Nearest-neighbor Projected-Distance Regression (NPDR) detects network interactions and controls for confounding and multiple testing

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

February 14, 2019

Abstract

Motivation: We develop a new feature selection technique that uses the generalized linear model (GLM) to perform regression between nearest-neighbor pair distances projected onto attributes to address a broad spectrum of statistical challenges in high-dimensional data, including interactions and confounding variables. Recently we developed STatistical Inference Relief (STIR), a pseudo t-test approach to estimate the statistical significance of Relief-based attribute scores for case-control (classification) problems, where the data may involve main effects and complex statistical interactions (network epistasis). However, efficient statistical inference methods are needed to detect complex network effects in more complicated modeling situations, including continuous outcomes, mixtures of categorical and continuous predictors, and correcting for potential confounding variables.

Methods: Our new Nearest-neighbors Projected-Distance Regression (NPDR) encompasses STIR for case-control data and extends its capabilities to statistically correct for covariates, which previously was not feasible for Relief-based methods. NPDR provides a novel and improved way to compute regression-based Relief scores (RRelief) for quantitative outcomes that also allows statistical corrections and various combinations of predictor data types such as genetic variants and gene expression. In addition, we implement a penalized version of NPDR and derive the theoretical constant-k approximation to the expected number of neighbors for spatially uniform radial neighborhoods.

Results: Using realistic simulations that include main effects and gene-network interactions, we show for that NPDR improves attribute estimates compared to standard Relief while also being able to compute statistical significance of attributes and adjust for covariates. We demonstrate that the NPDR framework has similar statistical properties to the pseudo t-test based method for case-control data with the added flexibility of regression modeling. We also compare glmnet with the penalized version of NPDR. We use RNA-Seq data from a study of major depressive disorder to show that NPDR with covariate adjustment removes spurious associations due to confounding by sex.

Availability: Code and data available at https://insilico.github.io/npdr/.

Contact: brett.mckinney@gmail.com

Supplementary information: Supplementary data are available online.

1 Introduction

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 [1–4]. Epistasis is a measure of the effect of mutations on a phenotype beyond what would be expected by their independent effects. There is evidence that these non-independent effects are pervasive [5] and that higher-order interactions also play an important role in genetics [6].

A similar effect can be observed in differential co-expression, where the phenotypic effect of one gene is modified depending on the expression of another gene [7, 8]. The embedding of these interactions in a regulatory network may lead to, not only pairwise interactions, but also higher-order epistasis network effects. Attempting to model these higher order interactions combinatorially would be computationally and statistically intractable [9]. Thus, computationally scalable feature selection methods, like Relief, are needed to capture these higher-order effects in high-dimensional data such as genome-wide association [10] and gene expression studies [11].

We recently introduced the STatistical Inference Relief (STIR) formalism [11] to address Relief-based methods' lack of a statistical distribution for hypothesis testing and the challenge of assessing the false positive rate of Relief-based scores. STIR extended Relief-based methods to compute statistical significance of attributes in case-control data by reformulating the Relief weight [12] as a pseudo t-test for the difference of means between neighbors with the same and opposite case-control status (hits and misses). The STIR group means are computed from the difference of projected-distance differences between neighbors onto a given attribute dimension. We showed that STIR is an effective approach with high power and low false-positive rates for main effects and interactions. STIR is applicable to any predictor data type (numeric/expression or categorical/SNP); however, being based on a t-test, it does not apply to quantitative trait data and does not correct for covariates. Prior to the current study, no Relief-based method includes covariate correction.

In the current study, we extend STIR to projected distance-based regression with the generalized linear model (GLM). This new Nearest-neighbors Projected-Distance Regression (NPDR) approach opens up Relief-based methods to statistical inference for a broad class of problems by treating the pairwise projected-distance differences among nearest neighbors as the observations in a GLM. The projected difference function between two instances for a given attribute may be simply an arithmetic difference for numeric attributes or may be tailored to the specific data type, such as genetic variants [10]. The NPDR formalism encompasses case-control (classification) problems (Eq. 17), leads to a new and seamless way of solving numeric outcome (regression) problems (Eq. 10) and adjust for covariates (Eq. 13) while detecting main effects and interactions.

Covariate adjustment is important throughout biomedical research to avoid confounding by demographic variables, like age [13] or population stratification [14], but adjusting for covariates is often neglected in machine learning. Relief-based methods also suffer from confounding, but this limitation has not been addressed previously. Some proposed methods for correcting machine learning algorithms include restricted permutation [15], inverse probability weighting of training samples [16] and penalized support vector machines [17]. 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 on an RNA-Seq study of major depressive disorder in which there is a strong signal in the expression data due to the sex of study participants.

For each attribute, the NPDR model fits a GLM of projected-distance differences (onto the attribute) between all pairs of nearest-instance neighbors. For the outcome

variable in the model, the projected distances between instances are binary match/mismatches for case-control data and arithmetic differences for quantitative outcomes. The model is linear for regression problems and logistic for case-control problems, and either model may include terms for covariates. The importance of an attribute is given by the regression coefficient and the statistical significance by its P value. The 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). NPDR is applicable to any predictor data types such as SNPs in GWAS or expression in RNA-Seq. It allows for hypothesis testing and adjustment for multiple testing. Further, we implement a penalized version of NPDR with the elasticnet penalty. We also derive the theoretical constant-k approximation to the expected number of neighbors for SURF and multiSURF radial neighborhoods.

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. For quantitative trait data, we use linear regression, which has theoretical similarities to RRelief, but has better statistical properties and allows for covariate correction. For case-control data, we use a directional logistic model in the GLM and along with covariate terms. We use the projected-distance regression formalism to implement a penalized version of NPDR based on elasticnet, and we derive the theoretical expected constant-k approximation to radius-based neighborhood methods SURF and multiSURF. In the Results, we use realistic simulations with main effects and network interactions to demonstrate the similarities and differences in power and false discoveries between NPDR, STIR and Relief methods. We demonstrate the feature selection similarities between NPDR and penalized NPDR for simulated data. We apply NPDR with covariate correction to a real RNA-Seq dataset from a study of major depressive disorder, showing that NPDR removes spurious associations due to confounding by sex.

2 Materials and Methods

In this section, we first develop the mathematical formalism needed to describe the projected distance regression in NPDR, and we derive the relationship between radial and fixed-k Relief neighborhood methods. We then describe the NPDR models for different situations, including continuous outcome, case-control and covariates, and we describe the simulated and real data sets for method validation.

2.1 Distance metrics and nearest neighbors

Because NPDR and other Relief-based feature selection methods are based on distances, we first describe the algorithms and notation for identifying nearest neighbors in the space all attributes.

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, and the generic $\mathrm{d}_{ij}(a)$ supports any type of attributes (e.g., numeric versus categorical). An example projected difference between two instances i and j for a continuous numeric ($\mathrm{d}^{\mathrm{num}}$) attribute a for NPDR is

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

We use a simplified $\mathrm{d}_{ij}(a)$ notation in place of the $\mathrm{diff}(a,(i,j))$ notation that is customary in Relief-based methods. We omit the division by $\max(a) - \min(a)$ used by Relief to constrain scores to the interval from -1 to 1. As we show in subsequent sections, NPDR scores are [standardized] regression coefficients with corresponding P values, so any scaling operation at this stage is unnecessary for comparing attribute scores. *Omit: The scaling may alleviate bias in the distance calculation. However, standardizing the data matrix X ($\hat{X}$) should have the same effect without division by* $\max(a) - \min(a)$*, which has usual distribution properties for distances (expand).* The numeric $\mathrm{d}_{ij}^{\mathrm{num}}(a)$ projection is simply 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 [10], 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 regression models below (Eqs. 10, 13, 17) will be fit for all nearest neighbors i and j in the defined neighborhood (discussed next).

2.1.2 Nearest-neighbor ordered pairs

For the original ReliefF approach for case-control data, two neighborhood sets are calculated: one for hits and one for misses. 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 [12, 18, 19]. The NPDR neighbors are chosen blind to the outcome variable and then pairs of instances are assigned to hit or miss groups (instances in the same class or different class) for case-control 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 $(i,j) \in \mathcal{N}$ for the projected-distance regression analysis. The ordered pairs constituting the 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 Derivation of expected k for multiSURF neighborhoods

The NPDR algorithm applies to any Relief neighborhood algorithm. In the applications in the current study, we use the multiSURF [19] adaptive radius neighborhood, which

varies with each instance, and we use a fixed-k neighborhood that well-approximates mulitSURF, derived below. The multiSURF radius for an instance is the mean of its distances to all other instances subtracted by $\alpha = 1/2$ of the standard deviation of this mean. Previously we showed empirically for balanced case-control datasets that a good constant-k approximation to the expected number of neighbors within the multiSURF radii is $k = m/6$ [11], where m is the number of samples. Here we derive a more exact theoretical mean that shows the mathematical connection between neighbor-finding methods.

Regardless of the predictor data type (numeric or categorical), the distribution of the p predictors (uniform, Gaussian, or binomial), or the metric used to compute distances (Manhattan or Euclidean), the $m(m-1)/2$ pairwise distances in the p-dimensional space are well approximated by a normal distribution. 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 $\alpha = 1/2$ [20].

The probability of the remaining $m-1$ instances being inside the α-radius of instance i (R_i^α) can be viewed as $m-1$ Bernoulli trials each with a probability of success q_α. Then the average average number of neighbors is given by

$$\bar{k}_\alpha = (m-1)q_\alpha, \tag{7}$$

from the mean of a binomial random variable. To calculate q_α, we assume the distribution of distances ($\{D_{ij}\}_{j \neq i}$) of neighbors of instance i is normal $N(\bar{D}_i, \sigma_{\bar{D}_i})$. Our empirical studies confirm a normal distribution and that it is robust to data type and metric. Extreme violations of independence of attributes (extreme correlations or interactions) will cause the distribution to be right skewed, but this effect is difficult to observe in real data. Thus, for a Gaussian pairwise distance distribution, the probability q_α for one instance $j \neq i$ to be in the neighborhood of i ($j \in N_i^\alpha$) is given by the area under the mean-centered ($\bar{D}_i$) Gaussian from $-\infty$ to R_i^α. **show Gaussian plot illustration?** This integral can be written in terms of the error function (erf):

$$q_\alpha = \frac{1}{2}\left(1 - \mathrm{erf}\left(\frac{\alpha}{\sqrt{2}}\right)\right). \tag{8}$$

And finally using Eqs. (7 and 8) we find

$$\bar{k}_\alpha = \lfloor \frac{m-1}{2}\left(1 - \mathrm{erf}\left(\frac{\alpha}{\sqrt{2}}\right)\right) \rfloor, \tag{9}$$

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 divides this formula by 2. For multiSURF ($\alpha = 1/2$), this formula gives $\bar{k}_{1/2}^{\mathrm{hit/miss}} = \frac{1}{2}\bar{k}_{1/2} = .154(m-1)$, which is very close to our previous empirical estimate $m/6$. 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 performance for simulated data sets. However, the best value for α is likely data-specific and may be determined through nested cross-validation and other parameter tuning methods.

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

2.2.1 Continuous outcomes: linear regression NPDR

Once the set of neighborhoods $\mathcal{N}$ is determined by the distance matrix D_{ij} (Eq. 1) and the choice of neighborhood method (*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 is the attribute's projected distances (e.g., Eq. 2 for numeric attributes) between all pairs of nearest-neighbor instances (Eq. 3), and the outcome is the numeric difference (for quantitative traits) or match/mismatch difference (for case/control data) between all nearest neighbors i and j. For quantitative outcomes, we find the parameters of the following model that minimize the least-squares error over $\forall(i,j) \in \mathcal{N}$:

$$\mathrm{d}_{ij}^{\mathrm{num}}(y) = \beta_o + \beta_a \mathrm{d}_{ij}(a) + \epsilon_{ij}. \tag{10}$$

The $\mathrm{d}_{ij}^{\mathrm{num}}(y)$ term on the left is the projected distances (diff) between instances i and j for 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. The NPDR test statistic is the β_a estimate with null and alternative hypotheses

$$\begin{aligned} H_0 &: \beta_a < 0 \\ H_1 &: \beta_a \geq 0. \end{aligned} \tag{11}$$

The β_a can be interpreted as the predicted amount the quantitative outcome changes between a pair of subjects when the projected difference of the attribute value a changes by one unit. *Should we pre-center the diffs and set $\beta_o = 0$??* The attribute weights in the original RRelief algorithm can be described as a weighted covariance between the attribute neighbor diffs, $\mathrm{d}_{ij}^{\mathrm{type}}(a)$, and the outcome neighbor diffs, $\mathrm{d}_{ij}^{\mathrm{num}}(y)$. The extra weighting in RRelief is an exponentially decaying function of the rank of the distance between neighbors. Because the NPDR attribute weight, β_a, is a regression coefficient, it can also be described as a weighted covariance between attribute and outcome neighbor diffs:

$$\beta_a = \frac{\mathrm{Cov}\left(\mathrm{d}(\mathbf{y}), \mathrm{d}(\mathbf{a})\right)}{\mathrm{Var}\left(\mathrm{d}(\mathbf{a})\right)}. \tag{12}$$

But unlike RRelief, the NPDR covariance is divided by the variance of the outcome diffs. Thus, there is a similarity between NPDR and RRelief for regression, but, as we show shortly, NPDR provides an improved attribute estimation in a flexible framework for correcting for additional sources of variation (i.e., confounding covariates) as well handling case-control outcomes.

2.2.2 Continuous outcomes 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 for the covariate attribute(s) between subjects on the neighborhood ($\forall(i,j) \in \mathcal{N}$) and include an additional projected distance term in the regression model:

$$\mathrm{d}_{ij}^{\mathrm{num}}(y) = \beta_0 + \beta_a \mathrm{d}_{ij}(a) + \vec{\beta}_{\mathrm{covs}}^{T} \mathrm{d}_{ij}(\vec{y}_{\mathrm{covs}}) + \epsilon_{ij}. \tag{13}$$

The above vector notation can be expanded as

$$\vec{\beta}_{\mathrm{covs}}^{T} = \left(\beta_{\mathrm{cov}_1}, \beta_{\mathrm{cov}_2}, \ldots, \beta_{\mathrm{cov}_{p_c}}\right) \tag{14}$$

for the regression coefficients of the p_c covariates and

$$\mathrm{d}_{ij}(\vec{y}_{\mathrm{covs}}) = \left(\mathrm{d}_{ij}^{\mathrm{type}_1}(y_{\mathrm{cov}_1}), \mathrm{d}_{ij}^{\mathrm{type}_2}(y_{\mathrm{cov}_2}), \ldots, \mathrm{d}_{ij}^{\mathrm{type}_{p_c}}(y_{\mathrm{cov}_{p_c}})\right)^T \tag{15}$$

for the projection differences between instances i and j for each of the p_c covariates with the appropriate projection type for each covariate data type (e.g., numeric or categorical). 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. (11).

2.2.3 Binary outcomes with covariates: directional logistic regression NPDR

We now apply the GLM formalism to enable NPDR to handle binary outcome data (e.g., case-control phenotype), which is what the original STIR was designed for. Nonetheless, NPDR can also adjust for covariates. We model 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 find the parameters of the following model that minimize the least-squares error over $\forall(i,j) \in \mathcal{N}$:

$$\mathrm{logit}(p_{ij}^{\mathrm{miss}}) = \beta_0 + \beta_a \mathrm{d}_{ij}(a) + \epsilon_{ij}, \tag{16}$$

or with covariates:

$$\mathrm{logit}(p_{ij}^{\mathrm{miss}}) = \beta_0 + \beta_a \mathrm{d}_{ij}(a) + \vec{\beta}_{\mathrm{covs}}^T \mathrm{d}_{ij}(\vec{y}_{\mathrm{covs}}) + \epsilon_{ij}, \tag{17}$$

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 being modeled by probability p_{ij}^{miss} is a binary diff between subjects for the phenotype (y):

$$\mathrm{d}_{ij}^{\mathrm{miss}}(\vec{y}) = \begin{cases} 0, & y_i == y_j \\ 1, & \text{else.} \end{cases} \tag{18}$$

The β_a importance score can be interpreted in the following way. For a unit increase in the difference in the value between two neighbors, we predict a change of e^{β_a} in the odds of the neighbors being in opposite classes. For case-control data, we are specifically interested in the alternative hypothesis that $\beta_a > 0$ because negative β_a values represent irrelevant attributes. Thus, we are interested in testing directional null and alternative hypotheses

$$\begin{aligned} H_0 &: \beta_a \leq 0 \\ H_1 &: \beta_a > 0. \end{aligned} \tag{19}$$

When there are no covariates for case-control data, the STIR (based on a pseudo t-test) and NPDR (based on regression with a directional logit model) are equivalent given reasonable distribution assumptions. But, of course, NPDR is a more flexible framework. In the current notation, the STIR null and alternative hypotheses would be

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

where

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

and the test statistic is a pseudo t-test (see [11]).

2.2.4 Regularized NPDR

We now propose a regularized NPDR approach that combines all of the attribute difference vectors into one design matrix and constrains the optimization of the coefficients with an elastic net penalty [21] for feature selection. Specifically, we minimize the vector of regression coefficients, $\vec{\beta}_A$ for all attributes $a \in A$, that is $\vec{\beta}_{\mathrm{A}}^T = (\beta_{a_1}, \beta_{a_2}, \ldots, \beta_{a_p})$, subject to the elastic net penalty:

$$\min_{\beta_o, \vec{\beta}_A} \frac{1}{|\mathcal{N}|} \sum_{i,j \in \mathcal{N}} \mathcal{L}\left(\mathrm{d}_{ij}^{\mathrm{miss}}(y), \beta_0 + \vec{\beta}_A^T \mathrm{d}_{ij}(A)\right) + \lambda \left[\alpha ||\vec{\beta}_A||_1 + ||\vec{\beta}_A||_2^2 (1-\alpha)/2\right].$$

$\mathcal{L}$ 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 $a \in A$:

$$\mathrm{d}_{ij}(A) = \left(\mathrm{d}_{ij}^{\mathrm{type}_1}(a_1), \mathrm{d}_{ij}^{\mathrm{type}_2}(a_2), \ldots, \mathrm{d}_{ij}^{\mathrm{type}_p}(a_p)\right)^T \tag{22}$$

The elastic-net parameter α mixes the amount of lasso (α=1) and ridge (α=0) penalty. Our implementation allows any value of α, but we use lasso as the default to choose one predictor (attribute diff vector) when a set of predictors is correlated. The overall penalty strength λ is chosen by cross-validation. *Could set constant λ as option.* For case-control, we use binomial link function for the hit/miss projected distances in the likelihood optimization, which leads to elastic net treating positive and negative β_{a_i} coefficients similarly when shrinking. One can simply remove any negative coefficients after shrinkage, but a better approach may be to modify the likelihood to ordinal or directional logistic regression.

2.2.5 Comparing NPDR to existing Relief-based methods

We examine the slight differences between the importance scores computed from NPDR, STIR and traditional Relief-based methods. Roughly speaking, the formulas for importance scores become more generalized over time (Table 1). Specifically, for binary outcome, STIR incorporates sample variance of the nearest neighbor distances to enable the calculation of statistical significance with the assumptions of a t-test, while NPDR assumes intra- and inter-class differences are randomly sampled from one distribution and computes the importance score from a logistic regression. This generalization enables NPDR to have desirable properties and be applicable to a wider range of problems.

	Traditional methods	STIR	NPDR
Importance score W[1]	$\bar{M}_a - \bar{H}_a$	$\frac{\bar{M}_a - \bar{H}_a}{S_p[M,H]\sqrt{\frac{1}{\lvert M\rvert} + \frac{1}{\lvert H\rvert}}}$	$\frac{\mathrm{Cov}\left(\mathrm{p}_y, \mathrm{d}_a\right)}{\mathrm{Var}\left(\mathrm{d}_a\right)}$[2]
W has a distribution	✗	✓	✓
Supports continuous outcome	✓	✗	✓
Supports covariates	✗	✗	✓

Table 1. Comparison of NPDR, STIR and traditional Relief-based methods.

[1] Considering a binary outcome problem.
[2] where $\mathrm{d}_a = [M_a, H_a]$ and $\mathrm{logit}(\mathrm{p}_y) = [M_y, H_y]$.

2.3 Real and simulated datasets

2.3.1 Simulation methods

To compare power and false positive performance for multiple feature selection methods, we use the simulation tool from our private Evaporative Cooling (privateEC) software [22] that was designed to simulate realistic main effects, correlations, and interactions found in gene expression or resting-state fMRI correlation data. In the current study, we first simulate main effect data with $m = 100$ subjects (50 cases and 50 controls) and $p = 1000$ real-valued attributes with 10% functional (true positive association with outcome). We chose a sample size consistent with real gene expression data but on the smaller end to demonstrate a more challenging scenario. Likewise, an effect size bias of $b = 0.8$ was selected to be sufficiently challenging with power approximately 40%.

For interactions, we use the differential co-expression network-based simulation tool in privateEC, which is described in Refs. [7, 22]. We first induce a co-expression network on an Erdős-Rényi random graph with 0.1 attachment probability, which is near the critical value for a giant component. We give connected genes a higher average correlation, approximately $r_{\text{connected}} = 0.8$. This connection correlation induces interaction effect because the interaction is created by disrupting the correlation of these connections. Finally, we simulate functional effects on the outcome of functional variables by permuting their values in cases but leaving the controls unpermuted, thereby leading to a final differential correlation network.

All resulting p-values (from STIR and NPDR) are adjusted for multiple testing using the [Benjamini-Hochberg] procedure [23]. Attributes with adjusted p-values less than 0.05 are counted as a positive test (null hypothesis rejected), else the test is negative. [Variables that are not shrunk to zero by regularization (glmnet and regularized NPDR) are counted as positive tests.] We remark that, even though the terminology used here is very similar to traditional sample classification problems, we instead focus on evaluating the scoring ability of each method on the attributes. We assess the 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 compared to non-functional ones, the precision and recall are useful and objective measures to assess how good the methods are in assigning higher scores to the correct functional attributes.

2.3.2 RNA-Seq dataset with confounding

To test the ability of NPDR to correct for confounding, we use the RNA-Seq study in Ref. [24] that consists of RNA-Seq measurements of 15,231 genes in 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 ($p = 3.89e-7$). We analyzed 4,570 genes that passed a low variation filter (genes with the lowest $[70]^{th}$ percentile variation removed). We applied NPDR with multiSURF neighborhood and tested the 4,570 high variation genes with and without sex as a covariate. We applied univariate logistic regression with and without sex as a covariate to these genes, and we directly tested each gene for association with sex. All P values were adjusted with the Bonferroni procedure.

3 Results

For case-control data, we compare the performance of NPDR using directional logistic model (Eq. 17) with the original STIR (which is based on a t-test). We simulate main effect and interaction effect data sets, and compare based on adjusted P values. We also

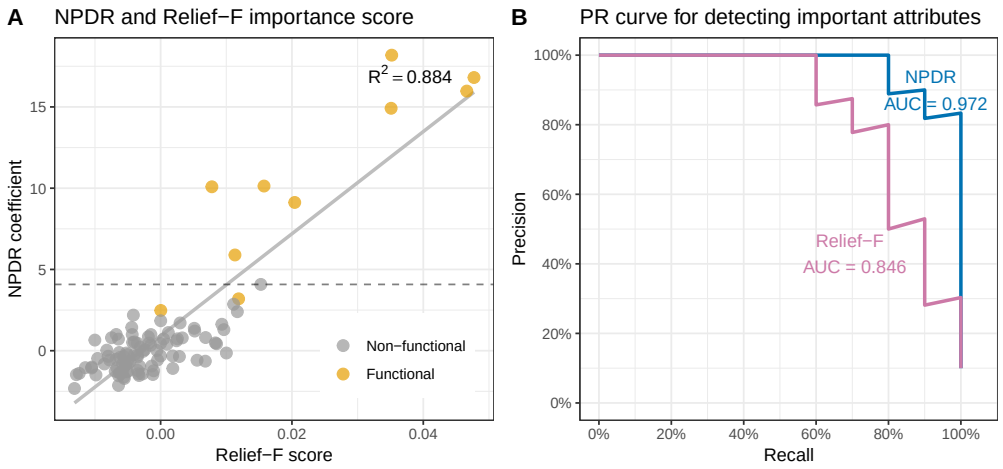

Fig 1. Comparison of NPDR and Relief-F results. Importance score computed from the Relief-F and NPDR algorithm (standardized β coefficient, Table 1) from one replicate of simulation with $m = 100$ samples and $p = 100$ attributes (A). Area under the precision-recall curve used to assess algorithm performance (B).

compared NPDR regression coefficients with standard Relief scores because the latter does not have an associated P value. We use the multiSURF neighborhood $\mathcal{N}_{\alpha=1/2}$ for these methods. Similarly, for continuous outcome simulations, we compare NPDR regression coefficients with standard RRelief scores. We use the fixed-k neighborhood that approximates multiSURF. Compare Glmnet cross-validated hyperparameters with glmnet-NPDR? Interactions have correlation built in; it would be interesting to do main effect simulations with and without correlation (maybe another paper to test the effect of correlation). Compare with random forest importance (fairly common in the machine learning field). Implement the new Cholesky correlation – maybe later. Include a Random Forest importance score comparison. I wouldn't mind showing Random Forest can't find interactions again (although it could because of the small number of background attributes).

3.1 Simulation comparison for case-control outcome data

In simulations with $m = 100$ samples and $p = 100$ attributes, while the logistic regression and t-test have slightly different assumptions on the distribution of samples, NPDR and STIR yielded very similar results ($R^2_P = 0.999$, see Supplementary Fig. S1). [NPDR vs STIR here]

In one replicate of simulation with $m = 100$ samples and $p = 100$ attributes, we show that NPDR correctly detects 8 out of 10 functional attributes with a Bonferroni cutoff (Fig 1). On the other hand, because Relief-F does not produce a threshold, in practice, this value is arbitrarily selected. While the importance scores computed from the two methods are similar ($R^2 = 0.884$), there is a greater separation between the functional and nonfunctional group on the NPDR coefficient compared to Relief-F score. In other words, if a vertical line was drawn as an arbitrary Relief-F score cutoff that yields a reasonable number of correct functional attributes, it will include at least several incorrect ones. The precision-recall curve clearly demonstrates NPDR outperforms Relief-F in this simulated dataset. Across 100 simulations, NPDR yields significantly higher auPRC ($P < 0.0001$, see Supplementary Fig. S2).

For a dataset of the size simulated in our study ($m = 100$ samples and $p = 100$ attributes), NPDR has a [X]-second runtime on a desktop with an Intel Xeon W-2104

CPU and 32GB of RAM.

3.2 Simulation comparison for continuous outcome data

For simulated data with continuous outcomes, we compare NPDR using a linear model (Eq. 10) with standard RReliefF from CORElearn, and univariate linear regression. Because CORElearn does not include the adaptive multiSURF neighborhood, we use a fixed-k neighborhood $\mathcal{N}_{\bar{k}_{1/2}}$ for NPDR and RReliefF. The value $\bar{k}_{1/2} = 30$ (Eq. 9) is the expected number of nearest neighbors corresponding to a multiSURF neighborhood. Just do main effects (simulating continuous outcomes with interactions is non-trivial) but with and without correlation? It is important to note CORElearn and other standard Relief-based methods do not have P values, which is one of the advantages of NPDR. Thus, we compare these methods based on their power to detect the functional variables based for a sliding threshold of top scores. A threshold makes it difficult to compare with glmnet because it is harder to specify a number of selected variables with glmnet. To address this and showcase P values, we could do a separate comparison with NPDR with P value, linear regression, and glmnet and regularized NPDR.

3.3 Real-world RNA-Seq data with confounding

TODO: - Supplementary table of genes with expression associated with sex ([339] adjusted P value less than .05). - Also show univariate analysis with and without adjustment in supplement?

With multiSURF neighborhood $\mathcal{N}_{\alpha=1/2}$, we apply NPDR with and without adjusting for the sex covariate. Several genes including UTY, PRKY, and USP9Y are removed from the list of MDD-associated genes when you adjust for the sex covariate, which indicates suprious associations between these genes and diagnostic phenotype due to confounding by sex. These genes are Y-linked (i.e., Y chromosome (male)) genes and mainly expressed in testis. For example, the RPS4Y2 ribosomal protein S4 Y-linked 2 has been shown by tissue specific studies to mainly express in prostate and testis [25]. Meanwhile, highly expressed in the brain as well as testis, the gene C2orf55 (KIAA1211L) is associated with sex but remains in the adjusted NPDR list ($p_{\text{adj}} < 0.05$) and may be relevant to MDD pathophysiology. [other genes here] In summary, this comparison suggests the NPDR with covariate adjustment effectively removes sex-related confounding variables.

Sensitivity to confounders go beyond nearest-neighbors feature selection algorithms. Lacking the ability to include covariates, even powerful methods like random forest are expected to produce attribute importance scores that are biased with respect to sex. Like previous feature selection methods before STIR, importance scores produced by random forest does not follow a distribution. Hence, instead of P values, we compare NPDR standardized β coefficient and random forest Gini-based importance score (Fig. 3). Most genes with high random forest importance score but low NPDR coefficient are highly associated with sex (dark pink - purple). However, there is a consensus between the two methods in giving high scores for genes such as [].

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 allows it to detect interacting attributes, an ability it shares with Relief-based methods, but NPDR is a departure from Relief in four ways. (1) For feature selection with a continuous outcome, it does not rely on the

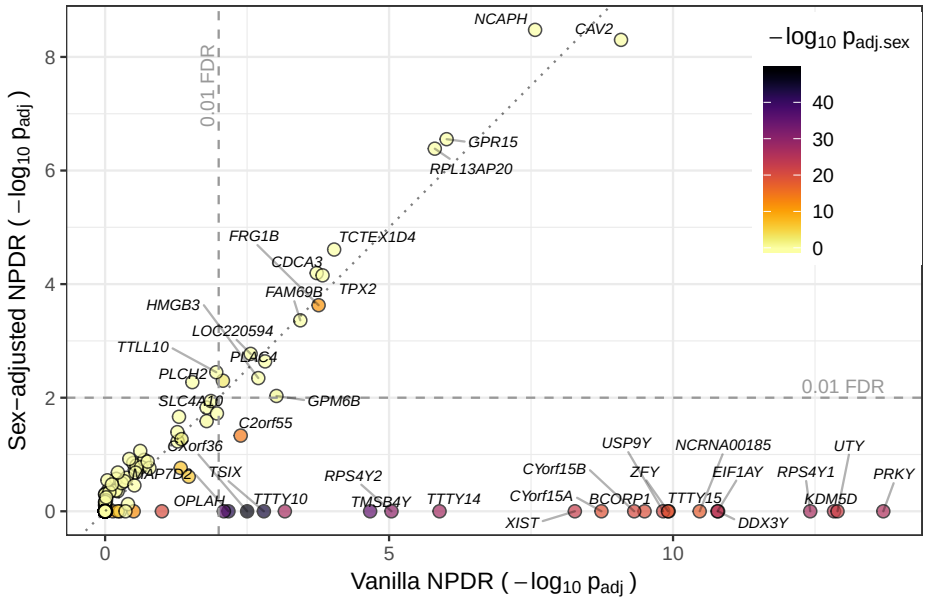

Fig 2. To adjust or not to adjust. Gene scatter plot of $-\log_{10}$ adjusted significance for association with major depressive disorder using NPDR without correction for sex (horizontal axis) and with correction for sex (vertical axis). NPDR without sex correction finds [47] genes that are significant at the Bonferroni-adjusted 0.05 level (above horizontal dashed line) [28] of which are also significantly associated with sex ($p^{\text{sex}}_{\text{adj}} < 0.05$). NPDR with adjustment for sex finds [23] genes that are significant at the Bonferroni-adjusted 0.05 level (right of vertical dashed line) only 3 of which are also significantly associated with sex, thus eliminating most of the confounding variables.

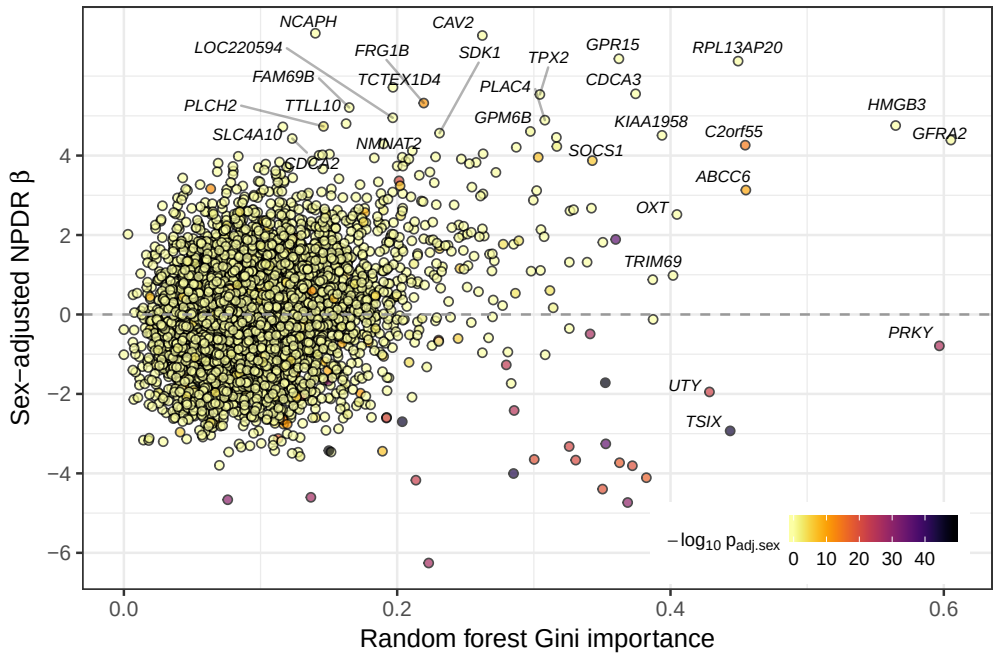

Fig 3. NPDR vs random forest feature importance. Gene scatter plot of importance score for association with major depressive disorder using random forest and NPDR with correction for sex. Many genes with high random forest importance score also have high association with sex. Genes are labeled if they are NPDR significance or have Gini importance > 0.4.

idea of a hit/miss group like current RRelief approaches [19]. Rather, NPDR simply does a regression between the outcome and attribute projected distances. (2) For feature selection with case-control outcomes, NPDR uses a directional 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 (case-control or regression) and predictor data type, NPDR computes the statistical significance of attribute importance scores, which allows for statistically based thresholds that can adjust for multiple hypothesis testing. (5) 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 regression formalism of NPDR is a novel way to perform RRelief by defining the attribute importance as the regression coefficient between the projected attribute differences and the numeric outcome differences between neighbors. For linear regression, NPDR shares some similarity with the original RRelief algorithm because the regression coefficient is related to the correlation coefficient. The RRelief importance score is a weighted correlation between attribute and outcome differences, while the NPDR regression coefficient is a covariance between attribute and outcome differences divided by the variance in the outcome. The NPDR model may also include an offset if the diffs are not centered and the model may include correction terms for covariates that may be confounding.

The original formulation of Relief was for case-control data, which lends itself naturally to nearest hit (same class) and miss (opposite class) neighbors. As discussed above, for regression problems, the original regression Relief (RRelief) score was cast as

a weighted correlation between outcome and the attribute diff [26]. In a different approach, Ref. [19] 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 and has 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 effects.

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 the original STIR, which is specific to case-control data. This indicates that NPDR, which models hit/miss differences between neighbors with a directional logit link, can be safely used instead of STIR with the added benefit of covariance correction and the analysis of quantitative traits. Comparison of NPDR for quantitative outcomes...

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. It is important to note that some genes removed due to a shared association with sex may be important for the pathophysiology of MDD (C2orf55 (KIAA1211L)) or for classifiers. Thus, covariate adjustment in NPDR is a useful option to inform a holistic analysis of a given dataset. Further improvements in the NPDR covariate adjustment may be achieved by correcting the distance matrix calculation. Adding a covariate term may not be sufficient in some cases because the neighborhood could be strongly influenced by confounding genes due to the curse of dimensionality. The curse of dimensionality can also affect confounding in distance matrix calculation.

We also derived the theoretical mapping between the average k expected for fixed-k Relief neighborhoods and radius-based Relief neighborhoods. One reason this relationship is useful is to compare with implementations of Relief that do not include the multiSURF neighborhood. 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 $k = k_{\max} = \lfloor (m-1)/2 \rfloor$. Real biological data will likely contain a mixture of main effects and epistasis network effects [27]. STIR feature selection could be embedded in the backwards elimination of private Evaporative Cooling (privateEC) for feature selection and classification [22] or embedded in a nested cross-validation approach. Nested CV and privateEC can also return classification and optimize α or k.

Note to self. glmnet NPDR and using the $\mathrm{d}_{ij}^{(\mathrm{type})}(A)$ design matrix of all attributes to create a $p \times p$ nearest-neighbor based gene network for hub and module based on the correlation between genes based on the projected distances vectors. Like STIR before it, NPDR P values suffer from deflation due to dependence among the nearest neighbors in the neighborhood, $\mathcal{N}$. The FDR procedure does a good job of selecting the true positives from the false positives (so I think it's not a really big concern), but it would be nice to address the dependence of neighbors in $\mathcal{N}$. One approach would be to create a batch-like indicator variable with m states (a lot), indicating from which instance i a batch of neighbors comes from. This neighbor batch could be used as a covariate in

NPDR. Really case and controls should be evenly distributed within each "batch," so I'm not sure this would solve the problem. Another thought is to create a weight variable that counts how often each pair happens and use this as a covariate. Another possible approach would be to use principle components of the design matrix (might require fixed-k) as covariates in the NPDR regression model. The hypothesis is that the PCs would capture the neighborhood structure within $\mathcal{N}$. Requires some thought on how to map down to PCs with m samples from all the neighbor pairs.

A related distance-based regression method is Multivariate Distance Matrix Regression (MDMR) [28]. 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, which makes it susceptible to missing interactions. Also MDMR focuses on sets of factors, whereas NPDR projects distances onto each attribute, allowing for hypothesis testing of individual attributes (i.e., perform feature selection). While NPDR uses the context of all attributes to compute nearest neighbors, it focuses on the projected regression of each attribute at a time and uses the nearest neighbors to allow for detection of interactions. The ability to remove imposters from the set of nearest neighbors illustrates the blessings of dimensionality for Relief-based methods, but this class of nearest-neighbor methods is still, of course, susceptible to the curses of dimensionality [29]. NPDR can also be used to compute the importance of sets of factors. The penalized version of NPDR uses the set of all attributes in a multiple projected-distance regression.

Multi-state categorical outcomes can be analyzed (similar to multi-state ReliefF) with NPDR by grouping all miss types together. This can be improved by using multinomial regression in NPDR. Application to GWAS data requires no additional modifications of the algorithm other than specification of a different diff function for categorical variables [10], and the covariate option allows for principle components to be included to adjust for population structure.

Acknowledgements

Funding

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

References

1. Ryan J. Urbanowicz, Melissa Meeker, William La Cava, Randal S. Olson, and Jason H. Moore. Relief-based feature selection: Introduction and review. *Journal of Biomedical Informatics*, 2018.

2. Igor Kononenko, Edvard Šimec, and Marko Robnik-Šikonja. Overcoming the Myopia of Inductive Learning Algorithms with RELIEFF. *Applied Intelligence*, 7(1):39–55, January 1997.

3. Brett A. McKinney, James E. Crowe, Jingyu Guo, and Dehua Tian. Capturing the spectrum of interaction effects in genetic association studies by simulated evaporative cooling network analysis. *PLoS genetics*, 5(3):e1000432, March 2009.

4. Kenji Kira and Larry A. Rendell. The feature selection problem: Traditional methods and a new algorithm. In *Proceedings Tenth National Conference on Artificial Intelligence*, pages 129–134. AAAI Press/The MIT Press, 1992.

5. MS Breen, C Kemena, PK Vlasov, C Notredame, and Kondrashov FA. Epistasis as the primary factor in molecular evolution. *Nature*, 490(7421):535–8, 2012.

6. DM Weinreich, Y Lan, e CS Wyli, and RB Heckendorn. Should evolutionary geneticists worry about higher-order epistasis? *Curr Opin Genet Dev.*, 23(6):700–7, 2013.

7. Caleb A Lareau, Bill C White, Ann L Oberg, and Brett A McKinney. 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, 2015.

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

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

10. Marziyeh Arabnejad, BD Dawkins, WS Bush, BC White, AR Harkness, and Brett A McKinney. Transition-transversion encoding and genetic relationship metric in relieff feature selection improves pathway enrichment in gwas. *BioData mining*, 11:23, 2015.

11. Trang T Le, Ryan J Urbanowicz, Jason H Moore, and Brett A McKinney. Statistical inference relief (stir) feature selection. *Bioinformatics*, page bty788, 2018.

12. Brett A McKinney, Bill C White, Diane E Grill, Peter W Li, Richard B Kennedy, Gregory A Poland, and Ann L Oberg. 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, 2013.

13. Trang T Le, Rayus T Kuplicki, Brett A McKinney, Hung-wen Yeh, Wesley K Thompson, and Martin P Paulus. A nonlinear simulation framework supports adjusting for age when analyzing brainage. *Frontiers in aging neuroscience*, 10, 2018.

14. Han Chen, Chaolong Wang, Matthew P Conomos, Adrienne M Stilp, Zilin Li, Tamar Sofer, Adam A Szpiro, Wei Chen, John M Brehm, Juan C Celedón, et al. 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, 2016.

15. Anil Rao, Joao M Monteiro, Janaina Mourao-Miranda, Alzheimer's Disease Initiative, et al. Predictive modelling using neuroimaging data in the presence of confounds. *NeuroImage*, 150:23–49, 2017.

16. Kristin A Linn, Bilwaj Gaonkar, Jimit Doshi, Christos Davatzikos, and Russell T Shinohara. Addressing confounding in predictive models with an application to neuroimaging. *The international journal of biostatistics*, 12(1):31–44, 2016.

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

18. Casey S. Greene, Nadia M. Penrod, Jeff Kiralis, and Jason H. Moore. Spatially Uniform ReliefF (SURF) for computationally-efficient filtering of gene-gene interactions. *BioData Mining*, 2:5, September 2009.

19. Ryan J. Urbanowicz, Randal S. Olson, Peter Schmitt, Melissa Meeker, and Jason H. Moore. Benchmarking relief-based feature selection methods for bioinformatics data mining. *Journal of Biomedical Informatics*, 85:168–188, 2018.

20. Delaney Granizo-Mackenzie and Jason H Moore. 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, 2013.

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

22. Trang T Le, W Kyle Simmons, Masaya Misaki, Jerzy Bodurka, Bill C White, Jonathan Savitz, and Brett A McKinney. Differential privacy-based evaporative cooling feature selection and classification with relief-f and random forests. *Bioinformatics*, 33(18):2906–2913, 2017.

23. Yoav Benjamini, Dan Drai, Greg Elmer, Neri Kafkafi, and Ilan Golani. Controlling the false discovery rate in behavior genetics research. *Behavioural brain research*, 125(1-2):279–284, 2001.

24. Sara Mostafavi, Alexis Battle, Xiaowei Zhu, James B Potash, Myrna M Weissman, Jianxin Shi, Kenneth Beckman, Christian Haudenschild, Courtney McCormick, Rui Mei, et al. Type i interferon signaling genes in recurrent major depression: increased expression detected by whole-blood rna sequencing. *Molecular psychiatry*, 19(12):1267, 2014.

25. Alexandra M Lopes, Ricardo N Miguel, Carole A Sargent, Peter J Ellis, António Amorim, and Nabeel A Affara. 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, 2010.

26. Marko Robnik-Šikonja and Igor Kononenko. Theoretical and empirical analysis of relieff and rrelieff. *Machine learning*, 53(1-2):23–69, 2003.

27. Brett McKinney and Nicholas Pajewski. Six degrees of epistasis: statistical network models for gwas. *Frontiers in genetics*, 2:109, 2012.

28. Nicholas J Schork and Matthew A Zapala. Statistical properties of multivariate distance matrix regression for high-dimensional data analysis. *Frontiers in genetics*, 3:190, 2012.

29. Naomi S. Altman and Martin Krzywinski. The curse(s) of dimensionality this-month. *Nature Methods*, 15(6):399–400, 6 2018.