Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

Data and text mining

Nearest-neighbor Projected-Distance Regression (NPDR) for detecting network interactions with adjustments for multiple tests and confounding

Trang T. Le[1], **Bryan A. Dawkins**[2] **and Brett A. McKinney**[2,3,*]

[1]Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, PA 19104
[2]Department of Mathematics, University of Tulsa, Tulsa, OK 74104
[3]Tandy School of Computer Science, University of Tulsa, Tulsa, OK 74104

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Efficient machine learning methods are needed to detect complex interaction-network effects in complicated modeling scenarios in high-dimensional data, such as GWAS, gene expression, eQTL, and structural/functional neuroimage studies for case-control or continuous outcomes. Many machine learning feature selection methods have limited ability to address the issues of controlling the false discovery rate and adjusting for covariates. To address these challenges, we develop a new feature selection technique called Nearest-neighbor Projected-Distance Regression (NPDR) that uses the generalized linear model (GLM) to perform regression between nearest-neighbor pair distances projected onto predictor dimensions. Motivated by the nearest-neighbor mechanism in Relief-based algorithms, NPDR captures the underlying interaction structure of the data, handles both dichotomous and continuous outcomes and various combinations of predictor data types, statistically corrects for covariates and permits regularization. We use realistic simulations with main effects and network interactions to show that NPDR outperforms standard Relief-based methods and random forest at detecting functional variables while also enabling covariate adjustment and multiple testing correction. Using RNA-Seq data from a study of major depressive disorder (MDD), we show that NPDR with covariate adjustment effectively removes spurious associations due to confounding. We apply NPDR to eQTL data to identify potentially interacting variants that regulate transcripts associated with MDD and demonstrate NPDR's utility for GWAS and continuous outcomes.

1 Introduction

Epistasis is a measure of the effect of two genetic variants on a phenotype beyond what would be expected by their independent effects. There is evidence that these non-independent effects are pervasive Breen *et al.* (2012) and that higher-order interactions also play an important role in genetics Weinreich *et al.* (2013). A similar interaction effect can be observed in differential co-expression, where the phenotypic effect of one gene is modified depending on the expression of another gene Lareau *et al.* (2015); De la Fuente (2010). The embedding of these interactions in a regulatory network may lead to, not only pairwise interactions, but also higher-order epistasis network effects. Explicit modelling of these higher-order interactions would be computationally and statistically intractable Riesselman *et al.* (2018). Thus, computationally scalable feature selection methods are needed to capture these higher-order effects in high-dimensional data such as genome-wide association Arabnejad *et al.* (2018) and RNA-Seq studies Le *et al.* (2018b).

Relief-based algorithms are efficient nearest-neighbor feature selection methods that are able to detect epistasis or statistical interaction effects in high-dimensional data without resorting to pairwise modeling of attributes Urbanowicz *et al.* (2018b); Kononenko *et al.* (1997); McKinney *et al.* (2009a); Robnik-Šikonja and Kononenko (2003a). We recently introduced the STatistical Inference Relief (STIR) formalism Le *et al.* (2018b) to address the lack of a statistical distribution for hypothesis testing and the related challenge of controlling the false positive rate of Relief-based scores. STIR extended Relief-based methods to compute statistical significance of attributes in dichotomous outcome data (e.g., case-control) by reformulating the Relief weight McKinney *et al.* (2013) as a pseudo t-test for the difference of means between projected-distances onto a given attribute. STIR is an effective approach with high power and low false-positive rates for data with main and interaction effects and is applicable to any predictor data type (continuous/gene expression or nominal/genetic variants). However, being based on a t-test, it does not apply to data with a continuous outcome (e.g., quantitative trait) and does not correct for covariates.

In the current study, we introduce a new Nearest-neighbor Projected-Distance Regression (NPDR) approach that extends the STIR formalism to regression of nearest neighbors with the generalized linear model (GLM). For each attribute, the NPDR model fits a GLM of projected-distances (onto the attribute) between all pairs of nearest-instance neighbors. The model of projected distances is linear for continuous outcomes and logistic for dichotomous outcomes, and either model may include corrections for covariates. The importance of an attribute is given by its standardized regression coefficient and the statistical significance by its P value. Prior to the current study, no Relief-based method has included covariate correction.

Covariate adjustment is often neglected in machine learning, yet many biological and clinical omic studies involve potentially confounding covariates such as sex, bmi, and age Le *et al.* (2018a) or population stratification Chen *et al.* (2016). Some proposed methods for correcting machine learning algorithms include restricted permutation Rao *et al.* (2017), inverse probability weighting of training samples Linn *et al.* (2016) and penalized support vector machines Li *et al.* (2011). Our NPDR framework leads to a natural way to control for covariates by including additional terms for the between-neighbor projected differences for each covariate within the GLM. We demonstrate the effectiveness of NPDR to correct for confounding in an RNA-Seq study of major depressive disorder (MDD) in which there is a strong signal in the expression data due to the sex of study participantsMostafavi *et al.* (2014).

The flexible GLM formalism of NPDR opens up Relief-based methods to statistical inference for a broad class of problems. It can detect main effects and interactions for dichotomous and continuous outcome studies while adjusting for covariates and multiple hypothesis testing. The models allow any predictor data type such as GWAS variants or RNA-Seq expression levels. NPDR also improves attribute importance estimation compared to other Relief methods because it includes dispersion of the projected distances. The projected distance model gives the appearance of being univariate but implicitly accounts for interactions with all other attributes via the neighborhood calculation in the space of all attributes (omnigenic). Further, we demonstrate that the NPDR formalism allows for Lasso and Ridge-like penalized feature selection for Relief-based methods.

The paper is organized as follows. In the Methods section, we develop the new formalism of NPDR to reformulate Relief-based scores as coefficients in a distanced-based GLM. We use the projected-distance regression formalism to implement a penalized version of NPDR. In the Results, we use realistic simulations with main effects and network interactions to demonstrate improved feature selection performance over standard Relief-based methods and random forest. We apply NPDR to an RNA-Seq and GWAS study of MDD. From the RNA-Seq data, we show that NPDR removes spurious associations with MDD due to confounding by sex and identifies biologically relevant genes. Combining the RNA-Seq with the GWAS data, we perform an eQTL analysis with NPDR to identify potentially interacting variants that regulate transcripts associated with MDD and demonstrate NPDR's utility for GWAS and continuous outcomes.

2 Materials and Methods

In this section, we develop the mathematical formalism needed to describe the projected distance regression in NPDR. We then construct the NPDR GLM models for common analysis situations, including continuous and dichotomous outcomes and adjustment for covariates. We also describe the simulation approach and real datasets for method validation.

2.1 Distance metrics and nearest neighbors

Because NPDR and other Relief-based feature selection methods are based on distances between instances, we first describe the algorithms and notation for identifying nearest neighbors in the space all attributes. We use the term attribute to refer to predictor variables, which may be continuous (e.g., expression) or categorical (e.g., variants). We use the term instance to refer to samples or subjects in a dataset.

2.1.1 Distances and projections onto attributes

The distance between instances i and j in the data set $X^{m\times p}$ of m instances and p attributes is calculated in the space of all attributes ($a \in A$, $|A| = p$) using a metric such as

$$D_{ij}^{(q)} = \left(\sum_{a\in A} |\mathrm{d}_{ij}(a)|^q \right)^{1/q}, \tag{1}$$

which is typically Manhattan ($q = 1$) but may also be Euclidean ($q = 2$). The quantity $\mathrm{d}_{ij}(a)$, known as a "diff" in Relief literature, is the projection of the distance between instances i and j onto the attribute a dimension. The function $\mathrm{d}_{ij}(a)$ supports any type of attribute (e.g., numeric/continuous versus categorical). For example, the projected difference between two instances i and j for a continuous numeric (d^{num}) attribute, a, may be defined as

$$\begin{aligned} \mathrm{d}_{ij}^{\text{num}}(a) &= \text{diff}(a,(i,j)) \\ &= |\hat{X}_{ia} - \hat{X}_{ja}|, \end{aligned} \tag{2}$$

where $\hat{X}$ represents the standardized data matrix X. We use a simplified $\mathrm{d}_{ij}(a)$ notation in place of the diff$(a,(i,j))$ notation that is customary in Relief-based methods. We also omit the division by $\max(a) - \min(a)$ that is used by Relief to constrain scores to the interval from -1 to 1. As we show in subsequent sections, NPDR attribute importance scores are standardized regression coefficients with corresponding P values, so any scaling of the projected distances is unnecessary for comparing attribute scores. Thus, for the numeric-data projection, $\mathrm{d}_{ij}^{\text{num}}(a)$, we simply use the absolute difference between row elements i and j of the data matrix $X^{m\times p}$ for the attribute column a.

This numeric projection function (Eq. 2) is appropriate for gene expression and other quantitative predictors and outcomes. For genome-wide association study (GWAS) data, where attributes are categorical, one simply modifies the type in the projection function Arabnejad *et al.* (2018), but the projected-distance regression methods will be otherwise unchanged. The $\mathrm{d}_{ij}(a)$ quantity is typically part of the metric to define the neighborhood, but it is also essential for computing the importance coefficients (Sec. 2.2.1). The projected-distance regression models below (Eqs. 8, 11, 14, and 15) will be fit for all nearest neighbors i and j in the defined neighborhood (discussed next in Eq. (3)).

2.1.2 Nearest-neighbor ordered pairs

In the original Relief-F approach for dichotomous outcome data, two neighborhood sets are calculated: one for hits (instances in the same class) and one for misses (instances in different classes). For NPDR, only one neighborhood is needed, regardless of whether the problem is classification or regression and regardless of whether a fixed-k or adaptive radius neighborhood method is used Greene *et al.* (2009); Urbanowicz *et al.* (2018a); McKinney *et al.* (2013). The NPDR neighbors are chosen blind to the outcome variable and then pairs of instances are assigned to hit or miss groups for dichotomous outcome data and assigned numeric differences for quantitative outcome data. This blinded selection leads to less overfitting of the neighborhood boundaries and less bias for imbalanced data.

We define the NPDR neighborhood set $\mathcal{N}$ of ordered-pair indices as follows. Instance i is a point in p dimensions, and we designate the topological neighborhood of i as N_i. This neighborhood is a set of other instances trained on the data $X^{m\times p}$ and depends on the type of Relief neighborhood method (e.g., fixed-k or adaptive radius) and the type of metric (e.g., Manhattan or Euclidean). If instance j is in the neighborhood of i ($j \in N_i$), then the ordered pair is in the overall neighborhood ($(i,j) \in \mathcal{N}$) for the projected-distance regression analysis. The ordered pairs constituting the overall neighborhood can then be represented as nested sets:

$$\mathcal{N} = \{\{(i,j)\}_{i=1}^{m}\}_{\{j\neq i : j\in N_i\}}. \tag{3}$$

The cardinality of the set $\{j \neq i : j \in N_i\}$ is k_i, the number of nearest neighbors for subject i.

2.1.3 Adaptive-radius and fixed-k Neighborhoods

The NPDR algorithm applies to any Relief neighborhood algorithm. In the applications in the current study, we use the multiSURF Urbanowicz *et al.* (2018a) adaptive radius neighborhood, which uses a different radius for each instance, and we use a fixed-k neighborhood that well-approximates mulitSURF, which is derived in Ref. Dawkins *et al.* (2019). The adaptive radius for an instance may be defined as the mean of its distances to all other instances subtracted by the fraction α of the standard deviation of this mean. More precisely, an instance j is in the adaptive α-radius neighborhood of i ($j \in N_i^{\alpha}$) under the condition

$$D_{ij} \leq R_i^{\alpha} \implies j \in N_i^{\alpha}, \tag{4}$$

where the threshold radius for instance i is

$$R_i^{\alpha} = \bar{D}_i - \alpha\,\sigma_{\bar{D}_i} \tag{5}$$

and

$$\bar{D}_i = \frac{1}{m-1}\sum_{j\neq i} D_{ij}^{(\cdot)} \tag{6}$$

is the average of instance i's pairwise distances (using Eq. 1) with standard deviation $\sigma_{\bar{D}_i}$. MultiSURF uses $\alpha = 1/2$ Granizo-Mackenzie and Moore (2013).

Previously we showed empirically for balanced dichotomous outcome datasets that a good constant-k approximation to the expected number of neighbors within the multiSURF radii is $k = m/6$ Le *et al.* (2018b), where m is the number of samples. In Ref. Dawkins *et al.* (2019) we derive a more exact theoretical mean that shows the mathematical connection between fixed-α and fixed-k neighbor-finding methods, which is given by

$$\bar{k}_{\alpha} = \left\lfloor \frac{m-1}{2}\left(1 - \mathrm{erf}\left(\frac{\alpha}{\sqrt{2}}\right)\right)\right\rfloor, \tag{7}$$

where we apply the floor to ensure the number of neighbors is integer. For data with balanced hits and misses in standard fixed-k Relief, one further divides this formula by 2, and then for multiSURF ($\alpha = 1/2$), we find $\bar{k}_{1/2}^{\text{hit/miss}} = \frac{1}{2}\bar{k}_{1/2} = 0.154(m-1)$, which is very close to our previous empirical estimate $m/6$. In the current study, when we compare multiSURF neighborhood methods with fixed-k neighborhoods, we use $\bar{k}_{1/2}$. Using this $\alpha = 1/2$ value has been shown to give good feature selection performance by balancing power for main effects and interaction effects. However, the best value for α or k is likely data-specific and may be determined through nested cross-validation and other parameter tuning methods Dawkins *et al.* (2019).

2.2 Nearest-neighbor Projected-Distance Regression (NPDR) with the generalized linear model

2.2.1 Continuous outcomes: linear regression NPDR

Once the neighborhood $\mathcal{N}$ (Eq. 3) is determined by the distance matrix D_{ij} (Eq. 1) and the neighborhood method is chosen (e.g., fixed number of neighbors k or adaptive radius), we can compute the NPDR test statistic and P value for the association of an attribute with the phenotype. The NPDR model predictor vector is the attribute's projected distances ($\mathrm{d}_{ij}(a)$) between all pairs of nearest-neighbor instances ($(i,j) \in \mathcal{N}$). For continuous outcome data (quantitative phenotypes), the NPDR model outcome vector is the numeric difference (Eq. 2) between all nearest neighbors i and j. We find the parameters of the following model that minimize the least-squares error over $\forall(i,j) \in \mathcal{N}$:

$$\mathrm{d}_{ij}^{\text{num}}(y) = \beta_o + \beta_a \mathrm{d}_{ij}(a) + \epsilon_{ij}. \tag{8}$$

The $\mathrm{d}_{ij}^{\text{num}}(y)$ term on the left is the projected distance (diff) between instances i and j for a numeric phenotype y (Eq. 2) and ϵ_{ij} is the error term for this random variable. The predictor attribute a may be numeric or categorical, which determines the "type" used in the diff function on the right hand side of Eq.(8). The NPDR test statistic for attribute a is the β_a estimate with a one-sided hypothesis

$$\begin{aligned} H_0 &: \beta_a < 0 \\ H_1 &: \beta_a \geq 0. \end{aligned} \tag{9}$$

The β_a can be interpreted as the predicted change in the difference of the quantitative outcome between a pair of subjects when the projected distance of the attribute, a, changes by one unit. The attribute weights in the original RRelief algorithm Robnik-Šikonja and Kononenko (2003a) can be described as a weighted covariance between the attribute neighbor projected distances, $\mathrm{d}_{ij}(a)$, and the outcome neighbor differences, $\mathrm{d}_{ij}^{\text{num}}(y)$. The extra weighting in RRelief is an exponentially decaying function of the rank of the distance between neighbors. The NPDR attribute weight is the standardized regression coefficient, β_a', which is the covariance of the projected distances divided by the variance of the outcome projected distances. When the regression contains no additional covariates, the NPDR attribute weight can be written as the correlation between outcome and attribute neighbor projected distances:

$$\beta_a' = \mathrm{corr}\left(\mathrm{d}(\mathbf{y}), \mathrm{d}(\mathbf{a})\right). \tag{10}$$

Thus, in the case of no covariates, NPDR for regression and RRelief have similar structure, but, as we show shortly, NPDR provides an improved attribute estimation, and the flexible NPDR framework can include additional sources of variation (i.e., adjust for confounding covariates).

2.2.2 Linear regression NPDR with covariates

Previous Relief-based methods do not include the ability to adjust for covariates. The regression formalism of NPDR makes adding covariates straightforward. We simply compute the projected difference values $\mathrm{d}_{ij}(\vec{y}_{\text{covs}})$ for the covariate attribute(s) between subjects on the neighborhood ($\forall(i,j) \in \mathcal{N}$) and include this as an additional projected distance term in the regression model:

$$\mathrm{d}_{ij}^{\text{num}}(y) = \beta_0 + \beta_a \mathrm{d}_{ij}(a) + \vec{\beta}_{\text{covs}}^T \mathrm{d}_{ij}(\vec{y}_{\text{covs}}) + \epsilon_{ij}. \tag{11}$$

The above vector notation for the regression coefficients of the p_c covariates can be expanded as

$$\vec{\beta}_{\text{covs}}^T = \left(\beta_{\text{cov}_1}, \beta_{\text{cov}_2}, \ldots, \beta_{\text{cov}_{p_c}}\right) \tag{12}$$

and the projection differences between instances i and j for each of the p_c covariates can be expanded as

$$\mathrm{d}_{ij}(\vec{y}_{\text{covs}}) = \left(\mathrm{d}_{ij}^{\text{type}_1}(y_{\text{cov}_1}), \mathrm{d}_{ij}^{\text{type}_2}(y_{\text{cov}_2}), \ldots, \mathrm{d}_{ij}^{\text{type}_{p_c}}(y_{\text{cov}_{p_c}})\right)^T. \tag{13}$$

The superscripts in the projection operators above indicate the appropriate operator type for each covariate data type (e.g., numeric or categorical). In addition, the predictor attribute a may be numeric or categorical, which determines the type used in $\mathrm{d}_{ij}(a)$. The NPDR test statistic is again β_a with alternative hypothesis $\beta_a \geq 0$ as in Eq. (9), and the standardized β'_a is the attribute importance score.

2.2.3 Dichotomous outcomes with covariates: logistic regression NPDR

The STIR method was designed for statistical testing in dichotomous data, but as we have seen NPDR can also handle continuous outcomes and adjust for covariates. Here we show that the GLM formalism also enables NPDR to handle dichotomous outcome data (e.g., case-control phenotype). For dichotomous outcomes, NPDR models the probability p_{ij}^{miss} that subjects i and j are in the opposite class (misses) versus the same class (hits) from the neighbor projected distances with a logit function. We estimate the parameters of the following model for neighbors $\forall(i,j) \in \mathcal{N}$:

$$\text{logit}(p_{ij}^{\text{miss}}) = \beta_0 + \beta_a \mathrm{d}_{ij}(a) + \epsilon_{ij}, \tag{14}$$

or if there are covariates,

$$\text{logit}(p_{ij}^{\text{miss}}) = \beta_0 + \beta_a \mathrm{d}_{ij}(a) + \vec{\beta}_{\text{covs}}^T \mathrm{d}_{ij}(\vec{y}_{\text{covs}}) + \epsilon_{ij}, \tag{15}$$

where p_{ij}^{miss} is the probability that subjects i and j have different phenotypes given the difference in their values for the attribute, a, and given the covariate differences. The outcome variable that is modeled by probability p_{ij}^{miss} is a binary difference between subjects for the phenotype ($\vec{y}$):

$$\mathrm{d}_{ij}^{\text{miss}}(\vec{y}) = \begin{cases} 0, & y_i == y_j \\ 1, & \text{else.} \end{cases} \tag{16}$$

The β_a statistic can be interpreted in the following way. For a unit increase in the difference in the value of the attribute between two neighbors, we predict a change of e^{β_a} in the odds of the neighbors being in opposite classes. For dichotomous outcome data, we are interested in the alternative hypothesis that $\beta_a > 0$ because negative β_a values represent attributes that are irrelevant to classification. Thus, like NPDR linear regression, we are interested in testing one-sided hypotheses

$$\begin{aligned} H_0 &: \beta_a \leq 0 \\ H_1 &: \beta_a > 0. \end{aligned} \tag{17}$$

Nominal outcomes can be analyzed (similar to multi-state Relief-F) with NPDR by grouping all misses of an instance as one group. This may be improved by using multinomial regression in NPDR.

2.2.4 Regularized NPDR

To complement the multiple-testing adjusted NPDR approach, we develop a regularized NPDR method that combines all of the attribute difference vectors into one design matrix and constrains the coefficients to be non-negative, similar to the one-tailed test we use in standard NPDR. Specifically, we minimize the vector of regression coefficients, $\vec{\beta}_A = (\beta_{a_1}, \beta_{a_2}, \ldots, \beta_{a_p})$, simultaneously for all attribute projections $a \in A$, subject to the coefficients being non-negative:

$$\begin{aligned} &\min_{\beta_o, \vec{\beta}_A} \frac{1}{|\mathcal{N}|} \sum_{i,j \in \mathcal{N}} \mathcal{L}\left(\mathrm{d}_{ij}^{\text{miss}}(y), \beta_0 + \vec{\beta}_A^T \mathrm{d}_{ij}(A)\right) + \lambda ||\vec{\beta}_A||_1 \\ &\quad \beta_{a_k} \geq 0, \quad k = 1, \ldots, p. \end{aligned} \tag{18}$$

$\mathcal{L}$ is the negative log-likelihood for each pair of instances i and j in neighborhood $\mathcal{N}$, and $\mathrm{d}_{ij}(A)$ represents the vector of diffs for fixed i and j for all attributes $a \in A$:

$$\mathrm{d}_{ij}(A) = \left(\mathrm{d}_{ij}^{\text{type}_1}(a_1), \mathrm{d}_{ij}^{\text{type}_2}(a_2), \ldots, \mathrm{d}_{ij}^{\text{type}_p}(a_p)\right)^T. \tag{19}$$

Our implementation uses a zero lower limit for the coefficients and the penalty strength $\lambda > 0$ is chosen by cross-validation Zou and Hastie (2005). For dichotomous outcomes, we use the binomial link function for the hit/miss projected distances in the likelihood optimization.

2.3 Properties of NPDR and existing Relief-based methods

Here we summarize the properties and capabilities of standard Relief-based methodsUrbanowicz *et al.* (2018b) and the generalizations STIRLe *et al.* (2018b) and NPDR (Table 1). When there are no covariates for dichotomous outcome data, STIR (based on a pseudo t-test) and NPDR (based on regression with a logit model) are approximately equivalent given reasonable distribution assumptions (see Supplementary Fig. S1). However, by design, the NPDR framework is more flexible and able to handle covariate adjustment and continuous outcomes. In the current notation, the STIR null and alternative hypotheses would be

$$\begin{aligned} H_0^{\text{stir}} &: \mu_M(a) - \mu_H(a) \leq 0 \\ H_1^{\text{stir}} &: \mu_M(a) - \mu_H(a) > 0, \end{aligned} \tag{20}$$

where

$$\begin{aligned} \mu_M(a) &= \bar{M}_a = E\left(\mathrm{d}_{ij}(a) \cdot \left(1 - \mathrm{d}_{ij}^{\text{miss}}(y)\right)\right) \\ \mu_H(a) &= \bar{H}_a = E\left(\mathrm{d}_{ij}(a) \cdot \mathrm{d}_{ij}^{\text{miss}}(y)\right) \end{aligned} \tag{21}$$

and $\mathrm{d}_{ij}^{\text{miss}}(y)$ is given by Eq.(16). The STIR test statistic is a pseudo t-test (see Ref Le *et al.* (2018b)).

For dichotomous outcomes, STIR improves attribute estimates over Relief weights ($\bar{M}_a - \bar{H}_a$) by incorporating sample variance of the nearest neighbor distances in the denominator, which also enables STIR to estimate statistical significance with the assumptions of a t-test (Table 1). NPDR assumes intra- and inter-class differences are randomly sampled from one distribution and computes the importance score from a logistic regression β'_a. This regression-based generalization improves NPDR's attribute estimates over Relief weights and enables statistical significance estimation for a wider range of problems than STIR.

2.4 Real and simulated datasets

2.4.1 Simulation methods

To compare power and false positive performance for NPDR and other feature selection methods, we use the simulation tool from our private Evaporative Cooling (privateEC) software Le *et al.* (2017) that was designed to simulate realistic main effects, correlations, and interactions

	Standard Relief-based	STIR	NPDR
Importance score (dichotomous)	$\bar{M}_a - \bar{H}_a$	$\dfrac{\bar{M}_a - \bar{H}_a}{S_p(M,H)\sqrt{\frac{1}{\lvert M\rvert}+\frac{1}{\lvert H\rvert}}}$	β'_a coefficient
Score has a null distribution	No	Yes	**Yes**
Supports continuous outcome	Yes	No	**Yes**
Supports covariates	No	No	**Yes**

Table 1. Properties of standard Relief-based methods and generalizations STIR and NPDR. The NPDR score is the standardized regression coefficient β'_a from a logistic model (Eq. 14) for dichotomous outcomes and from a linear model (Eq. 8) for continuous outcomes. The quantity S_p in STIR is the pooled standard deviation of the hit and miss means. Only the score for dichotomous (hit/miss) Relief is shown and STIR is limited to dichotomous outcomes Le et al. (2018b).

found in gene expression or resting-state fMRI correlation data. For continuous outcome data, we simulate main effects with $m = 200$ subjects and $p = 1000$ real-valued attributes with 10% functional (true positive association with outcome). We choose a sample size consistent with real gene expression data but on the smaller end to demonstrate a more challenging scenario. Likewise, the effect size parameter ($b = 0.8$) was selected to be sufficiently challenging with power approximately 40% Le *et al.* (2017).

For dichotomous outcome data (100 cases and 100 controls), we simulate network interactions using the differential co-expression network-based simulation tool in privateEC, which is described in Refs. Le *et al.* (2017); Lareau *et al.* (2015). We first create a co-expression network on an Erdős-Rényi random graph with 0.1 attachment probability, which is the critical value for a giant component. We give connected genes a higher average correlation, approximately $r_{\text{connected}} = 0.8$.This correlation is related to the interaction effect size because we disrupt the correlation of target genes in cases but maintain correlation within controls, thereby creating a final differential correlation network.

All resulting p-values (from STIR and NPDR) are adjusted for multiple testing. Attributes with adjusted p-values less than 0.05 are counted as a positive test (null hypothesis rejected), else the test is negative. We assess the feature-selection performance of each method by averaging the area under the precision-recall curve (auPRC) across 100 replicates of each simulation scenario. Assuming relatively few functional attributes (10% of 1,000 attributes) compared to non-functional ones, the precision and recall measures are robust to imbalanced data and are thus a useful assessment of a method's propensity to assign higher scores to the correct functional attributes. The auPRC is also a good comparison tool for methods that, unlike NPDR, do not have a statistical significance threshold. We remark that, even though the auPRC terminology used here is similar to traditional classification problems, we instead focus on evaluating the attribute score quality, not classification accuracy.

2.4.2 RNA-Seq data and NPDR adjustment for confounding factors

To test the ability of NPDR to correct for confounding, we apply NPDR to the RNA-Seq study in Ref. Mostafavi *et al.* (2014) that consists of 15,231 genes for 463 MDD cases and 452 controls. Of the 915 subjects, 641 are female and 274 are male. The chi-square between MDD and sex is 25.746 ($p = 3.89e - 7$), and there are 485 genes significantly associated with sex. Thus, there is high risk for confounding effects due to sex differences. We apply NPDR with a multiSURF neighborhood and compute importance scores of all genes with and without sex as a covariate to isolate confounding genes.

2.4.3 eQTL data and NPDR for GWAS and quantitative outcomes

We perform an eQTL analysis with NPDR feature selection to identify potentially interacting variants that regulate transcripts associated with MDD and demonstrate NPDR's utility for GWAS and continuous outcomes. The MDD RNA-Seq study described above includes 915 GWAS subjects genotyped with the Illumina Omni1-Quad microarrayMostafavi *et al.* (2014). We use NDPR to test for the cis- (1Mb from the gene's transcription start site) and trans-eQTL influence on one of the gene expression levels associated with MDD (SCAI gene). We included $281{,}648$ variants following GWAS filtering. We remove variants with a deviation from Hardy–Weinberg equilibrium (P < 0.0001 in controls) and a minor allele frequency (MAF) < 0.01, and we use linkage disequilibrium (LD) pruning to reduce the potential bias of correlation on interaction and distance calculations. SNPs are recursively removed within a sliding window along a given chromosome based on a pairwise LD of 0.5. We control for MDD status in the NPDR models to isolate more direct influence of variants on expression rather than MDD association. We use the Eq.(11) NPDR model and an allele mismatch operator for the SNP attribute projectionsArabnejad *et al.* (2018).

2.5 Software availability

Detailed simulation and analysis code needed to reproduce the results in this study is available at https://github.com/lelaboratoire/npdr-paper (R version 3.5.0). The *npdr* R package is available at https://insilico.github.io/npdr/.

3 Results

3.1 Simulation results

In simulated data with main effects and interactions for continuous and dichotomous outcomes, NPDR attribute estimates show improved precision and recall over standard Relief and random forest importance scores for detecting functional variables (Fig. 1). For one simulation, we illustrate the Precision Recall Curve (PRC) for a grid of attribute importance thresholds (Fig. 1A), which shows improved area under the PRC (auPRC) for NPDR for both continuous (left) and dichotomous (right) outcomes. Across 100 replicate simulations for each simulation type, NPDR shows significantly higher auPRC than random forest and Relief (both $P < 0.0001$, Fig. 1B). The auPRC values for all methods are higher in the interaction effect simulations relative to the main effect simulations because of a larger simulated effect size.

We use auPRC to compare other machine learning methods with NPDR because Relief and random forest lack a null distribution, whereas NPDR has an approximate distribution for hypothesis testing. NPDR correctly detects 57 out of 100 functional attributes in a continuous-outcome main effect simulation (Fig. 2A) and 86 out of 100 functional attributes in a dichotomous-outcome interaction simulation (Fig. 2B) using an adjusted P value threshold. Given any vertical score cutoff (Fig. 2), it is difficult for Relief or random forest to detect most of the functional attributes without including many more false positives than NPDR. As shown by the auPRC, NPDR tends to include fewer false positives than the other methods as it detects more functional attributes (Fig. 1).

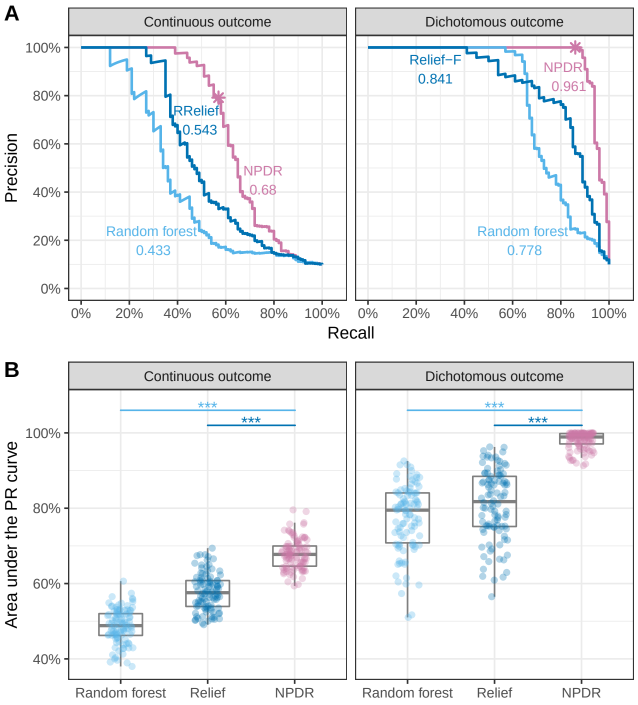

Fig. 1. Simulation comparison for detection of functional variables. For one replicate simulation (A, top row), precision-recall curves (PRC) for continuous outcome data with main effects (left) and dichotomous outcome data with interaction effects (right). The area under the PRC (auPRC) value is reported next to each method: NPDR (magenta), Relief-based (dark blue) and random forest (light blue). The magenta $*$ indicates the NPDR 0.05 adjusted cutoffs (scores shown in Fig. 2). For 100 replicate simulations (B, bottom row), the distributions of the auPRC values are compared for the methods. NPDR yields statistically significant higher auPRC than Relief or random forest ($***$ indicate P $<$ .0001). All simulations use $m = 200$ samples and $p = 1,000$ attributes with 100 functional.

NPDR importance scores are highly correlated with Relief-F scores with $r = 0.869$ for continuous outcome main effects and 0.848 for dichotomous outcome interaction effects (Fig. 2). This correlation is expected because both methods are nearest-neighbor based. The correlation of NPDR with random forest scores is lower than with Relief-F, where $r = 0.692$ for continuous outcome main effects and 0.578 for dichotomous outcome interaction effects. The correlation for interaction effect simulations decreases further because random forest underestimates the importance of interacting attributes as the attribute dimensionality becomes large compared to the number of functional attributes McKinney *et al.* (2009b); Winham *et al.* (2012).

Although our primary performance metric is auPRC, we also use the area under the Receiver Operating Characteristics curve (auROC), which also shows NPDR has statistically significant higher feature selection performance than random forest and Relief (both $P < 0.0001$, see Supplementary Fig. S2). We also compare NPDR (using a logistic model) with STIR (based on a t-test) for the dichotomous outcome data with interaction effects (Supplementary Fig. S1). While logistic regression and the t-test have slightly different assumptions on the distribution of samples, NPDR and STIR yield highly correlated scores for dichotomous data with interaction effects, where the correlation value r between the P values produced from the two methods ranges from 0.9827 to 0.9994 in 100 replications.

In the continuous and dichotomous simulations, respectively, we use NPDR with a linear model (Eq. 8) and logistic model (Eq. 14) to compare with standard RRelief and Relief-F (R package CORElearn) and with random forest regression and classification (permutation importance, R package randomforest). Because CORElearn does not include the adaptive multiSURF neighborhood, we use a fixed-k neighborhood $\mathcal{N}_{\bar{k}_{1/2}}$ for the Relief-based methods. The value $\bar{k}_{1/2} = 30$ (Eq. 7) is the expected number of nearest neighbors corresponding to a multiSURF neighborhood Dawkins *et al.* (2019). For a dataset of the size simulated in this study ($m = 200$ samples and $p = 1,000$ attributes with 100 functional), on a desktop with an Intel Xeon W-2104 CPU and 32GB of RAM, NPDR has a 24-second and 3-second runtime for dichotomous and continuous outcome data, respectively.

3.2 RNA-Seq NPDR analysis for MDD with confounding

NPDR with covariate adjustment effectively removes sex-related confounding genes in the RNA-Seq study of MDD in Ref. Mostafavi *et al.* (2014). We apply NPDR with the multiSURF neighborhood $\mathcal{N}_{\alpha=1/2}$ and an adjustment for the sex covariate (Eq. 15). This study contains numerous genes that are potentially confounded by sex differences. The sex variable is significantly associated with MDD, and 485 out of the 15,231 genes are associated with sex (Bonferroni-adjusted P value < 0.0001). The NPDR adjustment removes the genes that are most likely spurious MDD associations due to confounding (dark points below the horizontal 0.05 adjusted significance line in Fig. 3) compared to NPDR without adjustment. Not only do these removed genes have strong differential expression based on sex, but many of these genes, such as PRKY, UTY, and USP9Y, are Y-linked and mainly expressed in testis. For example, the RPS4Y2 ribosomal protein S4 Y-linked 2 has been shown by tissue

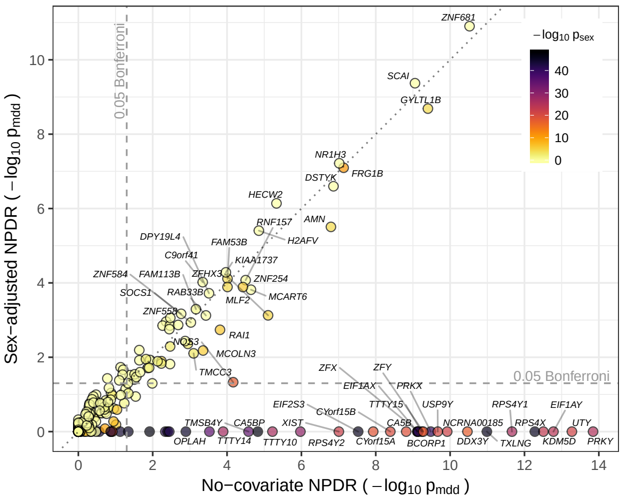

Fig. 3. Comparison of NPDR major depressive disorder associations with and without covariate adjustment. Gene scatter plot of $-\log_{10}$ significance using NPDR without correction for sex (horizontal axis) and with correction for sex (vertical axis). Genes with adjusted $p_{\text{mdd}} < 0.001$ by either method are labeled. NPDR without sex correction finds 87 genes associated with MDD at the Bonferroni-adjusted 0.05 level (right of vertical dashed line), 53 of which are also significantly correlated with sex (adjusted $p_{\text{sex}} < 0.05$). NPDR with adjustment for sex finds 56 genes associated with MDD at the Bonferroni-adjusted 0.05 level (above horizontal dashed line), 19 of which are significantly correlated with sex. The most highly associated genes with sex are eliminated by adjustment (dark genes below the horizontal dashed line) but remain in the non-adjusted set (right of dashed vertical line).

specific studies to mainly express in prostate and testis Lopes *et al.* (2010) while RPS4X (also associated with sex in the data) is most expressed in the ovary. The NPDR runtime for this RNA-Seq dataset ($m = 915$ samples and $p = 15,231$ attributes) was approximately 2.3 hours on a desktop with an Intel Xeon W-2104 CPU and 32GB of RAM.

3.3 eQTL analysis using NPDR with GWAS and quantitative outcomes

We perform an eQTL analysis with NPDR to demonstrate its ability to analyze continuous outcomes and SNP predictors from GWAS. We choose to test for eQTLs that influence the expression of SCAI (suppressor of cancer cell invasion, involved in cell migration and regulation of cell cycle on chromosome 9), which was one of the stronger NPDR associations with MDD (Fig. 3) and was found previously to have a modest cis-eQTL effect in this datasetMostafavi *et al.* (2014). One of the eQTLs found by NPDR (rs10997355) is an intron variant in CTNNA3 (catenin alpha-3 mediates cell–cell adhesion) on chromosome 10. In a study of schizophrenia, an intron variant near rs10997355 showed an interaction with maternal cytomegalovirus (CMV) status Børglum *et al.* (2014). Other studies of CTNNA3 and its nested gene LRRTM3 (encoding the Leucine-rich repeat transmembrane neuronal protein 3) have found associations with Alzheimer's disease Miyashita *et al.* (2007) and autism spectrum disorder Wang *et al.* (2009). The top 100 NDPR eQTLs for SCAI are provided as supplementary information. The NPDR runtime was 7.06 hours (on a high performance computing environment with an Intel Xeon E5-2650v3 CPU, 20 nodes and 32GB of RAM).

4 Discussion

NPDR is the first method to our knowledge to combine projected distances and nearest-neighbors into a generalized linear model regression framework to perform feature selection. The use of nearest neighbors enables its ability to detect interacting attributes, which is shared with Relief-based methods, but NPDR is a departure from Relief in at least five ways. (1) For feature selection with a continuous outcome, it does not rely on the idea of a hit/miss group like current RRelief approaches Urbanowicz *et al.* (2018a). Rather, NPDR simply performs a regression between the outcome and attribute projected distances. (2) For feature selection with dichotomous outcomes, NPDR uses a logistic model to fit pairwise projected distance regressors of hit and miss group. (3) This distance-based regression formalism provides a simple mechanism for NPDR to correct for covariates, which is often neglected in machine learning and has been a limitation of Relief-based methods. (4) For any outcome data type (dichotomous or continuous) and predictor data type, NPDR computes the statistical significance of attribute importance scores, which enables statistically based thresholds that can adjust for multiple hypothesis testing. (5) The NPDR attribute estimator (regression coefficient) includes more variation from the projected differences than other Relief-based methods and thereby improves attribute estimate quality. Moreover, we introduced a regularized NPDR that adds another layer of multivariate modeling to an already multi-dimensional nearest-neighbor method to shrink correlated projected attribute differences.

The novel NPDR importance score of an attribute is the standardized regression coefficient between the projected attribute differences and the numeric outcome differences between neighbors. For continuous outcomes, the regression Relief (RRelief) importance score Robnik-Šikonja and Kononenko (2003b) is a weighted correlation between attribute and outcome differences. This weighted correlation has a similar form to the NPDR standardized regression coefficient, which is a covariance between attribute and outcome differences divided by the variance in the outcome. However, we provided evidence that the NPDR standardized regression coefficient is a better attribute estimator than Relief and random forest, and we showed that the NPDR models can adjust for confounding covariates and other sources of variation.

We assessed NPDR's power and ability to control false positives using realistic simulations with main effects and network interactions. We showed that the statistical performance using NPDR P values is the same as STIR, which is limited to dichotomous outcome data. In other words, by modeling hit/miss differences between neighbors with a logit link, NPDR can be used safely instead of STIR with the added benefit of covariate correction and the analysis of quantitative traits. In a different Relief-based approach for continuous outcomes, Ref. Urbanowicz *et al.* (2018a) uses a standard deviation of the continuous outcome diffs to discretize the numeric outcome and make the RRelief algorithm compatible with the idea of hits and misses. However, discretization puts constraints on the variation in numeric data that increase the risk of losing power. NPDR uses the full variation in the continuous outcome variable, and the regression coefficient provides an interpretation in terms of variation explained while again providing flexibility for modeling additional sources of variation.

A related distance-based regression method is Multivariate Distance Matrix Regression (MDMR) Schork and Zapala (2012). The MDMR approach uses an F-statistic to test for the association of distance matrices between two sets of factors. The MDMR regression is performed for the distance matrix for all pairs of instances, not a subset of nearest neighbors like NPDR, which makes MDMR susceptible to missing interactions. Its use of local neighborhoods allows NPDR to remove imposter/irrelevant instances from the neighborhood and detect interactions in the higher dimensional space. The ability to remove imposters from the set of nearest neighbors illustrates the "blessings of dimensionality" for Relief-based methods Dawkins *et al.* (2019), but this class of nearest-neighbor methods is still, of course, susceptible to the curses of dimensionality Altman and Krzywinski (2018). Another distinction between methods is that NPDR projects distances onto each attribute, allowing for hypothesis testing of individual attributes (i.e., perform feature selection), whereas MDMR focuses on specified sets of attributes. NPDR uses the context of all attributes to compute nearest neighbors, but it focuses on the projected regression of each attribute at a time and uses the nearest neighbors to allow for detection of interactions. However, NPDR can also be used to compute the importance of sets of factors, like MDMR. An example of this is the penalized version of NPDR that uses the set of all attributes in a nearest-neighbor projected-distance multiple regression.

NPDR can use fixed-k Relief neighborhoods and radius-based Relief neighborhoods. For fixed-k neighborhoods, we expect the NPDR approach will handle imbalanced data in a less biased way than the original fixed-k methods, which focus on hit/miss neighborhoods separately. By identifying the nearest neighbors independently of hit/miss status, the neighborhood should naturally reflect the imbalance in the data. The hit/miss status of each pair is computed separately as a categorical outcome regression variable. This should make NPDR scores with fixed-k ($\mathcal{N}_{k_\alpha}$) similar to fixed radius ($\mathcal{N}_{R_\alpha}$) for balanced and imbalanced data. Power for detecting main effects is highest with the myopic maximum $\mathcal{N}_{k_{\text{max}}}$ ($k_{\text{max}} = \lfloor (m-1)/2 \rfloor$). Real biological data will likely contain a mixture of main effects and epistasis network effects McKinney and Pajewski (2012). NPDR feature selection can be embedded in the backwards elimination of private Evaporative Cooling (privateEC) or in a nested cross-validation for feature selection and classification Le *et al.* (2017) or for optimization of α or k to balance main effect and interaction detection.

A challenge in NPDR analysis is the inherent dependence between neighbors in the projection models, which violates distribution assumptions of regression and leads to artificially lower P-values. Our results indicate that the Bonferroni procedure effectively controls type I error despite deflated P-values. However, future studies are needed to investigate strategies to account for dependence between neighbor-pair observations such as regularization/shrinkage of the regression-coefficientsDe Hertogh *et al.* (2010). A related property of NPDR is that it is often the case that two instances will be in each other's neighborhood. In these cases, we provide an option to restrict $\mathcal{N}$ to contain only unique pairs. However, the extra sampling of neighbor projections may provide beneficial information by reinforcing tightly connected neighborhoods.

The ability to incorporate covariates into NPDR models addresses the important but often neglected issue of confounding factors in machine learning. We applied NPDR to a real RNA-Seq dataset for MDD to demonstrate the identification of biologically relevant genes and the removal of spurious associations by covariate correction. NPDR with sex as a covariate adjustment successfully removed X and Y-linked genes and genes highly expressed in sex organs. However, it is important to note that some genes removed due to a shared association with sex may be important for the pathophysiology of MDD or for classification accuracy. Thus, covariate adjustment in NPDR is a useful option to inform a holistic analysis of a given dataset.

Application to GWAS data required no additional modifications of the algorithm other than specification of a different diff projection operator for categorical variables Arabnejad *et al.* (2018), and the covariate option allows principal components to be included to adjust for population structure. One of the trans-eQTLs found by NPDR in CTNNA3 (rs10997355) for SCAI may suggest a gene-environment interaction due to maternal CMV infection. In addition, it has been suggested that exposure to CMV may increase mood disorder risk through interactions with susceptibility variants Kim *et al.* (2007). However, antibody titers for CMV were not available in this study. To expand variant discovery and demonstrate the ability of NDPR to analyze large GWAS data, we performed a conservative LD pruning that removes some correlation between variants while still leaving a substantial number of informative variants. Because of the omni-genic nature of NPDR, further investigation is needed to understand the effect of LD on the relative ranking of variants and the effect of correlation on nearest-neighbor calculations.

Acknowledgements

Funding

This work was supported in part by the National Institute of Health Grant Nos. GM121312 and GM103456 (to BAM).

References

Altman, N. and Krzywinski, M. (2018). The curse(s) of dimensionality this-month. *Nature Methods*, **15**(6), 399–400.

Arabnejad, M., Dawkins, B., Bush, W., White, B., Harkness, A., and McKinney, B. A. (2018). Transition-transversion encoding and genetic relationship metric in relieff feature selection improves pathway enrichment in gwas. *BioData mining*, **11**, 23.

Børglum, A., Demontis, D., Grove, J., Pallesen, J., Hollegaard, M. V., Pedersen, C., Hedemand, A., Mattheisen, M., Uitterlinden, A., Nyegaard, M., *et al.* (2014). Genome-wide study of association and interaction with maternal cytomegalovirus infection suggests new schizophrenia loci. *Molecular psychiatry*, **19**(3), 325.

Breen, M., Kemena, C., Vlasov, P., Notredame, C., and FA, K. (2012). Epistasis as the primary factor in molecular evolution. *Nature*, **490**(7421), 535–8.

Chen, H., Wang, C., Conomos, M. P., Stilp, A. M., Li, Z., Sofer, T., Szpiro, A. A., Chen, W., Brehm, J. M., Celedón, J. C., *et al.* (2016). Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed models. *The American Journal of Human Genetics*, **98**(4), 653–666.

Dawkins, B. A., Le, T. T., and McKinney, B. A. (2019). Blessings of dimensionality: Theoretical analysis of nearest-neighbor projected-distance methods for detecting interactions in high dimension. *Under Construction*.

De Hertogh, B., De Meulder, B., Berger, F., Pierre, M., Bareke, E., Gaigneaux, A., and Depiereux, E. (2010). A benchmark for statistical microarray data analysis that preserves actual biological and technical variance. *BMC bioinformatics*, **11**(1), 17.

De la Fuente, A. (2010). From differential expression to differential networking–identification of dysfunctional regulatory networks in diseases. *Trends in genetics*, **26**(7), 326–333.

Granizo-Mackenzie, D. and Moore, J. H. (2013). Multiple threshold spatially uniform relieff for the genetic analysis of complex human diseases. In *European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics*, pages 1–10. Springer.

Greene, C. S., Penrod, N. M., Kiralis, J., and Moore, J. H. (2009). Spatially Uniform ReliefF (SURF) for computationally-efficient filtering of gene-gene interactions. *BioData Mining*, **2**, 5.

Kim, J. J., Shirts, B. H., Dayal, M., Bacanu, S.-a., Wood, J., Xie, W., Zhang, X., Chowdari, K. V., Yolken, R., Devlin, B., *et al.* (2007). Are exposure to cytomegalovirus and genetic variation on chromosome 6p joint risk factors for schizophrenia? *Annals of medicine*, **39**(2), 145–153.

Kononenko, I., Šimec, E., and Robnik-Šikonja, M. (1997). Overcoming the Myopia of Inductive Learning Algorithms with RELIEFF. *Applied Intelligence*, **7**(1), 39–55.

Lareau, C. A., White, B. C., Oberg, A. L., and McKinney, B. A. (2015). Differential co-expression network centrality and machine learning feature selection for identifying susceptibility hubs in networks with scale-free structure. *BioData mining*, **8**(1), 5.

Le, T. T., Simmons, W. K., Misaki, M., Bodurka, J., White, B. C., Savitz, J., and McKinney, B. A. (2017). Differential privacy-based evaporative cooling feature selection and classification with relief-f and random forests. *Bioinformatics*, **33**(18), 2906–2913.

Le, T. T., Kuplicki, R. T., McKinney, B. A., Yeh, H.-w., Thompson, W. K., and Paulus, M. P. (2018a). A nonlinear simulation framework supports adjusting for age when analyzing brainage. *Frontiers in aging neuroscience*, **10**.

Le, T. T., Urbanowicz, R. J., Moore, J. H., and McKinney, B. A. (2018b). Statistical inference relief (stir) feature selection. *Bioinformatics*.

Li, L., Rakitsch, B., and Borgwardt, K. (2011). ccsvm: correcting support vector machines for confounding factors in biological data classification. *Bioinformatics*, **27**(13), i342–i348.

Linn, K. A., Gaonkar, B., Doshi, J., Davatzikos, C., and Shinohara, R. T. (2016). Addressing confounding in predictive models with an application to neuroimaging. *The international journal of biostatistics*, **12**(1), 31–44.

Lopes, A. M., Miguel, R. N., Sargent, C. A., Ellis, P. J., Amorim, A., and Affara, N. A. (2010). The human rps4 paralogue on yq11. 223 encodes a structurally conserved ribosomal protein and is preferentially expressed during spermatogenesis. *BMC molecular biology*, **11**(1), 33.

McKinney, B. and Pajewski, N. (2012). Six degrees of epistasis: statistical network models for gwas. *Frontiers in genetics*, **2**, 109.

McKinney, B. A., Crowe, J. E., Guo, J., and Tian, D. (2009a). Capturing the spectrum of interaction effects in genetic association studies by simulated evaporative cooling network analysis. *PLoS genetics*, **5**(3), e1000432.

McKinney, B. A., Crowe Jr, J. E., Guo, J., and Tian, D. (2009b). Capturing the spectrum of interaction effects in genetic association studies by simulated evaporative cooling network analysis. *PLoS genetics*, **5**(3), e1000432.

McKinney, B. A., White, B. C., Grill, D. E., Li, P. W., Kennedy, R. B., Poland, G. A., and Oberg, A. L. (2013). Reliefseq: a gene-wise adaptive-k nearest-neighbor feature selection tool for finding gene-gene interactions and main effects in mrna-seq gene expression data. *PLoS one*, **8**(12), e81527.

Miyashita, A., Arai, H., Asada, T., Imagawa, M., Matsubara, E., Shoji, M., Higuchi, S., Urakami, K., Kakita, A., Takahashi, H., *et al.* (2007). Genetic association of ctnna3 with late-onset alzheimer's disease in females. *Human molecular genetics*, **16**(23), 2854–2869.

Mostafavi, S., Battle, A., Zhu, X., Potash, J. B., Weissman, M. M., Shi, J., Beckman, K., Haudenschild, C., McCormick, C., Mei, R., *et al.* (2014). Type i interferon signaling genes in recurrent major depression: increased expression detected by whole-blood rna sequencing. *Molecular psychiatry*, **19**(12), 1267.

Rao, A., Monteiro, J. M., Mourao-Miranda, J., Initiative, A. D., *et al.* (2017). Predictive modelling using neuroimaging data in the presence of confounds. *NeuroImage*, **150**, 23–49.

Riesselman, A. J., Ingraham, J. B., and Marks, D. S. (2018). Deep generative models of genetic variation capture the effects of mutations. *Nature methods*, **15**(10), 816–822.

Robnik-Šikonja, M. and Kononenko, I. (2003a). Theoretical and empirical analysis of relieff and rrelieff. *Machine learning*, **53**(1-2), 23–69.

Robnik-Šikonja, M. and Kononenko, I. (2003b). Theoretical and empirical analysis of relieff and rrelieff. *Machine learning*, **53**(1-2), 23–69.

Schork, N. J. and Zapala, M. A. (2012). Statistical properties of multivariate distance matrix regression for high-dimensional data analysis. *Frontiers in genetics*, **3**, 190.

Urbanowicz, R. J., Olson, R. S., Schmitt, P., Meeker, M., and Moore, J. H. (2018a). Benchmarking relief-based feature selection methods for bioinformatics data mining. *Journal of Biomedical Informatics*, **85**, 168–188.

Urbanowicz, R. J., Meeker, M., Cava, W. L., Olson, R. S., and Moore, J. H. (2018b). Relief-based feature selection: Introduction and review. *Journal of Biomedical Informatics*.

Wang, K., Zhang, H., Ma, D., Bucan, M., Glessner, J. T., Abrahams, B. S., Salyakina, D., Imielinski, M., Bradfield, J. P., Sleiman, P. M., *et al.* (2009). Common genetic variants on 5p14. 1 associate with autism spectrum disorders. *Nature*, **459**(7246), 528.

Weinreich, D., Lan, Y., Wylie, C., and Heckendorn, R. (2013). Should evolutionary geneticists worry about higher-order epistasis? *Curr Opin Genet Dev.*, **23**(6), 700–7.

Winham, S. J., Colby, C. L., Freimuth, R. R., Wang, X., De Andrade, M., Huebner, M., and Biernacka, J. M. (2012). Snp interaction detection with random forests in high-dimensional genetic data. *BMC bioinformatics*, **13**(1), 164.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **67**(2), 301–320.

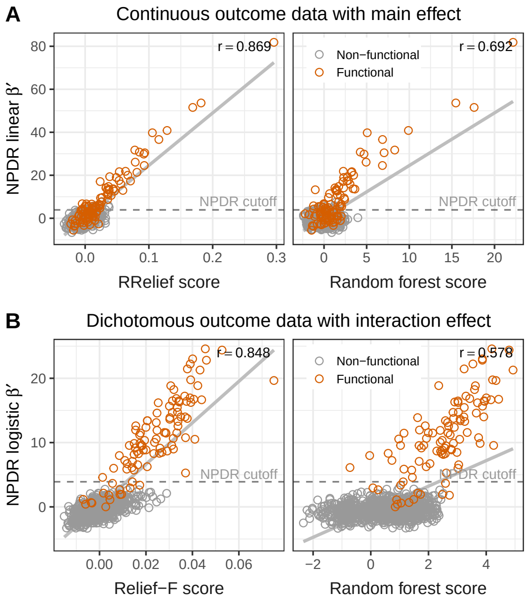

Fig. 2. Simulation comparison of importance scores. Scatter plots of NPDR versus Relief-based scores (left) and NPDR versus random forest scores (right) for representative simulations of continuous outcome data with main effects (A, top row) and dichotomous outcome data with interaction effects (B, bottom row). Simulations use $m = 200$ samples and $p = 1,000$ attributes with 100 functional (orange). For continuous outcome (A), importance scores computed by RRelief weight, random forest percent increase in MSE and NPDR standardized linear regression coefficient (β' from Eq. 8). For dichotomous outcome (B), scores computed by Relief-F, random forest mean decrease in accuracy and NPDR standardized logistic regression coefficient (β' from Eq. 14). A regression line between the scores with correlation r is displayed, and a 0.05 Bonferroni-adjusted cutoff (dashed) is shown for NPDR scores. There is no statistical threshold for Relief-based methods or random forest (area under the precision-recall curve (auPRC) is used to compare algorithm performance, see Fig. 1).