# Supplementary material for

# Exact hypothesis testing for shrinkage based Gaussian Graphical Models

**Authors:** Victor Bernal[1,2*], Rainer Bischoff[2], Victor Guryev[3], Marco Grzegorczyk[1], Peter Horvatovich[2]

[1] Bernoulli Institute, University of Groningen, Groningen, 9747 AG, The Netherlands.

[2] Department of Analytical Biochemistry, Groningen Research Institute of Pharmacy, University of Groningen, Groningen, 9713 AV, The Netherlands.

[3] European Research Institute for the Biology of Ageing, University Medical Center Groningen, University of Groningen, Groningen, 9713 AV, The Netherlands.

This document contains supporting material for the work *"Exact hypothesis testing for shrinkage based Gaussian Graphical Models"*. It includes some examples of the propagation of the shrinkage effects to the partial correlation, a detailed description of *p-value* estimation using Monte Carlo, and additional figures referenced in the main manuscript.

## Table of Contents

## S1 Shrinkage effect on the partial correlation coefficient- Examples

In this section we illustrate the shrinkage effects on the partial correlations in an analytical way. We chose two cases in which the correlation matrix is not invertible, and one case in which it is. The "shrunk" correlation matrix is computed, and inverted to find the matrix of partial correlation coefficient (i.e. the GGM) analytically. In particular, the third case where the correlation is already invertible, the two matrices (with and without shrinkage) are compared.

*Example 1:* Let's consider a 3 × 3 correlation matrix **R** where 2 random variables have maximum correlation, and the third one is independent from the others. Then

$$\mathbf{C} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Here **C** is not full ranked, its determinant is zero and it cannot be inverted. The shrunk correlation matrix is found by employing **Eq. 3**

$$\mathbf{R}^{\lambda} = \begin{pmatrix} 1 & (1-\lambda) & 0 \\ (1-\lambda) & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Its determinant is $1 - (1-\lambda)^2 \neq 0$, and it is invertible. The matrix of “shrunk” partial correlations is found with **Eq.1.**

$$\mathbf{P}^{\lambda} = \begin{pmatrix} 1 & (1-\lambda) & 0 \\ (1-\lambda) & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Illustrating the fact that the partial correlation is bounded by $(1-\lambda)$.

*Example 2:* Let’s consider a 3 × 3 correlation matrix **R** where all random variables have a correlation of 1. Then

$$\mathbf{R} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}.$$

Here **C** is not full ranked, its determinant is zero and it cannot be inverted. On the other hand, the shrunk correlation matrix is

$$\mathbf{R}^{\lambda} = \begin{pmatrix} 1 & (1-\lambda) & (1-\lambda) \\ (1-\lambda) & 1 & (1-\lambda) \\ (1-\lambda) & (1-\lambda) & 1 \end{pmatrix},$$

which has a determinant $1 - (1-\lambda)^2 \neq 0$, and it is invertible. The matrix of “shrunk” partial correlations is found with **Eq.1** and has off-diagonal elements $\mathbf{P}_{ij}^{\lambda} = \frac{(1-\lambda)}{(2-\lambda)}$ illustrating the fact that the partial correlation is bounded by $(1-\lambda)$.

*Example 3:* Let’s consider a 3 × 3 correlation matrix **R** where 2 random variables are the same, and the third one is independent from the others. Then

$$\mathbf{R} = \begin{pmatrix} 1 & 0.9 & 0 \\ 0.9 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Here **C** is full ranked, its determinant is $1 - (0.9)^2$, and it can be inverted. The shrunk correlation matrix is

$$\mathbf{R}^{\lambda} = \begin{pmatrix} 1 & (1-\lambda)0.9 & 0 \\ (1-\lambda)0.9 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Its determinant is $1 - \left((1-\lambda)0.9\right)^2$. The matrix of “shrunk” partial correlations is

$$\mathbf{P}^{\lambda} = \begin{pmatrix} 1 & (1-\lambda)0.9 & 0 \\ (1-\lambda)0.9 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Illustrating the fact that effect propagated gives $\mathbf{P}_{12}{}^{\lambda} = \mathbf{P}_{12} - 0.9\lambda$.

## S2 SHRUNK MLE

Given that the “shrunk” coefficients were inferred, a mixture of partial correlations is obtained (involving the null and real effects). Next, the goal is to test edge-wise the null hypothesis $H_0: \rho = 0$, however, when $n < p$ the degrees of freedom $k$ has no clear interpretation. Even due, the degrees of freedom $k$ in **Eq.7** can be estimated via Maximum Likelihood Estimation (MLE). Suppose the real data consists of $p$ variables and $n$ samples, then

1. Estimate the $\frac{p(p-1)}{2}$ “shrunk” partial correlations from the real data (SCHÄFER, et al., 2005).
2. Estimate the degrees of freedom $k^{\lambda}$ numerically
    a. Simulate $\frac{p(p-1)}{2}$ shrunk partial correlations from $H_0: \rho = 0$ (i.e. the precision matrix is the identity) with sample size $n$.
    b. Find $\widehat{k^{\lambda}}$ by maximizing the likelihood function for 2.a using **Eq.7.**
3. Using **Eq.7** with $\widehat{k^{\lambda}}$ (i.e. $f^{\lambda}(\rho, \widehat{k^{\lambda}})$) compute the *p-values* for the coefficients from step 1.

## S3 MONTE CARLO *P-VALUE* ESTIMATION

Suppose the real data consist on $p$ variables and sample size $n$. The *p-values* can be estimated by Monte Carlo (MC) as follows.

1. Estimate the $\frac{p(p-1)}{2}$ “shrunk” partial correlations $\rho_i$ from the real data (Schäfer and Strimmer, 2005).
2. Simulate data of length $n$ from $H_0: \rho = 0$ (i.e. the precision matrix is a $p \times p$ identity).
3. Reconstruct the “shrunk” partial correlations $\rho_{0i}$ from step 2 using (Schäfer and Strimmer, 2005) .
4. Compute the MC empirical *p-values.*
    a. For each $\rho_j$ (from step 1) its corresponding *p-value* estimator $\widehat{p_j}$ is defined as $\widehat{p_j} = \frac{1}{M}\sum_{i=1}^{M} \mathbb{1}_{|\rho_{0i}| \geq |\rho_j|}$.

where $\mathbb{1}$ denotes the indicator function, $|*|$ denotes the absolute value, and $M$ is the total number of coefficients (i.e. $\frac{p(p-1)}{2}$).

In other words, $\widehat{p_j}$ is the proportion of $|\rho_0\ |$s (from step 3) equal or greater than $|\rho_j|$ (from step 1).

## S4 SUPPLEMENTARY FIGURES AND TABLES

This section consists in figures and tables supporting the main manuscript.

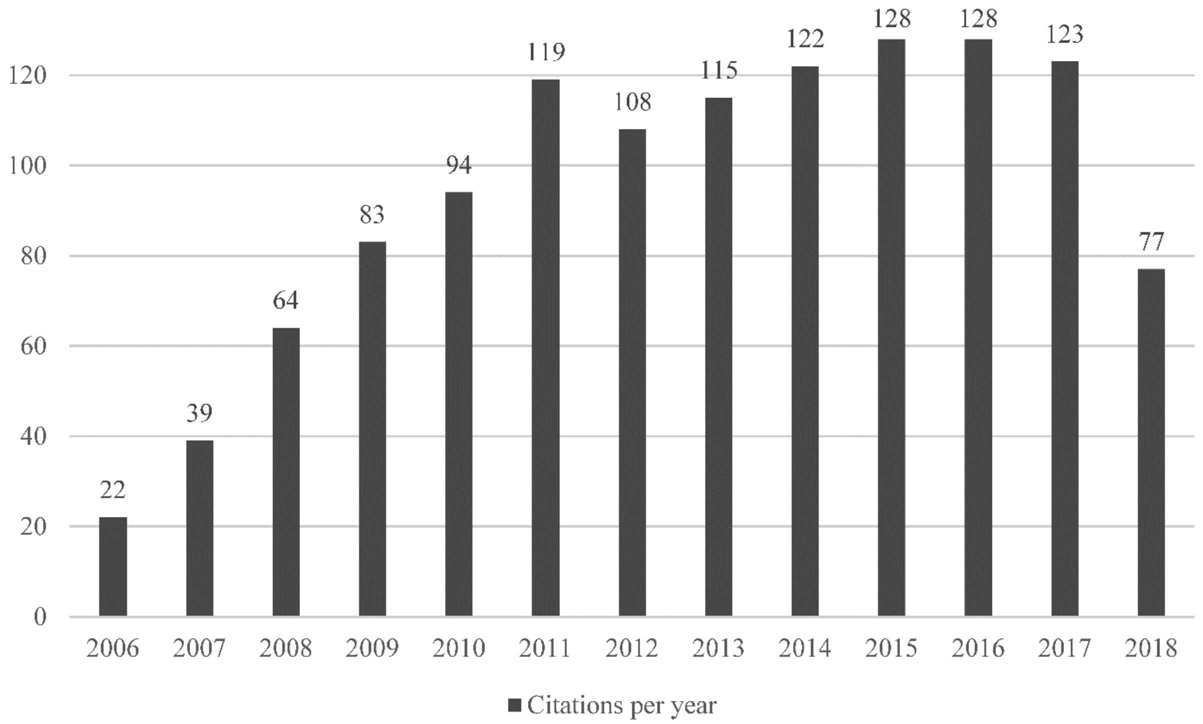

**Fig S1.-** *GeneNet's* citations per year. The histogram shows the number of citations per year of the R package GeneNet (Schäfer and Strimmer, 2005a). It has obtained more than 1200 citations to date (77 on the current year). Source: Google Scholar (9th of September 2018).

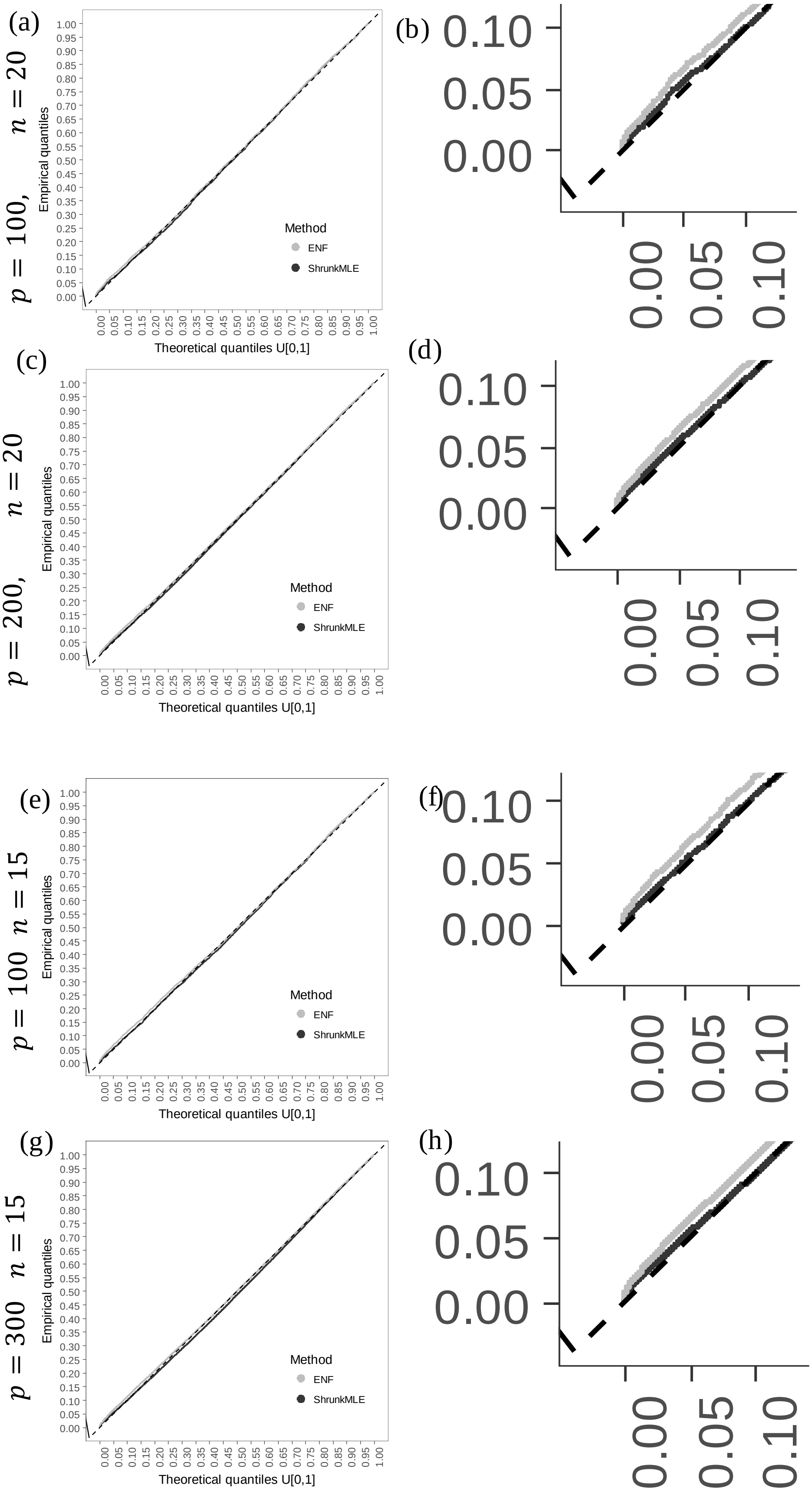

**Fig S2.-** Q-Q plots under $H_0$. This figure shows a comparison of the *p-values'* quantiles obtained with (i) ENF (grey), and (ii) Shrunk MLE (black), against the theoretical quantiles (i.e. uniform in [0,1]). Panels (a), (c), (e), and (g) show the Q-Q plots of the empirical *p-values*. Panels (b), (d), (f), and (g) display the respective zoom-in into the lower tails. The diagonal line (dashed black) is a graphical representation of a perfect agreement (in the case when inherent randomness is ignored). It can be seen that the quantiles from ENF are larger in the tails than the ones from Shrunk MLE. The simulations were performed with $p = 100,\ 200,\ 300$ and $n = 15,\ 20$.

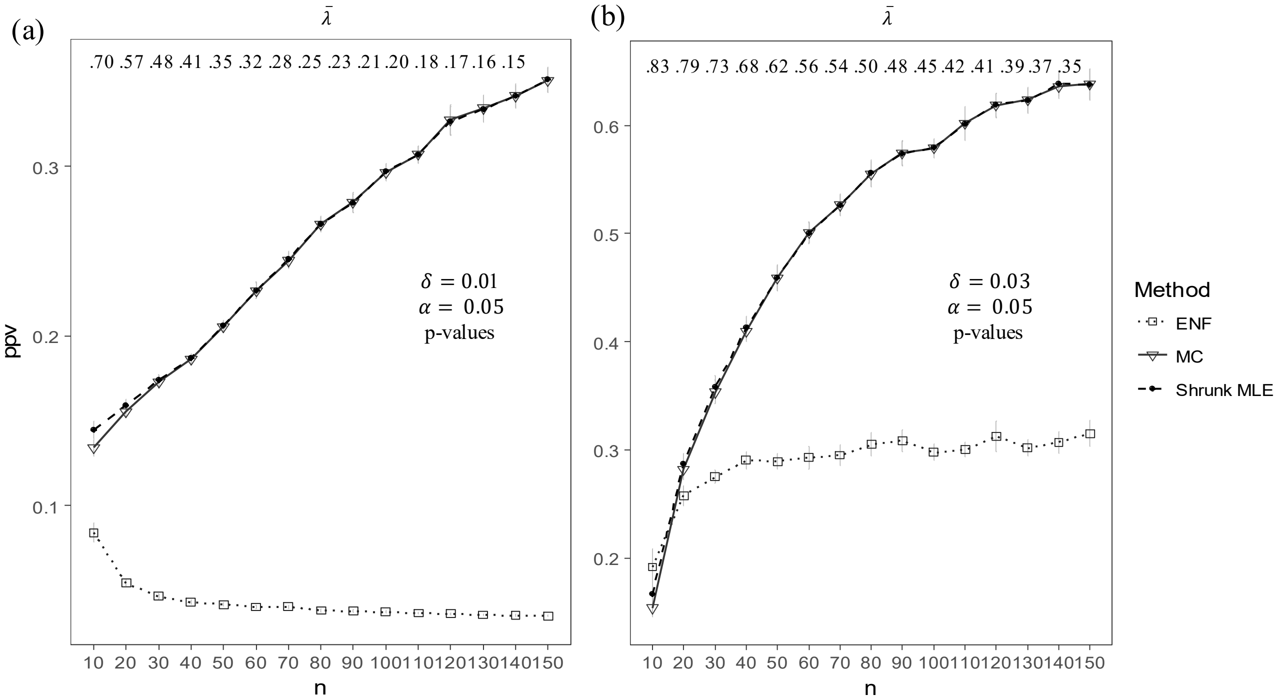

**Fig S3.-** Positive Predictive Value versus sample size using un-adjusted *p-values*. This figure shows the Positive Predictive Value (PPV) obtained with different sample sizes. The inference is carried out from simulated data for $p$=100 with $n$ ranging from 10 to 150 in steps of size 10, tested at $\alpha$=0.05. The panels (a) and (b) show the PPV with un-adjusted *p-values* with $\delta$=0.01 (or 49 correlations) and $\delta$=0.03 (or 148 correlations), respectively. Three approaches are compared: ENF (dot with dashed line), Shrunk MLE (square with dotted line), and MC with 15 iterations (triangle with continuous line). Symbols (and bars) represent the average (+/- 2 standard errors) over 25 repeated simulations. The upper horizontal axis shows the average shrinkage intensity $\bar{\lambda}$ rounded to two digits.

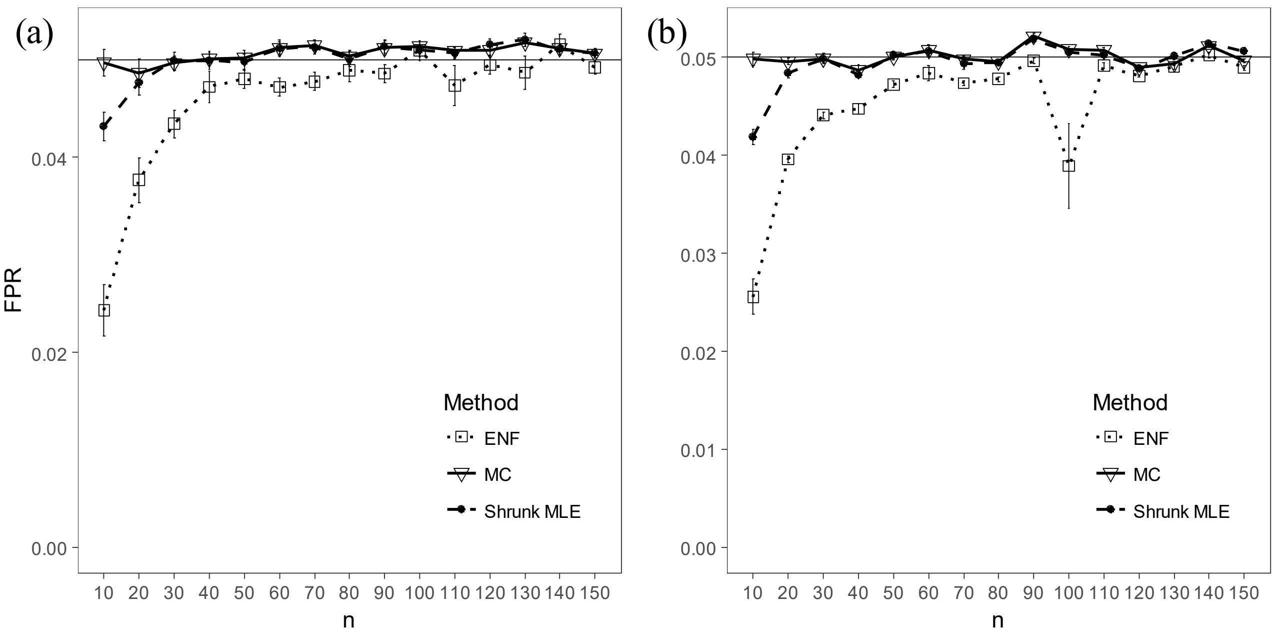

**Fig S4.-** Type I error versus sample size. This figure shows the Type I error (i.e. the false positives rate (FPRs)) under $H_0$ obtained with different $p$ and sample sizes $n$. Inference is carried out from simulated data for $p$=100, 200 in panels (a), and (b). The sample size $n$ ranges from 10 to 150 in steps of size 10 and the test is carried out at $\alpha = 0.05$. Three approaches are compared: ENF (dot with dashed line), Shrunk MLE (square with dotted line), and MC with 15 iterations (triangle with continuous line). Symbols (and bars) represent the average (+/- 2 standard errors) over 25, and 5 repeated simulations respectively.

**Table S1.** Number of simulated datasets per figure.

| Analysis | Number of (independent) simulated data sets | Place in the manuscript |
|---|---|---|
| Histograms | 25 × 2 = 50 | Page 4, Figure 2 |
| FP versus $n$ | 25 × 15× 2 = 750 | Page 5, Figure 3 |
| FP versus $p/n$ | 25 × 4 × 2 = 200 | Page 5, Figure 3 |
| Heat map of FPs | 9 × 9 × 10 × 2 = 1620 | Page 5, Figure 4 |
| PPV adjusted *p values* | 25 × 15× 2 = 750 | Page 5, Figure 5 |
| Q-Q plots (supplementary) | 4 | Page 5, Figure S2 |
| PPV un-adjusted *p values* (supplementary) | 25 × 15× 2 = 750 | Page 6, Figure S3 |
| Type-I error rate (supplementary) | 25 × 15 +5 ×15= 450 | Page 6, Figure S4 |
| Total | 4574 | |

**Table S2.** Quantities used to assess the GGM reconstruction.

| Quantity | Definition |
|---|---|
| True Positives (TP)/ True Negatives (TN) | Real/null effects correctly classified as such |
| False Positives (FP)/ False Negatives (FN) | Real/null effects incorrectly classified as such |
| Positive Predictive Value (PPV) | $\frac{TP}{TP+FP}$ |
| False Positive Rate (FPR) | $\frac{FP}{FP+TN}$ |
| True Positive Rate (TPR) | $\frac{TP}{FN+TP}$ |

**Table S3.** Comparison of computational times for the GGM reconstruction with different methods. Computational times for three methods: Shrunk MLE, ENF and MC as function of the number of variables $p$ in the GGM. Simulations were performed in an Intel(R) Core(TM) i5 CPU with 4 Gb of RAM (R version 3.4.3). We have used *R. 3.4.0,* and *GeneNet* version *1.2.13.* The computing time depends on the number of test (i.e. $p$) and not on the sample size.

| Method | Time (sec) | | | |
|---|---|---|---|---|
| | $p = 100$ | $p = 200$ | $p = 500$ | $p = 1000$ |
| Shrunk MLE | 0.41 | 1.29 | 2.51 | 8.00 |
| *GeneNet* | 0.07 | 0.28 | 0.62 | 1.89 |
| Monte Carlo* (MC) | 2.40 | 32.97 | 163.38 | 1256.70 |

*10 iterations.

**Table S4 a.** Most significant GOs with Shrunk MLE (*Escherichia coli*). The table reports the 10 most enriched GO pathways for the set of connected genes (at least one connection) in the GGM for *Escherichia coli*. The GGM is reconstructed with Shrunk MLE, and the enrichment is carried out using PANTHER Classification System (http://geneontology.org/) with ($FDR \leq 0.05$).

| GO biological process | fold Enrichment | raw *p-value* | (FDR) |
|---|---|---|---|
| cellular response to virus (GO:0098586) | 61.51 | 7.58 $10^{-5}$ | 5.61 $10^{-2}$ |
| phage shock (GO:0009271) | 61.51 | 7.58 $10^{-5}$ | 4.48 $10^{-2}$ |
| response to temperature stimulus (GO:0009266) | 8.35 | 1.51 $10^{-7}$ | 4.48 $10^{-4}$ |
| response to heat (GO:0009408) | 8.20 | 9.82 $10^{-6}$ | 1.45 $10^{-2}$ |
| response to stress (GO:0006950) | 2.57 | 3.92 $10^{-5}$ | 3.87 $10^{-2}$ |

**Table S4 b.** Most significant GOs with ENF (*Escherichia coli*). The table reports the 10 most enriched GO pathways for the set of connected genes (at least one connection) in the GGM, for *Escherichia coli*. The GGM is reconstructed with ENF (*GeneNet* version *1.2.13*), and the enrichment is carried out using PANTHER Classification System (http://geneontology.org/) with ($FDR \leq 0.05$).

| GO biological process | fold Enrichment | raw *p-value* | (FDR) |
|---|---|---|---|
| aerobic respiration (GO:0009060) | 9.76 | 3.59 $10^{-6}$ | 5.30 $10^{-3}$ |
| response to heat (GO:0009408) | 7.69 | 4.96 $10^{-6}$ | 4.89 $10^{-3}$ |
| response to temperature stimulus (GO:0009266) | 7.59 | 1.29 $10^{-7}$ | 3.81 $10^{-4}$ |

**Table S4 c.** Hubs in the GGM structure for *Escherichia coli* dataset. The table reports three central genes that are highly connected in the structure, as well as the most strongly connected genes to the central one ($\alpha$=0.001 and $\alpha$=0.01).

| Central gene | Connected at $\alpha = 0.001$ | Connected at $\alpha = 0.01$ |
|---|---|---|
| cspG* | pspA, pspB, cspA, yecO, lacA* | yedE, yaeM, lacY* |
| yhel | ycgX , dnaK, b1963, yedE | dnaG, atpD, folk |
| lacA* | lacY*, lacZ, asnaA, cspG* | yaeM |
| sucA | dnaJ, atpG | flgD, gltA, sucD, b1191, yhfV, yhDM, tnaA |

*it connects 2 hubs

**Table S5 a.** GGM comparative results for *Mus musculus* data set. Comparison of the number of connections obtained with BH-adjusted *p-values* $\alpha$=0.10. The assessment is in terms of True Positive Rate (TPR) and False Positive Rate (FPR) using STRING (https://string-db.org/) with a confidence strength of 0.90 for database results.

| Method | Connections | TPR | FPR | PPV(%)* |
|---|---|---|---|---|
| Shrunk MLE | 2433 edges (310 genes) | 0.32 | **0.05** | **0.57** |
| MC | 6350 (335 genes) | 0.50 | 0.12 | 0.37 |
| ENF | 12616 edges (456 genes) | 0.39 | **0.11** | **0.25** |

**Table S5 b.** Most significant GOs with Shrunk MLE (*Mus musculus*). The table reports the 10 most enriched GO pathways for the set of connected genes (at least one connection) in the GGM. The GGM is reconstructed with Shrunk MLE for *Mus musculus*, and the enrichment is carried out using PANTHER Classification System (http://geneontology.org/) with ($FDR \leq 0.05$).

| GO biological process | Fold enrichment | raw *p-value* | (FDR) |
|---|---|---|---|
| immune response (GO:0006955) | 4.76 | 1.95 $10^{-32}$ | 3.02 $10^{-28}$ |
| immune system process (GO:0002376) | 3.66 | 7.28 $10^{-32}$ | 5.63 $10^{-28}$ |
| response to external biotic stimulus (GO:0043207) | 5.35 | 1.23 $10^{-27}$ | 4.77 $10^{-24}$ |
| response to other organism (GO:0051707) | 5.37 | 1.10 $10^{-27}$ | 5.69 $10^{-24}$ |
| defense response to other organism (GO:0098542) | 6.63 | 6.86 $10^{-27}$ | 1.77 $10^{-23}$ |
| response to biotic stimulus (GO:0009607) | 5.19 | 6.31 $10^{-27}$ | 1.95 $10^{-23}$ |
| immune effector process (GO:0002252) | 6.54 | 1.33 $10^{-25}$ | 2.93 $10^{-22}$ |
| defense response (GO:0006952) | 4.31 | 2.15 $10^{-25}$ | 4.15 $10^{-22}$ |
| positive regulation of immune response (GO:0050778) | 5.86 | 1.32 $10^{-23}$ | 2.27 $10^{-20}$ |
| B cell receptor signaling pathway (GO:0050853) | 11.52 | 5.92 $10^{-23}$ | 9.16 $10^{-20}$ |
| immune response (GO:0006955) | 4.76 | 1.95 $10^{-32}$ | 3.02 $10^{-28}$ |

**Table S5 c.** Most significant GOs with ENF (*Mus musculus*). The table reports the 10 most enriched GO pathways for the set of connected genes (at least one connection) in the GGM for *Mus musculus*. The GGM is reconstructed with *ENF* (*GeneNet version 1.2.13*), and the enrichment is carried out using PANTHER Classification System (http://geneontology.org/) with ($FDR \leq 0.05$).

| GO biological process | fold enrichment | raw *p-value* | (FDR) |
|---|---|---|---|
| immune response (GO:0006955) | 4.95 | 6.54 $10^{-49}$ | 1.01 $10^{-44}$ |
| immune system process (GO:0002376) | 3.66 | 1.57 $10^{-44}$ | 1.21 $10^{-40}$ |
| response to external biotic stimulus (GO:0043207) | 5.73 | 1.25 $10^{-43}$ | 4.83 $10^{-40}$ |
| response to other organism (GO:0051707) | 5.74 | 1.06 $10^{-43}$ | 5.44 $10^{-40}$ |
| defense response to other organism (GO:0098542) | 7.17 | 9.20 $10^{-43}$ | 2.85 $10^{-39}$ |
| response to biotic stimulus (GO:0009607) | 5.55 | 1.51 $10^{-42}$ | 3.33 $10^{-39}$ |
| immune effector process (GO:0002252) | 7.26 | 1.45 $10^{-42}$ | 3.73 $10^{-39}$ |
| defense response (GO:0006952) | 4.46 | 9.45 $10^{-38}$ | 1.83 $10^{-34}$ |
| protein activation cascade (GO:0072376) | 11.77 | 1.31 $10^{-35}$ | 2.02 $10^{-32}$ |
| complement activation (GO:0006956) | 12.25 | 1.20 $10^{-35}$ | 2.06 $10^{-32}$ |
| immune response (GO:0006955) | 4.95 | 6.54 $10^{-49}$ | 1.01 $10^{-44}$ |

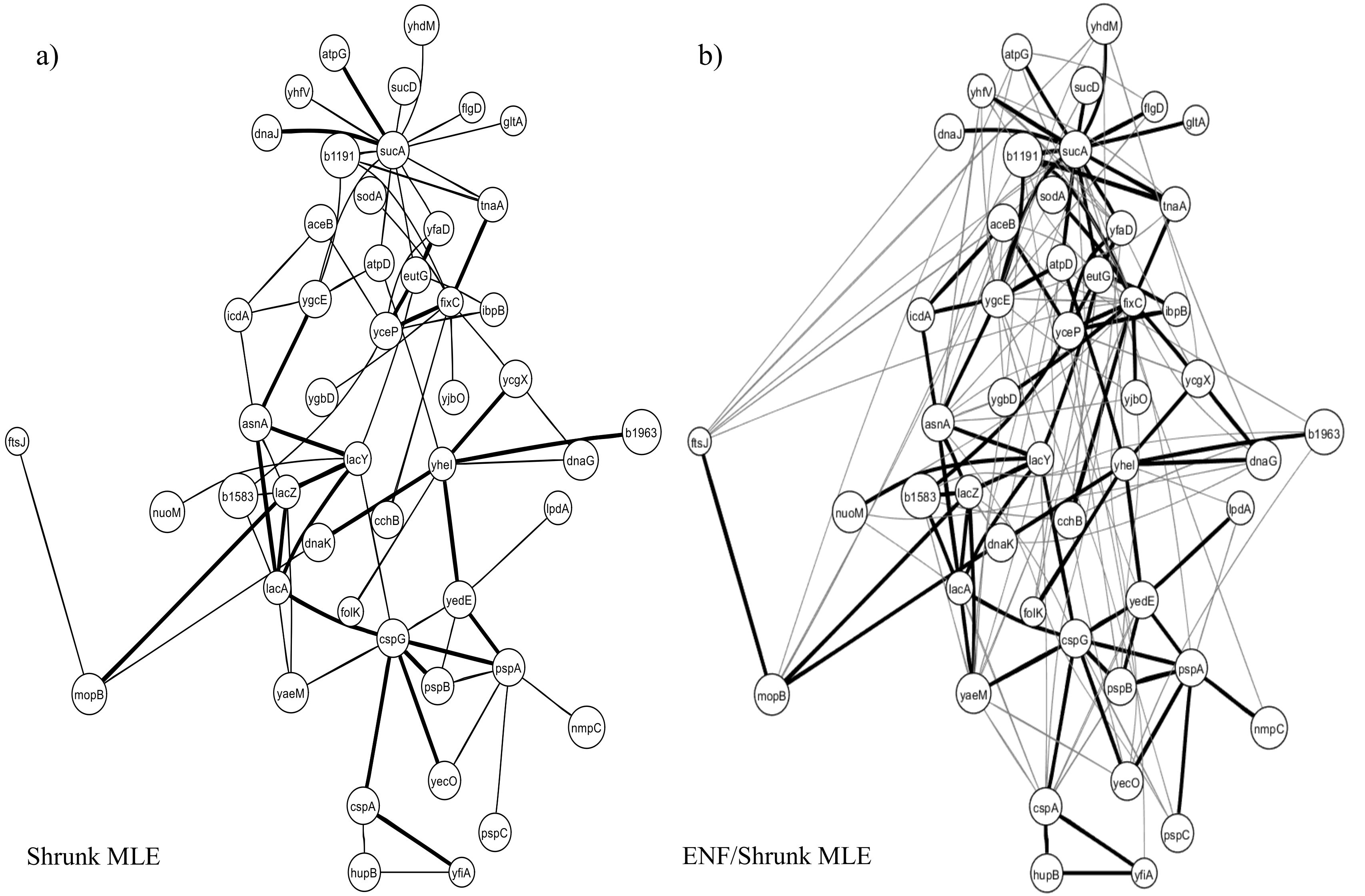

**Fig S5.-** GGM structure for *Escherichia coli*. The figure displays the GGM structure for *Escherichia coli for* the connected genes with Shrunk MLE at $\alpha = 0.01$. There are 49 genes at this level (unconnected genes are excluded from the figure). In the panel *a)* Shrunk MLE shows 48 significant edges at $\alpha = 0.001$ (thick black lines), and 152 edges at $\alpha = 0.01$ (thin black lines). In the panel *b)* ENF shows the same 49 genes with 152 significant edges from Shrunk MLE (thick black lines), and additional 330 edges (thick grey lines) both at $\alpha = 0.01$.

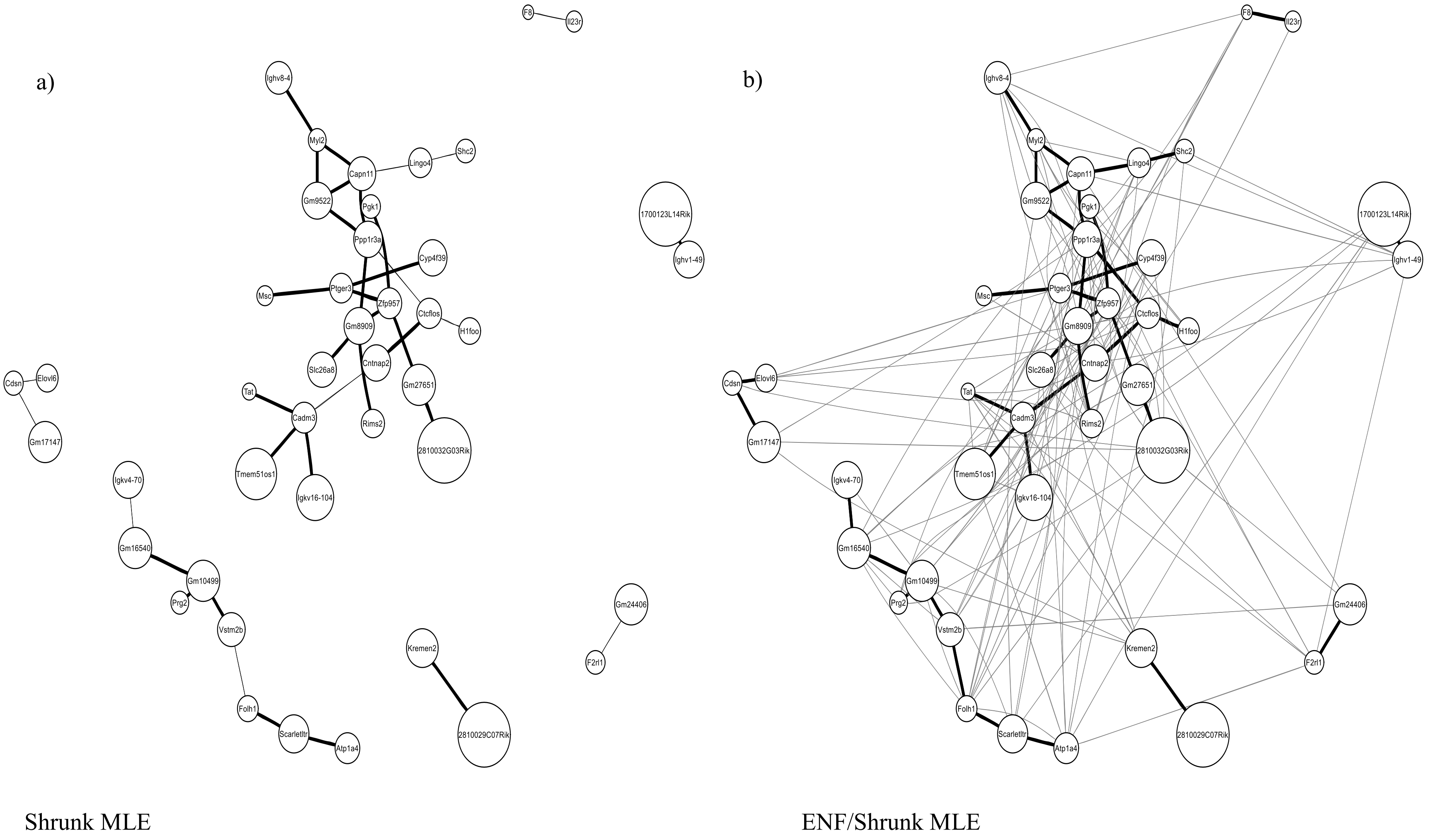

**Fig S6.**- GGM structure for *Mus musculus*. The figure displays the genes connected with Shrunk MLE at $\alpha = 10^{-11}$. There are 43 genes at this level (unconnected genes are excluded from the figure). In the panel *a)* 54 significant edges at $\alpha = 10^{-12}$ (thick black lines), and 76 edges at $\alpha = 10^{-11}$ (thin black lines) for Shrunk MLE. In the panel *b)* the same 43 genes with 76 significant edges from Shrunk MLE (thick black lines), and 302 edges from ENF (thin grey lines) at $\alpha = 10^{-11}$.