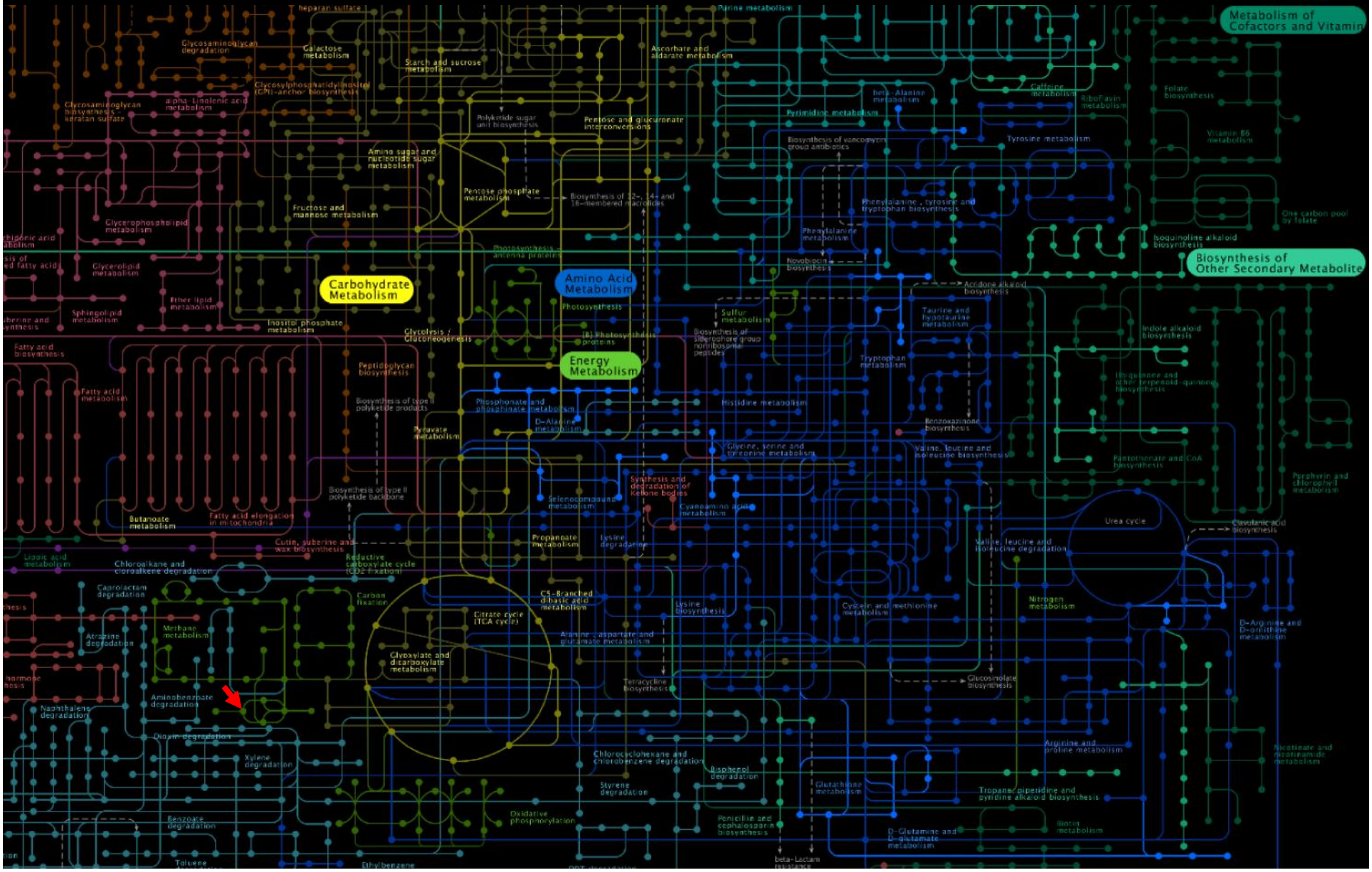
Advanced Strategies for Metabolomic Data Analysis

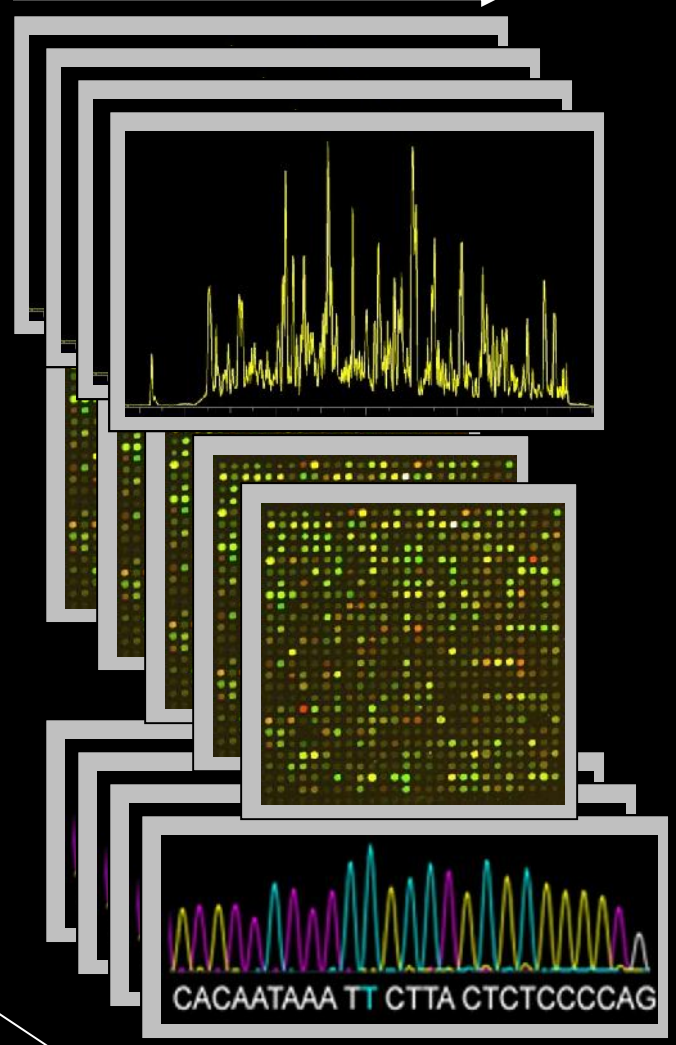
Dmitry Grapov, PhD

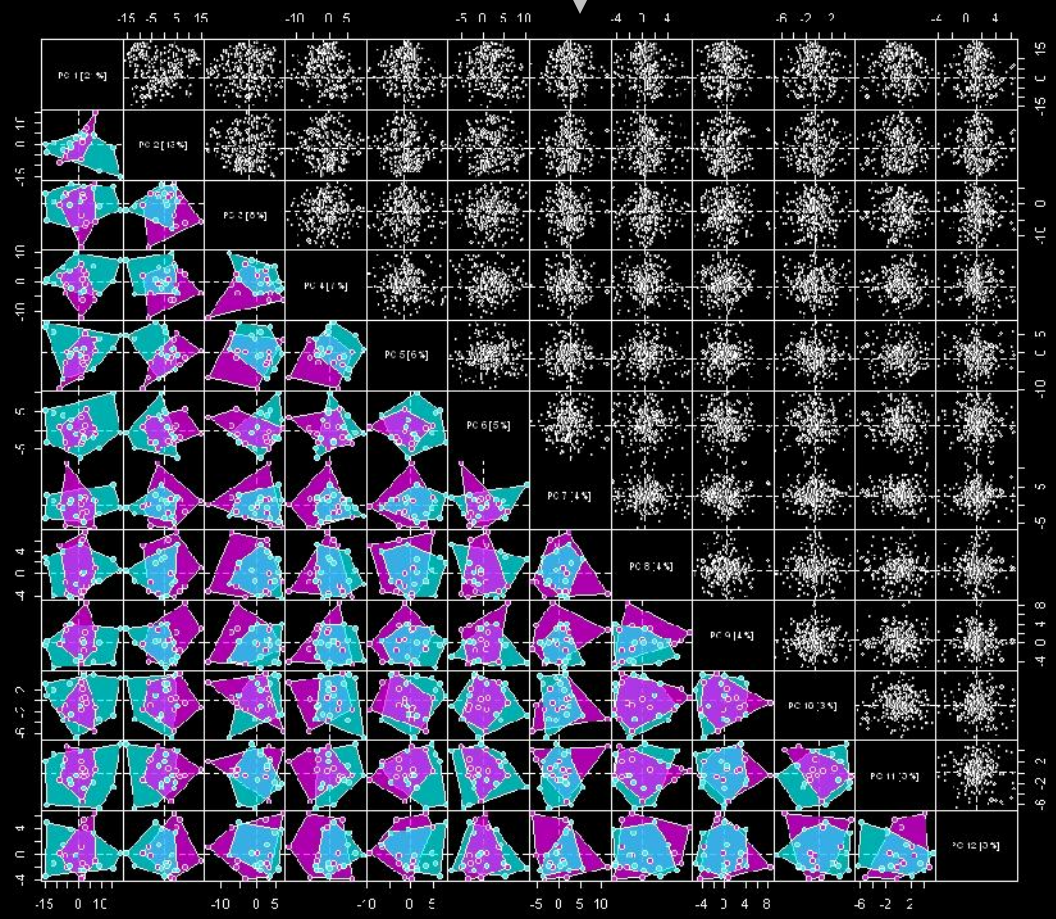
Analysis at the Metabolomic Scale



Multivariate Analysis

variables





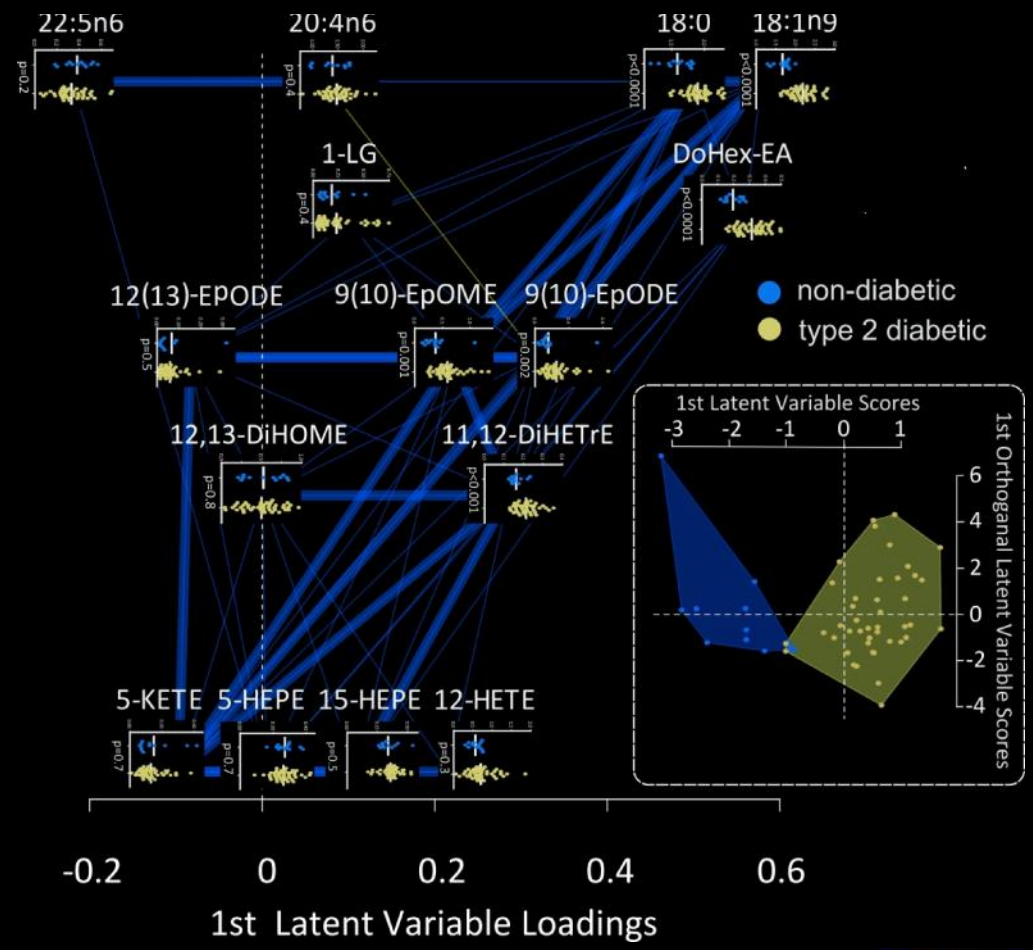
Samples

Metabolomics@ucdavis.edu Metabolomics.ucdavis.edu

Multivariate Analysis

Simultaneous analysis of many variables

- Visualization
- Clustering
- Projection
- Modeling
- Networks

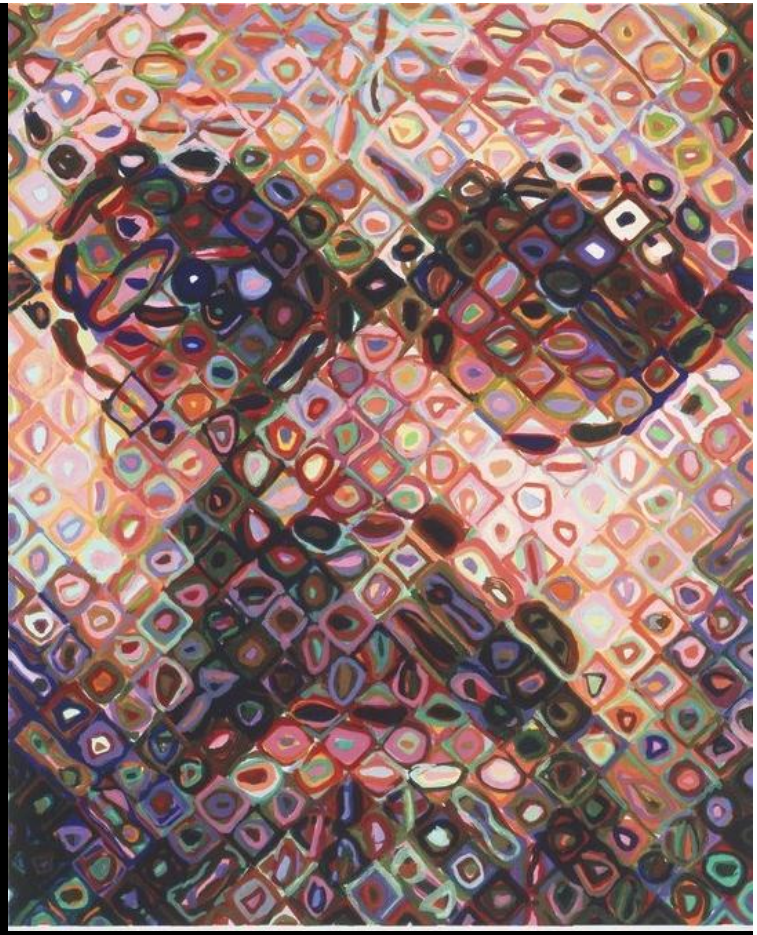


Clustering

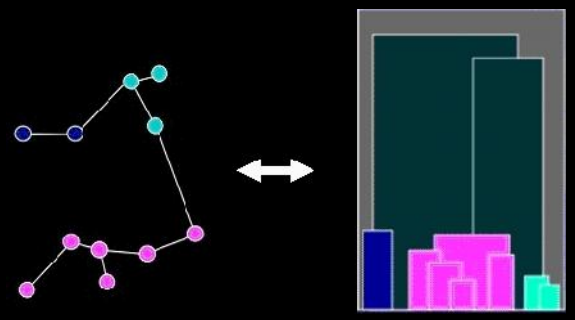
Identify

- **patterns**
- **group structure**
- **relationships**

- **Evaluate/refine hypothesis**
- **Reduce complexity**



Artist: Chuck Close



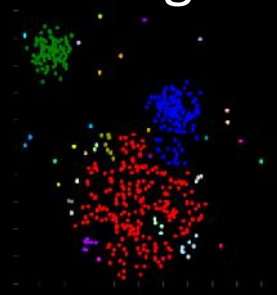
Cluster Analysis

Use the concept similarity/dissimilarity to group a collection of samples or variables

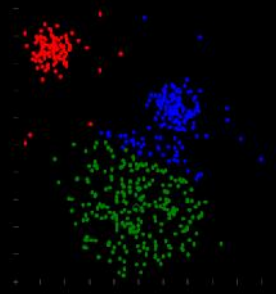
Approaches

- **hierarchical (HCA)**
- **non-hierarchical (k-NN, k-means)**
- **distribution (mixtures models)**
- **density (DBSCAN)**
- **self organizing maps (SOM)**

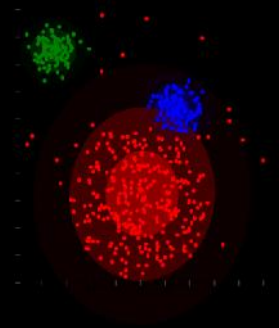
Linkage



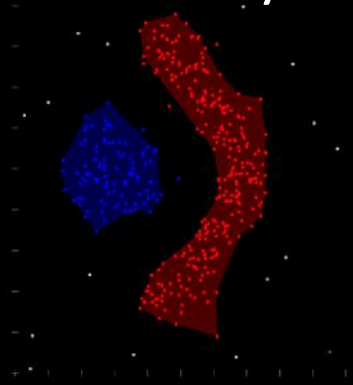
k-means



Distribution

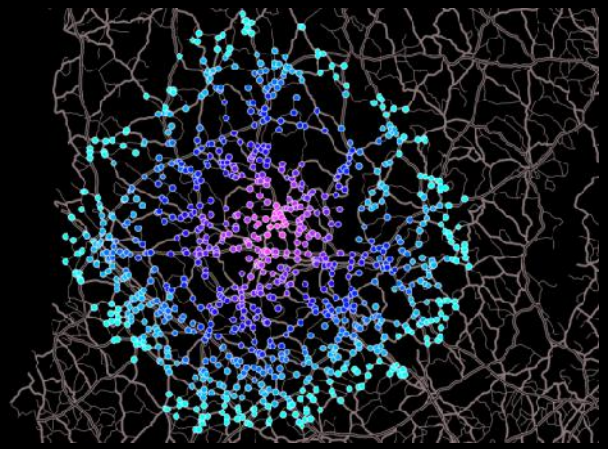


Density

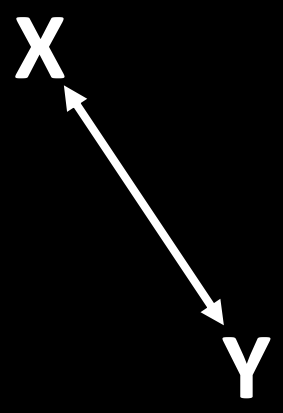


Hierarchical Cluster Analysis

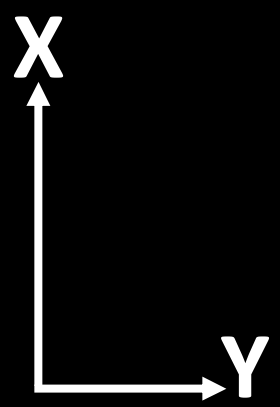
- similarity/dissimilarity defines "nearness" or distance



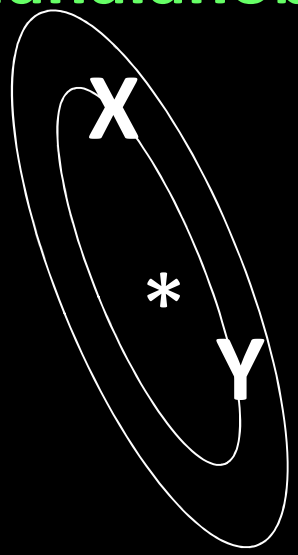
euclidean



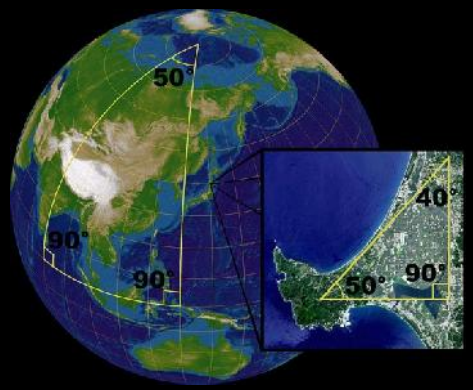
manhattan



Mahalanobis

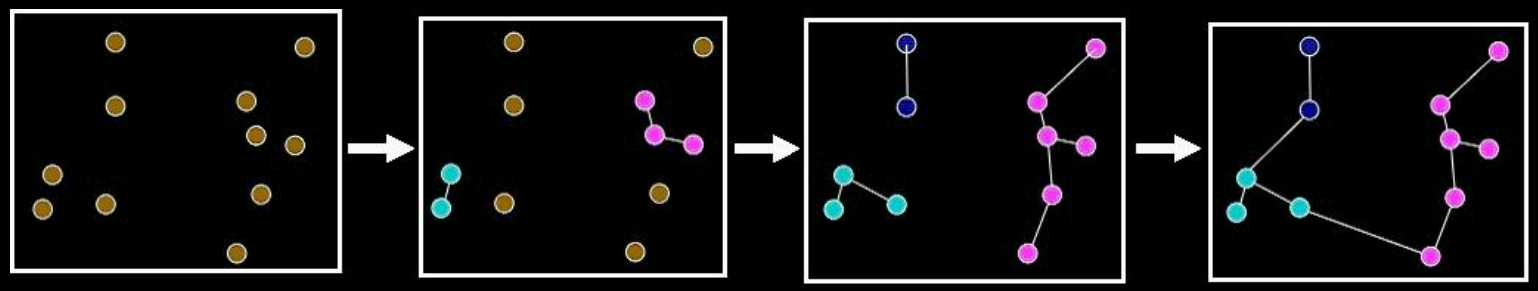


non-euclidean

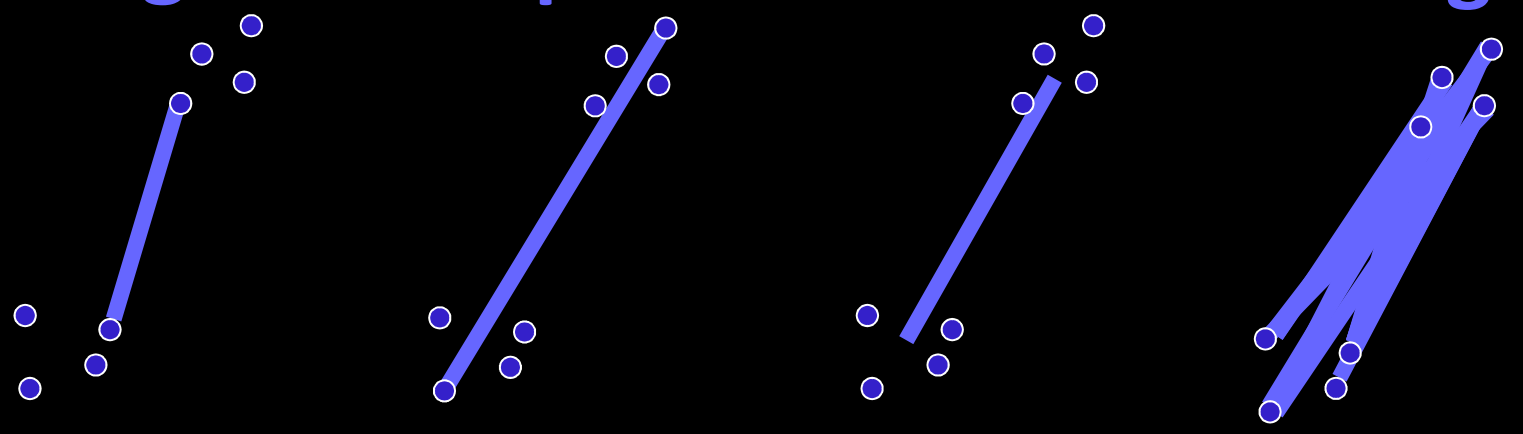


Hierarchical Cluster Analysis

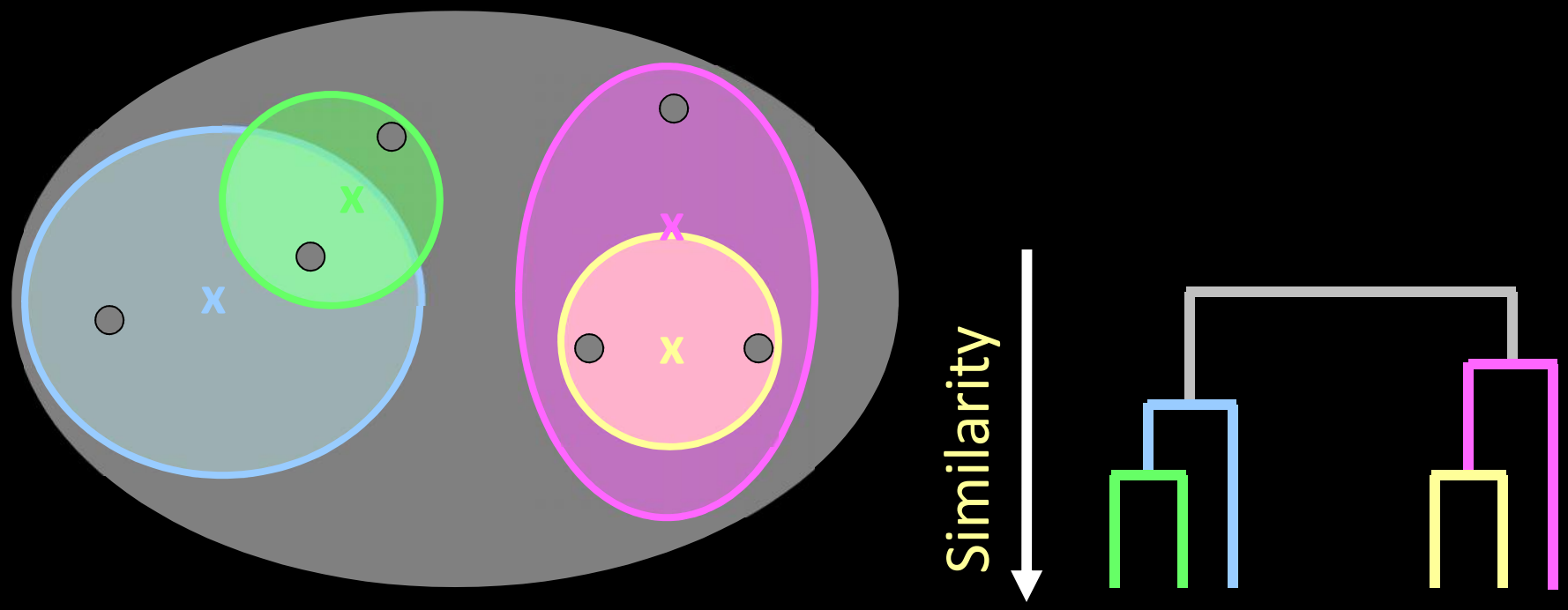
Agglomerative/linkage algorithm defines how points are grouped



single complete centroid average



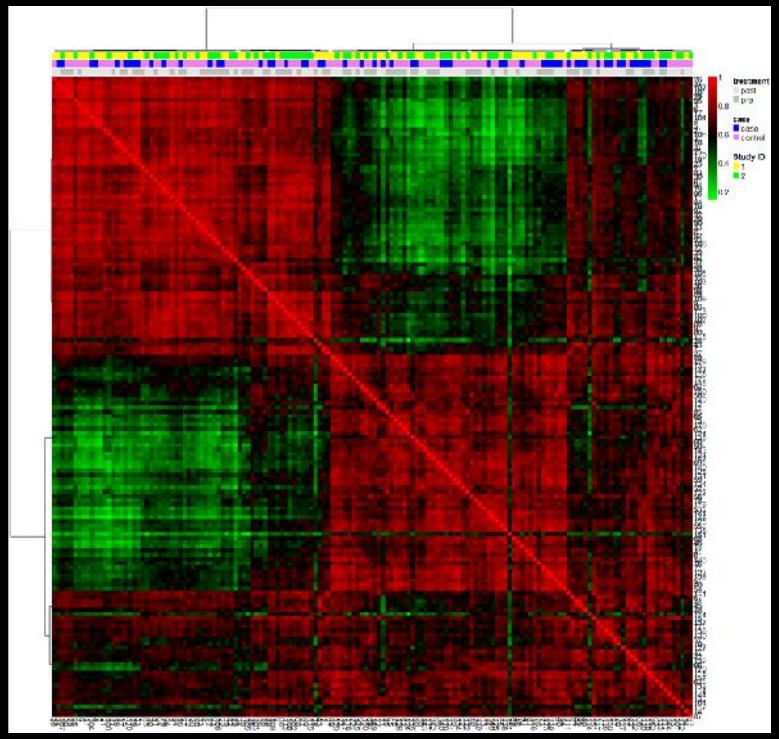
Visualization: Dendrogram



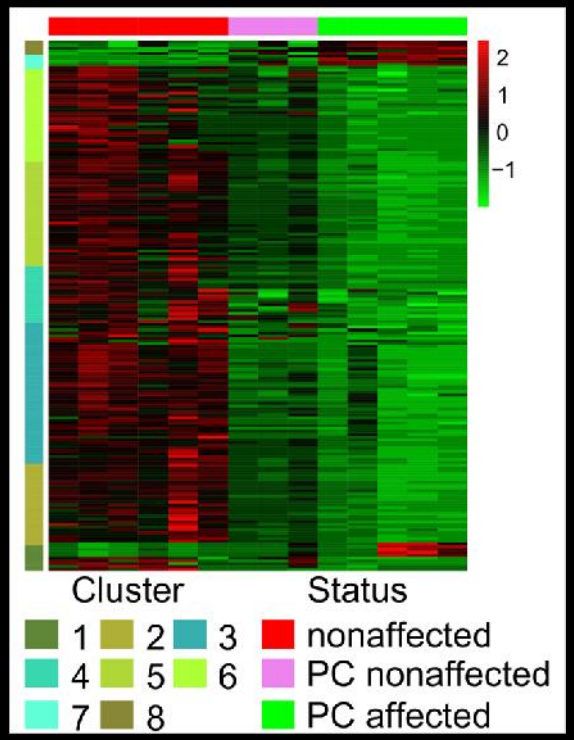
Implementation of Clustering

How does my metadata match my data structure?

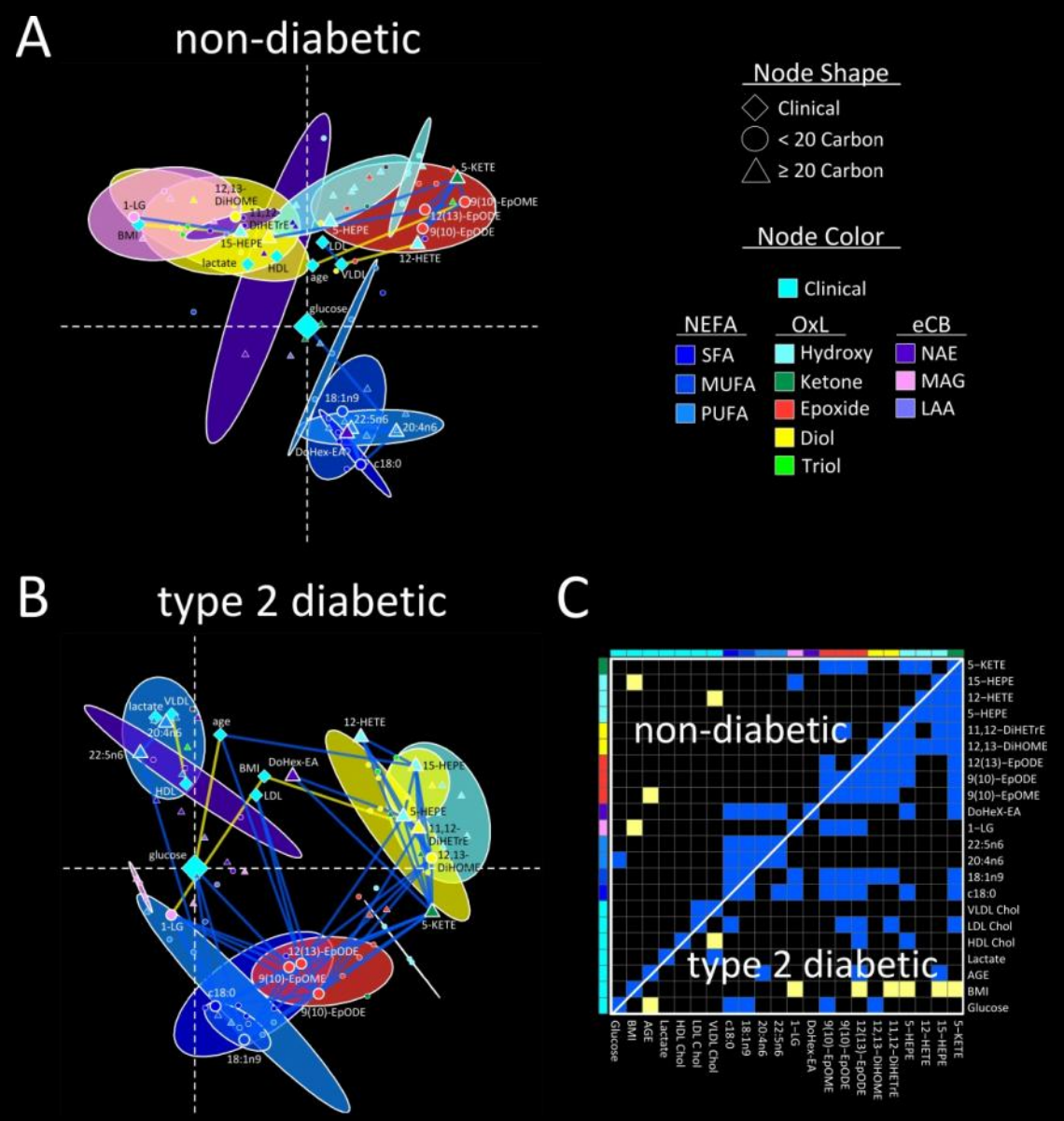
Overview



Confirmation



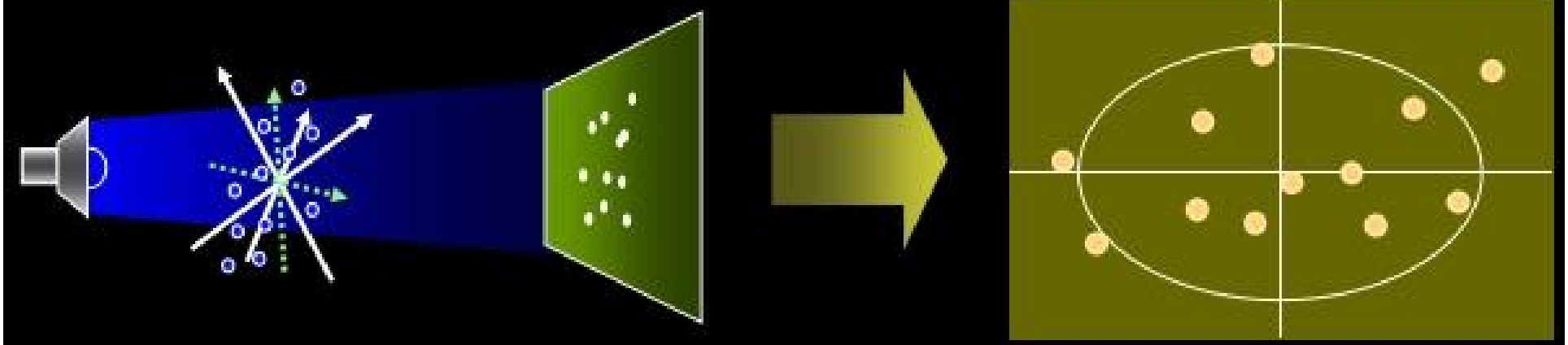
Multidimensional Scaling



PLoS ONE 7(11): e48852. doi:10.1371/journal.pone.0048852

Metabolomics@ucdavis.edu Metabolomics.ucdavis.edu

Projection of Data



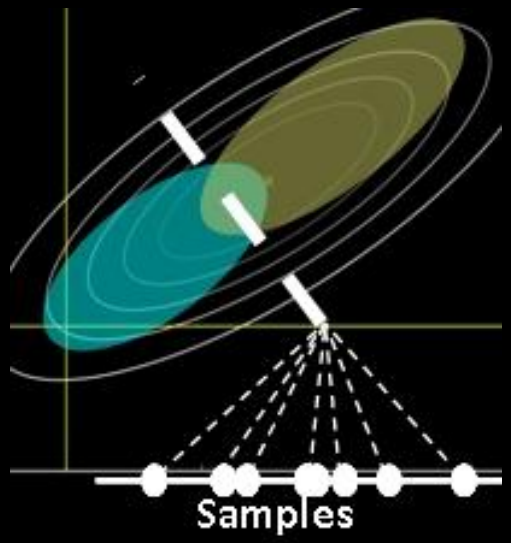
The algorithm defines the position of the light source

Principal Components Analysis (PCA)

- unsupervised
- maximize variance (X)

Partial Least Squares Projection to Latent Structures (PLS)

- supervised
- maximize covariance (Y ~ X)



PCA: Goals

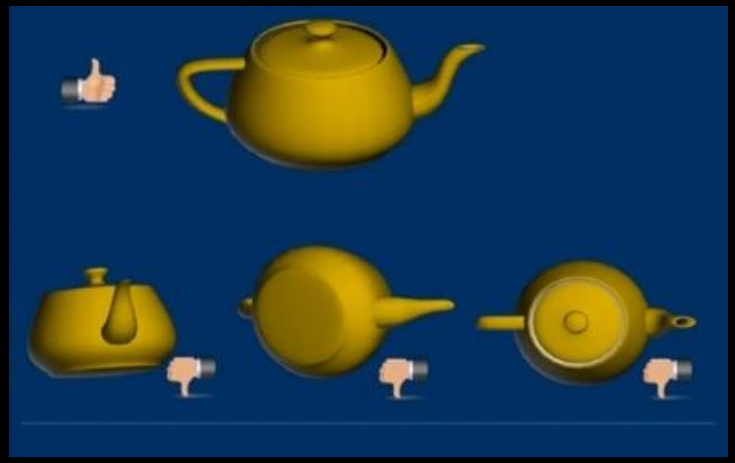
Non-supervised dimensional reduction technique

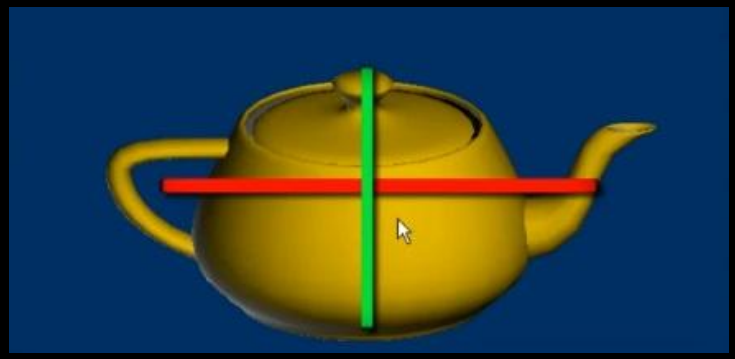
<u>Principal Components (PCs)</u>

- projection of the data which maximize variance explained

Results

- eigenvalues = variance explained
- scores = new coordinates for samples (rows)
- loadings = linear combination of original variables which

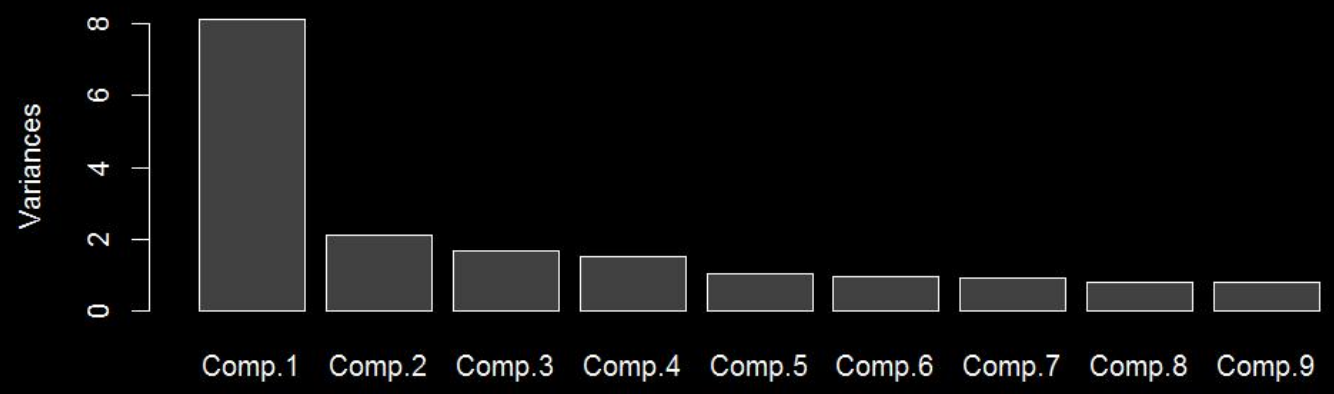




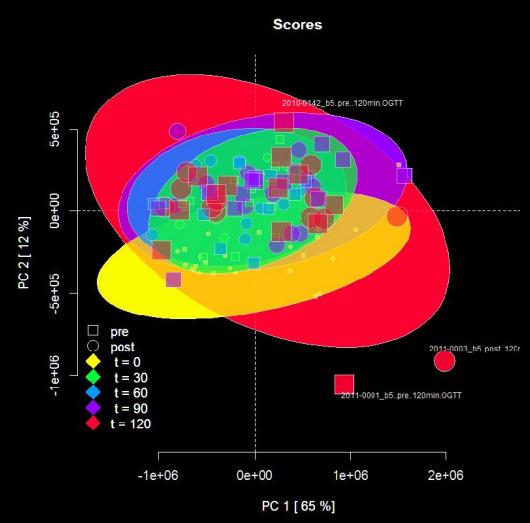
James X. Li, 2009, VisuMap Tech.

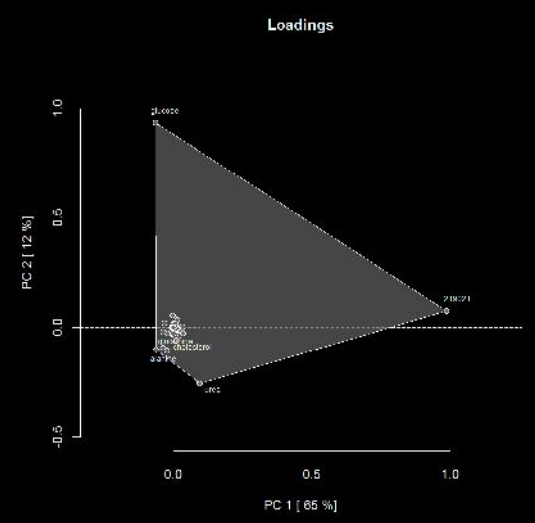
Interpreting PCA Results

Variance explained (eigenvalues)

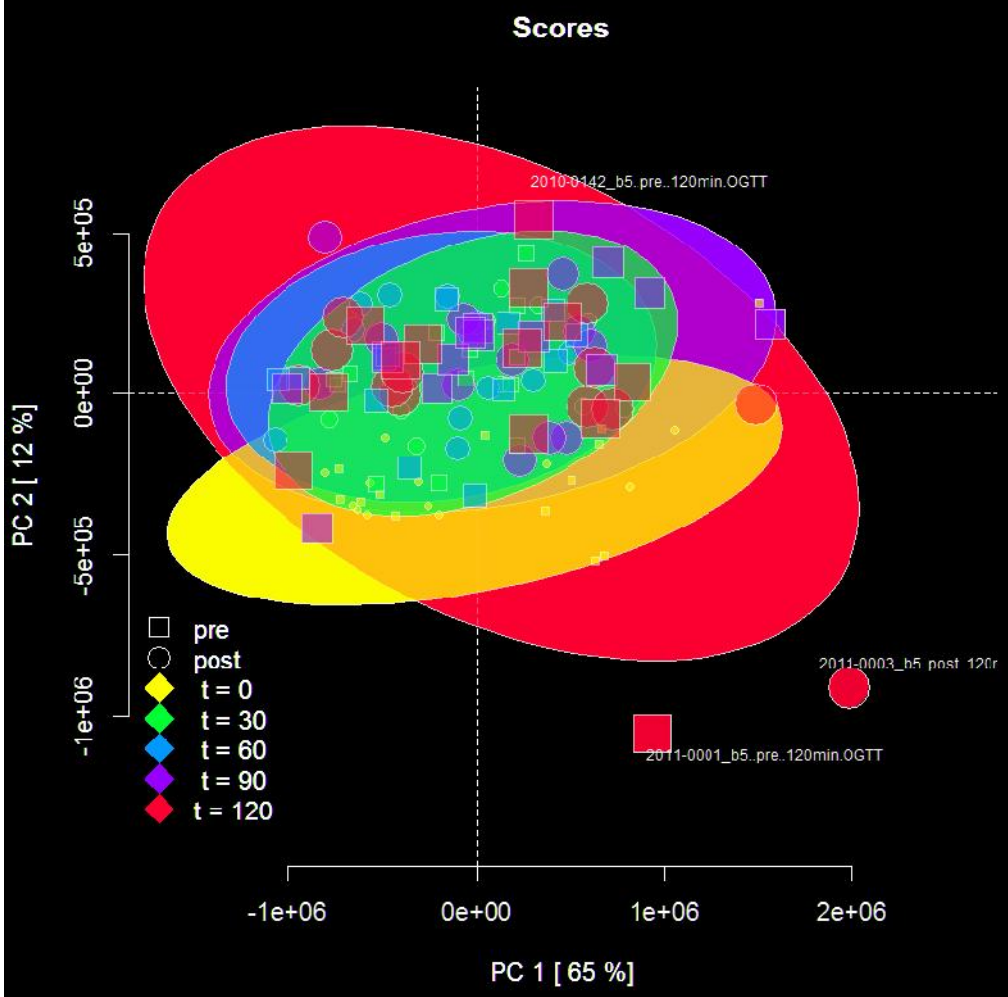


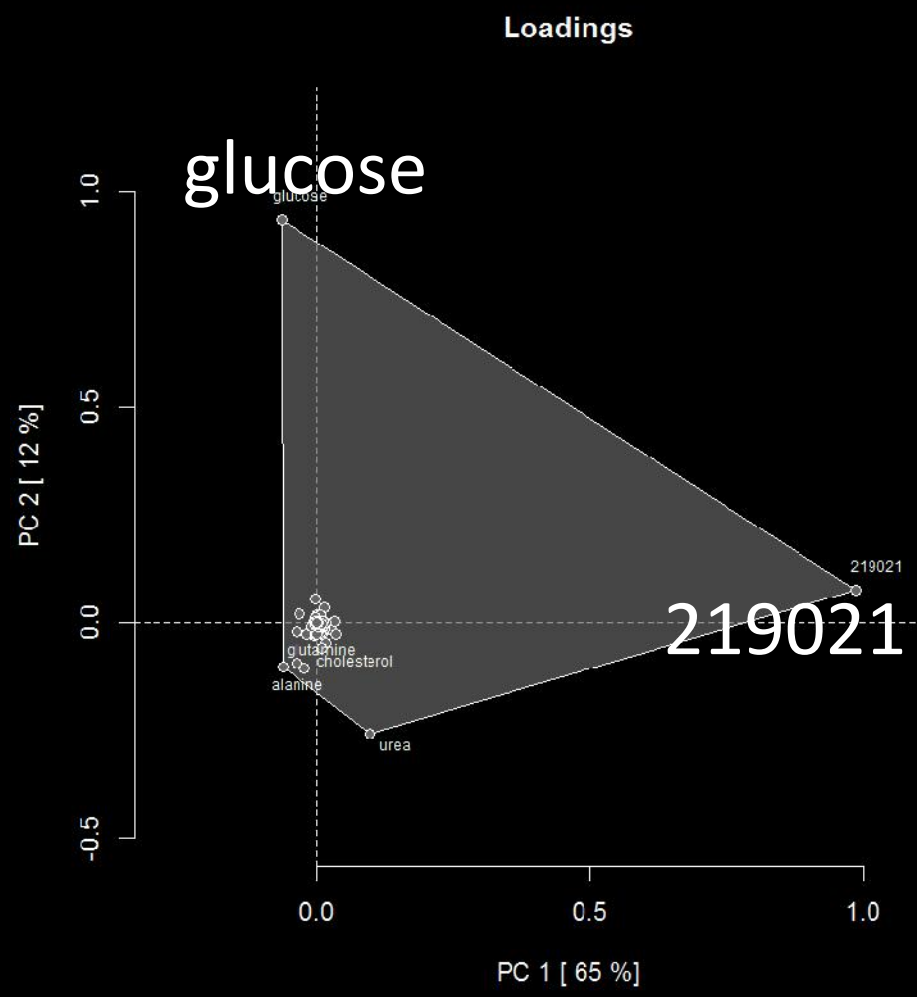
Row (sample) scores and column (variable) loadings





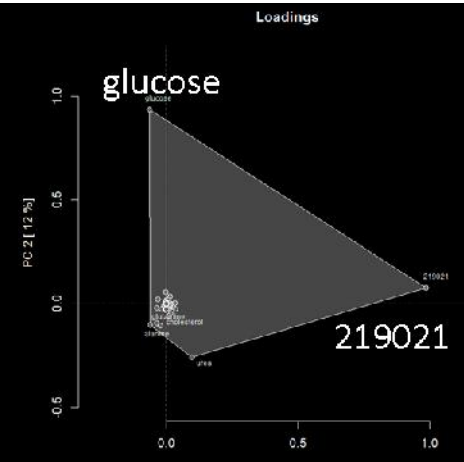
PCA Example

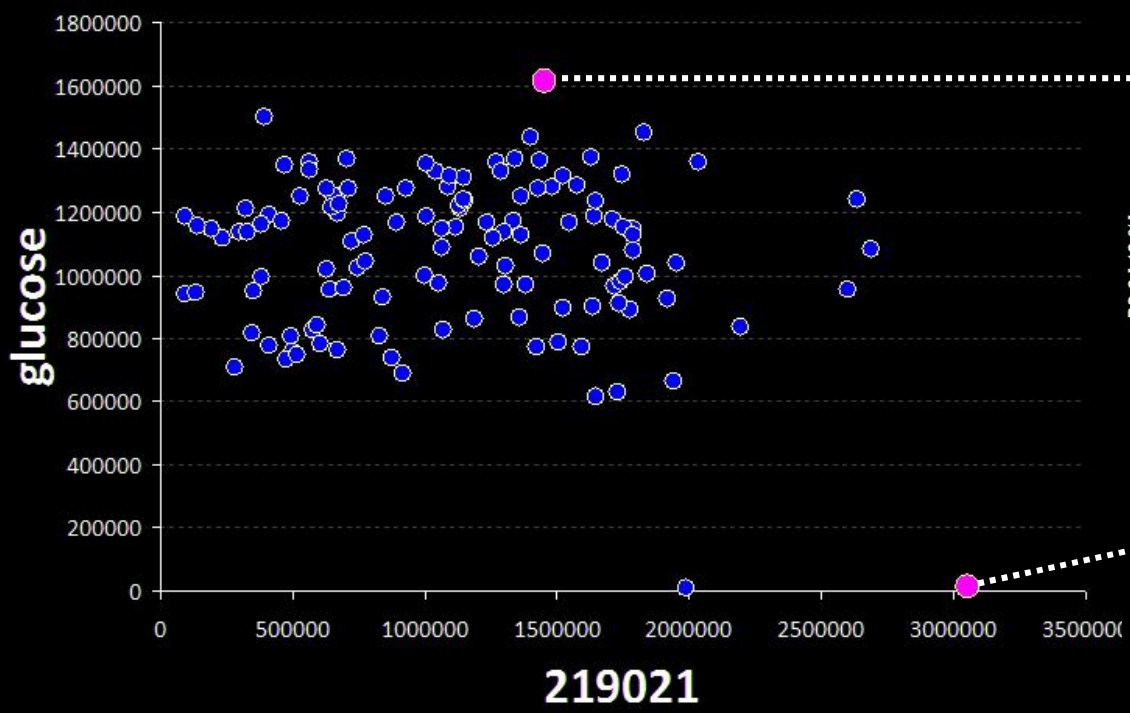


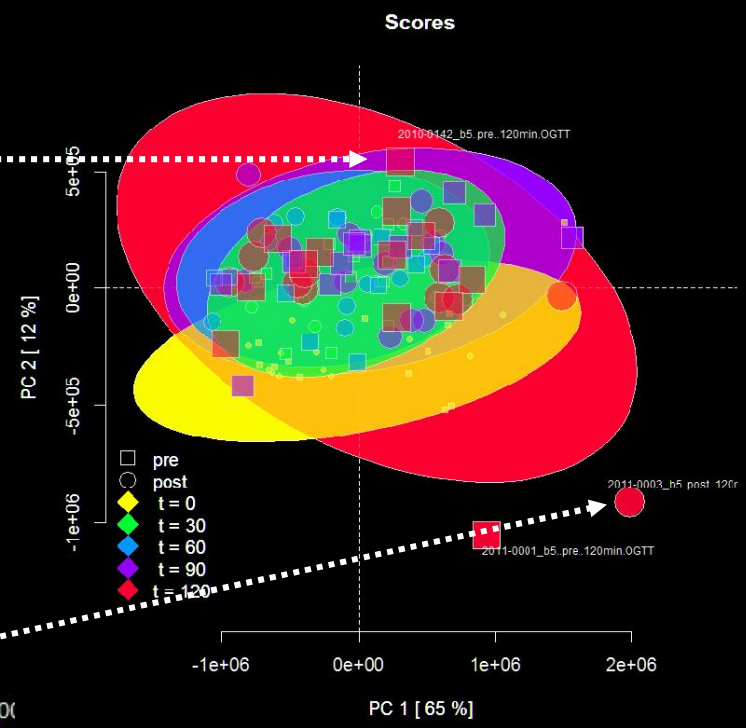


*no scaling or centering

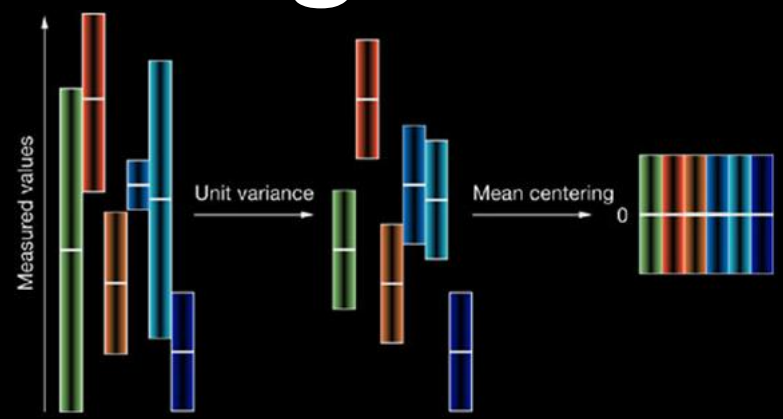
How are scores and loadings related?



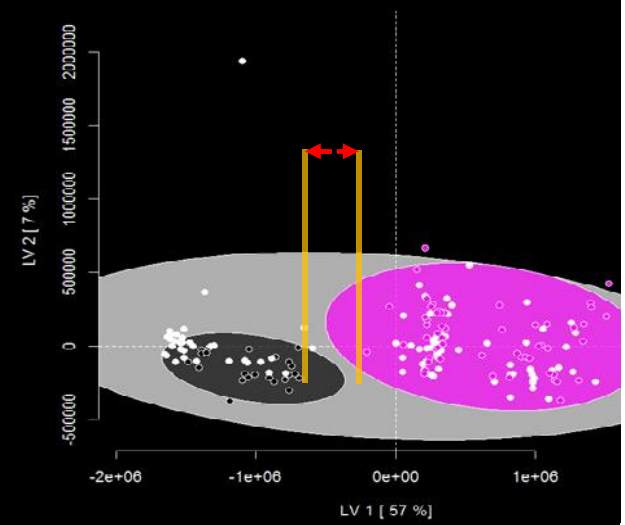


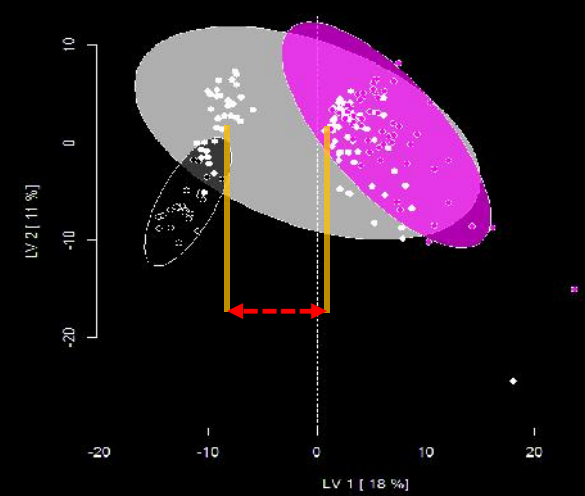


Centering and Scaling



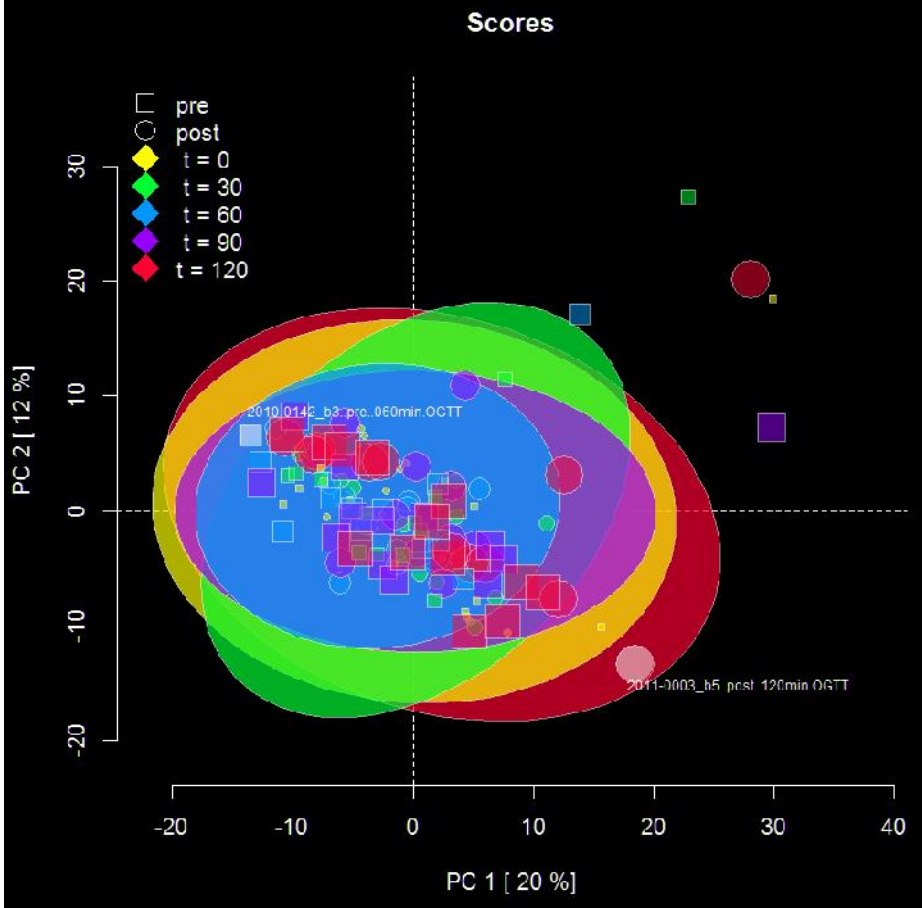
Method	Formula	Unit	Goal	Advantages	Disadvantages
Centering	$\tilde{x}_{ij} = x_{ij} - \bar{x}_i$	0	Focus on the differences and not the similarities in the data	Remove the offset from the data	When data is heteroscedastic, the effect of this pretreatment method is not always sufficient
Autoscaling	$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{s_i}$	(-)	Compare metabolites based on correlations	All metabolites become equally important	Inflation of the measurement errors
Range scaling	$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{\left(x_{i_{max}} - x_{i_{min}}\right)}$	(-)	Compare metabolites relative to the biological response range	All metabolites become equally important. Scaling is related to biology	Inflation of the measurement errors and sensitive to outliers
Pareto scaling	$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{\sqrt{s_i}}$	0	Reduce the relative importance of large values, but keep data structure partially intact	Stays closer to the original measurement than autoscaling	Sensitive to large fold changes
Vast scaling	$\tilde{x}_{ij} = \frac{\left(x_{ij} - \bar{x}_i\right)}{s_i} \cdot \frac{\bar{x}_i}{s_i}$	(-)	Focus on the metabolites that show small fluctuations	Aims for robustness, can use prior group knowledge	Not suited for large induced variation without group structure
Level scaling	$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{\bar{x}_i}$	(-)	Focus on relative response	Suited for identification of e.g. biomarkers	Inflation of the measurement errors
Log transformation	$\tilde{x}_{ij} = {}^{10}\log\left(x_{ij}\right)$ $\hat{x}_{ij} = \tilde{x}_{ij} - \bar{\tilde{x}}_i$	Log 0	Correct for heteroscedasticity, pseudo scaling. Make multiplicative models additive	Reduce heteroscedasticity, multiplicative effects become additive	Difficulties with values with large relative standard deviation and zeros
Power transformation	$\tilde{x}_{ij} = \sqrt{\left(x_{ij}\right)}$ $\hat{x}_{ij} = \tilde{x}_{ij} - \bar{\tilde{x}}_i$	0	Correct for heteroscedasticity, pseudo scaling	Reduce heteroscedasticity, no problems with small values	Choice for square root is arbitrary.

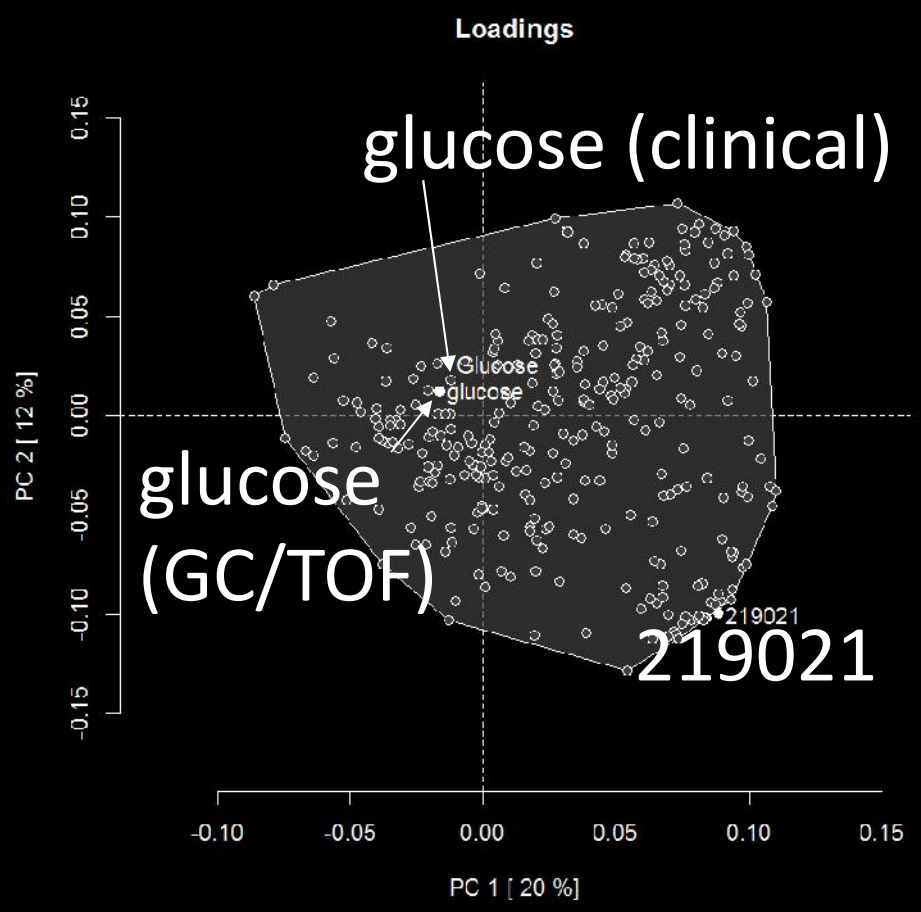




van den Berg RA, Hoefsloot HC, Westerhuis JA, Smilde AK, van der Werf MJ (2006) Centering, scaling, and transformations: improving the biological information content of metabolomics data. BMC Genomics 7: 142.

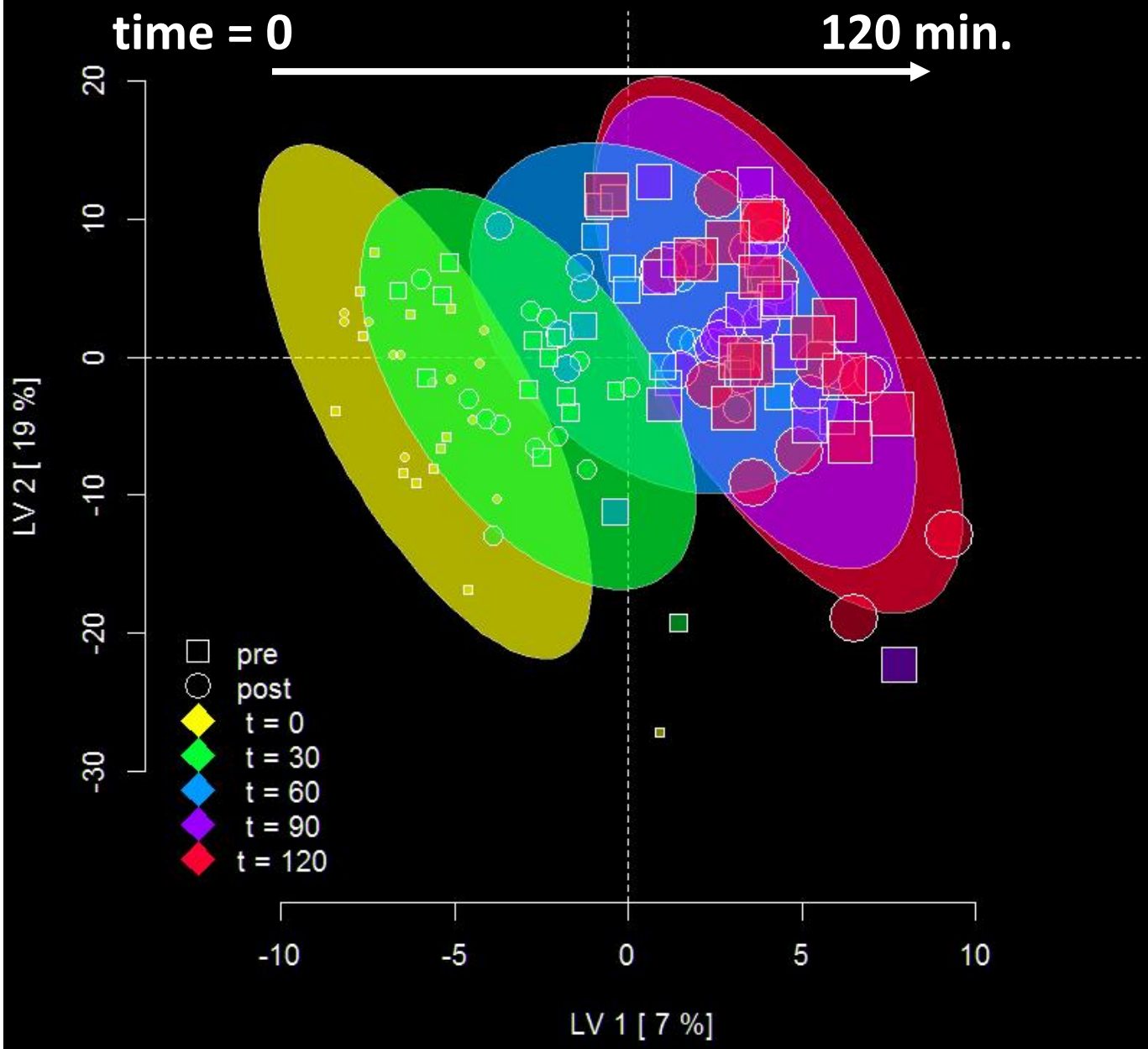
Data scaling is very important!





*autoscaling (unit variance and centered)

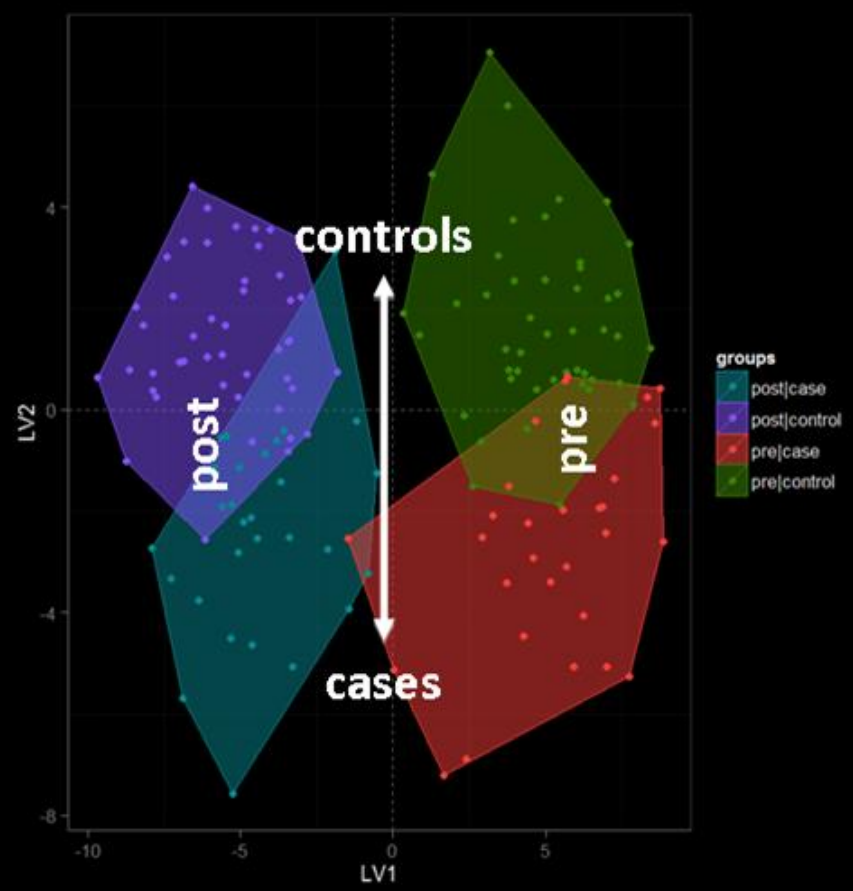
Use PLS to test a hypothesis



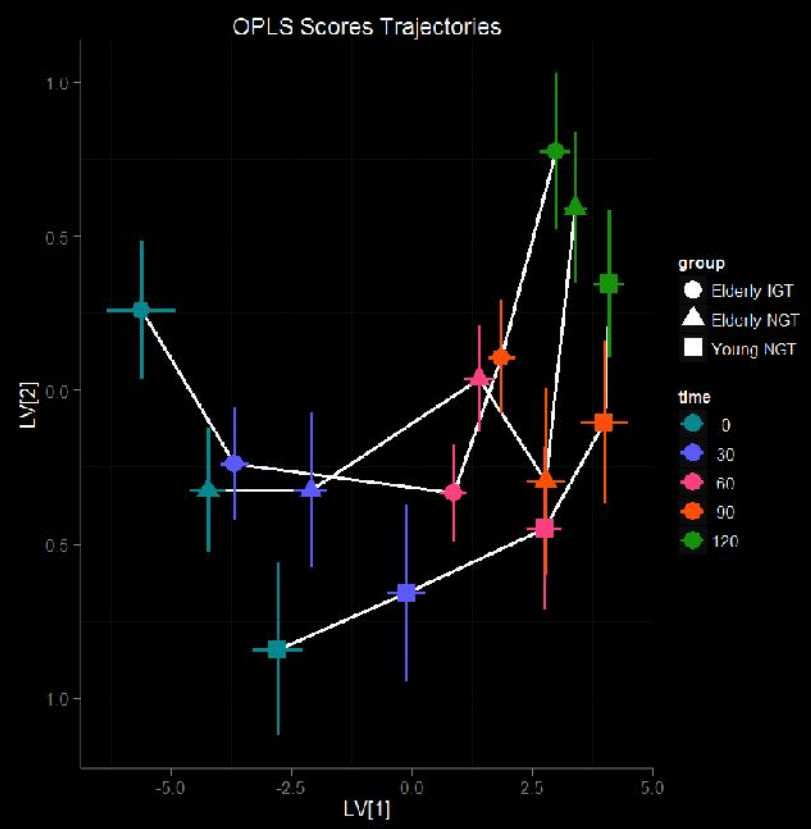
Loadings on the first latent variable (x-axis) can be used to interpret the multivariate changes in metabolites which are correlated with time

Modeling multifactorial relationships

~two-way ANOVA



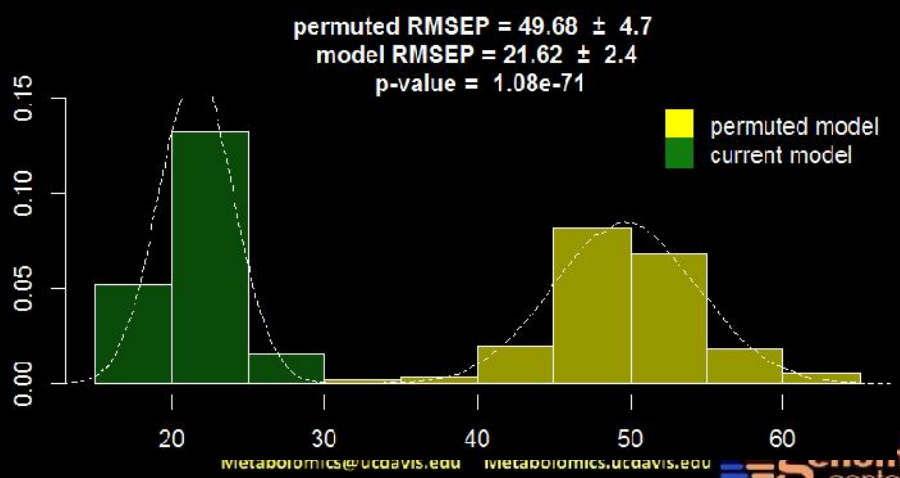
dynamic changes among groups



“goodness” of the model is all about the perspective

Determine in-sample (Q^2) and out-of-sample error (RMSEP) and compare to a random model

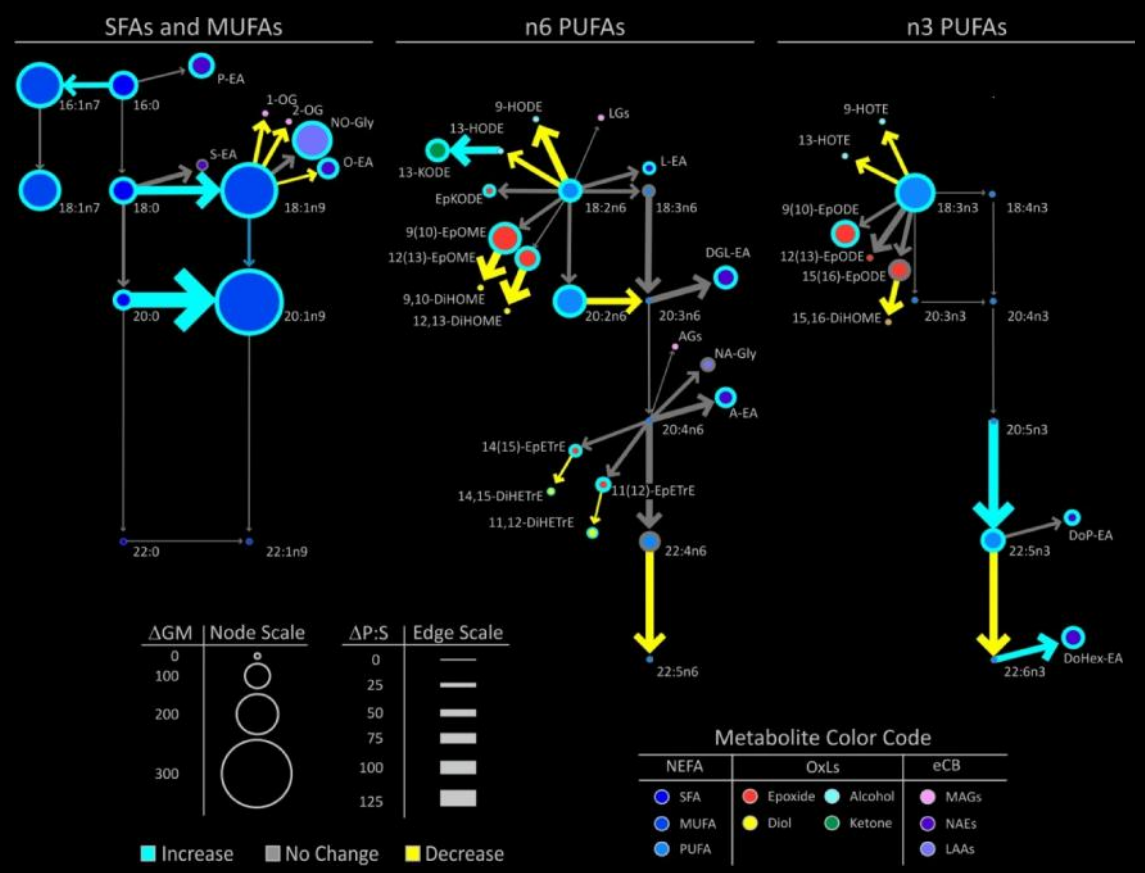
- permutation tests
- training/testing



Biological Interpretation

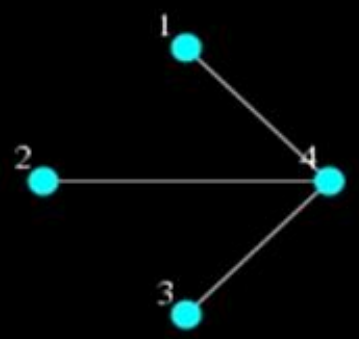
Projection or mapping of analysis results into a biological context.

- Visualization
- Enrichment
- Networks
  - biochemical
  - structural
  - empirical

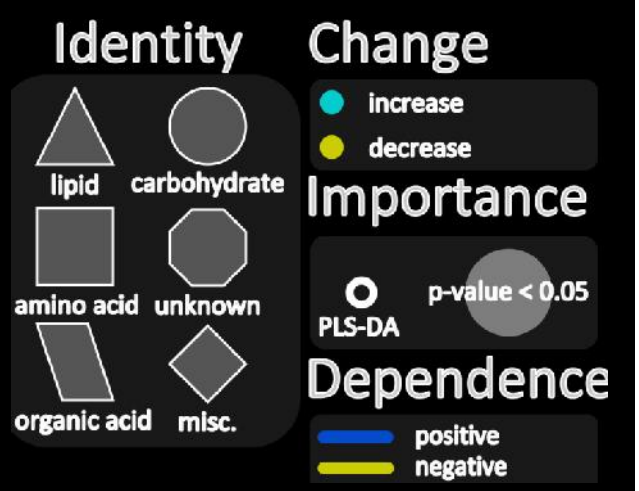


Ingredients for Network Mapping

1. Determine connections
   - Substrate/product (KEGG, biocyc)
   - chemical similarity (Tanimoto similarity)
   - dependency (partial correlation)
2. Determine vertex properties
   - magnitude
   - importance
   - direction
   - relationships
   - etc.

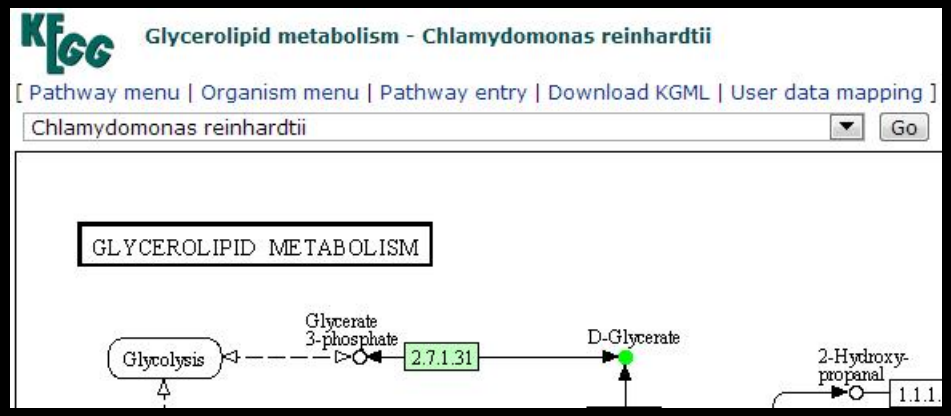


$$\begin{array}{c|cccc} & 1 & 2 & 3 & 4 \\ \hline 1 & 0 & 0 & 0 & 1 \\ 2 & 0 & 0 & 0 & 1 \\ 3 & 0 & 0 & 0 & 1 \\ 4 & 1 & 1 & 1 & 0 \end{array}$$



Making Connections Based on Biochemistry

- Organism specific biochemical relationships
- KEGG
  - paid API
  - download free KGML file
- BioCyc
  - Free API



Making Connections Based on Structural Similarity

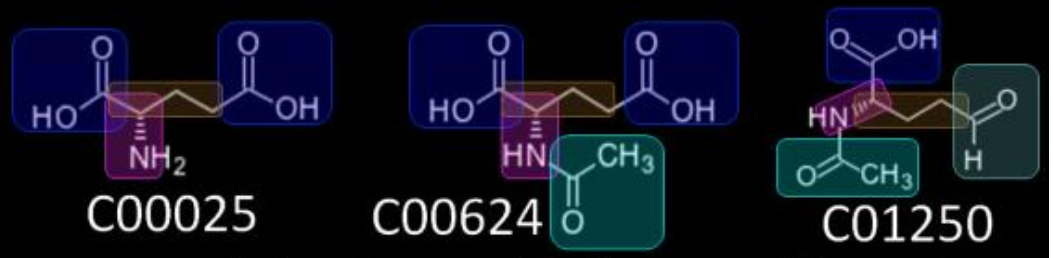
- Use structure to generate molecular fingerprint
- Calculate similarities between metabolites based on fingerprint
- PubChem service for similarity calculations

http://pubchem.ncbi.nlm.nih.gov//score_matrix/score_matrix.cgi

- Metamapp online tool for data formatting

http://uranus.fiehnlab.ucdavis.edu:8080/MetaMapp/homePage

Chemical mapping
of substructure comparison
using PubChem



substructure matrix decomposition and
Tanimoto chemical similarity calculations

BMC Bioinformatics 2012, **13**:99 doi:10.1186/1471-2105-13-99

Ingredients for Mapped Networks

1. edge list
   - biochemical
   - structural
   - empirical
2. vertex attributes
   - user-defined
   - based on analysis results
3. Visualization

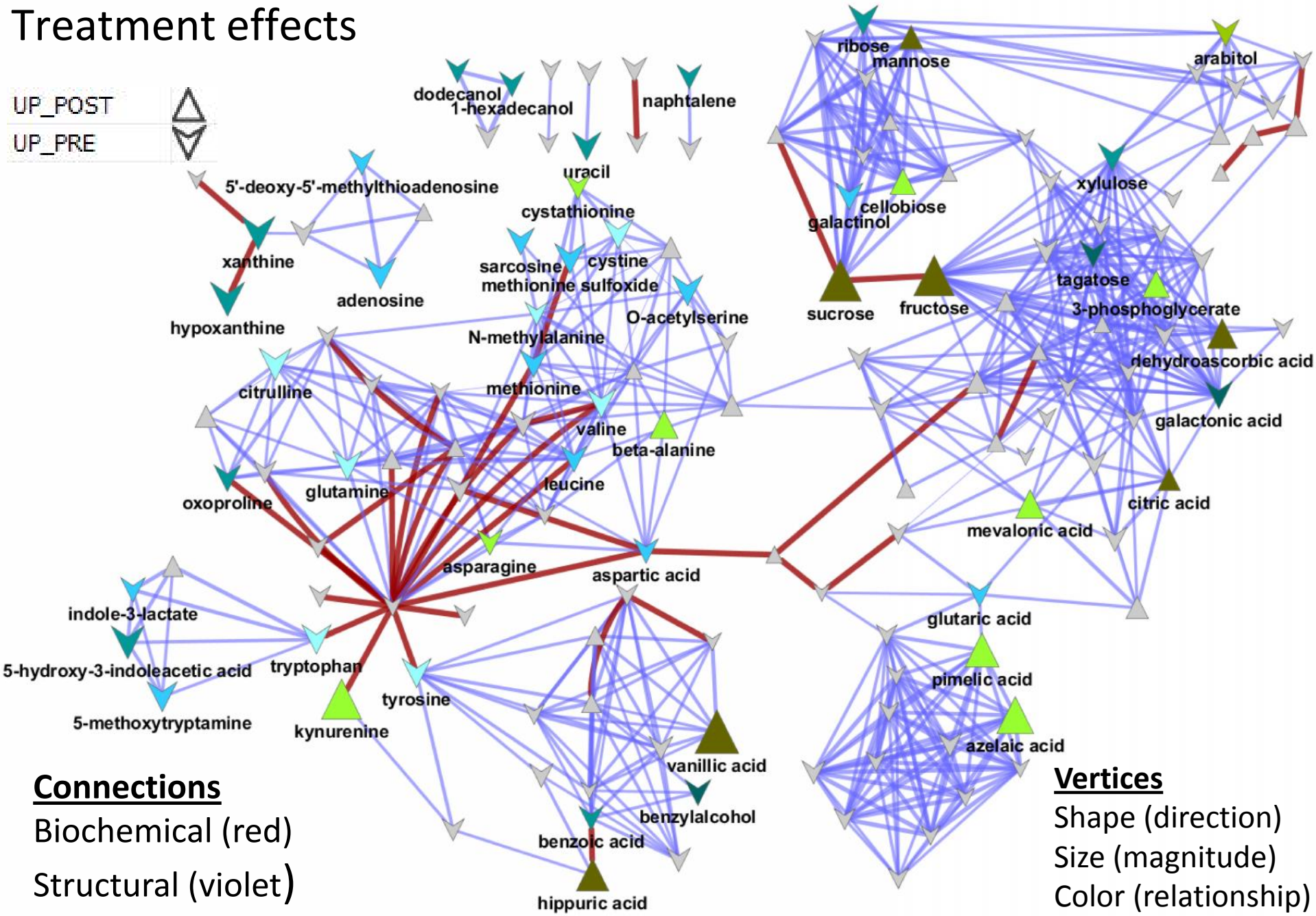
Edge list

Source	Target	Score
19	43	25
19	51	23
43	51	76

CID	names	fold change	p-values
19	2,3-dihydroxybenzoic acid	2.5	0.2118
43	2-hydroxyglutaric acid	1.4	0.0054
51	alpha ketoglutaric acid	1.3	0.3239
71	2-ketoadipic acid	4.6	0.1435
119	GABA	1.6	0.0001

Cytoscape

Treatment effects



Connections

Biochemical (red)

Structural (violet)

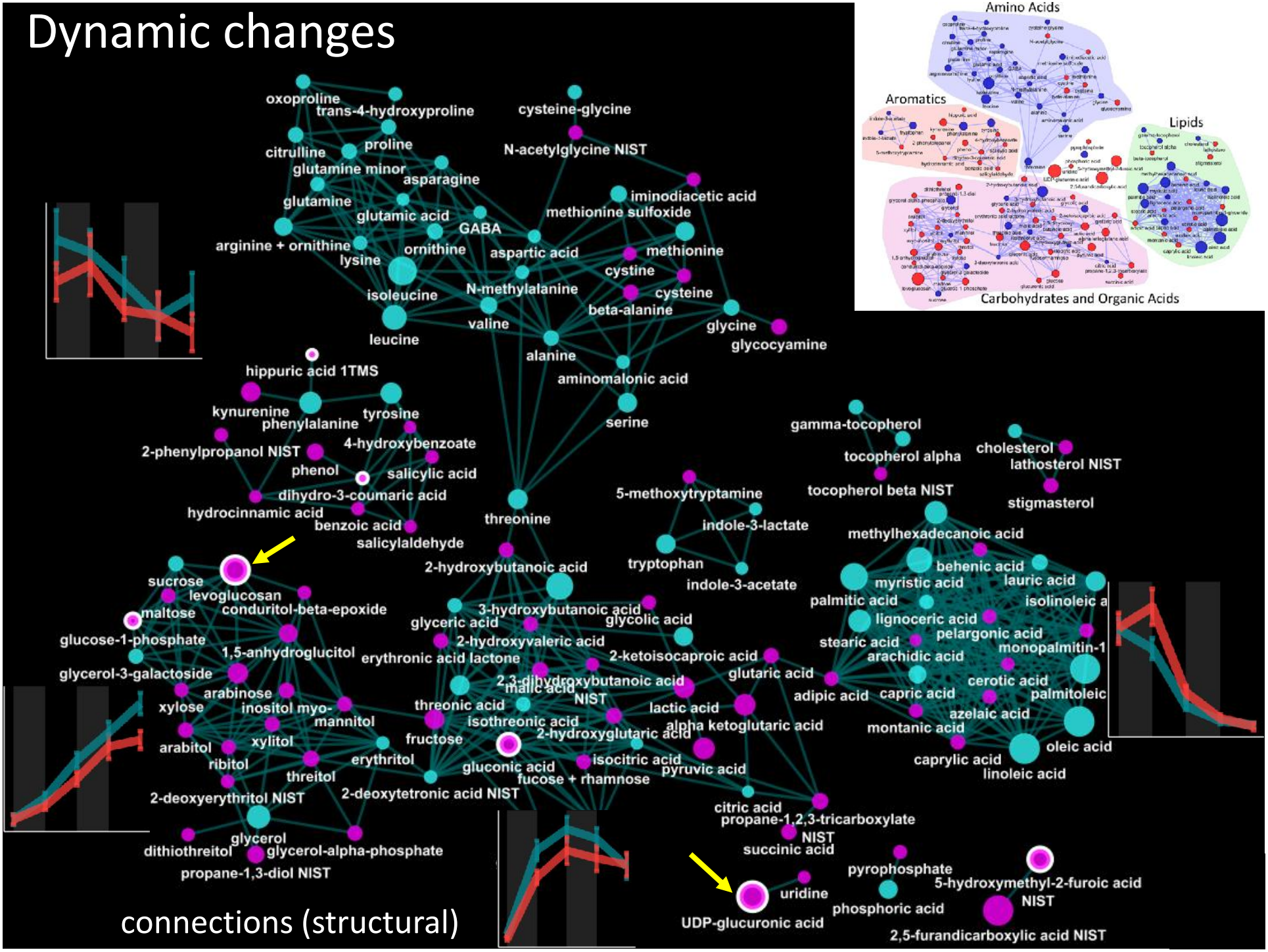
Vertices

Shape (direction)

Size (magnitude)

Color (relationship)

Dynamic changes



connections (structural)

Variable Relationships

Identity

lipid
carbohydrate
amino acid
unknown
organic acid
misc.

Change

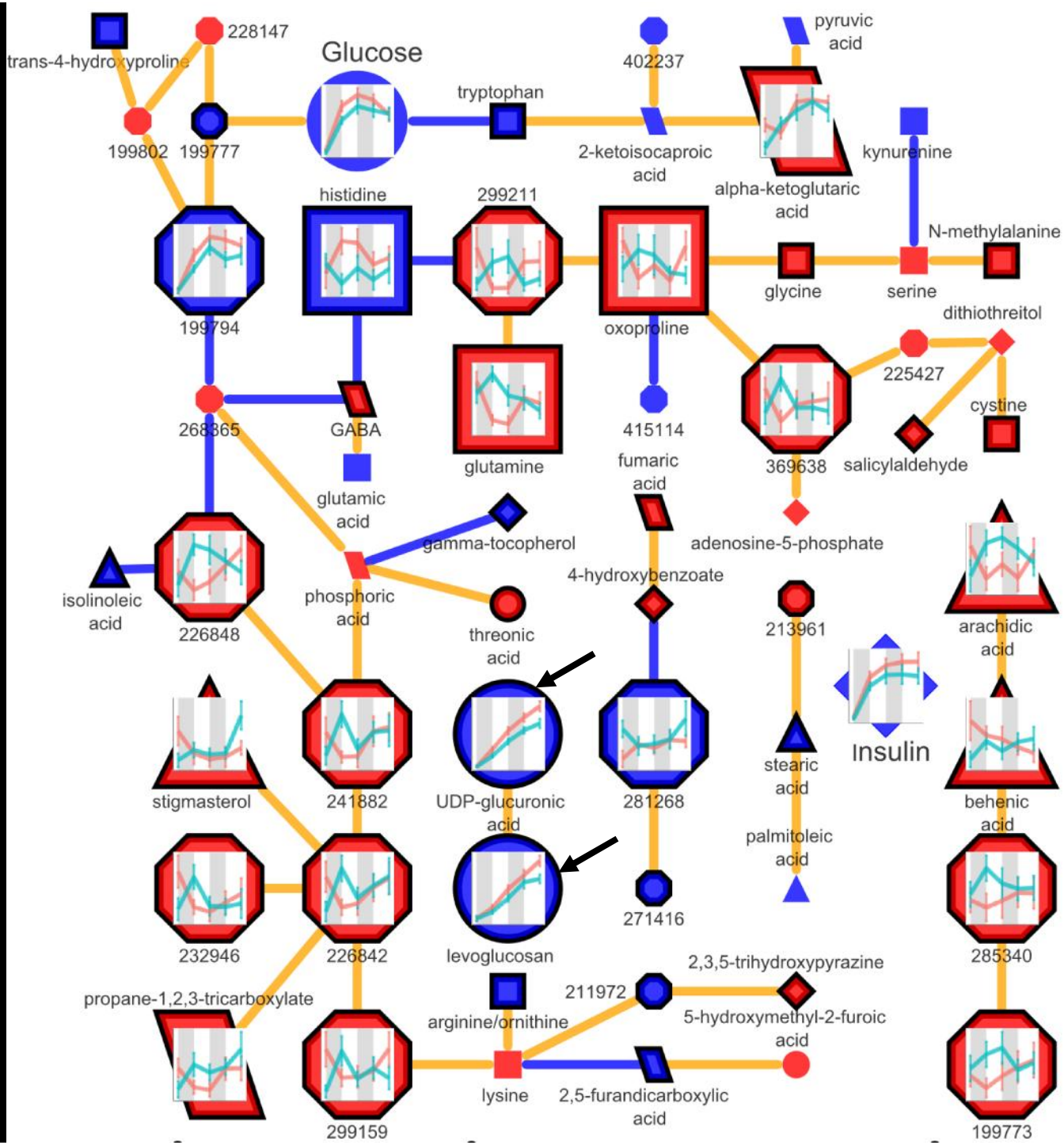
increase
decrease

Importance

PLS-DA
p-value < 0.05

Dependence

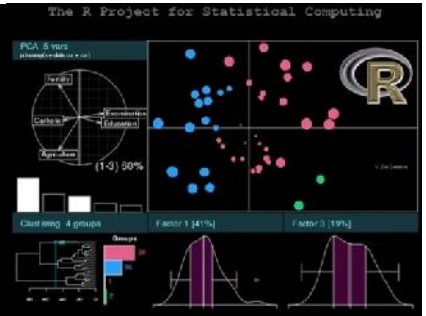
positive
negative



Summary

- Multivariate analysis is useful for
  - Visualization
  - Exploration and overview
  - Complexity reduction
  - Identification of multidimensional relationships and trends
  - Mapping to networks
  - Generating holistic summaries of findings

Resource



- Mapping tools (review)
  - Brief Bioinform (2012) doi: 10.1093/bib/bbs055
- Tutorials and Examples

  http://imdevsoftware.wordpress.com/category/uncategorized/
  https://github.com/dgrapov/TeachingDemos
- Chemical Translations Services
  - CTS: http://cts.fiehnlab.ucdavis.edu/
    - R-interface: https://github.com/dgrapov/CTSgetR
  - CIR: http://cactus.nci.nih.gov/chemical/structure
    - R-interface: https://github.com/dgrapov/CIRgetR