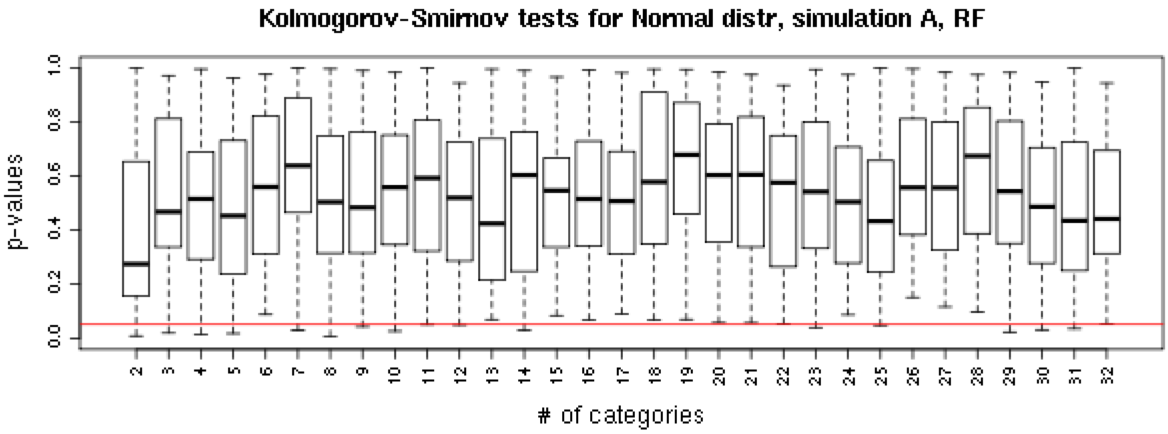
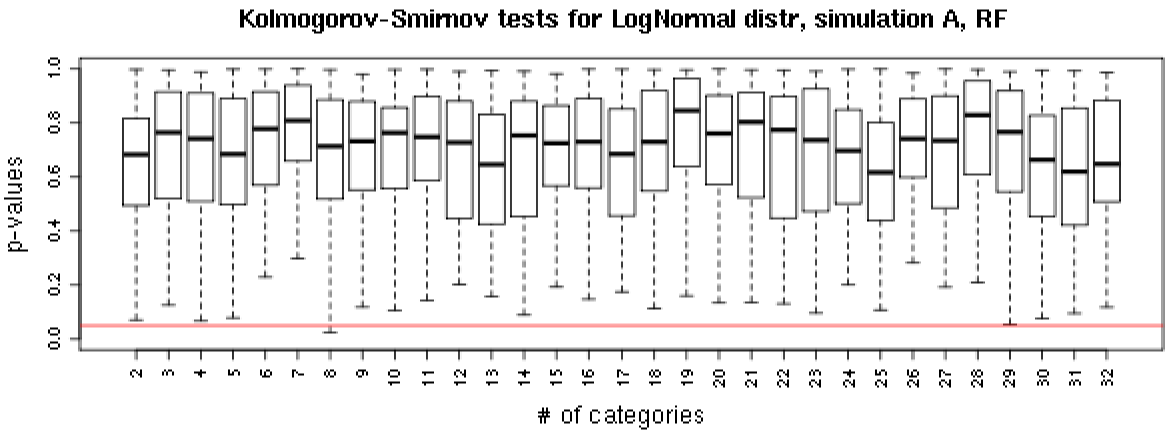
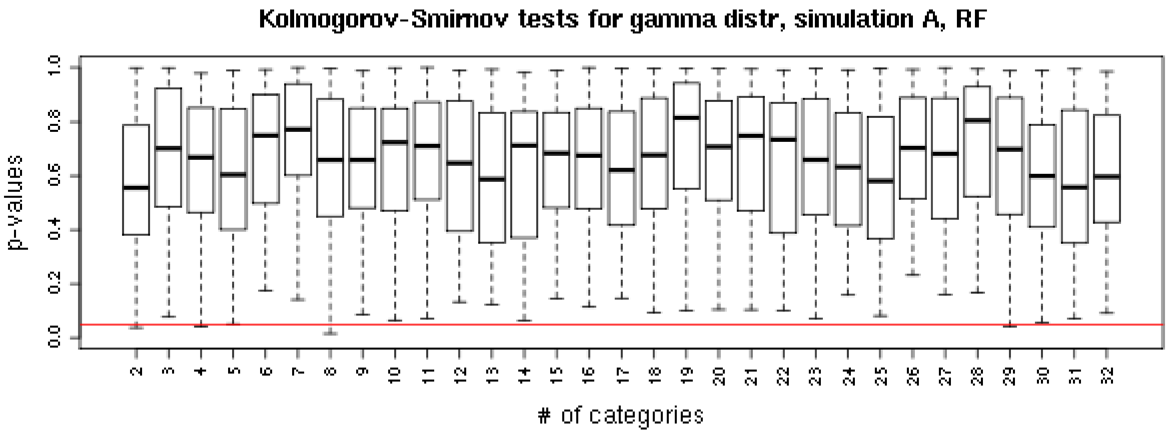
Supplementary material

Figure S1. Kolmogorov-Smirnov tests for the distribution of null importances in simulation A: **a)** RF; **b)** MI. High p-values indicate lack of evidence against a specific distribution. The red horizontal line shows the 0.05 significance level. In the case of RF importance, all distributions fit the null importances well. In the case of MI, gamma appears to approximate best the distribution of the null importances.

a)







b)

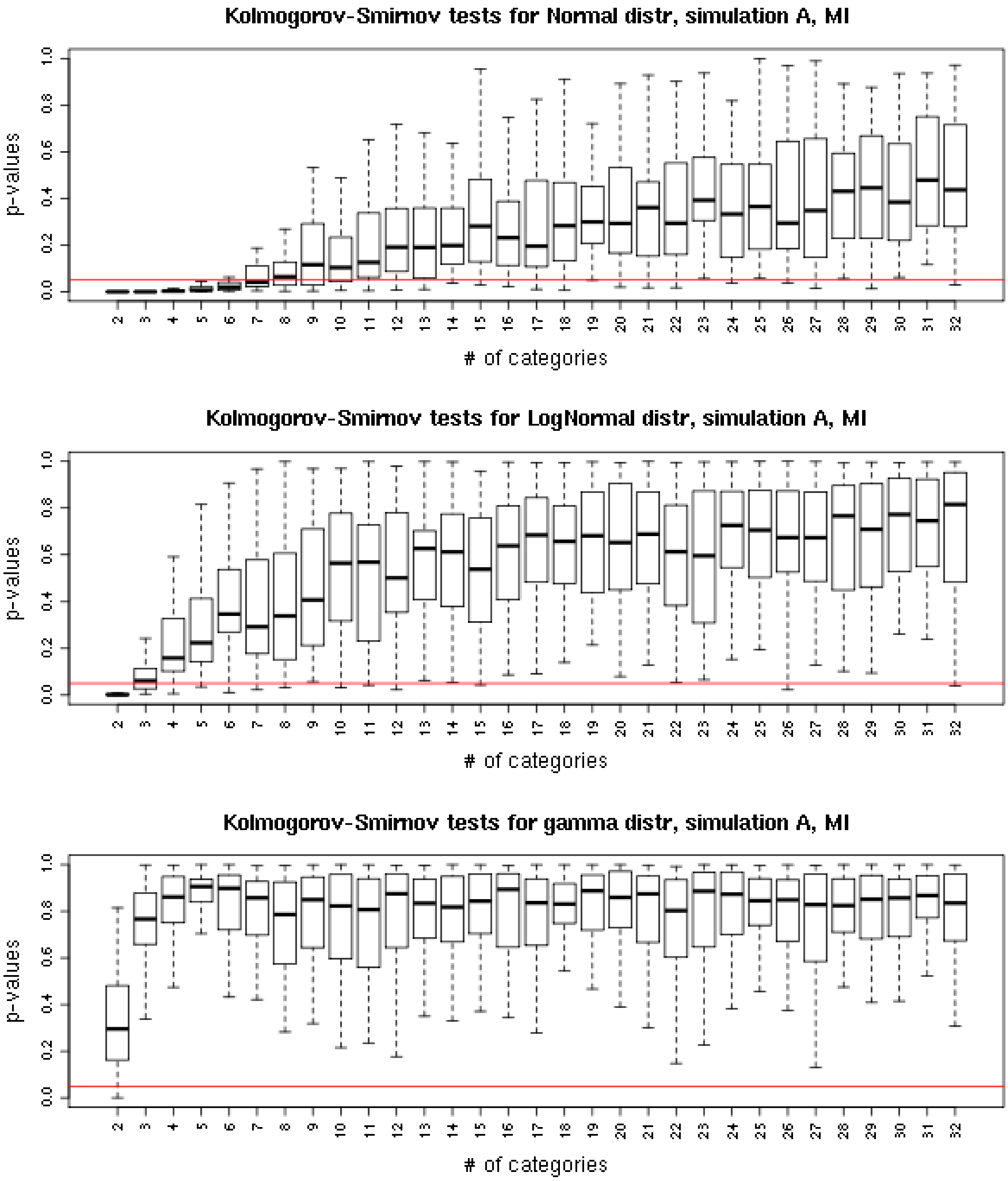
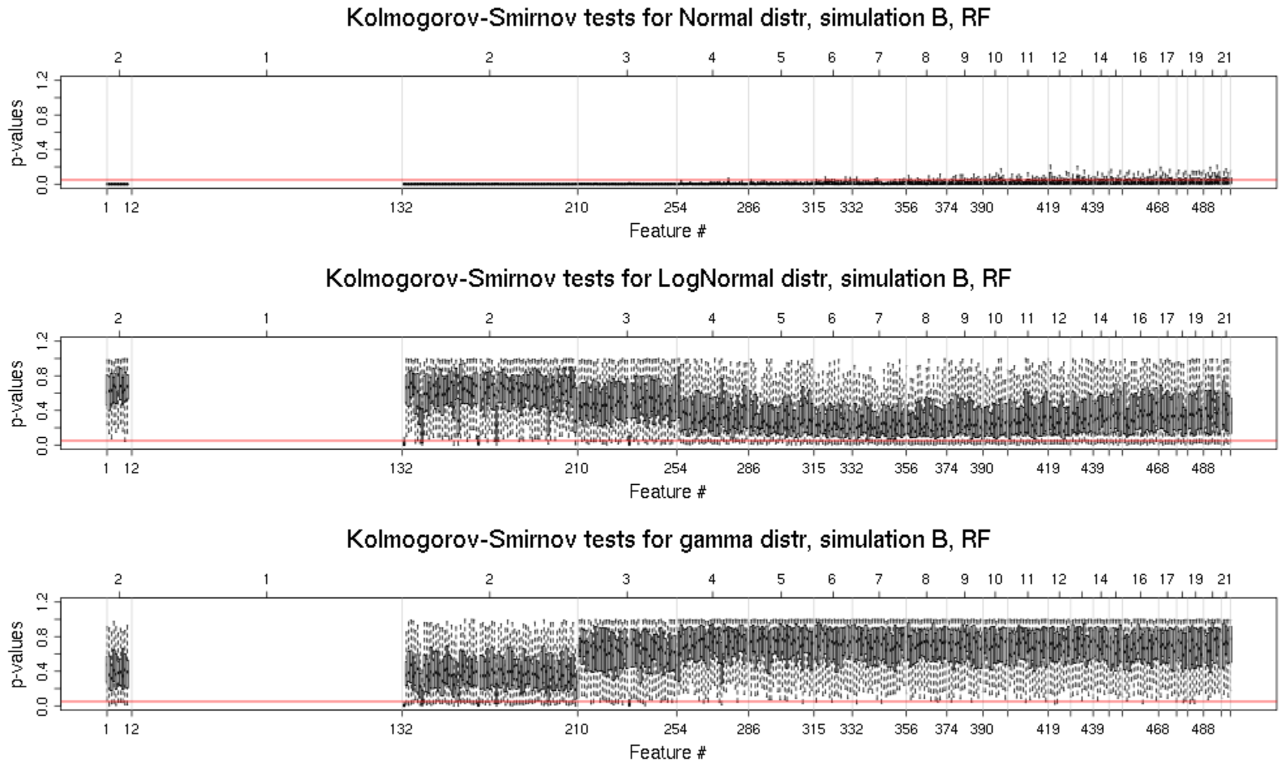

Kolmogorov-Smirnov tests for Normal distr, simulation A, MI
p-values
0.0 0.2 0.4 0.6 0.8 1.0
2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32
# of categories
Kolmogorov-Smirnov tests for LogNormal distr, simulation A, MI
p-values
0.0 0.2 0.4 0.6 0.8 1.0
2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32
# of categories
Kolmogorov-Smirnov tests for gamma distr, simulation A, MI
p-values
0.0 0.2 0.4 0.6 0.8 1.0
2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32
# of categories

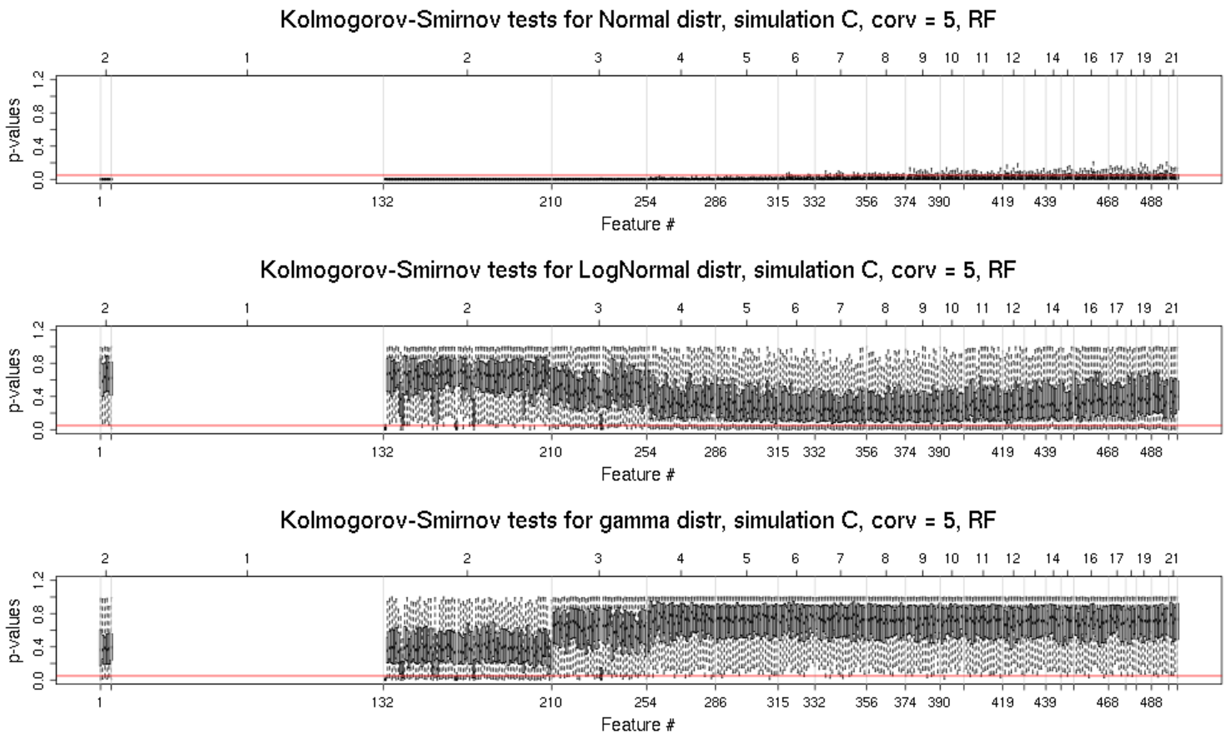
Figure S2. Boxplot for Kolmogorov-Smirnov tests for the distribution of null importances in simulation B. The red horizontal line depicts the 0.05 significance threshold. High p-values indicate lack of evidence against a particular distribution. The lognormal and gamma distributions appear to be good approximations for the null importances.



The features are presented in natural order. Features from 13 to 132 are constant (one category), their importance in the RandomForest models is constantly zero and therefore Kolmogorov-Smornov tests cannot be applied. The bottom horizontal axis shows the number of the feature and the top horizontal axis indicates the number of categories that the features have.

Figure S3. Boxplot for Kolmogorov-Smirnov tests for the distribution of null importances in simulation C: **a)** five correlated variables; **b)** 25 correlated variables. The red horizontal line depicts the 0.05 significance threshold. The lognormal and gamma distributions appear to be good approximations for the null importances. The features are presented in natural order. The bottom horizontal axis shows the number of the feature and the top horizontal axis indicates the number of categories that the features have.

a)



b)

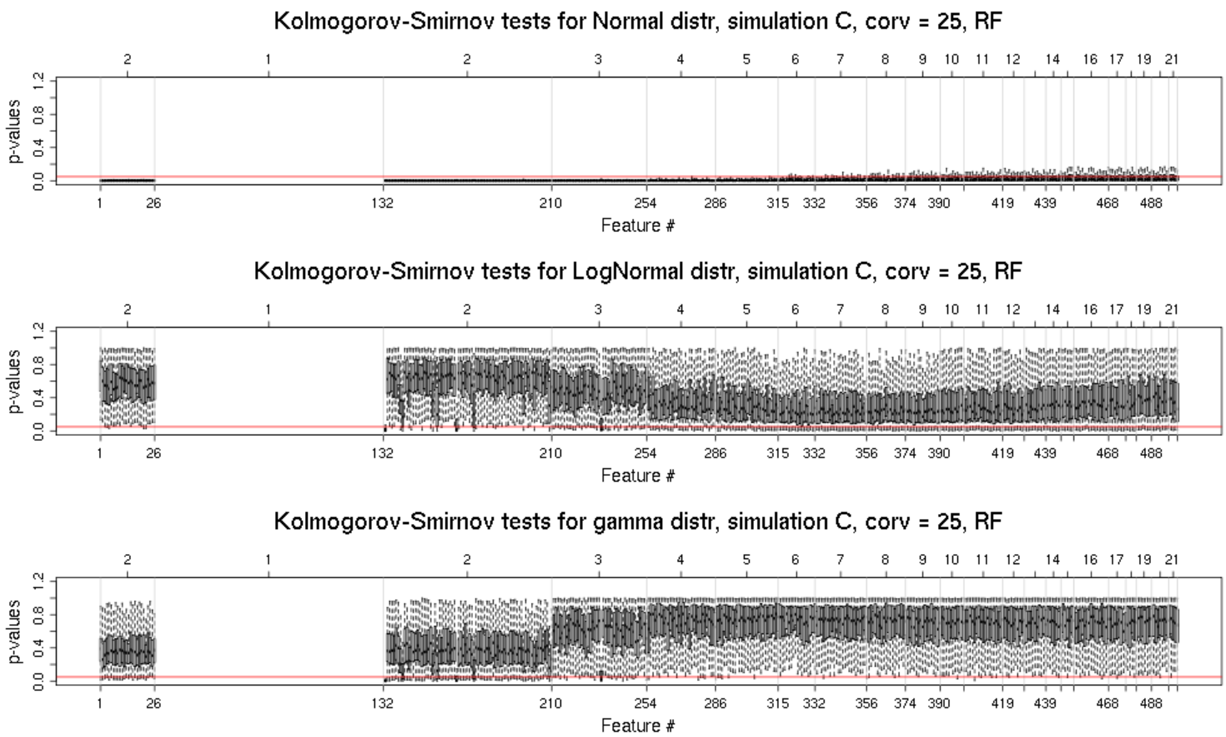
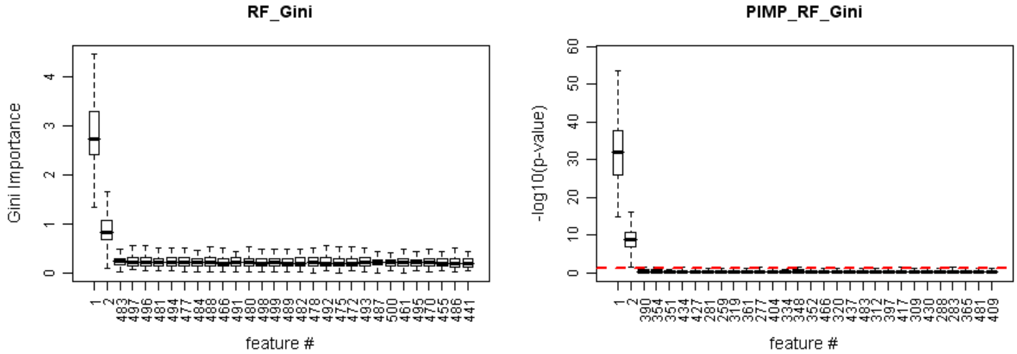
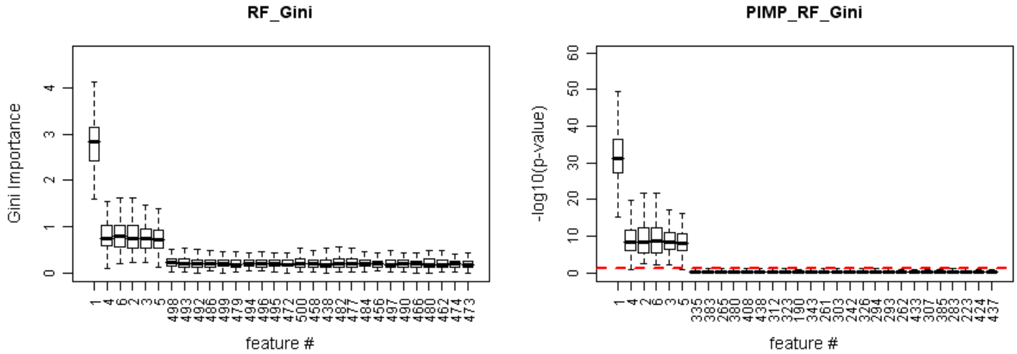


Figure S4. Discovery of relevant features in simulation C. Top ranking features are shown: in the left column RandomForest Gini importance, in the right column PIMP p-values. The red line depicts the 0.05 significance rate.

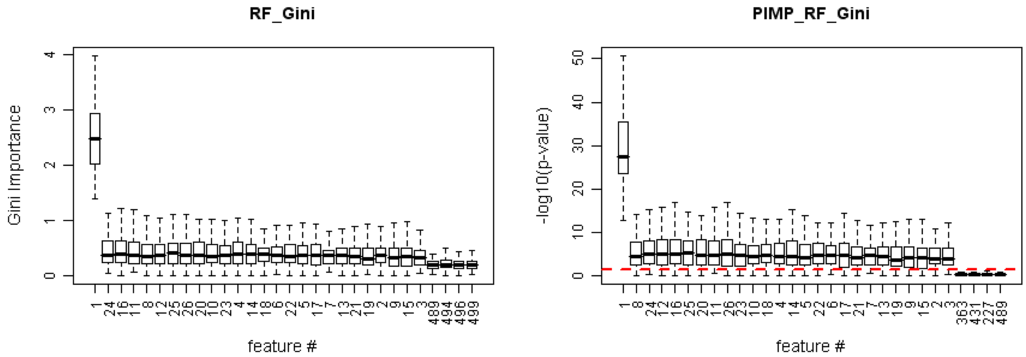
a) One variable in the second group.



b) Five correlated variables in the second group.



c) 25 correlated variables in the second group.



d) 50 correlated variables in the second group.

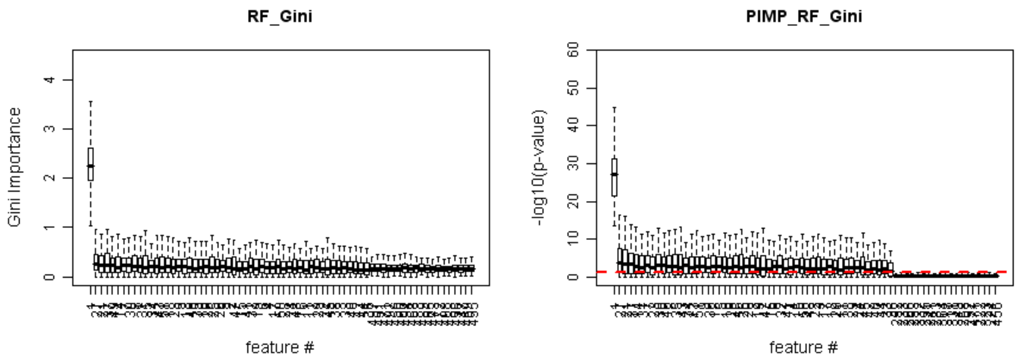


Figure S5. Boxplot for Kolmogorov-Smirnov (KS) tests for the distribution of null importances in the C-to-U dataset. The plot depicts the results of a KS test for normal distribution. The red horizontal line depicts the 0.05 significance threshold. High p-values indicate lack of evidence against a particular distribution. The normal distributions appear to be good approximations for the null importances.

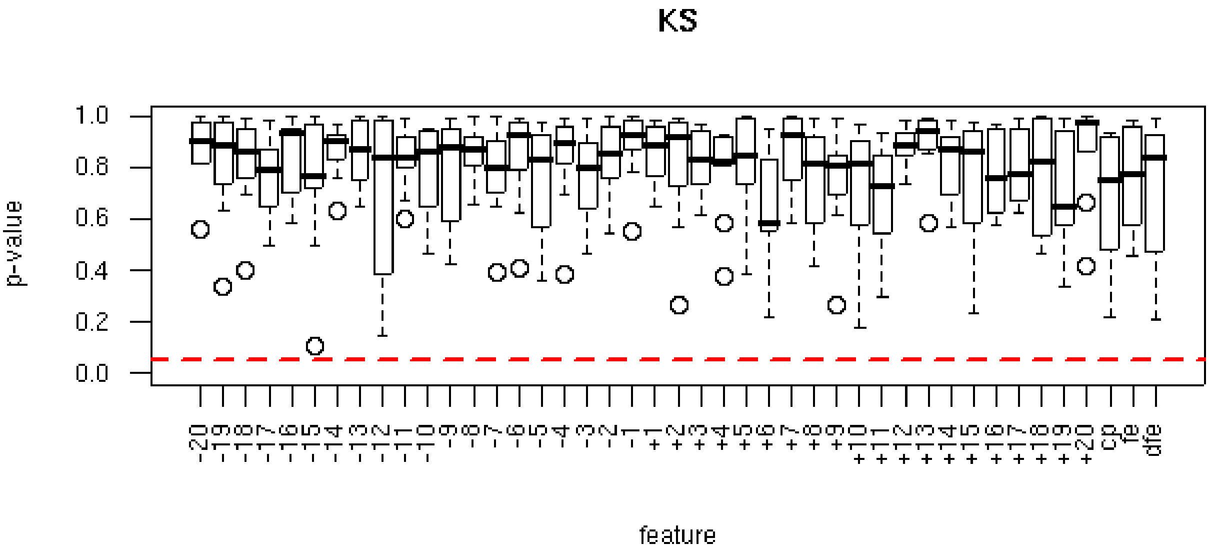
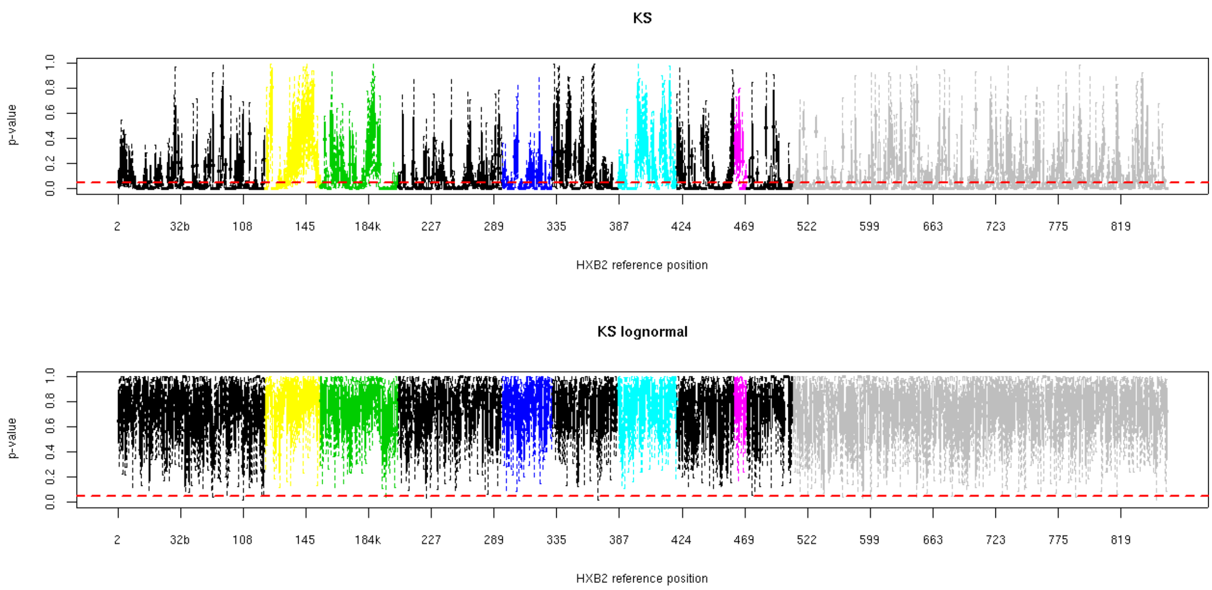


Figure S6. Boxplot for Kolmogorov-Smirnov (KS) tests for the distribution of null importances in the HIV dataset. The top (bottom) plot depicts the results of a KS test for normal (lognormal) distribution. Different colors indicate different regions of the Env gene (see main text for details). The red horizontal line depicts the 0.05 significance. High p-values indicate lack of evidence against a particular distribution. The lognormal distributions appear to be good approximations for the null importances.



Algorithm 1. The PIMP algorithm.

Algorithm 1 The permutation importance (PIMP) algorithm

Input: $P \in \mathbf{R}^{n \times p}$ (matrix of predictors),
p (the number of features),
l (response vector),
$VarImp$ (function to calculate variable importance),
s (number of permutations)

$\overrightarrow{\alpha} = VarImp(l, P),\ R \in \mathbf{R}^{s \times p}$
for $(i = 1; i \leq s; {+\!+}\, i)\{$
  $l' = permute(l)$
  $R_{i,*} = VarImp(l', P)$
$\}$
$\overrightarrow{\mu} \in \mathbf{R}^p, \overrightarrow{\sigma} \in \mathbf{R}^p$
for $(j = 1; j \leq p; {+\!+}\, j)\{$
  $\mu_j = mean(R_{*,j}), \sigma_j = sd(R_{*,j})$
$\}$
$\sigma_\mu = mean(\overrightarrow{\sigma}),\ \overrightarrow{\beta} \in \mathbf{R}^p$
for $(j = 1; j \leq p; {+\!+}\, j)\{$
  $\sigma' = \max\{\sigma_j, \sigma_\mu\}$
  $\beta_j = pnorm(\alpha_j, \mu_j, \sigma')$
$\}$
$return(\overrightarrow{\beta})$

The notation $R_{i,*}$ and $R_{*,j}$ refers to the i^{th} row and j^{th} column of the matrix R, respectively. The variable p denotes the number of different features (the number of columns of matrix P). The function $pnorm$ refers to the R-function that computes the probability of observing an importance of α_j or larger given a Gaussian distribution with mean μ_j and standard deviation σ'.