# BLOOD-BASED METABOLIC SIGNATURES IN ALZHEIMER'S DISEASE: SUPPLEMENTARY TEXT 2

# STATISTICAL ANALYZES: APPROACHES AND RESULTS

FRANCISCA A. DE LEEUW*†, CAREL F.W. PEETERS*†,
MAARTJE I. KESTER, AMY C. HARMS, EDUARD A. STRUYS,
THOMAS HANKEMEIER, HERMAN W.T. VAN VLIJMEN,
SVEN J. VAN DER LEE, CORNELIA M. VAN DUIJN, PHILIP SCHELTENS,
AYŞE DEMIRKAN, MARK A. VAN DE WIEL,
WIESJE M. VAN DER FLIER, AND CHARLOTTE E. TEUNISSEN

This supplementary text contains additional information on the data analysis. Section 1 contains information on processing the metabolite data for statistical analyzes. Section 2 then contains extensive information on the analyzes surrounding the differential expression, classification, and regulatory signatures. This section also contains all obtained results. Please note that, as this supplementary text is self-contained, there is some redundancy in presentation.

## 1. Data and Data Processing

1.1. **Data.** Plasma samples of 150 subjects with Alzheimer's disease (AD) and 150 subjects with subjective cognitive decline (SCD) were available. Subjects with SCD were used as cognitively normal controls in this study. Of these 300 subjects 263 (136 AD and 127 SCD) had their diagnosis confirmed by cerebral spinal fluid (CSF) markers (t-tau/A$\beta_{42}$ $>$ 0.52 for AD diagnosis). The 263 subjects with CSF-confirmed diagnosis were used for further study. Metabolite concentrations in four metabolite classes were determined using four different mass spectrometry platforms: amines (53), organic acids (22), lipids (120) and oxidative stress (40) compounds (see *Supplementary Text 1*).

1.2. **Data Processing.** Metabolites with more than 10% missing observations were removed, leading to the removal of 5 metabolites (the oxidative stress compound iPF2a-Unknown, and the lipids CE(18:1), TG(57:2), TG(58:3), and PE(O-38:7)). Three data samples (i.e., vectors of observed metabolite abundancies stemming from corresponding plasma samples) were removed as their (plasma) quality was deemed unsure. These samples had many (30 or more) concentrations below the limit of detection (LOD) that could not be attributed to instrumental errors. Twelve additional data samples were removed due to instrumental errors in one

* Shared first authorship.
† Corresponding author.

TABLE S2.1. List of clinical variables.

| Variable | Measurement |
|---|---|
| **Anthropometric**: | |
| Age | years at diagnosis |
| Sex | male or female |
| APOE $\epsilon 4$ allele status | at least one $\epsilon 4$ allele: yes, no |
| Mean arterial pressure | approximated by $\text{DBP} + \frac{1}{3}(\text{SBP} - \text{DBP})$ |
| Body mass index | $\text{weight}_{\text{kg}}/\text{height}^2_{\text{m}}$ |
| **Intoxications**: | |
| Smoking | status: current, former, never |
| Alcohol | current consumption of: yes, no |
| **Comorbidities**: | |
| Hypertension | present: yes, no |
| Diabetes Mellitus | present: yes, no |
| Hypercholesterolemia | present: yes, no |
| **Medication**: | |
| Cholesterol lowering medications | usage: yes, no |
| Antidepressant medications | usage: yes, no |
| Antiplatelet medications | usage: yes, no |

or more platforms. Hence, we only retain data samples that were free of instrumental errors across all four different mass spectrometry platforms. The remaining missing values are attributable to concentrations failing the LOD. These (feature-specific) missing values were imputed by half of the lowest observed value (for the corresponding feature). The final metabolic data set thus contained $n = 248$ data samples (127 AD and 121 SCD) and $p = 230$ metabolic features.

In addition to metabolomics data, phenotypic data (clinical and demographic characteristics such as height, weight, and APOE $\epsilon 4$ allele status) were evaluated for their possible confounding effects in the expression and classification signatures demarcating the AD and SCD groups. The missing observations on these variables (14 at most, for the height variable) were imputed. Continuous variables were imputed on the basis of Bayesian linear regression, polytomous variables were imputed on the basis of polytomous regression, and binary variables were imputed on the basis of logistic regression [S2.1]. To relief 'correctional stress' on the expression and classification signatures, certain aggregational clinical measures were calculated. The Body Mass Index (BMI) was calculated as $\text{weight}_{\text{kg}}/\text{height}^2_{\text{m}}$. In addition, the Mean Arterial Pressure (MAP) was approximated from the systolic blood pressure (SBP) and diastolic blood pressure (DBP) by $\text{DBP} + \frac{1}{3}(\text{SBP} - \text{DBP})$. See Table S2.1 for a full list of considered confounders.

## 2. SIGNATURES

### 2.1. **Differential Expression Signature.**

2.1.1. *Approach.* Differential metabolic expression between AD and SCD subjects was assessed by using nested linear models. We tested, for each individual metabolite, if its addition to a model containing the clinical characteristics (see Table S2.1) significantly contributed to model fit. One then assesses if, conditional on the effects of the clinical characteristics, metabolic expression does indeed differ between

the AD and SCD groups. Let $\mathrm{BG}_k$ represent the $k$th background or clinical variable and let $\mathrm{I}_{\mathrm{AD}}$ denote an indicator variable for AD group-membership. We are then interested in testing the following (abusing notation somewhat) nested models

$$\text{metabolite}_j = \beta_0 + \sum_{k=1}^{m} \beta_k \mathrm{BG}_k + \epsilon \tag{S2.1}$$

$$\text{metabolite}_j = \beta_0 + \sum_{k=1}^{m} \beta_k \mathrm{BG}_k + \beta_{(m+1)} \mathrm{I}_{\mathrm{AD}} + \epsilon, \tag{S2.2}$$

where the reduced model in (S2.1) is clearly nested in the full model (S2.2). This entails a test for nested models which, in this case, is equivalent to testing $H_0 : \beta_{(m+1)} = 0$ versus $H_a : \beta_{(m+1)} \neq 0$. The associated test statistic $F$ (see any standard statistics textbook) is distributed as $\mathcal{F}_{1,n-(m+2)}$ under the null hypothesis. The $p$-value for the observed test statistic can be obtained in reference to this distribution.

We have a multiple testing problem as we need to perform this test for each individual metabolite. Our approach to multiple testing is by controlling the False Discovery Rate (FDR), i.e., we aim to control "the expected proportion of falsely rejected hypotheses" [S2.2]. We control the FDR at .05.

2.1.2. *Results.* The metabolic features that survive multiple testing correction are listed in Table S2.2. The differential distributions for these features are depicted in Figures S2.1, S2.2, and S2.3. We see that all implicated features (except for SM(d18:1/20:1)) are underexpressed in the AD group relative to the control group. Table S2.3 contains, for purposes of comparison, the list of metabolic features that survive multiple testing correction when testing nested models in which only sex and age are used as possible confounders. We see that under less stringent corrections the list of potentially differentially expressed metabolites is longer. Substantive corrections harness against overoptimistic expression signatures.

## 2.2. **Classification Signature.**

2.2.1. *Approach.* Metabolic classification signatures for the prediction of group membership (AD or SCD) were constructed by way of penalized logistic regression with a Lasso-penalty [S2.3]. Two penalized settings were considered: (i) the Lasso selects among the metabolites while the clinical characteristics go unpenalized; and (ii) the Lasso selects among the metabolites without considering the clinical characteristics. The resulting models were compared to an unpenalized logistic regression model that (iii) considered only the clinical characteristics. Model estimation collides with minimizing the negative log-likelihood of the logistic model under an $\ell_1$-penalty. The general problem can then be stated as:

$$\underset{\beta_0, \boldsymbol{\beta}^u, \boldsymbol{\beta}^p}{\arg\min} \left\{ -\frac{1}{n} \mathcal{L}\left(\beta_0, \boldsymbol{\beta}^u, \boldsymbol{\beta}^p; \mathbf{y}, \mathbf{X}^u, \mathbf{X}^p\right) + \lambda_1 \|\boldsymbol{\beta}^p\|_1 \right\}, \tag{S2.3}$$

with $\mathcal{L}(\cdot)$ denoting the log-likelihood of the logistic model, $\mathbf{y}$ the binary $n$-dimensional response vector, $\mathbf{X}^u$ denoting the $(n \times m)$-dimensional matrix of clinical-predictors, $\mathbf{X}^p$ denoting the $(n \times p)$-dimensional matrix of metabolite-predictors, $\boldsymbol{\beta}^u$ an $m$-dimensional vector of unpenalized regression coefficients, $\boldsymbol{\beta}^p$ a $p$-dimensional vector of penalized regression coefficients, and with $\beta_0$ denoting an intercept. Lastly, $\lambda_1 \| \cdot \|_1$ indicates the $\ell_1$-norm with penalty parameter $\lambda_1$,

TABLE S2.2. Differentially expressed metabolites.

| Metabolite | Compound class | $p$-value | Adjusted $p$-value |
|---|---|---|---|
| 2-Aminoadipic acid | Amines | .0003110134 | .03071051 |
| TG(51:3) | Lipids: Triglycerides | .0005518318 | .03071051 |
| 3-Hydroxyisovaleric acid | Organic acids | .0005645572 | .03071051 |
| Tyrosine | Amines | .0006614924 | .03071051 |
| TG(54:6) | Lipids: Triglycerides | .0009136006 | .03071051 |
| TG(50:4) | Lipids: Triglycerides | .0010320542 | .03071051 |
| S-3-Hydroxyisobutyric acid | Organic acids | .0011085893 | .03071051 |
| TG(56:8) | Lipids: Triglycerides | .0012115039 | .03071051 |
| Methyldopa | Amines | .0012272994 | .03071051 |
| 8-iso-PGF2a (15-F2t-IsoP) | Oxidative stress: Isoprostane | .0013352397 | .03071051 |
| TG(48:3) | Lipids: Triglycerides | .0017977057 | .03703538 |
| O-Acetylserine | Amines | .0020920976 | .03703538 |
| TG(48:2) | Lipids: Triglycerides | .0025274816 | .03703538 |
| Methylmalonic acid | Organic acids | .0026395625 | .03703538 |
| TG(46:2) | Lipids: Triglycerides | .0027475334 | .03703538 |
| Valine | Amines | .0027973120 | .03703538 |
| TG(50:3) | Lipids: Triglycerides | .0031949553 | .03703538 |
| TG(52:4) | Lipids: Triglycerides | .0034264336 | .03703538 |
| TG(52:5) | Lipids: Triglycerides | .0034406945 | .03703538 |
| TG(56:7) | Lipids: Triglycerides | .0034900406 | .03703538 |
| TG(48:0) | Lipids: Triglycerides | .0035736266 | .03703538 |
| Ornithine | Amines | .0036249029 | .03703538 |
| SM(d18:1/23:0) | Lipids: Sphingomyelins | .0037035385 | .03703538 |
| SM(d18:1/20:1) | Lipids: Sphingomyelins | .0048428315 | .04491679 |
| TG(48:1) | Lipids: Triglycerides | .0049569537 | .04491679 |
| TG(58:10) | Lipids: Triglycerides | .0050775504 | .04491679 |

generally referred to as the Lasso-penalty. The Lasso-penalty enables estimation in our setting where the feature to observation ratio (230/248) is too high for standard logistic regression. It also achieves automatic model (i.e., feature) selection. The problem in S2.3 is generally stated, in the sense that it captures all situations of interest. In situation (iii) only the clinical predictors $\mathbf{X}^u$ are considered, such that the unpenalized parameters $\beta_0$ and $\boldsymbol{\beta}^u$ are estimated. In situation (ii) only the metabolite-predictors $\mathbf{X}^p$ are considered, such that, next to $\beta_0$, the penalized parameters in $\boldsymbol{\beta}^p$ are estimated. Situation (i) combines the former situations and, hence, considers the the general problem, estimating both unpenalized and penalized parameters. The optimal penalty parameter in the penalized models was determined on the basis of leave-one-out cross-validation (LOOCV) of the model likelihood. Predictive performance of all models was assessed by way of (the comparison of) Receiver Operating Characteristic (ROC) curves and Area Under the ROC Curves (AUCs). ROC curves and AUCs for all models were produced by 10-fold cross-validation.

Note that the metabolic features were scaled in the classification exercises. The (regularized) regression makes use of (in some sense) the covariance matrix of the features. However, the variability of the features may differ substantially. In such a situation the features with (relatively) extreme variability may drive the results. Hence, it is appropriate to perform regularization on the standardized scale.

2.2.2. *Results.* Model performances can be found in Figure S2.4. The prediction model carrying the clinical variables only resulted in an AUC of .736. The model

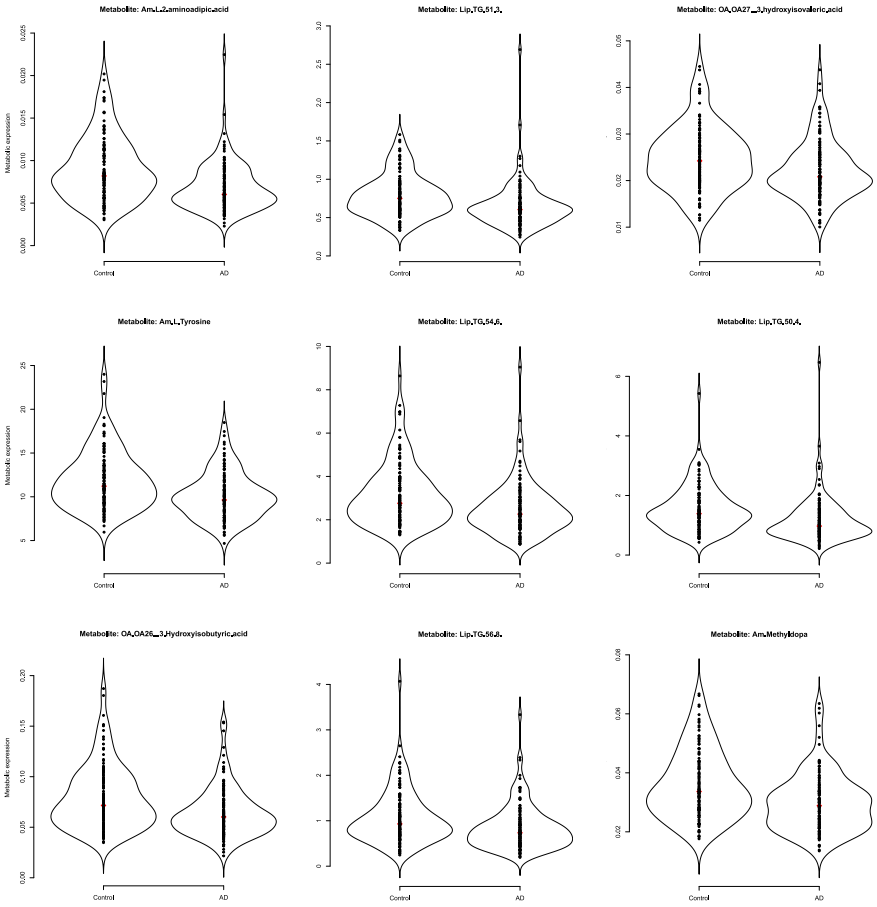


FIGURE S2.1. Violin plots of (a selection of) the metabolites that survive multiple testing correction. Violin plots [S2.4] combine the familiar box plot with a kernel density to better represent the distribution of the data. We see relative underexpression in the AD group for all depicted metabolites. The associated adjusted $p$-values can be found in Table S2.2. The remaining violin plots can be found in Figures S2.2 and S2.3.

that used the Lasso for selection amongst the metabolites sorts a comparable classification performance, yielding an AUC of approximately .7. The model that adds a (Lasso-based) selection of metabolites to the clinical variables then improves predictive performance along the full false positive rate range, sorting a AUC of .79. Table S2.4 contains the metabolites selected in the selection-amongst-metabolites-only situation. Table S2.5 then contains the metabolites selected in the selection-amongst-metabolites-whilst-clinical-variables-present situation. The highlighted features in these tables are also present in the differential expression signature (see Section 2.1). We see that the signs of their effects concur with the pattern of AD-associated under- and overexpression present in the differential expression signature.

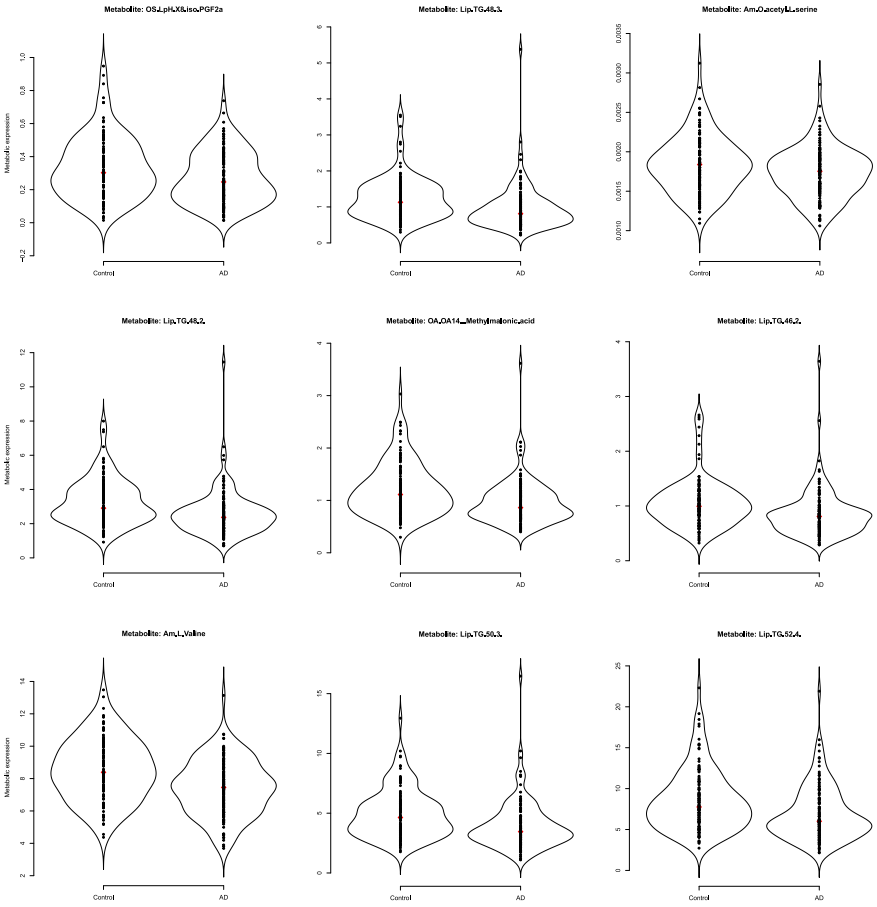


FIGURE S2.2. Violin plots of (a selection of) the metabolites that survive multiple testing correction. Violin plots [S2.4] combine the familiar box plot with a kernel density to better represent the distribution of the data. We see relative underexpression in the AD group for all depicted metabolites. The associated adjusted $p$-values can be found in Table S2.2. The remaining violin plots can be found in Figures S2.1 and S2.3.

Note that the Lasso selects features from all metabolite classes. To assess if the metabolite-class has predictive power a group-regularized logistic ridge regression [S2.5] was used in which the metabolite-class serves as co-data. This analysis indicated that the metabolite-class forms weakly informative co-data. This strengthens faith in the Lasso results. The strongest predictor amongst the clinical variables is (naturally) APOE $\epsilon 4$ allele status.

## 2.3. **Regulatory Signature.**

2.3.1. *Graphical Modeling.* A differential expression signature represents the features that are relatively under- or overexpressed in diagnostic groups of interest. This signature does not have to concur with the classification signature completely,

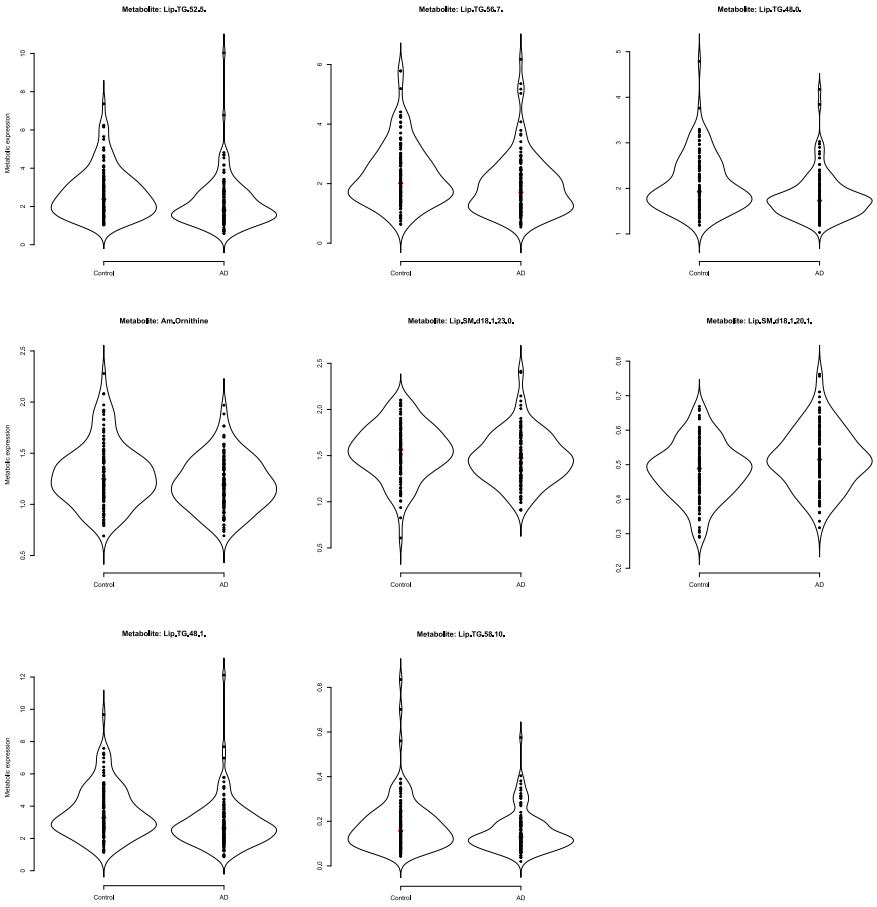


FIGURE S2.3. Violin plots of (a selection of) the metabolites that survive multiple testing correction. Violin plots [S2.4] combine the familiar box plot with a kernel density to better represent the distribution of the data. We see relative underexpression in the AD group for all depicted metabolites except SM(d18:1/20:1). The associated adjusted $p$-values can be found in Table S2.2. The remaining violin plots can be found in Figures S2.1 and S2.2.

as the latter (i) chooses amongst multicollinear features and (ii) emphasizes prediction rather than shifts in location. The classification signature, in turn, is limited in its capacity to represent complex dependencies amongst the metabolites (of interest). Hence, we seek to explore a third signature: The regulatory signature. This signature intends to uncover deregulation in metabolic biochemical pathways as pertaining to the AD disease process. A metabolic pathway can be thought of as a collection of metabolic features originating from all over the metabolome, that work interdependently to regulate some biochemical (disease) process. Hence, a pathway is a network. And a network can be represented by a graph. We thus take interest in graphical modeling.

TABLE S2.3. Differentially expressed metabolites when correcting for sex and age only.

| Metabolite | Compound class | $p$-value | Adjusted $p$-value |
|---|---|---|---|
| 2-Aminoadipic acid | Amines | 1.236871e−07 | 2.844803e−05 |
| Valine | Amines | 3.316148e−06 | 3.494500e−04 |
| Tyrosine | Amines | 4.558044e−06 | 3.494500e−04 |
| Methyldopa | Amines | 7.515081e−06 | 4.321172e−04 |
| Lysine | Amines | 2.749770e−05 | 1.264894e−03 |
| Methylmalonic acid | Organic acids | 4.222770e−05 | 1.390688e−03 |
| S-3-Hydroxyisobutyric acid | Organic acids | 4.232528e−05 | 1.390688e−03 |
| TG(48:0) | Lipids: Triglycerides | 7.904916e−05 | 2.272663e−03 |
| TG(50:4) | Lipids: Triglycerides | 9.542238e−05 | 2.438572e−03 |
| TG(48:2) | Lipids: Triglycerides | 1.144591e−04 | 2.472172e−03 |
| TG(51:3) | Lipids: Triglycerides | 1.182343e−04 | 2.472172e−03 |
| TG(54:6) | Lipids: Triglycerides | 1.305959e−04 | 2.503088e−03 |
| TG(50:3) | Lipids: Triglycerides | 1.629475e−04 | 2.882490e−03 |
| TG(50:2) | Lipids: Triglycerides | 1.754559e−04 | 2.882490e−03 |
| TG(50:1) | Lipids: Triglycerides | 2.217775e−04 | 3.400588e−03 |
| TG(48:1) | Lipids: Triglycerides | 2.421916e−04 | 3.481504e−03 |
| TG(52:4) | Lipids: Triglycerides | 3.157891e−04 | 4.141486e−03 |
| TG(48:3) | Lipids: Triglycerides | 3.241163e−04 | 4.141486e−03 |
| Leucine | Amines | 4.124135e−04 | 4.871236e−03 |
| LPC(18:1) | Lipids: Lysophosphatidylcholine | 4.235858e−04 | 4.871236e−03 |
| TG(46:2) | Lipids: Triglycerides | 5.465651e−04 | 5.986189e−03 |
| TG(50:0) | Lipids: Triglycerides | 6.518719e−04 | 6.815024e−03 |
| TG(52:5) | Lipids: Triglycerides | 7.216611e−04 | 7.216611e−03 |
| TG(52:3) | Lipids: Triglycerides | 1.149358e−03 | 1.101469e−02 |
| TG(51:2) | Lipids: Triglycerides | 1.393980e−03 | 1.252523e−02 |
| TG(56:8) | Lipids: Triglycerides | 1.451992e−03 | 1.252523e−02 |
| Isoleucine | Amines | 1.470353e−03 | 1.252523e−02 |
| 2-hydroxybutyric acid | Organic acids | 1.704211e−03 | 1.399888e−02 |
| 3-Hydroxyisovaleric acid | Organic acids | 1.845997e−03 | 1.464067e−02 |
| TG(51:1) | Lipids: Triglycerides | 1.996360e−03 | 1.530543e−02 |
| SM(d18:1/20:1) | Lipids: Sphingomyelins | 2.242748e−03 | 1.663974e−02 |
| TG(52:1) | Lipids: Triglycerides | 2.377869e−03 | 1.709093e−02 |
| 8-iso-PGF2a (15-F2t-IsoP) | Oxidative stress: Isoprostane | 2.670123e−03 | 1.860995e−02 |
| Proline | Amines | 3.197010e−03 | 2.162683e−02 |
| TG(54:5) | Lipids: Triglycerides | 3.389655e−03 | 2.227487e−02 |
| TG(56:7) | Lipids: Triglycerides | 3.846194e−03 | 2.401863e−02 |
| PGD2 | Lipids: Prostaglandins | 3.863867e−03 | 2.401863e−02 |
| TG(46:1) | Lipids: Triglycerides | 4.273714e−03 | 2.586722e−02 |
| PC(O-44:5) | Lipids: Plasmalogen Phosphatidylcholine | 4.595494e−03 | 2.683819e−02 |
| LPA C14:0 | Lyso-phosphatidic acid | 4.667512e−03 | 2.683819e−02 |
| PC(O-34:1) | Lipids: Plasmalogen Phosphatidylcholine | 5.876174e−03 | 3.296390e−02 |
| LPC(20:4) | Lipids: Lysophosphatidylcholine | 7.163584e−03 | 3.922915e−02 |
| SM(d18:1/24:2) | Lipids: Sphingomyelins | 7.371942e−03 | 3.943132e−02 |
| 8,12-iPF2a IV | Oxidative stress: Isoprostane | 8.208661e−03 | 4.290891e−02 |
| TG(46:0) | Lipids: Triglycerides | 8.513935e−03 | 4.351567e−02 |
| 5-iPF2a VI | Oxidative stress: Isoprostane | 9.211706e−03 | 4.589083e−02 |
| TG(52:2) | Lipids: Triglycerides | 9.377692e−03 | 4.589083e−02 |
| SM(d18:1/16:0) | Lipids: Sphingomyelins | 9.782507e−03 | 4.687451e−02 |
| TG(58:10) | Lipids: Triglycerides | 1.063227e−02 | 4.861745e−02 |
| Ornithine | Amines | 1.064486e−02 | 4.861745e−02 |
| Histidine | Amines | 1.078039e−02 | 4.861745e−02 |

Graphical modeling refers to a class of probabilistic models that use graphs to express conditional (in)dependence relations between random variables. We consider graphs $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ consisting of a finite set $\mathcal{V}$ of vertices and set of edges $\mathcal{E}$. The vertices of the graph correspond to a collection of random variables with probability distribution $\mathcal{P}$, i.e., $\{Y_1, \ldots, Y_p\} \sim \mathcal{P}$. Edges in $\mathcal{E}$ consist of pairs of distinct vertices such that $Y_j - Y_{j'} \in \mathcal{E}$. The basic assumption is: $\{Y_1, \ldots, Y_p\} \sim \mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma})$, with $\mathbf{\Sigma}$ positive definite. Hence, we focus on Gaussian graphical modeling.

In this Gaussian case, conditional independence between a pair of variables corresponds to zero entries in the precision matrix. Indeed, let $\hat{\mathbf{\Sigma}}^{-1} = \hat{\mathbf{\Omega}}$ denote a generic estimate of the precision matrix and consider its transformation to a partial correlation matrix $\hat{\mathbf{P}}$. Then the following relations can be shown to hold for all pairs $\{Y_j, Y_{j'}\} \in \mathcal{V}$ with $j \neq j'$ [see, e.g., S2.6]:

$$(\hat{\mathbf{P}})_{jj'} = 0 \iff (\hat{\mathbf{\Omega}})_{jj'} = 0 \iff Y_j \perp\!\!\!\perp Y_{j'} | \mathcal{V} \setminus \{Y_j, Y_{j'}\} \iff Y_j \not\sim Y_{j'},$$

where $\mathcal{V} \setminus \{\cdot\}$ denotes set-minus notation and where $\not\sim$ indicates the absence of an edge. In words: A zero partial correlation implies a zero precision matrix entry which in turn implies that the corresponding two variables are conditionally independent (given the remaining variables) which then implies the absence of an edge between these variables in the corresponding graph. Such a graph can thus be interpreted as a conditional independence graph.

2.3.2. *Approach.* Now, model selection efforts in Gaussian graphical models focus on determining the support of the precision matrix. Problematic is that in situations with $p \approx n$ or $p > n$ the sample covariance matrix $\hat{\mathbf{\Sigma}} = \mathbf{S}$ is ill-behaved or singular such that it's inverse ($\mathbf{S}^{-1}$, which would constitute an estimate of the precision matrix) is unstable or does not exist. Moreover, the metabolic features of interest are highly collinear (within the respective metabolite classes). Hence, we need a regularized estimate of the precision matrix. In addition, we want to take into account that our data consist of distinct classes of interest. The first distinction is, naturally, AD versus SCD.

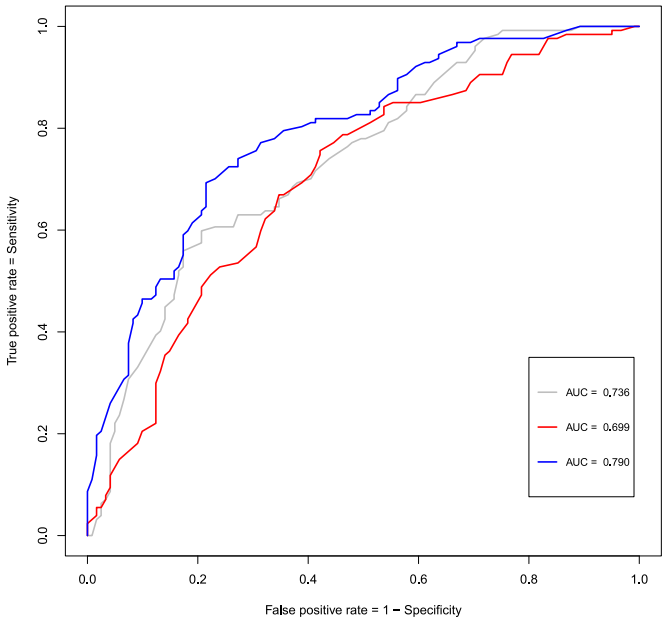


FIGURE S2.4. ROC curves for the classification models. The grey line represents the ROC curve for the unpenalized logistic regression model that entertains the clinical characteristics only. The red line represents the ROC curve for the logistic model in which the Lasso performed variable selection amongst the metabolites (and that does not consider the clinical characteristics). The blue line represents the ROC curve of the logistic model in which the clinical characteristics are present while the Lasso may select amongst the metabolites. The clinical variables are listed in Table S2.1. Appears as Figure 2 in the main text.

TABLE S2.4. Selected metabolites and parameter estimates when considering only metabolites as potential predictors.

| Metabolite | Compound class | $\hat{\beta}$ |
|---|---|---|
| LPC(18:1) | Lipids: Lysophosphatidylcholine | .431701077 |
| PGD2 | Oxidative stress: Prostaglandins | −.335722610 |
| 8,12-iPF2a IV | Oxidative stress: Isoprostane | −.321506682 |
| O-Acetylserine | Amines | −.290694860 |
| Methyldopa | Amines | −.226545678 |
| NO2-aLA (C18:3) | Oxidative stress: Nitro-Fatty acid | −.214894781 |
| Methylmalonic acid | Organic acids | −.211631150 |
| TG(51:3) | Lipids: Triglycerides | −.202508671 |
| Tyrosine | Amines | −.192600743 |
| Serine | Amines | .151993825 |
| LPC(20:4) | Lipids: Lysophosphatidylcholine | .143911633 |
| Arginine | Amines | .139364338 |
| SM(d18:1/23:0) | Lipids: Sphingomyelins | −.138378505 |
| Glyceric acid | Organic acids | −.107386135 |
| Lysine | Amines | −.101945440 |
| Glycolic acid | Organic acids | .096537107 |
| cLPA C18:0 | Oxidative stress: Cyclic-lyso-phosphatidic acid | −.072919587 |
| SM(d18:1/18:0) | Lipids: Sphingomyelins | .068838390 |
| LPA C22:4 | Oxidative stress: Lyso-phosphatidic acid | .065264732 |
| LPA C16 | Oxidative stress: Lyso-phosphatidic acid | −.064109038 |
| 3-Methoxytyramine | Amines | −.063622716 |
| TG(54:6) | Lipids: Triglycerides | −.062333261 |
| 2,3-dinor-8-iso-PGF2a | Oxidative stress: Isoprostane | −.059017772 |
| PC(O-34:3) | Lipids: Plasmalogen Phosphatidylcholine | −.048557268 |
| Cis-Aconitic acid | Organic acids | −.038573969 |
| 3-Hydroxyisovaleric acid | Organic acids | −.034646727 |
| LPA C14:0 | Oxidative stress: Lyso-phosphatidic acid | −.033059634 |
| 2-Aminoadipic acid | Amines | −.028720481 |
| PC(O-36:6) | Lipids: Plasmalogen Phosphatidylcholine | −.025814804 |
| PC(O-38:6) | Lipids: Plasmalogen Phosphatidylcholine | −.021959733 |
| Putrescine | Amines | −.017413981 |
| TG(48:0) | Lipids: Triglycerides | −.016328393 |
| Homoserine | Amines | −.015906351 |
| TG(O-50:0) | Lipids: Triglycerides | .014668718 |
| Carnosine | Amines | .012064246 |
| 5-iPF2a VI | Oxidative stress: Isoprostane | −.003633138 |
| Sarcosine | Amines | −.001072232 |

So, we are after jointly estimating multiple regularized precision matrices from (aggregated) high-dimensional data consisting of distinct classes. From a network perspective, molecular pathway-deregulation in the disease state is likely characterized by the loss of normal (wanted) molecular interactions and the gain of abnormal (unwanted) molecular interactions. One would thus expect the network topologies of our groups of interest to primarily share the same structure, while potentially differing in a number of (topological) locations of interest. Our regularized network extraction method takes this explicitly into account. Specifically, we employ a special case of targeted fused ridge estimation [S2.7], solving the following estimation

TABLE S2.5. Selected metabolites and parameter estimates when considering metabolites as potential predictors on top of the clinical variables.

| Metabolite | Compound class | $\hat{\beta}$ |
|---|---|---|
| PGD2 | Oxidative stress: Prostaglandins | −.47539723 |
| O-Acetylserine | Amines | −.45299269 |
| Methylmalonic acid | Organic acids | −.39360564 |
| NO2-aLA (C18:3) | Oxidative stress: Nitro-Fatty acid | −.32516790 |
| TG(51:3) | Lipids: Triglycerides | −.29844463 |
| SM(d18:1/20:1) | Lipids: Sphingomyelins | .29404906 |
| 3-Hydroxyisovaleric acid | Organic acids | −.27369915 |
| 8,12-iPF2a IV | Oxidative stress: Isoprostane | −.22711647 |
| PE(38:2) | Lipids: Phosphatidylethanolamine | −.16302211 |
| Gamma-glutamylalanine | Amines | .15344618 |
| LPC(18:1) | Lipids: Lysophosphatidylcholine | .15057771 |
| Methyldopa | Amines | −.14912088 |
| SM(d18:1/23:0) | Lipids: Sphingomyelins | −.14714005 |
| Putrescine | Amines | −.14137904 |
| LPC(20:4) | Lipids: Lysophosphatidylcholine | .12467937 |
| 8-iso-PGF2a (15-F2t-IsoP) | Oxidative stress: Isoprostane | −.12101061 |
| LPA C18:3 | Oxidative stress: Lyso-phosphatidic acid | −.10140491 |
| LPA C14:0 | Oxidative stress: Lyso-phosphatidic acid | −.09396454 |
| Uracil | Organic acids | .09348099 |
| Citrulline | Amines | −.08529403 |
| Histamine | Amines | .07351094 |
| Glyceric acid | Organic acids | −.06673703 |
| TG(56:8) | Lipids: Triglycerides | −.06152908 |
| TG(58:10) | Lipids: Triglycerides | −.05699948 |
| 2,3-dinor-8-iso-PGF2a | Oxidative stress: Isoprostane | −.04718506 |
| NO2-OA (C18:1) | Oxidative stress: Nitro-Fatty acid | −.04178329 |
| Glycolic acid | Organic acids | .04016606 |
| Carnosine | Amines | .03960024 |
| Serine | Amines | .03620446 |
| SM(d18:1/18:0) | Lipids: Sphingomyelins | .01812763 |
| S-3-Hydroxyisobutyric acid | Organic acids | −.01042997 |

problem:

$$\underset{\{\mathbf{\Omega}_g\}\in\mathcal{S}_{++}^p}{\arg\max}\left\{\mathcal{L}\left(\{\mathbf{\Omega}_g\};\{\mathbf{S}_g\}\right)-\frac{\lambda}{2}\sum_g\left\|\mathbf{\Omega}_g-\mathbf{T}\right\|_F^2-\frac{\lambda_f}{4}\sum_{g_1,g_2}\left\|\mathbf{\Omega}_{g_1}-\mathbf{\Omega}_{g_2}\right\|_F^2\right\}, \tag{S2.4}$$

where the $\mathbf{S}_g$ indicate group-specific sample covariance matrices, $\lambda$ denotes a strictly positive ridge penalty, $\lambda_f$ denotes a positive fusion penalty, and $\mathbf{T}$ denotes a target matrix. The penalty parameter $\lambda$ controls the rate of shrinkage of each precision $\mathbf{\Omega}_g$ towards the corresponding target $\mathbf{T}$, while $\lambda_f$ determines the retainment of entry-wise similarities between $\mathbf{\Omega}_{g_1}$ and $\mathbf{\Omega}_{g_2}$ for all class pairs $g_1 \neq g_2$. For given penalties the problem can be solved with an block coordinate ascent procedure [S2.7], resulting in an estimated precision matrix for each class $g$. In this case $g = 1, 2$.

We solve (S2.4) using the class-specific sample covariance matrices (i.e., the sample covariance matrices of the class-specific data) as the data entries. For the target $\mathbf{T}$ we choose the (weakly informative) $p$-dimensional identity matrix $\mathbf{I}_p$. The optimal penalty parameters were determined by the LOOCV procedure described in

[S2.7]. The optimal penalty values were found to be $\lambda^* = 2.742348$, and $\lambda_f^* =$ 9.867606e-22. These penalty values emphasize individual regularization over retainment of entry-wise similarities, indicating strong differences in class-specific precision matrices. The support of the estimated class precision matrices was determined by thresholding. For each class-specific matrix, the 100 strongest edges (in terms of absolute partial correlations) were retained. As metabolic networks are very dense, retaining the 100 strongest edges is assumed to give a more clear picture of the most influential regulatory players. The retained partial correlations range, in absolute value, from .1670877 to .6412267 over the respective classes. All analyzes were performed with the `rags2ridges` package [S2.8] in `R` [S2.9].

2.3.3. *Visualization.* The first idea regarding the network structures represented by our class-specific precision matrices can be obtained by simple visualization. Figure S2.5 contains the class-specific networks visualized with the Fruchterman-Reingold (FR) algorithm [S2.10]. These networks contain all metabolic features, even when they are not connected, resulting in 'hairball' networks: Networks that are too tangled to be effectively visualized. They do, however, convey that the strongest edges implicate metabolic features from all considered metabolite families.

Figure S2.6 contains the pruned class-specific networks visualized with the FR algorithm. That is, it retains only the class-specific connected components from Figure S2.5. The pruned networks more effectively represent the retained topologies.

To assess the topology more closely, it is beneficial to arrange the metabolic features in fixed coordinates over the respective groups of interest. Figure S2.7

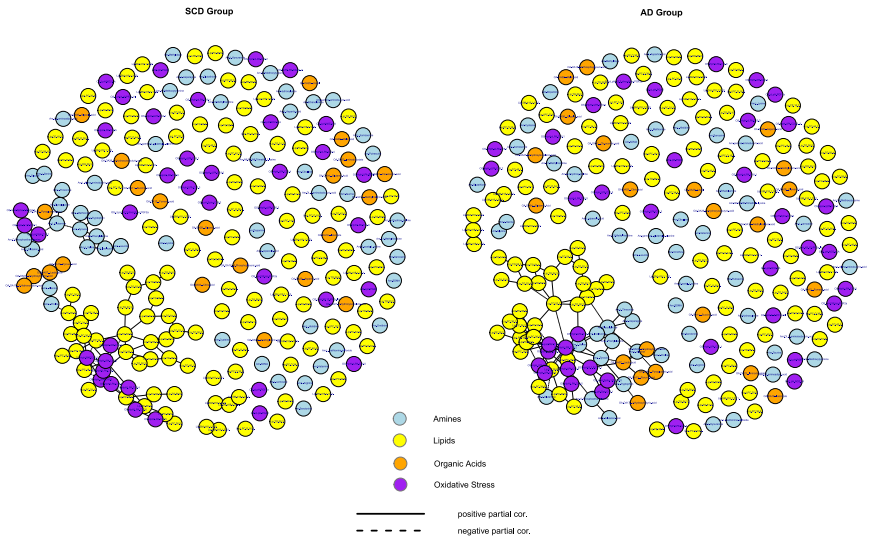


FIGURE S2.5. Class-specific networks visualized with the Fruchterman-Reingold algorithm. The left-hand panel contains the network for the SCD group. The right-hand panel contains the network for the AD group. The metabolite compounds are colored according to metabolite family: Blue for amines, yellow for lipids, orange for organic acids, and purple for oxidative stress. Solid edges represent positive partial correlations while dashed edges represent negative partial correlations.

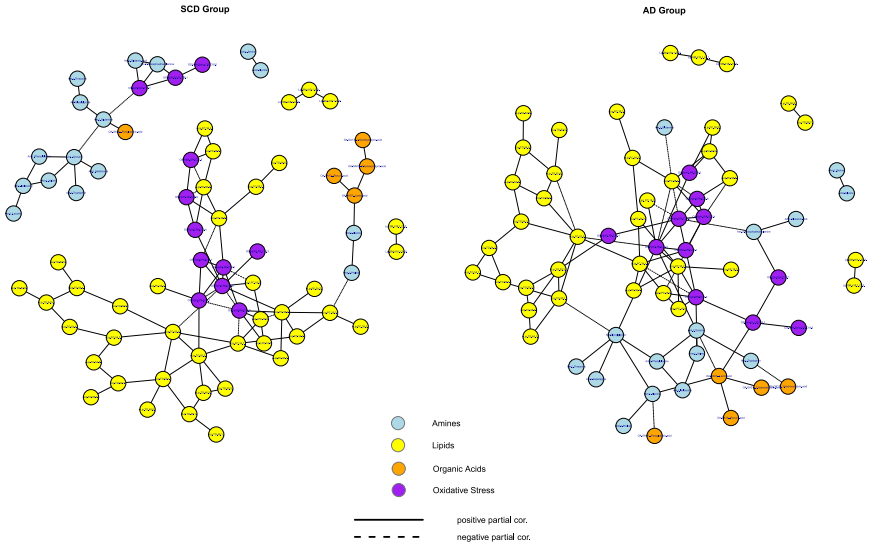


FIGURE S2.6. Class-specific *pruned* networks visualized with the Fruchterman-Reingold algorithm. The left-hand panel contains the network for the SCD group. The right-hand panel contains the network for the AD group. The metabolite compounds are colored according to metabolite family: Blue for amines, yellow for lipids, orange for organic acids, and purple for oxidative stress. Solid edges represent positive partial correlations while dashed edges represent negative partial correlations.

contains the semi-pruned class-specific networks. They are semi-pruned as in each class-specific topology all metabolites are depicted that are present in the union of connected metabolites over all class-specific topologies. This allows us to visualize the individual topologies with fixed metabolite-coordinates. The FR-based coordinates for the SCD group serve as the reference coordinates for all topologies. We see that the union of metabolic features is quite tight, suggesting that the core metabolic features for the SCD and AD groups overlap to a large extent. At first glance the diseased state indeed seem less locally connected. We now turn to numerical and graph theoretic assessments to support understanding of the topologies.

2.3.4. *Global Characteristics.* Here, we will assess some global characteristics of each class-related graph as given in Figure S2.7. Table S2.6 contains some global metrics for each topology. Please note that formal definitions of all terms relating to network science as used throughout this supplement can be found in, e.g., [S2.12]. Transitivity is a shape measure, with higher scores indicating stronger local connectivity. Transitivity for the SCD topology is approximately .24, which is higher than the transitivity score for the AD topology ($\approx .15$) and also higher than many other biological networks [S2.12, p. 200 & Section 8.6]. Hence, the SCD topology is stronger locally connected. Centrality [S2.13] is another shape measure, indicating the degree in which the topology resembles a maximally centralized graph (i.e., a star graph). The more centralized a network, the more vulnerable it is, in the sense that it's connectedness hinges upon few nodes. The centralization scores indicate that both the SCD and the AD topology are not very centralized.

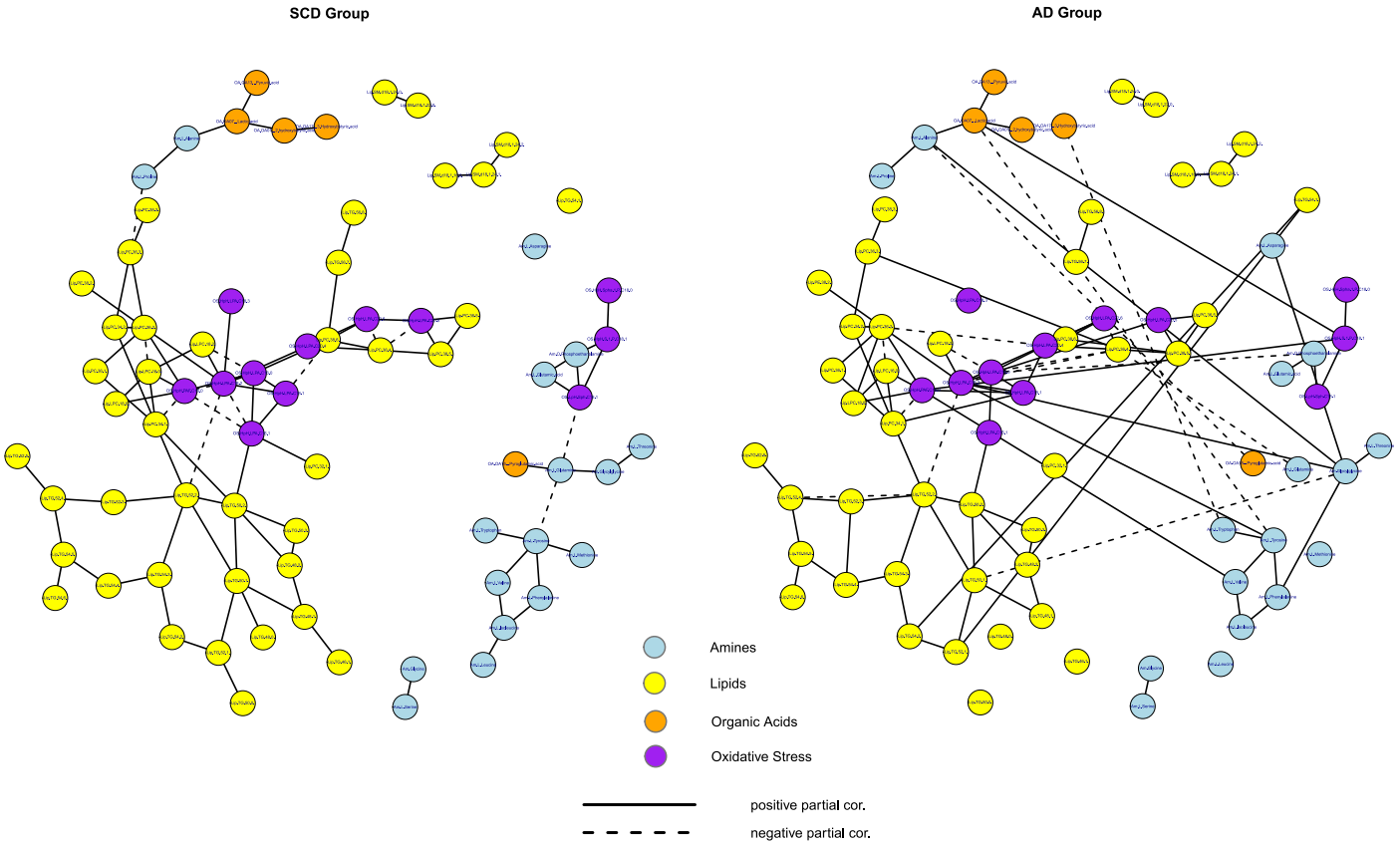


FIGURE S2.7. Class-specific *semi-pruned* networks visualized with the Fruchterman-Reingold algorithm. The left-hand panel contains the network for the SCD group. The right-hand panel contains the network for the AD group. The coordinates of the left-hand topology serve as the reference coordinates. The metabolite compounds are colored according to metabolite family: Blue for amines, yellow for lipids, orange for organic acids, and purple for oxidative stress. Solid edges represent positive partial correlations while dashed edges represent negative partial correlations.

In addition to shape metrics, there are cohesion metrics. One such cohesion metric is 'connectedness', which, for the topologies of interest, is also given in Table S2.6. Connectedness refers to the "proportion of pairs of nodes that can reach each other by a path of any length" [S2.14]. The AD topology has a higher connectedness score than the SCD topology. This is (in part) due to the fact that the SCD topology has a large disconnected component (consisting of amine compounds mostly). Figure S2.7 seems to indicate that the AD topology may be characterized by an increased connection density within amine compounds and between amine and oxidative stress compounds. This is reflected in the degree density and the relative outdegree [S2.15] for the respective topologies. The between-metabolite-family and within-metabolite-family degree densities (Figure S2.8) suggest that the AD topology is characterized by increased amine-connections (connections in which at least 1 amine compound is present). The relative outdegrees (Figure S2.9) imply that the AD topology is characterized by more connections between amines and organic acid compounds and more connections between amines and oxidative stress compounds.

TABLE S2.6. Global metrics for the two topologies of interest (see Figure S2.7).

| Topology | Transitivity | Connectedness | Centralization |
|---|---|---|---|
| SCD group | .2424242 | .4864865 | .0879304 |
| AD | .1458967 | .6187387 | .0879304 |

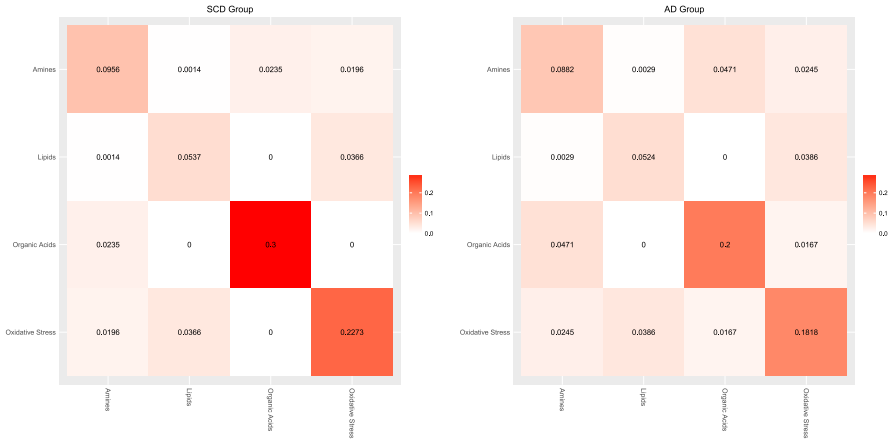


FIGURE S2.8. Heatmaps of degree densities for the SCD and AD networks. The reported numbers represent the degree density for the (combinations of) metabolite groups. Degree density represents the number of connections (edges) divided by the number of possible connections. For example, in the network for the SCD group the proportion of actual edges relative to the number of possible edges between Amines and oxidative stress compounds is .0196. Note that all heatmaps received the same color key. Hence, the color intensities (i.e., the color-spectrum representations of the cell-numbers) are comparable over the respective heatmaps.

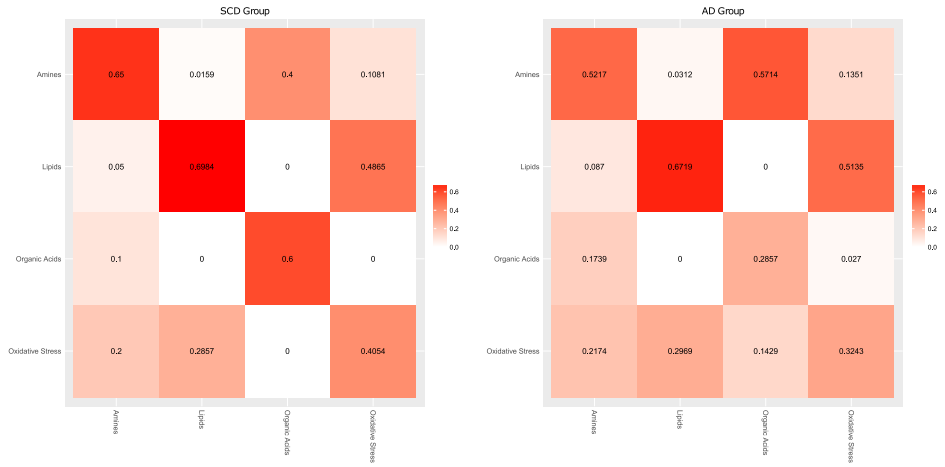


FIGURE S2.9. Heatmaps of relative outdegrees for the class-specific networks. The reported numbers represent the relative outdegree for the (combinations of) metabolite groups. The relative outdegree represents the number of connections (edges) between two metabolite groups divided by the number of 'outgoing' connections for one of these groups. For example, in the network for the SCD group the number of edges between lipid and oxidative stress compounds accounts for approximately 29% of all edges involving lipids and approximately 49% of all edges involving oxidative stress compounds. Note that all heatmaps received the same color key. Hence, the color intensities (i.e., the color-spectrum representations of the cell-numbers) are comparable over the respective heatmaps. Note that the column numbers should sum to unity.

2.3.5. *Node Characteristics.* In addition to global metrics, we assess certain characteristics of individual nodes within the topologies of interest, focusing especially on the notion of centrality. Centrality, in general, refers to metrics regarding the most central or (functionally) important nodes in a network. Several centrality measures are used whose formal definition can be found in, e.g., [S2.12]: degree centrality, betweenness centrality, closeness centrality, and eigenvalues centrality. Degree centrality simply indicates the number of connections in which a node takes part. It is indicative of the nodes that are central or influential in terms of the number of connections: more connections could imply deeper regulatory influence. Betweenness centrality [S2.16] measures centrality in terms of information flow. Under the assumption that information is passed over short(est) paths a node becomes central when the number of short(est) paths that pass through it is high. Closeness centrality indicates the mean distance of a node to other nodes. A node is central under the closeness centrality metric when it's mean distance to other nodes is low. Note that closeness as used here reflects the sum of inverse distances [S2.17], such that nodes that are close to many other nodes receive high closeness scores. The eigenvalue centrality [S2.12] is an extension of the degree centrality. A node's eigencentrality is based on the centrality of the nodes to which it is connected: connections to central others are weighted more heavily in the final eigencentrality score than connections to less central others. These various centrality scores thus

have different flavors and may indicate correspondingly flavored hubs (i.e., highly central nodes).

Figures S2.10 and S2.11 contain target plots [S2.18] depicting the various centrality properties of the metabolite compounds retained in the SCD and AD topologies. The target plots were created with the help of the `sna` and `igraph` packages in `R` [S2.19, S2.20, S2.21]. Tables S2.7 and S2.8 then contain, to accompany these figures, the top centrality metrics for the SCD and AD topologies, respectively. The strongest hub in the SCD topology is the oxidative stress compound LPA C18:2. This compound sorts the highest scores on all centrality measures. The top compounds, in terms of centrality, in the SCD topology all belong to either the oxidative stress or lipid family. Next to LPA C18:2, the Phosphatidylcholines PC(34:1) and PC(36:2) are consistently represented as hubs by all centrality measures. LPA C18:2 is also the strongest hub compound in the AD topology. The top compounds, in terms of centrality, of the AD topology indeed largely overlap with the top compounds of the SCD topology. Although the same compound may be a hub in both the AD and SCD topologies, it can still be wired very differently, i.e., it's connections may differ greatly between the two topologies (see Section 2.3.6). In addition, for the AD topology the amines Glycylglycine and Tyrosine are quite consistently indicated as central compounds by the degree, betweenness, and closeness metrics.

TABLE S2.7. Centrality measures for the SCD topology.

| Degree | | Betweenness | | Closeness | | Eigenvalue | |
|---|---|---|---|---|---|---|---|
| LPA C18:2 | 9 | LPA C18:2 | 483.00 | LPA C18:2 | .317 | LPA C18:2 | .415 |
| PC(36:2) | 7 | TG(52:2) | 454.06 | TG(52:2) | .295 | PAF C16:0 | .378 |
| PAF C16:0 | 7 | PC(36:3) | 300.00 | PAF C16:0 | .279 | LPA C16 | .331 |
| TG(50:2) | 6 | PC(36:2) | 291.50 | PC(36:2) | .276 | LPA C16:1 | .315 |
| TG(52:2) | 6 | Proline | 220.00 | PC(34:1) | .274 | PC(36:2) | .271 |
| PC(34:1) | 6 | TG(50:1) | 202.00 | TG(50:2) | .274 | PC(34:1) | .238 |
| LPA C16 | 6 | PC(34:1) | 180.85 | LPA C16:1 | .271 | LPA C18:1 | .237 |
| LPA C16:1 | 6 | Alanine | 180.00 | LPA C16 | .270 | LPC(16:0) | .235 |

TABLE S2.8. Centrality measures for the AD topology.

| Degree | | Betweenness | | Closeness | | Eigenvalue | |
|---|---|---|---|---|---|---|---|
| LPA C18:2 | 9 | LPA C18:2 | 661.87 | LPA C18:2 | .395 | LPA C18:2 | .353 |
| PC(36:2) | 8 | TG(52:2) | 533.69 | TG(52:2) | .359 | PC(36:2) | .331 |
| TG(52:2) | 7 | Glycylglycine | 328.92 | PC(36:2) | .345 | LPA C20:4 | .293 |
| PC(36:4) | 7 | Tyrosine | 249.87 | Glycylglycine | .336 | PAF C16:0 | .286 |
| PAF C16:0 | 7 | TG(50:1) | 220.17 | PC(36:4) | .331 | LPA C18:1 | .284 |
| Glycylglycine | 6 | PC(36:4) | 194.85 | PAF C16:0 | .327 | PC(36:4) | .270 |
| PC(34:1) | 6 | PC(36:2) | 180.92 | LPA C18:1 | .325 | PC(34:1) | .248 |
| LPA C16 | 6 | L-Lactic acid | 178.38 | PC(34:1) | .325 | LPA C16 | .218 |
| LPA C20:4 | 6 | PC(34:1) | 173.10 | Tyrosine | .322 | LPA C22:6 | .211 |

2.3.6. *Communities.* The nodes in a network often cluster in groups: collections of nodes that are more deeply connected to each other than to nodes outside their topological environment. There is thus interest in the detection of these groups. We approach the question of node-grouping from two angles. The first is the perspective of $k$-cores. A $k$-core of a network is the maximal connected subnetwork in which all nodes have a degree of at least $k$ [S2.12]. In this setting, the $k$-core decomposition of a topology is used as an indication of the core-periphery structure of a network. Figure S2.12 contains the $k$-core decomposition of the SCD and AD topologies

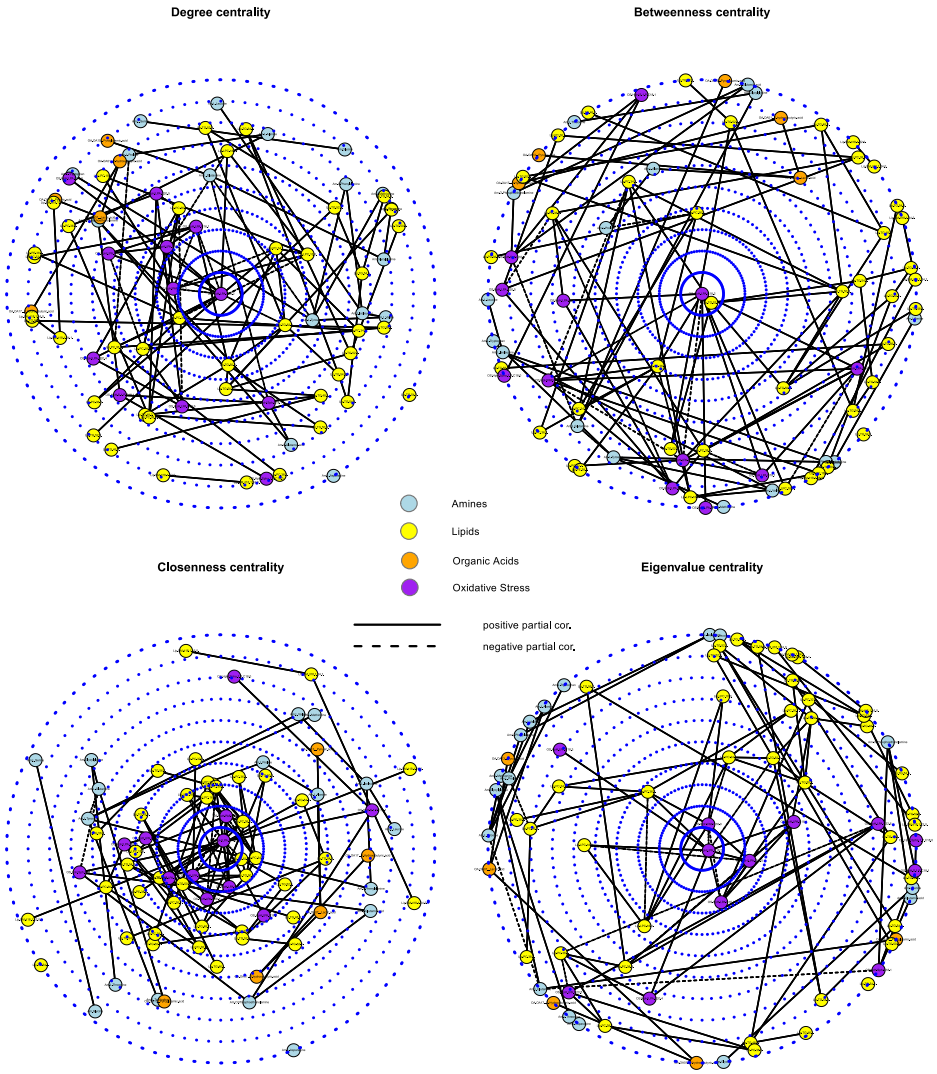


FIGURE S2.10. Target plots [S2.18] visualizing various centralities for the network representing the SCD group. The upper-left panel represents degree centralities. The upper-right panel represents betweenness centralities. The lower-left panel represents closeness centralities. The lower-right panel represents eigenvalue centralities. Note that, for each target plot, the network is the same as in the left-hand panel of Figure S2.7. The topology is now however plotted to represent metabolite features according to various centrality scores. For example, the oxidative stress compound LPA C18:2 has the highest degree centrality and, hence, is depicted in the center of the upper-left panel. The metabolite compounds are again colored according to metabolite family: Blue for amines, yellow for lipids, orange for organic acids, and purple for oxidative stress. Solid edges represent positive partial correlations while dashed edges represent negative partial correlations. The metabolite features attaining the highest centrality scores are given in Table S2.7.

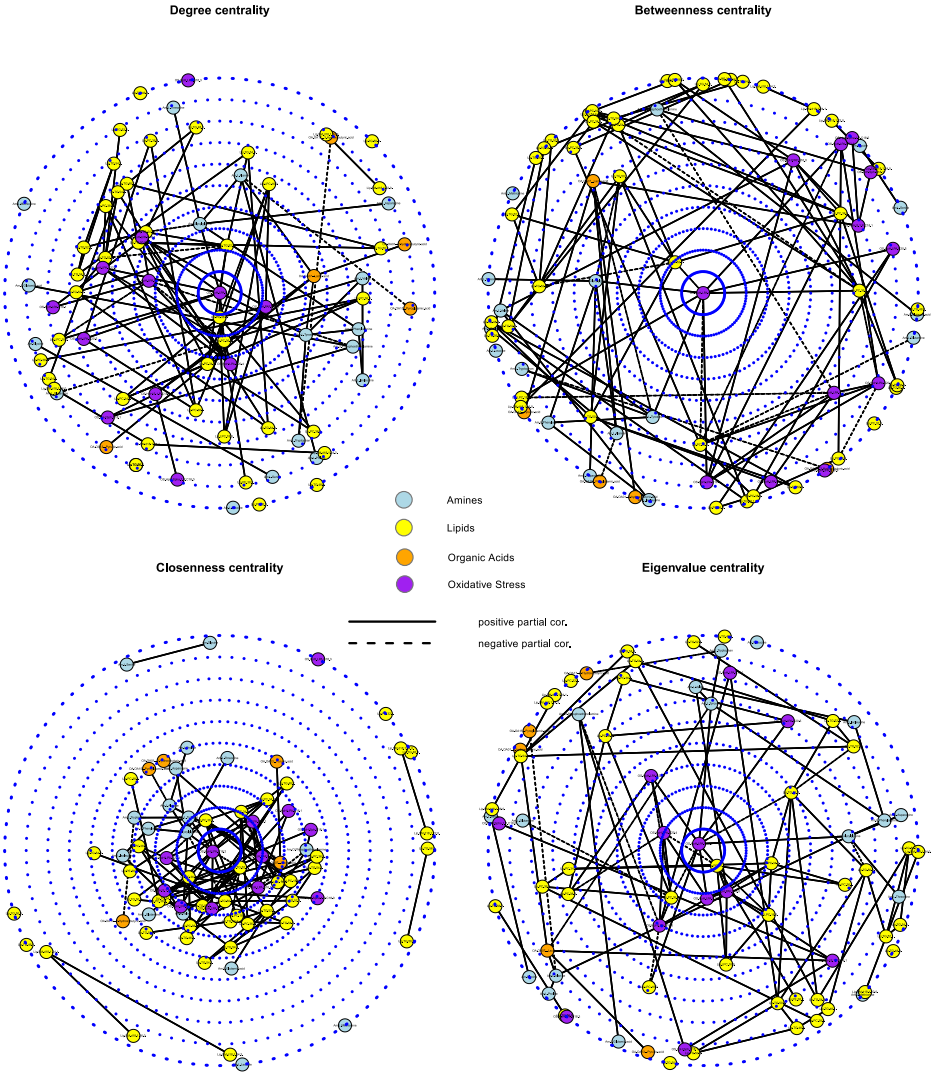


FIGURE S2.11. Target plots [S2.18] visualizing various centralities for the network representing the AD group. The upper-left panel represents degree centralities. The upper-right panel represents betweenness centralities. The lower-left panel represents closeness centralities. The lower-right panel represents eigenvalue centralities. Note that, for each target plot, the network is the same as in the right-hand panel of Figure S2.7. The topology is now however plotted to represent metabolite features according to various centrality scores. For example, the oxidative stress compound LPA C18:2 has the highest degree centrality and, hence, is depicted in the center of the upper-left panel. The metabolite compounds are again colored according to metabolite family: Blue for amines, yellow for lipids, orange for organic acids, and purple for oxidative stress. Solid edges represent positive partial correlations while dashed edges represent negative partial correlations. The metabolite features attaining the highest centrality scores are given in Table S2.8.

depicted in the radial layout of a target plot [S2.18]. For both panels the center represents the 3-core, the first ring of features around the center represents the 2-core, and the subsequent feature-rings represent the 1-core and 0-core, respectively. The $k$-cores of the SCD and AD topologies are similar, although the AD topology places the amine Glycylglycine in the 2-core instead of the 1-core.

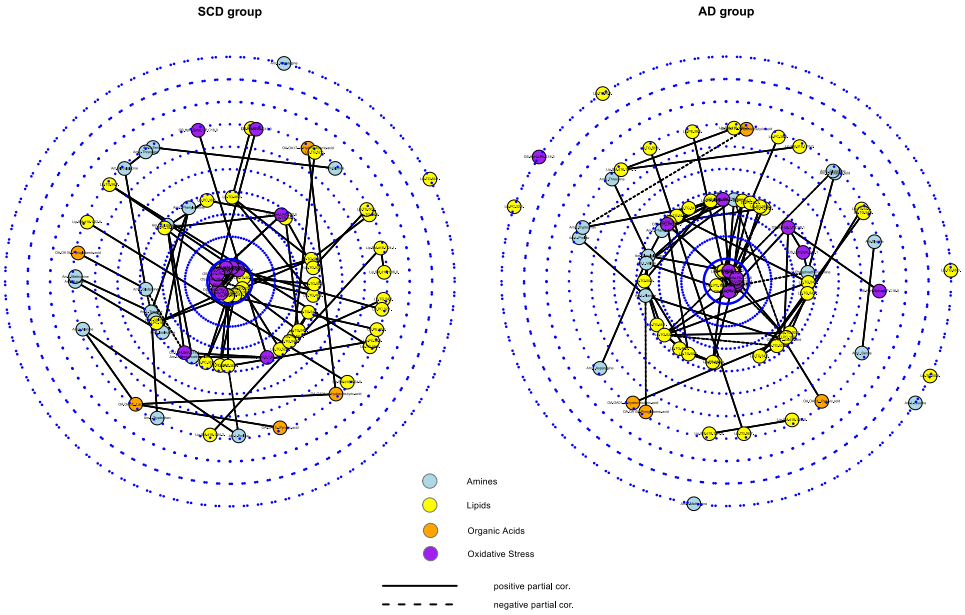


FIGURE S2.12. Target plots [S2.18] with the $k$-core decomposition of the SCD (left-hand panel) and AD (right-hand panel) networks. Note that the SCD and AD networks are the same as depicted in the panels of Figure S2.7. The respective topologies are now however plotted to represent $k$-coreness. The features in the middle of the radial layouts then represent features in the graph-core while features that are plotted further from the center represent the peripheral features. The metabolite compounds are again colored according to metabolite family: Blue for amines, yellow for lipids, orange for organic acids, and purple for oxidative stress. Solid edges represent positive partial correlations while dashed edges represent negative partial correlations.

The second perspective on finding node-groupings is community detection. Community detection, loosely speaking, refers to the "search for naturally occurring groups in a network" [S2.12, p. 371]. A betweenness-based method of community detection is used, commonly known as the Girvan-Newman algorithm [S2.22]. Figure S2.13 contains the same networks as Figure S2.7, but now they are visualized to express the community structure. The colored borders demarcate communities within the respective topologies. Most notably, the SCD topology seems to have 2 loosely connected amine components while the AD topology seems to have a larger and more densely connected amine component that has ties to oxidative stress components.

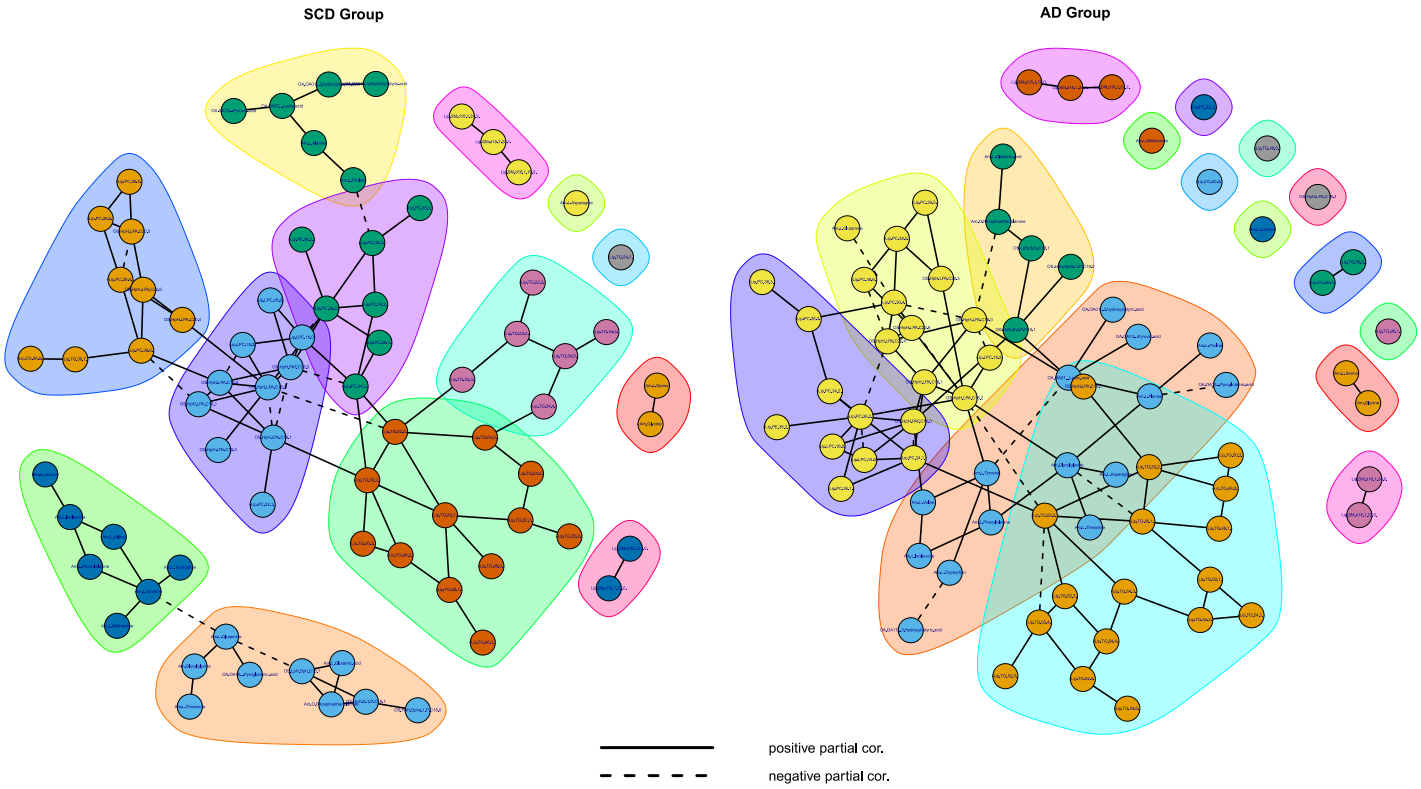


FIGURE S2.13. Class-specific *semi-pruned* networks visualized with the Fruchterman-Reingold algorithm and their community structure. The left-hand panel contains the network for the SCD group. The right-hand panel contains the network for the AD group. Solid edges represent positive partial correlations while dashed edges represent negative partial correlations.

2.3.7. *Differential Graphs.* Section 2.3.5 indicated that certain metabolites act as hubs in both the SCD and AD topologies. Section 2.3.6 then indicated that these hubs belong to different communities in the SCD and AD topologies. Thus, while the SCD and AD topologies may contain the same hub-compounds, these compounds may be connected very differently, implying differential regulatory functioning in the respective networks. Hence, we take interest in the networks of shared and differential connections over the SCD and AD topologies. Figure S2.14 depicts in the left-hand panel the connections that are shared between the SCD and AD networks, and in the right-hand panel the connections that are unique to either the SCD or the AD network. This figure is accompanied by Table S2.9, which contains the top degrees for the differential network (right-hand panel of Figure S2.14). These degrees indicate the compounds that are most differentially wired between the SCD and AD topologies. We see that the regulatory functioning (in terms of connections) of the oxidative stress compounds LPA C18:2 and PAF C16:0 – central to both the SCD and AD topologies – is different across the SCD and AD topologies. In addition, we see compounds that, although not central in either the SCD or AD topologies, are central in the differential network, such as the Phosphatidylcholine PC(38:6). Moreover, we see compounds whose wiring seems to be unique to either the SCD or AD topologies. For example, the amine Glutamine seems to be connected in the SCD topology mostly, while the amine Glycylglycine seems mostly unique to the AD topology, in which it connects with other amines and oxidative stress compounds. Overall, the most differentially wired metabolites across the SCD and AD topologies belong predominantly to the Lyso-phosphatidic acid oxidative stress compounds, the Phosphatidylcholines, and the amine family.

TABLE S2.9. Most differentially wired metabolic features in the differential network for the SCD group versus the AD group.

| Feature | Compound class | Degree |
|---|---|---|
| LPA C18:2 | Oxidative stress: Lyso-phosphatidic acid | 10 |
| Glycylglycine | Amines | 6 |
| PC(36:4) | Lipids: Phosphatidylcholine | 6 |
| PC(38:6) | Lipids: Phosphatidylcholine | 6 |
| LPA C16 | Oxidative stress: Lyso-phosphatidic acid | 6 |
| PAF C16:0 | Platelet activating factor | 6 |
| Glutamine | Amines | 5 |
| LPA C18:1 | Oxidative stress: Lyso-phosphatidic acid | 5 |
| LPA C20:4 | Oxidative stress: Lyso-phosphatidic acid | 5 |
| Tyrosine | Amines | 4 |
| LPA C16:1 | Oxidative stress: Lyso-phosphatidic acid | 4 |

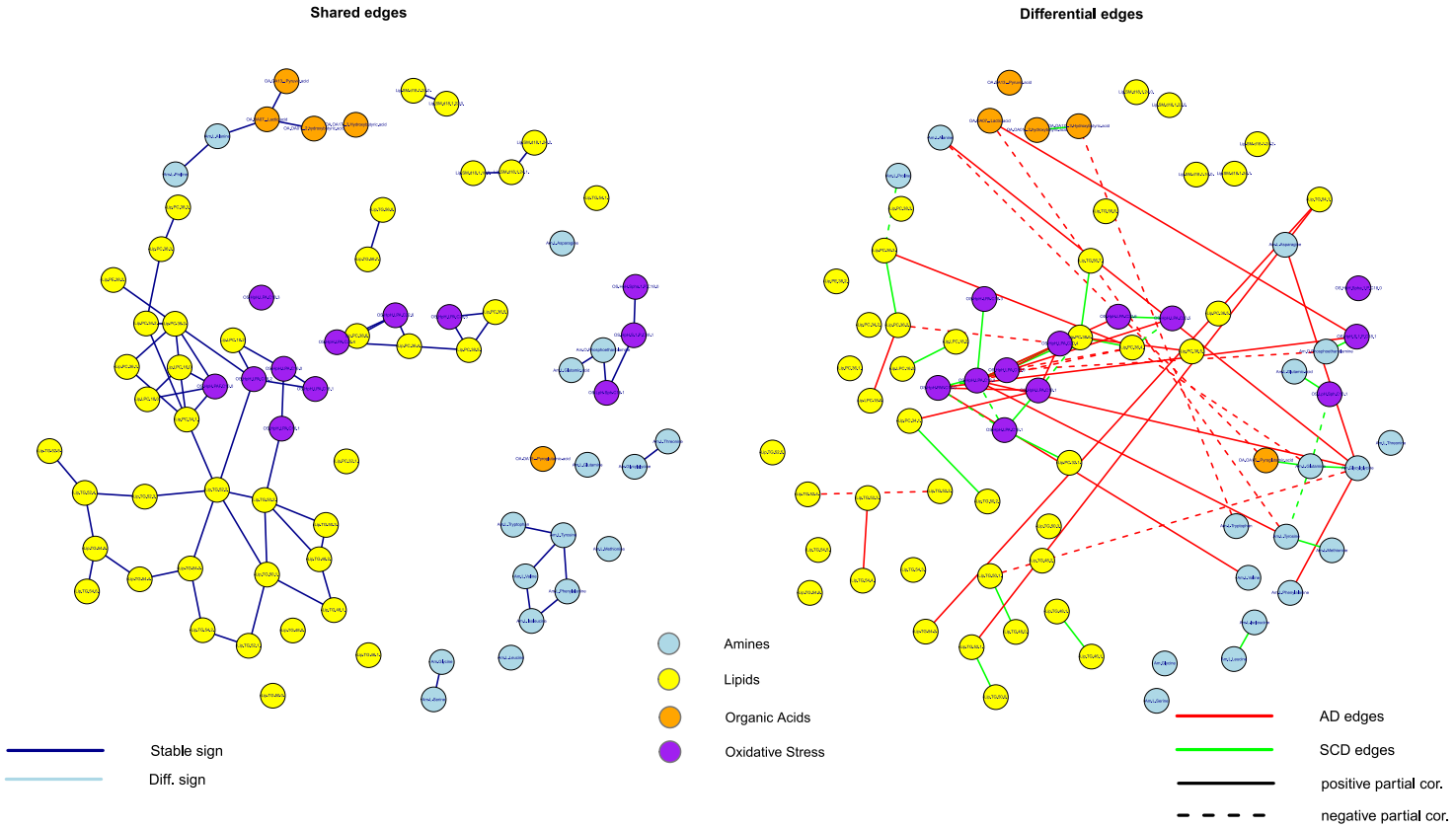


FIGURE S2.14. Common and differential networks for the SCD versus AD class. The left-hand panel contains the network consisting of the edges (solid and colored blue) that are shared between the SCD and AD groups. The right-hand panel contains the network consisting of the edges that are unique for either the SCD or the AD group. Red edges represent connections that are present in the AD group only. Green edges represent connections that are present in the SCD group only. Solid edges represent positive partial correlations while dashed edges represent negative partial correlations. The metabolite compounds are colored according to metabolite family: Blue for amines, yellow for lipids, orange for organic acids, and purple for oxidative stress. Note that the nodes in these networks have coordinates concordant with the node-placing of Figure S2.7.

## 2.4. **Regulatory Signature: Including APOE $\epsilon 4$ allele status.**

2.4.1. *Approach.* In our graphical modeling approach we take into account that our data consist of distinct classes of interest. In the preceding section we looked at the first natural distinction: AD versus SCD. The second distinction is having no APOE $\epsilon 4$ allele versus having at least 1 APOE $\epsilon 4$ allele. The APOE $\epsilon 4$ indicator proved influential in the classification signature. Moreover, having at least 1 APOE $\epsilon 4$ allele is strongly associated with the AD disease label: The Fisher exact test on Table S2.10 sorts a $p$-value of 1.553e-10, indicating that persons with AD are more likely to have at least 1 APOE $\epsilon 4$ allele. One may expect that AD with and AD without the APOE $\epsilon 4$ allele represent two (somewhat) distinct disease processes. Moreover, one could expect that for a portion of the SCD cases that have at least 1 APOE $\epsilon 4$ allele, certain metabolic changes indicative of looming AD may already be present. In this situation we thus have four classes or groups of interest (Table S2.10 gives the number of observations for each group).

TABLE S2.10. Number of observations in the cross-tabulation of AD disease status and APOE $\epsilon 4$ allele status.

| | At least 1 APOE $\epsilon 4$ allele | |
|---|---|---|
| | No | Yes |
| AD | 40 | 87 |
| SCD | 87 | 34 |

The graphical modeling approach is analogous to the strategy described in Section 2.3.2, but now we have $g = 1, \ldots, 4$. We solve (S2.4) using the class-specific sample covariance matrices (i.e., the sample covariance matrices of the class-specific data) as the data entries. For the target $\mathbf{T}$ we again choose the (weakly informative) $p$-dimensional identity matrix $\mathbf{I}_p$. The optimal penalty parameters were again determined by LOOCV [S2.7]. The optimal penalty values were found to be $\lambda^* = 10.02109$, and $\lambda_f^* = 3.970277\text{e-}17$. These penalty values again emphasize individual regularization over retainment of entry-wise similarities, indicating strong differences in class-specific precision matrices. For each class-specific matrix, the 100 strongest edges (in terms of absolute partial correlations) were retained. The retained partial correlations range, in absolute value, from .1376073 to .5791377 over the respective classes. All analyzes were again performed with the `rags2ridges` package [S2.8] in `R` [S2.9].

2.4.2. *Results.* Here, we state all results for the the group-specific networks stemming from the cross-tabulation of AD disease status APOE $\epsilon 4$ allele status. For detail on technical terms, see Section 2.3.

Figure S2.15 contains the class-specific pruned networks visualized with the FR algorithm [S2.10]. Figures S2.16 and S2.17 then contain the class-specific networks over the union of retained metabolites in which the FR-based coordinates of the network of the SCD group with no APOE $\epsilon 4$ allele serve as reference coordinates. As stated, the FR algorithm prefers coiled structures. From this perspective the network for the SCD group with no APOE $\epsilon 4$ allele ($\text{SCD}^{\neg\epsilon 4}$) and the network for the AD group with at least 1 APOE $\epsilon 4$ allele ($\text{AD}^{\epsilon 4}$) seem the most structured ones. Perhaps this is natural for the former group, given that these persons are,

in some sense, the least at risk for developing AD, and thus should represent the normal biochemical state. This might also be natural for the latter group, as this network may represent structured APOE $\epsilon 4$-driven changes in metabolism. The networks for the SCD group with at least 1 APOE $\epsilon 4$ allele ($SCD^{\epsilon 4}$) and the AD group with no APOE $\epsilon 4$ allele ($AD^{\neg \epsilon 4}$) appear more random, i.e., less modular or locally connected. As stated, one could expect that for a portion of the $SCD^{\epsilon 4}$ group, certain metabolic changes indicative of looming AD may already be present. Also, the $AD^{\neg \epsilon 4}$ group may represent various alternative AD disease processes. In short, both these latter groups are likely heterogeneous. Table S2.11 then contains some global characteristics of each class-related graph as given in Figures S2.16 and S2.17. These metrics corroborate to some degree the assessment made above: The topology for the $SCD^{\neg \epsilon 4}$ group is most strongly locally connected, and the topologies for the $SCD^{\neg \epsilon 4}$ group and the $AD^{\epsilon 4}$ group are most cohesive (overall). Figures S2.18 and S2.19 also indicate that the topologies for the $SCD^{\epsilon 4}$ group and the $AD^{\neg \epsilon 4}$ group are more diffuse. Moreover, they indicate that the topology for the $AD^{\epsilon 4}$ group can be characterized (vis-à-vis the $SCD^{\neg \epsilon 4}$ topology) by increased connection density within the amine compounds and between the amine and oxidative stress compounds.

TABLE S2.11. Global metrics for the four topologies of interest.

| Topology | Transitivity | Connectedness | Centralization |
|---|---|---|---|
| SCD group with no APOE $\epsilon 4$ allele | .2136986 | .7134809 | .1055901 |
| SCD group with at least 1 APOE $\epsilon 4$ allele | .142315 | .6201207 | .1496894 |
| AD group with no APOE $\epsilon 4$ allele | .1758621 | .5758551 | .1937888 |
| AD group with at least 1 APOE $\epsilon 4$ allele | .1480519 | .7126761 | .1496894 |

Figures S2.20, S2.21, S2.22, and S2.23 contain target plots depicting the various centrality properties of the metabolite compounds retained in the $SCD^{\neg \epsilon 4}$, $SCD^{\epsilon 4}$, $AD^{\neg \epsilon 4}$, and $AD^{\epsilon 4}$ topologies, respectively. Tables S2.13, S2.14, S2.15, and S2.16 then contain the top centrality metrics for these topologies. The top compounds, in terms of centrality, in the $SCD^{\neg \epsilon 4}$ topology, all belong to either the oxidative stress or lipid family. The $SCD^{\epsilon 4}$ and $AD^{\neg \epsilon 4}$ topologies are indeed, as also indicated in Table S2.11, more centralized, having more compounds with a high degree. The centrality picture for these topologies is more diffuse in terms of compound-families (i.e., all compound families are represented). The top compounds in the $AD^{\epsilon 4}$ topology largely overlap with the to compounds of the $SCD^{\neg \epsilon 4}$ topology. However, $AD^{\epsilon 4}$ topology the amines Glycylglycine and Tyrosine are consistently indicated as central compounds by all centrality measures. Interestingly, these amines are also consistently marked as central in the $SCD^{\epsilon 4}$ topology.

Figure S2.24 depicts the $k$-core decomposition of the topologies of interest. The inner-core of the $SCD^{\neg \epsilon 4}$ topology consists exclusively of lipid and oxidative stress compounds. The inner-core of the $AD^{\epsilon 4}$ topology includes the amines Glycylglycine, Tyrosine, Isoleucine, Threonine, and Valine. The inner-cores of the $SCD^{\epsilon 4}$ and $AD^{\neg \epsilon 4}$ topologies include metabolites from all compound-families and are thus less compound-family centered. Figure S2.25 contains the $SCD^{\neg \epsilon 4}$ and $SCD^{\epsilon 4}$ topologies visualized with their community structure. Figure S2.26 then contains the $AD^{\neg \epsilon 4}$ and $AD^{\epsilon 4}$ topologies visualized with their community structure. The colored borders demarcate communities within the respective topologies. The $SCD^{\neg \epsilon 4}$

and $AD^{\epsilon 4}$ topologies are indeed the most modular ones. The $SCD^{\neg\epsilon 4}$ has clear organic acid, triglyceride, Phosphatidylcholine, and amine communities. The Lyso-phosphatidic acids (oxidative stress compounds) largely form a community with the Phosphatidylcholines. It contains 2 loosely connected amine communities. The $AD^{\epsilon 4}$ topology, on the other hand, seems to have larger and more densely connected amine communities that include organic acid compounds and that have ties to oxidative stress compounds.

The $SCD^{\neg\epsilon 4}$ and $AD^{\epsilon 4}$ topologies represent the most structured graphs. Moreover, they represent (relatively) homogenous groups (in terms of AD pathology). In assessing differential graph structures, we thus focus on these two topologies. Figure S2.27 then contains in the left-hand panel the network of shared connections and in the right-hand panel the network of unique connections between the $SCD^{\neg\epsilon 4}$ and $AD^{\epsilon 4}$ groups. Table S2.12 then contains a list of metabolites with the highest degrees in the differential network (right-hand panel of S2.27). These differential degrees indicate the metabolites that change their regulatory function (in terms of differential connections) the most between the $SCD^{\neg\epsilon 4}$ and $AD^{\epsilon 4}$ groups. Overall, the most differentially wired metabolites across their topologies belong exclusively to the oxidative stress and amine compound-families. The oxidative stress compound LPA C18:2 is well-connected in both the $SCD^{\neg\epsilon 4}$ and $AD^{\epsilon 4}$ topologies, but is wired very differently between them. From this perspective LPA C18:2 is implied in the loss of normal and the gain of abnormal connections in the AD state driven by APOE $\epsilon 4$. The oxidative stress compound PAF C16:0 seems to be well-connected in the $SCD^{\neg\epsilon 4}$ group mostly. Hence, this compound is implied in the loss of normal regulatory connections in the AD state driven by APOE $\epsilon 4$. The amine Glycylglycine seems to be well-connected in the $AD^{\epsilon 4}$ group mostly. This compound is thus implied in the gain of abnormal regulatory connections in the AD state driven by APOE $\epsilon 4$. These connections are amongst amines predominantly. This latter observation also, to a lesser degree, holds for the amines Tyrosine and Glutamine.

TABLE S2.12. Most differentially wired metabolic features in the differential network for the SCD group with no APOE $\epsilon 4$ allele versus the AD group with at least 1 APOE $\epsilon 4$ allele.

| Feature | Compound class | Degree |
|---|---|---|
| LPA C18:2 | Oxidative stress: Lyso-phosphatidic acid | 13 |
| Glycylglycine | Amines | 11 |
| Tyrosine | Amines | 8 |
| PAF C16:0 | Oxidative stress: Platelet activating factor | 8 |
| Glutamine | Amines | 7 |

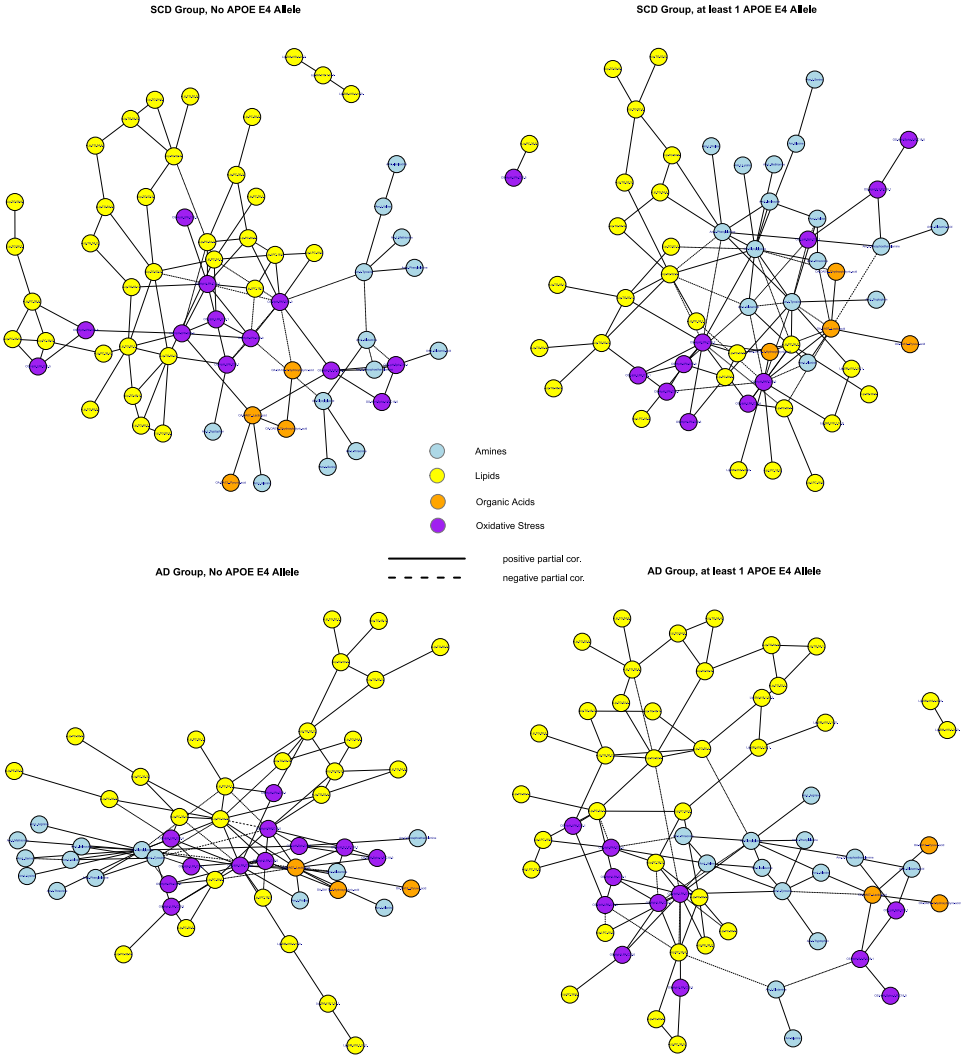


FIGURE S2.15. Class-specific *pruned* networks visualized with the Fruchterman-Reingold algorithm. The upper-left panel contains the network for the SCD group with no APOE $\epsilon 4$ allele. The upper-right panel contains the network for the SCD group with at least 1 APOE $\epsilon 4$ allele. The lower-left panel represents the network for the AD group with no APOE $\epsilon 4$ allele. The lower-right panel represents the network for the AD group with at least 1 APOE $\epsilon 4$ allele. The metabolite compounds are colored according to metabolite family: Blue for amines, yellow for lipids, orange for organic acids, and purple for oxidative stress. Solid edges represent positive partial correlations while dashed edges represent negative partial correlations. Appears as Figure 3 in the main text.

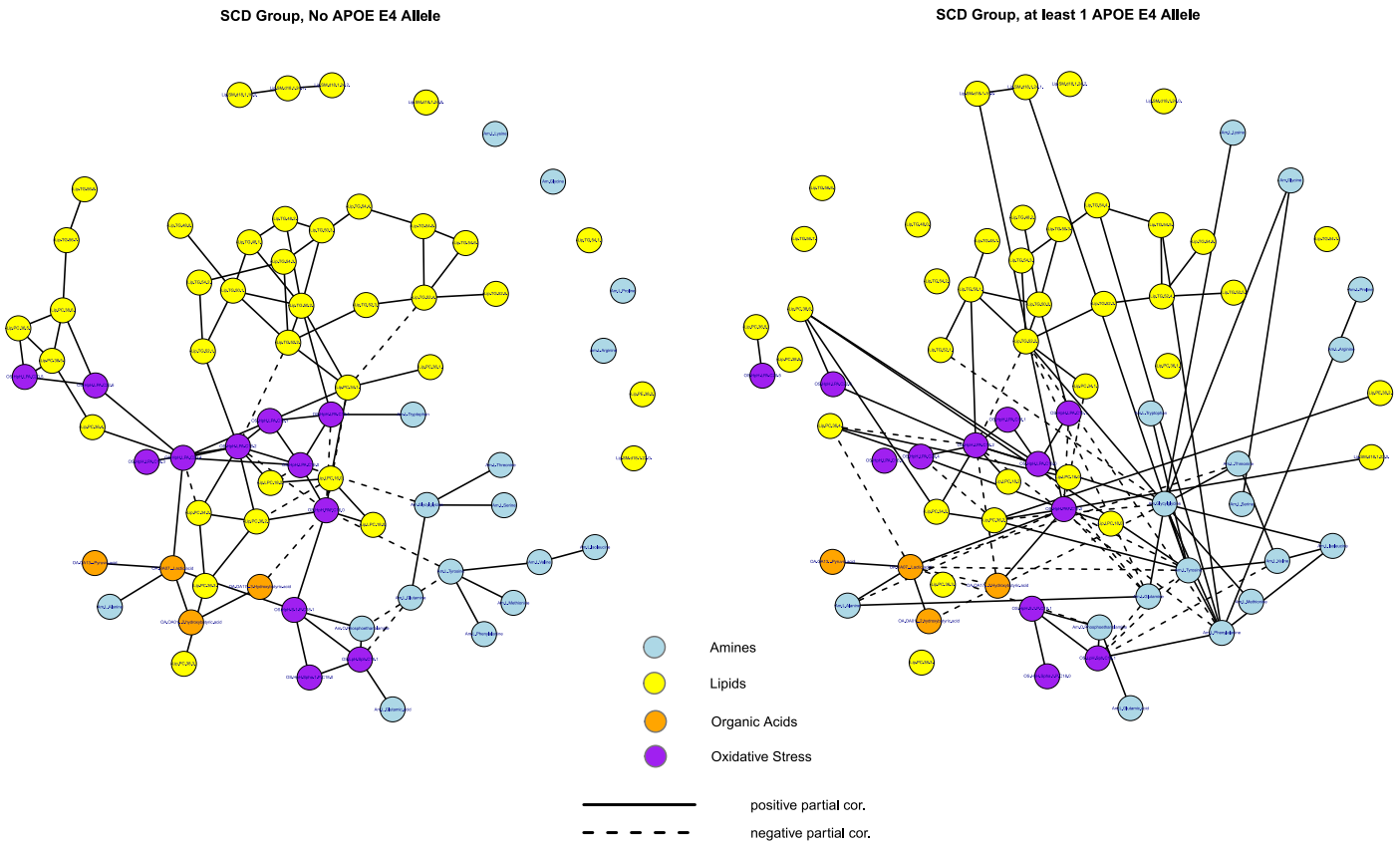


FIGURE S2.16. Class-specific *semi-pruned* networks visualized with the Fruchterman-Reingold algorithm. The left-hand panel contains the network for the SCD group with no APOE $\epsilon 4$ allele. The right-hand panel contains the network for the SCD group with at least 1 APOE $\epsilon 4$ allele. The coordinates of the left-hand topology serve as the reference coordinates. The metabolite compounds are colored according to metabolite family: Blue for amines, yellow for lipids, orange for organic acids, and purple for oxidative stress. Solid edges represent positive partial correlations while dashed edges represent negative partial correlations.

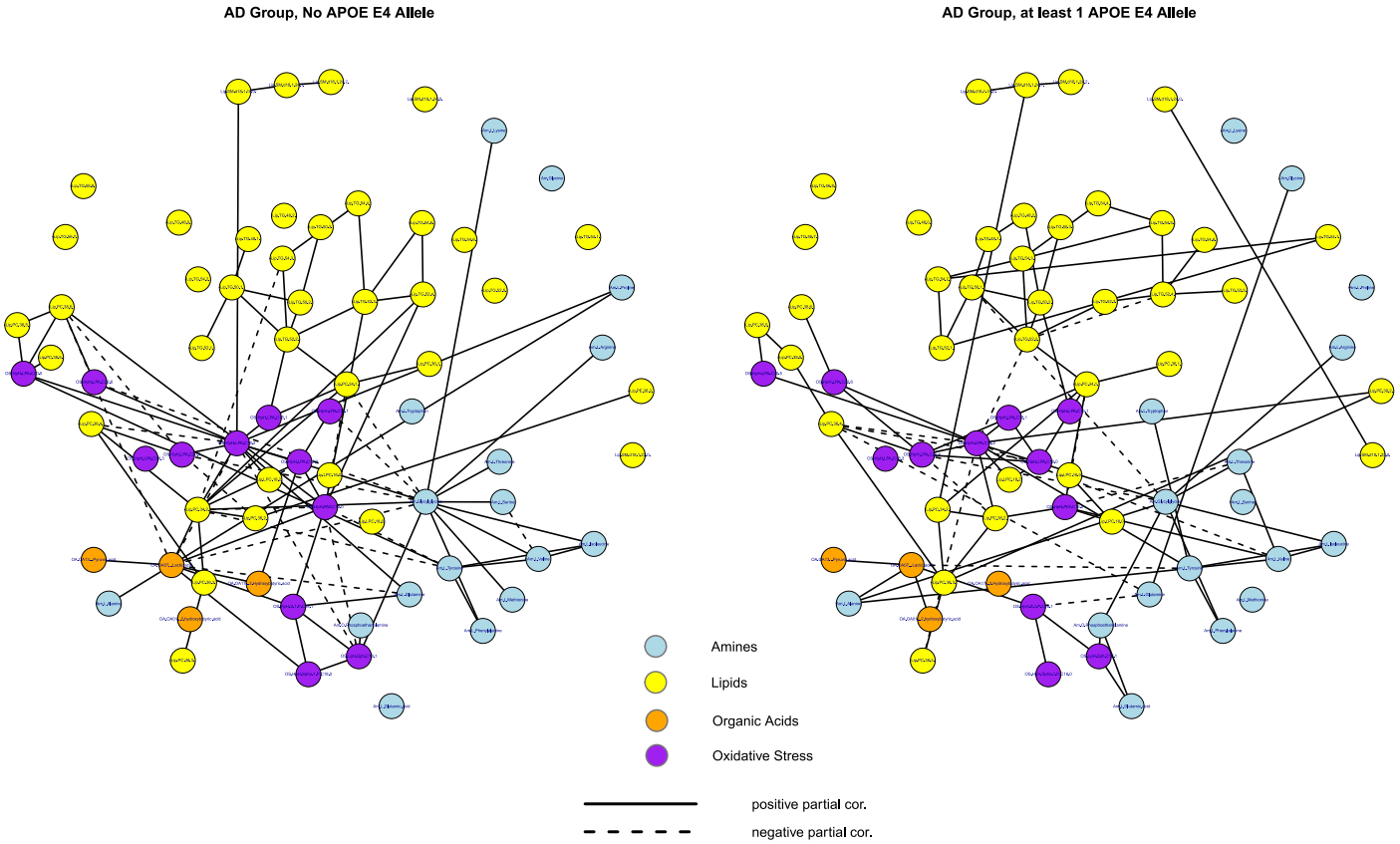


FIGURE S2.17. Class-specific *semi-pruned* networks visualized with the Fruchterman-Reingold algorithm. The left-hand panel contains the network for the AD group with no APOE $\epsilon 4$ allele. The right-hand panel contains the network for the AD group with at least 1 APOE $\epsilon 4$ allele. The coordinates of the left-hand topology in Figure S2.16 serve as the reference coordinates. The metabolite compounds are colored according to metabolite family: Blue for amines, yellow for lipids, orange for organic acids, and purple for oxidative stress. Solid edges represent positive partial correlations while dashed edges represent negative partial correlations.

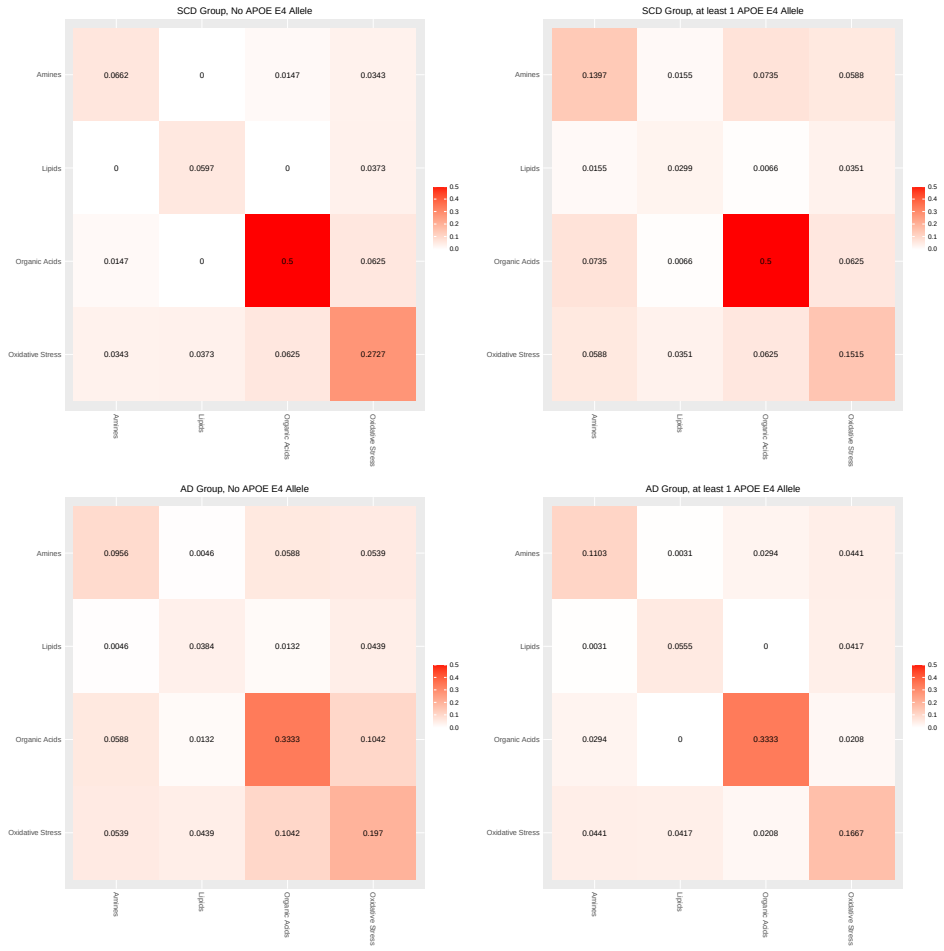


FIGURE S2.18. Heatmaps of degree densities for the class-specific networks. The reported numbers represent the degree density for the (combinations of) metabolite groups. Degree density represents the number of connections (edges) divided by the number of possible connections. For example, in the network for the SCD group with no APOE $\epsilon 4$ allele the proportion of actual edges relative to the number of possible edges between Amines and oxidative stress compounds is .0343. Note that all heatmaps received the same color key. Hence, the color intensities (i.e., the color-spectrum representations of the cell-numbers) are comparable over the respective heatmaps.

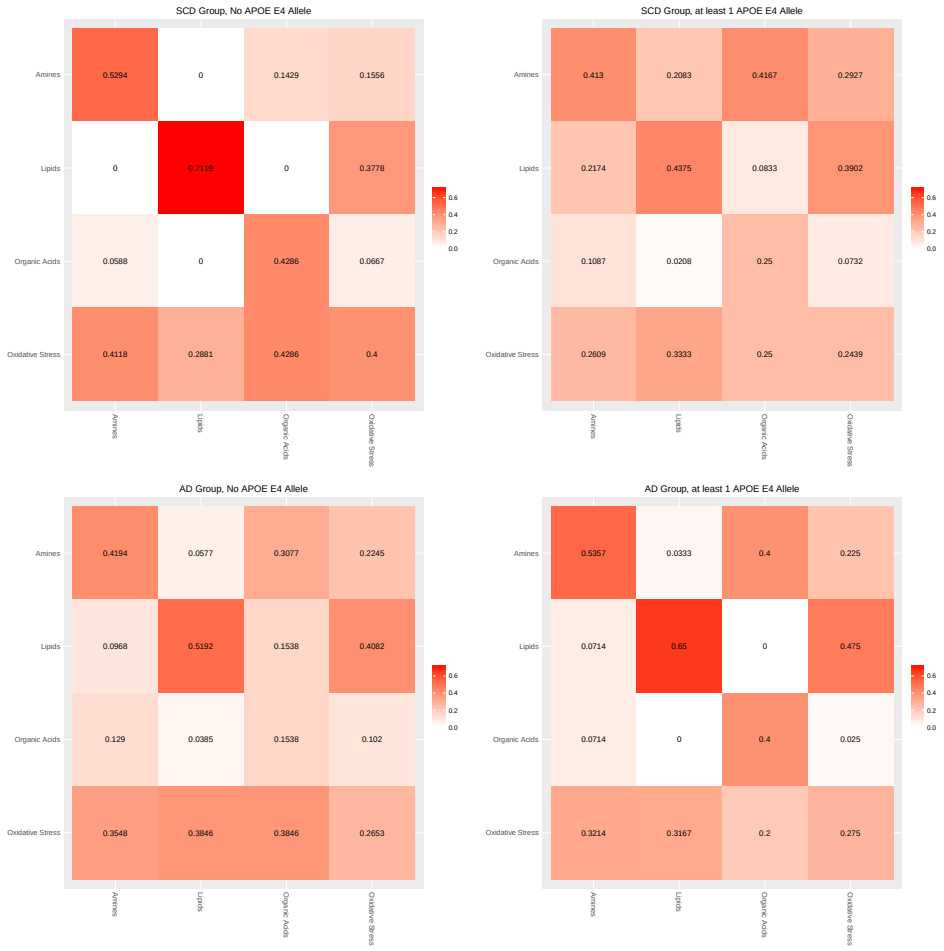


FIGURE S2.19. Heatmaps of relative outdegrees for the class-specific networks. The reported numbers represent the relative outdegree for the (combinations of) metabolite groups. The relative outdegree represents the number of connections (edges) between two metabolite groups divided by the number of 'outgoing' connections for one of these groups. For example, in the network for the SCD group with no APOE $\epsilon 4$ allele the number of edges between lipid and oxidative stress compounds accounts for approximately 29% of all edges involving lipids and approximately 38% of all edges involving oxidative stress compounds. Note that all heatmaps received the same color key. Hence, the color intensities (i.e., the color-spectrum representations of the cell-numbers) are comparable over the respective heatmaps. Note that the column numbers sum to unity.

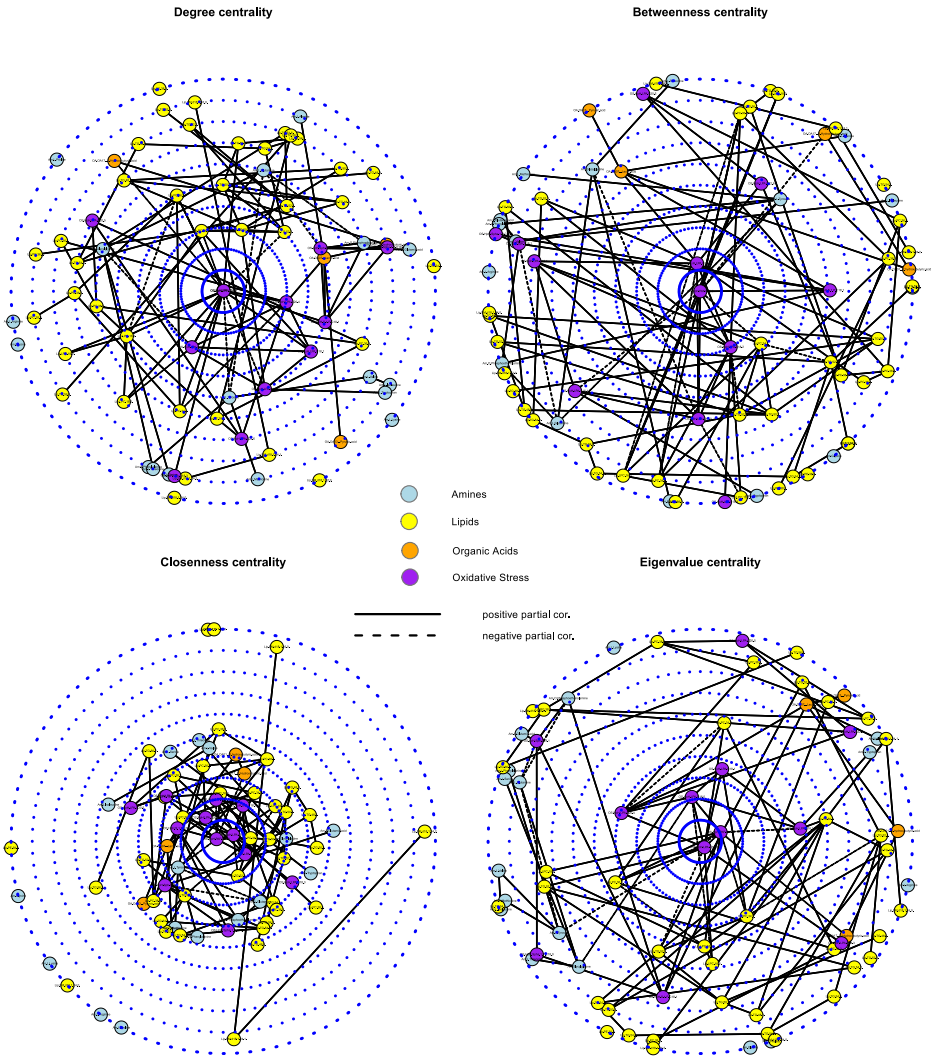


FIGURE S2.20. Target plots visualizing various centralities for the network representing the SCD group with no APOE $\epsilon 4$ allele. The upper-left panel represents degree centralities. The upper-right panel represents betweenness centralities. The lower-left panel represents closeness centralities. The lower-right panel represents eigenvalue centralities. Note that, for each target plot, the network is the same as in the left-hand panel of S2.16. The topology is now however plotted to represent metabolite features according to various centrality scores. For example, the oxidative stress compounds LPA.C18.2 and PAF.C16.0 have the highest degree centrality and, hence, are depicted in the center of the upper-left panel. The metabolite compounds are again colored according to metabolite family: Blue for amines, yellow for lipids, orange for organic acids, and purple for oxidative stress. Solid edges represent positive partial correlations while dashed edges represent negative partial correlations. The metabolite features attaining the highest centrality scores are given in Table S2.13.

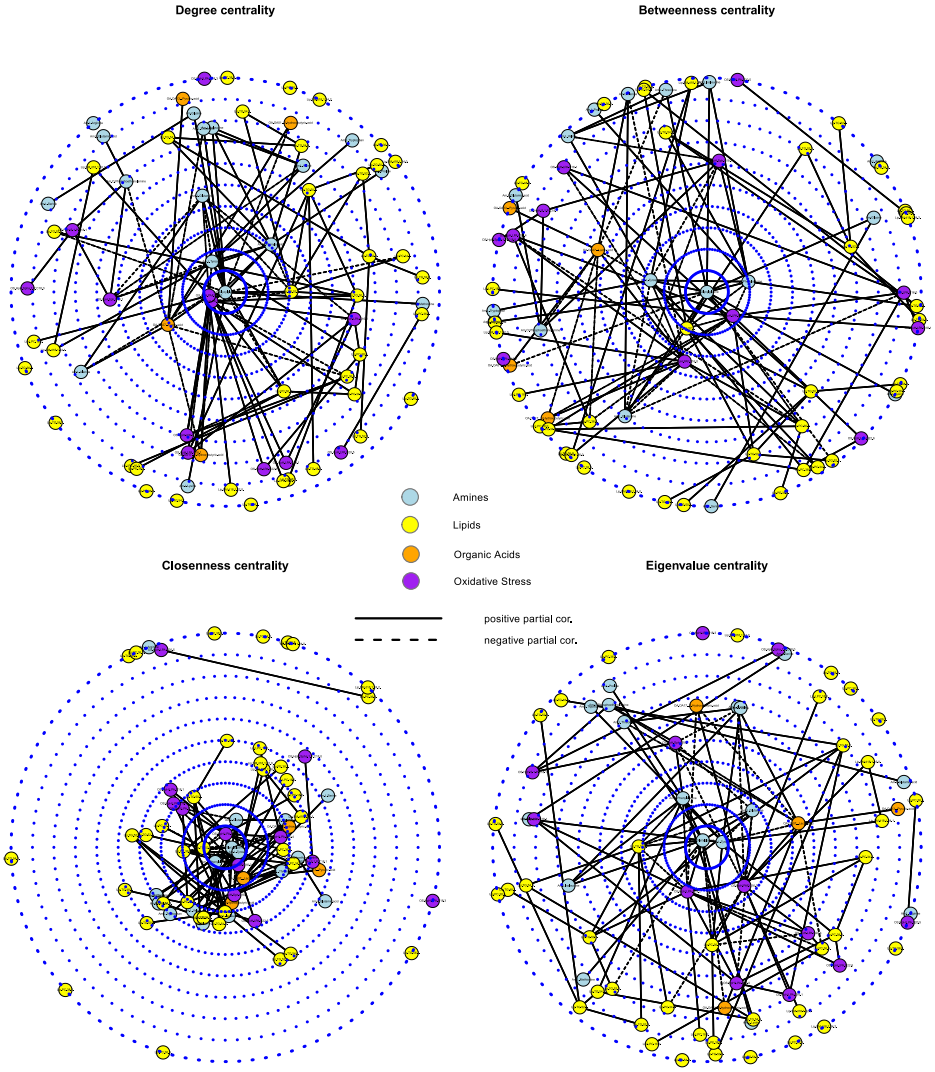


FIGURE S2.21. Target plots visualizing various centralities for the network representing the SCD group with at least 1 APOE $\epsilon 4$ allele. The upper-left panel represents degree centralities. The upper-right panel represents betweenness centralities. The lower-left panel represents closeness centralities. The lower-right panel represents eigenvalue centralities. Note that, for each target plot, the network is the same as in the right-hand panel of S2.16. The topology is now however plotted to represent metabolite features according to various centrality scores. For example, the Amine Glycylglycine has the highest degree centrality and, hence, it is depicted in the center of the upper-left panel. The metabolite compounds are again colored according to metabolite family: Blue for amines, yellow for lipids, orange for organic acids, and purple for oxidative stress. Solid edges represent positive partial correlations while dashed edges represent negative partial correlations. The metabolite features attaining the highest centrality scores are given in Table S2.14.

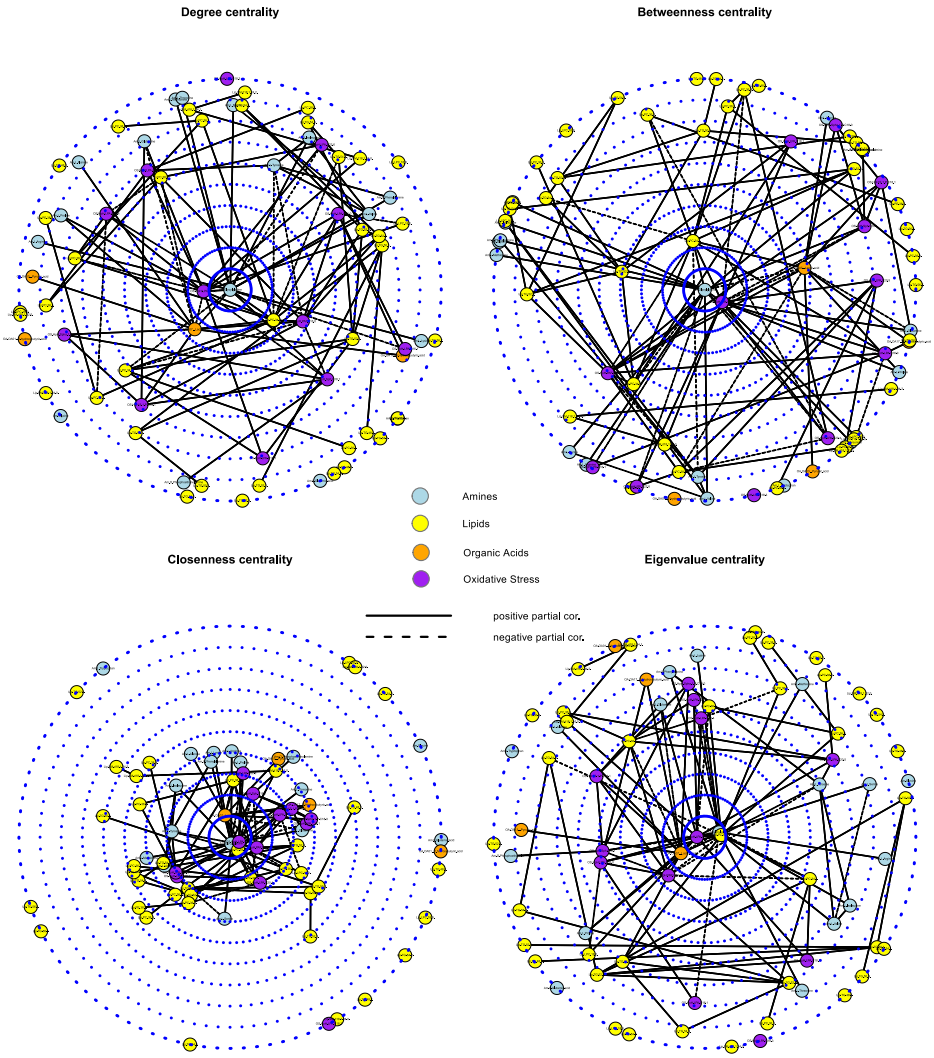


FIGURE S2.22. Target plots visualizing various centralities for the network representing the AD group with no APOE $\epsilon 4$ allele. The upper-left panel represents degree centralities. The upper-right panel represents betweenness centralities. The lower-left panel represents closeness centralities. The lower-right panel represents eigenvalue centralities. Note that, for each target plot, the network is the same as in the left-hand panel of S2.17. The topology is now however plotted to represent metabolite features according to various centrality scores. For example, the Amine Glycylglycine has the highest degree centrality and, hence, it is depicted in the center of the upper-left panel. The metabolite compounds are again colored according to metabolite family: Blue for amines, yellow for lipids, orange for organic acids, and purple for oxidative stress. Solid edges represent positive partial correlations while dashed edges represent negative partial correlations. The metabolite features attaining the highest centrality scores are given in Table S2.15.

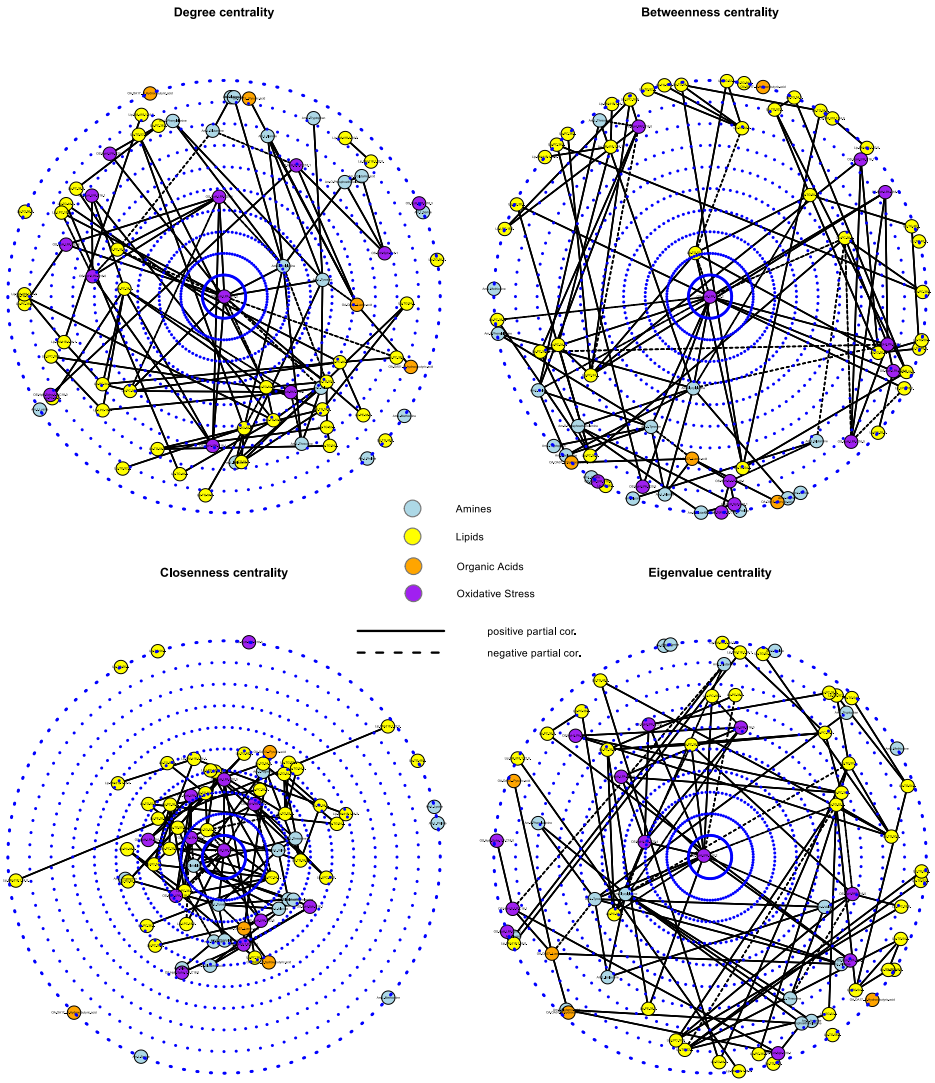


FIGURE S2.23. Target plots visualizing various centralities for the network representing the AD group at least 1 APOE $\epsilon 4$ allele. The upper-left panel represents degree centralities. The upper-right panel represents betweenness centralities. The lower-left panel represents closeness centralities. The lower-right panel represents eigenvalue centralities. Note that, for each target plot, the network is the same as in the right-hand panel of S2.17. The topology is now however plotted to represent metabolite features according to various centrality scores. For example, the oxidative stress compound LPA.C18.2 has the highest degree centrality and, hence, it is depicted in the center of the upper-left panel. The metabolite compounds are again colored according to metabolite family: Blue for amines, yellow for lipids, orange for organic acids, and purple for oxidative stress. Solid edges represent positive partial correlations while dashed edges represent negative partial correlations. The metabolite features attaining the highest centrality scores are given in Table S2.16.

TABLE S2.13. Centrality measures for the SCD group with no APOE $\epsilon 4$ allele.

| Degree | | Betweenness | | Closeness | | Eigenvalue | |
|---|---|---|---|---|---|---|---|
| LPA C18:2 | 10 | PAF C16:0 | 585.19 | PAF C16:0 | 0.416 | LPA C18:2 | .405 |
| PAF C16:0 | 10 | LPA C20:4 | 507.93 | LPA C18:2 | 0.406 | PAF C16:0 | .376 |
| TG(50:2) | 7 | LPA C18:2 | 409.96 | LPA C16 | 0.379 | LPA C16 | .330 |
| PC(34:1) | 7 | PC(34:1) | 363.92 | PC(34:1) | 0.375 | LPA C18:1 | .270 |
| LPA C16 | 7 | Tyrosine | 248.58 | LPA C20:4 | 0.371 | LPA C20:4 | .254 |
| LPA C20:4 | 7 | LPA C22:6 | 243.67 | TG(52:2) | 0.354 | PC(34:1) | .245 |
| TG(52:2) | 6 | LPA C16 | 232.80 | LPA C18:1 | 0.348 | PC(36:2) | .233 |

TABLE S2.14. Centrality measures for the SCD group with at least 1 APOE $\epsilon 4$ allele.

| Degree | | Betweenness | | Closeness | | Eigenvalue | |
|---|---|---|---|---|---|---|---|
| Glycylglycine | 13 | Glycylglycine | 365.75 | Glycylglycine | .443 | Glycylglycine | .360 |
| LPA C18:2 | 12 | LPA C18:2 | 305.64 | LPA C18:2 | .429 | Tyrosine | .346 |
| PAF C16:0 | 12 | TG(52:2) | 294.73 | Tyrosine | .427 | LPA C18:2 | .290 |
| Tyrosine | 11 | Phenylalanine | 292.56 | Phenylalanine | .419 | PAF C16:0 | .281 |
| Phenylalanine | 9 | Tyrosine | 267.47 | TG(52:2) | .411 | Phenylalanine | .279 |
| L-Lactic acid | 9 | PAF C16:0 | 241.17 | PAF C16:0 | .410 | Glutamine | .272 |
| TG(52:2) | 9 | L-Lactic acid | 166.26 | Glutamine | .394 | TG(52:2) | .255 |
| Glutamine | 7 | SPH C18:1 | 140.86 | L-Lactic acid | .381 | L-Lactic acid | .209 |
| PC(36:2) | 6 | TG(50:2) | 120.13 | PC(36:2) | .368 | PC(36:2) | .204 |
| SPH C18:1 | 6 | Glutamine | 111.86 | SPH C18:1 | .354 | SPH C18:1 | .184 |

TABLE S2.15. Centrality measures for the AD group with no APOE $\epsilon 4$ allele.

| Degree | | Betweenness | | Closeness | | Eigenvalue | |
|---|---|---|---|---|---|---|---|
| Glycylglycine | 16 | Glycylglycine | 418.71 | Glycylglycine | .455 | LPA C18:2 | .369 |
| LPA C18:2 | 14 | LPA C18:2 | 376.70 | LPA C18:2 | .446 | Glycylglycine | .356 |
| L-Lactic acid | 12 | PC(34:2) | 319.04 | PC(34:2) | .436 | PC(34:2) | .353 |
| PC(34:2) | 12 | TG(52:2) | 250.23 | L-Lactic acid | .419 | L-Lactic acid | .331 |
| PAF C16:0 | 10 | L-Lactic acid | 216.33 | PAF C16:0 | .405 | PAF C16:0 | .288 |
| Tyrosine | 6 | PC(34:1) | 184.65 | PC(34:1) | .367 | SPH C18:1 | .196 |
| TG(52:2) | 6 | PAF C16:0 | 165.12 | SPH C18:1 | .360 | LPA C16 | .189 |
| PC(34:1) | 6 | TG(50:1) | 103.00 | LPA C16 | .348 | PC(34:1) | .178 |
| PC(38:6) | 6 | SM(d18:1/16:0) | 102.00 | LPA C20:4 | .348 | LPA C20:4 | .167 |
| SPH C18:1 | 6 | LPA C18:1 | 78.31 | PC(38:6) | .342 | PC(38:6) | .161 |
| LPA C16 | 6 | SPH C18:1 | 77.82 | Tyrosine | .339 | S1P C18:1 | .158 |

TABLE S2.16. Centrality measures for the AD group with at least 1 APOE $\epsilon 4$ allele.

| Degree | | Betweenness | | Closeness | | Eigenvalue | |
|---|---|---|---|---|---|---|---|
| LPA C18:2 | 13 | LPA C18:2 | 682.98 | LPA C18:2 | .452 | LPA C18:2 | .473 |
| Glycylglycine | 9 | TG(52:2) | 537.91 | Glycylglycine | .399 | LPA C20:4 | .336 |
| Tyrosine | 7 | Glycylglycine | 388.77 | TG(52:2) | .398 | Glycylglycine | .279 |
| TG(52:2) | 7 | TG(50:1) | 232.91 | LPA C20:4 | .385 | PC(36:4) | .236 |
| PC(34:1) | 7 | PC(34:1) | 231.83 | Tyrosine | .363 | PC(36:2) | .230 |
| LPA C20:4 | 7 | Tyrosine | 231.23 | PC(34:1) | .354 | LPA C18:1 | .218 |
| PC(36:4) | 6 | PC(36:4) | 217.96 | PC(36:4) | .354 | Tyrosine | .211 |
| PAF C16:0 | 6 | PC(34:2) | 179.58 | Valine | .349 | Valine | .206 |

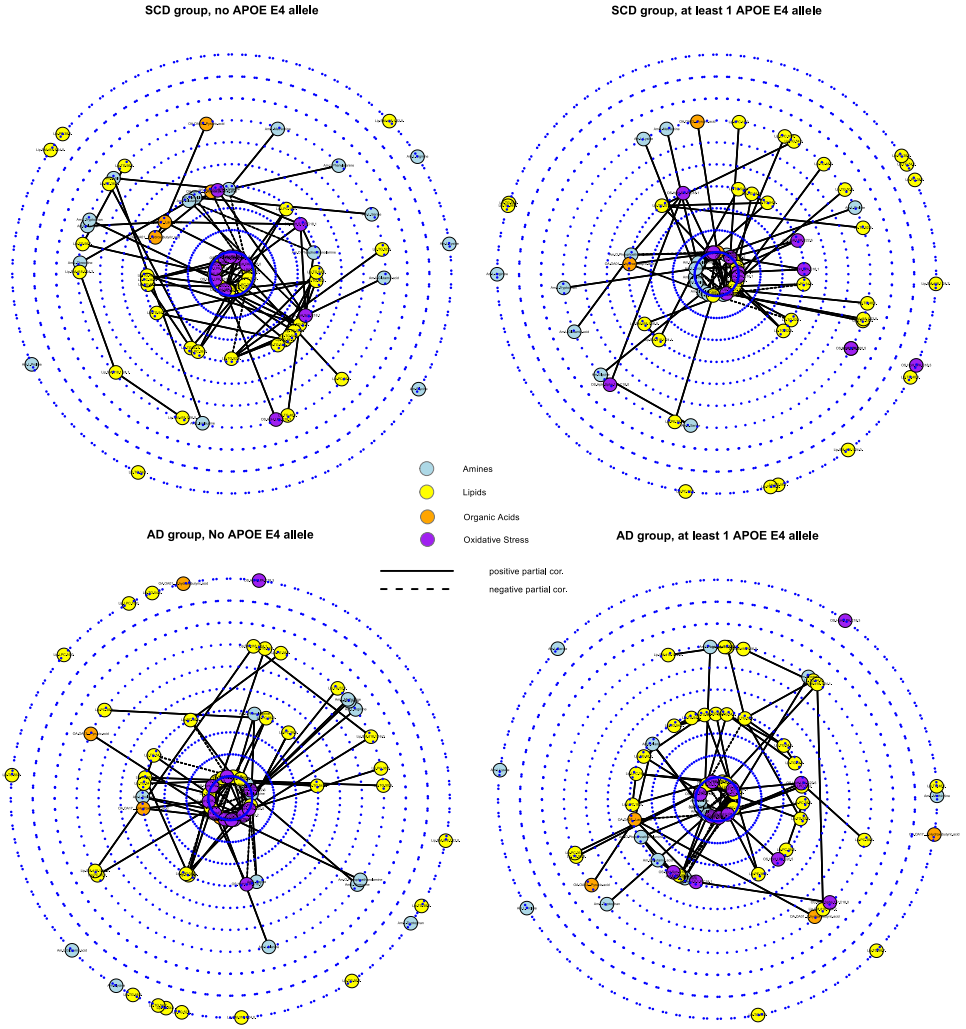


FIGURE S2.24. Target plots [S2.18] depicting $k$-core decompositions. The upper-left panel contains the network for the SCD group with no APOE $\epsilon 4$ allele. The upper-right panel contains the network for the SCD group with at least 1 APOE $\epsilon 4$ allele. The lower-left panel represents the network for the AD group with no APOE $\epsilon 4$ allele. The lower-right panel represents the network for the AD group with at least 1 APOE $\epsilon 4$ allele. Note that the respective topologies are now plotted to represent coreness. The features in the middle of the radial layouts then represent features in the graph-core while features that are plotted further from the center then represent the peripheral features. The metabolite compounds are again colored according to metabolite family: Blue for amines, yellow for lipids, orange for organic acids, and purple for oxidative stress. Solid edges represent positive partial correlations while dashed edges represent negative partial correlations.

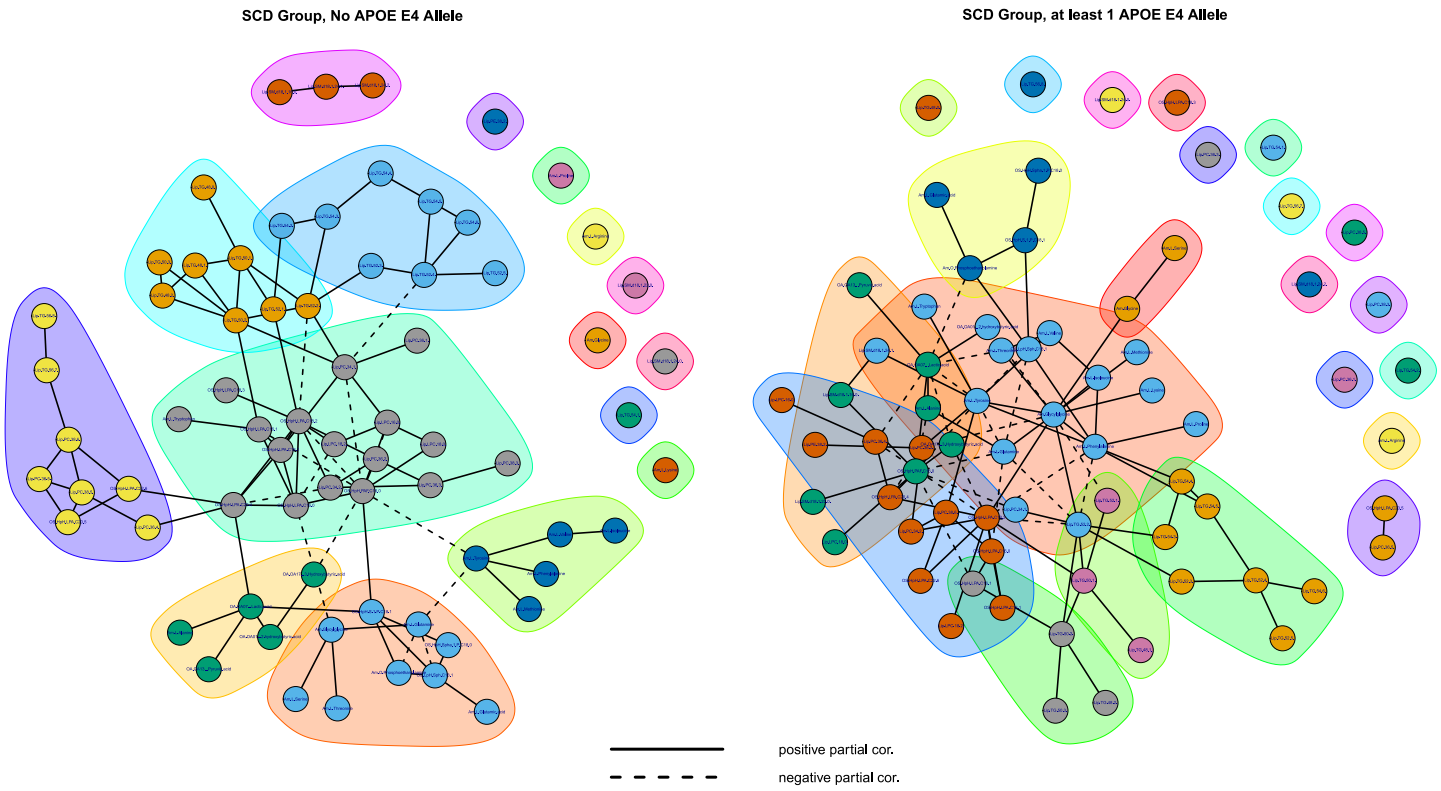


FIGURE S2.25. Class-specific *semi-pruned* networks visualized with their community structure. The left-hand panel contains the network for the SCD group with no APOE $\epsilon 4$ allele. The right-hand panel contains the network for the SCD group with at least 1 APOE $\epsilon 4$ allele. Solid edges represent positive partial correlations while dashed edges represent negative partial correlations.

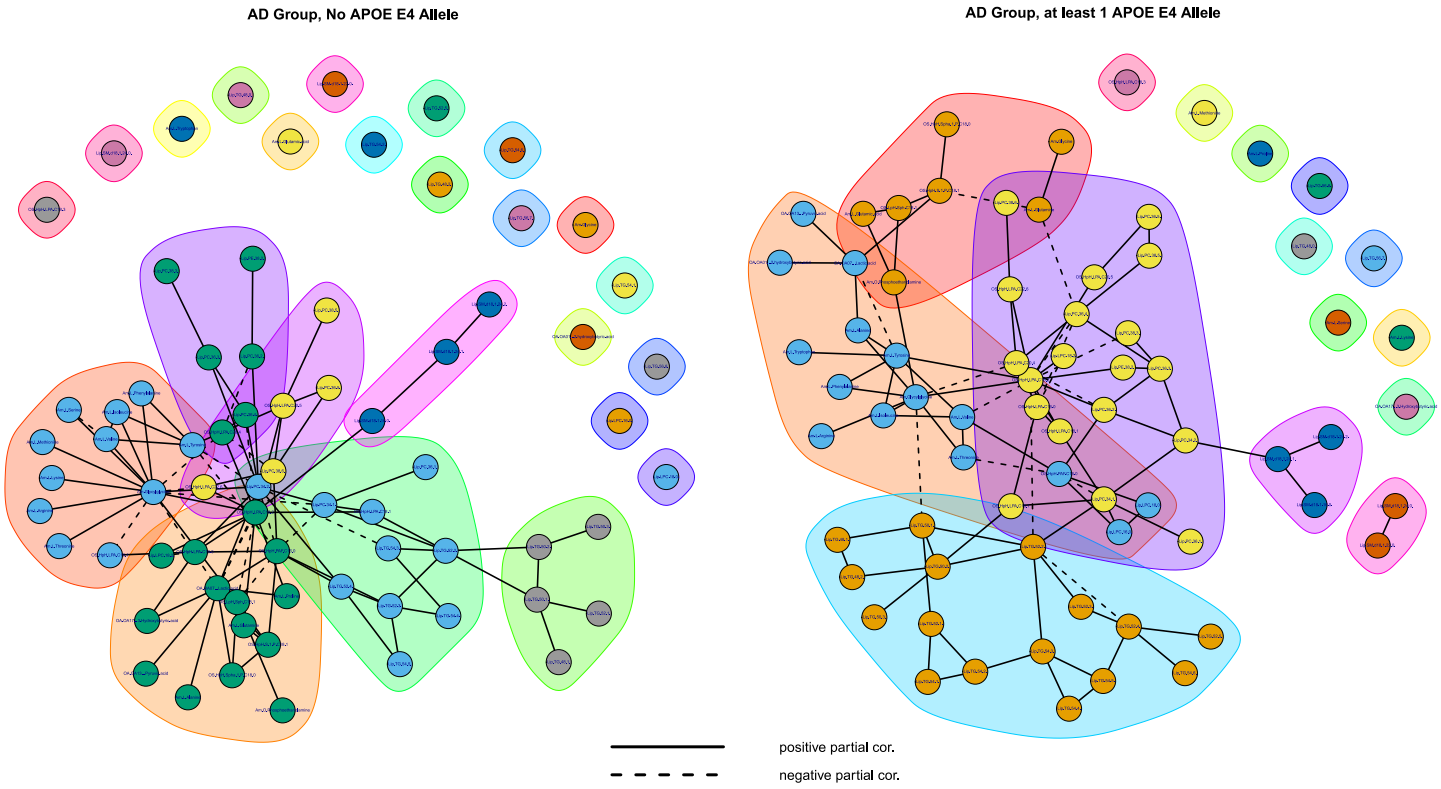


FIGURE S2.26. Class-specific *semi-pruned* networks visualized with their community structure. The left-hand panel contains the network for the AD group with no APOE $\epsilon 4$ allele. The right-hand panel contains the network for the AD group with at least 1 APOE $\epsilon 4$ allele. Solid edges represent positive partial correlations while dashed edges represent negative partial correlations.

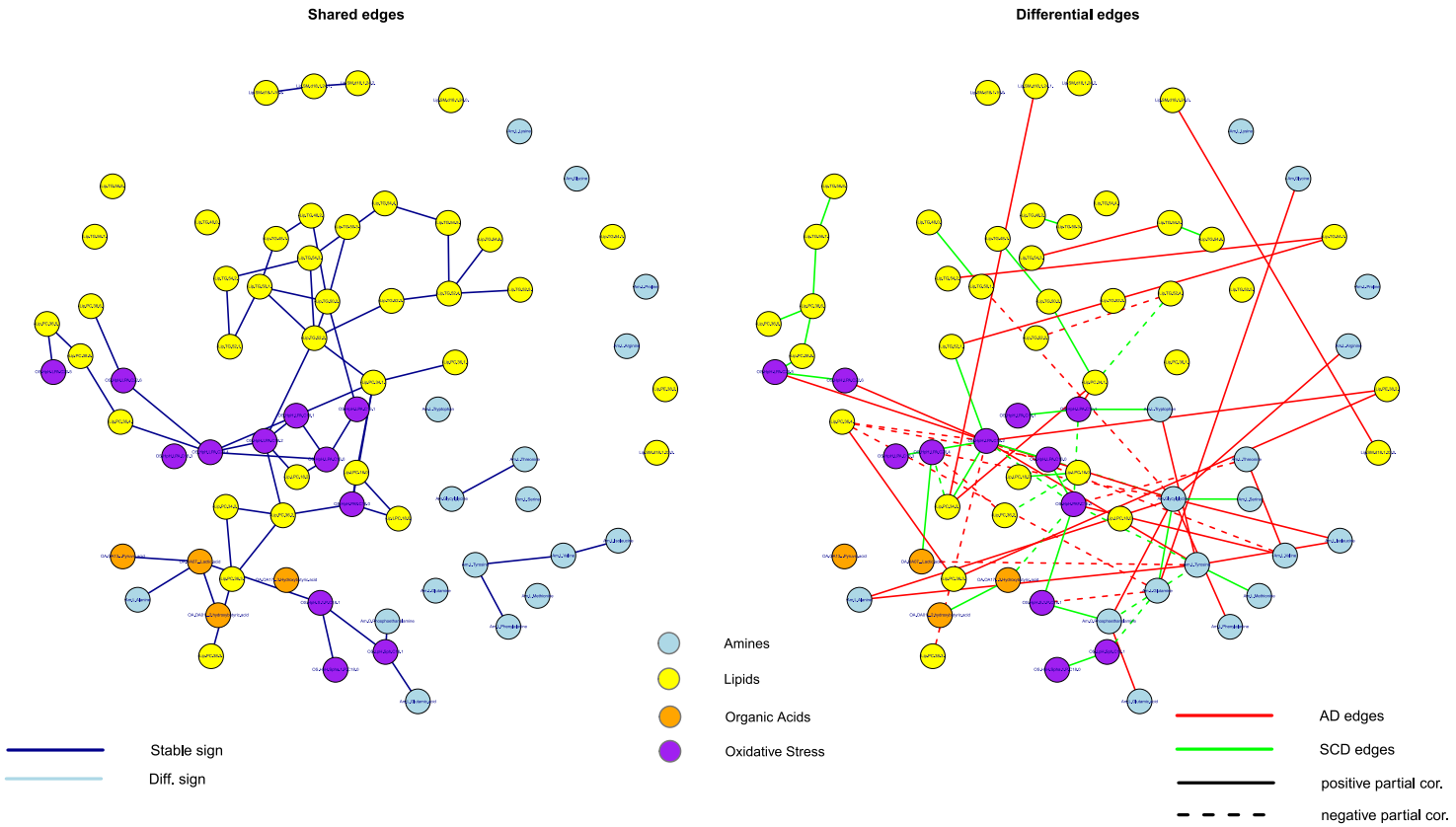


FIGURE S2.27. Common and differential networks for the SCD group with no APOE $\epsilon 4$ allele versus the AD group with at least 1 APOE $\epsilon 4$ allele. The left-hand panel contains the network consisting of the edges (solid and colored blue) that are shared between these groups. The right-hand panel contains the network consisting of the edges that are unique for either of the groups. Red edges represent connections that are present in the AD group with at least 1 APOE $\epsilon 4$ allele only. Green edges represent connections that are present in the SCD group with no APOE $\epsilon 4$ allele only. Solid edges represent positive partial correlations while dashed edges represent negative partial correlations. The metabolite compounds are colored according to metabolite family: Blue for amines, yellow for lipids, orange for organic acids, and purple for oxidative stress. Note that the nodes in these networks have coordinates concordant with the node-placing of Figure S2.16. Appears as Figure 4 in the main text.

## References


[S2.1] van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 45(3).

[S2.2] Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society, Series B, 57: 289–300.

[S2.3] Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society, Series B, 58: 267–288.

[S2.4] Hintze, J.L., & Nelson, R.D. (1998). Violin Plots: A Box Plot-Density Trace Synergism. The American Statistician, 52: 181–184.

[S2.5] van de Wiel, M.A., Lien, T.G., Verlaat, W., van Wieringen, W.N., & Wilting, S.M. (2016). Better prediction by use of co-data: Adaptive group-regularized ridge regression. Statistics in Medicine, 35: 368–381.

[S2.6] Whittaker, J. (1990). Graphical models in applied multivariate statistics. John Wiley & Sons Ltd., Chichester.

[S2.7] Bilgrau, A.E., Peeters, C.F.W., 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].

[S2.8] 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. Available from: `http://cran.r-project.org/packages=rags2ridges`

[S2.9] `R` Development Core Team (2011). `R`: A Language and Environment for Statistical Computing. `R` Foundation for Statistical Computing, Vienna, Austria. `http://www.R-project.org/`. ISBN 3-900051-07-0.

[S2.10] Fruchterman, T.M.J., & Reingold, E.M. (1991). Graph Drawing by Force-Directed Placement. Software: Practice & Experience, 21: 1129–1164.

[S2.11] Villa-Vialaneix, N., Liaubet, L., & SanCristobal, M. (2016). Depicting Gene Co-expression Networks Underlying eQTLs, in: Kadarmideen, H.N. (Ed.). Systems Biology in Animal Production and Health, Vol. 2. Springer International Publishing.

[S2.12] Newman, M.E.J. (2010). Networks: An introduction. Oxford, Oxford University Press.

[S2.13] Freeman, L.C. (1978). Centrality in social networks: Conceptual clarification. Social Networks, 1: 215–239.

[S2.14] Borgatti, S.P., Everett, M.G., & Johnson, J.C. (2013). Analyzing social networks. SAGE Publications Limited.

[S2.15] Krumsiek, J., Suhre, K., Illig, T., Adamski, J., & Theis, F.J. (2011). Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. BMC Systems Biology, 5: a21.

[S2.16] Freeman, L.C. (1977). A set of measures of centrality based on betweenness. Sociometry. 40: 35–41.

[S2.17] Sinclair, P.A. (2009). Network centralization with the Gil Schmidt power centrality index. Social Networks, 31: 214–219.

[S2.18] Brandes, U., Kenis, P., & Wagner, D. (2003). Communicating Centrality in Policy Network Drawings. IEEE Transactions on Visualization and Computer Graphics, 9: 241–253.

[S2.19] Butts, C.T. (2008). Social Network Analysis with `sna`. Journal of Statistical Software, 24: a6.

[S2.20] Csárdi, G., & Nepusz, T. (2006). The `igraph` software package for complex network research, InterJournal Complex Systems: 1695.

[S2.21] Kolaczyk, E.D., & Csárdi, G. (2014). Statistical Analysis of Network Data with `R`. Springer-Verlag New York.

[S2.22] Girvan, M., & Newman, M.E.J. (2002). Community structure in social and biological networks. Proceedings of the National Academy of Sciences of the United States of America, 99: 7821–7826.

(Francisca A. de Leeuw) Alzheimer Center and Dept. of Neurology, Amsterdam Neuroscience, VU University Medical Center Amsterdam, Amsterdam, The Netherlands; and Dept. of Clinical Chemistry, VU University Medical Center Amsterdam, Amsterdam, The Netherlands
*E-mail address*: f.deleeuw@vumc.nl

(Carel F.W. Peeters) Dept. of Epidemiology & Biostatistics, Amsterdam Public Health research institute, VU University medical center Amsterdam, Amsterdam, The Netherlands
*E-mail address*: cf.peeters@vumc.nl

(Maartje I. Kester) Alzheimer Center and Dept. of Neurology, Amsterdam Neuroscience, VU University Medical Center Amsterdam, Amsterdam, The Netherlands
*E-mail address*: m.kester@vumc.nl

(Amy C. Harms) Division of Analytical Biosciences, Leiden Academic Centre for Drug Research, Leiden University, Leiden, The Netherlands
*E-mail address*: a.c.harms@lacdr.leidenuniv.nl

(Eduard A. Struys) Dept. of Clinical Chemistry, VU University medical center Amsterdam, Amsterdam, The Netherlands
*E-mail address*: E.Struys@vumc.nl

(Thomas Hankemeier) Division of Analytical Biosciences, Leiden Academic Centre for Drug Research, Leiden University, Leiden, The Netherlands
*E-mail address*: hankemeier@lacdr.leidenuniv.nl

(Herman W.T. van Vlijmen) Discovery Sciences, Janssen Research and Development, Beerse, Belgium; and Division of Medicinal Chemistry, Leiden Academic Centre for Drug Research, Leiden University, Leiden, The Netherlands
*E-mail address*: hvvlijme@its.jnj.com

(Sven J. van der Lee) Genetic Epidemiology Unit, Dept. of Epidemiology, Erasmus MC, Rotterdam, The Netherlands; and Alzheimer Center, VU University Medical Center Amsterdam, Amsterdam, The Netherlands
*E-mail address*: s.j.vanderlee@vumc.nl

(Cornelia M. van Duijn) Genetic Epidemiology Unit, Dept. of Epidemiology, Erasmus MC, Rotterdam, The Netherlands
*E-mail address*: c.vanduijn@erasmusmc.nl

(Philip Scheltens) Alzheimer Center and Dept. of Neurology, Amsterdam Neuroscience, VU University Medical Center Amsterdam, Amsterdam, The Netherlands
*E-mail address*: p.scheltens@vumc.nl

(Ayşe Demirkan) Genetic Epidemiology Unit, Dept. of Epidemiology, Erasmus MC, Rotterdam, The Netherlands; and Dept. of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands
*E-mail address*: a.demirkan@erasmusmc.nl

(Mark A. van de Wiel) Dept. of Epidemiology & Biostatistics, Amsterdam Public Health research institute, VU University medical center Amsterdam, Amsterdam, The Netherlands; and Dept. of Mathematics, VU University Amsterdam, Amsterdam, The Netherlands
*E-mail address*: mark.vdwiel@vumc.nl

(Wiesje M. van der Flier) Alzheimer Center and Dept. of Neurology, Amsterdam Neuroscience, VU University Medical Center Amsterdam, Amsterdam, The Netherlands; and Dept. of Epidemiology & Biostatistics, Amsterdam Public Health research institute, VU University medical center Amsterdam, Amsterdam, The Netherlands
*E-mail address*: WM.vdFlier@vumc.nl

(Charlotte E. Teunissen) Neurochemistry Laboratory and Biobank, Dept. of Clinical Chemistry, Amsterdam Neuroscience, VU University medical center Amsterdam, Amsterdam, The Netherlands
*E-mail address*: c.teunissen@vumc.nl