Supporting Information: Text

Extended Results from *Arabidopsis thaliana* QTL Study

In the main text, we apply our nonparametric variable selection approach to a quantitative trait loci (QTL) association mapping study focused on the characterization of complex traits in *Arabidopsis thaliana*. Specifically, the data consists of $n = 420$ F6 plants from a Bay-0 $\times$ Shahdara recombinant inbred lines (RILs) population that were genotyped for $p = 69$ microsatellite markers and phenotyped for twenty different quantitative traits (Loudet et al., 2002). We limit the scope of our analyses to six of these outcomes including: amino acid (AA) content, shoot dry matter (DM), flowering time in long days (FLOLD), nitrate content (NO), nitrogen content percentage (NP), and sulfate content (SO). Overall, the Gaussian process (GP) regression model identified a total of 31 of the 69 genetic markers to have some moderately significant "RelATive cEntrality" (RATE) measures with at least one of the six analyzed phenotypes. We detail the biologically relevant findings here:

Amino Acid (AA) Content. Our nonparametric variable selection method identified 13 genetic markers as central covariates when analyzing amino acid content. The three most important of these variants are IND2188 and T27K12 on the first chromosome, and ATHCHIB2 on the third chromosome. Note that while the QTL marked by T27K12 has been previously detected by single-test linear models as having a significant association with this particular trait (Loudet et al., 2003), this is not the case for variants IND2188 and ATHCHIB2. The variant IND2188 has been suggested to be associated with primary root length (Bouteillé, 2011), and ATHCHIB2 is linked to the control of variation in sulfate content (Loudet et al., 2007; Koprivova et al., 2013). However, it is worth mentioning that there is believed to be a strong correlation between these two functionalities and amino acid makeup (e.g. Leustek, 2002). This is contextually relevant in this data set as well (see Figure S18).

Shoot Dry Matter (DM). In this trait, neither the GP regression model nor the Bayesian variable selection method had any cohesion or obvious patterns in the genetic markers that they determined to be significant for shoot dry matter. This is most likely due to the fact that shoot dry matter is the least heritable trait among the six phenotypes. Nonetheless, according to our simulations, we know our method to outperform simple linear models in prioritizing covariates during these situations. Much like what has been indicated in previous studies (e.g. Loudet et al., 2003), our approach found significantly associated QTL to be distributed on all chromosomes. Specifically, these were marked by variables ATHCHIB2, MSAT305754, MSAT4.18, IND2188, MSAT3.21, and T27K12.

Flowering Time in Long Days (FLOLD). Flowering time was initially utilized as an ideal trait for which to study genotypic variation in the Bay-0 $\times$ Shahdara RILs population (Loudet et al., 2002). In the main text, the competing GP regression, Lasso regularization, and Bayesian variable selection methods were shown to have the most overlap in their results for this phenotype. According to the GP regression model, the three markers with the greatest centrality measures when explaining the variance in flowering time are located on the fourth chromosome: MSAT4.8, MSAT4.39, and MSAT4.43, respectively. The genetic distance interval corresponding these markers was identified in the original Bay-0 $\times$ Shahdara RILs study as being the region containing the extensively validated gene *FRIGIDA (FRI)* (Johanson et al., 2000). Contextually, *FRI* is known to have an epistatic interaction with the gene *FLOWERING LOCUS C (FLC)* (Caicedo et al., 2004). More specifically, *FRI* has been shown to repress flowering ability by promoting the expression of the floral repressor *FLC* (Lee et al., 1994; Sheldon et al., 1999). Alternatively, the process of vernalization then accelerates flowering by down-regulating *FLC* expression and implicitly antagonizing the effect of *FRI* (Shindo et al., 2005). This collective evidence greatly suggests that these findings by our nonparametric approach may be a true positives.

Nitrate Content (NO) and Nitrogen Content Percentage (NP). Analyses for the two nitrogen based traits revealed the exact same three most central markers, all of which are located on the third and fourth chromosomes (MSAT305754, MSAT3.21, and MSAT4.18). This result comes as no surprise since the

genetic correlation between the nitrogen content and percentage has been estimated to be approximately 0.80 (again see Figure S18). The genetic variant MSAT3.21 has been previously indicated to be significantly associated with variation in *Arabidopsis thaliana* nitrogen percentage (Loudet et al., 2003). Similarly, the marker MSAT4.18 is in the same region that neighbors the locus corresponding to the primary root length gene *PRL3* (Reymond et al., 2006). This is also relevant here since nitrogen content is an essential soil nutrient for plant growth.

Sulfate Content (SO). The GP regression model determined 13 different variables to be strongly associated with sulfate content. A good proportion of these identified genetic markers are on the third chromosome and headlined by the variants: MSAT305754, ATHCHIB2, dCAPsAPR2, and MSAT3.99. These results are consistent with previous analyses of this trait where the enrichment of this particular region has been linked to the sulfate assimilation gene *ATPS1* (Koprivova et al., 2013) and experimentally validated as a factor that is responsible for the variation in sulfate levels (Loudet et al., 2007). The next two significant centrality measures were marked on the first chromosome by variants IND2188 and NGA128. The former has been suggested to be associated with primary root length (Bouteillé, 2011). The latter, has been experimentally validated as a factor explaining the variation in nitrogen content (Loudet et al., 2003).

Supporting Information: Algorithmic Overview

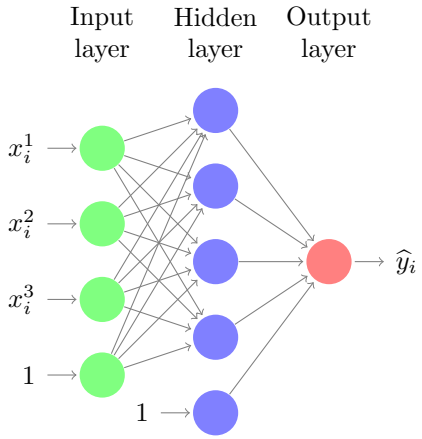
Algorithm 1 Gaussian Process Regression (GPR)

Select a positive definite covariance function $k(\mathbf{x}_i, \mathbf{x}_j)$ where $\mathbf{x}_i$ and $\mathbf{x}_j$ are n-dimensional vectors from the design matrix.
Construct the $n \times n$ covariance matrix $\mathbf{K}$.
Define the full model where $\mathbf{y} = \mathbf{f} + \boldsymbol{\varepsilon}$ and $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \tau^2\mathbf{I})$.
Specify the prior distributions $\mathbf{f} \sim \mathcal{N}(\mathbf{0}, \mathbf{K})$ and $\tau^2 \sim$ Scale-Inv-$\chi(a, b)$.
Run the Gibbs Sampler (T Iterations).
for $t = 1 \to T$ **do**
    $\mathbf{f} \,|\, \tau^2, \mathbf{y} \sim \mathcal{N}(\mathbf{m}^*, \mathbf{V}^*)$ where $\mathbf{m}^* = \mathbf{K}(\mathbf{K} + \tau^2\mathbf{I})^{-1}\mathbf{y}$ and $\mathbf{V}^* = \mathbf{K} - \mathbf{K}(\mathbf{K} + \tau^2\mathbf{I})^{-1}\mathbf{K}$;
    $\tau^2 \,|\, \mathbf{f}, \mathbf{y} \sim$ Scale-Inv-$\chi^2(a^*, b^*)$ where $a^* = a + n$ and $b^* = a^{*-1}[ab + (\mathbf{y} - \mathbf{f})^\intercal(\mathbf{y} - \mathbf{f})]$;
    $\widetilde{\boldsymbol{\beta}} = \mathbf{X}^\dagger\mathbf{f}$.
end for
Calculate the empirical mean, covariance, and precision of the posterior distribution $p(\widetilde{\boldsymbol{\beta}} \,|\, \mathbf{y})$ as $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, and $\boldsymbol{\Lambda}$, respectively.
Compute the centrality of every p predictor via Kullback-Leibler Divergence (KLD).
for $j = 1 \to p$ **do**
    $\text{KLD}(\widetilde{\beta}_j) = \frac{1}{2}\left[-\log(|\boldsymbol{\Sigma}_{-j}\boldsymbol{\Lambda}_{-j}|) + tr(\boldsymbol{\Sigma}_{-j}\boldsymbol{\Lambda}_{-j}) + 1 - p + \alpha_j(\widetilde{\beta}_j - \mu_j)^2\right]$.
end for
Scale each centrality measure for the p predictors to determine their relative importance.
for $j = 1 \to p$ **do**
    $\text{RATE}(\widetilde{\beta}_j) = \text{KLD}(\widetilde{\beta}_j) / \sum \text{KLD}(\widetilde{\beta}_\ell)$.
end for

Algorithm 2 Bayesian Kernel Ridge Regression (BKRR)

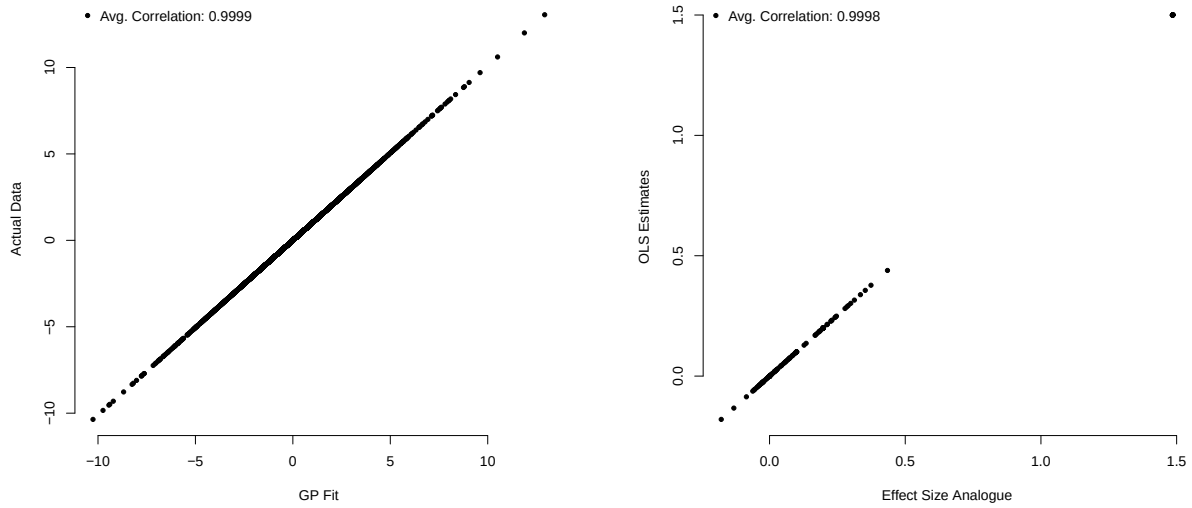
Select a positive definite covariance function $k(\mathbf{x}_i, \mathbf{x}_j)$ where $\mathbf{x}_i$ and $\mathbf{x}_j$ are n-dimensional vectors from the design matrix.
Construct the $n \times n$ covariance matrix $\mathbf{K}$.
Define the full model where $\mathbf{y} = \mathbf{K}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$ and $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \tau^2\mathbf{I})$.
Specify the prior distributions $\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{K}^{-1})$ and $\sigma^2, \tau^2 \sim$ Scale-Inv-$\chi(a, b)$.
Run the Gibbs Sampler (T Iterations).
for $t = 1 \to T$ **do**
    $\boldsymbol{\theta} \mid \sigma^2, \tau^2, \mathbf{y} \sim \mathcal{N}(\mathbf{m}^*, \mathbf{V}^*)$ with $\mathbf{m}^* = \tau^{-2}\mathbf{V}^*\mathbf{K}^\intercal\mathbf{y}$ and $\mathbf{V}^* = \tau^2\sigma^2(\tau^2\mathbf{K}^{-1} + \sigma^2\mathbf{I}_q)^{-1}$;
    $\sigma^2 \mid \boldsymbol{\theta}, \tau^2, \mathbf{y} \sim$ Scale-inv-$\chi^2(a^*_\sigma, b^*_\sigma)$ where $a^*_\sigma = a + q$ and $b^*_\sigma = a^{*-1}_\sigma(ab + \boldsymbol{\theta}^\intercal\mathbf{K}^{-1}\boldsymbol{\theta})$;
    $\tau^2 \mid \boldsymbol{\theta}, \sigma^2, \mathbf{y} \sim$ Scale-inv-$\chi^2(a^*_\tau, b^*_\tau)$ where $a^*_\tau = a + n$ and $b^*_\tau = a^{*-1}_\tau(ab + \boldsymbol{\epsilon}^\intercal\boldsymbol{\epsilon})$ where $\boldsymbol{\epsilon} = \mathbf{y} - \mathbf{K}\boldsymbol{\theta}$;
    $\widetilde{\boldsymbol{\beta}} = \mathbf{X}^\dagger\mathbf{K}\boldsymbol{\theta}$.
end for
Calculate the empirical mean, covariance, and precision of the posterior distribution $p(\widetilde{\boldsymbol{\beta}} \,|\, \mathbf{y})$ as $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, and $\boldsymbol{\Lambda}$, respectively.
Compute the centrality of every p predictor via Kullback-Leibler Divergence (KLD).
for $j = 1 \to p$ **do**
    $\text{KLD}(\widetilde{\beta}_j) = \frac{1}{2}\left[-\log(|\boldsymbol{\Sigma}_{-j}\boldsymbol{\Lambda}_{-j}|) + tr(\boldsymbol{\Sigma}_{-j}\boldsymbol{\Lambda}_{-j}) + 1 - p + \alpha_j(\widetilde{\beta}_j - \mu_j)^2\right]$.
end for
Scale each centrality measure for the p predictors to determine their relative importance.
for $j = 1 \to p$ **do**
    $\text{RATE}(\widetilde{\beta}_j) = \text{KLD}(\widetilde{\beta}_j) / \sum \text{KLD}(\widetilde{\beta}_\ell)$.
end for

Algorithm 3 Bayesian Neural Network (BNN)



Specify the architecture of the neural network (e.g. see above), operating over input/output pairs $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$. Denote the output of the network for a given x_i as $\widehat{y}_i$.
Specify a prior distribution $\pi(\cdot)$ over all parameters (e.g. weights and biases) in the network, summarized in a vector $\boldsymbol{\theta}$.
Use an MCMC sampler or any approximate Bayesian method to obtain a set of T samples $\{\widehat{\mathbf{y}}^{(t)}\}_{t=1}^T$ from the posterior predictive distribution $p(y_1^*, \ldots, y_n^* \,|\, \mathcal{D})$.
for $t = 1 \to T$ **do**
    $\widetilde{\boldsymbol{\beta}}^{(t)} = \mathbf{X}^\dagger \widehat{\mathbf{y}}^{(t)}$.
end for
Using the samples for $\widetilde{\boldsymbol{\beta}}$, calculate the empirical mean, covariance, and precision of the posterior distribution $p(\widetilde{\boldsymbol{\beta}} \,|\, \mathcal{D})$ as $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, and $\boldsymbol{\Lambda}$, respectively.
Compute the centrality of every p predictor via Kullback-Leibler Divergence (KLD).
for $j = 1 \to p$ **do**
    $\text{KLD}(\widetilde{\beta}_j) = \frac{1}{2}\left[-\log(|\boldsymbol{\Sigma}_{-j}\boldsymbol{\Lambda}_{-j}|) + tr(\boldsymbol{\Sigma}_{-j}\boldsymbol{\Lambda}_{-j}) + 1 - p + \alpha_j(\widetilde{\beta}_j - \mu_j)^2\right]$.
end for
Scale each centrality measure for the p predictors to determine their relative importance.
for $j = 1 \to p$ **do**
    $\text{RATE}(\widetilde{\beta}_j) = \text{KLD}(\widetilde{\beta}_j) / \sum \text{KLD}(\widetilde{\beta}_\ell)$.
end for

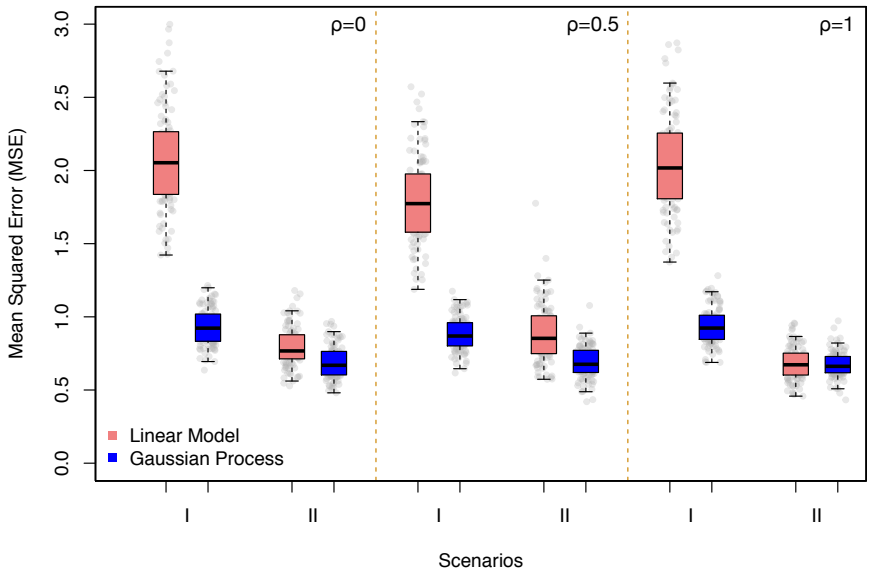
Supporting Information: Figures



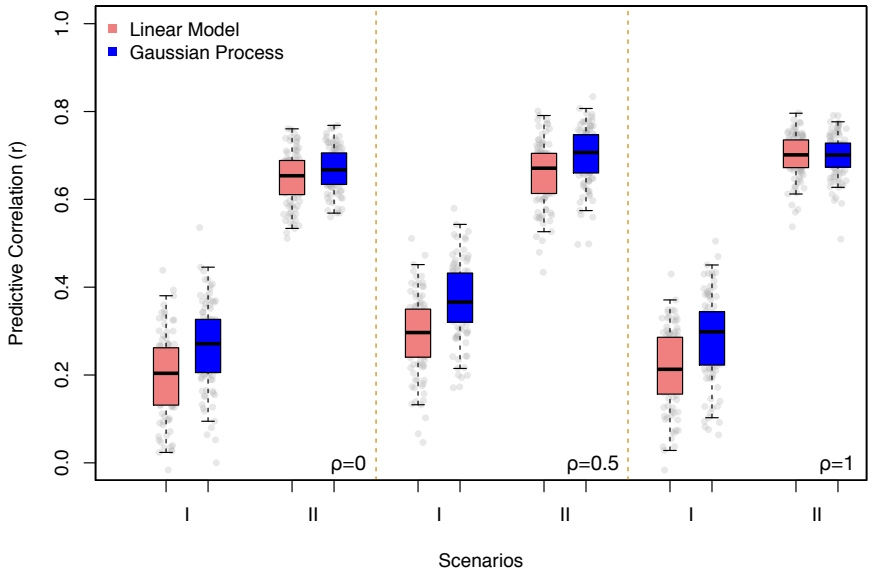
(a) GP Fitted vs. Observed

(b) OLS vs. Effect Size Analogue

Figure S1: Assessment of OLS estimates derived by a standard linear model and the posterior mean of the effect size analogues as computed by the Gaussian process (GP) regression method. Continuous outcomes were generated by a model taking on the form: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + (\mathbf{x}_1 \circ \mathbf{x}_2)\gamma + \boldsymbol{\varepsilon}$, with $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \tau^2\mathbf{I})$. We set $\tau^2 = 0.1$ and use $\mathbf{x}_1 \circ \mathbf{x}_2 = [\mathbf{x}_{11}\mathbf{x}_{21}, \ldots, \mathbf{x}_{1n}\mathbf{x}_{2n}]^\intercal$ to denote the element-wise interaction between the two vectors. We consider two scenarios. In Scenario I, we consider a purely additive model involving all covariates with $\gamma = 0$. Scenario II utilizes a sparse regression model, where we assume that only the first two predictors are causal and interact. Figure (a) assesses model fit while using the effect size analogue estimated by GP regression in both scenarios. Figure (b) shows the equivalence between the coefficients estimated via OLS and the effect size analogue in both scenarios.

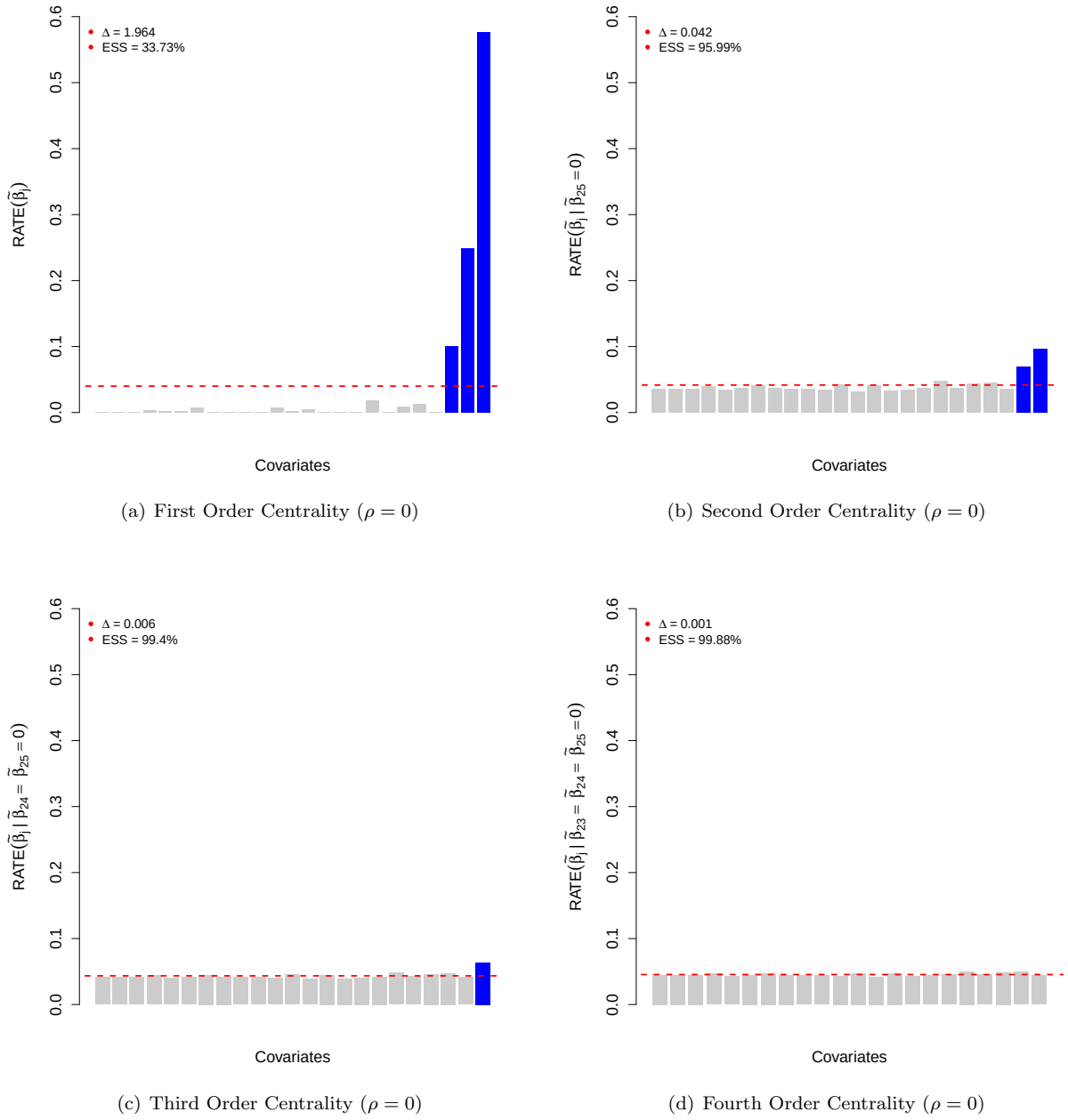


(a) Mean Squared Error (MSE)



(b) Predictive Correlation (r)

Figure S2: Comparisons of the out-of-sample predictive mean squared errors (MSE) and predictive correlations (r) for the linear regression model using the standard OLS estimates and the GP regression method using the effect size analogue. Scenarios I and II correspond to response variables being generated according to signal-to-noise ratios $V_{\mathbf{x}} = \{0.25, 0.75\}$ with control parameter $\rho = \{0, 0.5, 1\}$. Here, $(1 - \rho)$ is used to determine the proportion of signal that is contributed by interaction effects. Figure (a) corresponds to MSE results, while Figure (b) depicts results for predictive correlation. Results are based on 100 replicates in each case.



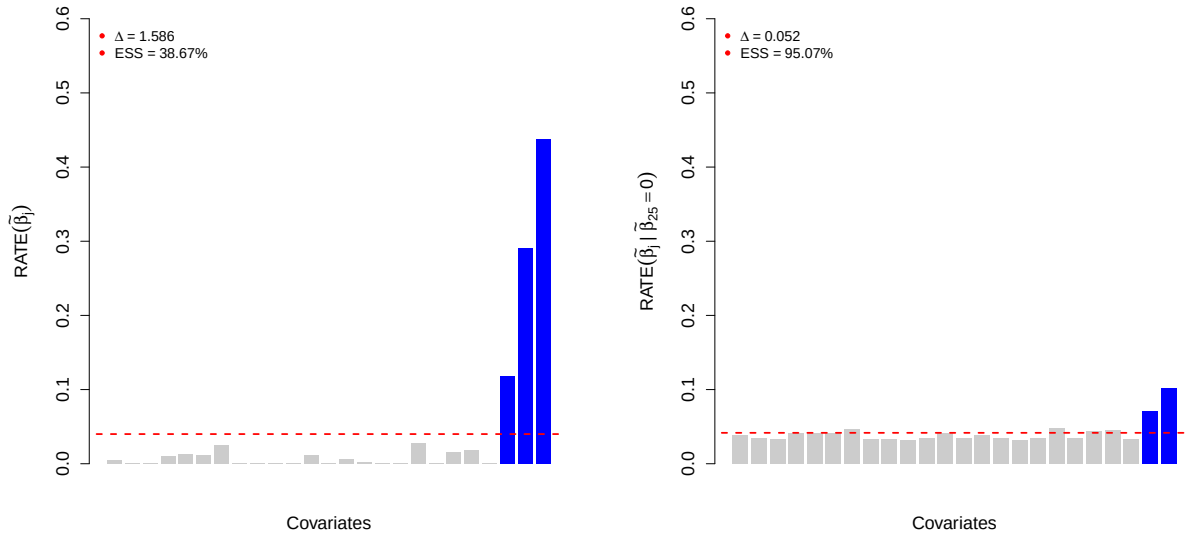
(a) First Order Centrality ($\rho = 0$)

(b) Second Order Centrality ($\rho = 0$)

(c) Third Order Centrality ($\rho = 0$)

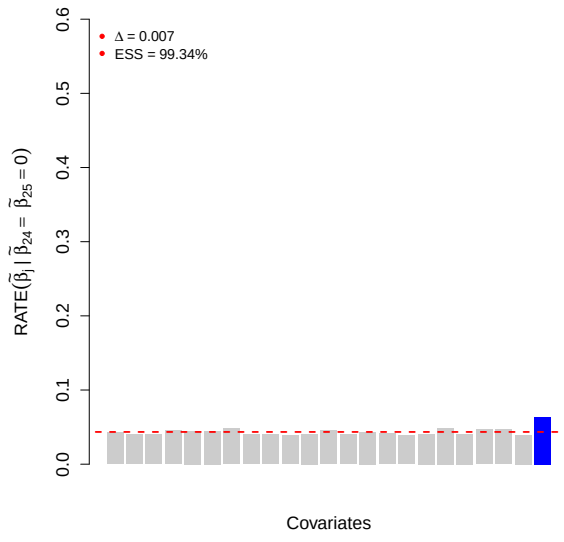
(d) Fourth Order Centrality ($\rho = 0$)

Figure S3: Demonstrating different orders of distributional centrality via RATE measures. Data are simulated such that only the effects of the last three covariates $p^* = \{23, 24, 25\}$ (blue) are nonzero with $\beta_{25} > \beta_{24} > \beta_{23}$. Outcomes are generated using a signal-to-noise ratio $V_{\mathbf{x}} = 0.75$ with $\rho = 0$. Here, $(1 - \rho)$ is used to determine the proportion of signal that is contributed by interaction effects. The x-axis of each figure shows the index of the different predictors, while the y-axis gives their relative centrality measures. The red dashed line is drawn at the level of relative equivalence (i.e. $1/p$). Figure (a) depicts the first order centrality across all predictors. Figures (b)-(d) illustrate scenarios where the most significantly associated covariates are iteratively nullified. These figures present results for the sets: (b) $\{\text{RATE}(\widetilde{\beta}_j \mid \widetilde{\beta}_{25} = 0)\}_{j=1}^{24}$; (c) $\{\text{RATE}(\widetilde{\beta}_j \mid \widetilde{\beta}_{25} = \widetilde{\beta}_{24} = 0)\}_{j=1}^{23}$; and (d) $\{\text{RATE}(\widetilde{\beta}_j \mid \widetilde{\beta}_{25} = \widetilde{\beta}_{24} = \widetilde{\beta}_{23} = 0)\}_{j=1}^{22}$, respectively. We also report values representing the degree to which the landscape of RATEs begin to look uniform: (i) the entropic difference Δ, and (ii) the corresponding empirical ESS estimate.

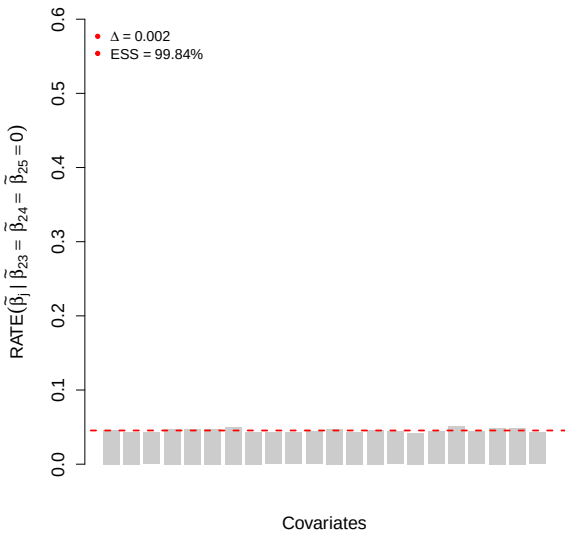


(a) First Order Centrality ($\rho = 0.5$)

(b) Second Order Centrality ($\rho = 0.5$)

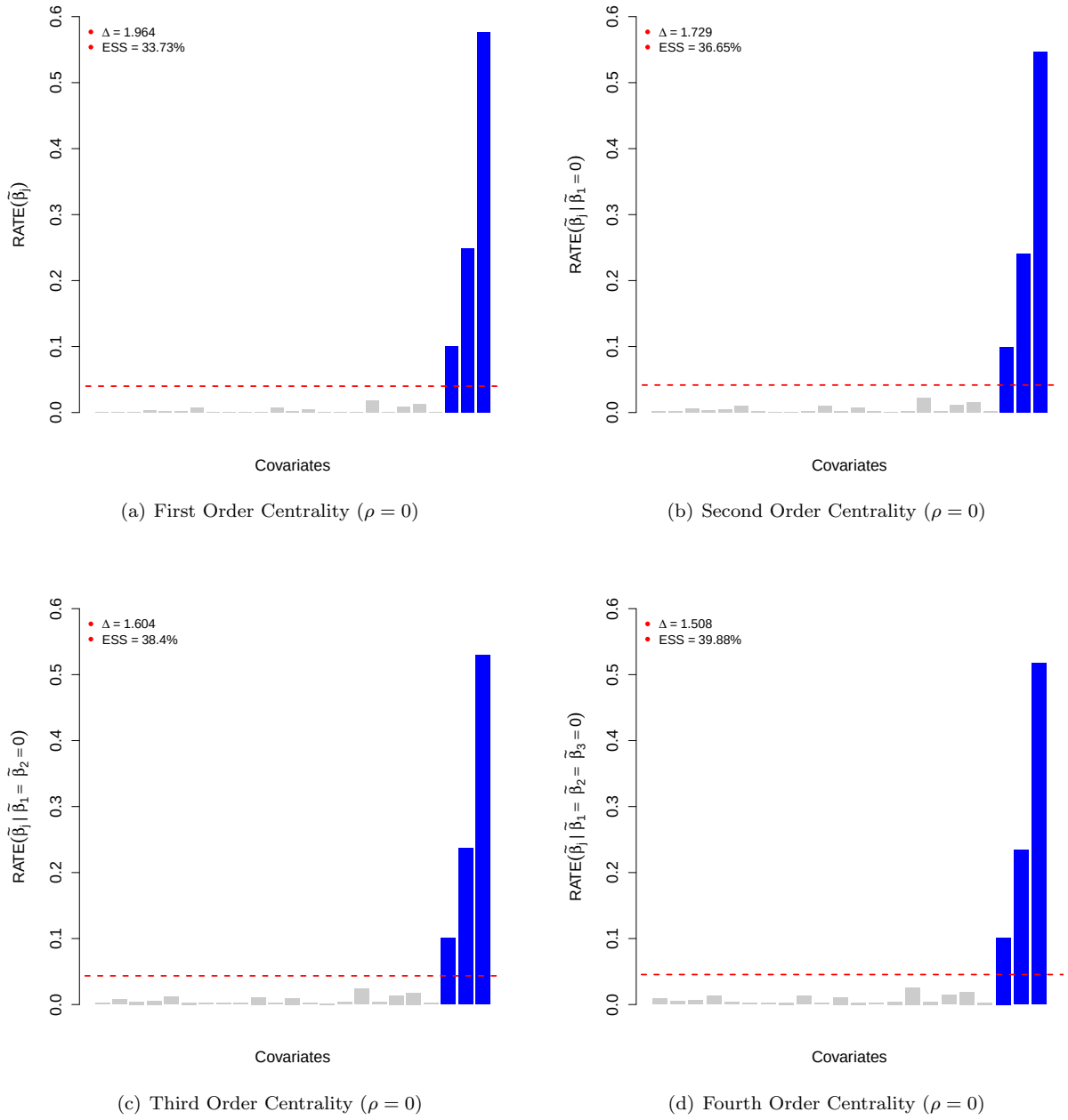


(c) Third Order Centrality ($\rho = 0.5$)



(d) Fourth Order Centrality ($\rho = 0.5$)

Figure S4: Demonstrating different orders of distributional centrality via RATE measures. Data are simulated such that only the effects of the last three covariates $p^* = \{23, 24, 25\}$ (blue) are nonzero with $\beta_{25} > \beta_{24} > \beta_{23}$. Outcomes are generated using a signal-to-noise ratio $V_{\mathbf{x}} = 0.75$ with $\rho = 0.5$. Here, $(1 - \rho)$ is used to determine the proportion of signal that is contributed by interaction effects. The x-axis of each figure shows the index of the different predictors, while the y-axis gives their relative centrality measures. The red dashed line is drawn at the level of relative equivalence (i.e. $1/p$). Figure (a) depicts the first order centrality across all predictors. Figures (b)-(d) illustrate scenarios where the most significantly associated covariates are iteratively nullified. These figures present results for the sets: (b) $\{\text{RATE}(\widetilde{\beta}_j \mid \widetilde{\beta}_{25} = 0)\}_{j=1}^{24}$; (c) $\{\text{RATE}(\widetilde{\beta}_j \mid \widetilde{\beta}_{25} = \widetilde{\beta}_{24} = 0)\}_{j=1}^{23}$; and (d) $\{\text{RATE}(\widetilde{\beta}_j \mid \widetilde{\beta}_{25} = \widetilde{\beta}_{24} = \widetilde{\beta}_{23} = 0)\}_{j=1}^{22}$, respectively. We also report values representing the degree to which the landscape of RATEs begin to look uniform: (i) the entropic difference Δ, and (ii) the corresponding empirical ESS estimate.



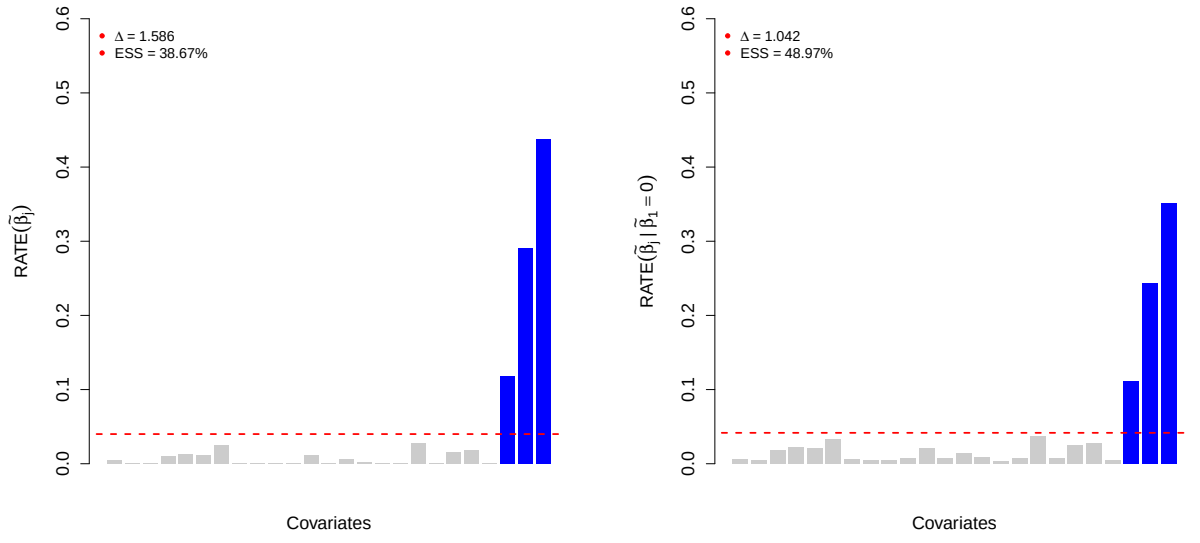
(a) First Order Centrality ($\rho = 0$)

(b) Second Order Centrality ($\rho = 0$)

(c) Third Order Centrality ($\rho = 0$)

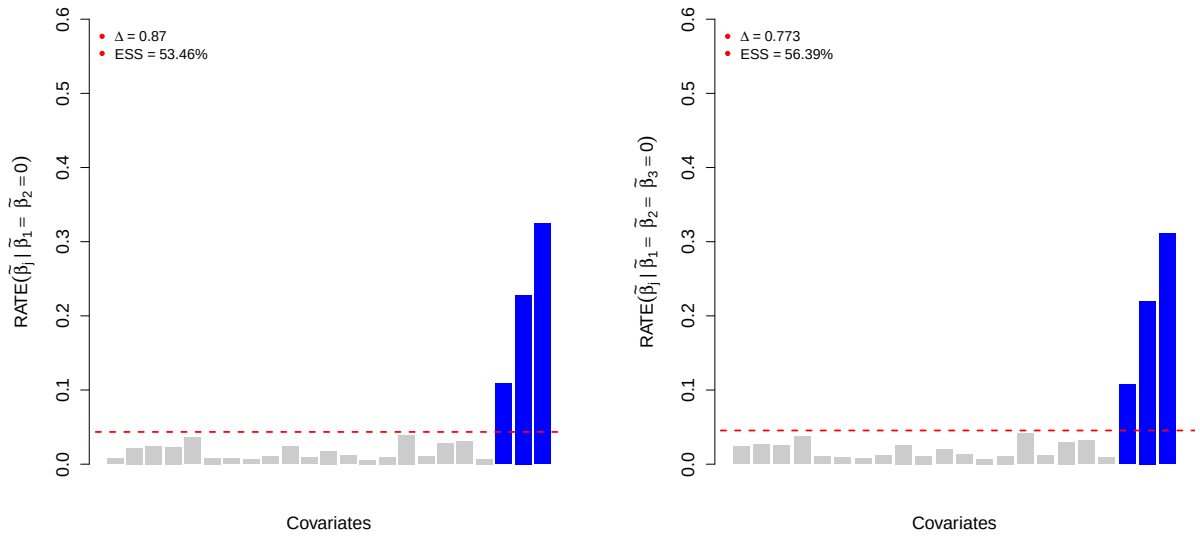
(d) Fourth Order Centrality ($\rho = 0$)

Figure S5: Distributional centrality and false positives with RATE measures. Data are simulated such that only the effects of the last three covariates $p^* = \{23, 24, 25\}$ (blue) are nonzero with $\beta_{25} > \beta_{24} > \beta_{23}$. Outcomes are generated using a signal-to-noise ratio $V_{\mathbf{x}} = 0.75$ with $\rho = 0$. Here, $(1-\rho)$ is used to determine the proportion of signal that is contributed by interaction effects. The x-axis of each figure shows the index of the different predictors, while the y-axis gives their relative centrality measures. The red dashed line is drawn at the level of relative equivalence (i.e. $1/p$). Figure (a) depicts the first order centrality across all predictors. Figures (b)-(d) illustrate scenarios where known nonsignificant predictors #1-3 are iteratively nullified. These figures present results for the sets: (b) $\{\text{RATE}(\widetilde{\beta}_j \mid \widetilde{\beta}_1 = 0)\}_{j \neq 1}$; (c) $\{\text{RATE}(\widetilde{\beta}_j \mid \widetilde{\beta}_1 = \widetilde{\beta}_2 = 0)\}_{j \neq (1,2)}$; and (d) $\{\text{RATE}(\widetilde{\beta}_j \mid \widetilde{\beta}_1 = \widetilde{\beta}_2 = \widetilde{\beta}_3 = 0)\}_{j \neq (1,2,3)}$, respectively. We also report values representing the degree to which the landscape of RATEs begin to look uniform: (i) the entropic difference Δ, and (ii) the corresponding empirical ESS estimate.



(a) First Order Centrality ($\rho = 0.5$)

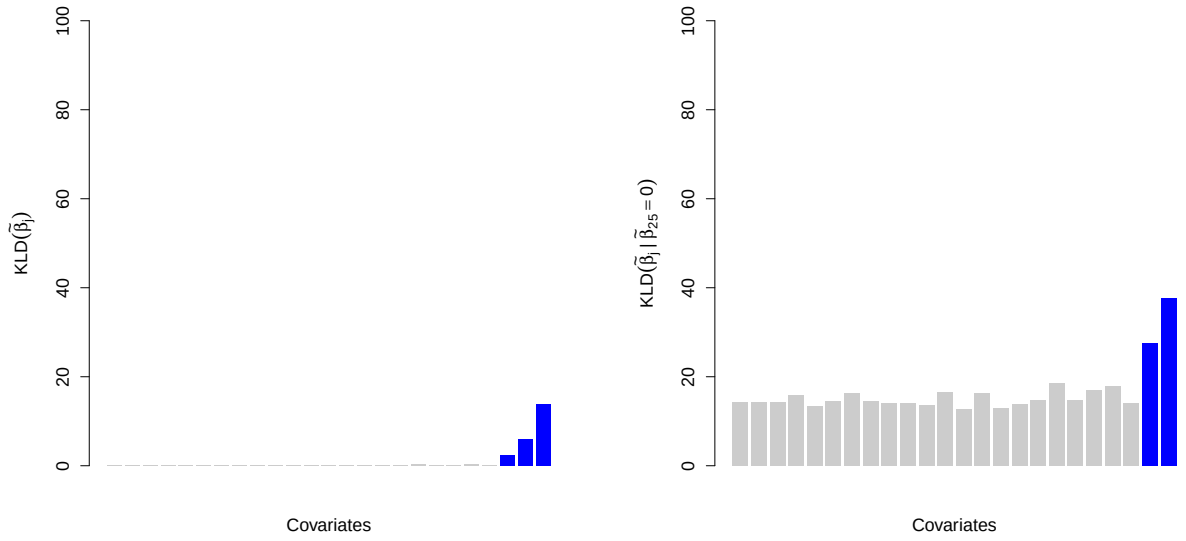
(b) Second Order Centrality ($\rho = 0.5$)



(c) Third Order Centrality ($\rho = 0.5$)

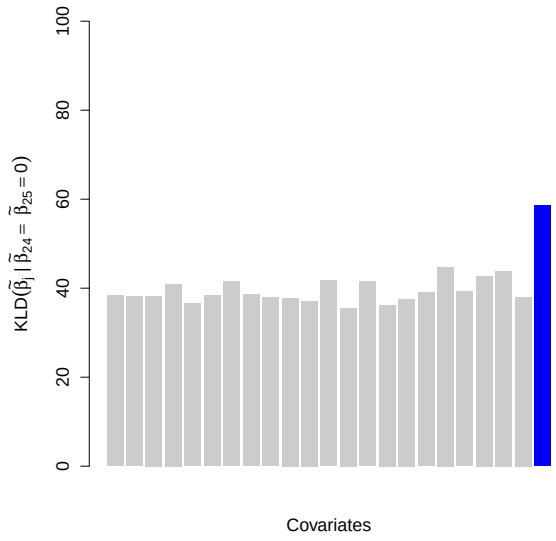
(d) Fourth Order Centrality ($\rho = 0.5$)

Figure S6: Distributional centrality and false positives with RATE measures. Data are simulated such that only the effects of the last three covariates $p^* = \{23, 24, 25\}$ (blue) are nonzero with $\beta_{25} > \beta_{24} > \beta_{23}$. Outcomes are generated using a signal-to-noise ratio $V_{\mathbf{x}} = 0.75$ with $\rho = 0.5$. Here, $(1-\rho)$ is used to determine the proportion of signal that is contributed by interaction effects. The x-axis of each figure shows the index of the different predictors, while the y-axis gives their relative centrality measures. The red dashed line is drawn at the level of relative equivalence (i.e. $1/p$). Figure (a) depicts the first order centrality across all predictors. Figures (b)-(d) illustrate scenarios where known nonsignificant predictors #1-3 are iteratively nullified. These figures present results for the sets: (b) $\{\text{RATE}(\widetilde{\beta}_j \mid \widetilde{\beta}_1 = 0)\}_{j \neq 1}$; (c) $\{\text{RATE}(\widetilde{\beta}_j \mid \widetilde{\beta}_1 = \widetilde{\beta}_2 = 0)\}_{j \neq (1,2)}$; and (d) $\{\text{RATE}(\widetilde{\beta}_j \mid \widetilde{\beta}_1 = \widetilde{\beta}_2 = \widetilde{\beta}_3 = 0)\}_{j \neq (1,2,3)}$, respectively. We also report values representing the degree to which the landscape of RATEs begin to look uniform: (i) the entropic difference Δ, and (ii) the corresponding empirical ESS estimate.

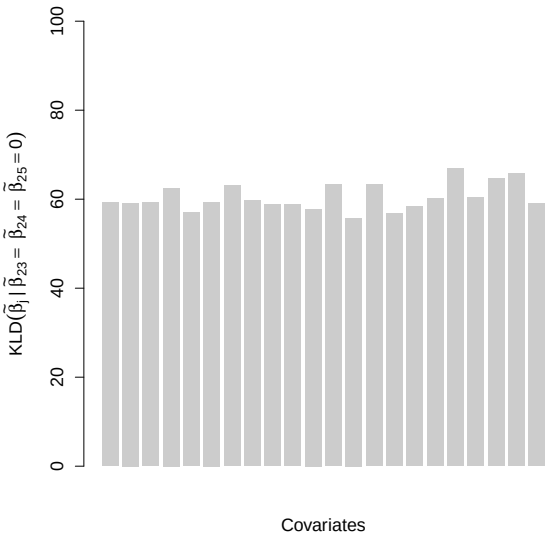


(a) First Order Centrality ($\rho = 0$)

(b) Second Order Centrality ($\rho = 0$)

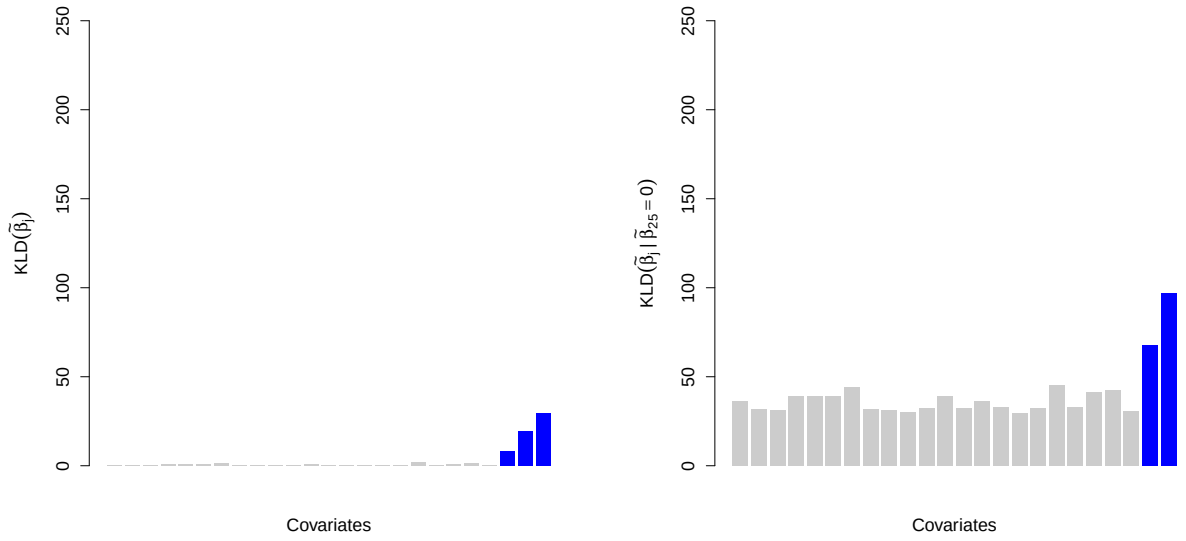


(c) Third Order Centrality ($\rho = 0$)



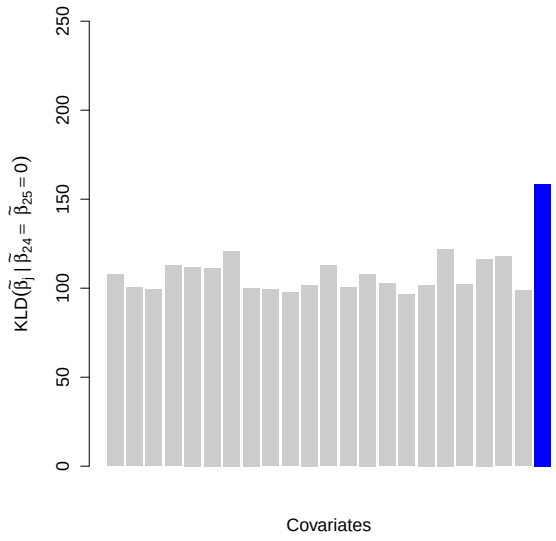
(d) Fourth Order Centrality ($\rho = 0$)

Figure S7: Demonstrating different orders of distributional centrality via the raw and unscaled Kullback-Leibler divergence (KLD) measures. Data are simulated such that only the effects of the last three covariates $p^* = \{23, 24, 25\}$ (blue) are nonzero with $\beta_{25} > \beta_{24} > \beta_{23}$. Outcomes are generated using a signal-to-noise ratio $V_{\mathbf{x}} = 0.75$ with $\rho = 0$. Here, $(1 - \rho)$ is used to determine the proportion of signal that is contributed by interaction effects. The x-axis of each figure shows the index of the different predictors, while the y-axis gives their overall centrality measures. Figure (a) depicts the first order centrality across all predictors. Figures (b)-(d) illustrate scenarios where the most significantly associated covariates are iteratively nullified. These figures present results for the sets: (b) $\{\text{KLD}(\widetilde{\beta}_j \mid \widetilde{\beta}_{25} = 0)\}_{j=1}^{24}$; (c) $\{\text{KLD}(\widetilde{\beta}_j \mid \widetilde{\beta}_{25} = \widetilde{\beta}_{24} = 0)\}_{j=1}^{23}$; and (d) $\{\text{KLD}(\widetilde{\beta}_j \mid \widetilde{\beta}_{25} = \widetilde{\beta}_{24} = \widetilde{\beta}_{23} = 0)\}_{j=1}^{22}$, respectively.

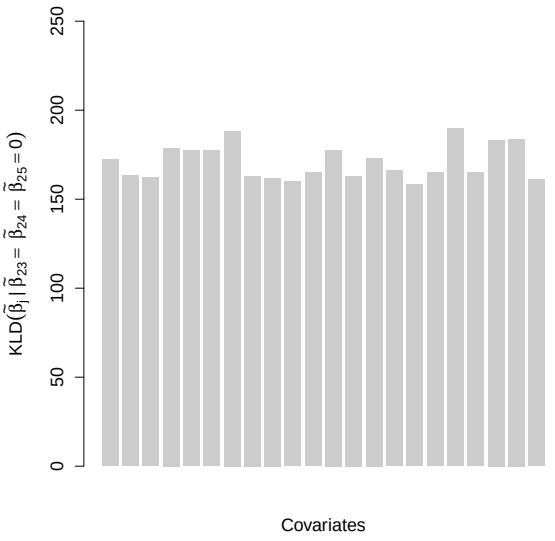


(a) First Order Centrality ($\rho = 0.5$)

(b) Second Order Centrality ($\rho = 0.5$)

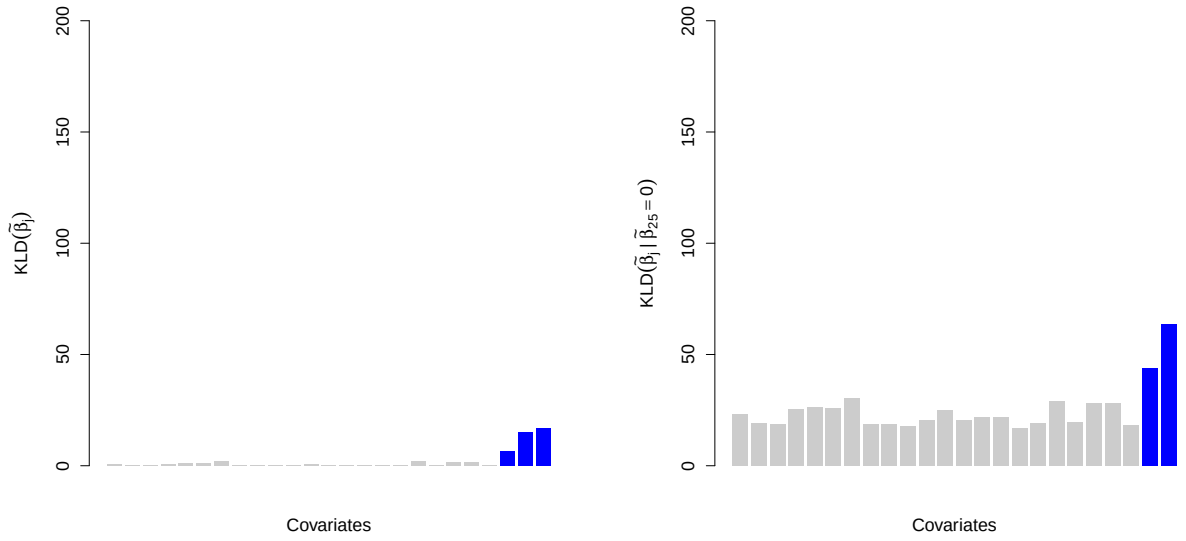


(c) Third Order Centrality ($\rho = 0.5$)



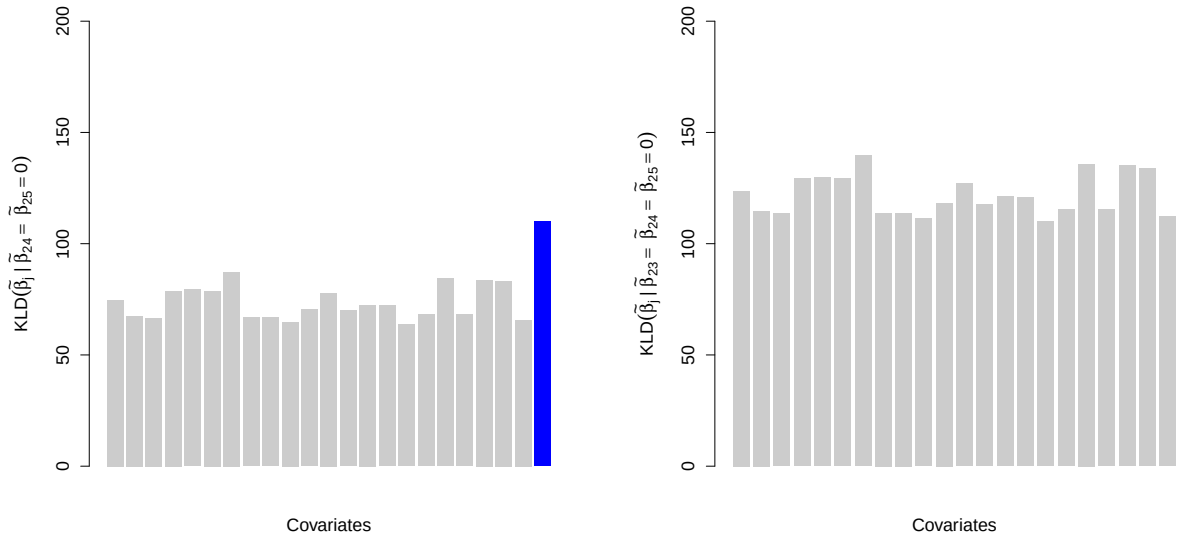
(d) Fourth Order Centrality ($\rho = 0.5$)

Figure S8: Demonstrating different orders of distributional centrality via the raw and unscaled Kullback-Leibler divergence (KLD) measures. Data are simulated such that only the effects of the last three covariates $p^* = \{23, 24, 25\}$ (blue) are nonzero with $\beta_{25} > \beta_{24} > \beta_{23}$. Outcomes are generated using a signal-to-noise ratio $V_{\mathbf{x}} = 0.75$ with $\rho = 0.5$. Here, $(1 - \rho)$ is used to determine the proportion of signal that is contributed by interaction effects. The x-axis of each figure shows the index of the different predictors, while the y-axis gives their overall centrality measures. Figure (a) depicts the first order centrality across all predictors. Figures (b)-(d) illustrate scenarios where the most significantly associated covariates are iteratively nullified. These figures present results for the sets: (b) $\{\text{KLD}(\widetilde{\beta}_j \mid \widetilde{\beta}_{25} = 0)\}_{j=1}^{24}$; (c) $\{\text{KLD}(\widetilde{\beta}_j \mid \widetilde{\beta}_{25} = \widetilde{\beta}_{24} = 0)\}_{j=1}^{23}$; and (d) $\{\text{KLD}(\widetilde{\beta}_j \mid \widetilde{\beta}_{25} = \widetilde{\beta}_{24} = \widetilde{\beta}_{23} = 0)\}_{j=1}^{22}$, respectively.



(a) First Order Centrality ($\rho = 1$)

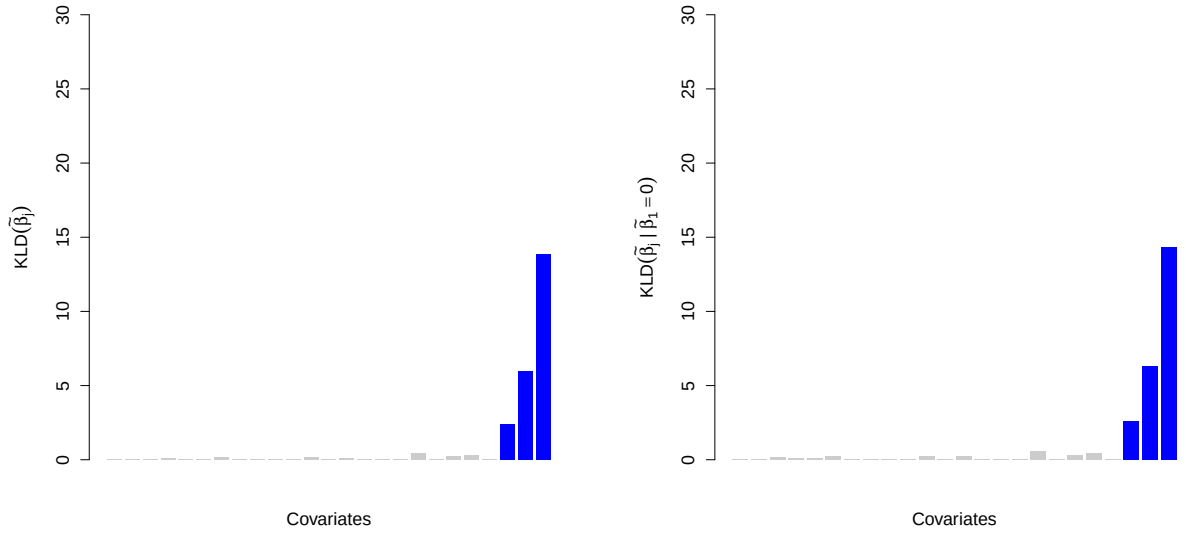
(b) Second Order Centrality ($\rho = 1$)



(c) Third Order Centrality ($\rho = 1$)

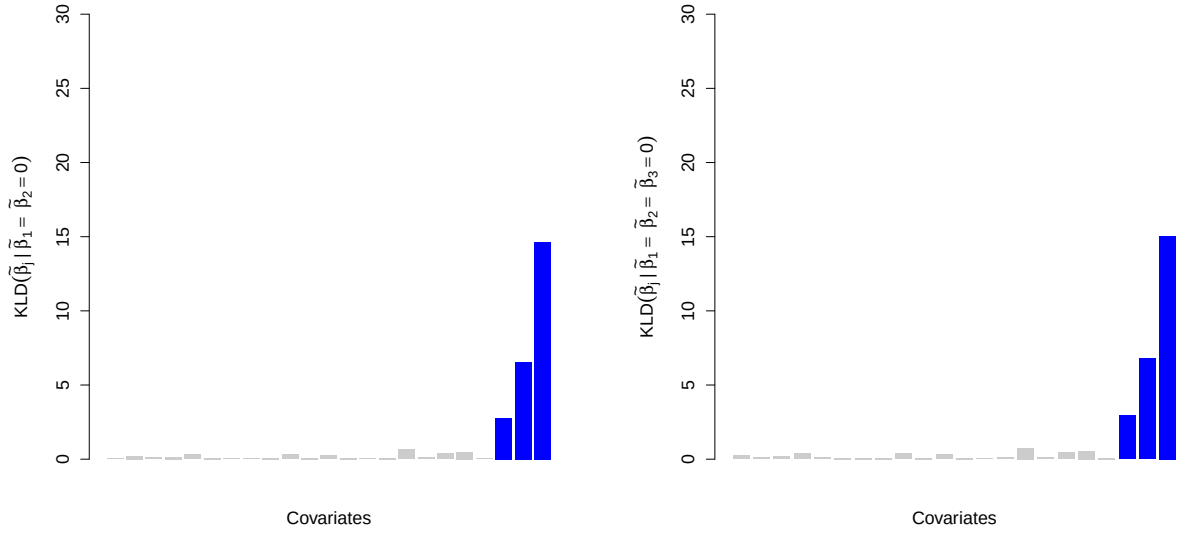
(d) Fourth Order Centrality ($\rho = 1$)

Figure S9: Demonstrating different orders of distributional centrality via the raw and unscaled Kullback-Leibler divergence (KLD) measures. Data are simulated such that only the effects of the last three covariates $p^* = \{23, 24, 25\}$ (blue) are nonzero with $\beta_{25} > \beta_{24} > \beta_{23}$. Outcomes are generated using a signal-to-noise ratio $V_{\mathbf{x}} = 0.75$ with $\rho = 1$. Here, $(1 - \rho)$ is used to determine the proportion of signal that is contributed by interaction effects. The x-axis of each figure shows the index of the different predictors, while the y-axis gives their overall centrality measures. Figure (a) depicts the first order centrality across all predictors. Figures (b)-(d) illustrate scenarios where the most significantly associated covariates are iteratively nullified. These figures present results for the sets: (b) $\{\text{KLD}(\widetilde{\beta}_j \mid \widetilde{\beta}_{25} = 0)\}_{j=1}^{24}$; (c) $\{\text{KLD}(\widetilde{\beta}_j \mid \widetilde{\beta}_{25} = \widetilde{\beta}_{24} = 0)\}_{j=1}^{23}$; and (d) $\{\text{KLD}(\widetilde{\beta}_j \mid \widetilde{\beta}_{25} = \widetilde{\beta}_{24} = \widetilde{\beta}_{23} = 0)\}_{j=1}^{22}$, respectively.



(a) First Order Centrality ($\rho = 0$)

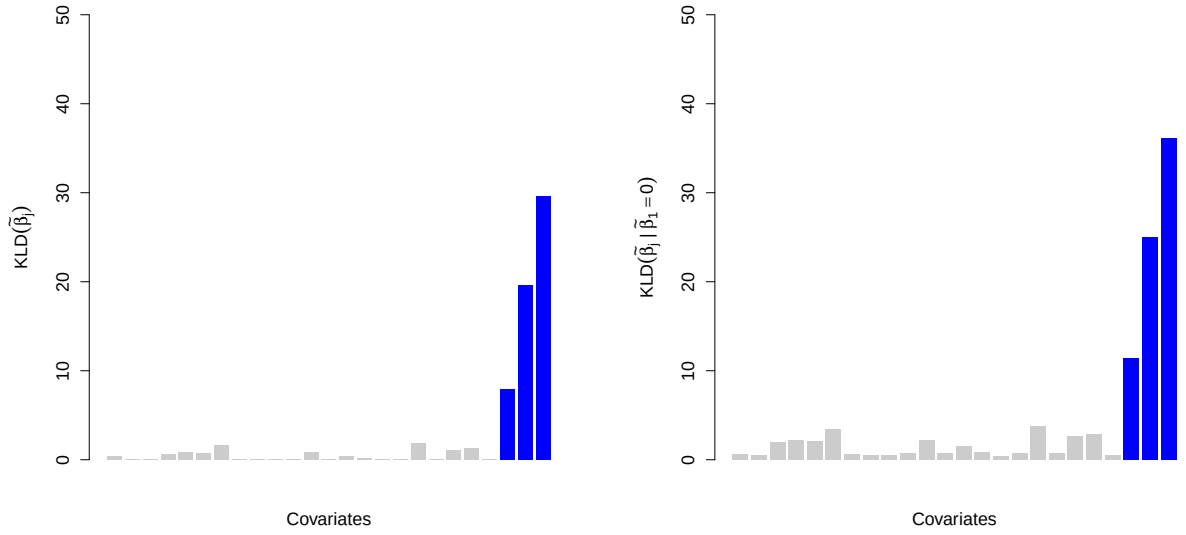
(b) Second Order Centrality ($\rho = 0$)



(c) Third Order Centrality ($\rho = 0$)

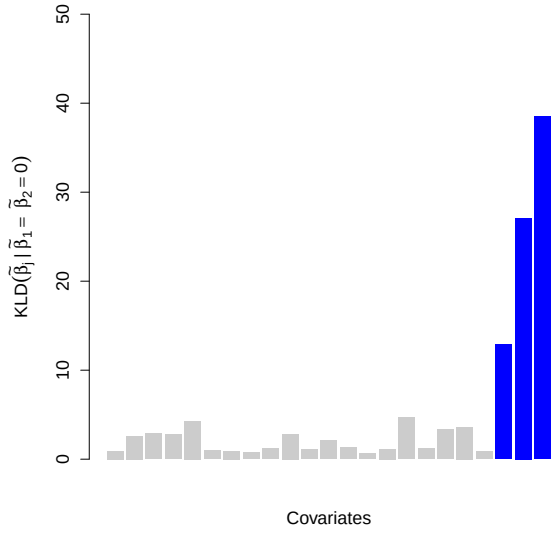
(d) Fourth Order Centrality ($\rho = 0$)

Figure S10: Distributional centrality and false positives with Kullback-Leibler divergence (KLD) measures. Data are simulated such that only the effects of the last three covariates $p^* = \{23, 24, 25\}$ (blue) are nonzero with $\beta_{25} > \beta_{24} > \beta_{23}$. Outcomes are generated using a signal-to-noise ratio $V_{\mathbf{x}} = 0.75$ with $\rho = 0$. Here, $(1 - \rho)$ is used to determine the proportion of signal that is contributed by interaction effects. The x-axis of each figure shows the index of the different predictors, while the y-axis gives their overall centrality measures. Figure (a) depicts the first order centrality across all predictors. Figures (b)-(d) illustrate scenarios where known nonsignificant predictors #1-3 are iteratively nullified. These figures present results for the sets: (b) $\{\text{KLD}(\widetilde{\beta}_j \mid \widetilde{\beta}_1 = 0)\}_{j \neq 1}$; (c) $\{\text{KLD}(\widetilde{\beta}_j \mid \widetilde{\beta}_1 = \widetilde{\beta}_2 = 0)\}_{j \neq (1,2)}$; and (d) $\{\text{KLD}(\widetilde{\beta}_j \mid \widetilde{\beta}_1 = \widetilde{\beta}_2 = \widetilde{\beta}_3 = 0)\}_{j \neq (1,2,3)}$, respectively.

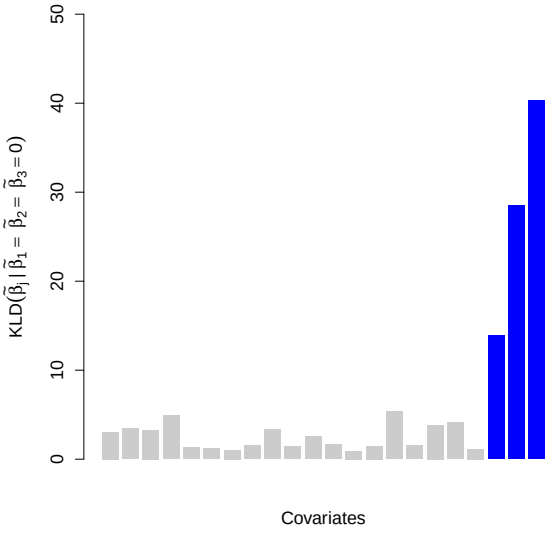


(a) First Order Centrality ($\rho = 0.5$)

(b) Second Order Centrality ($\rho = 0.5$)

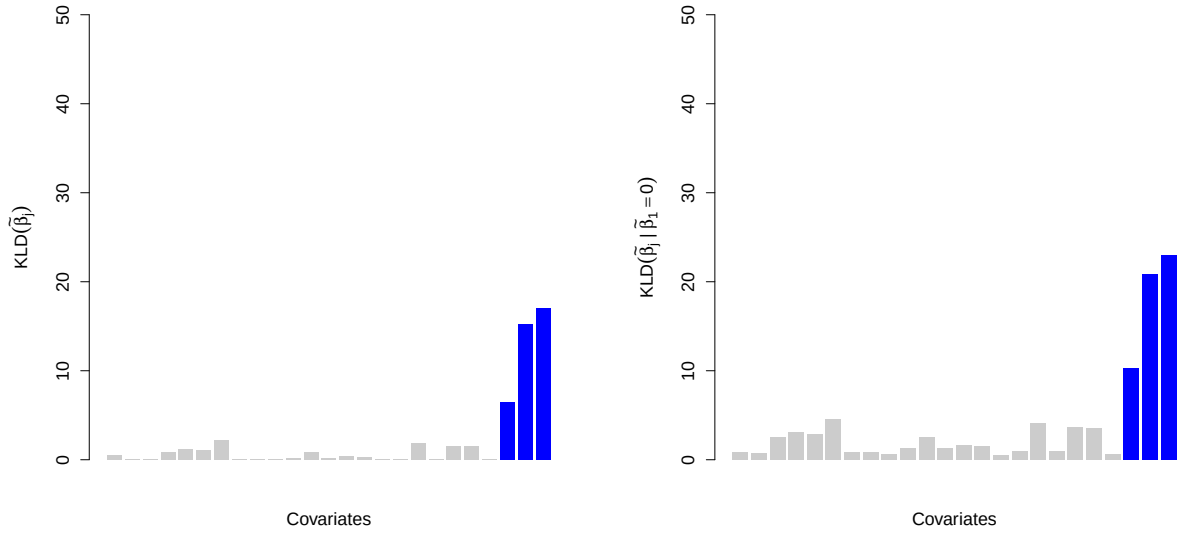


(c) Third Order Centrality ($\rho = 0.5$)



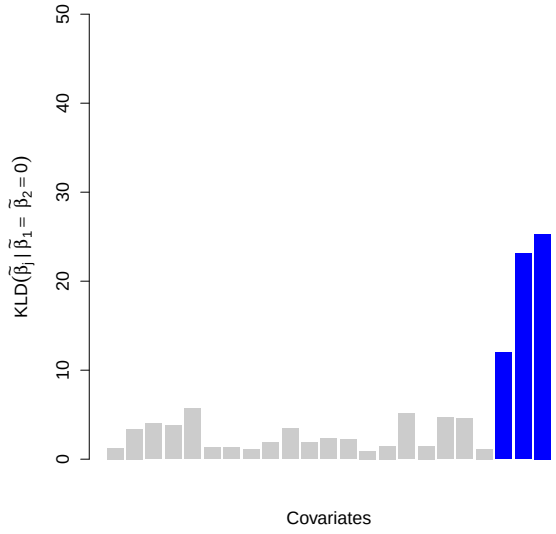
(d) Fourth Order Centrality ($\rho = 0.5$)

Figure S11: Distributional centrality and false positives with Kullback-Leibler divergence (KLD) measures. Data are simulated such that only the effects of the last three covariates $p^* = \{23, 24, 25\}$ (blue) are nonzero with $\beta_{25} > \beta_{24} > \beta_{23}$. Outcomes are generated using a signal-to-noise ratio $V_{\mathbf{x}} = 0.75$ with $\rho = 0.5$. Here, $(1 - \rho)$ is used to determine the proportion of signal that is contributed by interaction effects. The x-axis of each figure shows the index of the different predictors, while the y-axis gives their overall centrality measures. Figure (a) depicts the first order centrality across all predictors. Figures (b)-(d) illustrate scenarios where known nonsignificant predictors #1-3 are iteratively nullified. These figures present results for the sets: (b) $\{\text{KLD}(\widetilde{\beta}_j \,|\, \widetilde{\beta}_1 = 0)\}_{j \neq 1}$; (c) $\{\text{KLD}(\widetilde{\beta}_j \,|\, \widetilde{\beta}_1 = \widetilde{\beta}_2 = 0)\}_{j \neq (1,2)}$; and (d) $\{\text{KLD}(\widetilde{\beta}_j \,|\, \widetilde{\beta}_1 = \widetilde{\beta}_2 = \widetilde{\beta}_3 = 0)\}_{j \neq (1,2,3)}$, respectively.

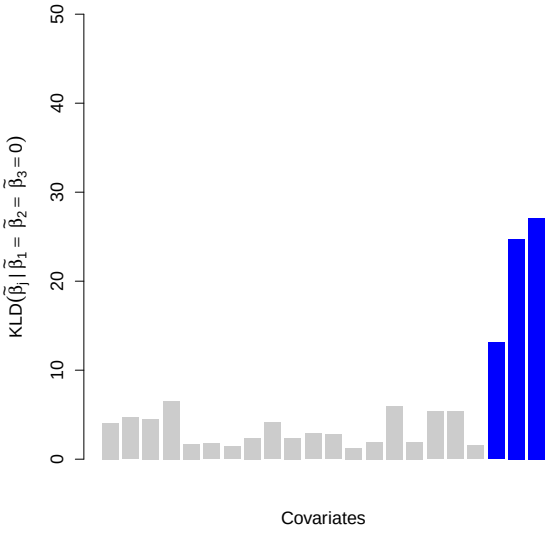


(a) First Order Centrality ($\rho = 1$)

(b) Second Order Centrality ($\rho = 1$)

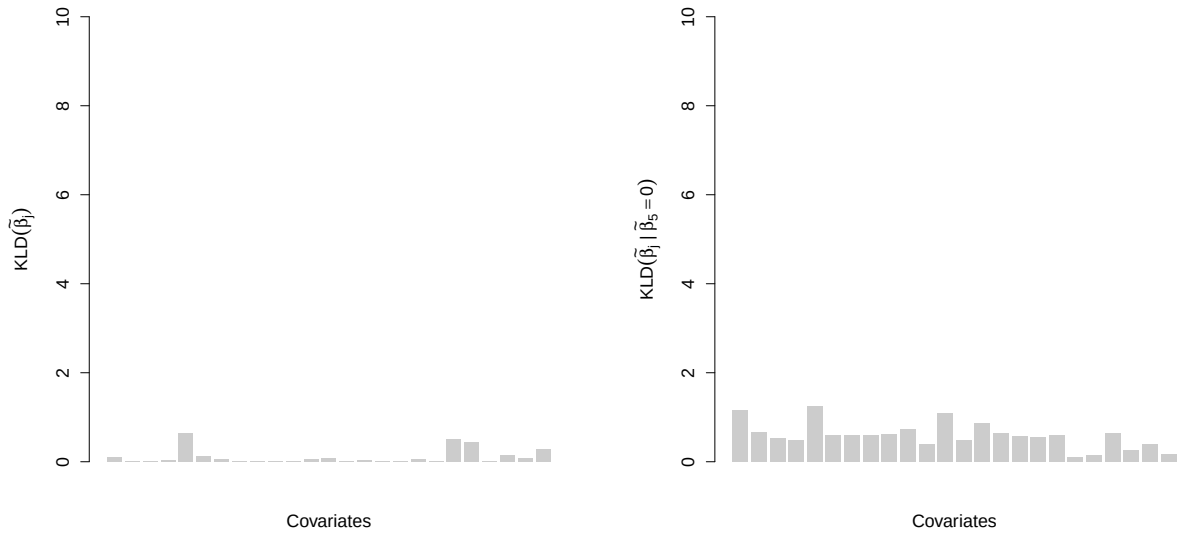


(c) Third Order Centrality ($\rho = 1$)



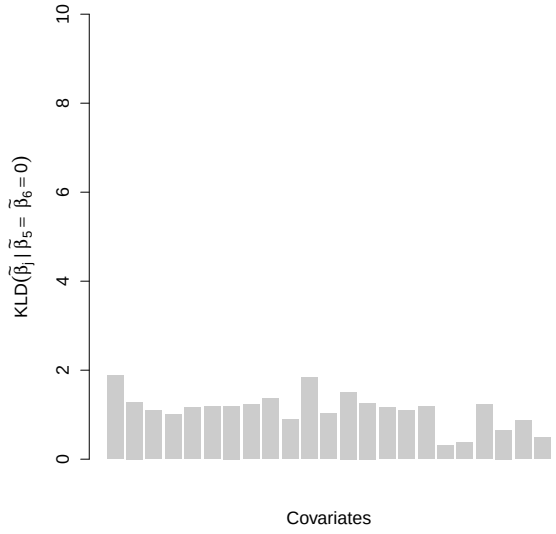
(d) Fourth Order Centrality ($\rho = 1$)

Figure S12: Distributional centrality and false positives with Kullback-Leibler divergence (KLD) measures. Data are simulated such that only the effects of the last three covariates $p^* = \{23, 24, 25\}$ (blue) are nonzero with $\beta_{25} > \beta_{24} > \beta_{23}$. Outcomes are generated using a signal-to-noise ratio $V_{\mathbf{x}} = 0.75$ with $\rho = 1$. Here, $(1 - \rho)$ is used to determine the proportion of signal that is contributed by interaction effects. The x-axis of each figure shows the index of the different predictors, while the y-axis gives their overall centrality measures. Figure (a) depicts the first order centrality across all predictors. Figures (b)-(d) illustrate scenarios where known nonsignificant predictors #1-3 are iteratively nullified. These figures present results for the sets: (b) $\{\text{KLD}(\widetilde{\beta}_j \,|\, \widetilde{\beta}_1 = 0)\}_{j \neq 1}$; (c) $\{\text{KLD}(\widetilde{\beta}_j \,|\, \widetilde{\beta}_1 = \widetilde{\beta}_2 = 0)\}_{j \neq (1,2)}$; and (d) $\{\text{KLD}(\widetilde{\beta}_j \,|\, \widetilde{\beta}_1 = \widetilde{\beta}_2 = \widetilde{\beta}_3 = 0)\}_{j \neq (1,2,3)}$, respectively.

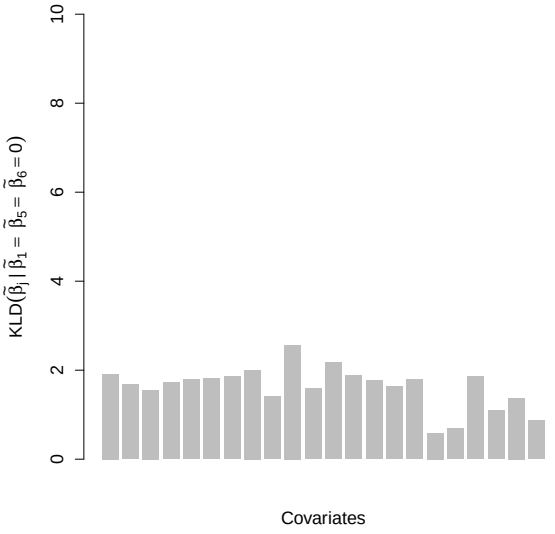


(a) First Order Centrality

(b) Second Order Centrality

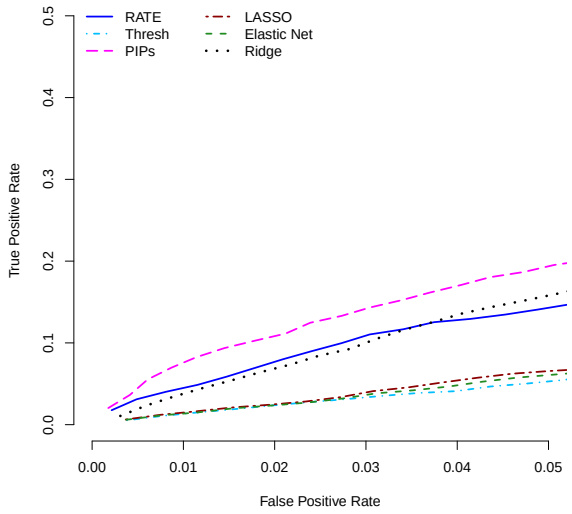


(c) Third Order Centrality

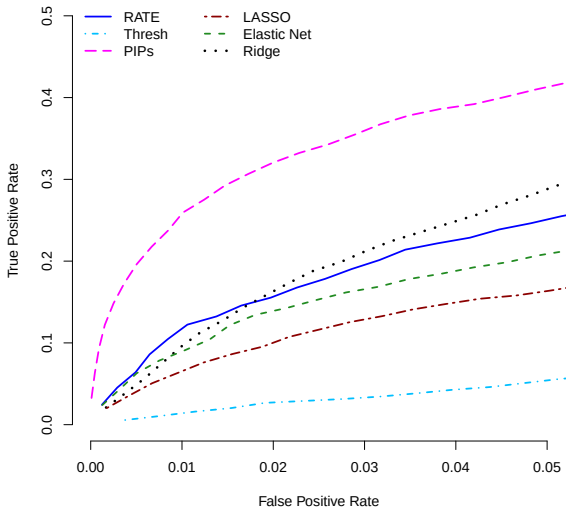


(d) Fourth Order Centrality

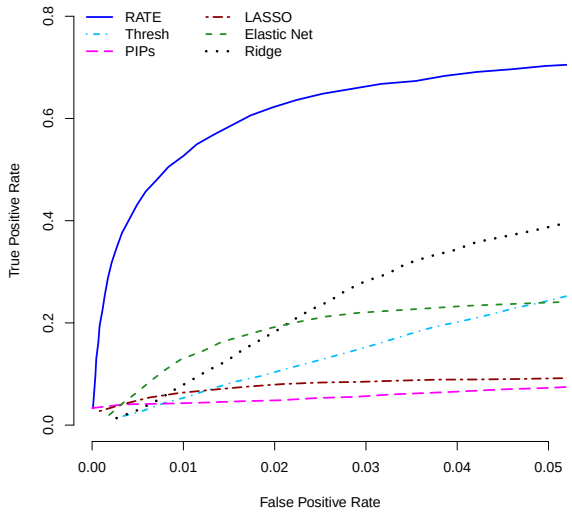
Figure S13: Demonstrating the behavior of Kullback-Leibler divergence (KLD) measures in the presence of noise. The outcome is made up exclusively of noise and data are simulated such that none of the input covariates are associated. The x-axis of each figure shows the index of the different predictors, while the y-axis gives their overall centrality measures. Figure (a) depicts the first order centrality across all predictors. Figures (b)-(d) illustrate scenarios where the most significantly associated covariates are iteratively nullified. These figures present results for the sets: (b) $\{\text{KLD}(\widetilde{\beta}_j \mid \widetilde{\beta}_5 = 0)\}_{j \neq 5}$; (c) $\{\text{KLD}(\widetilde{\beta}_j \mid \widetilde{\beta}_5 = \widetilde{\beta}_6 = 0)\}_{j \neq (5,6)}$; and (d) $\{\text{KLD}(\widetilde{\beta}_j \mid \widetilde{\beta}_1 = \widetilde{\beta}_5 = \widetilde{\beta}_6 = 0)\}_{j \neq (1,5,6)}$, respectively.



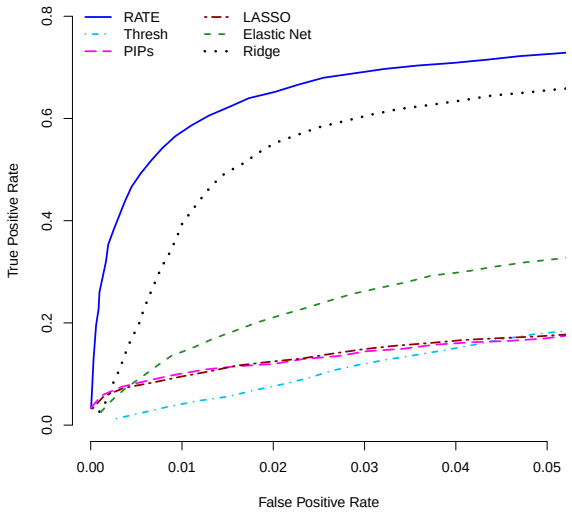
(a) Independent Predictors ($V_{\mathbf{x}} = 0.25$; $\rho = 0.5$)



(b) Independent Predictors ($V_{\mathbf{x}} = 0.75$; $\rho = 0.5$)

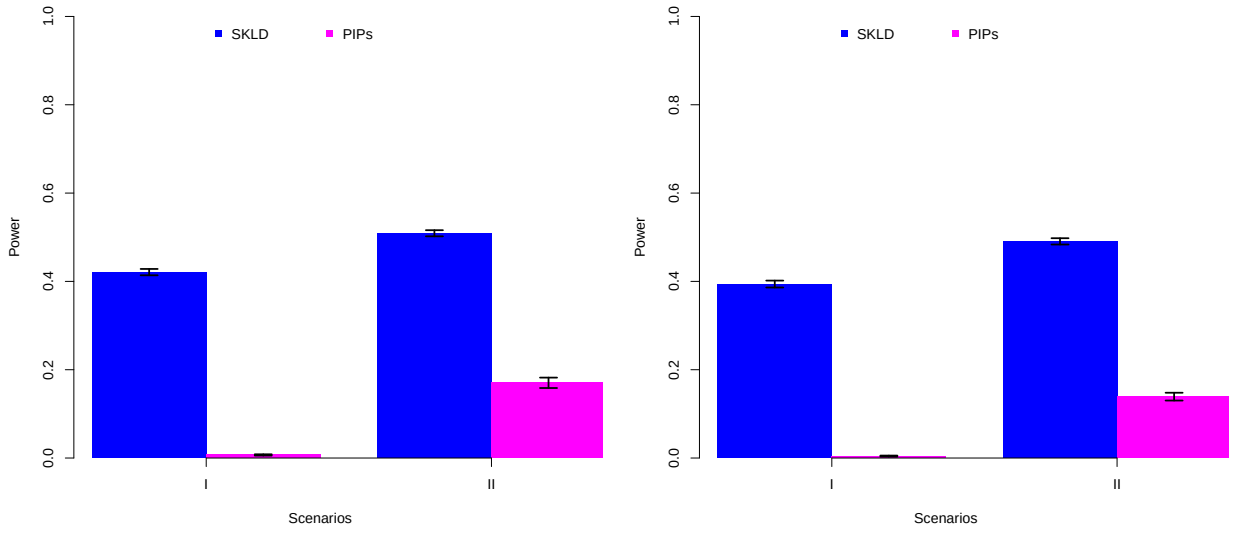


(c) Correlated Predictors ($V_{\mathbf{x}} = 0.25$; $\rho = 0.5$)



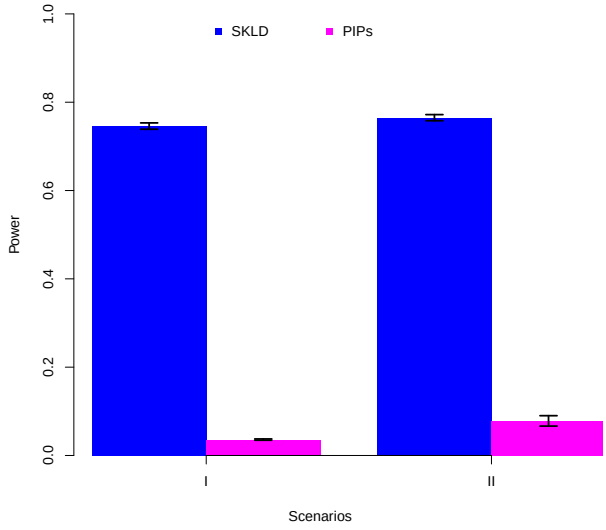
(d) Correlated Predictors ($V_{\mathbf{x}} = 0.75$; $\rho = 0.5$)

Figure S14: Power analysis for prioritizing causal covariates. We compare the association mapping ability of our nonparametric centrality measures (RATE) to the standard significance metrics provided by ridge regression, lasso regression, the elastic net, Bayesian variable selection method with a spike and slab prior (PIPs), and a Bayesian linear model with a normal-exponential-gamma prior and thresholded effect sizes (Thresh). Figures (a) and (b) display results when data is simulated using independent predictors; while Figures (c) and (d) correspond to results under correlated predictors. In either case, outcomes are generated using a signal-to-noise ratio $V_{\mathbf{x}} = \{0.25, 0.75\}$ with $\rho = 1$. Here, $(1 - \rho)$ is used to determine the proportion of signal that is contributed by interaction effects. The x-axis shows the false positive rate, while the y-axis gives the rate at which true causal variables were identified. Results are based on 100 replicates in each case.

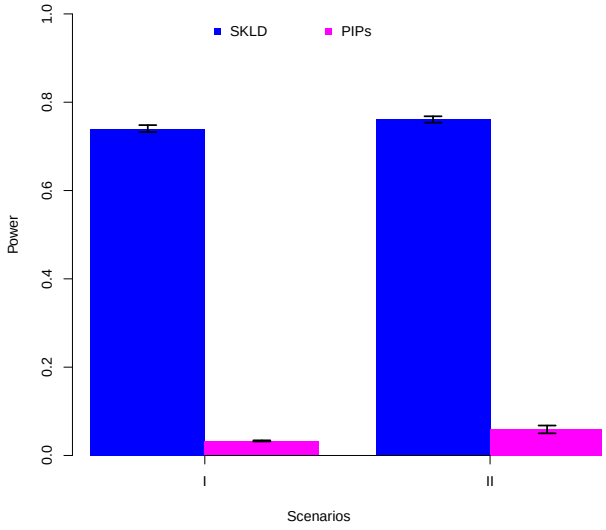


(a) Independent Predictors ($\rho = 0.5$)

(b) Independent Predictors ($\rho = 1$)

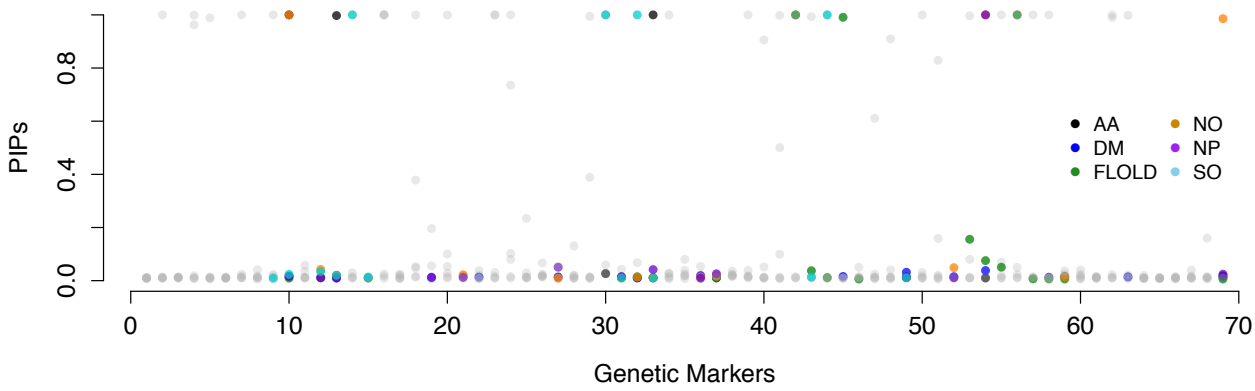


(c) Correlated Predictors ($\rho = 0.5$)

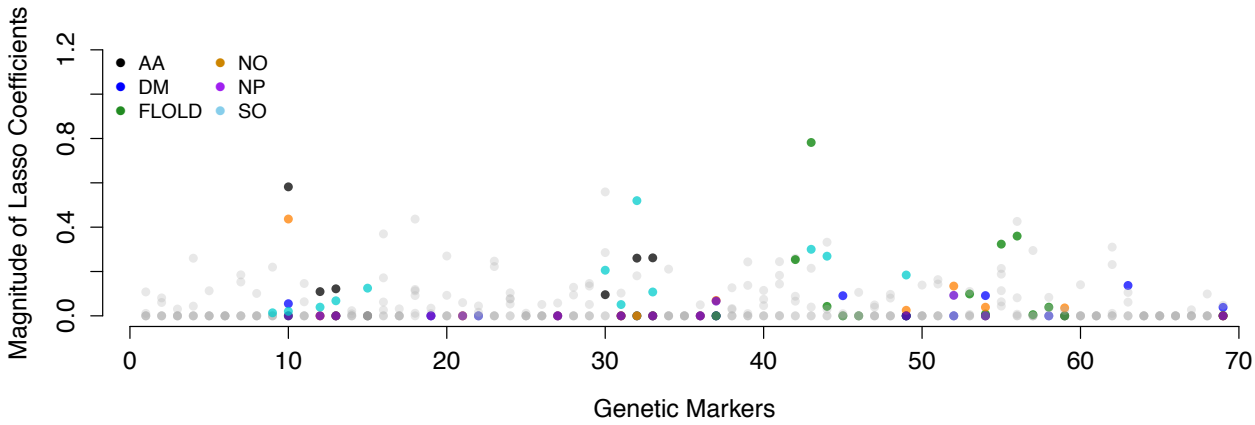


(d) Correlated Predictors ($\rho = 1$)

Figure S15: Power analysis for prioritizing causal covariates under the "optimal" model criterion. We assess the proportion of true positives with nonparametric centrality measures greater than relative equivalence (i.e. RATEs $> 1/p$). This power is compared to the proportion of true positives identified by the Bayesian "median probability model" as computed via the spike and slab prior method (i.e. PIPs > 0.5). Scenarios I and II correspond to response variables being generated according to signal-to-noise ratios $V_{\mathbf{x}} = \{0.25, 0.75\}$ with some parameter ρ. Here, $(1 - \rho)$ is used to determine the proportion of signal that is contributed by interaction effects. Figures (a) and (b) display results when data is simulated using independent predictors with $\rho = 0.5$ and 1, respectively; while Figures (c) and (d) correspond to results with correlated predictors under the same respective conditions. Results are based on 100 replicates in each case.



(a) Bayesian Variable Selection with Spike and Slab Prior



(b) Lasso Regression

Figure S16: Genetic map wide scan using competitive methods on all six traits analyzed in the *Arabidopsis thaliana* QTL mapping study. Here, we consider the standard lasso regression and Bayesian variable selection method with a spike and slab prior (PIPs) are also considered. The six traits analyzed include: amino acid (AA) content, shoot dry matter (DM), flowering time in long days (FLOLD), nitrate content (NO), nitrogen content percentage (NP), and sulfate content (SO). All microsatellite genetic markers are plotted in order of their positions along the genome. To ease the comparisons, points in color represent genetic markers with significant distributional centrality measures above the line of relative equivalence according to GP regression (i.e. RATEs $> 1/p$). Note that each color corresponds to a different phenotypic trait.

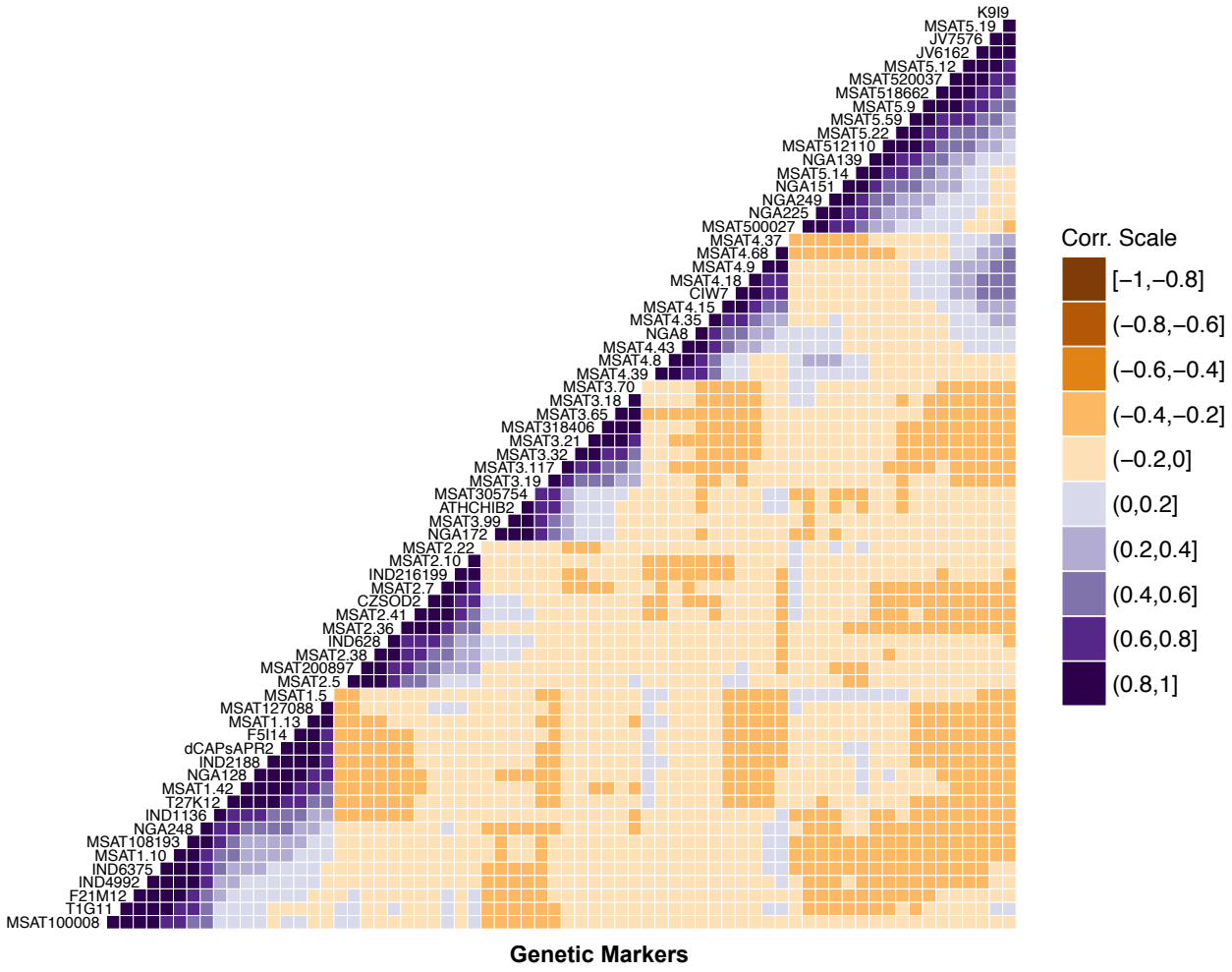


Figure S17: Lower-triangular heat map illustrating the correlation structure of the genotyped microsatellite markers in the *Arabidopsis thaliana* QTL mapping study. The legend represents a correlation scale on an [-1,1] interval that has been evenly divided into ten shorter subintervals. Note that there appears to be an underlying covarying structure between groups of markers located on different chromosomes.

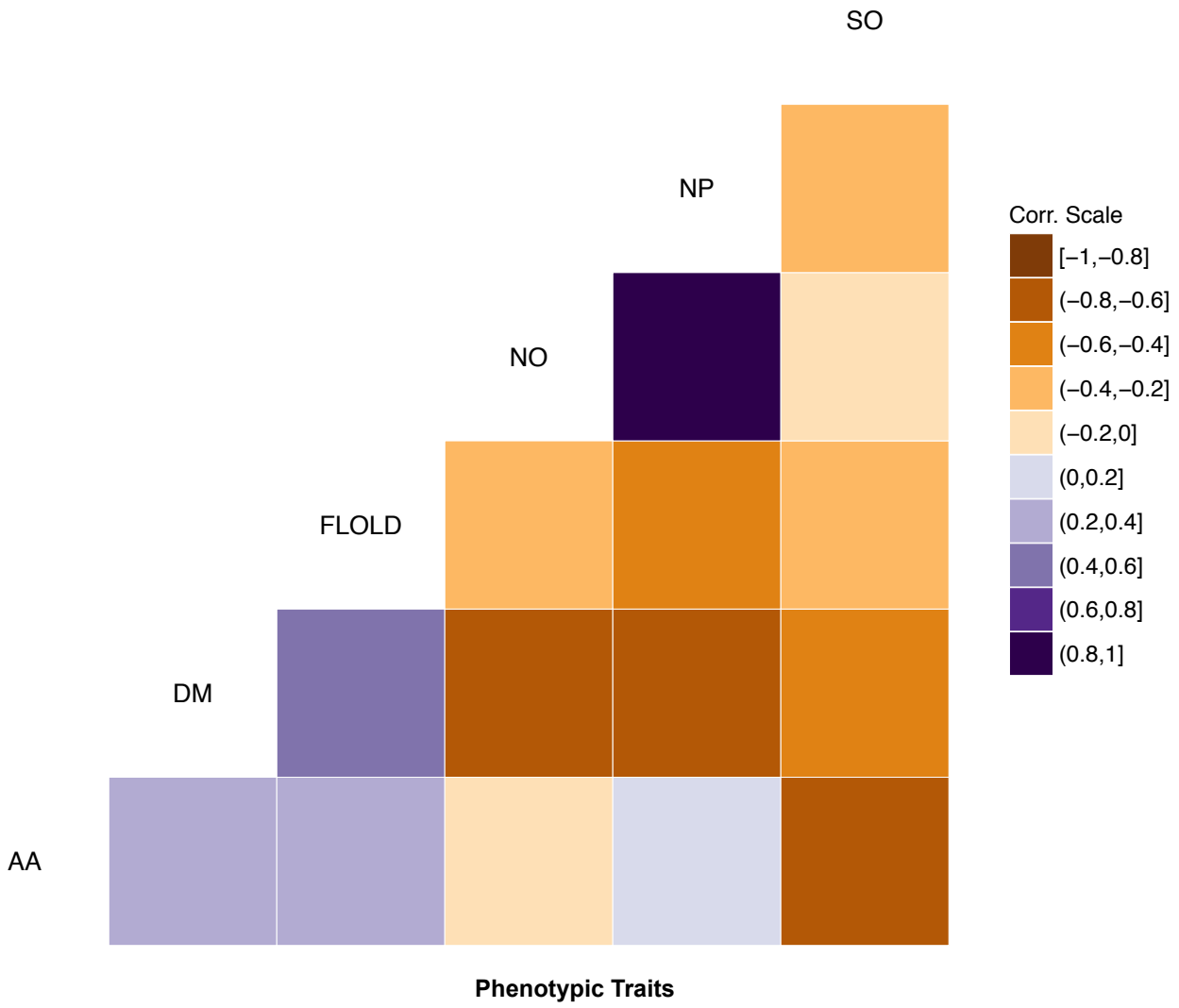


Figure S18: Lower-triangular heat map illustrating the correlation structure between the six phenotypes in the *Arabidopsis thaliana* QTL mapping study. The six traits analyzed include: amino acid (AA) content, shoot dry matter (DM), flowering time in long days (FLOLD), nitrate content (NO), nitrogen content percentage (NP), and sulfate content (SO). The legend represents a correlation scale on an [-1,1] interval that has been evenly divided into ten shorter subintervals.

Supporting Information: Tables

Table S1: A table that lists a description of the six quantitative phenotypes that are analyzed in the *Arabidopsis thaliana* QTL mapping study. The six traits analyzed include: amino acid (AA) content, shoot dry matter (DM), flowering time in long days (FLOLD), nitrate content (NO), nitrogen content percentage (NP), and sulfate content (SO). (XLSX)

Table S2: Table of all genetic markers and their distributional centrality measures for each of the six traits in the *Arabidopsis thaliana* QTL mapping study. Listed are the relative centrality (RATE) measures for each variant, along with their L1-regularized effect sizes as computed by lasso regression and the posterior inclusion probabilities (PIPs) derived from the Bayesian variable selection model. All microsatellite genetic markers are given in order of their positions along the genome. (XLSX)

	Average Time (sec)				
Computing Cores	$p = 50$	$p = 100$	$p = 500$	$p = 1000$	$p = 2500$
$n = 1$	0.13 (0.08)	0.68 (0.09)	247.77 (0.43)	3445.91 (7.42)	93723.20 (533.90)
$n = 4$	0.07 (0.01)	0.19 (0.05)	63.36 (1.42)	894.81 (27.35)	26405.56 (186.23)
$n = 8$	0.07 (0.01)	0.13 (0.05)	35.25 (0.45)	489.28 (9.46)	14805.76 (311.97)
$n = 16$	0.09 (0.01)	0.11 (0.05)	18.91 (0.33)	273.74 (2.47)	7608.53 (32.00)
$n = 32$	0.14 (0.01)	0.17 (0.04)	11.71 (0.29)	197.43 (2.38)	5097.67 (23.75)

Table S3: Computational complexity for calculating RATE as a function of the number of covariates that are present within the data and the number of available computing clusters for parallelization. Each entry represents the mean computation time (in seconds). Computations were performed using the Athena computing cluster at the Center for Statistical Sciences at Brown University. To create synthetic data for these simulations, we generated $p = \{50, 100, 500, 1000, 2000\}$ predictor variables respectively. Sample sizes were fixed at $n = 1000$. Values in the parentheses are the standard deviations of the estimates.

References

Bouteillé, M. (2011). *Control of shoot and root growth by water deficit in Arabidopsis thaliana: A parallel analysis using artificial and natural mapping populations*. Ph. D. thesis, Institut National d'Etudes Supérieures Agronomiques de Montpellier, FRA.

Caicedo, A. L., J. R. Stinchcombe, K. M. Olsen, J. Schmitt, and M. D. Purugganan (2004). Epistatic interaction between *Arabidopsis FRI* and *FLC* flowering time genes generates a latitudinal cline in a life history trait. *Proceedings of the National Academy of Sciences of the United States of America 101*(44), 15670–15675.

Johanson, U., J. West, C. Lister, S. Michaels, R. Amasino, and C. Dean (2000). Molecular analysis of *FRIGIDA*, a major determinant of natural variation in *Arabidopsis* flowering time. *Science 290*(5490), 344–347.

Koprivova, A., M. Giovannetti, P. Baraniecka, B.-R. Lee, C. Grondin, O. Loudet, and S. Kopriva (2013). Natural variation in the ATPS1 isoform of ATP sulfurylase contributes to the control of sulfate levels in Arabidopsis. *Plant Physiology 163*(3), 1133–1141.

Lee, I., S. D. Michaels, A. S. Masshardt, and R. M. Amasino (1994). The late-flowering phenotype of *FRIGIDA* and mutations in *LUMINIDEPENDENS* is suppressed in the landsberg *erecta* strain of *Arabidopsis*. *The Plant Journal 6*(6), 903–909.

Leustek, T. (2002). *Sulfate Metabolism*, Volume 1 of *e0017*. The American Society of Plant Biologists.

Loudet, O., S. Chaillou, C. Camilleri, D. Bouchez, and F. Daniel-Vedele (2002). Bay-0 × Shahdara recombinant inbred line population: A powerful tool for the genetic dissection of complex traits in Arabidopsis. *Theoretical and Applied Genetics 104*(6), 1173–1184.

Loudet, O., S. Chaillou, P. Merigout, J. Talbotec, and F. Daniel-Vedele (2003). Quantitative trait loci analysis of nitrogen use efficiency in Arabidopsis. *Plant Physiology 131*(1), 345–358.

Loudet, O., V. Saliba-Colombani, C. Camilleri, F. Calenge, V. Gaudon, A. Koprivova, K. A. North, S. Kopriva, and F. Daniel-Vedele (2007). Natural variation for sulfate content in *Arabidopsis thaliana* is highly controlled by APR2. *Nature Genetics 39*, 896 EP.

Reymond, M., S. Svistoonoff, O. Loudet, L. Nussaume, and T. Desnos (2006). Identification of QTL controlling root growth response to phosphate starvation in *Arabidopsis thaliana*. *Plant, Cell & Environment 29*(1), 115–125.

Sheldon, C. C., J. E. Burn, P. P. Perez, J. Metzger, J. A. Edwards, W. J. Peacock, and E. S. Dennis (1999). The *FLF* MADS box gene: A repressor of flowering in *Arabidopsis* regulated by vernalization and methylation. *The Plant Cell 11*(3), 445–458.

Shindo, C., M. J. Aranzana, C. Lister, C. Baxter, C. Nicholls, M. Nordborg, and C. Dean (2005). Role of *FRIGIDA* and *FLOWERING LOCUS C* in determining variation in flowering time of *Arabidopsis*. *Plant Physiology 138*(2), 1163–1173.