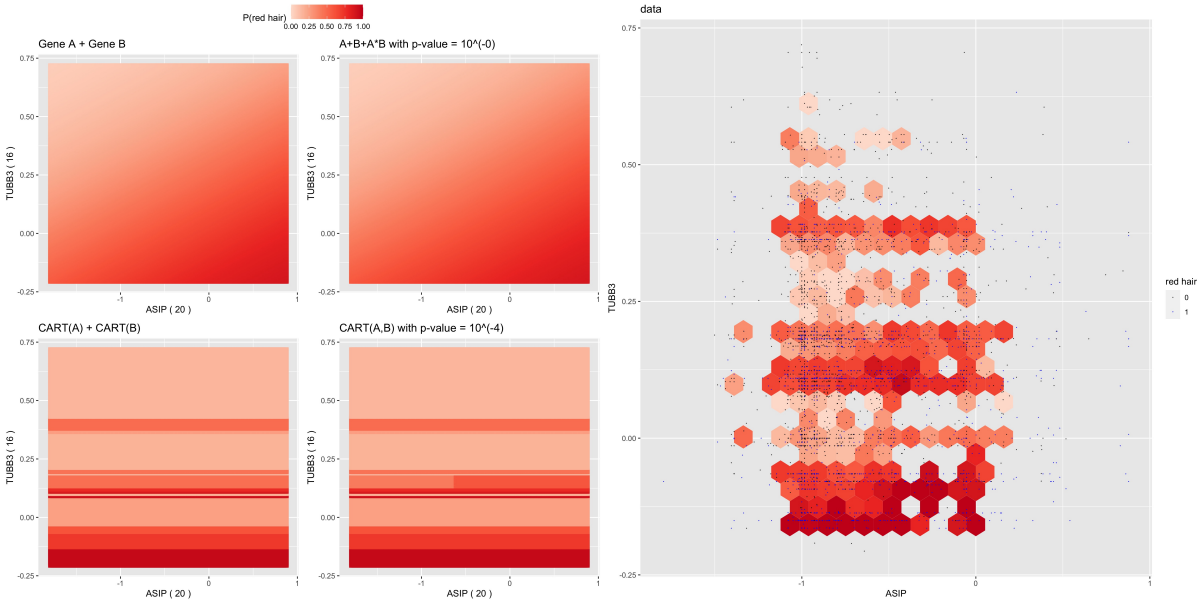

Figure 6: Response surface for *ASIP* - *TUBB3*, right plot: smoothed test data (hexagons are drops when they contain less than 5 data points), the full test data is shown as blue (red-hair) and black (non-red hair) dots ; left plot: response surfaces for fitted models; top: logistic regression model, bottom: CART based model, right: epistasis model, left: non-epistasis model.

Supplementary Information to manuscript Learning epistatic polygenic phenotypes with Boolean interactions

S1 Supplemental Material and Methods

S1.1 epiTree test based on PCS epistasis inference and PCS p-value

Traditional methods of statistical hypothesis testing evaluate uncertainty relative to a null distribution that describes the hypothesized data generating process. Intuitively, these methods address questions of the form: how likely is the observed test statistic if data were drawn from the specified null distribution. A statistic (and hence data) that is not likely under the null provides evidence against the corresponding null model. In practice, null distributions are often used without justification as to why they are a "reasonable" baseline for such comparisons. This can lead to small p-values that provide strong evidence against a "straw-man" null hypothesis — i.e. a null hypothesis that is *a priori* known or widely believed to be a poor description of the data generating process.

To address this issue, the PCS inference framework [32] introduced a prediction screening step that evaluates model accuracy on hold-out test set. This step allows one to filter out "straw-man" null hypotheses and focus inference on those that provide a reasonable fit to the data. In order to evaluate the accuracy of red hair phenotype predictions, we split data into training and hold-out test sets, learn models corresponding to epistasis and non-epistasis on the training set, and carry out statistical hypothesis testing on the hold-out test set. By splitting our data in this way, we introduce an additional layer of uncertainty — inference is defined relative to one of many possible sample splits. We address this extra layer of uncertainty through the PCS inference framework.

Specifically, we build on general PCS inference ideas to assess the strength of an epistatic effect. Intuitively, our inference framework captures both the uncertainty quantified by traditional statistical inference and the uncertainty surrounding data perturbations (sample split). In other words, the PCS p-values we propose consider not only distributional assumptions of a null model — as in traditional statistical inference — but also uncertainty surrounding the range of sample splits that could have been considered but were not. By "inflating" p-values to account for a wider range of uncertainty, our approach helps ensure that evidence for a given interaction is more robust to a broader range of modeling decisions. We call our approach epiTree test, because it is for **epi**stasis discovery and uses decision-**tree**s for both epistasis and no-epistasis models.

S1.1.1 Inference setting

For simplicity, we assume that a candidate interaction consists of two features: A and B. We provide details for the analogous case of general higher-order interactions in Section S1.1.9. Features A and B could be either continuous-valued gene expression features or discrete-valued SNP features. In the following, we focus on the gene expression setting. The case of SNP features is similar and described in Section S1.1.10. Recall that in our epiTree pipeline, candidate interactions are obtained by running iRF on the training data. The PCS epistasis inference for these candidate interactions, as outlined below, uses the same training data to obtain individual (for each interaction), tree-based null (no-epistasis) and alternative (epistasis) models. On the separate hold-out test data, our aim is to evaluate evidence (i.e. a p-value) for the null hypothesis

$$H_0 : (A, B) \text{ is not epistatic for phenotype } Y, \tag{1}$$

relative to the alternative hypothesis

$$H_1 : (A, B) \text{ is epistatic for phenotype } Y, \tag{2}$$

using test data consisting of n sample points $Y = (y_1, \ldots, y_n)^\top$, $A = (a_1, \ldots, a_n)^\top$, $B = (b_1, \ldots, b_n)^\top$. The individual response values y_i and features a_i, b_i may be continuous valued or discrete. Here, we consider the binary phenotype of red hair, $y_i \in \{0, 1\}$, where $y_i = 1$ indicates red hair, and continuous-valued gene expression features $a_i, b_i \in \mathbb{R}$.

S1.1.2 Recap of standard logistic regression

The classical translation of Fisher's original definition of epistasis for binary phenotypes seeks evidence against null H_0 relative to alternative hypothesis H_1 under a simple logistic regression model. That is, the data are assumed to have come from a logistic regression model, which is a strong assumption that often leads to unrealistic uncertainty assessments in practice, especially for large samples of data. Models for the null and alternative can be written as,

$$\begin{aligned} H_0 &: \text{logit}(P(y = 1|a, b)) = \beta_A a + \beta_B b + \beta_0, \\ H_1 &: \text{logit}(P(y = 1|a, b)) = \beta_A a + \beta_B b + \beta_{AB} ab + \beta_0, \end{aligned} \tag{3}$$

where β_j, $j \in \{A, B, AB, 0\}$ represent coefficients describing the contribution of A, B, the interaction term AB, and an intercept respectively. In some cases, correction for population structure is necessary. Employing standard chi-square approximations for the likelihood ratio test of H_0 vs. H_1, it is straightforward to obtain a p-value for the inference problem in (3). This (or some slight variation) is

the standard p-value typically reported as a measure of an interaction's significance, e.g., with p-value < 0.05/number of interactions tested, reported as significant (with Bonferroni correction).

S1.1.3 Scaling of the response variable

The logit scaling in (3) has a considerable influence on the way an interaction is defined. Clearly, a relation which is multiplicative (non-linear, hence epistatic) on one scale, can be additive (non-epistatic) in another. For instance, a log transform of a multiplicative interaction becomes additive. Hence, Fisher's commonly accepted definition of *epistasis* critically depends on the selected scaling. In fact, the situation is even more extreme. It is demonstrated in [9] that any multivariate, real valued function with compact support can be written as an additive function for some appropriate scaling. In other words, without fixing a response scale, every function can be rewritten in an additive form. Thus, Fisher's definition of epistasis is only well-defined relative to a selected scale.

The intrinsic scaling problem for epistasis has been pointed out by many authors, see e.g., [4, 5, 22, 24]. Out of statistical convenience, logistic regression models still dominate standard analyses. However, there is no biological justification for the logistic scaling. It has even been reported in empirical studies that this scaling does not always reflect biological function [22, 8].

Acknowledging that any epistatic results is only well-defined with respect to a specific scaling, it is natural (see e.g., discussion and references in [4]) to select a canonical scaling for this purpose: the penetrance $P(y = 1 \mid a, b)$ itself (instead of a logit scaling that transforms the penetrance via the function $f(x) = log(x/(1-x))$ as in logistic regression) [10, 5]. We articulate Fisher's definition in the penetrance scale as

$$H_0 : P(y = 1 \mid a, b) = f_A(a) + f_B(b) \quad \text{vs.} \quad H_1 : P(y = 1 \mid a, b) = f_{AB}(a, b), \tag{4}$$

where $f_j : \mathbb{R} \to [0, 1]$, $j \in \{A, B, AB\}$ are functions that describe the relationship between genes or interactions and penetrance.

S1.1.4 Form of additive and interaction models using training data

While raw gene expression data are often represented as counts (e.g. RNA-Sequencing), commonly available data are typically preprocessed. For example, PrediXcan estimates/imputes inverse normal transformed gene expression rather than raw expression value. In practice, there are a range of standard transforms and pre-processing steps, such that the actual scaling of gene features A and B can be rather arbitrary (see e.g., [1] for some general discussion). These transforms/pre-processing steps are generally not linear, although they are typically monotonic. As a result, we favor models that are invariant to monotone transformations of the data (resulting from potential pre-processing steps). In addition, we

want to allow more flexible mappings beyond linear (as in 3) to account for non-linearities that pervade biological systems. There are many different forms of non-linear functional relationships, which may each be useful in describing different epistatic behavior. A particularly flexible class of functions are decision trees, which have the benefit of being both simple to interpret and invariant to monotone feature transformations. In addition, as mentioned previously, it is well known that many biological processes exhibit thresholding dynamics that are mimicked in the Boolean rules used by decision trees [19, 14, 18, 16]. Moreover, for the classical epistasis model H_1 in (3) the multiplicative functional form lacks biological justification. Collectively, these considerations motivate our use of decision trees in both the additive, null model H_0 and non-additive, alternative model H_1. All individual decision trees are obtained via a CART [3] fit on the training data, using backfitting for the additive model. In summary, we translate Fisher's epistasis definition into a (non-parametric) hypothesis inference setting as

$$H_0 : P(y = 1 \mid a, b) = CART_A(a) + CART_B(b) \quad \text{vs.} \quad H_1 : P(y = 1 \mid a, b) = CART_{AB}(a, b). \tag{5}$$

Decision trees $CART_A, CART_B, CART_{AB}$ are fit on the training data set (that is, the same data that was used in the iRF candidate selection step) via the CART algorithm (using the R Package rpart) as follows: The complexity parameter (`cp_max` in `rpart`) (controlling the minimum improvement in the model needed at each node) for the interaction model $CART_{AB}$ is chosen adaptively, namely, small enough such that the interaction model included all genes (but not smaller than 10^{-3}). More precisely, we start with `cp_max` $= 0.01$ (which is the default parameter in the `rpart` implementation) and then, as long as not all features get split on in the tree, we replace the current `cp_max` value by `cp_max /1.1`. For the additive components, $CART_A$ and $CART_B$, the complexity parameter is chosen to be the same as for the interaction component. In this way it is guaranteed that the tree for the interaction model, $CART_{AB}$, is sufficiently deep to capture potential interaction behavior. Moreover, we prevent overfitting by not making `cp_max` smaller than necessary for all features to appear in the tree and generally not smaller than 10^{-3}. At the same time, this guarantees that the complexity of all trees, $CART_{AB}, CART_A, CART_B$, is comparable. To fit the additive regression model in H_0, we applied backfitting [11], where one recursively re-fitts additive components on the respective residuals with the remaining components. We stopped the recursion when none of the predicted values changed by more than 1% compared to the previous iteration. Finally, we conduct inference using a hold-out test set, following the prediction principle (see next section).

S1.1.5 Predictability principle from the PCS framework

Classical p-values, as obtained via logistic regression-tests for model (3), compare the goodness-of-fit for the null and the alternative hypothesis using all available data. In other words, the null and the

alternative models are fit on the same data that are used to evaluate the quality of each fit. In contrast, we learn null and alternative models on one (often randomly sampled) training data set and compare their prediction accuracy on a disjoint (often randomly sampled) set of test or hold-out data. This sample splitting approach is an old idea in statistics and has been used recently to deal with post-selection and so called universal inference problems, see, e.g., [29, 15, 2, 7, 30] and is widely used by the machine learning community to guard against overfitting.

Fitted null and alternative models correspond to hypotheses for non-epistasis (null, additive) and epistasis (alternative, non-additive) respectively, which leads to a simple hypothesis testing problem on the hold-out test data (since we fix both, the null and the alternative model from training data). We define predictions from the non-epistasis and epistasis models on the test data with gene expression (or SNP) values (a_i, b_i), for $i = 1, \ldots, n$, as:

$$p_0(a_i, b_i) = CART_A(a_i) + CART_B(b_i) \quad \text{vs.} \quad p_1(a_i, b_i) = CART_{AB}(a_i, b_i). \tag{6}$$

For simplicity, we denote $p_0, p_1 \in \mathbb{R}^n$ as the n-vectors with components $p_{0,i} := p_0(a_i, b_i)$ for and $p_{1,i} := p_1(a_i, b_i)$.

In principle, one could apply a Neyman-Pearson test (with likelihood ratio test statistic) in order to obtain a p-value (conditioned on p_0 and p_1). However, the models p_0 and p_1 will generally be (at least slightly) mis-specified. For large sample size this can lead to unrealistic extremely small p-values (e.g., in the red hair data example as small as 10^{-100}). This astronomically small p-value is due to the fact that the chi-squared asymptotic distributional approximation to the likelihood ratio test-statistics does not take into account model misspecification, which cannot be ignored for sufficiently large sample sizes (as in the UK biobank data).

In the following we use the PCS inference ideas proposed in [32] to explicitly take finite sample variability into account, via a bootstrapping approach on the test data, and to address the misspecification problem through prediction error evaluation on the test set. Instead of using the simple null hypothesis of the Neyman-Pearson test — the underlying penetrance is exactly p_0 — we use a more flexible null hypothesis. Specifically, we inflate the p_0 null distribution with the (centered) empirical sample distribution of the test statistic (via bootstrap sampling). As a result, our approach to inference takes into account the empirical variation in the likelihood ratio test statistic on the test data.

S1.1.6 epiTree test details: computation of CART based PCS p-value with bootstrap sampling of test data

Classically, p-values in a logistic regression model are calculated using a chi-squared approximation of the likelihood ratio test statistic under the null distribution, derived without accounting for model

misspecification. Our approach does not assume a classical probabilistic model. Instead, it generates a null hypothesis based on the PCS framework, where the null distribution is approximated directly via bootstrap samples of the test data. To conduct inference using this approach, we need to specify a test statistic and null distribution, which we detail in the following two paragraphs.

Test statistic. CART models (5) estimate penetrance for both the no-epistasis (null) and epistasis (alternative) hypotheses (using training data). Given this pair of fitted CART models, we estimate penetrance in the test data as p_0 (no-epistasis) and p_1 (epistasis) based on test set genotype features. Note that both, p_0 and p_1, are independent of the observed response Y in the test data used for inference. Given the observed test data (Y, A, B), we consider a canonical test statistic, the log-likelihood ratio

$$T(Y) = T(Y, A, B) = \ln\left(\frac{P(Y \mid p_0(A, B))}{P(Y \mid p_1(A, B))}\right), \tag{7}$$

with

$$P(Y|p_\cdot) = \prod_{i=1}^{n} p_{\cdot i}^{y_i}(1 - p_{\cdot i})^{1-y_i},$$

where $p_\cdot$ can be p_0 or p_1, and we have dropped dependence on A, B in our notation for simplicity. For given labels Y, $T(Y)$ measures how much more likely observed labels are under the null (p_0) compared to the alternative (p_1). Smaller values of $T(Y)$ correspond to greater evidence for the epistasis model.

PCS Prediction screening. Note that both, p_0 and p_1, were obtained from training data only. Evaluating $P(Y \mid p_0)$ and $P(Y \mid p_1)$ on hold-out test data provides a measure of how accurately the null and alternative models generalize to new samples. When $T(Y) \geq 0$, or equivalently $P(Y \mid p_0) \geq P(Y \mid p_1)$, then the alternative model p_1 (epistasis) provides no increase in prediction accuracy — as measured by the likelihood — on the test data compared to the null model p_0 (no epistasis). In this case, we conclude that there is not evidence for epistasis and epiTree formally reports a p-value of 1, that is

$$\text{PCS p-value} = 1 \quad \text{if } P(Y \mid p_0) \geq P(Y \mid p_1). \tag{8}$$

When $T(Y) < 1$, i.e. the epistasis model yields an increase in prediction accuracy, we quantify the uncertainty surrounding this improvement by pairing a simulated null data perturbation with bootstrap sub-sampling, as outlined below.

PCS Null perturbation. The PCS framework uses *null perturbations* to simulate data that respect known structure and compares observed results to results obtained from these simulated data. Recall that p_0 in (6) denotes our estimated penetrance for the test data under the hypothesis of no epistatic

	CART based PCS p-value	logistic regression p-value
Scale for response	penetrance scale $P(Y \mid A, B)$	Logit scale: $\log\left(\frac{P(Y \mid A,B)}{1-P(Y \mid A,B)}\right)$
Null model (no epistasis)	$CART(A) + CART(B)$	$\beta_A A + \beta_B B + \beta_0$
Alternativ model (epistasis)	$CART(A, B)$	$\beta_A A + \beta_B B + \beta_{AB} AB + \beta_0$
Data used for model fitting	(separate) training data	full data
P-value calculation	Bootstrap from null perturbation	Distributional approximation (chi-square)

Table S1: High-level summary of comparison between p-value from logistic regression with polynomial model vs. PCS p-value with CART model for epistasis testing.

interaction. We simulate responses under the no epistasis null hypothesis as

$$Y_0 \sim \text{Bernoulli}(p_0), \tag{9}$$

which we refer to as the null perturbation. Below, we outline our approach for constructing a reference null distribution of the test statistic.

PCS p-value To obtain a PCS p-value, we use bootstrap re-sampling over the test data to evaluate

$$\text{PCS p-value} = P_{\text{bootstrap}}(T(Y) > T(Y_0)), \tag{10}$$

where $P_{\text{bootstrap}}$ indicates randomness relative to bootstrap sampling of the test data. More precisely, we generate bootstrap samples $m = 1, \ldots, M$, each of size n (= sample size of the test data) drawn i.i.d. with replacement from the empirical distribution of our observed test data. For each bootstrap sample we obtain $p_0|m \in [0, 1]^n$, a vector of estimated penetrance under the null. Using $p_0|m$, we simulate null responses for bootstrap sample m under the null perturbation (9) as

$$Y_0|m \sim \text{Bernoulli}(p_0|m).$$

In addition to the null responses above, we obtain $Y|m \in \{0, 1\}^n$, an n-vector consisting of the observed responses for the mth bootstrap sample. The PCS p-value is then given by

$$\text{PCS p-value} = \frac{1}{M} \sum_{m=1}^{M} \mathbb{1}\{T(Y|m) > T(Y_0|m)\}. \tag{11}$$

In practice, one can obtain a simple analytic approximation of (11) as outlined in Section S1.1.8. By construction, Y and Y_0 have the same distribution under the null hypothesis, making our proposed PCS inference a valid p-value (conditioned on p_0). We summarize the general test procedure epiTree in terms of the computation of the CART based PCS p-values in Figure 7 of the main text. Table S1 shows a high level comparison of the CART based PCS p-values and p-values obtained from logistic regression.

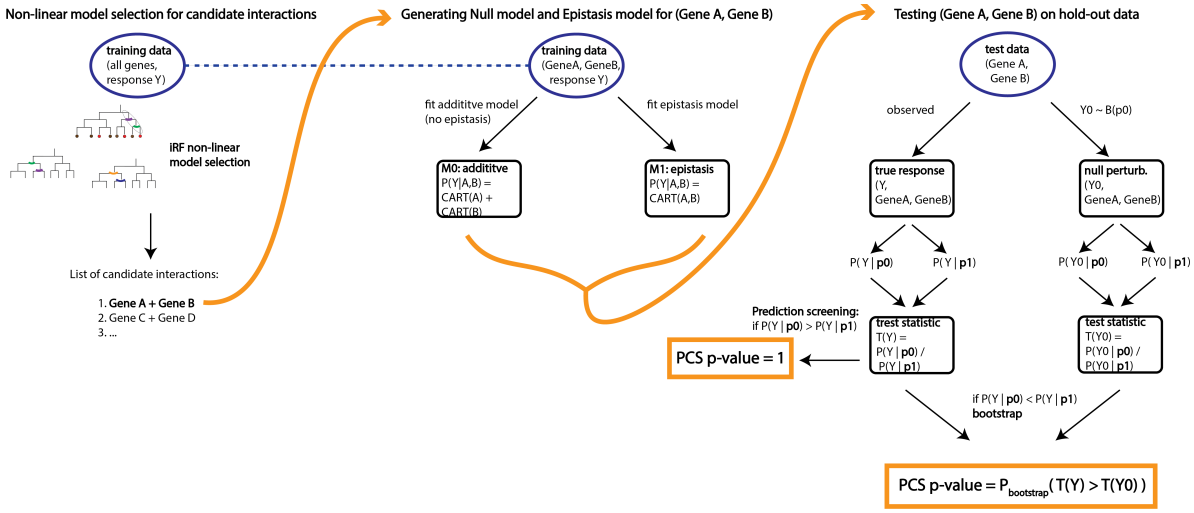

Figure S1: Illustration of epiTree pipeline.

S1.1.7 Comparison between PCS p-value and classical p-value calculations

In the following, we provide insights into the major differences between calculations of a PCS p-value and a classical p-value. For a given test statistic $T(Y)$, null hypothesis $H_0 : Y \sim F_0$ with null distribution F_0, and alternative hypothesis H_1 where smaller values of T correspond to stronger evidence against the null (as e.g., for T as in (7) for the epiTree test), a classical p-value is given by

$$\text{p-value } = P_{F_0}(T(Y_0) < t), \tag{12}$$

where $t = T(Y)$ is the observed test statistic and the probability is taken over the randomness of Y_0 with distribution F_0. The classical p-value in 12 does not take into account empirical fluctuation of $T(Y)$ among different subsamples of data — randomness only enters through Y_0.

In contrast, the PCS p-value explicitly takes the empirical variability of the test data (Y, A, B) into account through bootstrap sampling, as in (11). This is necessary because sample splitting leads to a particular test set. Bootstrap samples of the test set mimic the other possible test sets that could have been resulted from a different sample split. More precisely, let $\hat{F}_n$ denote the empirical distribution of the test data. As the number of bootstrap samples M in (11) tends to infinity,

$$\text{PCS p-value } = P_{F_0 \times \hat{F}_n}(T(Y_0) < T(Y)). \tag{13}$$

Thus, asymptotically, when the bootstrap distribution (empirical distribution) converges to the true underlying distribution of Y, denoted as F, the PCS p-value is equivalent to

$$\text{PCS p-value } \approx P_{F_0 \times F}(T(Y_0) < T(Y)), \tag{14}$$

where Y and Y_0 are independent. For the given observed value t, as in (12), we can rewrite (14) to obtain

$$\text{PCS p-value } \approx P_{F_0 \times F}(T(Y_0) + (t - T(Y)) < t), \tag{15}$$

where t is fixed and the probability is taken as in (14) jointly over (Y_0, Y) with distribution $F_0 \times F$. Note that the difference between (12) and (15) is that $T(Y_0)$ is replaced by $T(Y_0) + (t - T(Y))$. This means that the effective null distribution of the PCS p-value corresponds to a convolution of the proposed null distribution, $T(Y_0)$, and a centered version of the observed distribution $T(Y)$. In particular, (15) does not just take into account the distribution of the probabilistic null model F_0, but also the observed sample distribution of Y. Hence the PCS p-value is more stable with respect to the observed data and possibly also model mis-specifications. This centered, convoluted null distribution also prevents artificially small p-values that are often obtained in the classical setting (12) with large n and slightly missspecified null distributions.

For the CART based PCS p-value, as in the epiTree pipeline, the test statistic $T(Y)$ as in (7) corresponds to the log-likelihood ratio statistic for the simple null hypothesis $y_i \sim Bernoulli(p_{0,i})$ against the simple alternative $y_i \sim Bernoulli(p_{1,i})$. A classical p-value in this setting corresponds to a Neyman-Pearson test for these simple null and alternative hypothesis. In particular, for the classical Neyman-Pearson p-value the variance of the test statistic $T(Y)$ under the null distribution $H_0 : Y \sim F_0$ is completely independent of the actual observed responses Y. In contrast, the PCS p-value considers the *inflated* null distribution of $T(Y_0) - T(Y)$, with $Y_0 \sim F_0$ and $Y \sim \hat{F}_n$. Hence, the PCS p-value does not just incorporate the distributional variance of F_0 but also the observed empirical variance of the response Y. We stress that the general concept of the PCS p-value does not depend on the particular choice of statistic $T(Y)$ in (7). As a result, we can calculate PCS p-values not just for the CART based model that we employed for the epiTree pipeline, but also for any other hypothesis testing problem, with an arbitrary test statistic $T(Y)$ and null distribution F_0.

In the supplemental Section S1.5, we provide a detailed toy example of a simple linear regression model without intercept, where analytic forms of a classical p-value, as in (12), and PCS p-value, as in (14), can easily be derived analytically. Although this setup is overly simplistic, it provides concrete analytical insights into situations when a PCS p-value is beneficial compared with the a classical p-value. As shown in Section S1.5, the classical p-value has higher power when the data are exactly generated from the hypothetical alternative distribution H_1. In this case, the PCS p-value's lower detection power originates from the additional uncertainty quantification in the bootstrap sub-sampling — PCS p-value explicitly takes the empirical variation of the observed data sample into account. However, in arguably most real data applications (and certainly for epistasis testing, as stressed throughout this paper), the data generating process cannot be specified exactly for either the null or the alternative model. In such

settings, the PCS p-value can be more robust towards such misspecification. Specifically, we show (both, analytically and in simulations) that classical p-values can result in severe false positive rates when the data are generated from a slight variation of the hypothesised null distribution. On the other hand, the PCS p-value is generally more robust and, in contrast to the classical p-value, does not result in an increased type 1 error. Moreover, we also provide an explicit example in Section S1.5 where the data are generated from a slight variation of the hypothesised alternative distribution. In this case, we observe that the PCS p-value can have a higher detection power compared to the classical p-value. This originates form the fact that the PCS p-value explores the full empirical distribution of the observed responses, and is therefore able to detect deviations from the hypothesized null distribution of the statistic $T(Y)$, which typically cannot be detected from the a single observed t, as in the classical p-value. In summary, while the standard p-values work well in settings where everything is specified correctly, when models are misspecified PCS p-values can be favorable – they provide more stable type 1 error control and can have a higher detection power in the presence of outliers.

S1.1.8 Analytic approximation of PCS p-value in the epiTree test

In order to obtain an analytic approximation of (11) note that

$$\begin{aligned} T(Y) - T(Y_0) &= \log\left(\frac{P(Y|p_0)}{P(Y|p_1)}\right) - \log\left(\frac{P(Y_0|p_0)}{P(Y_0|p_1)}\right) \\ &= \sum_{i=1}^{n} (y_i - y_{0,i})(\log(p_{1,i}) - \log(1-p_{1,i}) - \log(p_{0,i}) + \log(1-p_{0,i})) \\ &= \sum_{i=1}^{n} \delta_i, \end{aligned}$$

where $p_{1,i}$ and $p_{0,i}$ denote the ith component of the vectors p_1 and p_0, respectively, and $y_{0,i}$ denotes the ith component of the n-vector Y_0, and

$$\delta_i := (y_i - y_{0,i})(\log(p_{1,i}) - \log(1-p_{1,i}) - \log(p_{0,i}) + \log(1-p_{0,i})). \tag{16}$$

For each bootstrap sample m, we draw n i.i.d. indexes $I(m)_1, \ldots, I(m)_n$ uniformly from the set $\{1, \ldots, n\}$. Thus we can write

$$\mathbb{1}\{T(Y|m) > T(Y_0|m)\} = \mathbb{1}\left\{\sum_{i=1}^{n} \delta_{I(m)_i} > 0\right\}. \tag{17}$$

Conditioned on the data (with the only randomness coming from the bootstrap sampling via the random indexes $I(m)_1, \ldots, I(m)_n$), the M different terms in equation (11) are independent and identically distributed Bernoulli random variables. Hence, by the law of large numbers and using equation (17), we

get that for infinitely many bootstrap samples ($M \to \infty$)

$$\text{PCS p-value} = P_{(I_1,\ldots,I_n)}\left(\sum_{i=1}^{n} \delta_{I_i} > 0\right), \tag{18}$$

where randomness is over the n i.i.d. random indexes $I_1, \ldots, I_n$ drawn uniformly at random from the set $\{1, \ldots, n\}$. Again, conditioned on the data, we have that the n random variables $\delta_{I_1}, \ldots, \delta_{I_n}$ are independent and identically distributed, each following a uniform distribution on the set $\{\delta_1, \ldots, \delta_n\}$. Thus, they have mean

$$E_{I_1}(\delta_{I_1}) = \frac{1}{n}\sum_{i=1}^{n} \delta_i =: \mu \tag{19}$$

and variance

$$Var(\delta_{I_1}) = E_{I_1}((\delta_{I_1} - \mu)^2) = \frac{1}{n}\sum_{i=1}^{n}(\delta_i - \mu)^2 =: \sigma^2, \tag{20}$$

where $E_{I_1}(\cdot)$ mean that we take the expectation w.r.t. the random bootstrap sample. Hence, it follows from the central limit theorem that $X := \sqrt{n}(\frac{1}{n}\sum_{i=1}^{n} \delta_{I_i} - \mu)/\sigma$ converges in distribution to a standard Gaussian as $n \to \infty$ (in our case $n = 4$K). Thus,

$$\text{PCS p-value} = P(X > -\sqrt{n}\mu/\sigma), \tag{21}$$

$$\text{with } |P(X > -\sqrt{n}\mu/\sigma) - \Phi(\sqrt{n}\mu/\sigma)| \to 0, \text{ as } n \to \infty, \tag{22}$$

where Φ denotes the cumulative distribution function of the standard normal distribution. Note that it follows from (21) that PCS p-value < 0.5 if any only if $\mu < 0$, which is equivalent to $T(Y) < T(Y_0)$. Thus the PCS p-value can only be significant when the improvement in prediction (of the epistasis model over the non-epistasis model) for the observed data are greater than for the null perturbation. Further, note one can upper bound the rate of convergence in (21) using the nonuniform Berry–Esseen theorem:

Theorem 1. *[21, Theorem 3] Let $Z_1, \ldots, Z_n$ be i.i.d. random variables with $E(Z_1) = 0$, $Var(Z_1) = 1$, and $c_3 := E(|Z_1|^3) < \infty$. Then there exists some constant L such that for all $z \geq 0$*

$$\left|P\left(\sum_{i=1}^{n} Z_i > z\sqrt{n}\right) - (1 - \Phi(z))\right| \leq \frac{Lc_3}{\sqrt{n}(1 + |z|^3)}.$$

Therefore, from (18) and Theorem 1 we get that when $\mu < 0$

$$|\text{PCS p-value} - \Phi(\sqrt{n}\mu/\sigma)| = \left|P\left(\sum_{i=1}^{n}(\delta_{I_i} - \mu)/\sigma > n(-\mu/\sigma)\right) - (1 - \Phi(\sqrt{n}(-\mu)/\sigma))\right| \leq \frac{C}{n^2}, \tag{23}$$

with $C = \frac{L \cdot c_3}{|\mu/\sigma|^3} < \infty$ and $c_3 = \frac{1}{n} \cdot \sum_{i=1}^{n} |\delta_i - \mu|^3$. Thus, the approximation error from the CLT for the PCS p-value decreases of order n^{-2}. In our case $n = 4,000$, i.e., n^{-2} is of order 10^{-7}. We note, however, that one also has to take the exact value of C into account (which depends on the data).

Averaging over null perturbation For the PCS p-value approximation above, we conditioned on both the observed responses Y as well as the (simulated) null perturbation Y_0 (see definition of δ_i in (16)). In practice, the reported PCS p-value should be independent of any randomness introduced by Y_0. Below, we detail two different ways to achieve this:

1. Average the PCS p-value over infinitely many random realizations of Y_0. In other words, when computing the expectation and standard deviation of δ_{I_i} in equation (19) and (20), we average over the randomness of Y_0, which is equivalent to replacing the $y_{0,i}$ term with its expectation $p_{0,i}$. This gives

$$E_{I_1,Y_0}(\delta_{I_1}) = \frac{1}{n}\sum_{i=1}^{n} E_{y_{i,0}}(\delta_i) = \frac{1}{n}\sum_{i=1}^{n} \overline{\delta}_i =: \overline{\mu},$$

with

$$\overline{\delta}_i = (y_i - p_{0,i})(\log(p_{1,i}) - \log(1 - p_{1,i}) - \log(p_{0,i}) + \log(1 - p_{0,i})), \tag{24}$$

and

$$Var(\delta_{I_1}) = E_{I_1,Y_0}((\delta_{I_1} - \overline{\mu})^2) = \frac{1}{n}\sum_{i=1}^{n} E_{y_{i,0}}((\delta_i - \overline{\mu})^2) = \frac{1}{n}\sum_{i=1}^{n} \left((\overline{\delta}_i - \overline{\mu})^2 + \omega_i^2 p_{i,0}(1 - p_{i,0})\right) = \overline{\sigma}^2,$$

with

$$\omega_i := \log(p_{1,i}) - \log(1 - p_{1,i}) - \log(p_{0,i}) + \log(1 - p_{0,i}),$$

where $E_{I_1,Y_0}(\cdot)$ means that we take the expectation w.r.t. the random bootstrap index I_1 and the random null perturbation Y_0. The PCS p-value approximation is then given by $\Phi(\sqrt{n}\overline{\mu}/\overline{\sigma})$.

2. Alternatively, for a given bootstrap sample m in (17), instead of comparing $T(Y|m)$ to $T(Y_0|m)$ for some particular random realization of $Y_0|m$, one can compare $T(Y|m)$ to the average realization of $T(Y_0|m)$, that is, $E_{Y_0}(T(Y_0|m))$. This is equivalent to replacing δ_i by $\overline{\delta}_i$ from (24) when computing the mean μ and standard deviation σ in (19) and (20). Note that the only different between δ_i and $\overline{\delta}_i$ is that the random quantity $y_{0,i}$ gets replaced by its expectation $p_{0,i}$. In particular, conditioned on the observed data (Y, A, B), the quantity δ_i is random, as it depends on the random null perturbation $y_{0,i}$, and $\overline{\delta}_i$ is not random.

Note that the only difference between the two different approaches is in the variance term σ^2, where the former approach results in a slightly larger variance via the additional $\frac{1}{n}\sum_{i=1}^{n}(\omega_i^2 p_{i,0}(1 - p_{i,0}))$ term.

For the red hair analysis, we found that both approaches give the same magnitude for the p-values. The p-values presented in the results section for the red hair analysis correspond to the latter approximation.

S1.1.9 Higher order interactions

In the following, we discuss how to extend inference from pairwise to higher-order interactions. Inference for an interaction in the logistic regression model is based on the respective coefficients in a polynomial interaction model. Thus, for higher-order interactions one typically adds higher-order polynomial terms to the model. For example, to test an order-three interaction among genes A, B, C, the logistic regression hypothesis problem in equation (3) translates to

$$H_0 : \text{logit}(P(y = 1 \mid a, b, c)) = \beta_A a + \beta_B b + \beta_C c + \beta_{BC} bc + \beta_{AC} ac + \beta_{AB} ab + \beta_0, \tag{25}$$

$$H_1 : \text{logit}(P(y = 1 \mid a, b, c)) = \beta_A a + \beta_B b + \beta_C c + \beta_{BC} bc + \beta_{AC} ac + \beta_{AB} ab + \beta_{ABC} abc + \beta_0. \tag{26}$$

If we replace the polynomial terms in (25) with CART interaction terms and replace the logistic scale by the penetrance scale, we obtain

$$\begin{aligned} H_0 : P(y = 1 \mid a, b, c) &= CART_A(a) + CART_B(b) + CART_C(c) \\ &\quad + CART_{BC}(b, c) + CART_{AC}(a, c) + CART_{AB}(a, b) \\ H_1 : P(y = 1 \mid a, b, c) &= CART_A(a) + CART_B(b) + CART_C(c) \\ &\quad + CART_{BC}(b, c) + CART_{AC}(a, c) + CART_{AB}(a, b) + CART_{ABC}(a, b, c) \end{aligned} \tag{27}$$

Note, however, that in (27), without loss of generality, the lower order CART terms can be incorporated into the respective higher-order CART terms[1]. Hence, for the CART model the direct comparison to (25) is to rewrite equation (5) as

$$\begin{aligned} H_0 : P(y = 1 \mid a, b, c) &= CART_{AB}(a, b) + CART_{BC}(b, c) + CART_{AC}(a, c), \\ H_1 : P(y = 1 \mid a, b, c) &= CART_{ABC}(a, b, c). \end{aligned} \tag{28}$$

More generally, the null model H_0 to test for an order-d interaction can be written as an additive model of d CARTs each taking a subset of size $d - 1$ and the alternative epistasis model can be written as a single CART taking all the d features as input.

From a biological perspective, it is not clear whether H_0 in (28) — where each of the genes interact with every other gene — should be interpreted as no d^{th} order epistasis. Across the subcellular, cellular, tissue and organismal levels, biological processes are often found to be subdivided into multiple com-

[1] To see this, note that any additive CART model of the form $CART_A(a) + CART_{AB}(a, b)$ can be rewritten as a single CART term $\tilde{CART}_{AB}(a, b)$, where $\tilde{CART}_{AB}(a, b)$ is the same as $CART_{AB}(a, b)$ but with additional splits at the tip nodes that correspond to the splits of $CART_A(a)$. The same holds true for higher order interaction CART models.

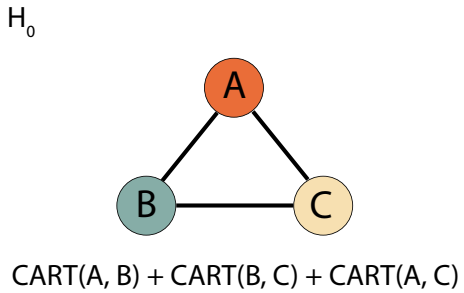

Figure S2: Illustration of the null model H_0 in (28).

ponent proteins, such as an enzyme complex or signaling cascade [23]. Considering the genetic basis of human phenotypes, the linkage of multiple genetic variants spanning protein subunits but influencing a specific pathway is an increasingly common finding [6, 26, 12]. Modeling the relationship between all d components in H_0 of (28) yields the opportunity to more directly capture the redundancy and non-linear behavior of biological pathways and protein signaling which characterizes observations of epistasis in complex organisms [25, 27]. To illustrate this, consider the relationship between the different features as a graph, where the nodes of the graph correspond to the individual features and edges between features correspond to pairwise interactions. For the null model (no d^{th} order epistasis) H_0 in (28), each pair of features is connected via an edge in the graph, as illustrated in Figure S2. From the perspective of a biological pathways, S2 implies that all genes belong to a shared pathway, which arguably represents a dth order biological interaction one may be interested in detecting.

By framing inference w.r.t. to prediction, the epiTree p-values have greater flexibility derived from learning null and alternative models on separate training data. In particular, for the epiTree p-value null model, instead of considering a null model as in (28) and Figure S2, we can learn a partition of the tested interaction features, such that features between different groups do not interact. More precisely, we consider all partitions of the features into two disjoint sets (corresponding to two unconnected graphs), such that features within each set can interact, but features between the two sets can not. The final null model H_0 is then selected via a prediction screening, i.e., the partition with the highest prediction accuracy on the test data is selected as the null model H_0 which is tested against the interaction model H_1 of d-th order epistasis. More precisely, we consider the CART null and alternative model to test a 3rd-order interaction as follows

$$\begin{aligned} &H_0 : P(y = 1 \mid a, b, c) = CART_1(I_1) + CART_2(I_2), \text{ with } I_1 \cup I_2 = \{a, b, c\}, I_1, I_2 \neq \emptyset, \text{ and } I_1 \cap I_2 = \emptyset, \\ &H_1 : P(y = 1 \mid a, b, c) = CART_{ABC}(a, b, c). \end{aligned} \tag{29}$$

We illustrate this case of 3rd-order epistasis null model H_0 in Figure S3 and 4th-order epistasis null

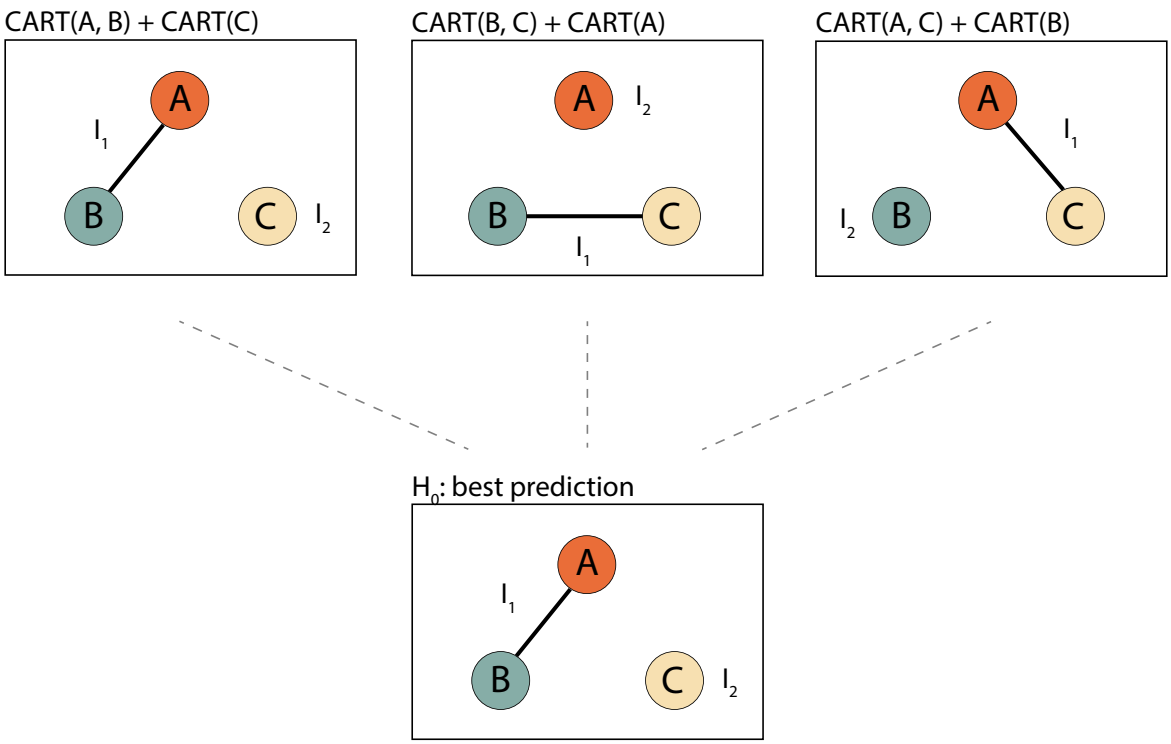

Figure S3: Illustration of the null model H_0 in (29), as it is considered for the epiTree p-value.

model H_0 in Table S2. Similarly, for a dth-order interaction with features $A_1, \ldots, A_d$, we consider

$$H_0 : P(y = 1 \mid a_1, \ldots, a_d) = CART_1(I_1) + CART_2(I_2),$$

$$\text{with } I_1 \cup I_2 = \{a_1, \ldots, a_d\}, I_1, I_2 \neq \emptyset, \text{ and } I_1 \cap I_2 = \emptyset,$$

$$H_1 : P(y = 1 \mid a_1, \ldots, a_d) = CART_{A_1 \ldots A_d}(a_1, \ldots, a_d),$$

where I_1 and I_2 correspond to the partition which yield the highest prediction accuracy on the test data.[2] Finally, note that in the case of pairwise interactions (i.e. $S = \{A, B\}$) the definitions (28) and (29) are the same.

[2] We evaluate prediction accurarcy analog as in the prediction screening step (8) via the likelihood.

Candidates for the 4th order no-epistasis H_0 model	Final 4th order no-epistasis null model H_0	Alternative 4th order epistasis model H_1
$M_1 : CART_A(a) + CART_{BCD}(b, c, d)$ $M_2 : CART_{A,B}(a, b) + CART_{CD}(c, d)$ $M_3 : CART_{A,B,C}(a, b, c) + CART_D(d)$ $M_4 : CART_B(b) + CART_{ACD}(a, c, d)$ $M_5 : CART_{BC}(b, c)) + CART_{AD}(a, d)$ $M_6 : CART_{BD}(b, d) + CART_{AC}(a, c)$ $M_7 : CART_C(c) + CART_{ABD}(a, b, d)$ $M_8 : CART_D(d) + CART_{ABC}(a, b, c)$	$H_0 : M_i$ with $P(\text{test data}\|M_i) > P(\text{test data}\|M_j)$ for all $j \neq i$	$H_1 : CART_{A,B,C,D}(a, b, c, d)$

Table S2: Illustration of the null model H_0 in (29), but for a 4th order interaction.

S1.1.10 SNP features

Due to the discreteness of SNP features, there is no need to consider different scaling functions f_A, f_B, f_{AB} as in equation (4). This is because any (bijective) scaling will simply map the three different possible SNP values $\{0, 1, 2\}$ to some other distinct values. Thus, equation (3) becomes

$$\begin{aligned} H_0 : \text{logit}(P(y = 1 \mid a, b)) =& \beta_0 + \sum_{i=1}^{2} \beta_{Ai} \mathbb{1}_{a=i} + \sum_{j=1}^{2} \beta_{Bj} \mathbb{1}_{b=j} \\ H_1 : \text{logit}(P(y = 1 \mid a, b)) =& \beta_0 + \sum_{i=1}^{2} \beta_{Ai} \mathbb{1}_{a=i} + \sum_{j=1}^{2} \beta_{Bj} \mathbb{1}_{b=j} + \sum_{i=1}^{2} \sum_{j=1}^{2} \beta_{ABij} \mathbb{1}_{a=i,b=j} \end{aligned}$$

and equation (5) becomes

$$\begin{aligned} H_0 : \text{logit}(P(y = 1 \mid a, b)) =& \beta_0 + \sum_{i=1}^{2} \beta_{Ai} \mathbb{1}_{a=i} + \sum_{j=1}^{2} \beta_{Bj} \mathbb{1}_{b=j} \\ H_1 : \text{logit}(P(y = 1 \mid a, b)) =& \beta_0 + \sum_{i=1}^{2} \beta_{Ai} \mathbb{1}_{a=i} + \sum_{j=1}^{2} \beta_{Bj} \mathbb{1}_{b=j} + \sum_{i=1}^{2} \sum_{j=1}^{2} \beta_{ABij} \mathbb{1}_{a=i,b=j} \end{aligned}$$

For more than two genes the situation is analog as in Section S1.1.9, where the null model allows for interactions up to order $K-1$, when an interaction of order K gets tested. That means, for an interaction of order K the alternative model (epistasis) has 3^K parameters and the null model (no-epistasis) has $3^K - 2^K$ parameters.

S1.2 Further discussion on comparison between CART models and logistic regression for specific red hair interactions

In the results section of the main text we investigated the difference between CART based models and logistic regression on the response surfaces of the null (no-epistasis) and alternative (epistasis) models for two exemplary cases: *ASIP - TUBB3* and *ASIP-DEF8*. Here we provide a similar discussion for the other three pairwise interactions with significant p-values from either PCS or logistic regression.

The response surface for a putative interaction among *ASIP - DBNDD1* is shown in Figure S7. Here, the PCS p-value ($= 10^{-11}$) is much smaller than the logistic regression p-value ($= 10^{-1}$). However, the prediction error for the CART alternative model (epistasis) (cross entropy of 0.528) is slightly worse than the prediction error of the logistic regression null model (no-epistasis) (cross entropy of 0.516). Looking at the response surface, we find that the thresholding behavior observed for many of the other interactions (recall the results section in the main text) is less present for these features. As a consequence

the smooth response surface of the additive logistic regression model provides a similar data fit as the non-additive CART interaction model, such that overall the interaction behavior is less clear in this case and would require further investigation on the SNP level. Figure S8 shows the response surface for *ASIP - VS9D1*. Here, the data show a clear nonlinear thresholding behavior, which is reflected in the smaller prediction error of the CART models (cross entropy at around 0.5) compared to the logistic regression model (cross entropy at around 0.7). The final response surface shown in Figure S9 considers an interaction among *ASIP - GAS8*. Here, the logistic regression p-value ($= 10^{-7}$) is much smaller than the PCS p-value. However, all four models (CART null and alternative, logistic regression null and alternative) have a similar prediction error (cross entropy at 0.589, 0.596, 0.589, 0.585, respectively). From this, we conclude that there is less evidence for epistasis in this case, as the additive CART model can explain the data just as well as a multiplicative interaction term.

S1.3 An alternative approach to imputed gene expression dimension reduction via sub-batching

The extremely high dimensionality of the SNP features prohibited us from running iRF on the SNP data directly. Therefore, in the epiTree pipeline we followed a biologically inspired dimension reduction step via imputed gene expression and then first searched for interactions on the gene level, before going back to the SNP level data. In the following, we describe an alternative approach which we also explored.

We performed an initial screen for important variants across sub-batches of SNPs. First, we partitioned the full dataset into batches of 10,000 SNPs and fit a RF to predict the red hair phenotype from each batch. We used the RF feature importance – mean decrease in Gini impurity (MDI) – to rank SNPs and fitted separate iRFs using the top: 100, 500, 1000, and 2500 SNPs. This strategy parallels marginal screening approaches, which filter a large set of variants based on associations between a single gene and the target response. However, by using RF feature importance to screen SNPs, we maintain the possibility of including interacting variants with weak main effects since decision paths reflect the importance of a feature conditioned on previous splits in the path.

In principle, one advantage of the sub-batch over gene expression screening by iRF is that important SNPs that do not influence gene expression are more likely to pass the initial screening procedure. That is, SNPs that are not important in the PrediXcan model have a better chance of being picked up by iRF. However, we found that regions with particularly strong effects (like MC1R for red hair) tend to mask SNPs from regions with weaker effects. Aggregating SNPs at the gene level gives weaker SNPs a higher chance to pass the initial screen. Interpreting interactions directly on the gene level also has practical advantages. Follow up experiments are much easier to be performed on the gene level then on the SNP level. Therefore, we focused on the gene-expression approach in our analysis.

S1.4 Details on parameter choices for iRF, ranger, and penalized logistic regression

In the following, we provide further specific details on the choice of implementation and software parameters made in our analysis of the red hair phenotype for UK Biobank data with the epiTree pipeline.

Penalized logistic regression: The penalized logistic regression model was fit with the R package `glmnet`, using default parameters for classification. The tuning parameter `lambda` was selected via cross validation using the `cv.glmnet` function of the `glmnet` R package.

Random forest: The RF model was implemented in the R package `ranger` with default parameters for classification. We note that our iRF model was also fit using `ranger`. As a result, differences between iRF and RF can be directly attributed to iterative feature re-weighting, which is a form of "soft" regularization (see more details below).

iRF: For the iRF model, for $k = 10, 50, 100, 500, 1000, 2500$, we ran the first iteration with default parameters for classification. We then applied hard thresholding on the top k features, with k selected by minimizing out-of-bag error, followed by three iterations of iterative re-weighting (i.e. soft-thresholding). For the gene level analysis $k = 50$, while the subsequent SNP level analysis used $k = 1,000$. After the final iteration, we then searched for interactions using RIT, where we grew many shallow trees, in order to obtain a large set of candidate interactions ($ntree = 5,000, depth = 3, child = 5$). To evaluate stability of candidate interactions, we performed $n.bootstrap = 50$ bootstrap replicates and only included candidate interactions that were consistently filtered in at least 50% of the replicates.

S1.5 Illustrative example of PCS p-value for simple linear regression with one predictor

Here we compare classical and PCS p-values using a toy example: a linear model with a single feature or predictor variable, given unit variance, and no intercept. In this simple setting, one can easily derive closed form solutions for the classical p-value and the PCS p-value. We find that in situations where the data is generated exactly from the hypothesized alternative model, the PCS p-value has reasonable detection power, but the classical p-value is superior with fewer false negatives than the PCS p-value. However, when models are misspecified, the PCS p-value is more robust, leading to fewer false positives under a misspecified null model and fewer false negatives under a misspecified alternative model. In most practical situations the hypothesised models are misspecified to some extent, especially in the big data era. This suggests that the PCS p-value is favorable for most real data applications.

We consider a test data set $\{(y_1, x_1), \ldots, (y_n, x_n)\}$ with response values $y_i \in \mathbb{R}$ and a single, fixed

covariate $x_i \in \mathbb{R}$, along with a separate training data set $\{(\tilde{y}_1, \tilde{x}_1), \ldots, (\tilde{y}_n, \tilde{x}_n)\}$, where we have assumed for simplicity that the training and test data are of the same size. Further, to ease notation, in the following we assume that $\sum_{i=1}^n x_i^2 = \sum_{i=1}^n \tilde{x}_i^2 = 1$. Our goal is to obtain and compare classical and PCS p-values for the hypothesis testing problem

$$H_0 : \; y_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1) \quad \text{vs.} \quad H_1 : \; y_i \overset{ind.}{\sim} \mathcal{N}(c\, x_i, 1) \text{ for some } c > 0. \tag{30}$$

Test statistic: We consider the test statistic as in a classical *z-test*, that is

$$T(Y) = \sum_{i=1}^n x_i y_i, \tag{31}$$

with $Y = (y_1, \ldots, y_n)^\top$. The classical p-value from the *z-test* is given by

$$\text{p-value } = \Phi\left(-T(Y)\right) = \Phi\left(-\sum_{i=1}^n x_i y_i\right), \tag{32}$$

where, again, Φ denotes the cumulative distribution function of the standard normal distribution. In the following examples we compare this classical p-value to the respective PCS p-value. For the PCS p-value we first obtain an estimate for both, the hypothesis model and the alternative model, from the training data. For the alternative H_1 we obtain an estimate for the coefficient c from the training data $\{(\tilde{y}_i, \tilde{x}_i), i = 1, \ldots, n\}$ as

$$\hat{c} = \sum_{i=1}^n \tilde{x}_i \tilde{y}_i. \tag{33}$$

For the null hypothesis H_0 there are no unknown parameters that need to be estimated.

PCS Prediction screening: In the prediction screening step, we use the test data to evaluate the prediction error from the null model and the alternative model, where here we estimate the alternative model using the training data as in (33). For the loss function, we consider the negative log-likelihood, which equals the squared loss in the Gaussian case (30). More precisely, the prediction screening as in (8) yields

$$\text{PCS p-value } = 1 \text{ if } \sum_{i=1}^n y_i^2 \leq \sum_{i=1}^n (y_i - \hat{c} x_i)^2 \quad \Leftrightarrow \quad \sum_{i=1}^n x_i y_i \leq \frac{1}{2}\hat{c}. \tag{34}$$

PCS Null perturbation: When the prediction screening does not result in a failure to reject the null hypothesis, we simulate from the null perturbation. In this simple setting, the null hypothesis

corresponds to i.i.d. Gaussian white noise. Therefore, we perturb our true responses y_i as follows:

$$Y_0 = (y_{0,1}, \ldots, y_{0,n})^\top \sim \mathcal{N}((0, \ldots, 0)^\top, I_{n\times n}), \tag{35}$$

where $I_{n\times n}$ denotes the $n \times n$ identity matrix.

PCS p-value: We obtain a PCS p-value with test statistic $T(Y)$ as in (31) and null perturbation Y_0 as in (35) via bootstrap samples as in (11). Note that in this case we can approximate the PCS p-value as in (14) such that

$$\text{PCS p-value } \approx P\left(\sum_{i=1}^{n} x_i\,(y_i - y_{0,i}) < 0\right). \tag{36}$$

S1.5.1 Simulation studies

In the following simulations, we compare the classical p-value as in (32) and the PCS p-value as in (36) under different probabilistic generating models for the responses y_i (and $\tilde{y}_i$, respectively). We note that the PCS p-value requires a separate training data set, which is not required by the classical p-value. In order to provide a fair comparison in all of our simulations studies , we use all available data as test data for the classical p-value and for the PCS p-value we randomly split all available data into a training and a test data sets of equal size. Below, we summarize our results. More details, as well as further examples, can be found in a supplementary R Markdown file, available at `https://github.com/merlebehr/epiTree/example_pcs_pvalues_linear_regression.R`.

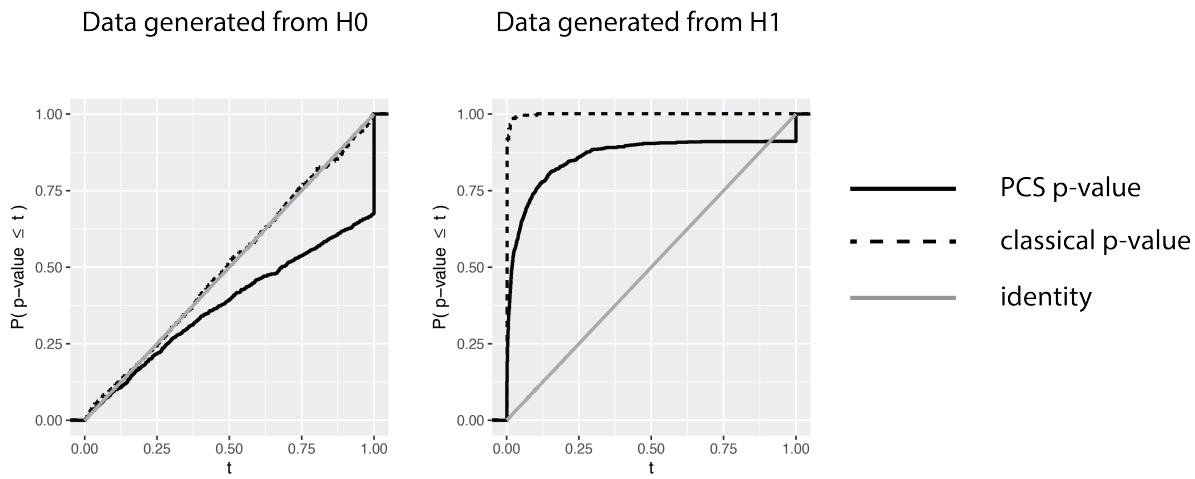

Figure S4: Distribution of classical p-value and PCS p-value for the hypothesis testing problem in (30) with test statistic as in (31) for $n = 1,000$ and a single dependent variables x_i drawn independently from a uniform distribution on $[0, 1]$. The black solid line corresponds to the PCS p-value, the black dotted lines to the classical p-value, and the gray solid line to the identity. Results are obtained from $1,000$ Monte Carlo runs. Left: responses y_i as in H_0 in (30); Right: responses y_i as in H_1 in (30) with $c = 4$.

Classical setting – correctly specified model. Here we compare PCS and classical p-values under the correctly specified model where data are generated exactly as indicated by (i) the null hypothesis H_0 and (ii) the alternative hypothesis H_1. In these settings, the classical p-value outperforms the PCS p-value. That is, the classical p-value has higher power under the alternative and maintains the correct control over type I error. Of course, we note that in real data settings the "true" model is rarely known and thus this performance is not expected in practice.

First, we consider the setting where the responses are generated from the H_0 distribution as in (30) with $y_i \sim \mathcal{N}(0,1)$ independent for all $i = 1, \ldots, n$. As seen in the left plot in Figure S4, both, the classical p-value and the PCS p-value follow a sub-uniform distribution, indicated by the fact that their cumulative distribution function (cdf) is smaller or equal to the identity line. The PCS p-value is a bit more conservative than the classical p-value under the exact null distribution, indicated by the fact that the cdf of the PCS p-value is smaller or equal to the cdf of the classical p-value. However, for smaller quantiles (that are of dominant interest in practice) the different between the PCS p-value and the classical p-value distributions is almost negligible under the exact null distribution. For example, in our Monte Carlo simulations we found that the PCS p-value is smaller than 0.05 in exactly 5% of cases, which coincides with the classical p-value.

Second, we consider the situation where the responses follow a model as in H_1 in (30) with $y_i \sim \mathcal{N}(c\, x_i, 1)$ independent for all $i = 1, \ldots, n$, for some fixed $c > 0$. The right plot in Figure S4 shows the respective p-value distributions for $c = 4$. In this case, we find that the classical p-value outperforms the PCS p-value, although the PCS p-value has high power. For example, at a nominal level of $\alpha = 0.05$ the classical p-value correctly rejects the null hypothesis in 99% of cases, but the PCS p-value rejects in only 64% of cases. Similar, at nominal level $\alpha = 0.1$ the classical p-value rejects in 99.7% of cases, but the PCS p-value in only 74.9% of cases. Intuitively, the reason for this is that exploring the full distribution of the test statistic $T(Y)$ via the bootstrap sampling results in an additional variance term, which leads to a weaker detection power when the observed responses exactly follow the specified alternative model H_1.

Misspecified model Here we compare PCS and classical p-values when data are generated under a model that deviates slightly from the null H_0 via a misspecified variance. That is, we assume that the variance of observations Y_i in (30) is not 1 but $\sigma^2 > 1$. In this setting, we argue that it is preferable *not* to reject H_0. We find that the PCS p-value is robust to this model misspecification and behaves largely the same way as in the correctly specified setting. In contrast, the classical p-value is highly sensitive to this model misspecification, even in the simple linear model considered in our simulations.

We let responses follow the null distribution H_0 but with a larger variance $\sigma^2 \gg 1$. That is, $y_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$. The right plot in Figure S5 shows the respective p-value distributions when $\sigma = 10$ (results for

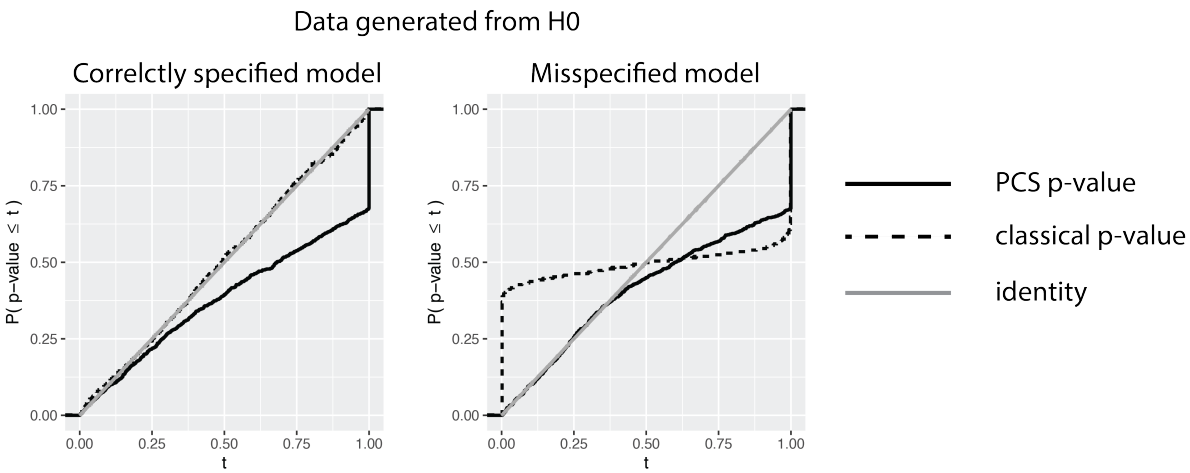

Figure S5: Same as Figure S4, but with the right plot such that responses $y_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$ with $\sigma = 10$.

different values of σ are shown in the R markdown file at `https://github.com/merlebehr/epiTree/example_pcs_pvalues_linear_regression.R`). In this case, we find that the classical p-value yields severe false positives due to the model misspecification. For example, at a nominal level of $\alpha = 0.05$ the classical p-value falsely rejects the null hypothesis in 42% of cases. In contrast, the PCS p-value, via the prediction screening as well as the bootstrap sampling, is robust to this misspecification. E.g., at nominal level $\alpha = 0.05$ the PCS p-value falsely rejects the null hypothesis in only 5% of cases, as is expected under the null. Similar, at nominal level $\alpha = 0.01$ the PCS p-value rejects in %1 of cases, but the classical p-value still rejects in as much as 40% of cases. We stress that even when the null model is extremely misspecified with a very large σ remarkably the PCS p-value is robust to this with no increase in the type 1 error (see the supplementary R markdown file at `https://github.com/merlebehr/epiTree/example_pcs_pvalues_linear_regression.R`).

In summary, in our simple simulation set-up, the PCS p-value is found to be more robust to a misspecified data generating process as it is almost always the case in practice, even though, as expected, the classical p-value is found to work better than the PCS p-value when the data generative models are specified correctly. In particular, the use of PCS p-values results in fewer false positives under the null. Moreover, as detailed in the supplementary R markdown file at `https://github.com/merlebehr/epiTree/example_pcs_pvalues_linear_regression.R`, the PCS p-value can result in fewer false negatives when the alternative is misspecified. Of course, our simulation set-up is very simple. We nevertheless believe that the lessons learned hold in general. It is our current research to provide more extensive evidence to support our belief that in practice the PCS p-value should be preferred over the classical p-value, unless the very precise forms of H_0 and H_1 are carefully backed up by domain knowledge and quantitative empirical evidence.

S1.6 Comparison between PCS and other inference methods

We compared the performance of epiTree (PCS epistasis p-values) against (i) logistic regression-based inference and (ii) permutation random forests (pRF) [17]. Logistic regression has been widely used to estimate epistasis (e.g. [13, 20, 28, 31]), but is formulated to detect multiplicative interactions after controlling for linear main effects. Permutation random forests (pRF) leverage random forests, as epiTree does, to detect interactions of arbitrary form. However, pRF is designed to detect the presence of pairwise, rather than higher-order interactions. We note that an extension of pRF to high-order interactions could be formulated in the context of PCS inference, with permutations serving as the null perturbation, but we do not explore this extension here.

We evaluated the performance of epiTree and pRF under four models with responses generated from the gene expression data used in our red hair case study (subset to the 50 genes with highest MDI importance). For each model, we generated Bernoulli responses from sets of active interactions S_k, $k = 1, 2$ (described below), identified candidate interactions using iRF, and computed p-values from each method for all candidate interactions. We note that each method uses a different inference technique for computing p-values (logistic regression: classical, epiTree: PCS, pRF: permutation-based). The probability of success π for Bernoulli responses in the four models was given, respectively, by:

$$
\begin{aligned}
\pi^{(AND)} &= \pi_A \left(\prod_{j \in S_1} \mathbb{1}(x_j > t_j) \right) + \pi_I \left(1 - \prod_{j \in S_1} \mathbb{1}(x_j > t_j) \right), \\
\pi^{(OR)} &= \pi_A \left(\max_{k=1,2} \prod_{j \in S_k} \mathbb{1}(x_j > t_j) \right) + \pi_I \left(1 - \max_{k=1,2} \prod_{j \in S_k} \mathbb{1}(x_j > t_j) \right), \\
\pi^{(ADD)} &= \pi_A \left(\prod_{j \in S_1} \mathbb{1}(x_j > t_j) + \prod_{j \in S_2} \mathbb{1}(x_j > t_j) \right) \\
&+ \pi_I \cdot \mathbb{1} \left\{ \left(\prod_{j \in S_1} \mathbb{1}(x_j > t_j) + \prod_{j \in S_2} \mathbb{1}(x_j > t_j) \right) = 0 \right\}, \\
\pi^{(MULT)} &= \pi_A \cdot \sigma \left(\prod_{j \in S_k} x_j \right),
\end{aligned}
$$

where σ denotes the sigmoid function and π_A, π_I baseline success probabilities for active and inactive samples respectively. We set $\pi_A = 0.75$ and $\pi_I = 0.25 \cdot \{\text{proportion of active samples}\}$ for all models. We selected features in active interactions S_k using an iterative sampling strategy to prevent high levels of correlation between interacting features. Specifically, the first feature was sampled with uniform probability and subsequent features with probability inversely proportional to their maximum absolute correlation over previously selected features. For all models, we set the size of each interaction component to $|S_k| = 3, k = 1, 2$ (i.e. an order-3 interaction).

We note that the different response generating models represent a range of interaction forms. The AND model ($\pi^{(AND)}$) corresponds to a single interaction that is either active or inactive for all observations. The OR model ($\pi^{(OR)}$) corresponds to two distinct interactions that can be viewed as heterogeneous mechanisms governing response behavior. The ADD model ($\pi^{(ADD)}$) corresponds to two interactions terms that do not interact with one another. Finally, the multiplicative model ($\pi^{(MULT)}$) corresponds to a multiplactive interactions.

For each model above, we generated a list of candidate interactions using iRF and computed p-values using logistic regression, pRF, and epiTree. To assess the variability in p-values, we repeated the inference step 25 times on new responses generated from the same set of active interactions. Figure S20 reports the results from each response generating model. We report the distribution of $-\log_{10}(p)$ for the following groups: (i) active interactions — the true response generating interaction (i.e. $S = S_1$ or $S = S_2$) (ii) active interaction subsets — a strict subset of the true response generating interaction (i.e. $S \subset S_k$) (iii) active features — features used in active interactions but not all part of the same interaction term (i.e. $S \subseteq S_1 \cup S_2, S \not\subseteq S_1, S \not\subseteq S_2$) (iv) inactive interaction — includes features that do not appear in any interaction term (i.e. $S \cap (S_1 \cup S_2) \neq \emptyset$). We note that the resolution of p-values for pRF depends on the number of permutations run, which we set to 1000 in our simulations. For consistent comparison across all methods, we threshold reported p-values below at 10^{-4}. Below we summarize several key findings.

- **AND/OR models:** All methods consistently detect (at significance level 0.05) active interactions. epiTree and logistic regression also detects subsets of active interactions while pRF generally does not. We find that active interactions fall at the top of the p-value ranked list for epiTree and pRF but not for logistic regression. In other words, the interactions deemed most important (with respect to p-value) by epiTree and pRF included only features that were truly interacting.

- **AND model:** pRF consistently detects false positives in the AND model (average p-value of inactive interactions ~ 0.025). We find that this is due to the fact that pRF p-values are generally low for interactions that contain both an active interaction and other non-interacting features. In contrast, epiTree and LR rarely detect these false positive interactions. The average p-values for epiTree and LR are consistently > 0.25.

- **ADD model:** epiTree is the only method that consistently detects active interactions (and their subsets) in the ADD model. We find that active interactions are at the top of the p-value ranked list for epiTree in the ADD model but not for pRF or logistic regression.

- **MULT model:** Logistic regression is the only method that consistently detects subsets of multiplicative interactions. We note that no method computes p-values for the active interaction since it was not detected by iRF, and thus not contained in the set of candidate interactions. While

epiTree does not detect subsets of the active multiplicative interaction at a significance level of 0.05, we find that such subsets still fall at the top of the p-value ranked list for epiTree.

In summary, epiTree performs comparably to or better than competing methods under the AND, OR, and ADD models. Specifically, epiTree consistently recovers true interactions with generally lower false positive rates than pRF or logistic regression. While logistic regression slightly outperforms epiTree under the multiplicative model, all methods perform poorly in this setting.

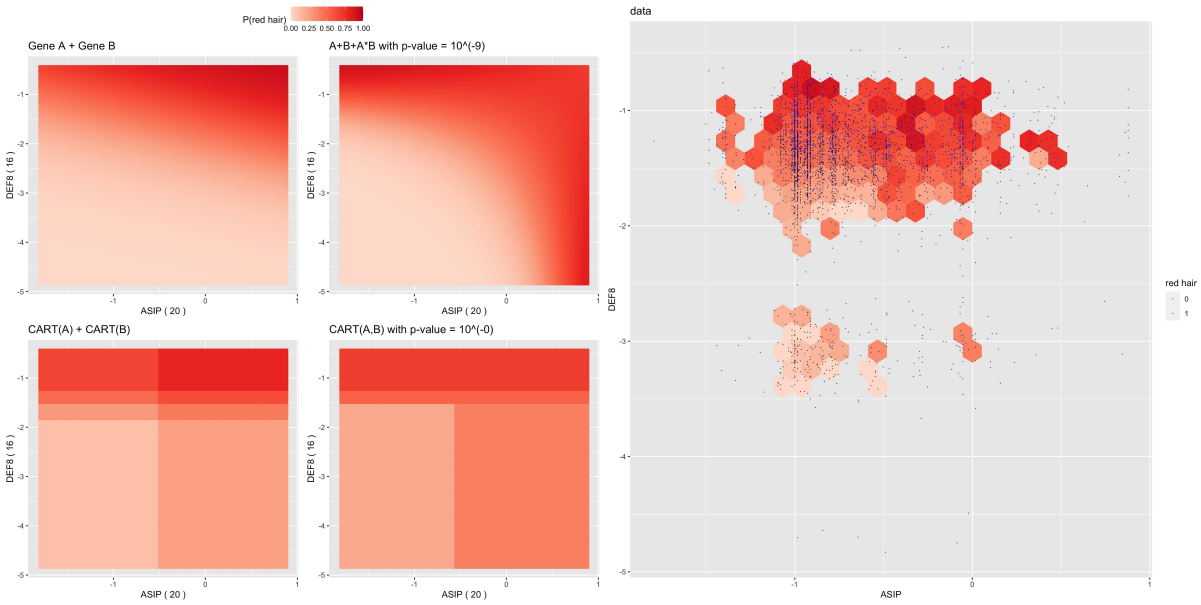

Figure S6: Response surface for *ASIP* - *DEF8*, otherwise as Figure 6

S2 Supplemental Figures

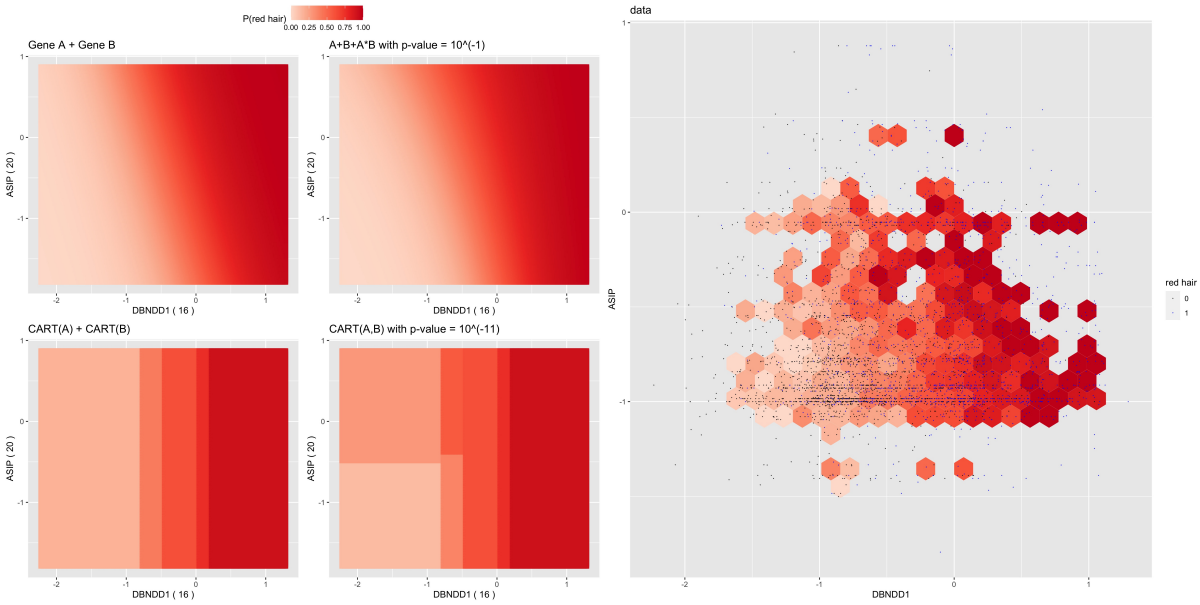

Figure S7: Response surface for *ASIP* - *DBNDD1*, otherwise as Figure 6

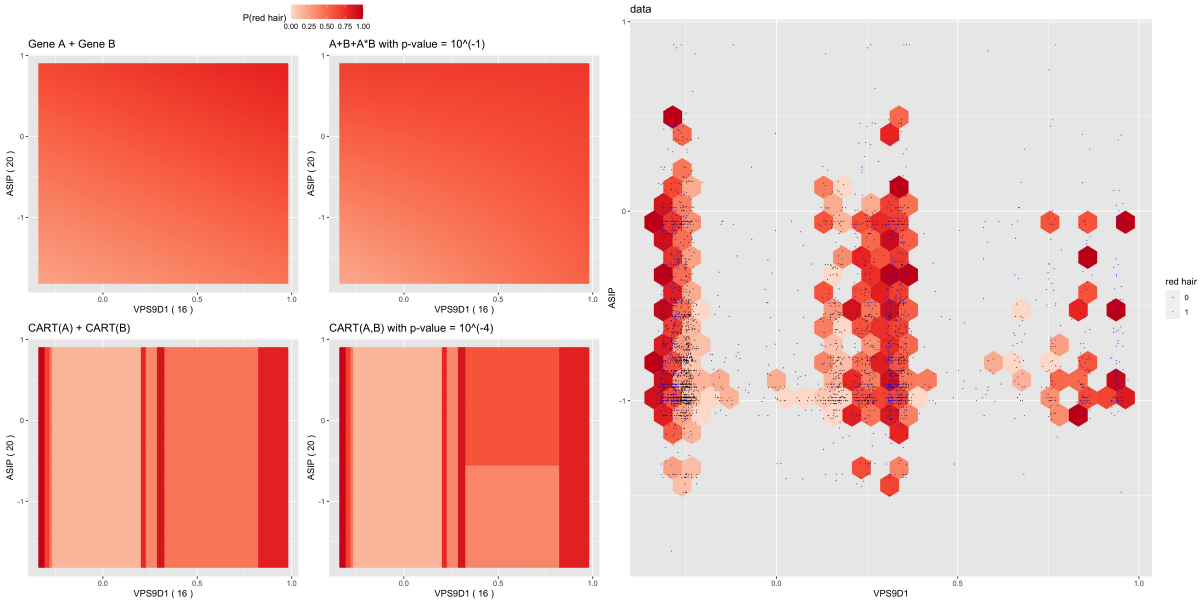

Figure S8: Response surface for *ASIP* - *VPS9D1*, otherwise as Figure 6

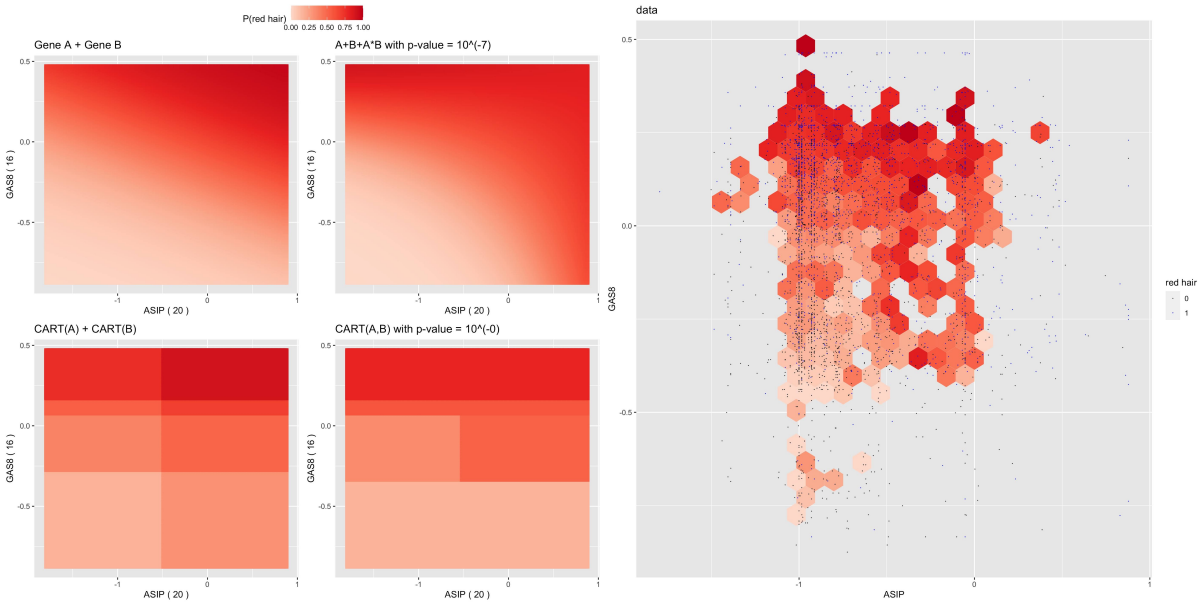

Figure S9: Response surface for *ASIP* - *GAS8*, otherwise as Figure 6

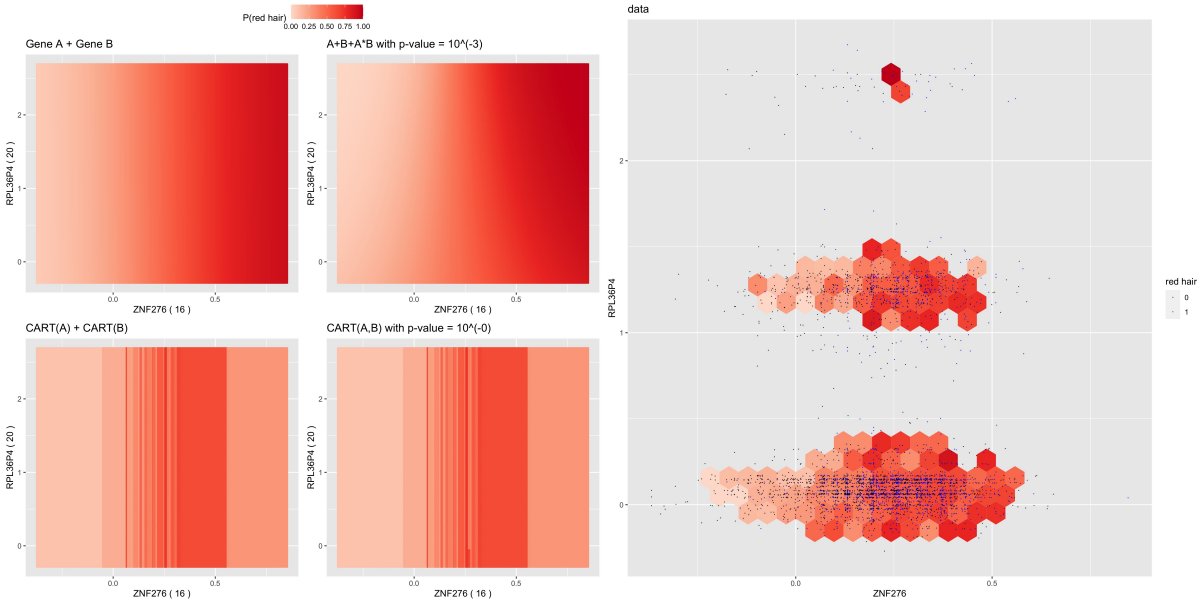

Figure S10: Response surface for *ZNF276* - *RPL36P4*, otherwise as Figure 6

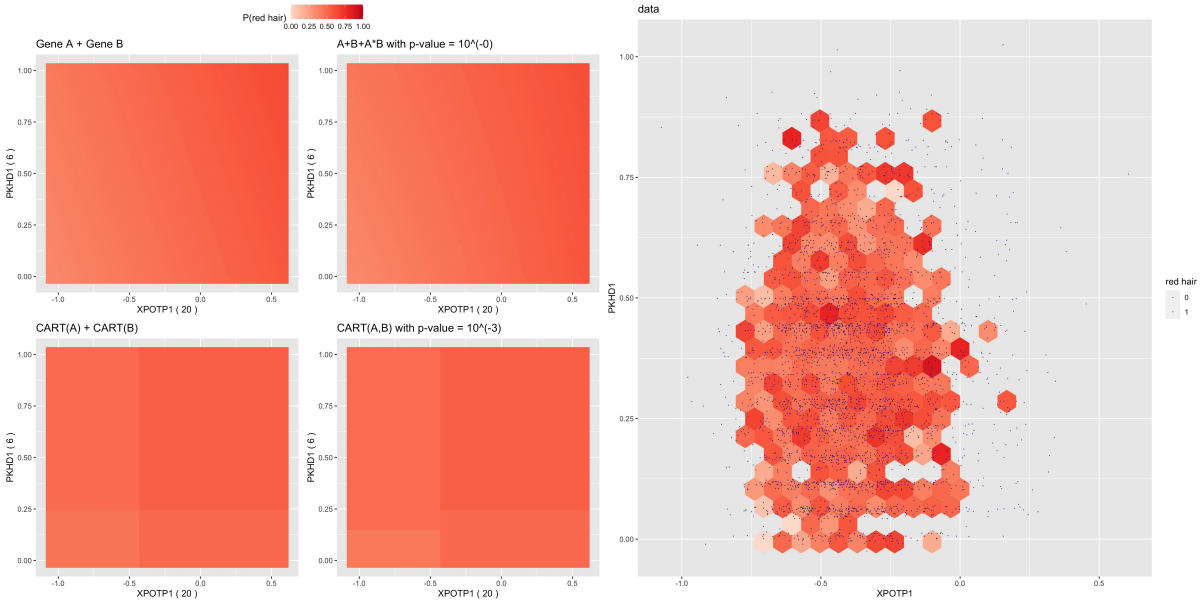

Figure S11: Response surface for *XPOTP1* - *PKHD1*, otherwise as Figure 6

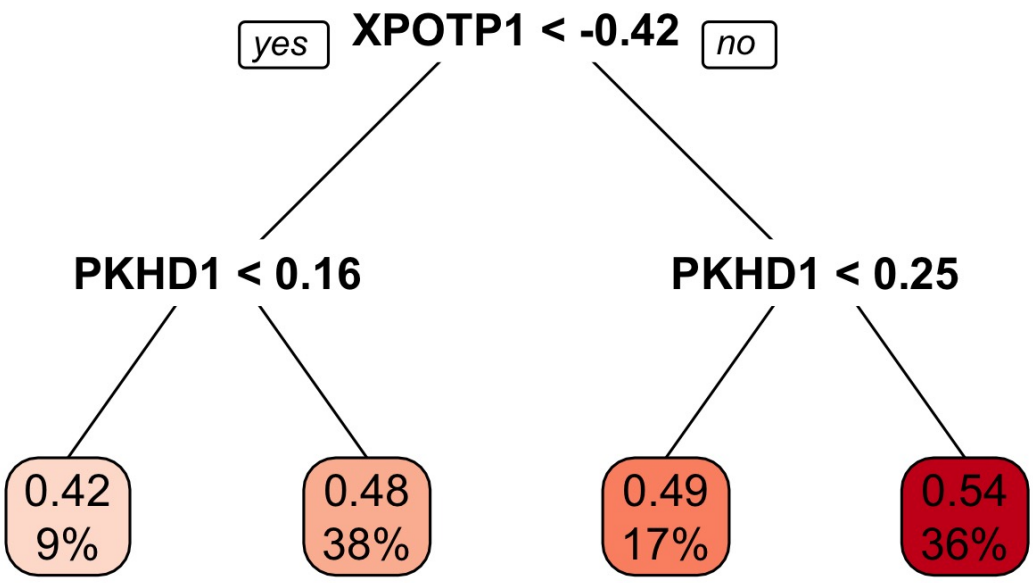

Figure S12: Decision trees for *XPOTP1* - *PKHD1* interaction. The decimal digits at the tip nodes correspond to the predicted probability of red hair. The percentage at the tip nodes corresponds to the percentage of training observations falling into this tip node.

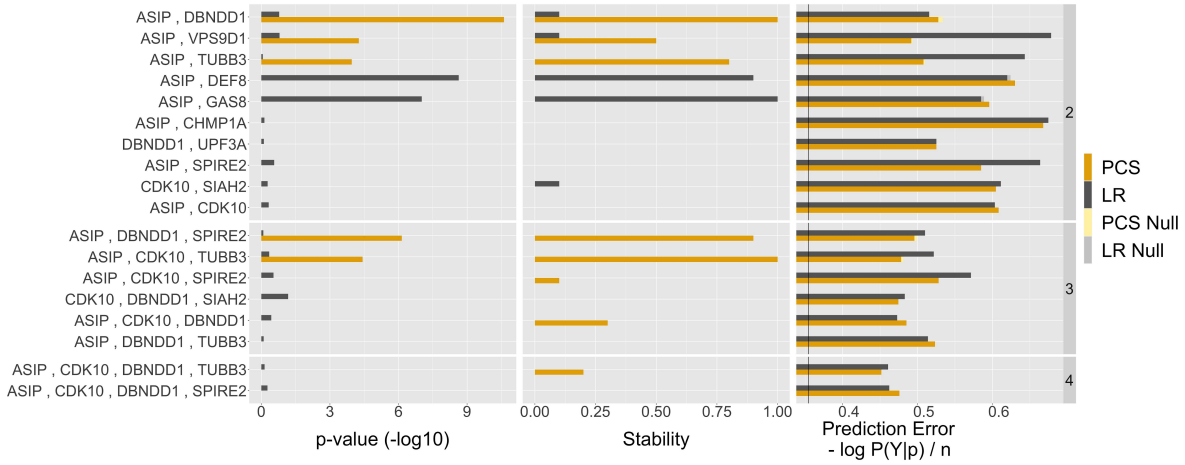

Figure S13: List of stable gene level interactions found by iRF (stability score > 0.5). The first column shows the PCS p-value (orange) and p-value from logistic regression (gray) on a -log10 scale. The second column shows percentage of bootstrap replicates of significant p-values (significant level 5% with Bonferroni correction). The third column shows the prediction error (cross-entropy) on the test data of the learned CART (orange) and logistic regression (gray) models for both, no-epistasis (light color) and epistasis (dark color). The black vertical line shows the prediction error achieved by iRF using all the gene features simultaneously.

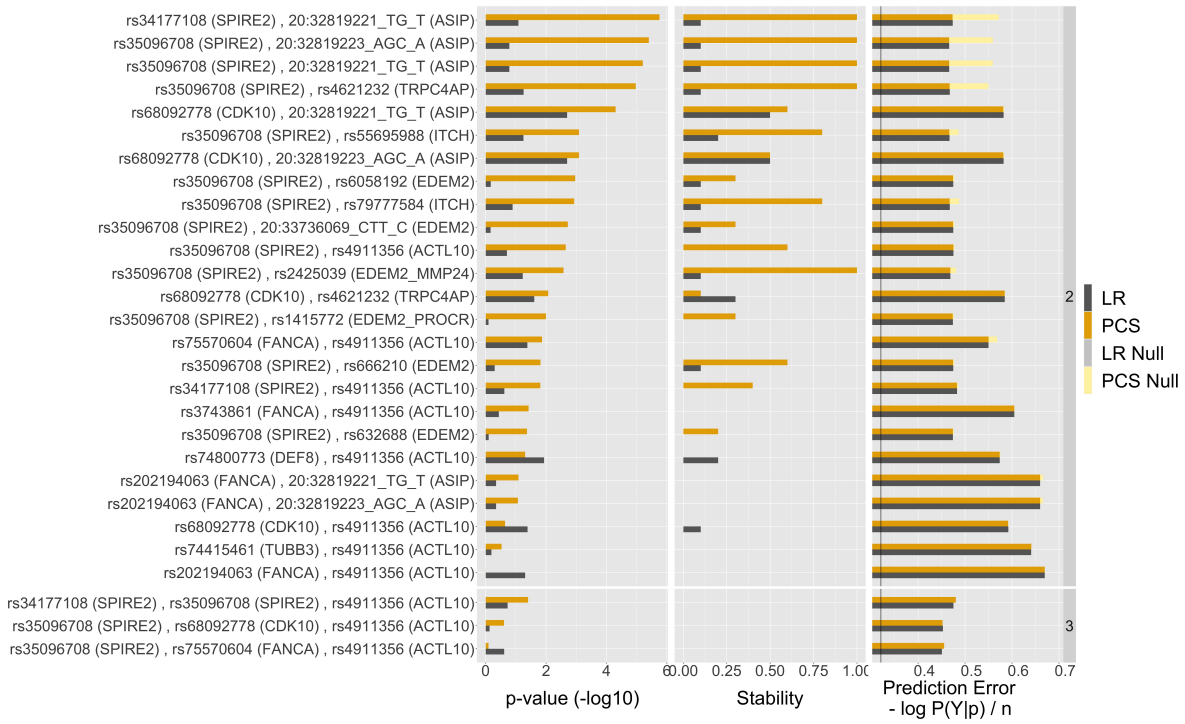

Figure S14: Same as Figure S13, but for the variant level.

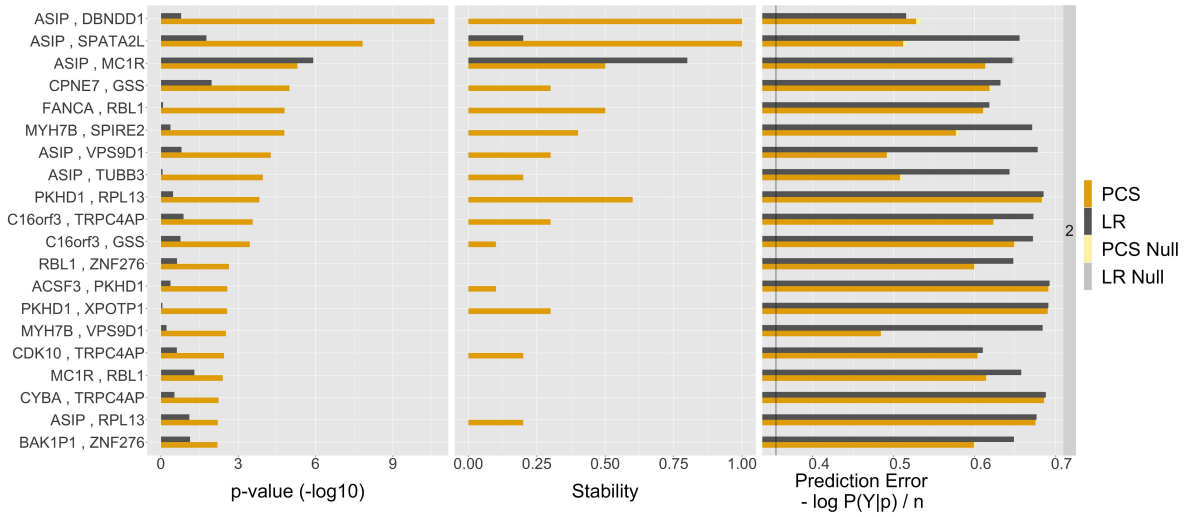

Figure S15: Same as Figure S13, but for the top 20 PCS p-values among inter chromosome pairwise brute force search of top 50 iRF genes (in terms of Gini importance).

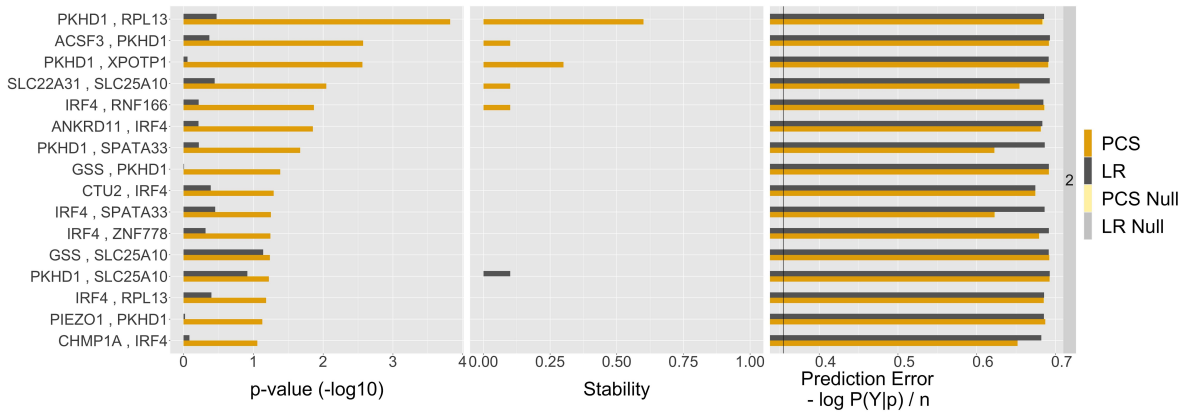

Figure S16: Same as Figure S15 but restricted those interactions which are not between chromosome 16 and chromosome 20 genes.

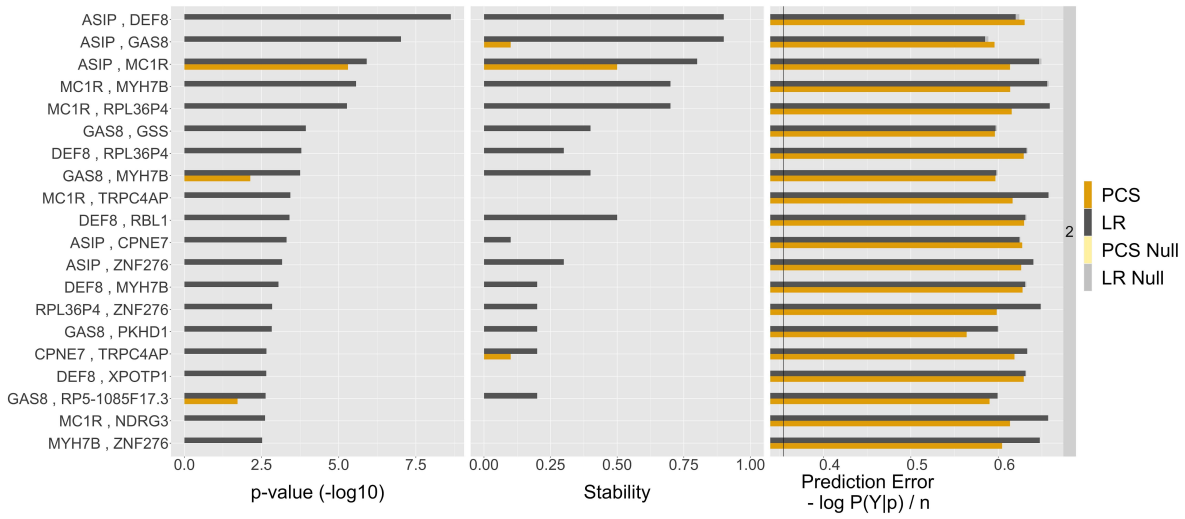

Figure S17: Same as Figure S15 but for top p-values from logistic regression.

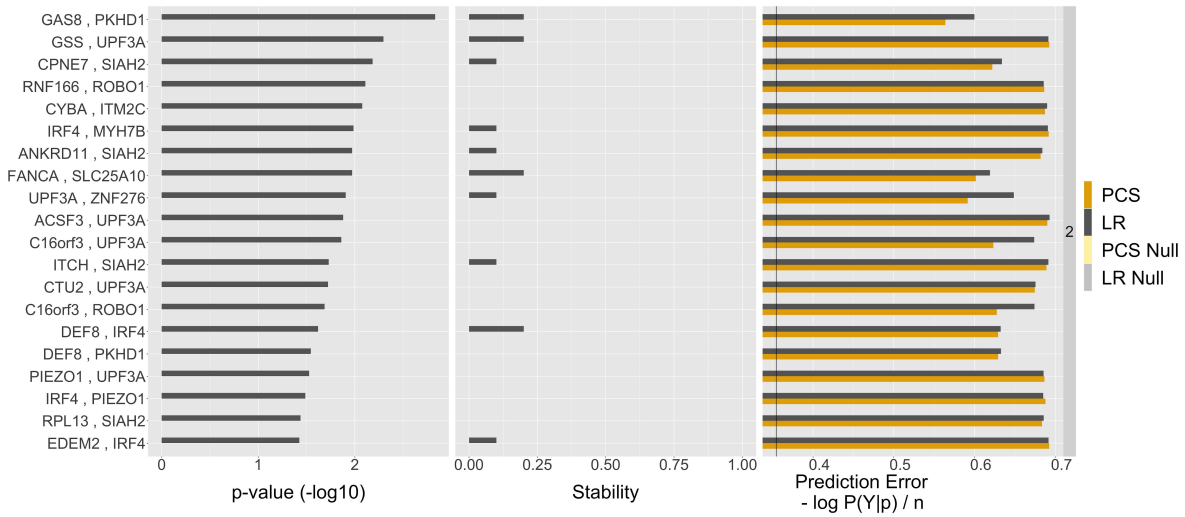

Figure S18: Same as Figure S16 but for top p-values from logistic regression.

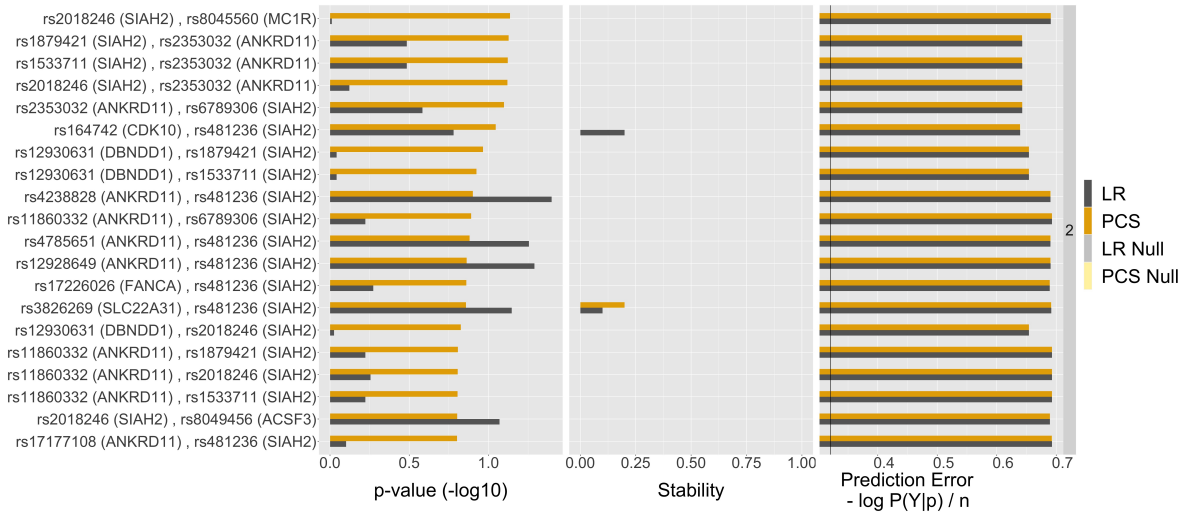

Figure S19: Same as Figure S13, but for the top 20 PCS p-value among all pairs of variants that enter the PrediXcan model for *CDK10* and *SIAH2*.

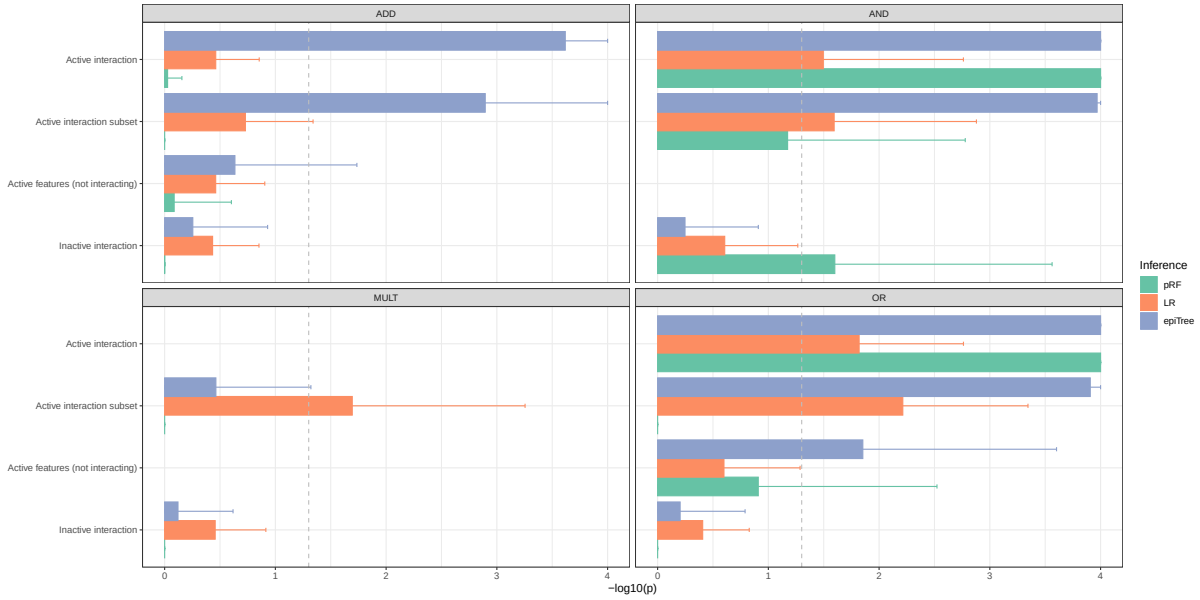

Figure S20: Summary of average p-values across 25 simulation replicates by response model and inference method (pRF: permutation random forests, LR: logistic regression, epiTree: PCS epistasis p-values). Inference was conducted on interactions recovered by iRF. Note: iRF did not detect the full active interaction for the multiplicative model. Candidate interactions are grouped by whether they correspond to (i) a true interaction (ii) a subset of the true interaction (iii) a non-active interaction that includes only active features (iv) a non-active interaction that includes inactive features. Error bars show 1 standard deviation. The grey dashed line corresponds to a significance threshold of 0.05.

References

[1] T. M. Beasley, S. Erickson, and D. B. Allison. Rank-based inverse normal transformations are increasingly used, but are they merited? *Behavior genetics*, 39(5):580–595, 2009.

[2] R. Berk, et al. Misspecified mean function regression: Making good use of regression models that are wrong. *Sociological Methods & Research*, 43(3):422–451, 2014.

[3] L. Breiman, et al. *Classification and Regression Trees*. Chapman and Hall, New York, 1984.

[4] H. J. Cordell. Epistasis: What it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics*, 11(20):2463–2468, 2002.

[5] H. J. Cordell, et al. Statistical modeling of interlocus interactions in a complex disease: Rejection of the multiplicative model of epistasis in type 1 diabetes. *Genetics*, 158(1):357–367, 2001.

[6] A. Dumitriu, et al. Integrative analyses of proteomics and RNA transcriptomics implicate mitochondrial processes, protein folding pathways and GWAS loci in Parkinson disease. *BMC Medical Genomics*, 9(1):5, December 2015.

[7] J. J. Faraway. Does data splitting improve prediction? *Statistics and Computing*, 26(1-2):49–60, 2016.

[8] R. Foraita, K. Bammann, and I. Pigeot. Modeling gene-gene interactions using graphical chain models. *Human Heredity*, 65(1):47–56, 2008.

[9] F. Girosi and T. Poggio. Representation properties of networks: Kolmogorov's theorem is irrelevant. *Neural Computation*, 1(4):465–469, 1989.

[10] I. B. Hallgrímsdóttir and D. S. Yuster. A complete classification of epistatic two-locus models. *BMC Genetics*, 9(1), 2008.

[11] T. Hastie and R. Tibshirani. Generalized additive models. *Statistical Science*, 1(3):297–318, 1986.

[12] J. Hathaway, et al. Diagnostic yield of genetic testing in a heterogeneous cohort of 1376 HCM patients. *BMC Cardiovascular Disorders*, 21(1):126, December 2021.

[13] Y. Huang, S. Wuchty, and T. M. Przytycka. eQTL epistasis – challenges and computational approaches. *Frontiers in Genetics*, 4:51, 2013.

[14] O. Kobiler, et al. Quantitative kinetic analysis of the bacteriophage genetic network. *Proceedings of the National Academy of Sciences*, 102(12):4470–4475, 2005.

[15] H. Leeb. Conditional predictive inference post model selection. *The Annals of Statistics*, 37(5B):2838–2876, 2009.

[16] E. Levine and T. Hwa. Small RNAs establish gene expression thresholds. *Current Opinion in Microbiology*, 11(6):574–579, 2008.

[17] J. Li, et al. Detecting gene-gene interactions using a permutation-based random forest method. *BioData mining*, 9(1):1–17, 2016.

[18] J. W. Little. Threshold effects in gene regulation: When some is not enough. *Proceedings of the National Academy of Sciences*, 102(15):5310–5311, 2005.

[19] J. W. Little, D. P. Shepley, and D. W. Wert. Robustness of a gene regulatory circuit. *The EMBO Journal*, 18(15):4299–4307, 1999.

[20] M. D. Morgan, et al. Genome-wide study of hair colour in UK Biobank explains most of the SNP heritability. *Nature Communications*, 9:5271, 2018.

[21] S. V. Naoaev. Some limit theorems for large deviation. *Theory of Probability and its Applications*, 10(2):214–235, 1965.

[22] B. V. North, D. Curtis, and P. C. Sham. Application of logistic regression to case-control association studies involving two causative loci. *Human Heredity*, 59(2):79–87, 2005.

[23] S. Pudewell, et al. Accessory proteins of the RAS-MAPK pathway: Moving from the side line to the front line. *Communications Biology*, 4(1):696, December 2021.

[24] Z. R. Sailer and M. J. Harms. Detecting high-order epistasis in nonlinear genotype-phenotype maps. *Genetics*, 205(3):1079–1088, 2017.

[25] S. H. Spiezio, et al. Genetic divergence and the genetic architecture of complex traits in chromosome substitution strains of mice. *BMC Genetics*, 13(1):38, December 2012.

[26] C. Tcheandjieu, et al. A phenome-wide association study of 26 mendelian genes reveals phenotypic expressivity of common and rare variants within the general population. *PLOS Genetics*, 16(11):e1008802, November 2020.

[27] A. T. Timberlake, et al. Two locus inheritance of non-syndromic midline craniosynostosis via rare SMAD6 and common BMP2 alleles. *eLife*, 5:e20125, September 2016.

[28] M. Ueki and H. J. Cordell. Improved statistics for genome-wide interaction analysis. *PLOS Genetics*, 8(4):e1002625, 2012.

[29] G. Wahba. Bayesian "confidence intervals" for the cross-validated smoothing spline. *Journal of the Royal Statistical Society: Series B (Methodological)*, 45(1):133–150, 1983.

[30] L. Wasserman, A. Ramdas, and S. Balakrishnan. Universal inference. *Proceedings of the National Academy of Sciences*, 117(29):16880–16890, 2020.

[31] X. Wu, et al. A novel statistic for genome-wide interaction analysis. *PLoS Genetics*, 6(9):e1001131, 2010.

[32] B. Yu and K. Kumbier. Veridical data science. *Proceedings of the National Academy of Sciences*, 117(8):3920–3929, 2020.