Statistical Analysis of High-Throughput Metabolomics Data

Differential Signatures Pertaining to Alzheimer's Disease

Carel F.W. Peeters
Dept. of Epidemiology & Biostatistics
VU University medical center, Amsterdam
cf.peeters@vumc.nl

Alzheimercyclus
VUmc Alzheimercentrum
VU University Medical Center, Amsterdam
October 21, 2016

Contributors

Francien de Leeuw
Alzheimer Center, VUMC

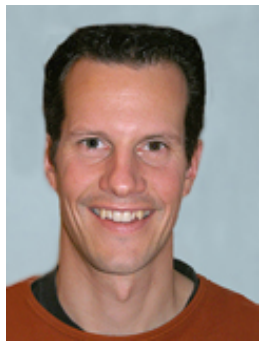
Charlotte Teunissen
Neurology Laboratory, Dept. of Clinical Chemistry, VUMC

Wiesje van der Flier
Alzheimer Center, VUMC
Dept. of Epimiology & Biostatistics, VUMC

Contributors

Wessel N. van Wieringen
Dept. of Epimiology & Biostatistics, VUMC
Dept. of Mathematics, VU University Amsterdam

Anders E. Bilgrau
Novo Nordisk
Dept. of Mathematical Sciences, Aalborg University



Mark A. van de Wiel
Dept. of Epimiology & Biostatistics, VUMC
Dept. of Mathematics, VU University Amsterdam

Contributors

Thomas Hankemeier
Division for Analytical Biosciences, Leiden University

Herman van Vlijmen
Dept. of Medicinal Chemistry, Leiden University
Molecular Sciences div., Janssen Pharmaceutica

Cornelia van Duijn
Netherlands Institute for Health Sciences
Dept. of Genetic Epimiology, Erasmus MC

Funding

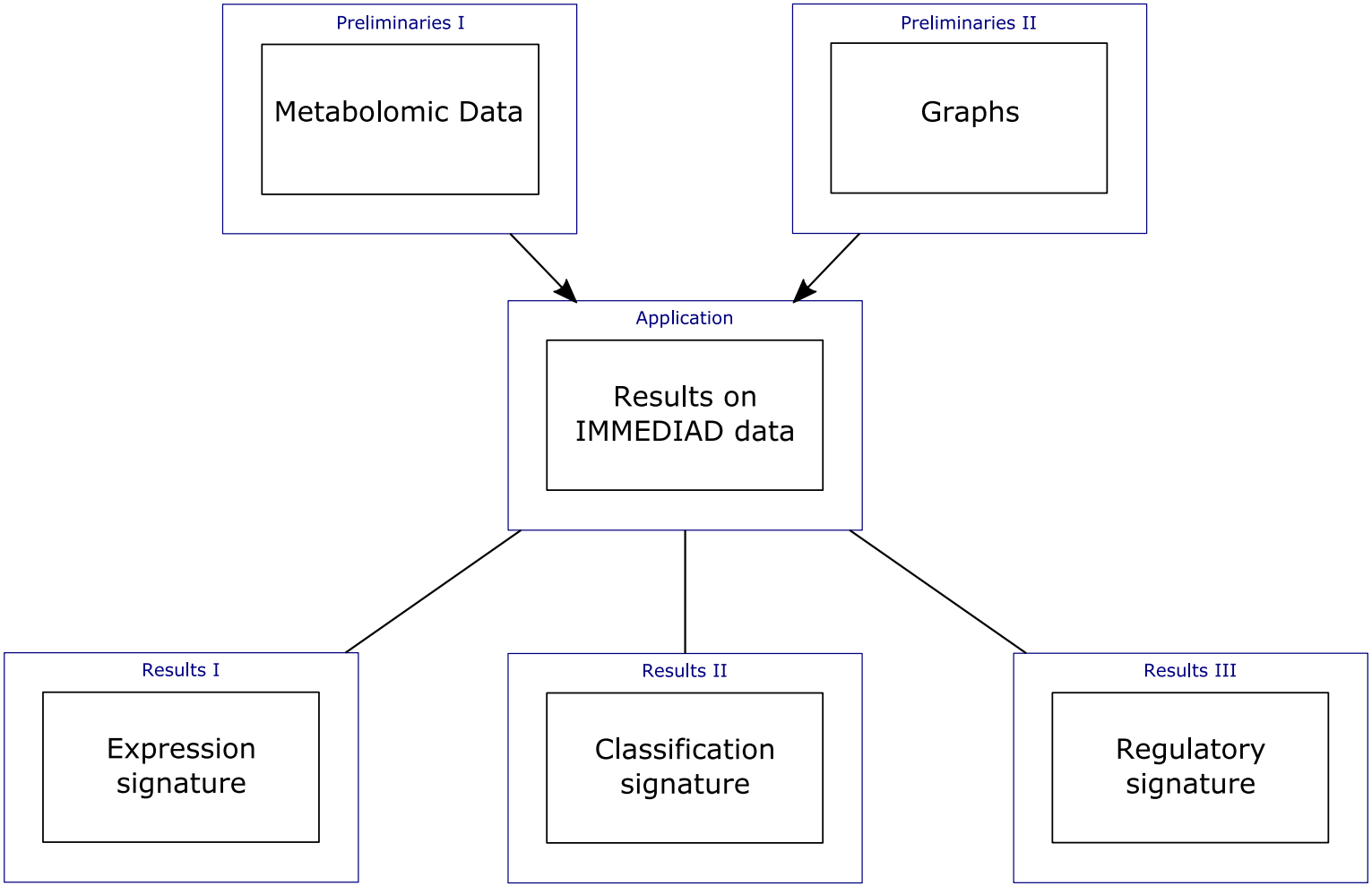
Methodological Developments

European Community Seventh Framework Programme (FP7): grant agreement No. FP7-269553

Data and Analyzes

Janssen Pharmaceutica Stellar funded project (IMMEDIAD): Stellar Neurodegeneration Collaboration Project, Call 2, No. 3

Outline


Preliminaries I
Metabolomic Data
Preliminaries II
Graphs
Application
Results on IMMEDIAD data
Results I
Expression signature
Results II
Classification signature
Results III
Regulatory signature

Omics and omics data

-ome

A totality of some (molecular biological) sort

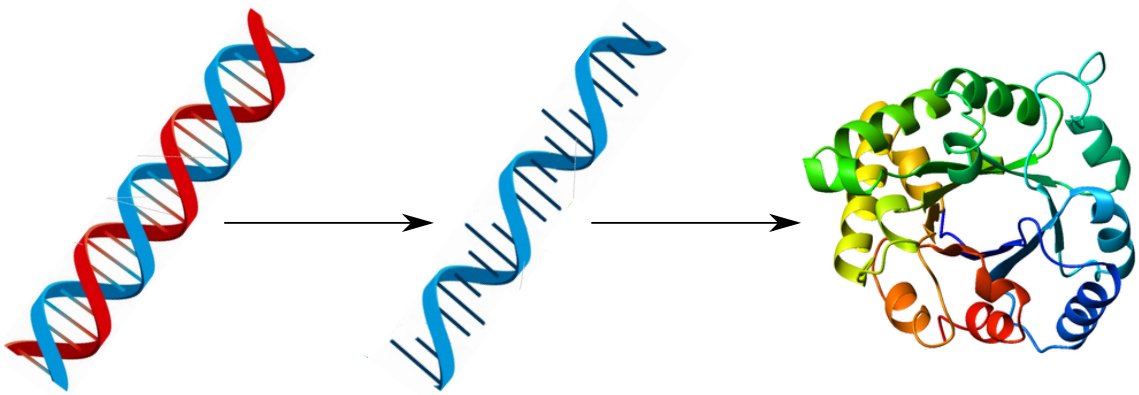
-omics

Collective quantification of some pool of molecular molecules

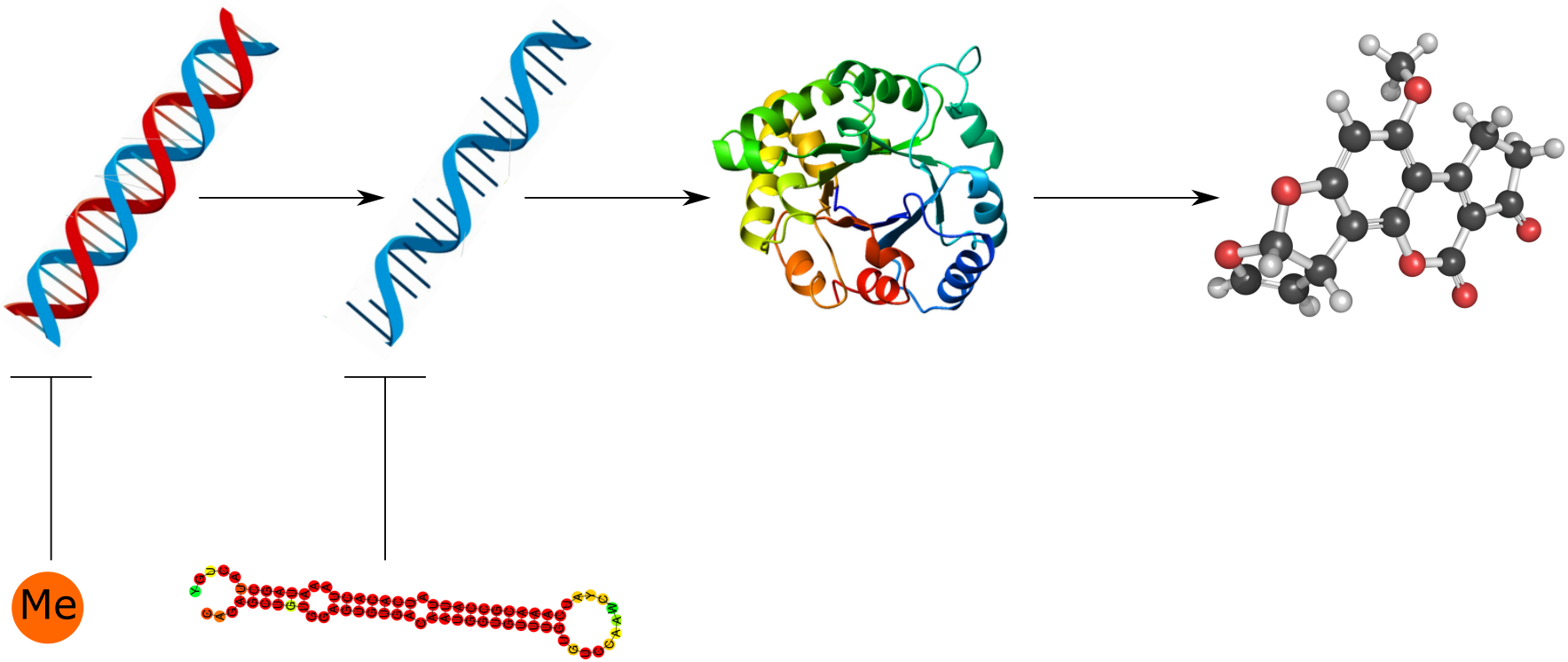
Genomics

The omics of the genome (of some organism)

The omic cascade



The omic cascade



Metabolite quantification

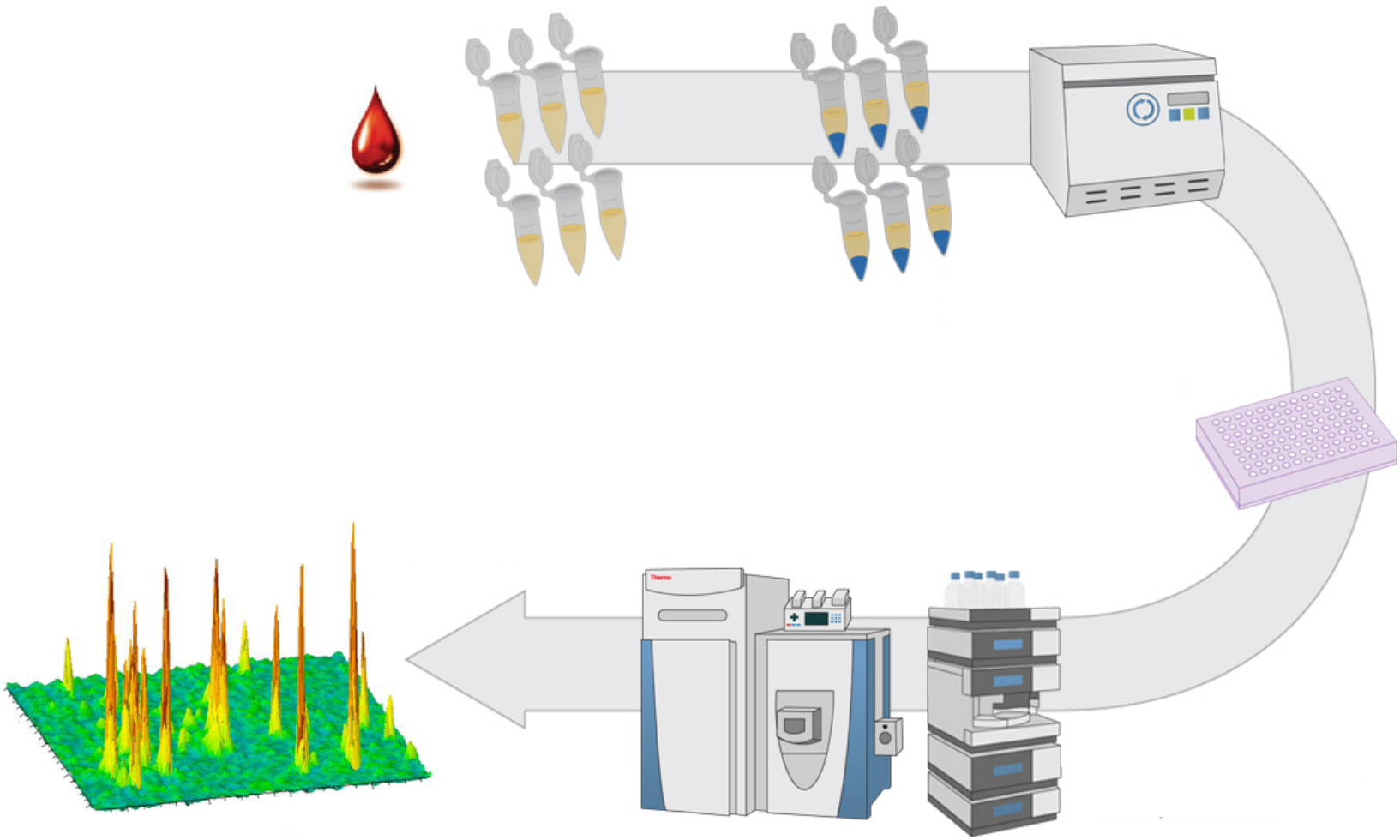
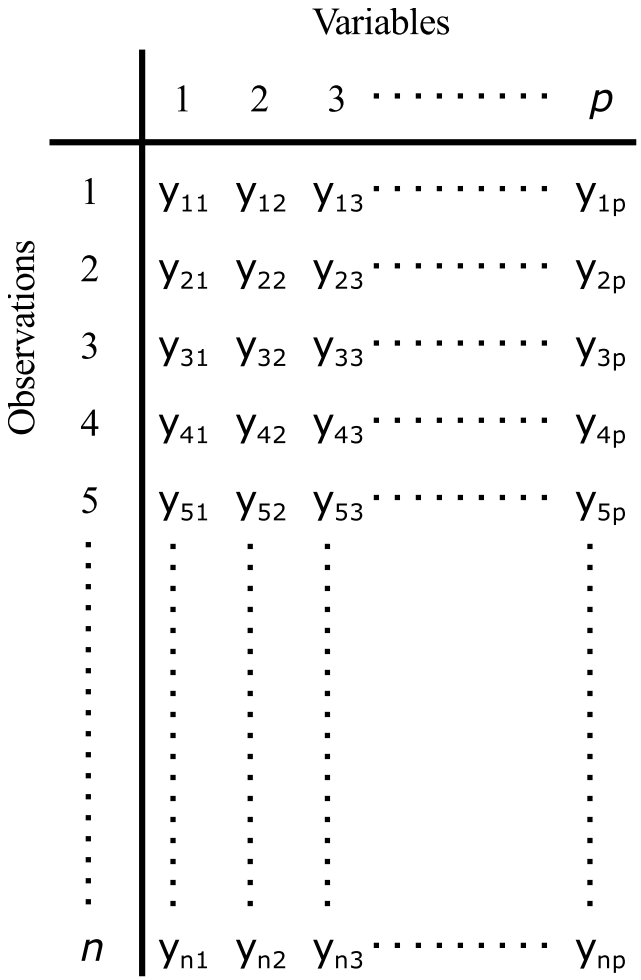


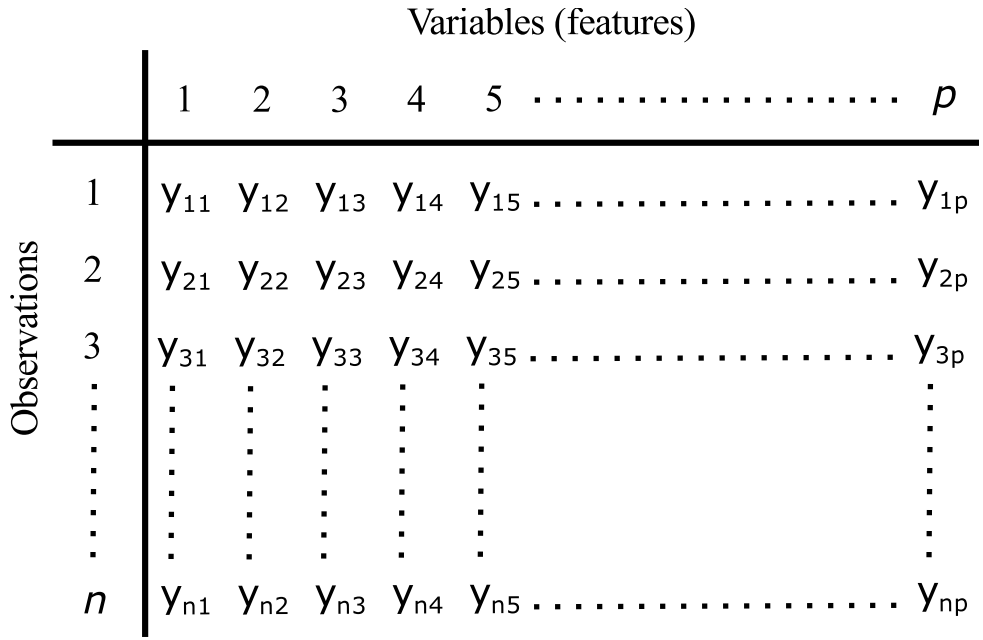
Illustration adapted from: http://planetorbitrap.com/untargeted-metabolomics#.Vzw6yfmLRaQ &

http://metabolomicsplatform.com/metabolomics-overview/

Challenge: Dimensionality metabolomic data

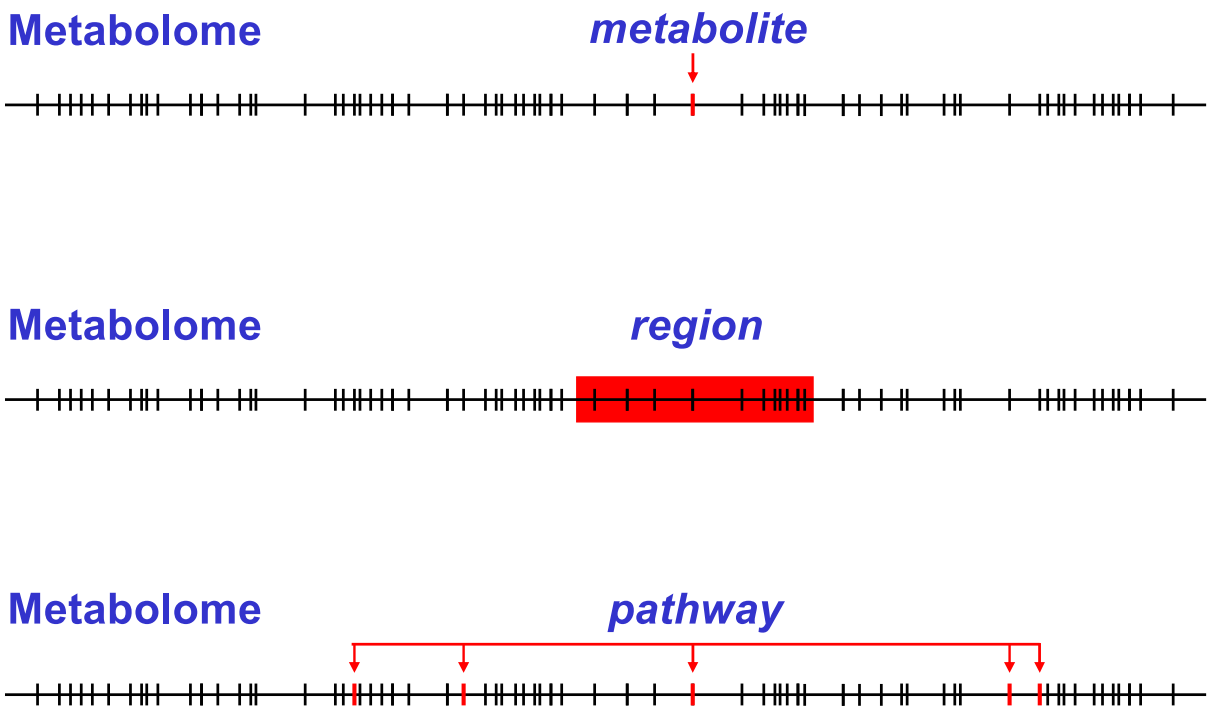


Regular data: $n > p$



Metabolomic data: $p > n$ or $p >> n$

Unit of analysis



Graphs

Representation

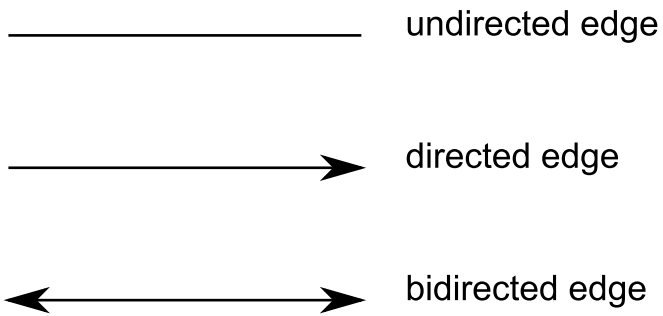
Pathways are represented by a *graph* (or *network*)

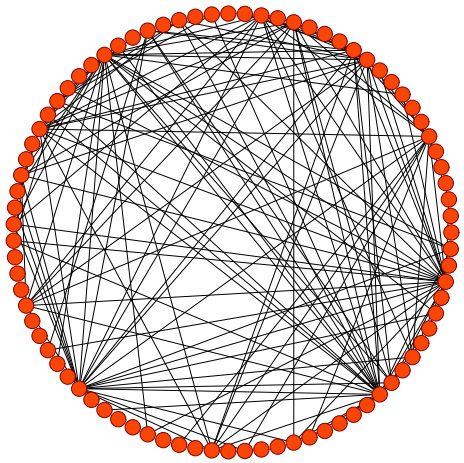
Vertices

- *Node* or *vertex* represents molecular feature

Edges

Edge or *arrow* represents some functional relation





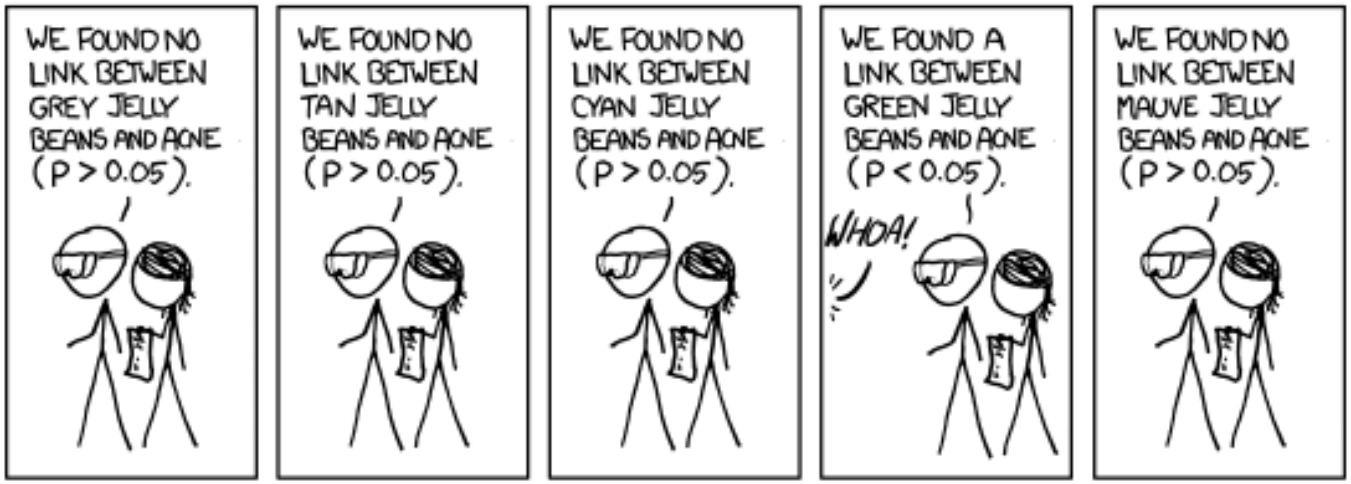
Nested models

Nested models

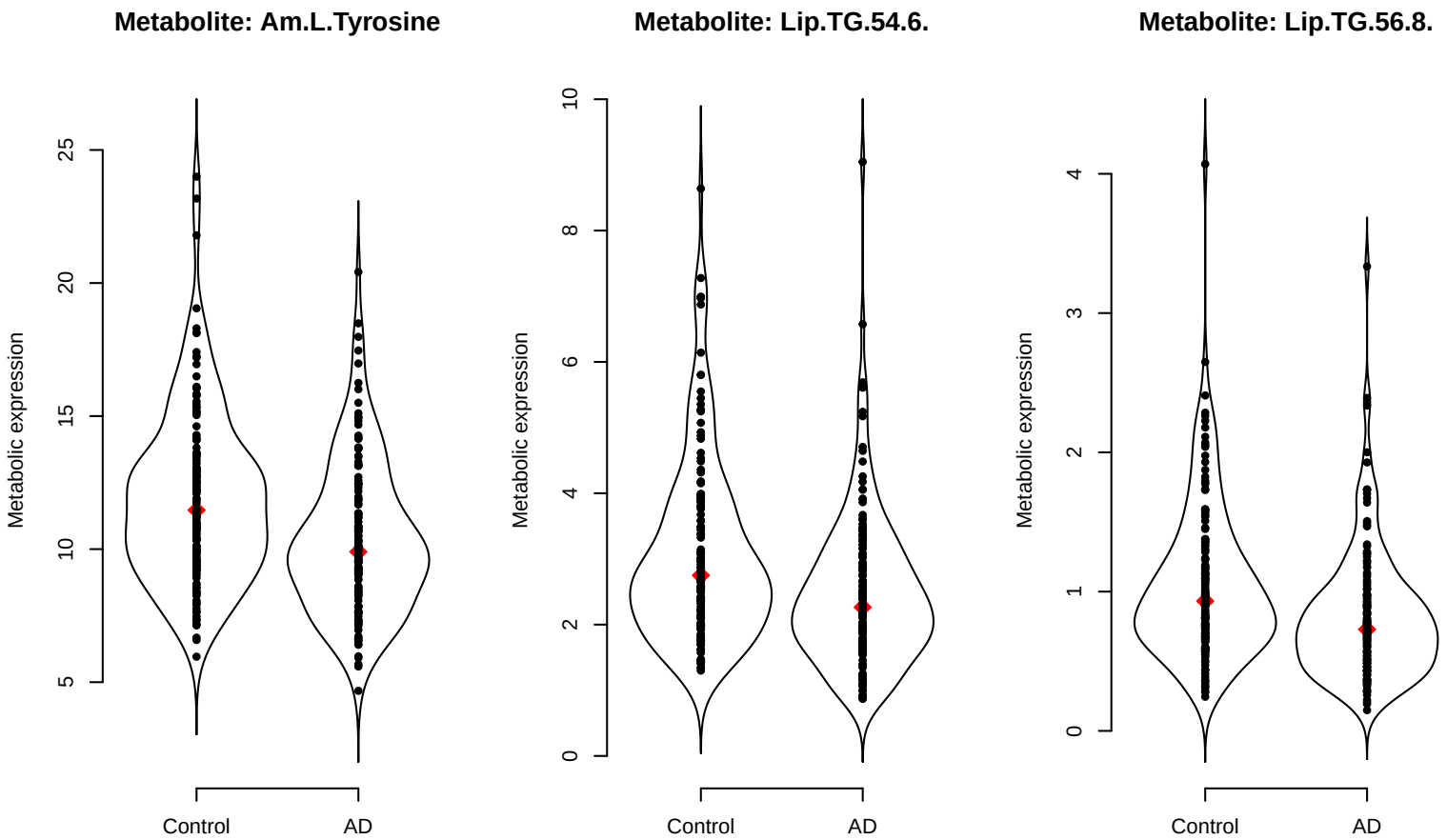
$$\text{Metabolite expression} = \beta_0 + \beta_1 \text{SBP} + \beta_2 \text{ApoE} + \epsilon$$

$$\text{Metabolite expression} = \beta_0 + \beta_1 \text{SBP} + \beta_2 \text{ApoE} + \beta_3 \text{AD} + \epsilon$$

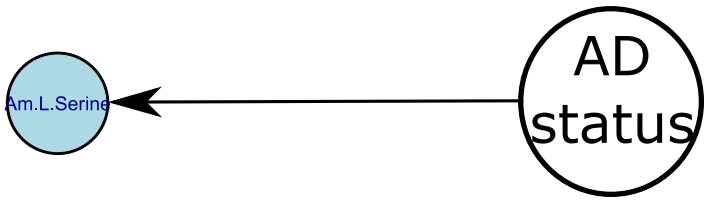
Multiple testing

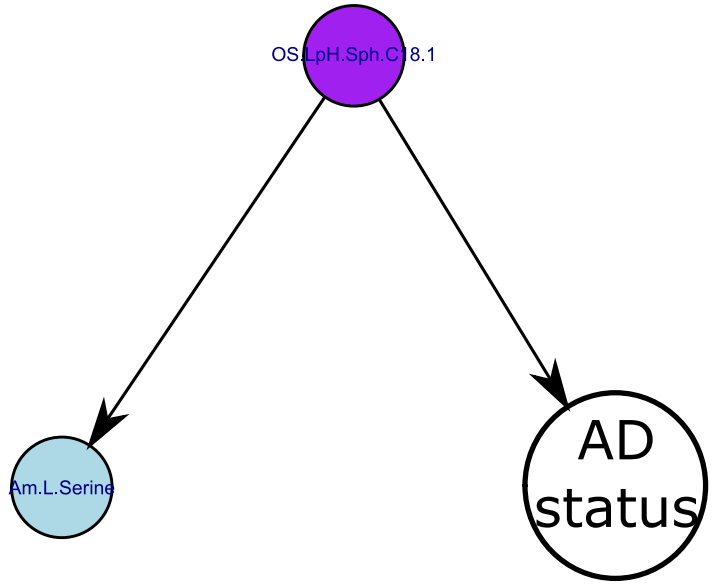


Results

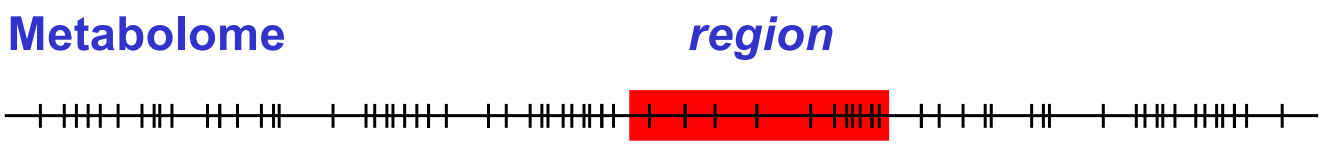

Metabolite: Am.L.Tyrosine
Metabolic expression
5
10
15
20
25
Control
AD
Metabolite: Lip.TG.54.6.
Metabolic expression
0
2
4
6
8
10
Control
AD
Metabolite: Lip.TG.56.8.
Metabolic expression
0
1
2
3
4
Control
AD

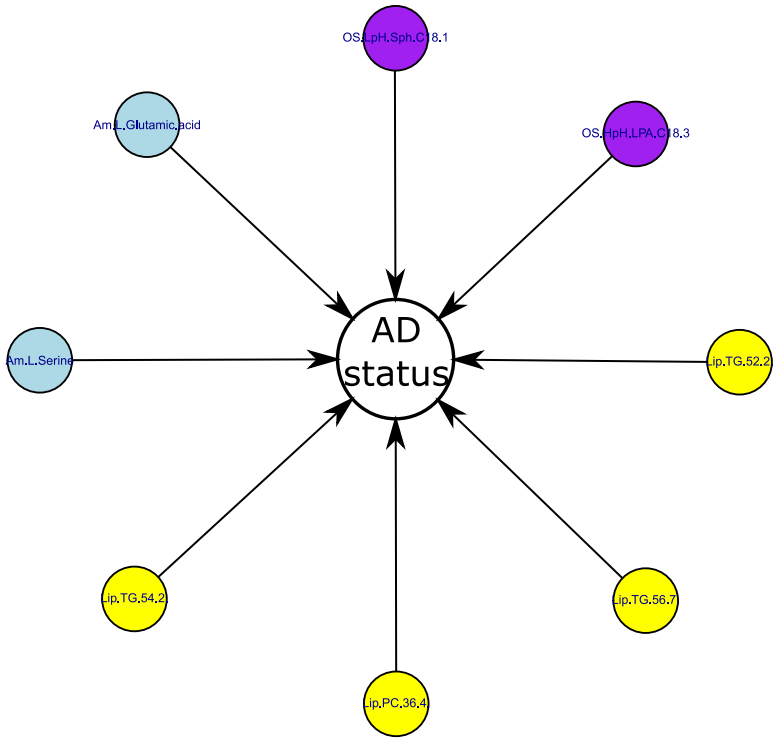
Not the complete picture



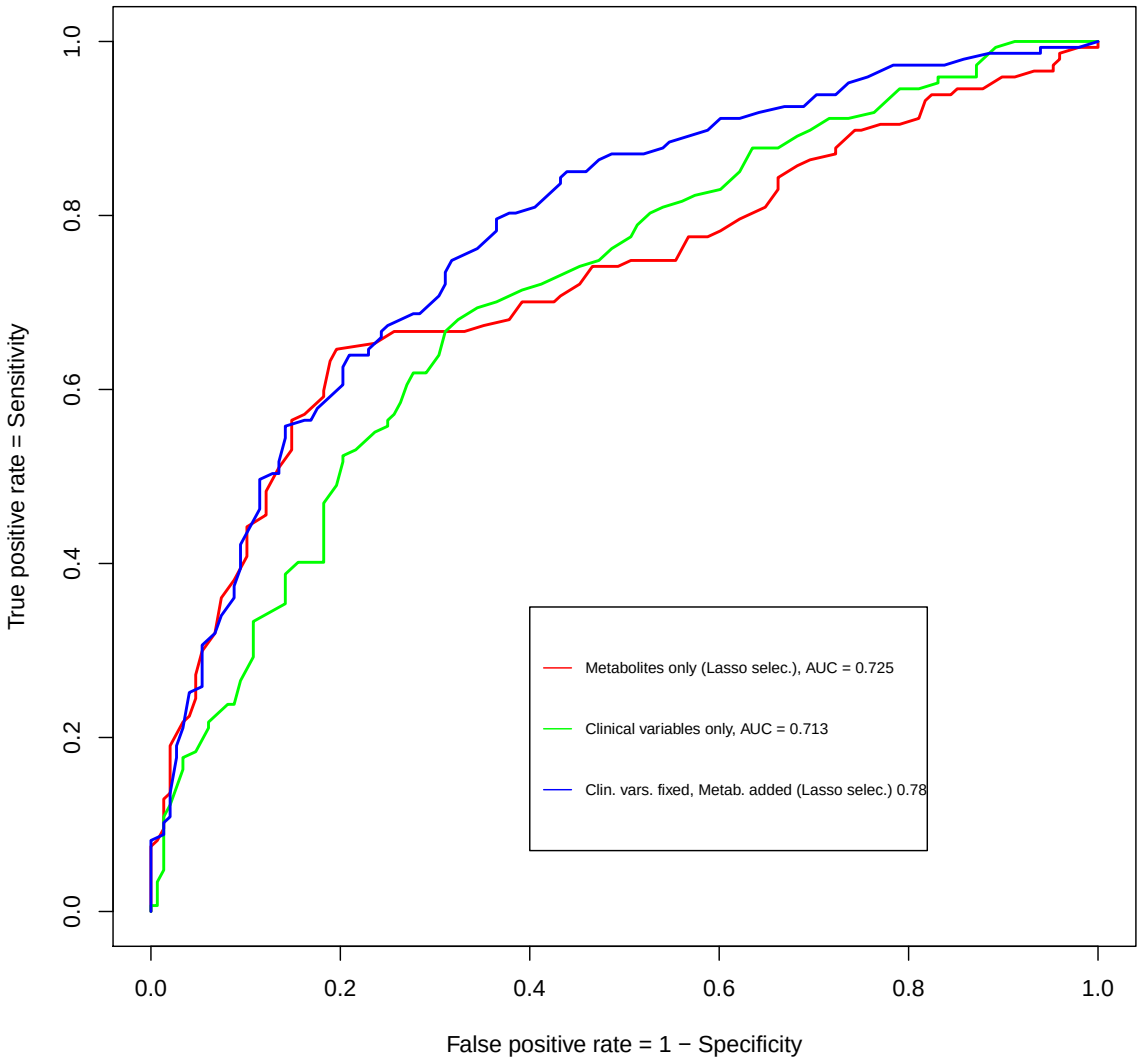


The prediction model

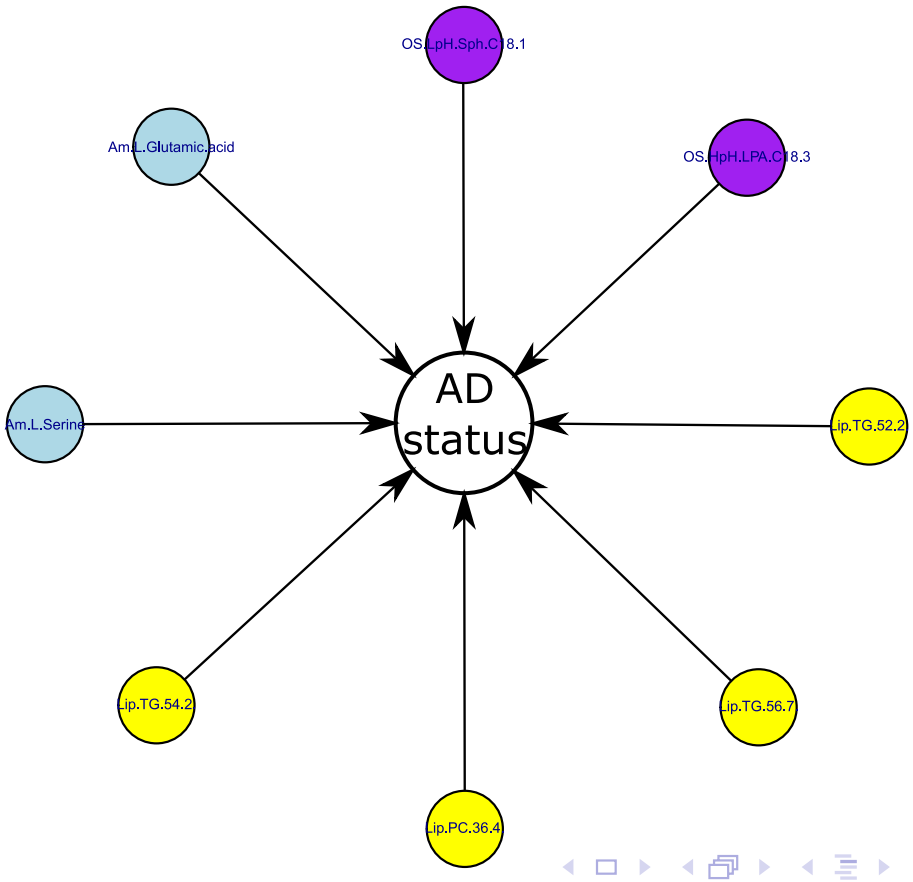




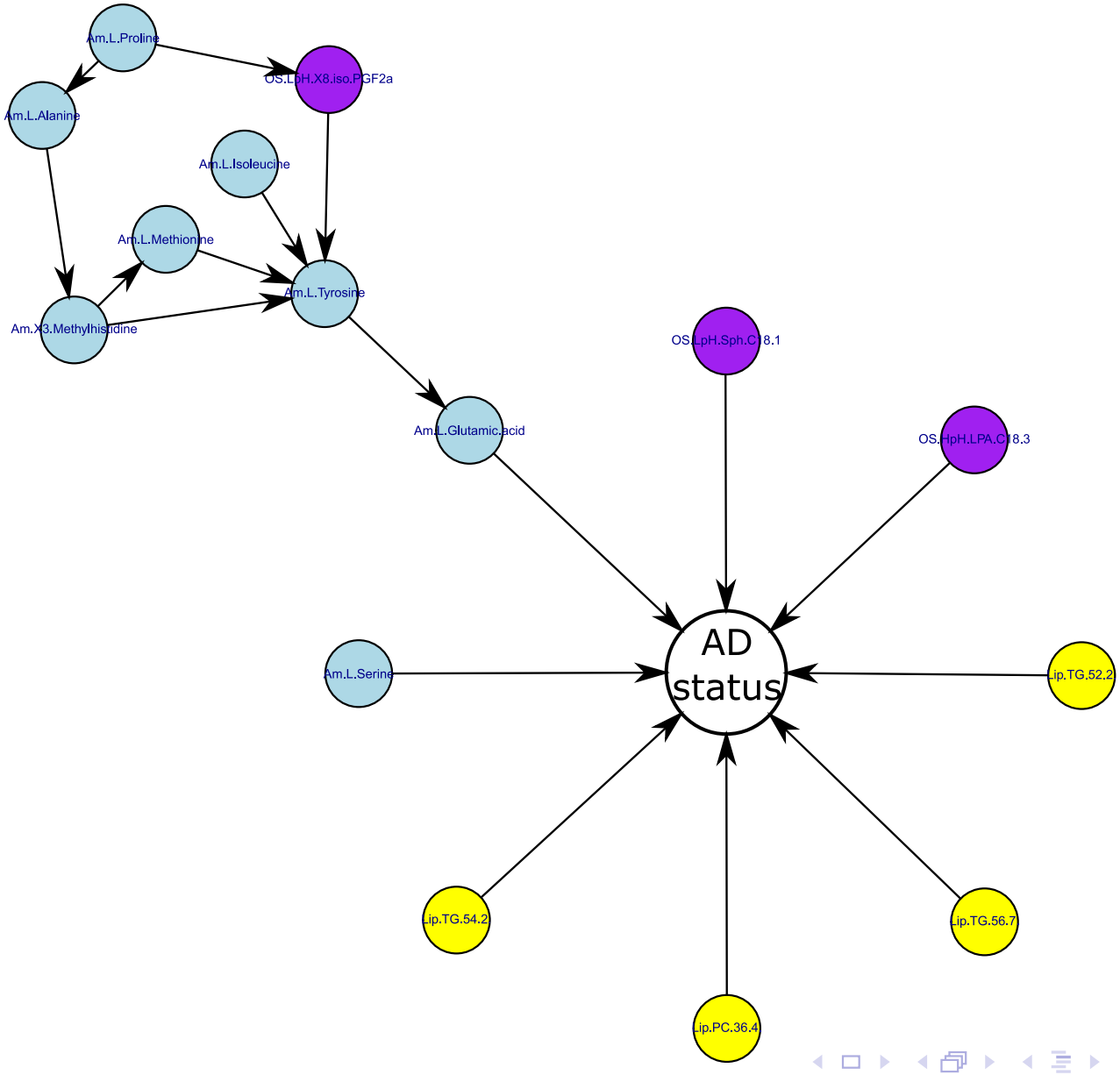
Results


True positive rate = Sensitivity
False positive rate = 1 − Specificity
0.0
0.2
0.4
0.6
0.8
1.0
Metabolites only (Lasso selec.), AUC = 0.725
Clinical variables only, AUC = 0.713
Clin. vars. fixed, Metab. added (Lasso selec.) 0.78

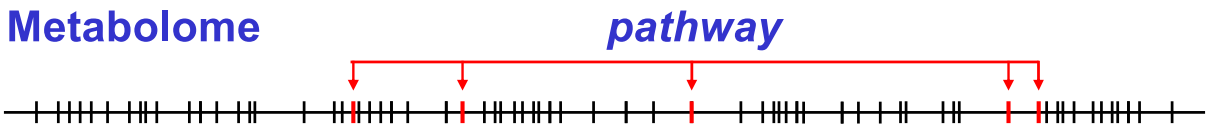
Not the complete picture

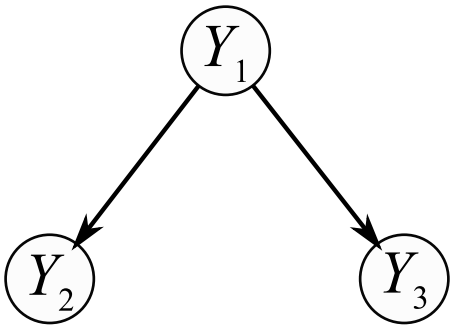


Not the complete picture

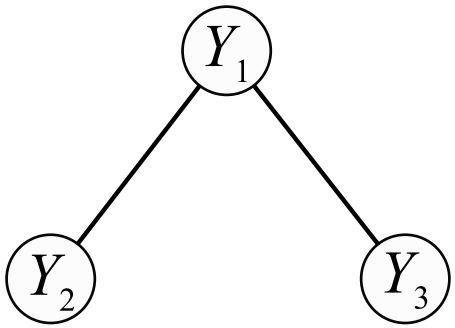


Graphical modeling

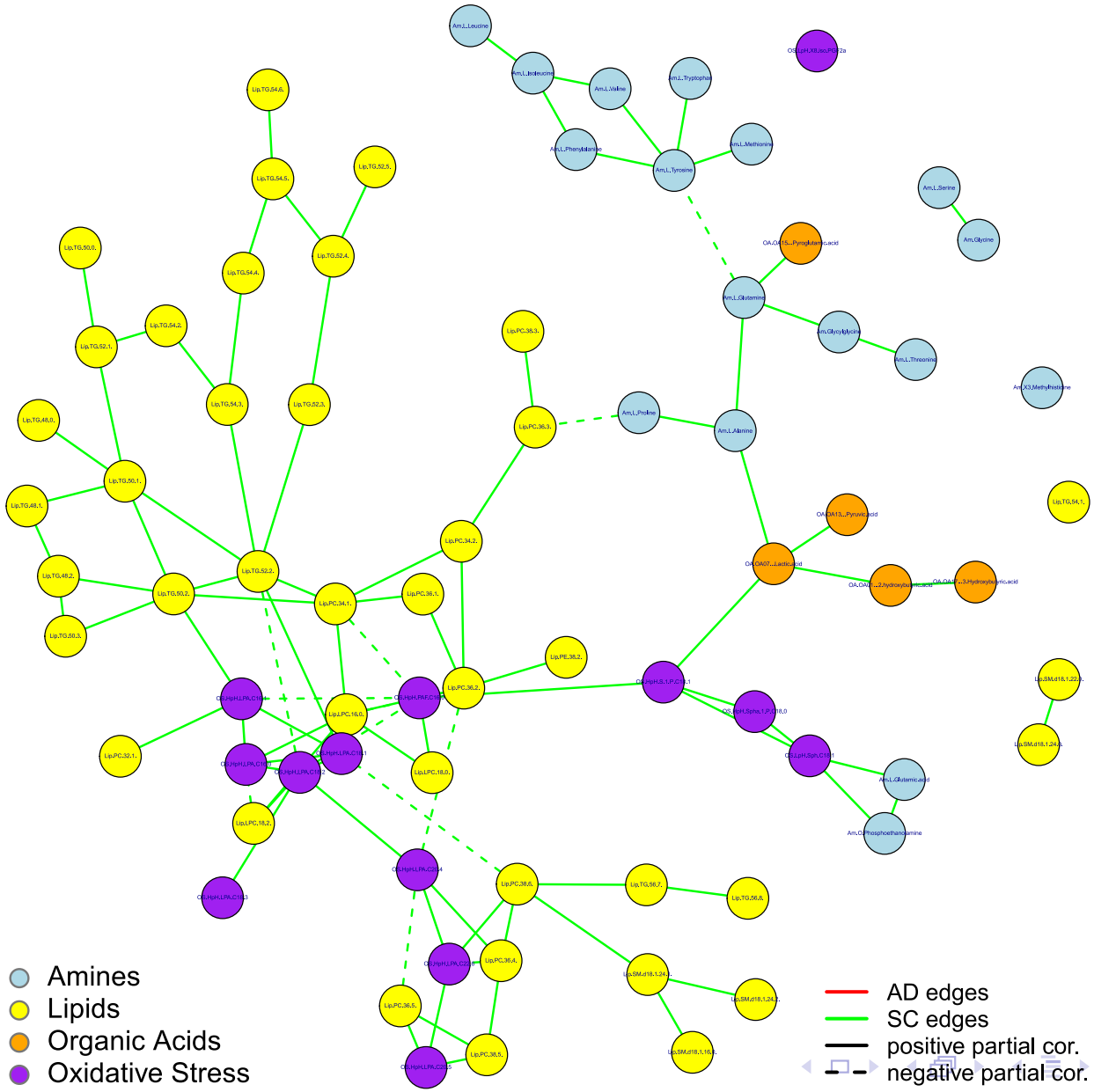




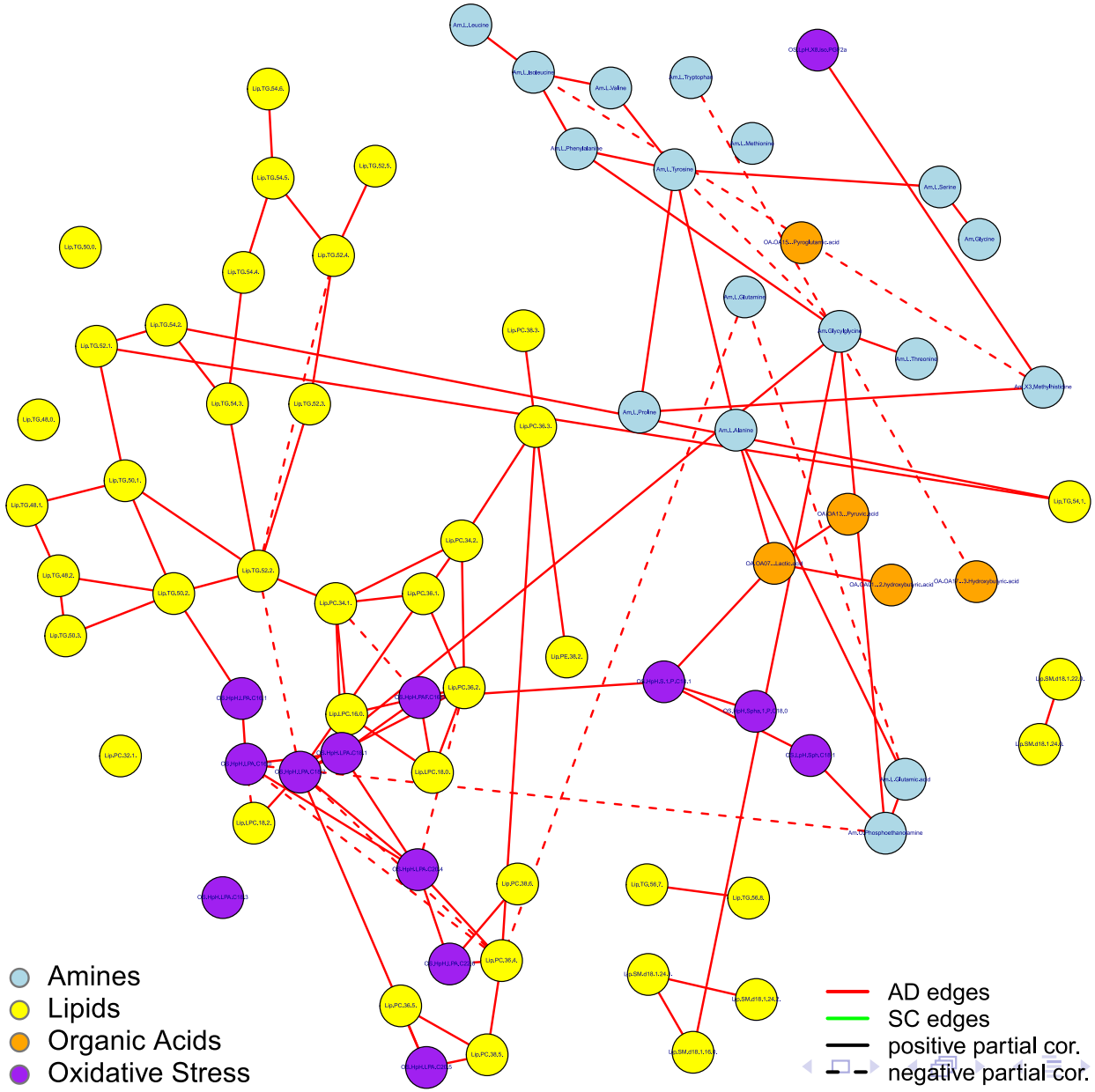
$$\text{cor}(Y_1, Y_2 | Y_3) \neq 0$$
$$\text{cor}(Y_1, Y_3 | Y_2) \neq 0$$
$$\text{cor}(Y_2, Y_3 | Y_1) = 0$$



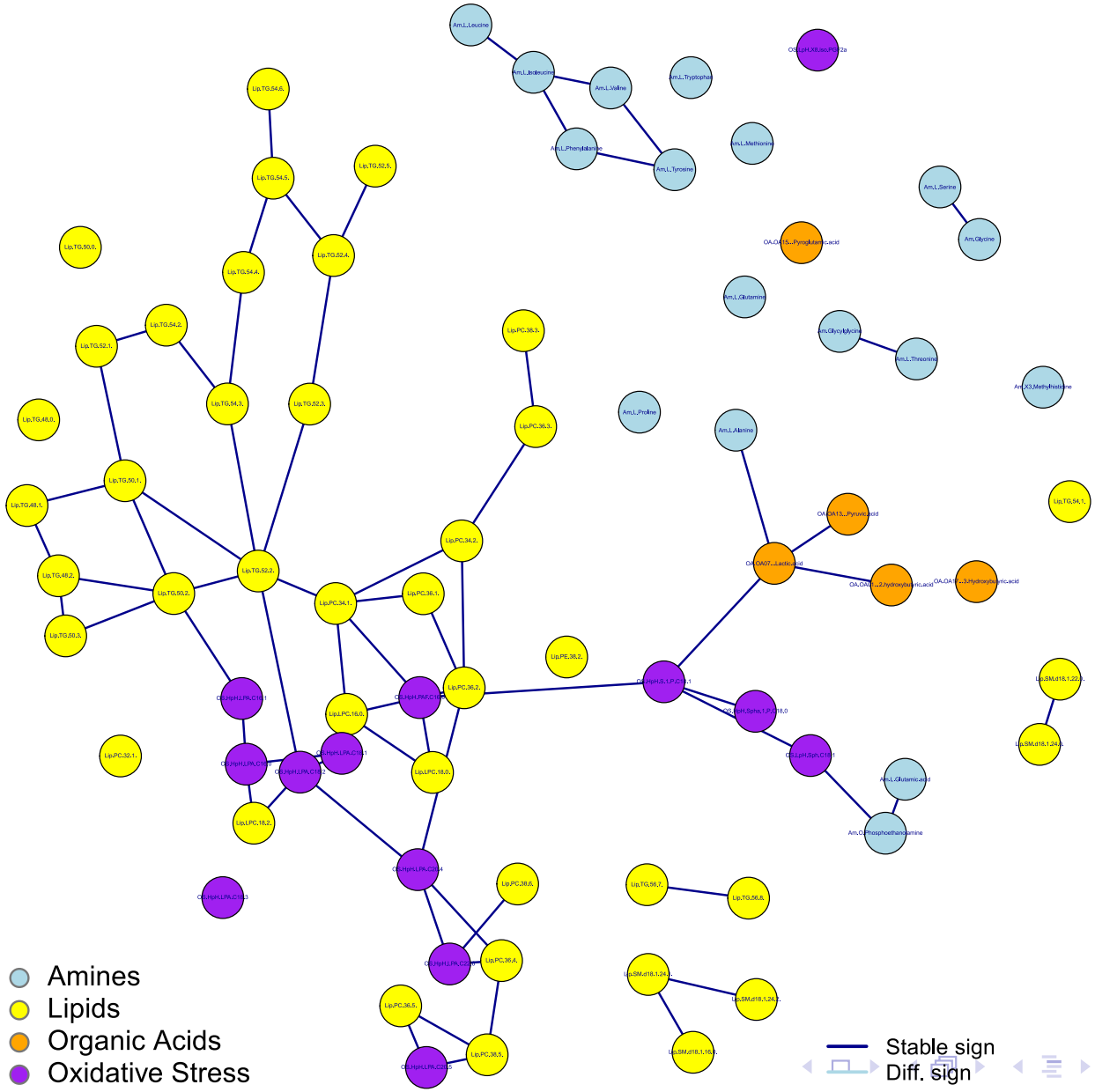
Control connections



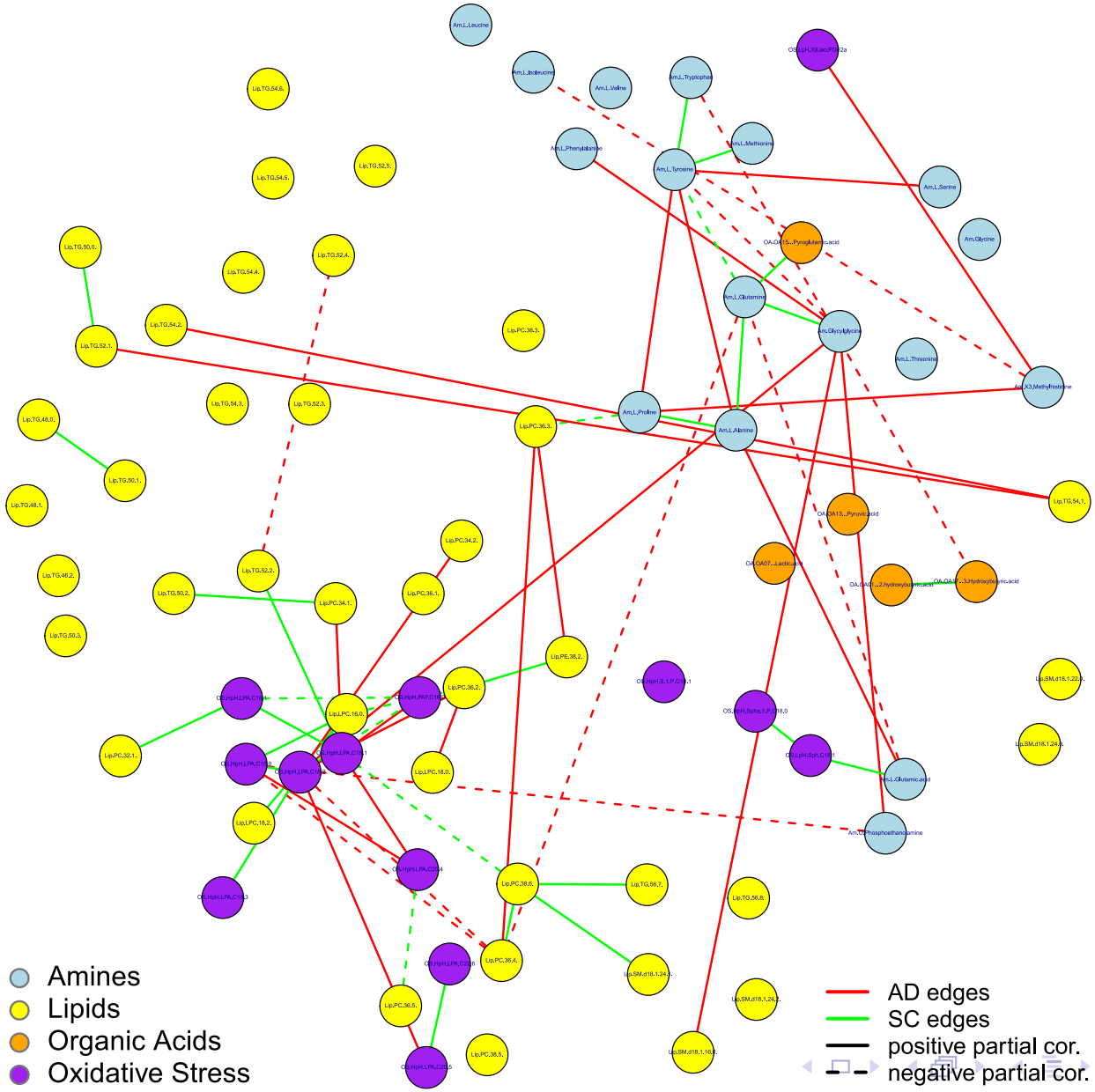
AD connections



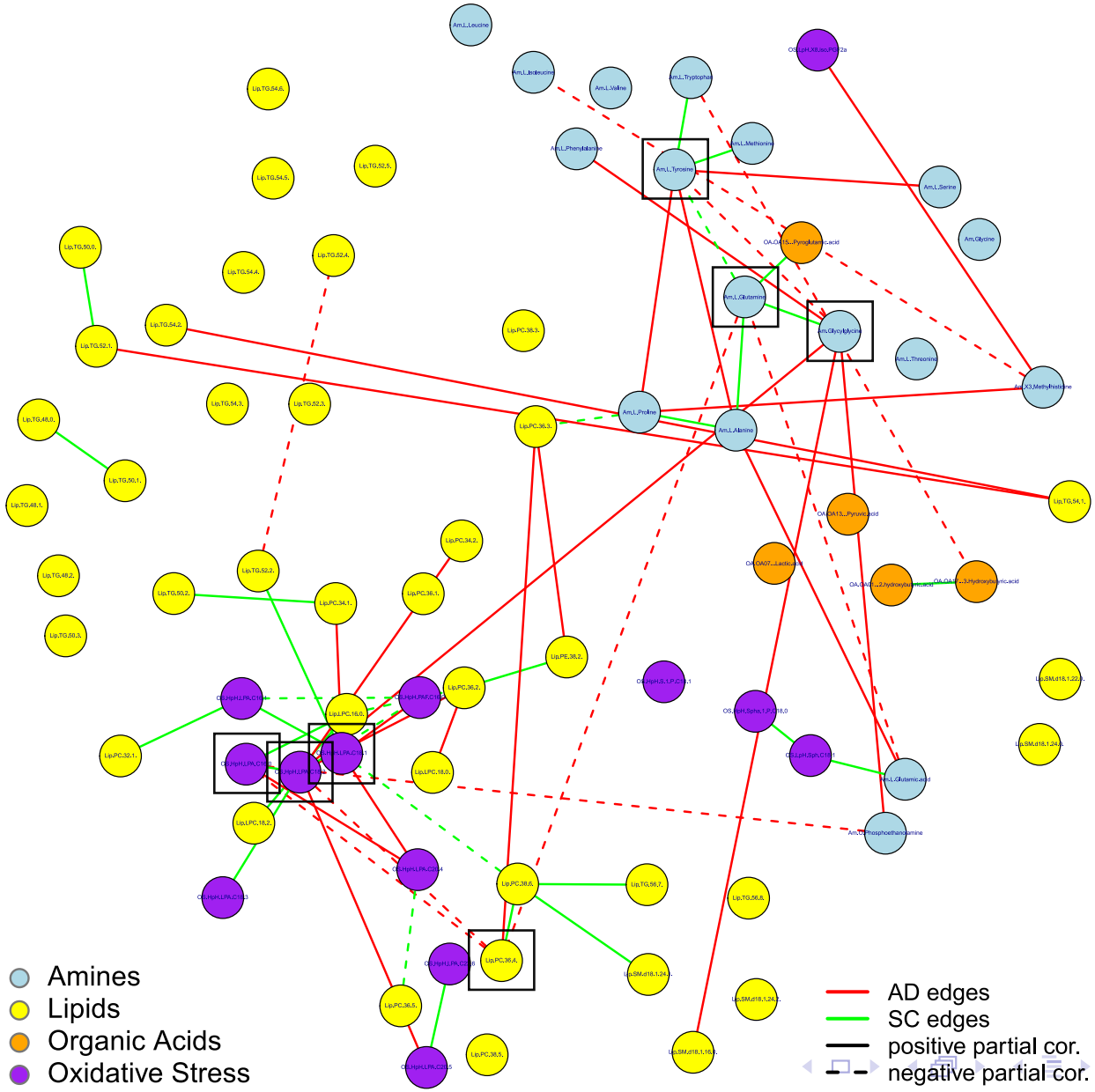
Shared connections



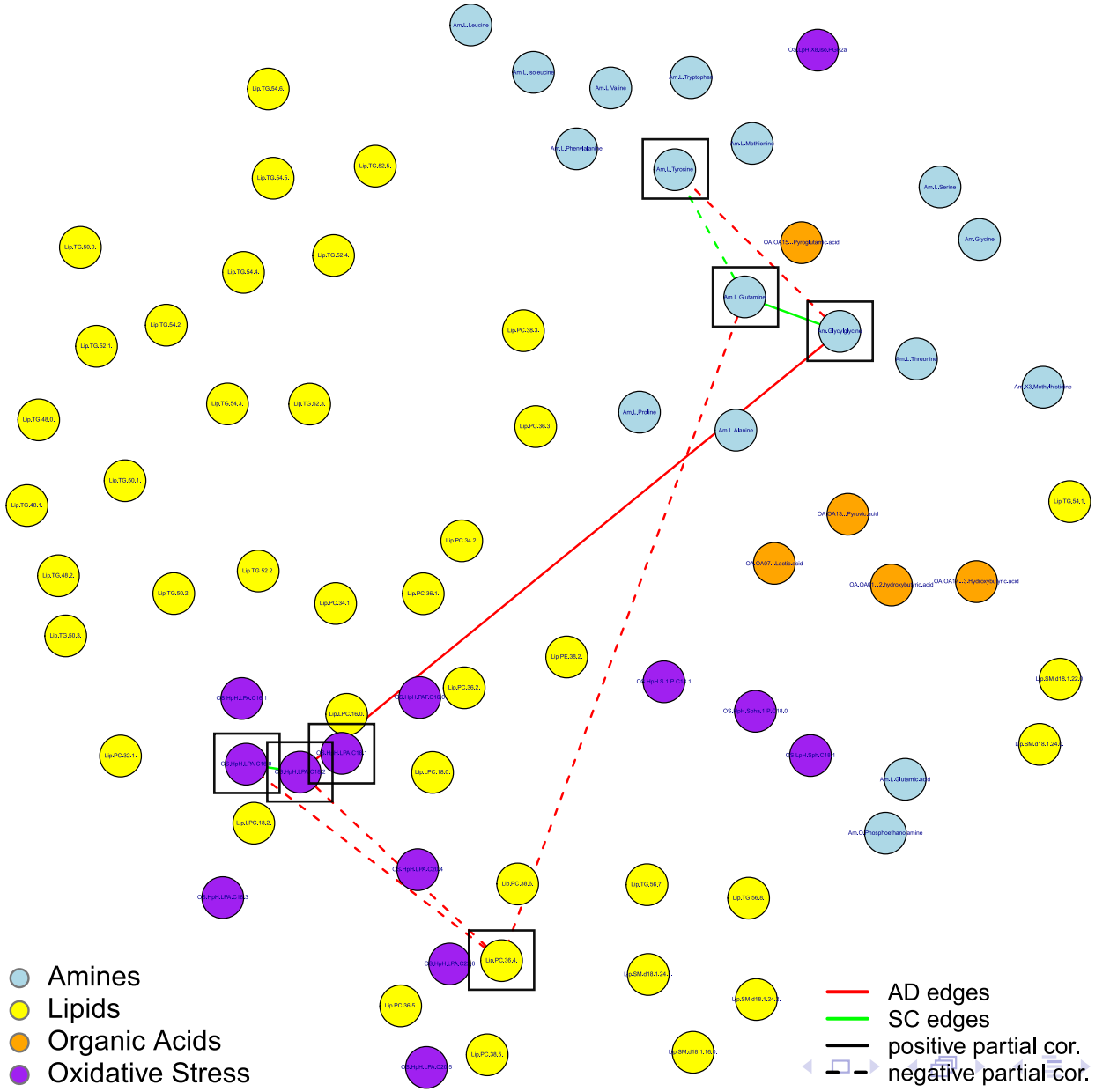
Differential connections



Differential connections



Differential connections



Concluding

Messages

- 3 metabolic signatures
- Each signature comes with different (complementary) information
- Tyrosine influential in all signatures

Now what?

- Follow-up studies
- Validation of findings
- Translational efforts

Manual/Download

- Peeters, C.F.W., Bilgrau, A.E., & van Wieringen, W.N. (2016). "`rags2ridges`: Ridge Estimation of Precision Matrices from High-Dimensional Data". R package, version 2.1.1 URL: https://cran.r-project.org/package=rags2ridges.

Theory/Methodology

- Peeters*, C.F.W., Bilgrau*, A.E., Eriksen, P.S., Boegsted, M., & van Wieringen, W.N. (2015). "Targeted Fused Ridge Estimation of Inverse Covariance Matrices from Multiple High-Dimensional Data Classes". arXiv:1509.07982v1 [stat.ME].
- Peeters, C.F.W., van Wieringen, W.N., & van de Wiel, M.A. (in preparation). "Directed Cyclic Mixed Graph Modeling for High-Dimensional Genomic Data Integration".
- van Wieringen, W.N. & Peeters, C.F.W. (2016). "Ridge Estimation of Inverse Covariance Matrices from High-Dimensional Data". Computational Statistics & Data Analysis, 103: 284-303. arXiv:1403.0904v3 [stat.ME].

Software

- Peeters, C.F.W., van de Wiel, M.A., & van Wieringen, W.N. (2016) "The Spectral Condition Number Plot for Regularization Parameter Determination". arXiv:1608.04123v1 [stat.CO].
- van Wieringen, W.N. & Peeters, C.F.W. (2015). "Application of a New Ridge Estimator of the Inverse Covariance Matrix to the Reconstruction of Gene-Gene Interaction Networks". In: di Serio, C., Lio, P., Nonis, A., and Tagliaferri, R. (Eds.) 'Computational Intelligence Methods for Bioinformatics and Biostatistics'. Lecture Notes in Computer Science, vol. 8623. Springer, pp. 170–179.

Table: List of clinical confounders/variables

Anthropometric:
Age
Sex
ApoE $\epsilon 4$ allele status (at least one allele $\epsilon 4$ yes,no)
Systolic blood pressure
Diastolic blood pressure
Height
Weight
Smoking (yes, no, quit)
Alcohol (yes, no)

Comorbidities (binary):
Hypertension
Diabetes Mellitus
Hypercholesterolemia

Medication use (binary):
Cholesterol lowering medications
Antidepressant medications
Antiplatelet medications
Antidiabetic medication
Antiepileptic medication
Antiparkinson medication
Antipsychotics

Table: Metabolites selected for prediction model

Am.X1.Methylhistidine
Am.X3.Methylhistidine
Am.Citrulline
Am.Cysteine
Am.gamma.L.glutamyl.L.alanine
Am.Histamine
Am.L.carnosine
Am.L.Tyrosine
Am.Methyldopa
Am.O.acetyl.L.serine
Am.Putrescine
OA.OA14Methylmalonic.acid
OA.OA173.Hydroxybutyric.acid
OA.OA263.Hydroxyisobutyric.acid
OA.OA273.hydroxyisovaleric.acid
OA.OA28Glyceric.acid
Lip.TG.51.3.
Lip.TG.56.8.
Lip.TG.58.10.
Lip.LPC.18.1.
Lip.LPC.20.3.
Lip.LPC.20.4.
Lip.PC.O.34.3.
Lip.PC.O.36.5.
Lip.PC.O.38.6.
Lip.SM.d18.1.18.2.
Lip.SM.d18.1.20.1.
Lip.SM.d18.1.23.0.
OS.LpH.X8.12.iPF2a.IV
OS.LpH.X..5.iPF2a.VI
OS.LpH.NO2.aLA
OS.LpH.NO2.OA
OS.LpH.PGD2
OS.LpH.Spha.C18.0
OS.cLpH.PGA2
OS.cLpH.X2.3.dinor.8.iso.PGF2a
OS.HpH.LPA.C14.0

Explaining the inverse

The scalar inverse

- Let a denote a number (excluding 0)
- The inverse is then the number b such that $a \times b = 1$
- Clearly, $b = \frac{1}{a}$

Matrix

A matrix is a generalization of a number, an array of numbers

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pp} \end{bmatrix}$$

Explaining the inverse

The Matrix Inverse

Consider the matrix $\mathbf{A}$. Its inverse $\mathbf{B} = \mathbf{A}^{-1}$ is defined such that

$$\mathbf{AB} = \mathbf{I},$$

where

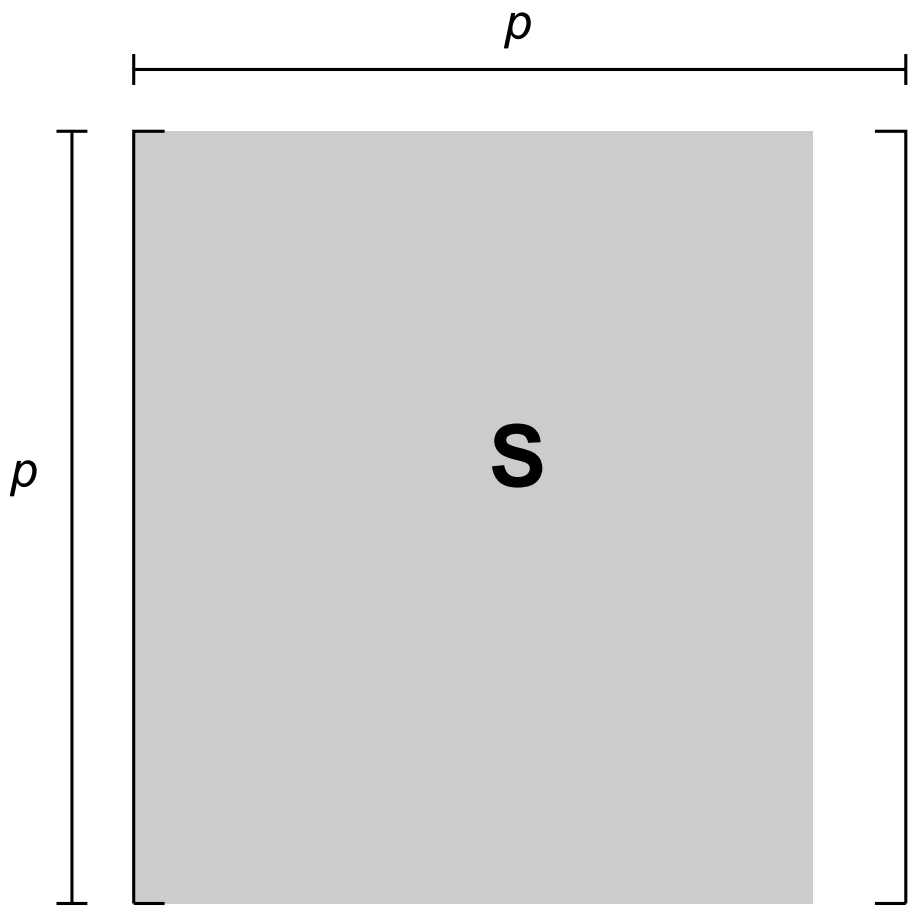
$$\mathbf{I} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

Solution

$$\mathbf{A}^{-1} = \begin{bmatrix} \mathbf{A}_{11}^{-1} + \mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{Q}^{-1}\mathbf{A}_{21}\mathbf{A}_{11}^{-1} & -\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{Q}^{-1} \\ -\mathbf{Q}^{-1}\mathbf{A}_{21}\mathbf{A}_{11}^{-1} & \mathbf{Q}^{-1} \end{bmatrix},$$

with $\mathbf{Q} = \mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}$ denoting the Schur complement.

Singularity



Ridge estimation

Maximize

$$\underbrace{\ln|\boldsymbol{\Omega}| - \text{tr}(\mathbf{S}\boldsymbol{\Omega})}_{\log-\text{likelihood}} - \underbrace{\frac{\lambda}{2}\|\boldsymbol{\Omega} - \mathbf{T}\|_2^2}_{\ell_2-\text{penalty}}$$

- $\mathbf{T}$ denotes a p.d. symmetric target matrix
- $\lambda \in (0, \infty)$ denotes a penalty parameter

Analytic penalized ML estimator

$$\hat{\boldsymbol{\Omega}}(\lambda) = \left\{ \left[\lambda \mathbf{I}_p + \frac{1}{4}(\mathbf{S} - \lambda\mathbf{T})^2 \right]^{1/2} + \frac{1}{2}(\mathbf{S} - \lambda\mathbf{T}) \right\}^{-1}$$

Properties

Behavior

i. $\hat{\boldsymbol{\Omega}}(\lambda) \succ 0$, for all $\lambda \in (0, \infty)$;

ii. $\lim_{\lambda \to 0^+} \hat{\boldsymbol{\Omega}}(\lambda) = \mathbf{S}^{-1}$ if $p < n$;

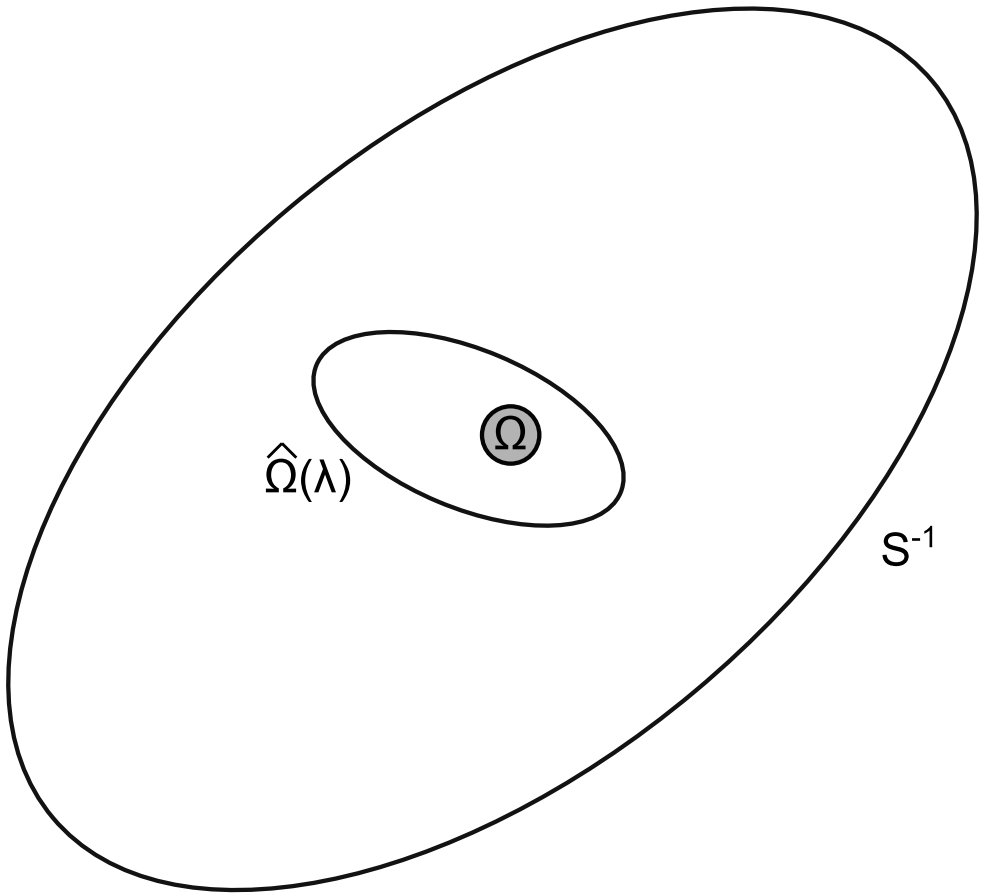
iii. $\lim_{\lambda \to \infty} \hat{\boldsymbol{\Omega}}(\lambda) = \mathbf{T}$.

Consistency

i. $\lim_{n \to \infty} \mathbb{E}\left[\hat{\boldsymbol{\Omega}}_n(\lambda_n)\right] \longrightarrow \lim_{n \to \infty} \mathbb{E}\left(\mathbf{S}_n^{-1}\right) = \boldsymbol{\Omega}$;

ii. $\lim_{n \to \infty} \mathbb{E}\left(\|\hat{\boldsymbol{\Omega}}_n(\lambda_n) - \boldsymbol{\Omega}\|_F^2\right) = 0$.

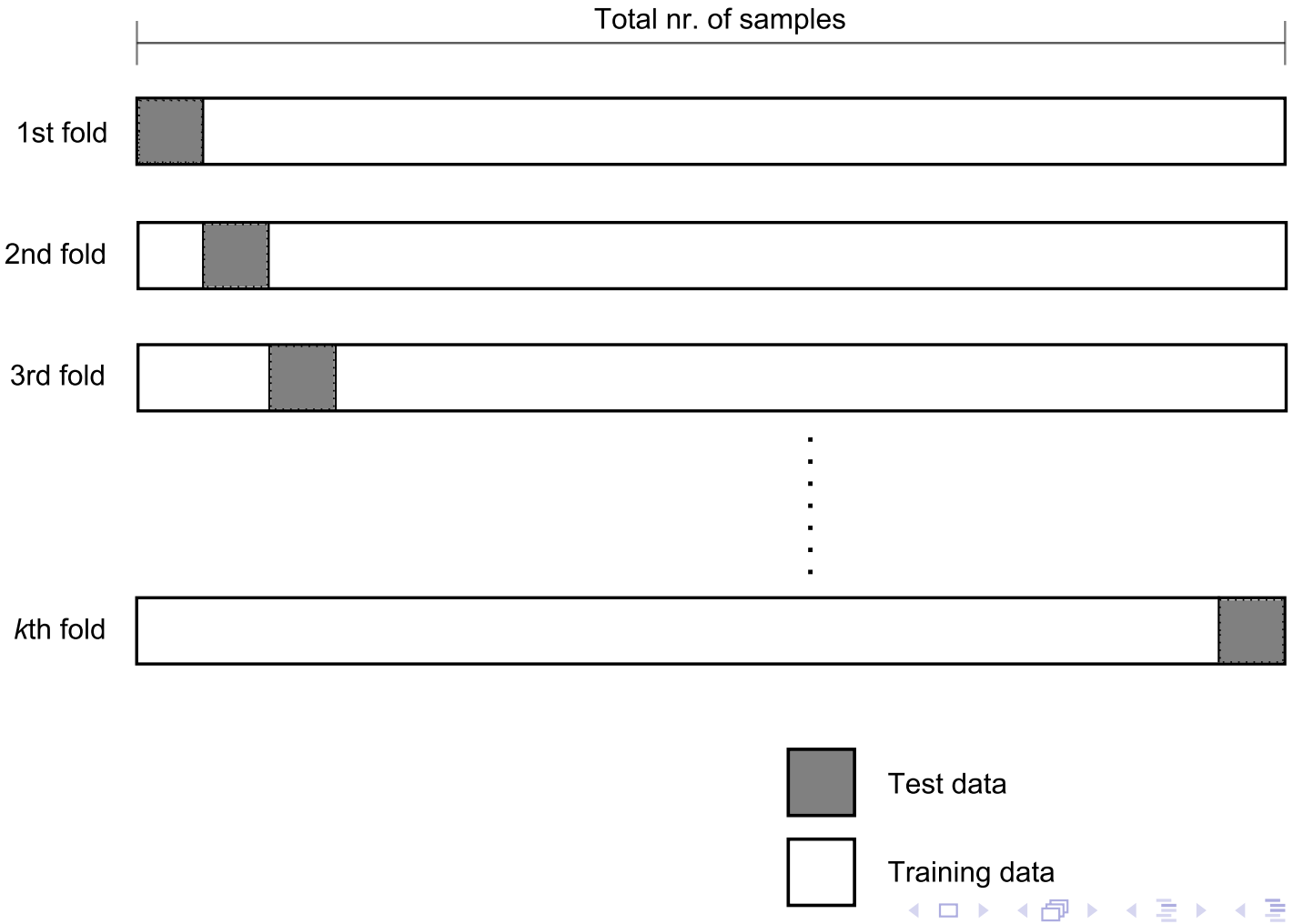
Visual explanation



Choosing the penalty value

K-fold cross-validation (CV)

Single iteration of *K*-fold CV



Choosing the penalty value

K-fold CV score

$$\varphi^K(\lambda) = \sum_{k=1}^{K} n_k \left\{ -\ln |\hat{\boldsymbol{\Omega}}(\lambda)_{-k}| + \text{tr}[\hat{\boldsymbol{\Omega}}(\lambda)_{-k}\mathbf{S}_k] \right\},$$

n_k is the size of subset k, for $k = 1, \ldots, K$ disjoint subsets;
$\mathbf{S}_k$ denotes the sample covariance matrix on kth test set;
$\hat{\boldsymbol{\Omega}}(\lambda)_{-k}$ denotes the estimated regularized precision matrix on kth training set

Highest predictive accuracy

Choose $n_k = 1$, such that $K = n$ (known as leave-one-out CV - LOOCV)

Problem

K-fold CV is computationally demanding for large p and/or large K

Solution

Computationally efficient approximate LOOCV score

Support determination

Scaling

$\hat{\mathbf{P}}(\lambda)$: Regularized precision estimate scaled to partial correlation form

Assume

Nonredundant off-diagonal partial correlation coefficients (indexed by $j < j'$) follow a mixture distribution:

$$f\left\{[\hat{\mathbf{P}}(\lambda^*)]_{jj'}\right\} = \eta_0 f_0\left\{[\hat{\mathbf{P}}(\lambda^*)]_{jj'}; \kappa\right\} + (1-\eta_0) f_{\mathcal{E}}\left\{[\hat{\mathbf{P}}(\lambda^*)]_{jj'}\right\}$$

- $\eta_0 \in [0,1]$ is the mixture weight
- $f_0\{\cdot\}$ denotes the distribution of a null-edge
- $f_{\mathcal{E}}\{\cdot\}$ denotes the distribution of a present edge

Determine

$$P\left(Y_j \neq Y_{j'} | [\hat{\mathbf{P}}(\lambda^*)]_{jj'}\right)$$

Situation

Data

- G classes of $(n_g \times p)$-dimensional data
- Classes defined by data sets and/or (subtypes of) diseases

Assumption

Precision matrices of constituent classes chiefly share the same structure but potentially differ in a number of locations of interest

Desire

Integrative or meta-analytic Gaussian graphical modeling

Targeted fused ridge estimation: General Formulation

Maximize

$$\underbrace{\mathcal{L}\left(\{\boldsymbol{\Omega}_g\};\{\mathbf{S}_g\}\right)}_{\text{log-likelihood}} - \sum_g \underbrace{\frac{\lambda_{gg}}{2}\left\|\boldsymbol{\Omega}_g - \mathbf{T}_g\right\|_F^2}_{\text{ridge-penalty}} - \sum_{g_1,g_2} \underbrace{\frac{\lambda_{g_1g_2}}{4}\left\|(\boldsymbol{\Omega}_{g_1} - \mathbf{T}_{g_1}) - (\boldsymbol{\Omega}_{g_2} - \mathbf{T}_{g_2})\right\|_F^2}_{\text{fusion-penalty}}$$

- $\mathbf{T}_g$ indicate class-specific target matrices
- $\lambda_{gg} \in (0, \infty)$ denote class-specific ridge penalty parameters
- $\lambda_{g_1g_2} \in [0, \infty)$ denote pair-specific fusion penalty parameters, $\lambda_{g_1g_2} = \lambda_{g_2g_1}$

Penalty matrix

All penalties can be collected into a non-negative symmetric matrix $\boldsymbol{\Lambda} = [\lambda_{g_1g_2}]$

Targeted fused ridge estimation

Maximizing argument for class g_0

$$\hat{\boldsymbol{\Omega}}_{g_0}\left(\boldsymbol{\Lambda}, \{\boldsymbol{\Omega}_g\}_{g\neq g_0}\right) = \left\{\left[\bar{\lambda}_{g_0}\mathbf{I}_p + \frac{1}{4}\left(\bar{\mathbf{S}}_{g_0} - \bar{\lambda}_{g_0}\mathbf{T}_{g_0}\right)^2\right]^{1/2} + \frac{1}{2}\left(\bar{\mathbf{S}}_{g_0} - \bar{\lambda}_{g_0}\mathbf{T}_{g_0}\right)\right\}^{-1},$$

where

$$\bar{\mathbf{S}}_{g_0} = \mathbf{S}_{g_0} - \sum_{g\neq g_0}\frac{\lambda_{gg_0}}{n_{g_0}}(\boldsymbol{\Omega}_g - \mathbf{T}_g), \quad \text{and} \quad \bar{\lambda}_{g_0} = \frac{\sum_g \lambda_{gg_0}}{n_{g_0}}$$

Properties

Behavior

i. $\hat{\boldsymbol{\Omega}}_g \succ \mathbf{0}$ for all $\lambda_{gg} \in (0, \infty)$;

ii. $\lim_{\lambda_{gg} \to 0^+} \hat{\boldsymbol{\Omega}}_g = \mathbf{S}_g^{-1}$ if $\sum_{g' \neq g} \lambda_{gg'} = 0$ and $p \leq n_g$;

iii. $\lim_{\lambda_{gg} \to \infty} \hat{\boldsymbol{\Omega}}_g = \mathbf{T}_g$ if $\lambda_{gg'} < \infty$ for all $g' \neq g$;

iv. $\lim_{\lambda_{g_1 g_2} \to \infty} (\hat{\boldsymbol{\Omega}}_{g_1} - \mathbf{T}_{g_1}) = \lim_{\lambda_{g_1 g_2} \to \infty} (\hat{\boldsymbol{\Omega}}_{g_2} - \mathbf{T}_{g_2})$ if $\lambda_{g_1' g_2'} < \infty$ for all $\{g_1', g_2'\} \neq \{g_1, g_2\}$.

Block coordinate ascent

Input:
Sufficient data: $(\mathbf{S}_1, n_1), \ldots, (\mathbf{S}_G, n_G)$
Penalty matrix: $\boldsymbol{\Lambda}$
Convergence criterion: $\varepsilon > 0$
Output:
Estimates: $\hat{\boldsymbol{\Omega}}_1, \ldots, \hat{\boldsymbol{\Omega}}_G$
procedure RIDGEP.FUSED$(\mathbf{S}_1, \ldots, \mathbf{S}_G, n_1, \ldots, n_G, \boldsymbol{\Lambda}, \varepsilon)$
    Initialize: $\hat{\boldsymbol{\Omega}}_g^{(0)}$ for all g.
    for $c = 1, 2, 3, \ldots$ **do**
        for $g = 1, 2, \ldots, G$ **do**
            Update $\hat{\boldsymbol{\Omega}}_g^{(c)} := \hat{\boldsymbol{\Omega}}_g\left(\boldsymbol{\Lambda}, \hat{\boldsymbol{\Omega}}_1^{(c)}, \ldots, \hat{\boldsymbol{\Omega}}_{g-1}^{(c)}, \hat{\boldsymbol{\Omega}}_{g+1}^{(c-1)}, \ldots, \hat{\boldsymbol{\Omega}}_G^{(c-1)}\right)$
        end for
        if $\max_g \left\{ \frac{\|\hat{\boldsymbol{\Omega}}_g^{(c)} - \hat{\boldsymbol{\Omega}}_g^{(c-1)}\|_F^2}{\|\hat{\boldsymbol{\Omega}}_g^{(c)}\|_F^2} \right\} < \varepsilon$ **then**
            return $\left(\hat{\boldsymbol{\Omega}}_1^{(c)}, \ldots, \hat{\boldsymbol{\Omega}}_G^{(c)}\right)$
        end if
    end for
end procedure