Supplementary Materials

VisuNet: an interactive tool for rule network visualization of rule-based learning models

Karolina Smolinska[1], Mateusz Garbulowski[1], Klev Diamanti[1,2], Xavier Davoy[4], Stephen O. O. Anyango[1], Fredrik Barrenäs[1], Susanne Bornelöv[3], Jan Komorowski[1,5,6,7]

[1]Science for Life Laboratory, Department of Cell and Molecular Biology, Uppsala University, Uppsala, Sweden

[2]Science for Life Laboratory, Department of Immunology, Genetics and Pathology, Uppsala University, Uppsala, Sweden

[3]Cancer Research UK Cambridge Institute, University of Cambridge, England

[4]The Grenoble Institute of Technology– Phelma, Grenoble, France

[5]Swedish Collegium for Advanced Study, Uppsala, Sweden

[6]The Institute of Computer Science, Polish Academy of Sciences, Warszawa, Poland

[7]Washington National Primate Research Center, Seattle, WA, USA

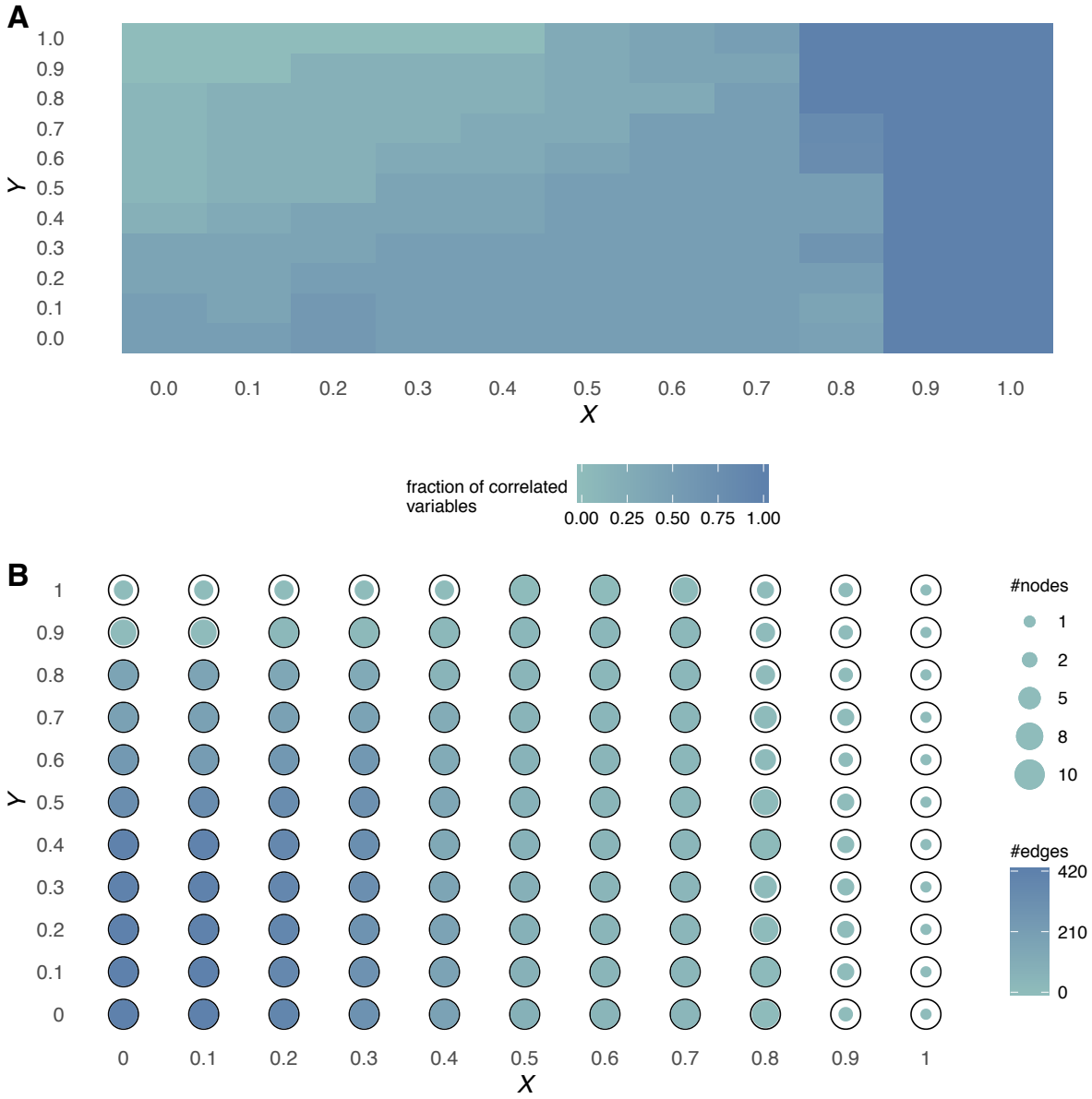
Supplementary Methods

Type 2 diabetes study on ULSAM

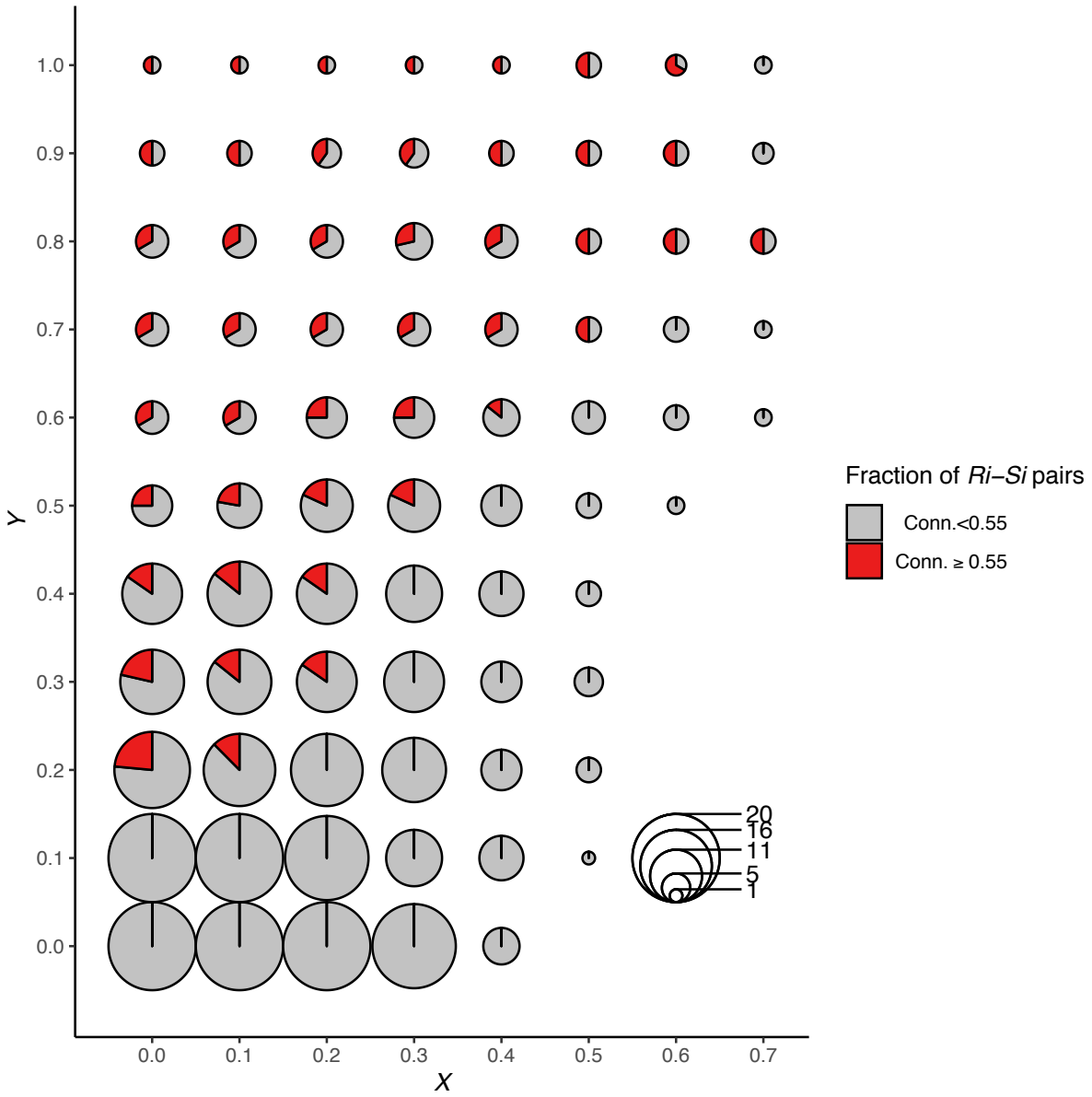
The final dataset of 888 samples and 207 features was corrected for the anthropometric measurements body-mass index, waist-hip ratio and age, that could influence associations among metabolites and decision. Correction was performed retrieving the residuals from a regression model with the R package limma [1]. Next, we used the R feature selection tool rMCFS that uses decision trees to select non-linearly associated variables to the outcome [2, 3]. rMCFS was run with the following settings: projections = 10000, projectionSize = 50, cutoffPermutations = 100, featureFreq = 1000, cutoffMethod = 'criticalAngle', balance = 8_balance, buildID = TRUE, finalCV = TRUE, finalRuleset = TRUE, finalCVSetSize = 1000, seed = 2019, threadsNumber = 2, splitSetSize = 0, mode = 1. rMCFS selected 18 significant variables based on the criticalAngle cutoff method.

We applied a histogram's gap approach that allowed the identification of three outliers based on the assumption that there is a gap in a histogram that represents normal samples and outliers. Next, each variable was discretized from an equal frequency binning approach (n=3) that was applied to the values not further than two standard deviations from the mean and cut-values were obtained. Variable-specific cut-values were then applied to discretize the dataset in three classes: LOW, MEDIUM and HIGH. We then ran R.ROSETTA on the discretized dataset of 885 samples and 18 variables with the following settings: classifier = "ObjectTrackingVoter", discrete = TRUE, roc = TRUE, clroc = "T2D", fallBack = TRUE, fallBackClass = "T2D", underSample = TRUE, underSampleNum = 100, JohnsonParam = list(Fraction=0.9), pAdjustMethod = "BH", seed = 2019. Finally, we ran the R.ROSETTA function recalculateRules on the rule-set from R.ROSETTA, we filtered rules on $p \leq 0.05$ [4].

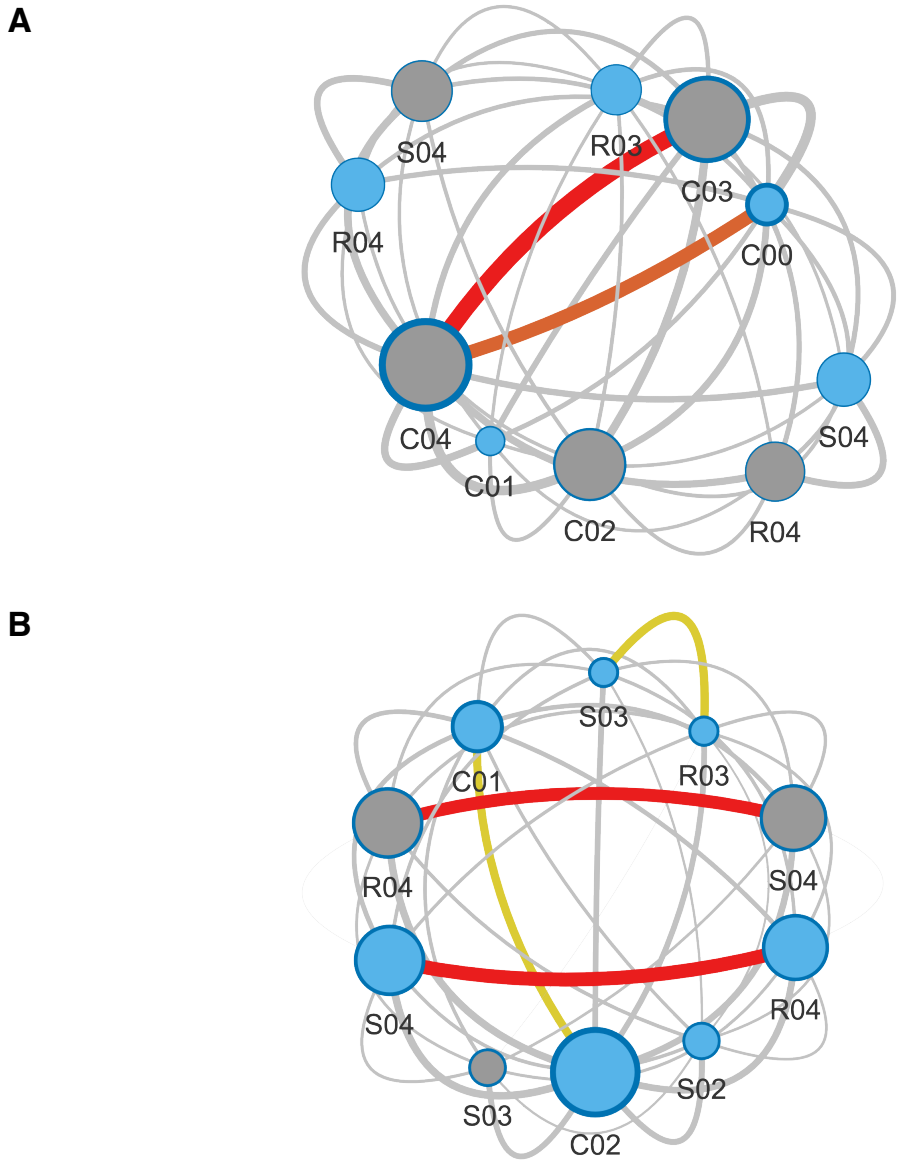
Supplementary Figures



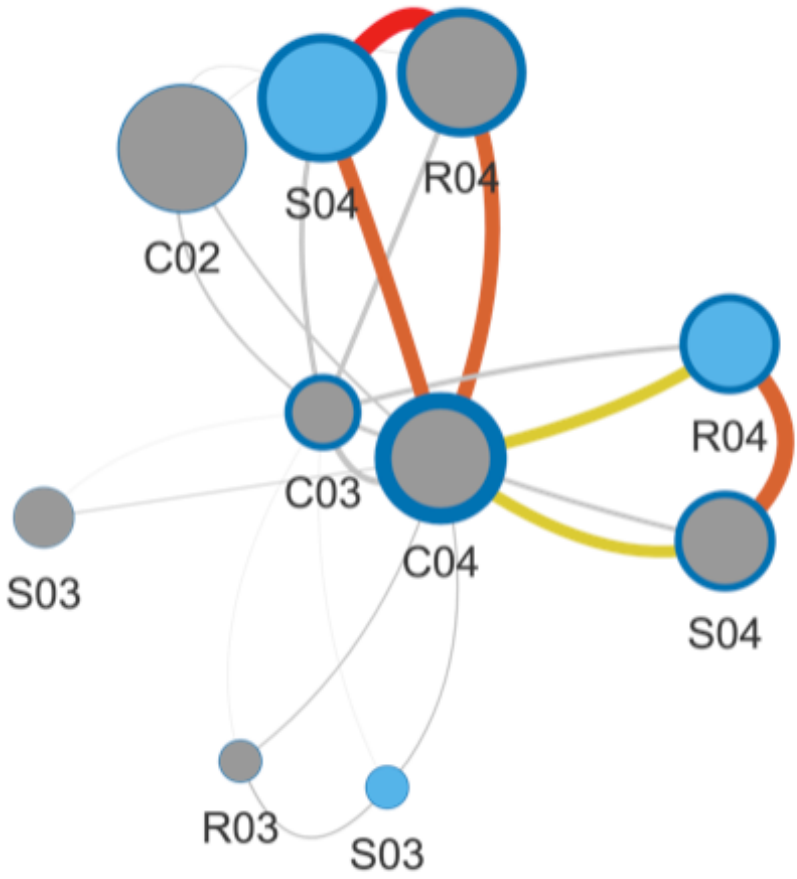
Supplementary Figure 1. The overview of RNs for the simulated data for all combinations of the *X* and *Y* values. A) The heatmap of fractions of SFCD (*C*00 – *C*04 features) on RNs. B) The plot of the structure of RNs. The size of the nodes indicated the number of nodes on RN and the intensity of the color corresponded to the number of edges on RN.



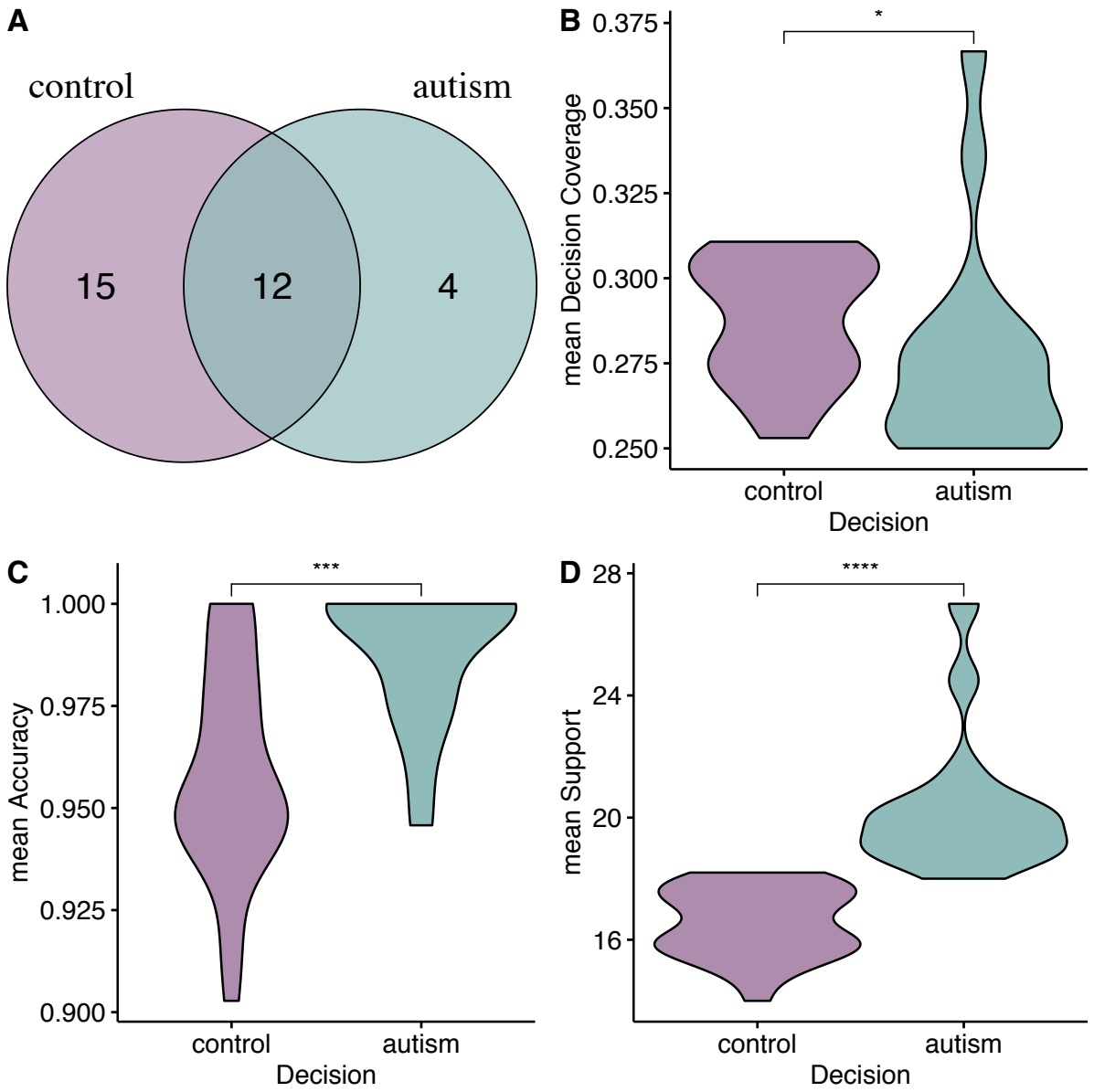
Supplementary Figure 2. The set of piecharts for the simulated data for all combinations of the *X* and *Y* values. The size of piecharts was correlated with the number of interdependencies between CPFD on RN. The fraction of the strongest interdependences (connection value equal or higher than 0.55) between CPFD on RN were assigned as red. Gray part of the circle corresponded to the interdependences between CPFD for connection values below 0.55 on RN.



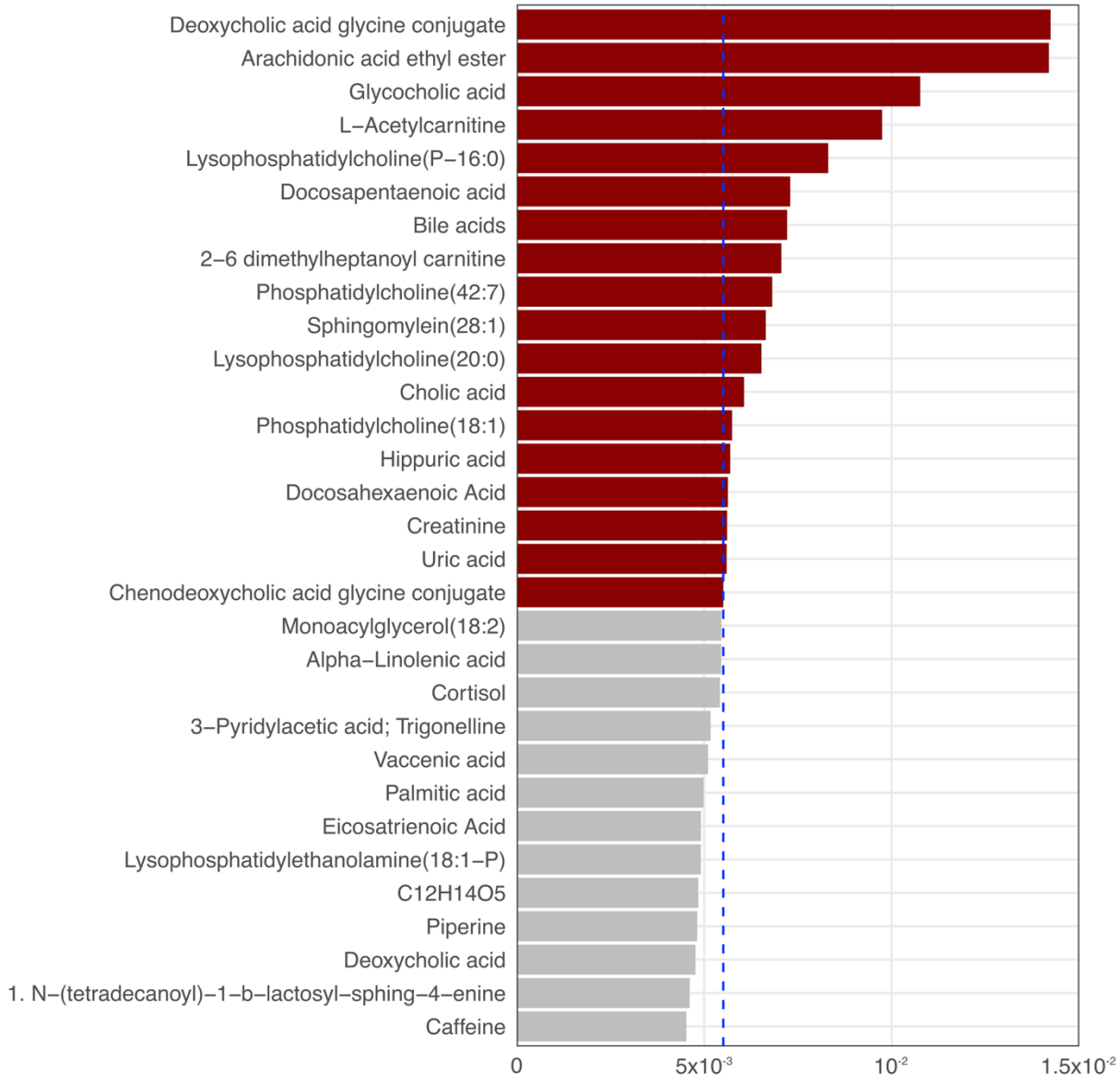
Supplementary Figure 3. RNs generated for a simulated data for X=0.2 and Y=0.2. A) Original RN and B) RN generated after removing the 2 most correlated features (C03 and C04) from RBM.



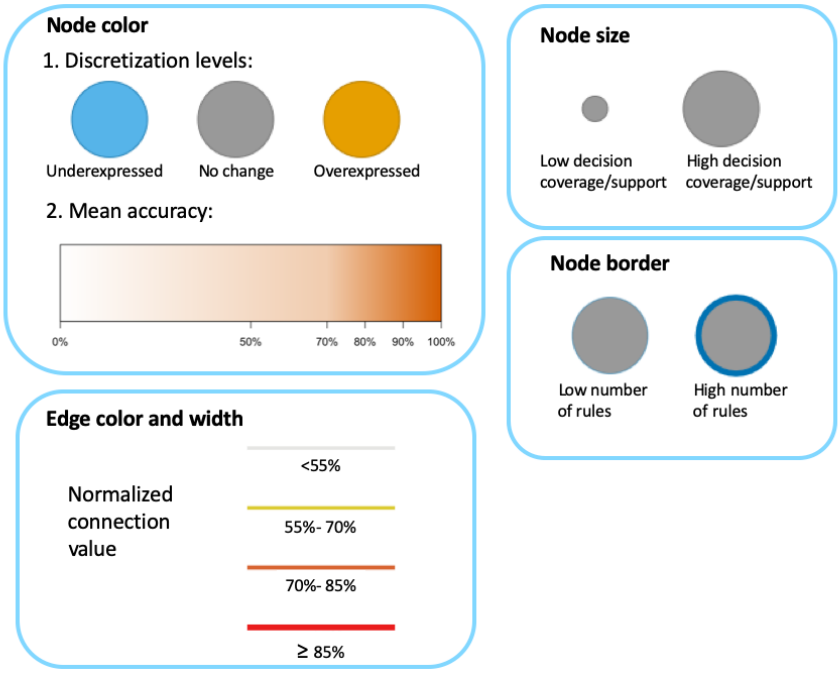
Supplementary Figure 4. RN obtained by using the ARBM for the simulated data for X=0.6 and Y=0.8. See Figure 2 for the description of labels.



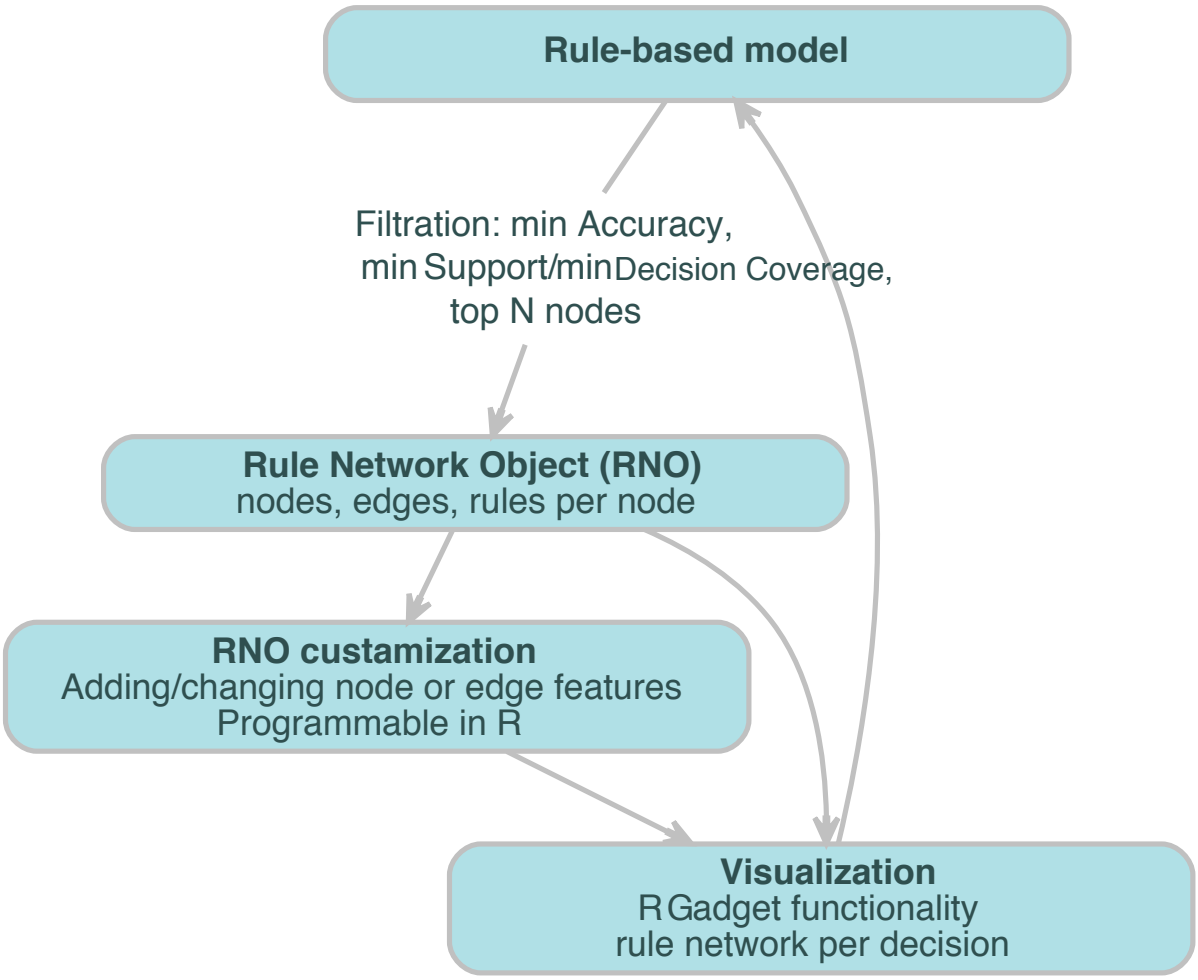
Supplementary Figure 5. A subnetworks comparison for the case-control study of autism. A) An intersection of gene on subnetworks for autism and control. Comparisons of distributions of the B) mean decision coverage, C) mean accuracy values and D) mean supports of nodes in RN between control and autism decisions. The stars indicated the significance of differences between the mean values distributions: * for *P*-value ≤ 0.05, ** for *P*-value ≤ 0.01 and *** for $P \leq 0.001$.



Supplementary Figure 6. Relative importance of the top-30 variables ranked by rMCFS. Colored in dark red are the ones marked as significant to perform the classification task, while in grey the non-important ones.



Supplementary Figure 7. The RN legend. The parameters used to display RN, could be specify from the command line in R or in the interface (See Methods for details).



Supplementary Figure 8. A VisuNet algorithm used to create a RN visualization and customization.

Bibliography

1. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK: **limma powers differential expression analyses for RNA-sequencing and microarray studies.** *Nucleic acids research* 2015, **43:**e47-e47.

2. Dramiński M, Koronacki J: **rmcfs: an R package for Monte Carlo feature selection and interdependency discovery.** *Journal of Statistical Software* 2018, **85:**1-28.

3. Dramiński M, Rada-Iglesias A, Enroth S, Wadelius C, Koronacki J, Komorowski J: **Monte Carlo feature selection for supervised classification.** *Bioinformatics* 2008, **24:**110-117.

4. Garbulowski M, Diamanti K, Smolińska K, Stoll P, Bornelöv S, Øhrn A, Komorowski J: **R. ROSETTA: a package for analysis of rule-based classification models.** *bioRxiv* 2019**:**625905.