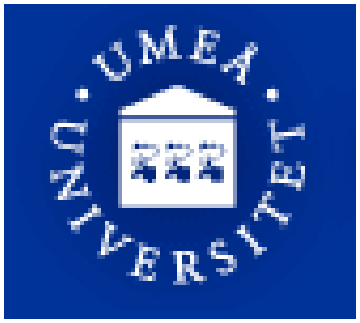


A view of metabolomics from a chemometrics perspective

Rui CLIMACO PINTO

with

Johan TRYGG

- Umeå University – Sweden -

Overview

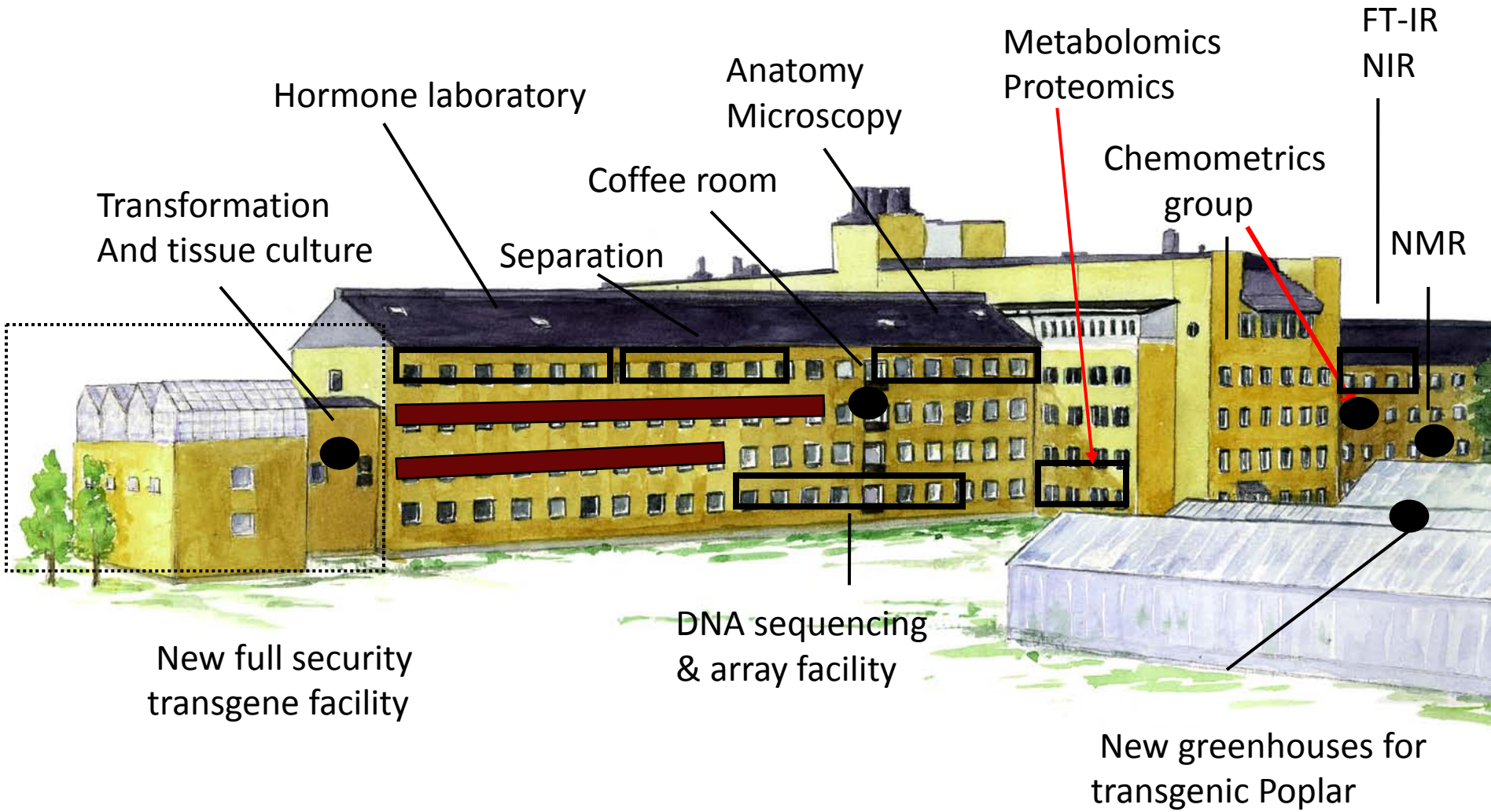
- Umeå chemometrics/bioinformatics group CLiC
- Metabolomics
- Integration of chemometrics in metabolomics
- Multivariate regression / Discriminant analysis
- OPLS and O2PLS framework
- Examples of chemometrics in metabolomics

Umeå, Sweden

University built in 1965
25 000 students / 4000 staff

Collaboration: Umeå Plant Science Center

Excellence centre in plant biology



UMEÅ – CLiC

Objectives:

- Stimulate, organize and advance computer based modelling, tools and strategies to understand complex biological systems and e-bioscience
- Be the critical and missing link to the ongoing strong experimental research at Umeå University (UPSC, UCFB, FuncFiber, MIMS, UCMR and UCMM centers)
-Establish a unique bioinformatics/e-science profile in Umeå

Main research areas:

- Omics-technologies (mainly transcriptomics, proteomics, metabolomics)
- Network modeling, databases and visualization
- Structural biology and sequence analysis

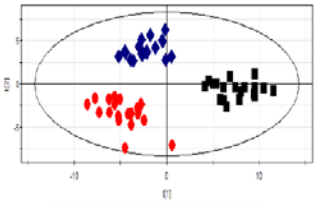
Group leaders:

Antti, Henrik (Assoc. Prof.) - Predictive and Human Metabolomics
Hedenström, Mattias (Ass.Prof.) - Characterization of plant material and biofluids using NMR spectroscopy
Hvidsten, Torgeir (Ass.Prof) -A systems biology approach to model the transcriptional network in trees
Linusson Jonsson, Anna (Ass.Prof) - Probing molecular interactions of protein-ligand complexes guided by an integration of chemometrics and molecular modelling
Rydén, Patrik (Ass.Prof.) - Pathogenicity of Francisella tularensis
Sauer, Uwe (Assoc.Prof) - BioCrystallography and BioInformatics
Sjöström, Michael (Prof.) - Multivariate quantitative structure activity relationships (M-QSAR)
Stenberg, Per (Ass.Prof) - Mining functional DNA elements in eukaryotic genomes
Trygg, Johan (Assoc.Prof) -Chemometrics in metabolomics, 'omics profiling and systems biology

Trygg group's chemometrics in 'Bio-'

Tree biology: Functional genomics in transgenic Poplar trees

- Umeå Plant Science Center



Disease diagnosis & biomarker identification

- Rheumatoid Arthritis, Diabetes 1 & 2, Huntington, etc...

Urine test to monitor kidney-transplant rejection

A urine test that diagnoses acute rejection without the need for an invasive biopsy

Medicine (Post operative surgery): Kidney transplant

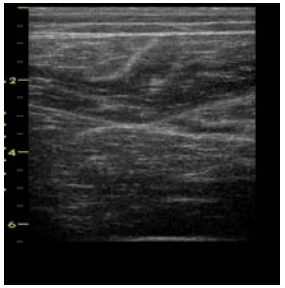
- Monitor immune suppression vs toxicity with NMR spectroscopy

Dietary: Functional foods

- Health effect from food supplement with NMR & GC-MS spectroscopy

Medical imaging by ultrasound

- Study muscle tissue physiology and function in rehabilitation



Metabolomics

Metabolomics - definitions

Supporting thesis: Functional status of a complex biological system resides in the quantitative and qualitative pattern of metabolites in body fluids

- **Metabolome** - Complete set of metabolites to be found within a biological sample
- **Metabolite**
  - Small biological molecules, intermediates and products of metabolism
  - Primary: main functions (growth, development, reproduction)
  - Secondary: ecological function (ex. antibiotics and pigments)

- **Metabolomics** – systematic study of the unique chemical fingerprints that specific cellular processes leave behind (MS)
- **Metabonomics** - quantitative measurement of the dynamic multiparametric metabolic response of living systems to pathophysiological stimuli or genetic modification (NMR)

Metabolomics

- Instrumental analysis
  - Mainly GC-MS, LC-MS, NMR
  - Also Raman, FTIR
  - Large amounts of data
- Use of chemometrics
- **Disease diagnosis**, **functional genomics**, toxicology, **plant science**, **nutrition**, **pharmaceutical** and environmental research, **personalized medicine**
- Today - trend in biological interpretation rather than only classify samples

Personalised medicine

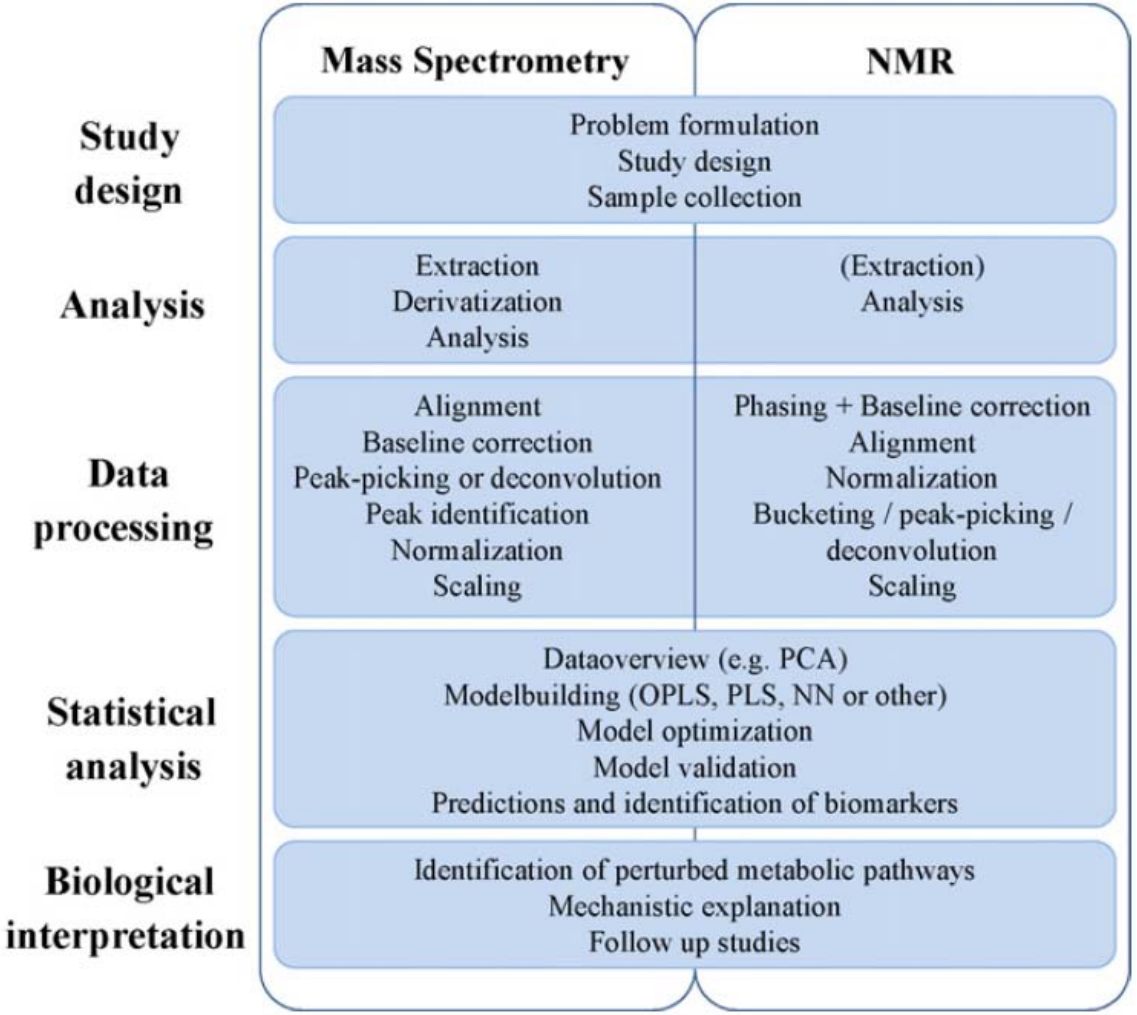
- Personalized medicine – refine the empirical approach used in most clinical trials by incorporating powerful new diagnostics that can identify individual predictive characteristics and better control variability

Metabolomics in personalised medicine

- Drug metabolism pathways
- Definition of disease subsets
- Definition of groups of patients
- **Monitoring treatment response**
- Prevention
- Drug safety

J. Woodcock, Clinical Pharmacology & Therapeutics, 81 (2007) 164

Metabolomics – steps

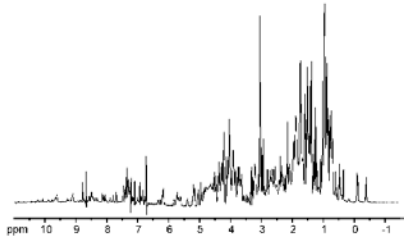


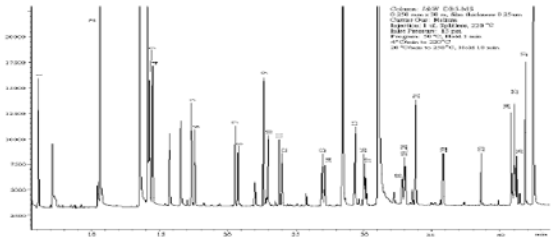
Madsen, R.; Lundstedt, T.; Trygg, J.; **Chemometrics in metabolomics - A review in human disease diagnosis,** Analytica chimica acta, in print.

Metabolomics – Simplified view

Biological samples

Biochemical analysis of endogenous metabolites

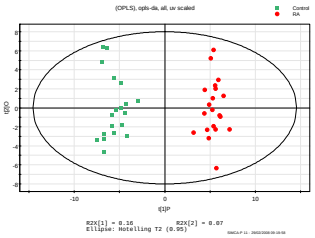


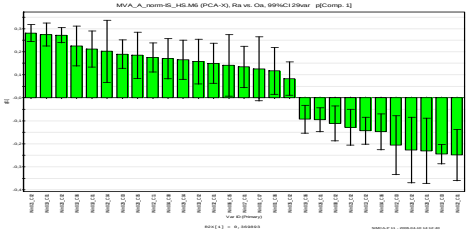


Data

Challenge in modern biology:

maximizing information





GC-TOF/MS-based metabolomics platform

Sample preparation

+ 40 ul heptane

Silylation. 30ul MSTFA + 1h
N-Methyl-*N*-trimethylsilyltrifluoroacetamide

Methoxymation
vortex mix 10 min + 16h RT
30 ul methoxyamine/pyridine

Speed Vac Concentrator
200 ul supernatant

Mix / ultrasonicate
vibration mill / centrifuge

Add methanol
700 µl

Add water
200 µl

Human plasma
100 µl

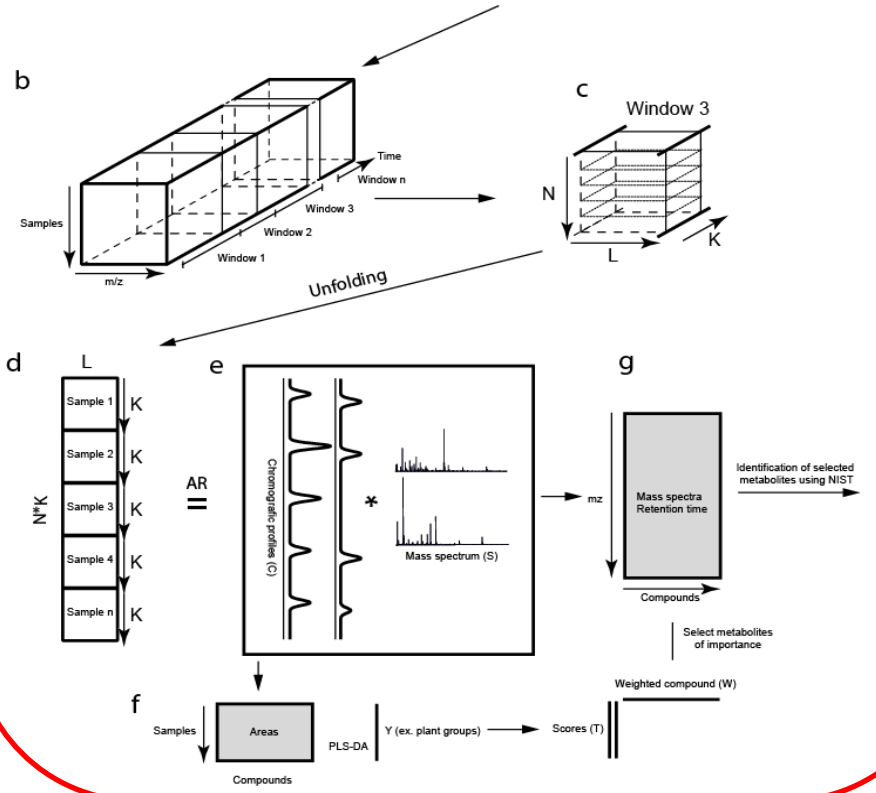
1ul
aliquot

Analysis + Data processing

Agilent 6980 GC
Agilent 7683 Autosampler

GC-TOFMS

Scans ~ 15000
Masses ~ 750
Samples ~ 50



Gullberg, J.; Jonsson, P.; Nordström, A.; Sjöström, M.; Moritz, T.; **Design of experiments: an efficient strategy to identify factors influencing extraction and derivatization of *Arabidopsis thaliana* samples in metabolomic studies with gas chromatography/mass spectrometry**, Analytical Biochemistry, **2004,** 331, 283-295.

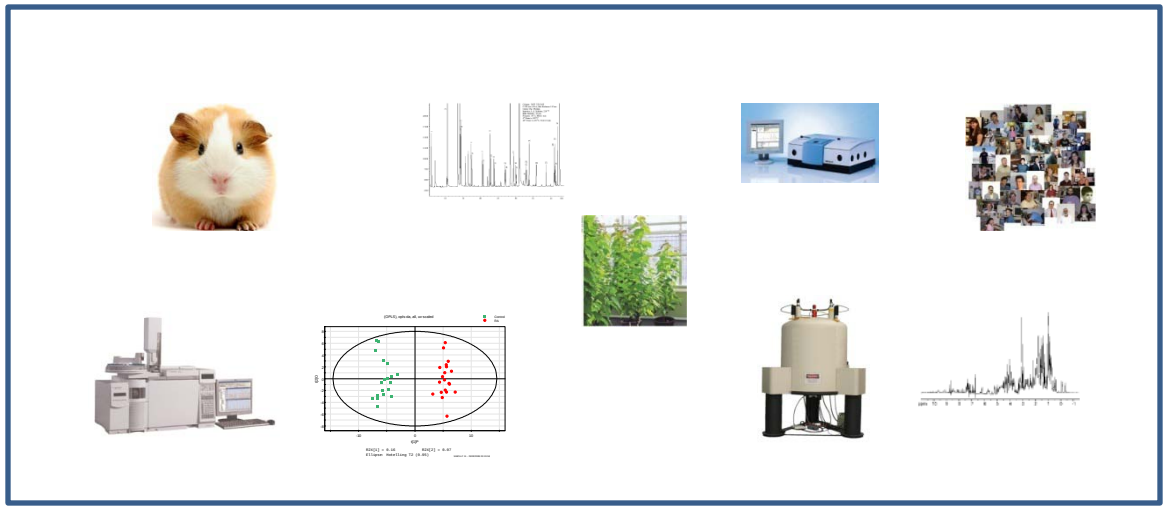
Jiye A.;Trygg, A.;Gullberg, J.; Johansson, A.; Jonsson, P.; Antti, H.; Marklund, S.; Moritz, T.; **Extraction and GC/MS Analysis of the Human Blood** Plasma Metabolome, Analytical Chemistry, **2005**, 77, 8086-8094.

NMR and GC / LC-MS methods - Umeå

- Trees
- Arthrytis in human (>300) and rats (>200)
- Diabetes (2 mouse models, >200 samples)
- Huntington disease
- LC-MS methods for amino-acids in final phase of development. Adding compounds
- LC-MS for lipids and hormones in preparation
- Bacterian and human cell cultures for analysis in GC / LC

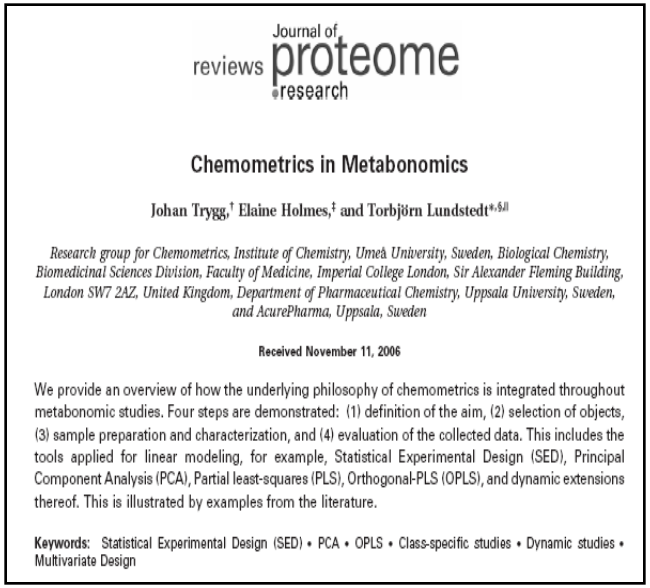
Integration of Chemometrics in metabolomics

- DOE, MVD
- PCA
- MCR
- OPLS (OPLS, O2PLS, OPLS-DA)



REVIEW: metabolomics literature 2002-2006

- **Chemometrics – reduced to a data modelling tool**
  - ANOVA- analysis of variance (hypothesis testing)
  - Overview of data (Principal component analysis)
  - Two class discrimination (PLS-DA, SIMCA)
- **Metabolomics – reduced to NMR/MS based technique**
  - ... with many interesting case studies, samples
- **Chemometrics + Metabonomics**
  - Samples + NMR/MS based characterisation + PCA/PLS-DA



Journal of reviews proteome research

Chemometrics in Metabonomics

Johan Trygg,† Elaine Holmes,‡ and Torbjörn Lundstedt*,§,||

Research group for Chemometrics, Institute of Chemistry, Umeå University, Sweden, Biological Chemistry, Biomedicinal Sciences Division, Faculty of Medicine, Imperial College London, Sir Alexander Fleming Building, London SW7 2AZ, United Kingdom, Department of Pharmaceutical Chemistry, Uppsala University, Sweden, and AcurePharma, Uppsala, Sweden

Received November 11, 2006

We provide an overview of how the underlying philosophy of chemometrics is integrated throughout metabonomic studies. Four steps are demonstrated: (1) definition of the aim, (2) selection of objects, (3) sample preparation and characterization, and (4) evaluation of the collected data. This includes the tools applied for linear modeling, for example, Statistical Experimental Design (SED), Principal Component Analysis (PCA), Partial least-squares (PLS), Orthogonal-PLS (OPLS), and dynamic extensions thereof. This is illustrated by examples from the literature.

Keywords: Statistical Experimental Design (SED) • PCA • OPLS • Class-specific studies • Dynamic studies • Multivariate Design

–Is this enough?

Not many papers had been published...

...that aim for the the whole chain of planning, sampling, experimental characterisation, modelling, visualisation and interpretation...

... especially, regarding validating the hypothesis made based on models.

Trygg, J.; Holmes, E.; Lundstedt, T.; **Chemometrics in Metabonomics**, Journal of Proteome Research, **2007**, 6 (2), 469-479.

Integration Chemometrics/Metabolomics

providing information for studying complex systems

1. **Define the aim**
   - What do we want?
   - What is known already / what more knowledge is needed?
2. **Selection of objects**
   - Design of Experiments (DOE)
     - Samples, time points, replicates...
3. **Sample preparation and characterisation**
   - Experimental protocol optimization
     - Extraction, derivatization, instruments parameters optimization...
     - Randomization of samples for GC/LC/NMR analysis by day, disease/control...
   - Data processing
     - Align peaks, correct baseline, curve resolution, normalisation , scaling
4. **Evaluation/Validation of collected data**
   - Exploratory analysis
   - Multivariate design
   - Interpretation & Visualization
   - Class-specific study
   - Dynamic study

1. Define the aim

– What do we want? Example for disease diagnostics:

	Metabolomics / metabonomics	**Metabolic fingerprinting**	**Metabolite profiling**
Description	Comprehensive analysis with identification and quantification of as many metabolites as possible in a biological system, done in an unbiased way	Fast classification of samples based on metabolite data, without necessarily quantifying or identifying the individual metabolites.	Quantification of a number of pre-defined metabolites
Potential use	Diagnosis + biomarker discovery + biological understanding	Diagnosis method	Diagnosis + biomarker discovery + biological understanding

– What is known already / what more knowledge is needed?

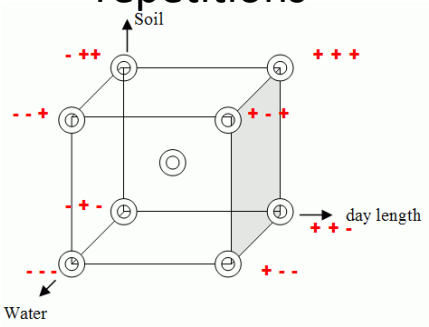
- Literature review
- Known biomarkers
- Other extraction procedures, solvents, instruments

Madsen, R.; Lundstedt, T.; Trygg, J.; **Chemometrics in metabolomics - A review in human disease diagnosis,** Analytica chimica acta, in print.

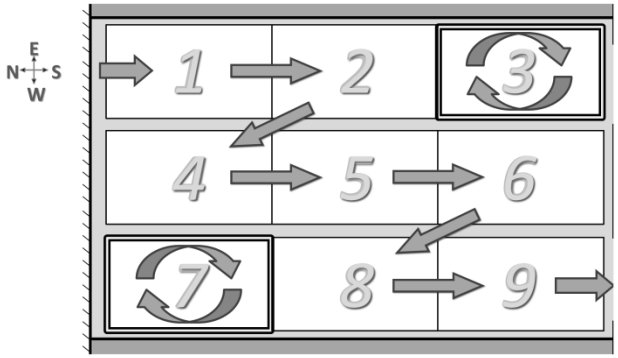
2. Selection of objects

- Design of Experiments (DOE)

Define samples, repetitions



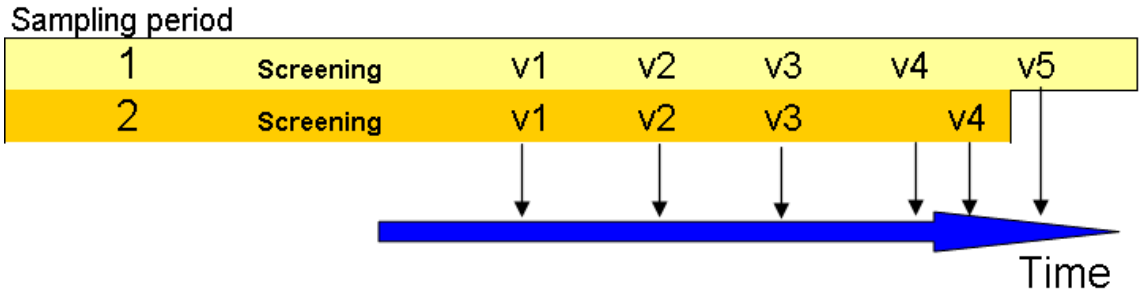
Reduce residual variability



Dynamic studies

- Allow slow/fast responders
- Different sampling times

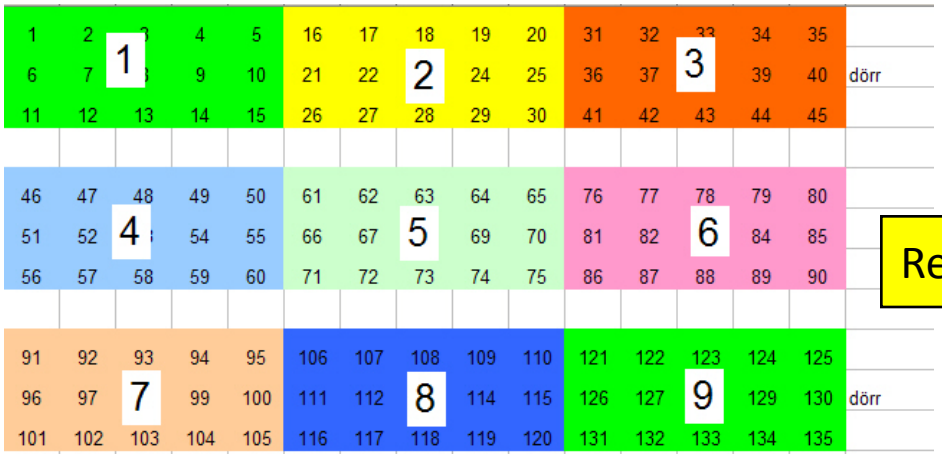
Study design



DoE: Greenhouse design study

”biological variation”

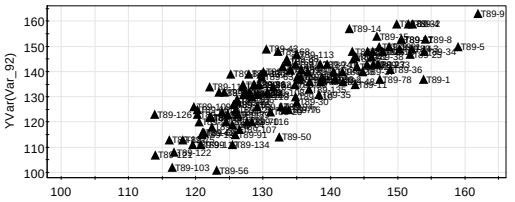
- Experimental design
  - Initial conditions
  - Growth conditions
  - Position in greenhouse
  - Harvesting conditions
  - Grinding / Storage
  - Sample preparation



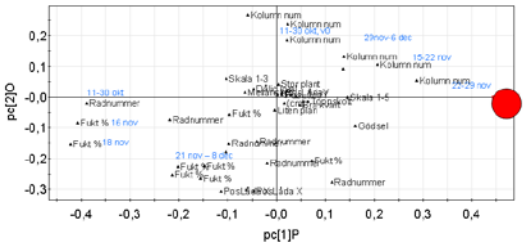
Greenhouse overview

Results

Observed vs Predicted height (cm)



Variable influence



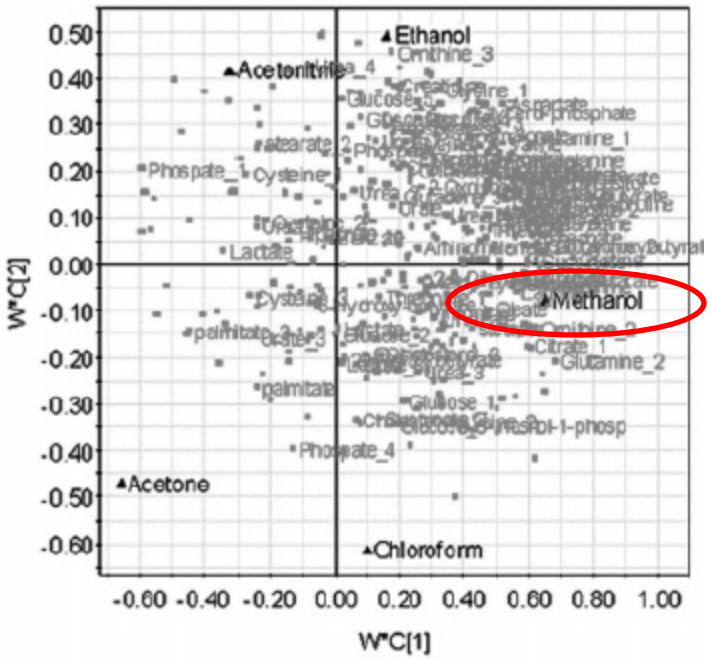
3. Sample preparation and characterization

3.1. Experimental protocol optimization

- Solvents for extraction, derivatization, instruments parameters optimization...
- Randomization of samples for GC/LC/NMR analysis by day, disease/control...

Solvent DoE

ID no	run order	amount, μL					
		methanol	ethanol	acetonitrile	acetone	chloroform	plasma
N1[a]	20, 14, 18	800	0	0	0	0	
N2	6	0	800	0	0	0	
N3	31	0	0	800	0	0	
N4	11	0	0	0	800	0	
N5	24	600	0	0	0	200	
N6	18	0	600	0	0	200	
N7[b]	3	0	0	600	0	200	
N8[c]	19, 12, 21	0	0	0	600	200	
N9	1	0	0	0	735	65	
N10	23	0	0	535	265	0	
N11	15	0	0	265	535	0	
N12	30	0	535	0	265	0	
N13	25	0	265	0	535	0	
N14	16	0	535	265	0	0	
N15	27	0	265	535	0	0	
N16[a]	29, 9, 17	665	0	0	0	135	
N17	4	535	0	0	265	0	
N18	5	265	0	0	535	0	
N19	13	535	0	265	0	0	
N20	22	265	0	535	0	0	
N21	32	535	265	0	0	0	
N22	8	265	535	0	0	0	
N23	7	0	235	235	235	100	
N24[d]	26, 2, 33, 10	200	200	200	200	0	



3. Sample preparation and characterization

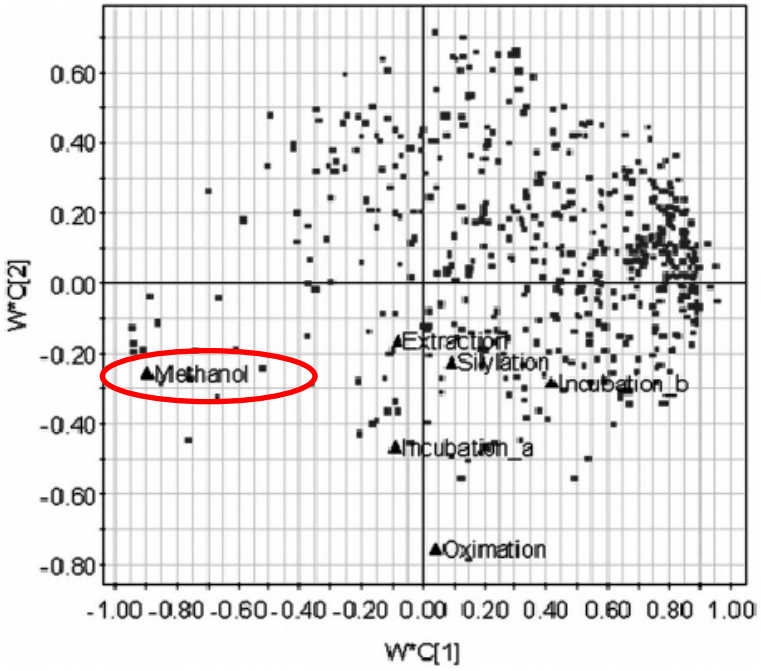
3.1. Experimental protocol optimization

- Solvents for extraction, derivatization, instruments parameters optimization...
- Randomization of samples for GC/LC/NMR analysis by day, disease/control...

Derivatization DoE

expt no.	methanol, μL	incubation[a] °C, min	extraction min	incubation[b] °C, min	oximation °C, h	silylation °C, h
N1	700	0, 10	1	0, 10	20, 16	20, 1
N2	700	70, 30	1	0, 10		
N3	700	0, 10	3	0, 10		
N4	700	70, 30	3	0, 10		
N5	700	0, 10	1	−20, 120		
N6	700	70, 30	1	−20, 120		
N7	700	0, 10	3	−20, 120		
N8	700	70, 30	3	−20, 120		
N9	900	0, 10	1	0, 10		
N10	900	70, 30	1	0, 10		
N11	900	0, 10	3	0, 10		
N12	900	70, 30	3	0, 10		
N13	900	0, 10	1	−20, 120		
N14	900	70, 30	1	−20, 120		
N15	900	0, 10	3	−20, 120		
N16	900	70, 30	3	−20, 120		
N17	800	0, 10	2	0, 10		
N18	800	0, 10	2	0, 10		
N19	800	0, 10	2	0, 10		
N20	800	0, 10	2	0, 10		

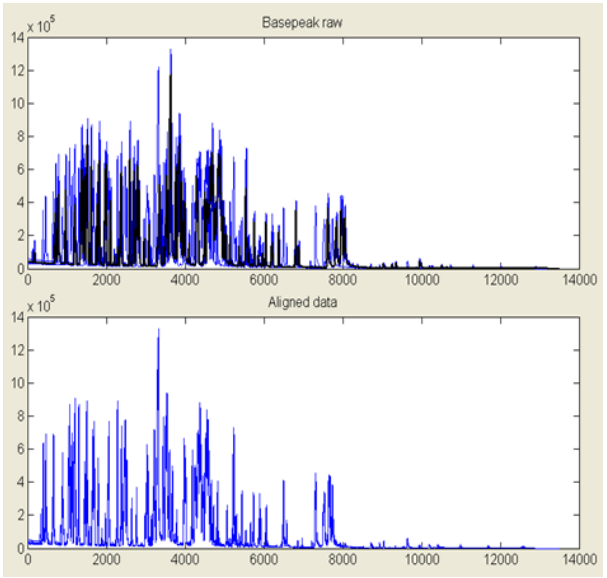
[a] Temperature and duration before extraction. [b] Temperature and duration after extraction.

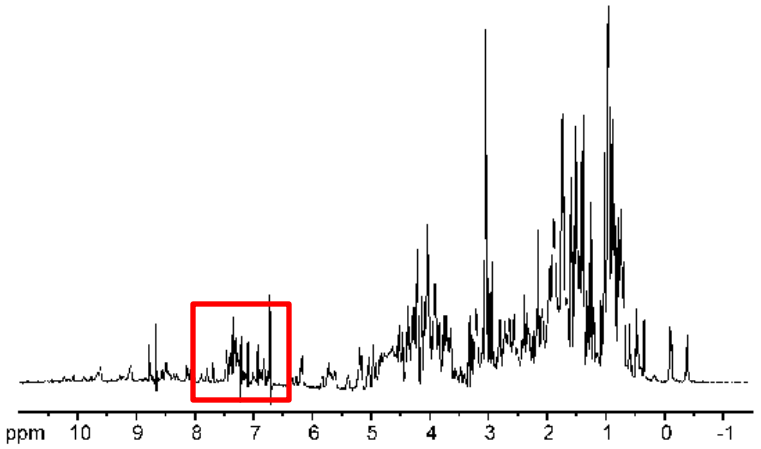


3. Sample preparation and characterization

3.2. Data processing

- Align peaks by a reference spectrum
- Region selection
- Baseline correction
- Normalisation
- Scaling
- Multivariate curve resolution (ex: GC-MS)





Data pre-processing
Methods in GC-MS, LC-MS, NMR

- Baseline correction
- Alignment
- Time-window setting (GC-MS, LC-MS)
- MCR

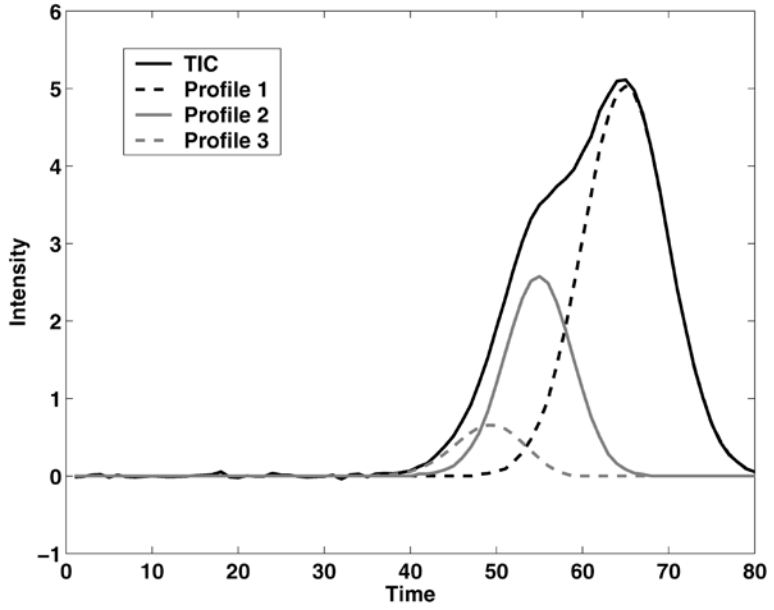
Multivariate curve resolution

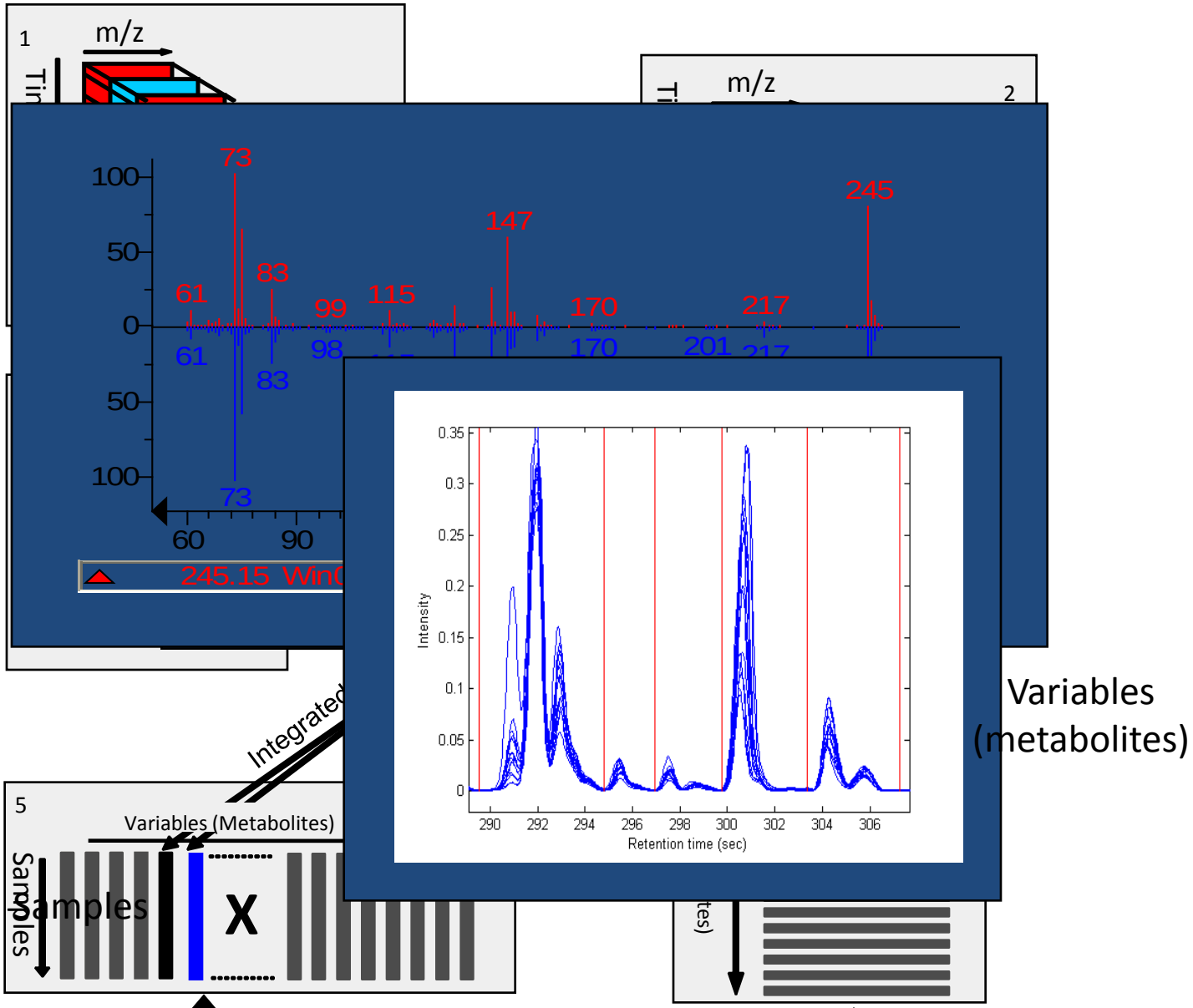
resolve hyphenated data into chromatographic and spectral profiles.

Solves: $\mathbf{X=C*S^T+E}$

C=Chromatographic profiles
S=Spectroscopic profiles

$\mathbf{X}=\mathbf{C}*\mathbf{S}^T=C_1*S_1^T+C_2*S_2^T+....+C_n*S_n^T+E$



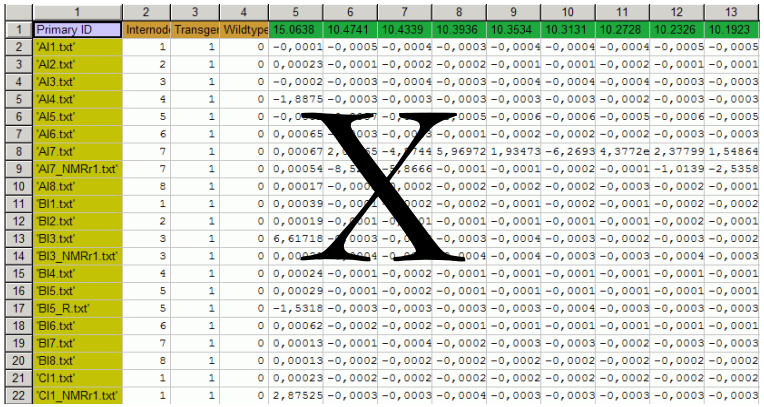


Jonsson, P.; Johansson, A. I. et al. High-Throughput Data Analysis for Detecting and Identifying Differences between Samples in GC/MS-Based Metabolomic Analyses. *Analytical Chemistry* **2005,** 77, (17), 5635-5642.

4. Evaluation/validation of collected data

K

N



	1	2	3	4	5	6	7	8	9	10	11	12	13
1	Primary ID	Internod	Transge	Wildtype	15.0638	10.4741	10.4339	10.3936	10.3534	10.3131	10.2728	10.2326	10.1923
2	'AI1.txt'	1	1	0	-0,0001	-0,0005	-0,0004	-0,0003	-0,0004	-0,0004	-0,0004	-0,0005	-0,0005
3	'AI2.txt'	2	1	0	0,00023	-0,0001	-0,0002	-0,0002	-0,0001	-0,0001	-0,0002	-0,0001	-0,0001
4	'AI3.txt'	3	1	0	-0,0002	-0,0003	-0,0004	-0,0003	-0,0004	-0,0004	-0,0004	-0,0003	-0,0003
5	'AI4.txt'	4	1	0	-1,8875	-0,0003	-0,0003	-0,0003	-0,0003	-0,0003	-0,0002	-0,0003	-0,0003
6	'AI5.txt'	5	1	0	-0,0	-0,0	-0,0	-0,0005	-0,0006	-0,0006	-0,0005	-0,0006	-0,0005
7	'AI6.txt'	6	1	0	0,00065	-0,0003	-0,0	-0,0001	-0,0002	-0,0002	-0,0002	-0,0003	-0,0003
8	'AI7.txt'	7	1	0	0,00067	2,4	-4,4744	5,96972	1,93473	-6,2693	4,3772e	2,37799	1,54864
9	'AI7_NMRr1.txt'	7	1	0	0,00054	-8,5	-0,8666	-0,0001	-0,0001	-0,0002	-0,0001	-1,0139	-2,5358
10	'AI8.txt'	8	1	0	0,00017	-0,000	-0,0002	-0,0002	-0,0002	-0,0002	-0,0003	-0,0002	-0,0001
11	'BI1.txt'	1	1	0	0,00039	-0,000	-0,0002	-0,0002	-0,0001	-0,0002	-0,0001	-0,0002	-0,0002
12	'BI2.txt'	2	1	0	0,00019	-0,0	-0,0	-0,0001	-0,0001	-0,0001	-0,0001	-0,0002	-0,0001
13	'BI3.txt'	3	1	0	6,61718	-0,0	-0,0	-0,0003	-0,0004	-0,0003	-0,0002	-0,0003	-0,0002
14	'BI3_NMRr1.txt'	3	1	0	0,0	-0,0	-0,0	-0,0004	-0,0004	-0,0003	-0,0003	-0,0004	-0,0003
15	'BI4.txt'	4	1	0	0,00024	-0,0001	-0,0002	-0,0001	-0,0001	-0,0001	-0,0001	-0,0001	-0,0001
16	'BI5.txt'	5	1	0	0,00029	-0,0001	-0,0002	-0,0001	-0,0001	-0,0001	-0,0001	-0,0001	-0,0001
17	'BI5_R.txt'	5	1	0	-1,5318	-0,0003	-0,0003	-0,0003	-0,0003	-0,0004	-0,0003	-0,0003	-0,0003
18	'BI6.txt'	6	1	0	0,00062	-0,0002	-0,0002	-0,0001	-0,0001	-0,0001	-0,0001	-0,0001	-0,0001
19	'BI7.txt'	7	1	0	0,00013	-0,0001	-0,0004	-0,0002	-0,0003	-0,0003	-0,0002	-0,0003	-0,0003
20	'BI8.txt'	8	1	0	0,00013	-0,0002	-0,0002	-0,0002	-0,0002	-0,0002	-0,0002	-0,0002	-0,0002
21	'CI1.txt'	1	1	0	0,00023	-0,0002	-0,0002	-0,0002	-0,0002	-0,0002	-0,0002	-0,0002	-0,0002
22	'CI1_NMRr1.txt'	1	1	0	2,87525	-0,0003	-0,0003	-0,0004	-0,0003	-0,0003	-0,0003	-0,0003	-0,0003

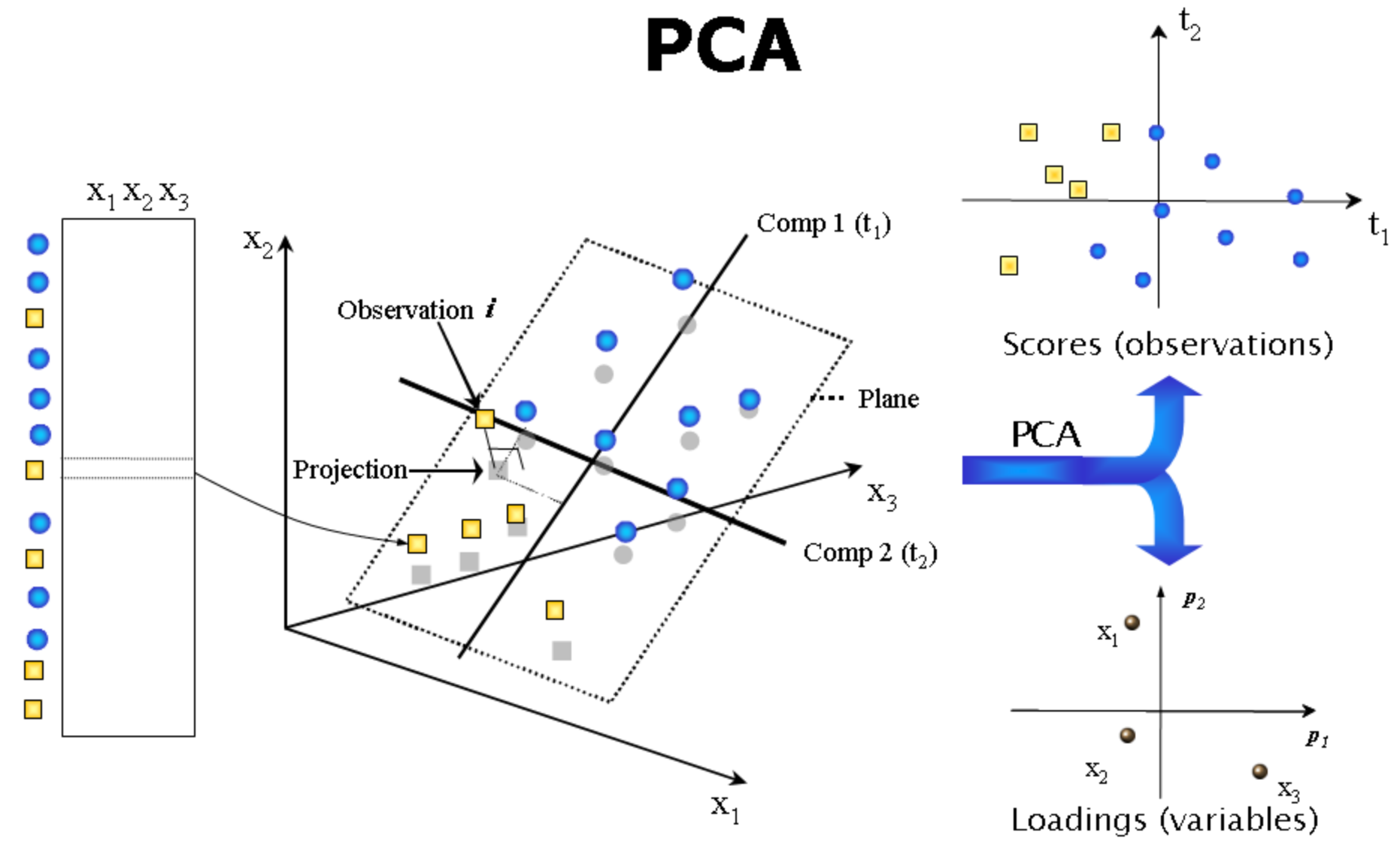
X

What to do?

- Overview of data
- Exploratory analysis
- Multivariate design
- Class-specific study
- Dynamic study
- Visualization
- Interpretation

Principal Component Analysis

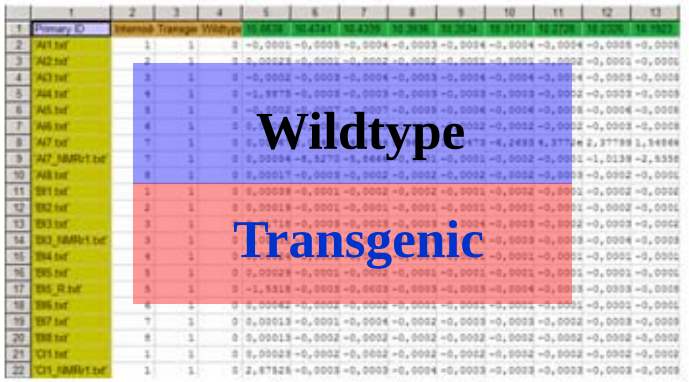
Overview, outliers, groups, tendencies



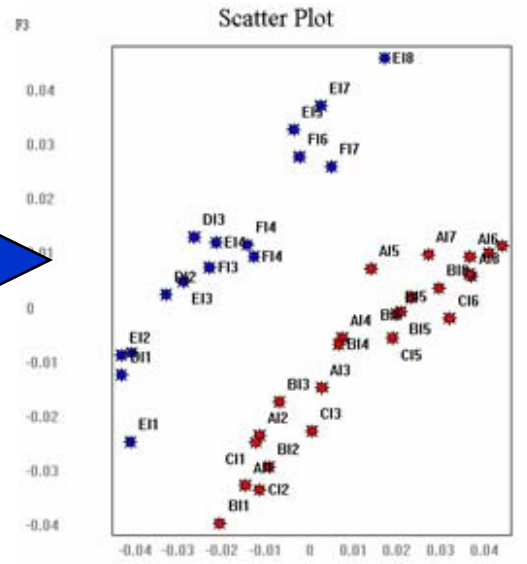
Overview of data

GC/MS metabolite profiles

Samples

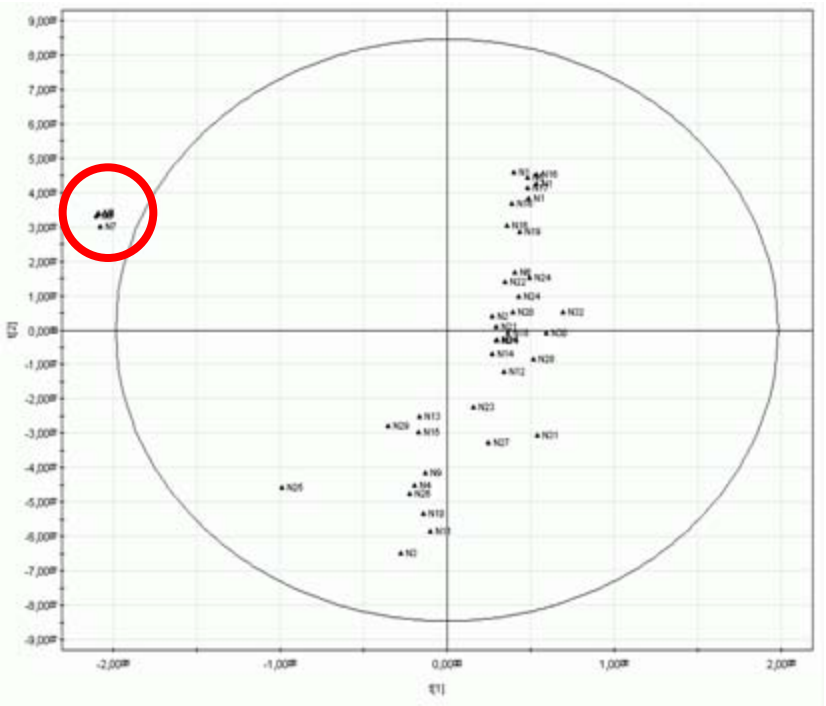


PCA



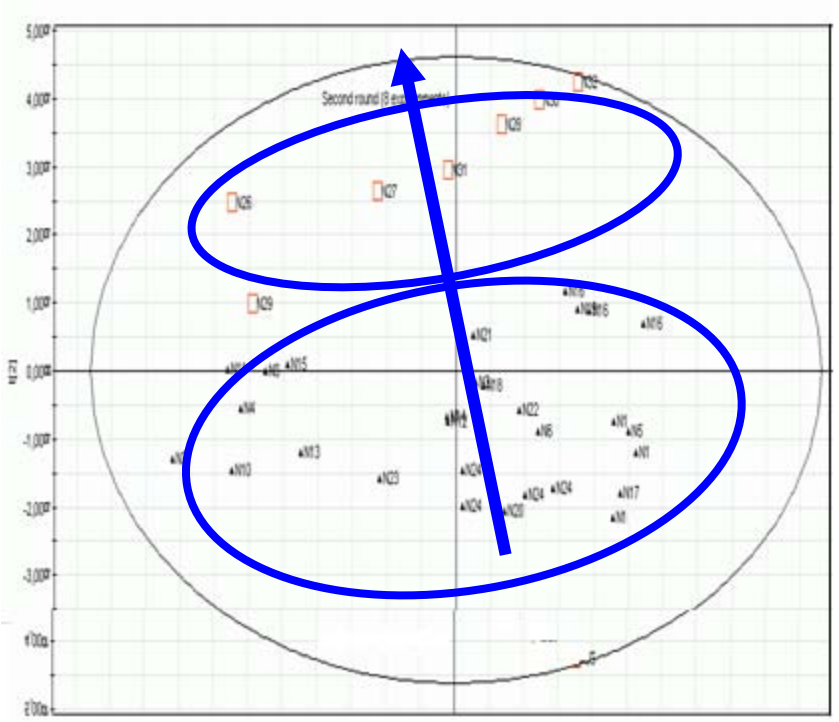
Overview & data exploration
Example: PCA on GC/MS spectra on human plasma

Outlier detection



Two phase problem
Chloroform /Acetone

- Drift / robustness



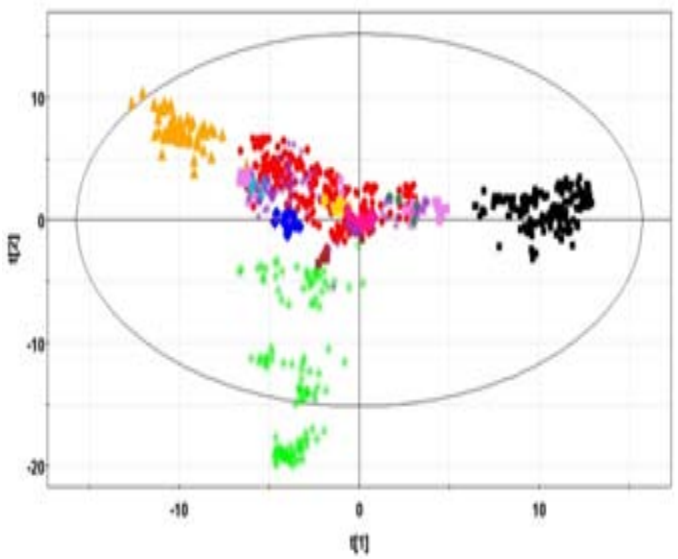
Tendences observed

PCA for Multivariate design

Example for choice of calibration and validation sets

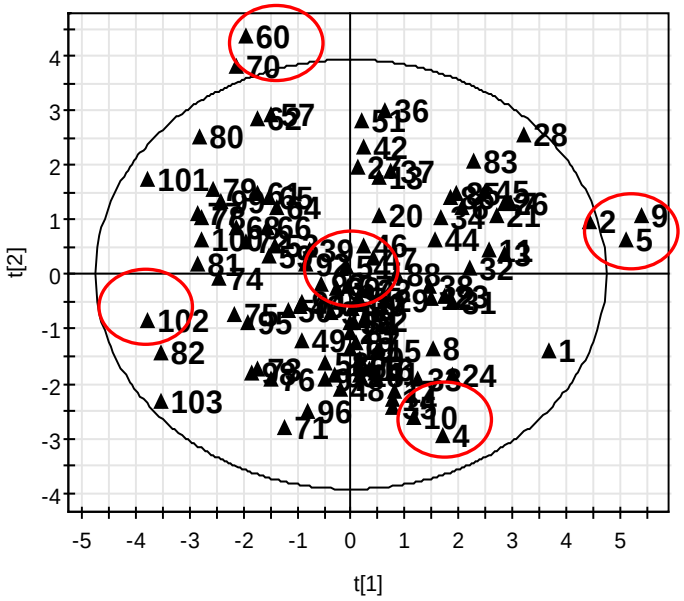
Groupings in data

Select subset from each meaningful cluster



Selection from a database

Diverse selection



Multivariate method – Get results

Many different methods to choose from

Linear methods

Full rank methods

- Multiple Linear Regression (MLR)
- Stepwise MLR
- Ridge Regression

Latent variable regression methods

- Principal Component Regression (PCR)
- **Partial Least Squares (PLS)**
- ***Orthogonal Projections to Latent Structures (OPLS)***

Non-Linear methods

- Neural Networks (NN)
- Support Vector Machines (SVM)
- Regression trees

Validation = f(Prediction,Interpretation)

- **Prediction is part of the *statistical validation*, many tools exist**
  - External predictions (RMSEP value), cross-validation
  - Many are familiar with these

Examples:
1. Predict concentration of active substance in tablet production with NIR spectroscopy
2. Predict viscosity in pulp using NIR spectroscopy
3. Predict severity of coronary heart disease (CHD) on biofluids with NMR
4. Predict biological activity from amino acid sequence (QSAR)

- **Interpretation is part of the *chemical / biological validation* (what does it mean?)**
  - No direct quantifiable measure as RMSEP exists
  - Model interpretation (e.g. regression coefficients)
    - Pure constituent spectrum
    - ”Sequence motif”
    - ”Functional profile”
  - Not as common, requires much more effort (communication between disciplines)

Both are related & complementary in validating models/results

Validation in disease diagnostics

	Statistical **results valid from statistical point of view**	**Biological** **results are relevant to study**
Minimum	- Prediction of **validation dataset** (not CV). - **3 classes**: Controls, disease and related disease control group. - Realistic measure for the error in the classification of new samples from the **same patient population**. Will **NOT guard against sampling bias nor drift** in analytical instruments.	- **Identification** of differentially regulated metabolites and their associated metabolic pathways. - Establish whether the results are in accordance with known facts or are spurious, e.g. products of uncontrolled factors.
Recommended	**- Follow-up study** in a **separate population**, analyzed separately in a **different lab**. - Realistic measure of the expected error in classification of **new patients.** - **Guard against sampling bias and drift** in analytical instruments.	- **Follow-up study** in a separate population analyzed separately in a **different lab**. - **Only reliable way** to reveal whether the observed metabolic perturbations are in fact a **product of the investigated disease**, or a product of sample bias.

Madsen, R.; Lundstedt, T.; Trygg, J.; **Chemometrics in metabolomics - A review in human disease diagnosis,** Analytica chimica acta, in print.

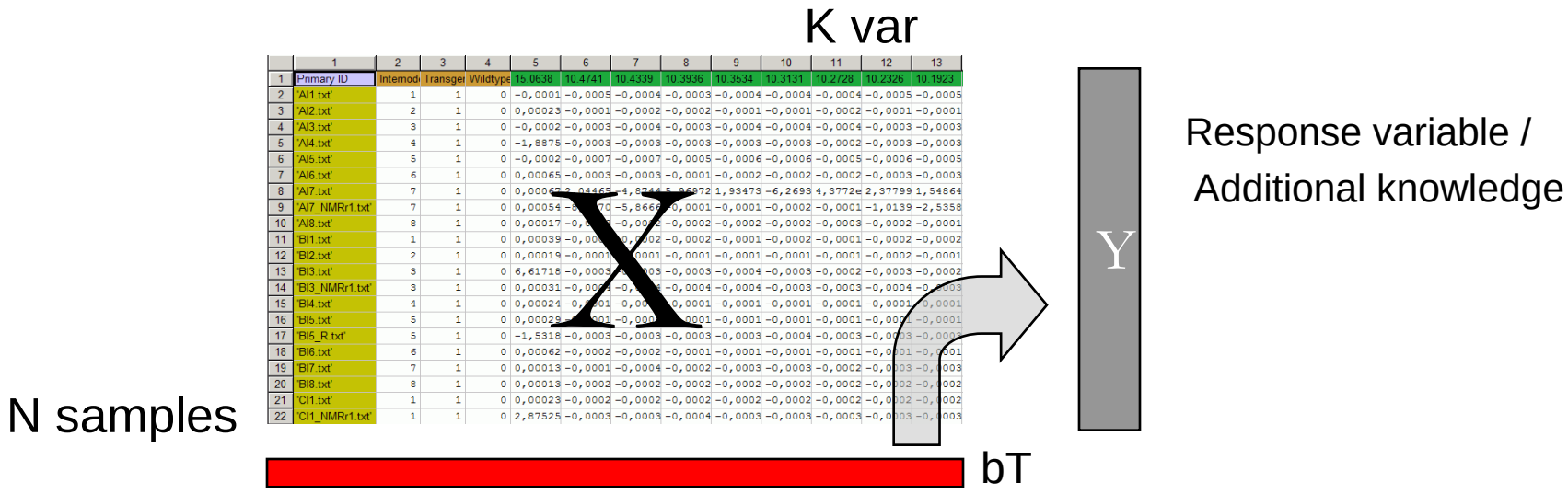
Multivariate calibration
Discriminant analysis / classification

Multivariate calibration, MC

Model the relation between two blocks of data

Samples - Powders, molecules, industrial process samples, plasma, tissue…

Sample characterisation - Spectrometers (NIR,UV, IR, NMR, MS), chromatography, chemical descriptors, gene-arrays, metabolites



	1	2	3	4	5	6	7	8	9	10	11	12	13
1	Primary ID	Internod	Transger	Wildtype	15.0638	10.4741	10.4339	10.3936	10.3534	10.3131	10.2728	10.2326	10.1923
2	'AI1.txt'	1	1	0	-0,0001	-0,0005	-0,0004	-0,0003	-0,0004	-0,0004	-0,0004	-0,0005	-0,0005
3	'AI2.txt'	2	1	0	0,00023	-0,0001	-0,0002	-0,0002	-0,0001	-0,0001	-0,0002	-0,0001	-0,0001
4	'AI3.txt'	3	1	0	-0,0002	-0,0003	-0,0004	-0,0003	-0,0004	-0,0004	-0,0004	-0,0003	-0,0003
5	'AI4.txt'	4	1	0	-1,8875	-0,0003	-0,0003	-0,0003	-0,0003	-0,0003	-0,0002	-0,0003	-0,0003
6	'AI5.txt'	5	1	0	-0,0002	-0,0007	-0,0007	-0,0005	-0,0006	-0,0006	-0,0005	-0,0006	-0,0005
7	'AI6.txt'	6	1	0	0,00065	-0,0003	-0,0003	-0,0001	-0,0002	-0,0002	-0,0002	-0,0003	-0,0003
8	'AI7.txt'	7	1	0	0,000673	0,04165	4,8744	5,26072	1,93473	-6,2693	4,3772e	2,37799	1,54864
9	'AI7_NMRr1.txt'	7	1	0	0,00054	-0,0	-5,8666	-0,0001	-0,0001	-0,0002	-0,0001	-1,0139	-2,5358
10	'AI8.txt'	8	1	0	0,00017	-0,0	-0,0	-0,0002	-0,0002	-0,0002	-0,0003	-0,0002	-0,0001
11	'BI1.txt'	1	1	0	0,00039	-0,0	-0,0	-0,0002	-0,0002	-0,0001	-0,0002	-0,0001	-0,0002
12	'BI2.txt'	2	1	0	0,00019	-0,0	-0,0	-0,0001	-0,0001	-0,0001	-0,0001	-0,0002	-0,0001
13	'BI3.txt'	3	1	0	6,61718	-0,0003	-0,0	-0,0003	-0,0004	-0,0003	-0,0002	-0,0003	-0,0002
14	'BI3_NMRr1.txt'	3	1	0	0,00031	-0,0004	-0,0	-0,0004	-0,0004	-0,0003	-0,0003	-0,0004	-0,0003
15	'BI4.txt'	4	1	0	0,00024	-0,0	-0,0	-0,0001	-0,0001	-0,0001	-0,0001	-0,0001	-0,0001
16	'BI5.txt'	5	1	0	0,00029	-0,0	-0,0	-0,0001	-0,0001	-0,0001	-0,0001	-0,0001	-0,0001
17	'BI5_R.txt'	5	1	0	-1,5318	-0,0003	-0,0003	-0,0003	-0,0003	-0,0004	-0,0003	-0,0003	-0,0003
18	'BI6.txt'	6	1	0	0,00062	-0,0002	-0,0002	-0,0001	-0,0001	-0,0001	-0,0001	-0,0001	-0,0001
19	'BI7.txt'	7	1	0	0,00013	-0,0001	-0,0004	-0,0002	-0,0003	-0,0003	-0,0002	-0,0003	-0,0003
20	'BI8.txt'	8	1	0	0,00013	-0,0002	-0,0002	-0,0002	-0,0002	-0,0002	-0,0002	-0,0002	-0,0002
21	'CI1.txt'	1	1	0	0,00023	-0,0002	-0,0002	-0,0002	-0,0002	-0,0002	-0,0002	-0,0002	-0,0002
22	'CI1_NMRr1.txt'	1	1	0	2,87525	-0,0003	-0,0003	-0,0004	-0,0003	-0,0003	-0,0003	-0,0003	-0,0003

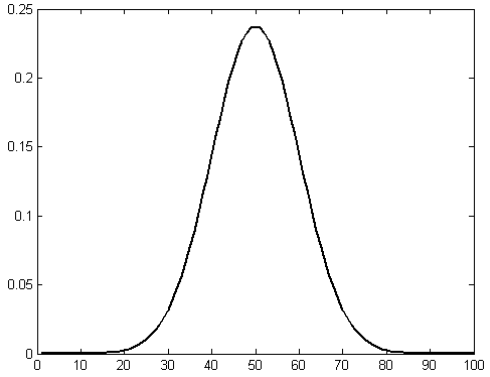
- Focus modelling towards known information (concentration, groupings)
- Model the relation between blocks of data (same samples, different spectra)

Linear prediction model: $y = Xb + f$

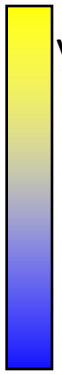
Focus: How to solve for b?

Objective: Provide good fit to estimate y, good predictions for future samples

Example: One component system

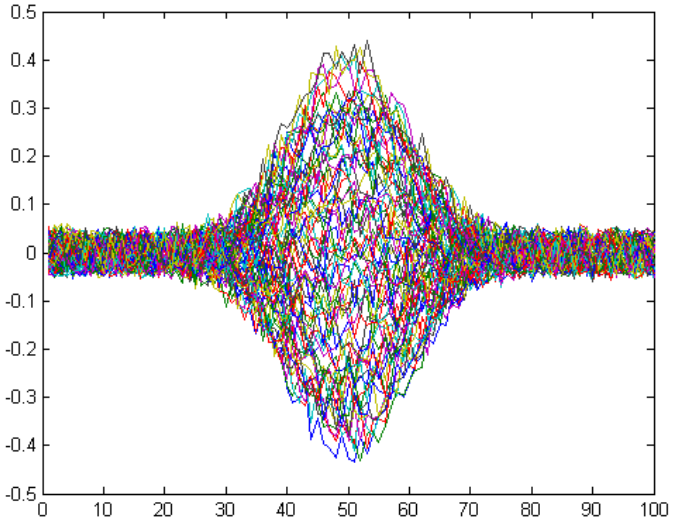


Spectral profile of
Y-predictive component



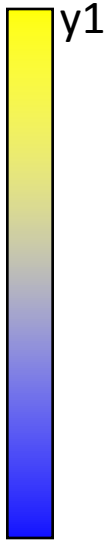
X matrix

100



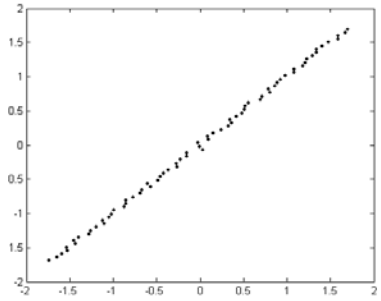
70

PLS
RR
NN

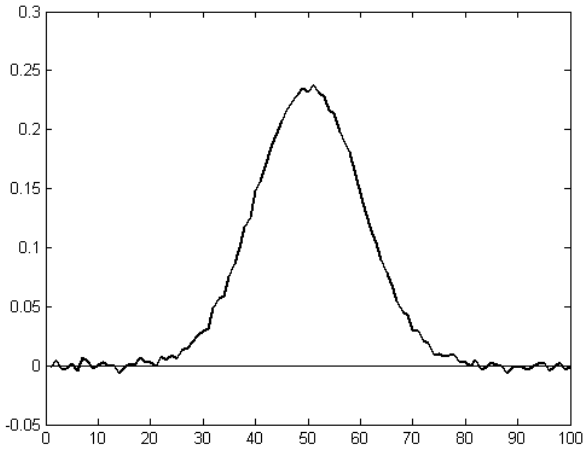


Example: Modeling 1-component model

PLS regression

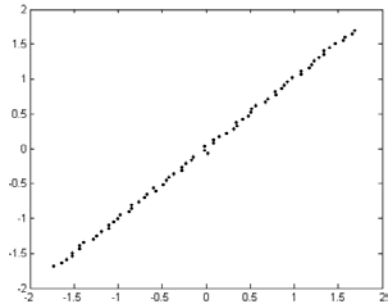


Observed vs Predicted

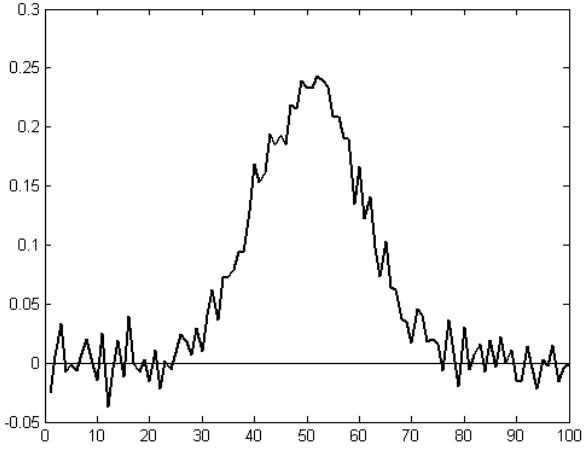


b coefficients

Ridge Regression

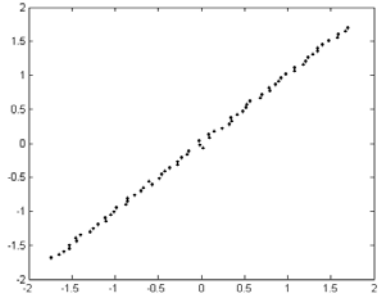


Observed vs Predicted

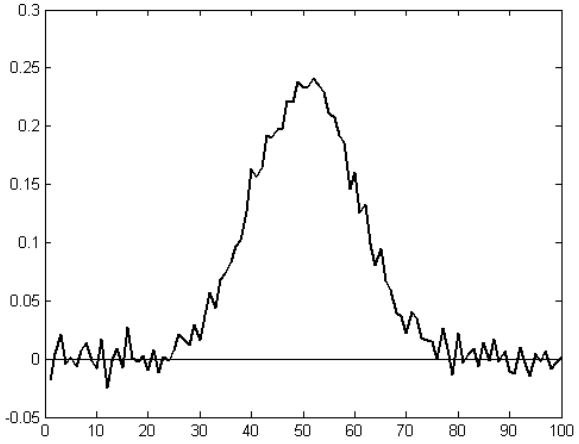


b coefficients

Linear Neural Net



Observed vs Predicted



b coefficients

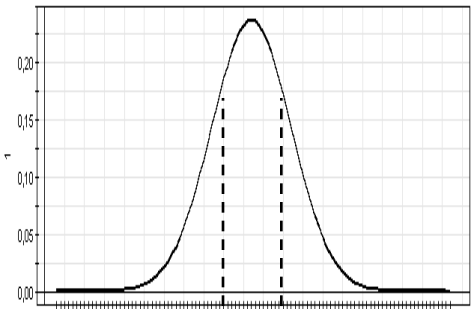
But... Chemical / biological data are complex

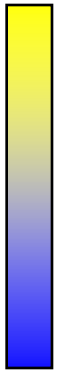
- Lots of unknown systematic variation – mostly due to poor knowledge…
  - strong dietary, environmental, hormonal variations, etc…
  - Experimental variation, sampling, instrumental variation
  - Input material varies with supplier

- Measured signal is the sum of many contributing factors
  - Pharmaceutical tablet formulation (e.g. binders, fillers, active drug, lubricant)
  - Human urine sample (e.g. genetics, diet, gender, age, stress, disease)
  - Plant biotech / Pulp & paper (e.g. wood species, cellulose & lignin content, water, age)
  - In QSAR the molecular descriptor profile is a function of its chemical and biological property/activity/function

Example: simulation with two component system (overlap)

Spectral profile of
Y-predictive component

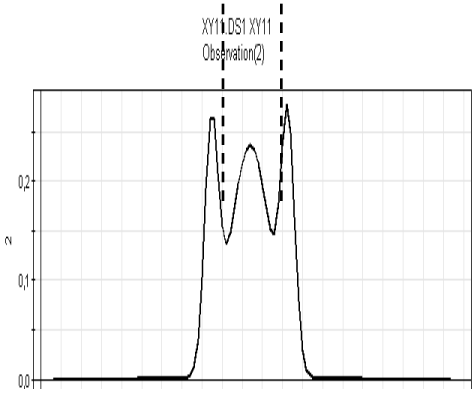




y1

y1 ⊥ y2

y2

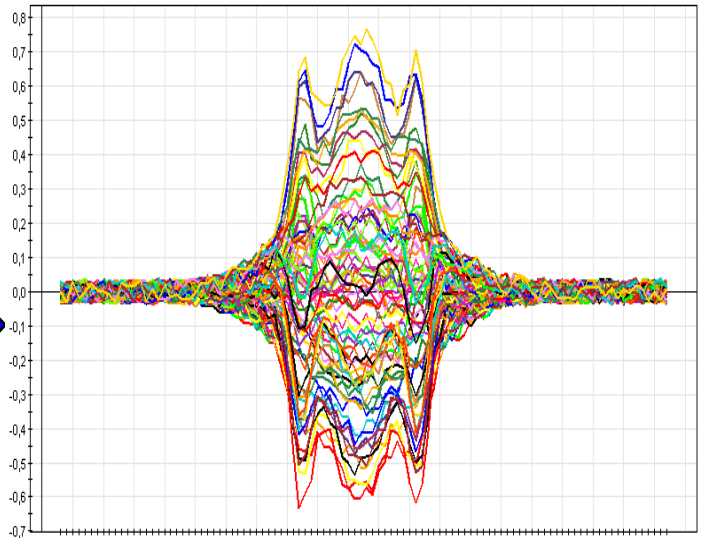


Spectral profile of
Y-orthogonal component

X matrix

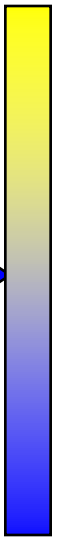
100

70



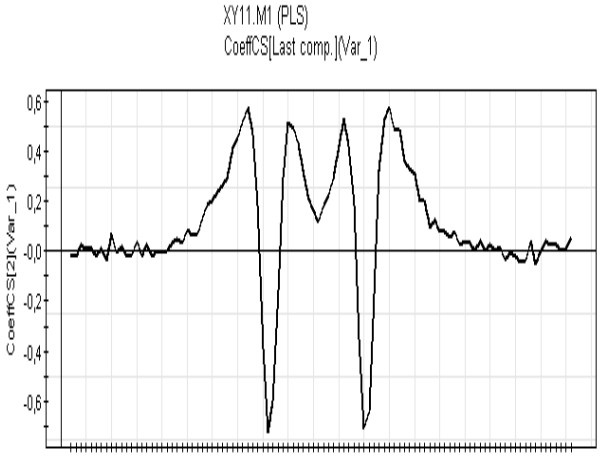
PLS

y1



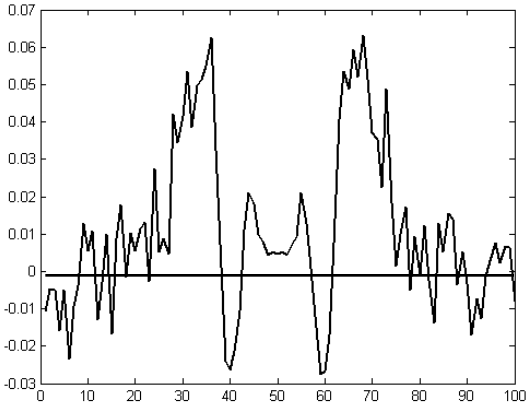
Example: Two Gaussian peaks
Model interpretation by coefficient profile

PLS regression



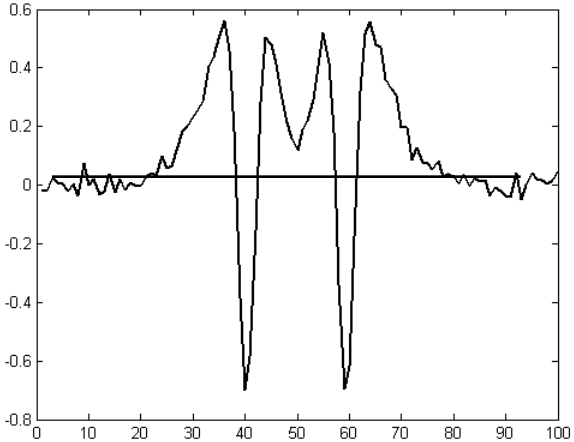
Coefficients b

Ridge Regression



Coefficients b

Linear Neural Net



Coefficients b

Negative dips observed!

PLS - Regression coefficients [$\mathbf{b}_1$ $\mathbf{b}_2$ …], one for each Y-variable

what do they mean?

$$y_1 = Xb_1 + f_1$$

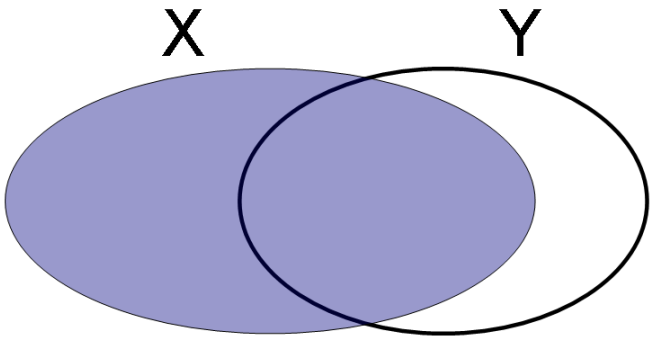
1. The regression coefficient vector **b** *does not* represent the estimated pure constituent spectrum

2. Its profile must be ***orthogonal to all other known and unknown*** constituents in **X**

(Otherwise it will not be good for prediction)

Model overview

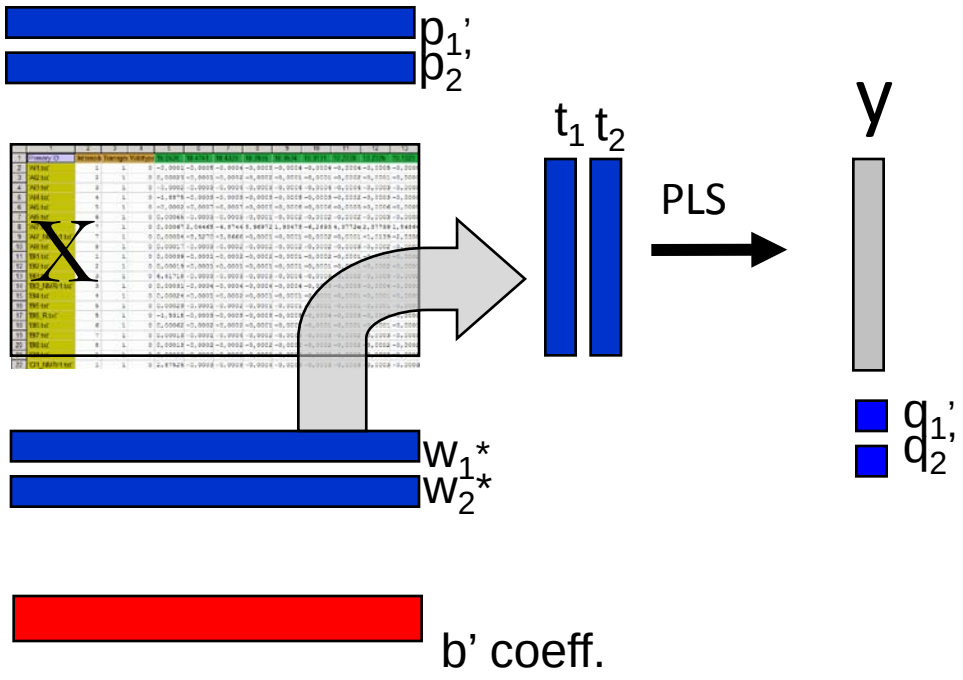
PLS, MLR, PCR, RR etc...



PLS NIPALS (1980's)
Wold, Martens and colleagues

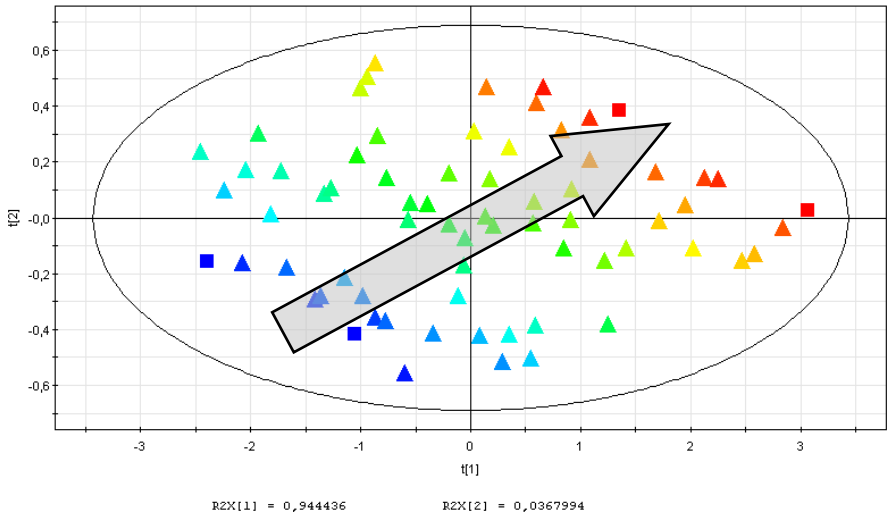
$X=TP' + E$

$y=Tc' + f$

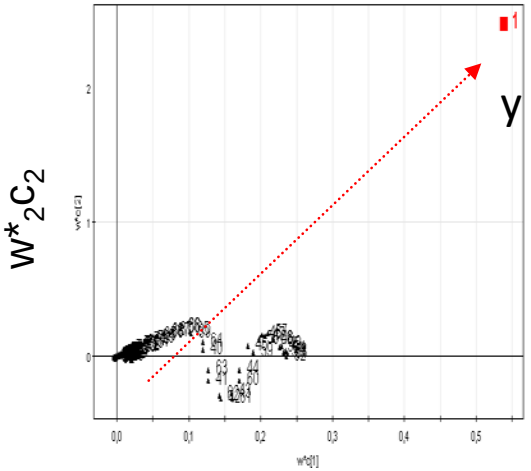


PLS model

Example: Single-Y, two component system



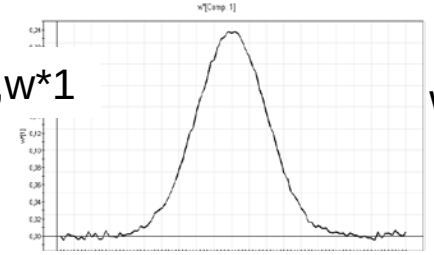
Loading plot



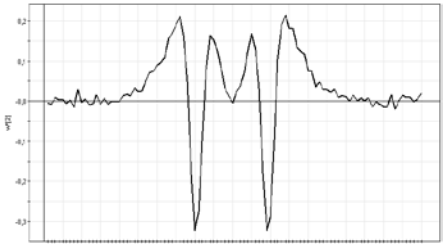
94% variation

3.7% variation

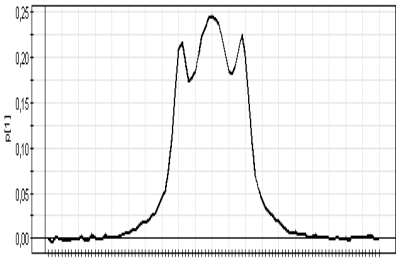
w1,w*1



w*2

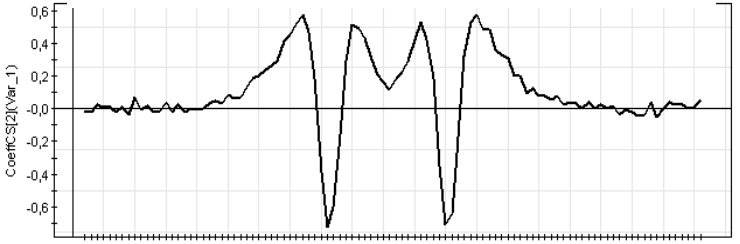


p1



p2

Regression coefficients, b



What to do, and interpret?

1. Use preprocessing filters
   - MSC, SNV, 1,2nd derivatives, wavelet, Fourier, etc
     - Can remove pertinent information, loadings...
2. Avoid this variation
   - Improve instrument, sample preparation, and so on …
     - Requires much knowledge, often not realistic
3. Why not …

   Separately model the Y-predictive and Y-orthogonal variation?
   - Understand what's going on!!
   - Orthogonal signal correction method [Wold S et *al.* 1998]
   - OPLS method [Trygg J & Wold S. 2002]

The O-PLS framework

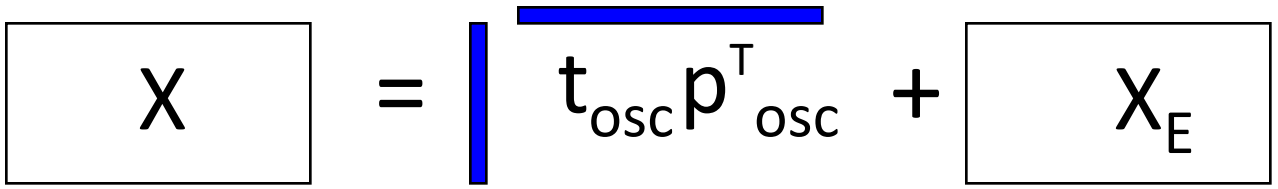
Orthogonal Signal Correction (OSC)

OSC, Wold et al. (1998), Sjöblom et al. (1998), DOSC, Westerhuis et al. (2001)

POSC, Trygg et al. (2001), OSC, Fearn (2000), Höskuldsson(2001)

- Basic idea, perform an ”inverse PLS model” :
  Remove *structured noise* (i.e. systematic) from **X** not correlated to **Y**

$$\mathbf{X} = \mathbf{t}_{osc}\,\mathbf{p}^{T}_{osc} + \mathbf{X}_{E} \qquad \text{(i.e. } \mathbf{Y}^{T}\mathbf{t}_{osc}=\mathbf{0}\text{)}$$



Estimate calibration model (e.g. PLS) based on the filtered $\mathbf{X}_{E}$

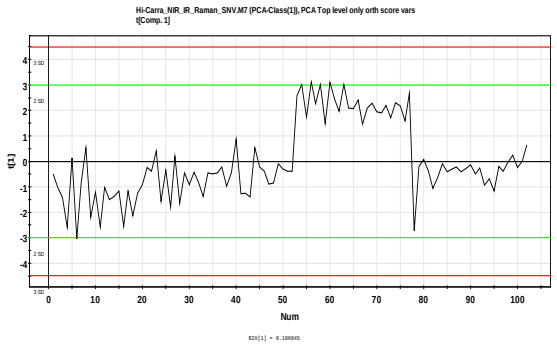
Y-Orthogonal variation, what is it?

”Impact of nothingness” – Gottfries et al.

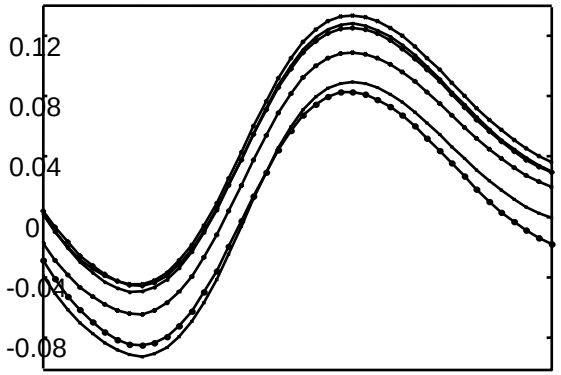
For example...

- Experimental problems
- Side reactions causing biproducts
- Non-linearities (e.g. kinetics)
- Within class variation
- Sampling issues
- and so on...

Drift



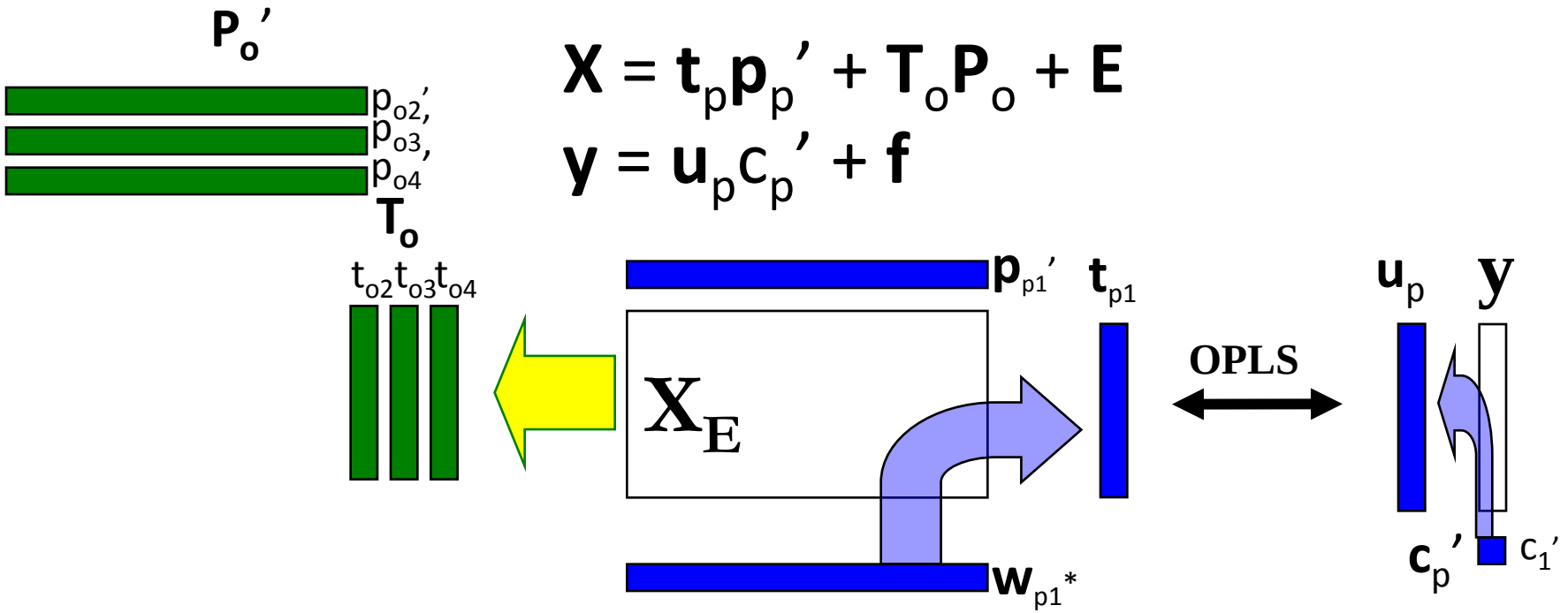
Baseline effects



Gottfries, J.; Johansson, E.; Trygg, J.; **On the impact of uncorrelated variation in regression mathematics**. Journal of chemometrics, **2008**, 22, 565-570.

Orthogonal PLS (OPLS)

Focus modelling towards known information



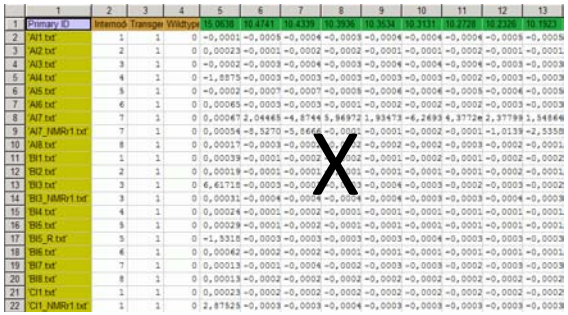
- Only a single Y-related component
- Used for two-class discriminant analysis

Trygg J.; Wold, S.; **Orthogonal projections to latent structures (O-PLS)**, Journal of Chemometr.ics, **2002**, **16**, 119-128

Multi-block modeling

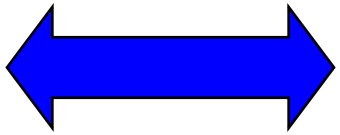
- Compare & Integrate **X** and **Y** in terms of….
  - Analytical platforms, Experimental conditions, Process step, Time (drift), Replication, Pre-treatments, ...
- Understand…
  - Overlap? What is jointly related?
  - What is unique for X, for Y?

Transcriptomics

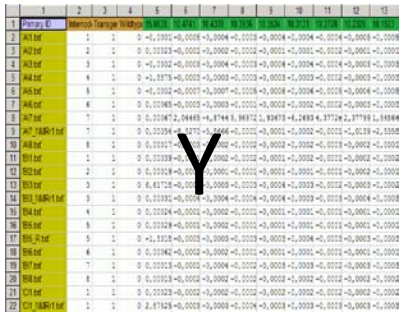


X

O2PLS modelling



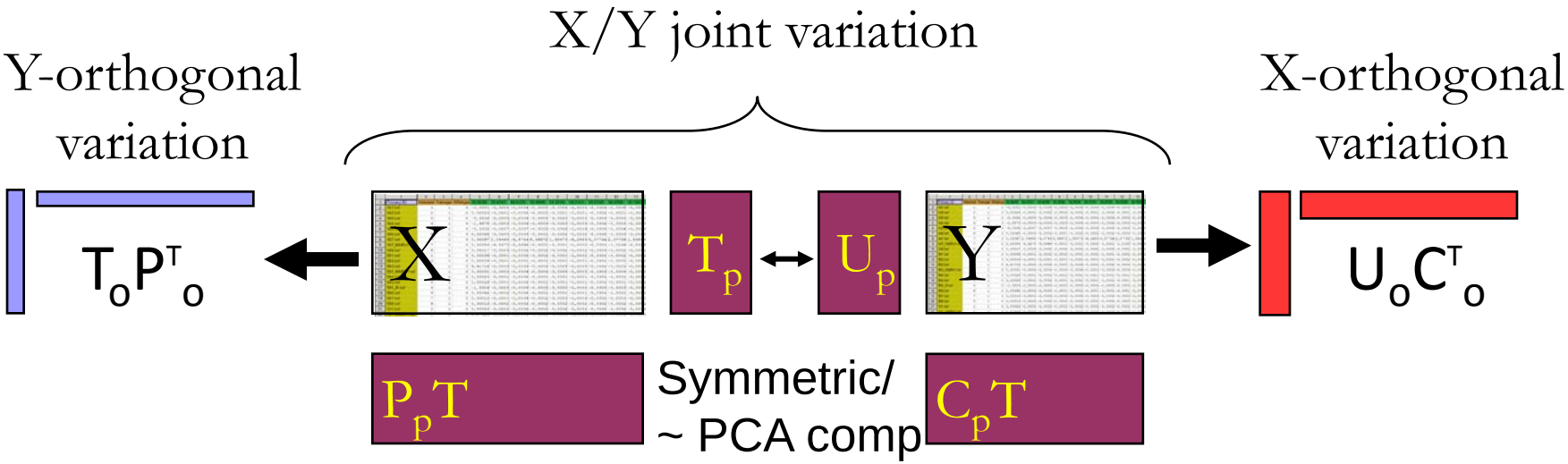
Metabolomics



Y

Two block modeling

The O2-PLS model

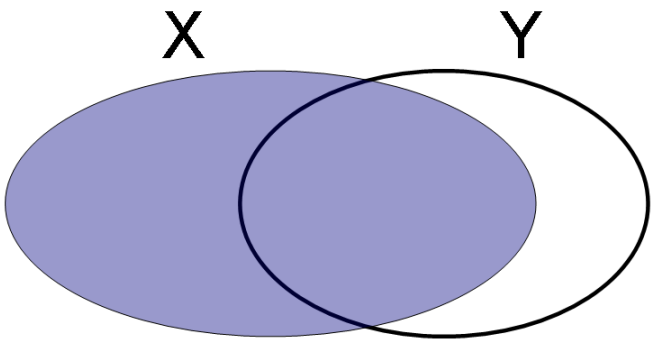


Trygg J.; **O2-PLS for qualitative and quantitative analysis in multivariate calibration**, Journal of Chemometrics, **2002**, 16, 283-293.

Trygg, J.; Wold, S.; **O2-PLS, a two-block (X-Y) latent variable regression (LVR) method with an integral OSC filter**. *Journal of Chemometrics*, **2003**, 17, 53-64.

PLS modeling vs OPLS modeling

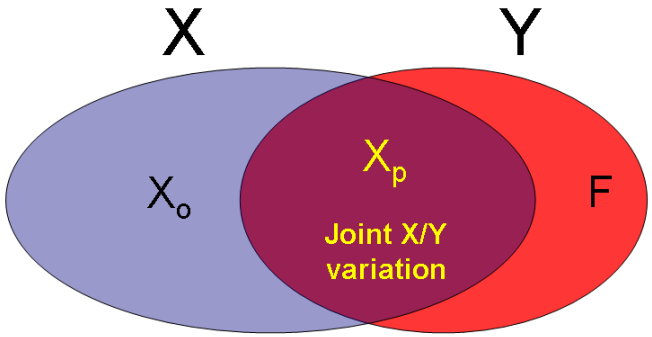
PLS, MLR, PCR, RR etc...



- Mixes Y-orthogonal and Y-predictive variation

- Uni-directional, Models Y FROM X

OPLS



- Separates Orthogonal and Predictive variation
(e.g. 'between block' from 'within block')

- Bi-directional, Models X AND Y

-*O*nly uses predictive variation for modeling Y

Benefits of OPLS modeling

- ✓ **Model diagnostics:**
  - $R^2(XY)$: How much variation in X is correlated to Y, and vice versa?
  - $R2(X_{yo})$: How much is not correlated to Y? (to X?)

- ✓ **Model interpretation**
  - More focussed components (plots) & easier interpretation
    - Predictive components ($\mathbf{T}_p\mathbf{P}_p^T$)
    - Y-orthogonal components ($\mathbf{T}_o\mathbf{P}_o^T$)
  - Pure profile estimation

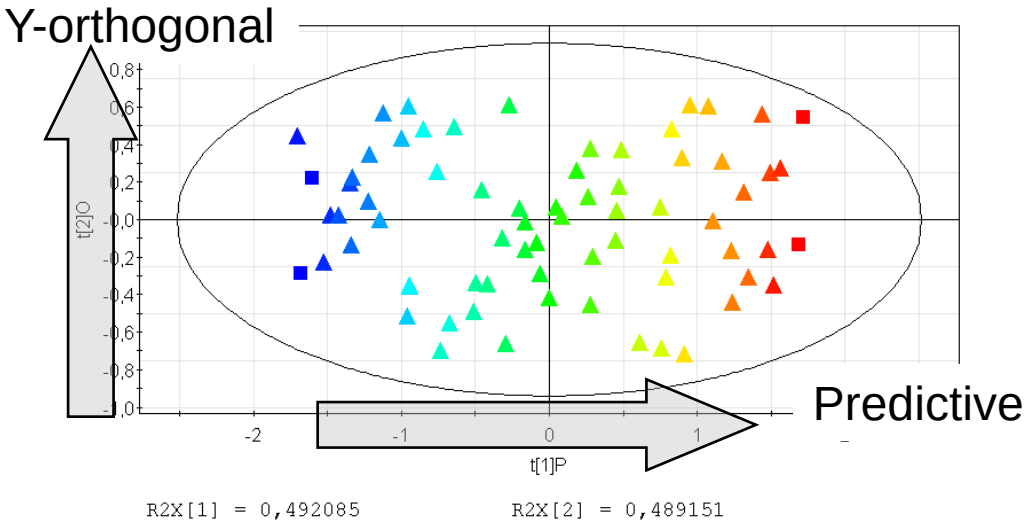
- ✓ **Model ($_{\text{prediction}}$):**
  - Understand & correct for faults/mistakes found in Y-orthogonal components
  - e.g. experimental, sampling

- **Multi-block modeling** (X←→Y)
  - Integrate, compare and filter multiple data tables

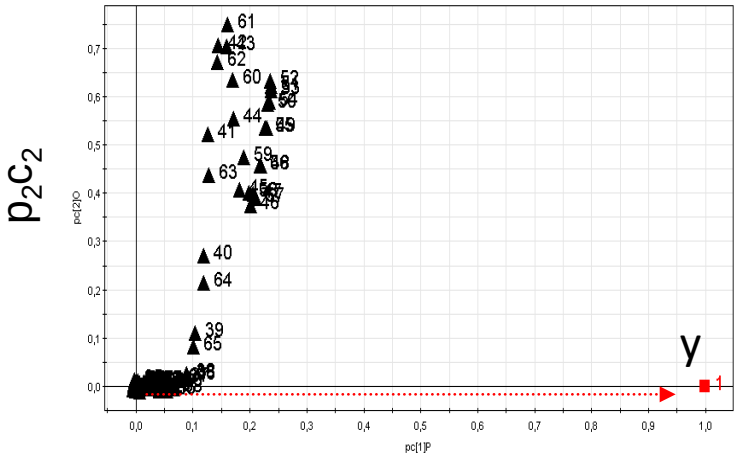
OPLS model

Example: Single-Y, two component system

Scores plot



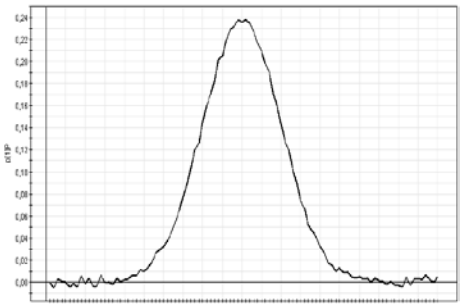
Loading plot



p_1c_1

49% variation

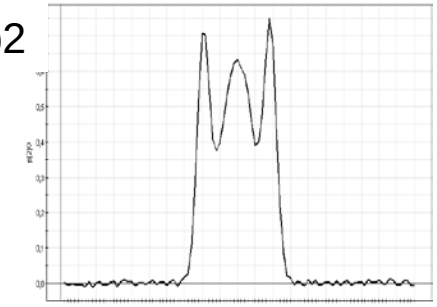
p1



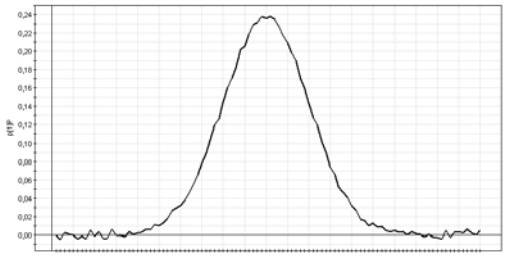
Predictive profile

49% variation

p2



Y-orthogonal profile



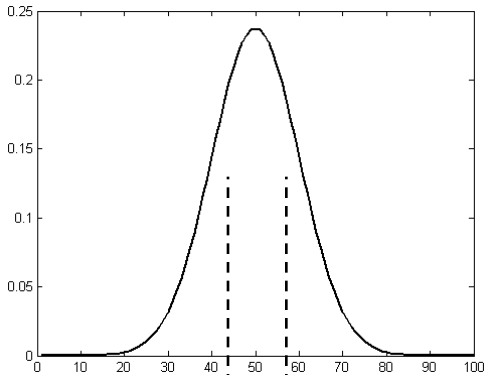
Predictive profile

OPLS model

Example: Two component system,

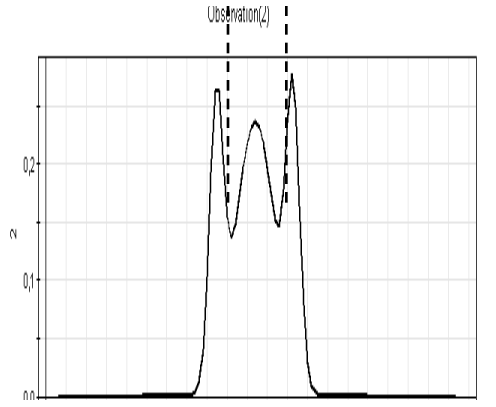
where unknown variation is correlated to known y

y1



Spectral profile of
Y-predictive component

y2

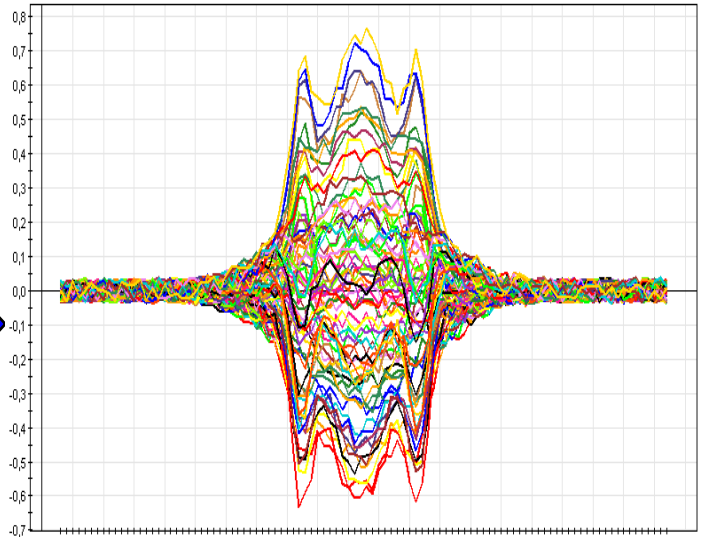


Spectral profile of
Y-orthogonal component

y1 and y2
have 0.7 correlation

X matrix

100



70

PLS

OPLS

y1

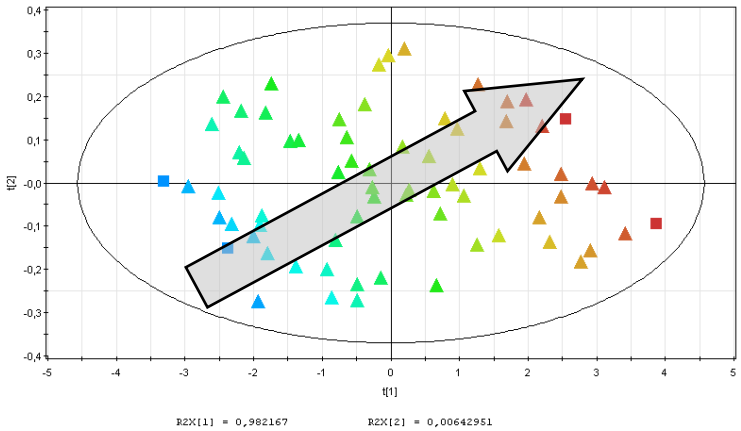
Single y

PLS x O-PLS

Example: Two component system,

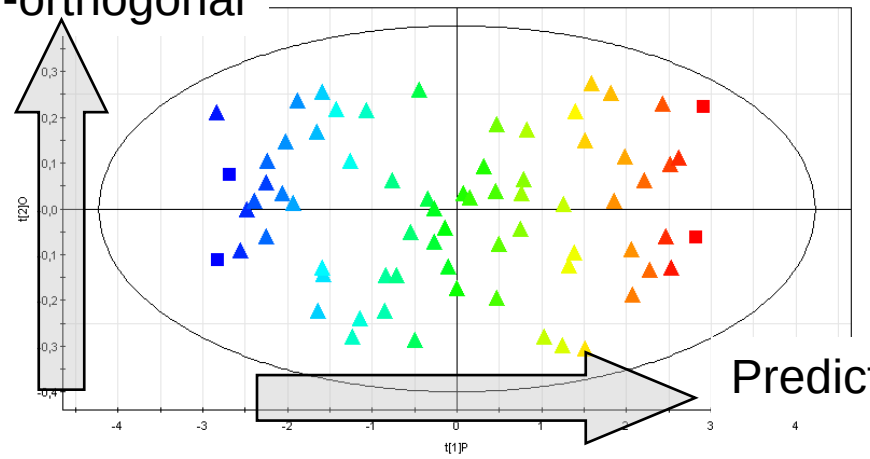
*where unknown variation is **strongly** correlated to known y*

PLS



OPLS

Y-orthogonal

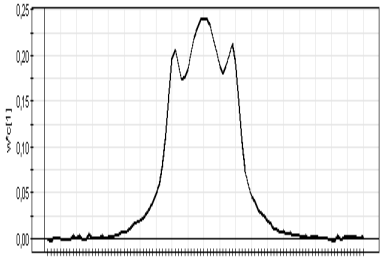


Predictive

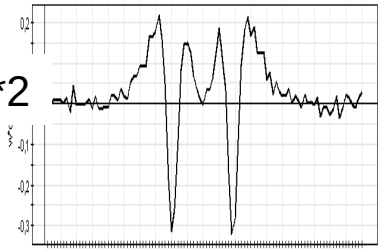
98 % variation

1 % variation

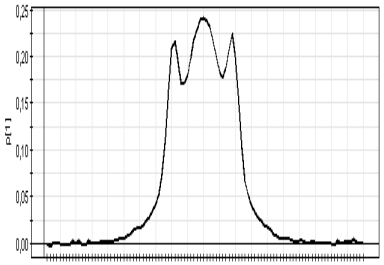
w*1



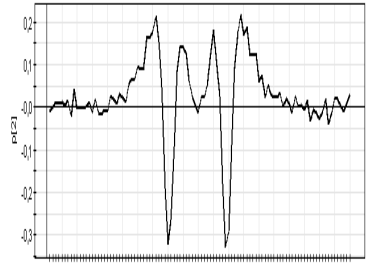
w*2



p1



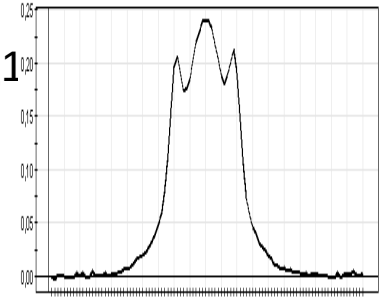
p2



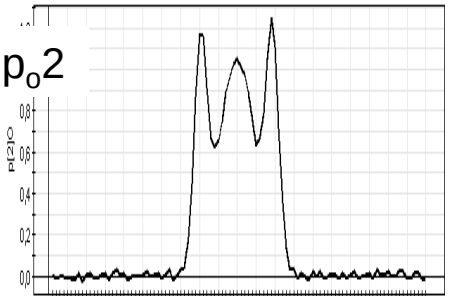
84 % variation

15 % variation

p_p1



p_o2



Predictive profile

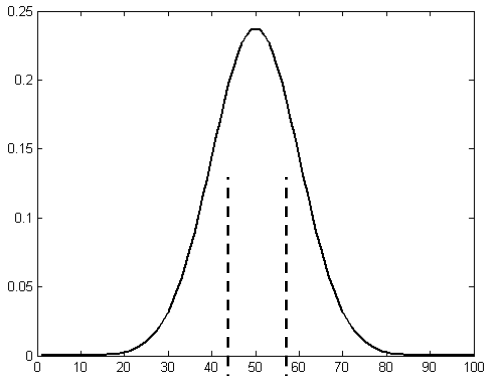
Y-orthogonal profile

Difficult to relate PLS loadings to the variation it represents

OPLS

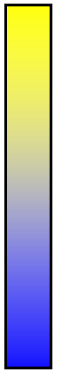
Example: Two component system,

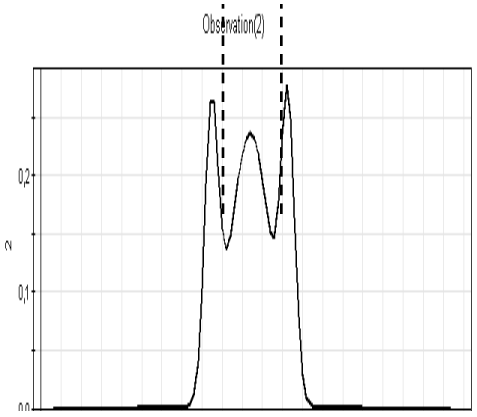
where unknown variation is correlated to known y



Spectral profile of
Y-predictive component

y1



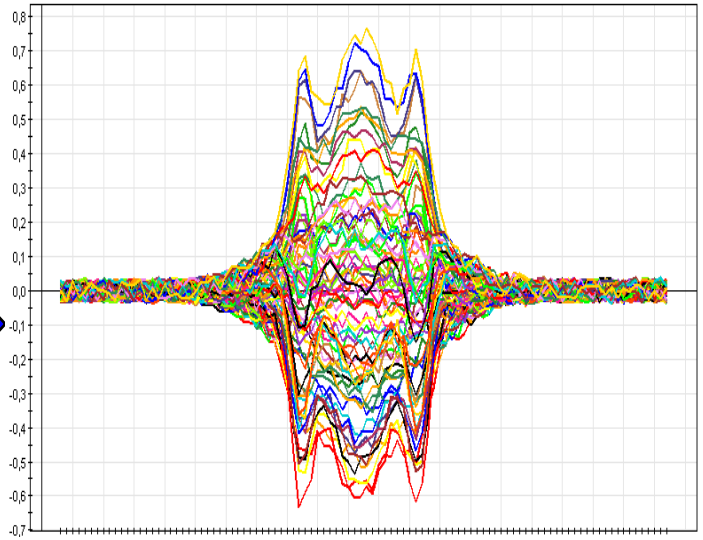


Spectral profile of
Y-orthogonal component

y2

y1 and y2 have
0.7 correlation

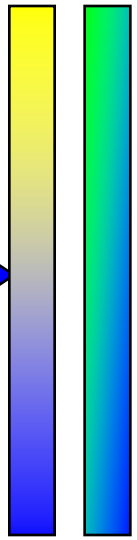
X matrix



PLS

OPLS

y1 y2



Multiple y

Use 2 separate OPLS models?

OR

Use 1 OPLS model with multi-y?

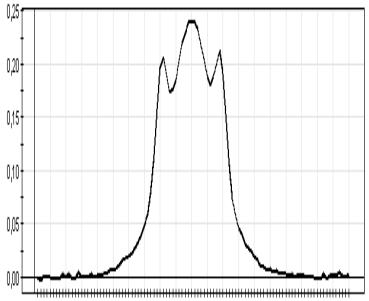
Single-Y vs multi-Y OPLS models

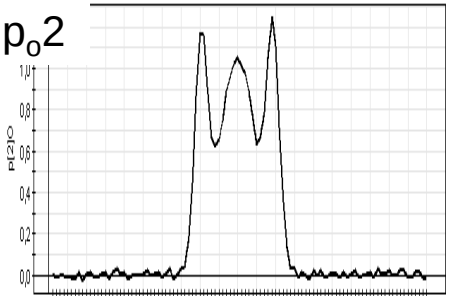
Two single-Y OPLS models

84 % variation

15 % variation

y1





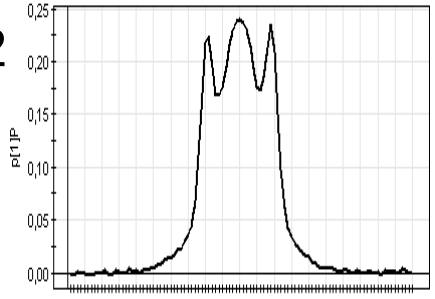
Predictive profile

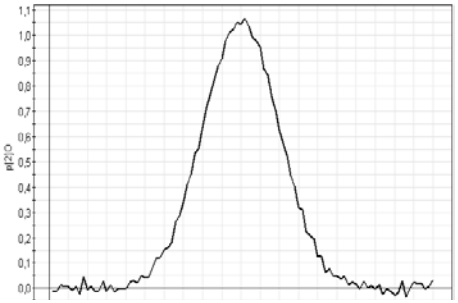
Y-orthogonal profile

84 % variation

15 % variation

y2





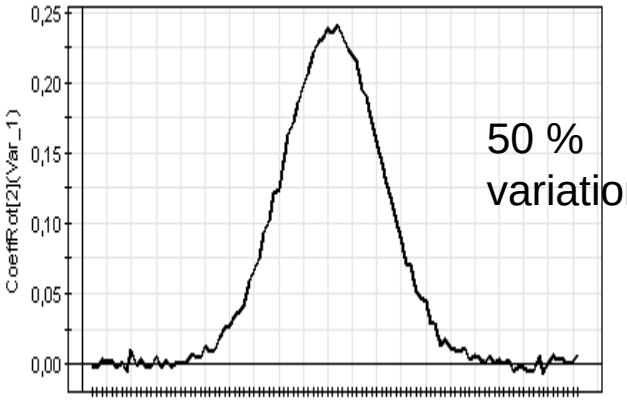
Predictive profile

Y-orthogonal profile

Multi-Y OPLS regression

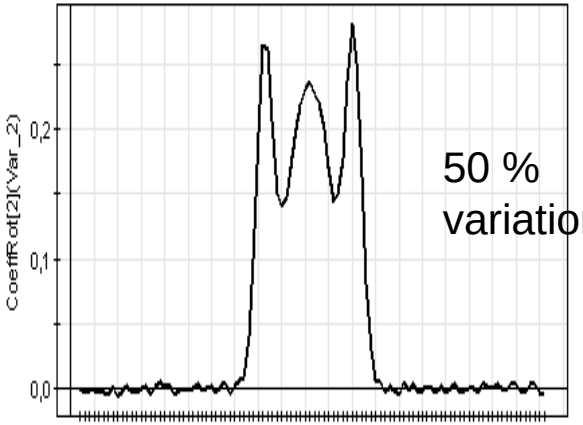
$K=B_p(B_pTB_p)-1$

y1



Predictive profile

y2



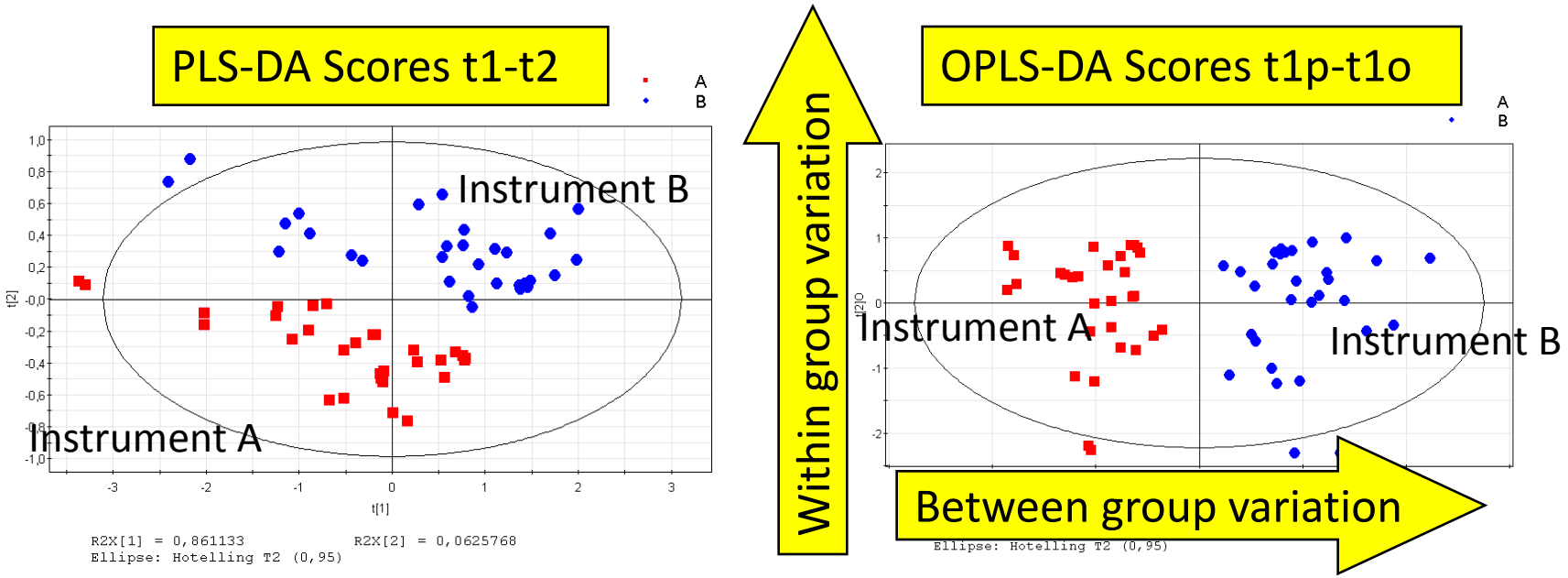
Predictive profile

Trygg J.; **Prediction and spectral profile estimation in multivariate calibration**, Journal of Chemometrics, **2004**, 18, 166-172.

OPLS as a filter

Example: Calibration transfer of near infrared spectra

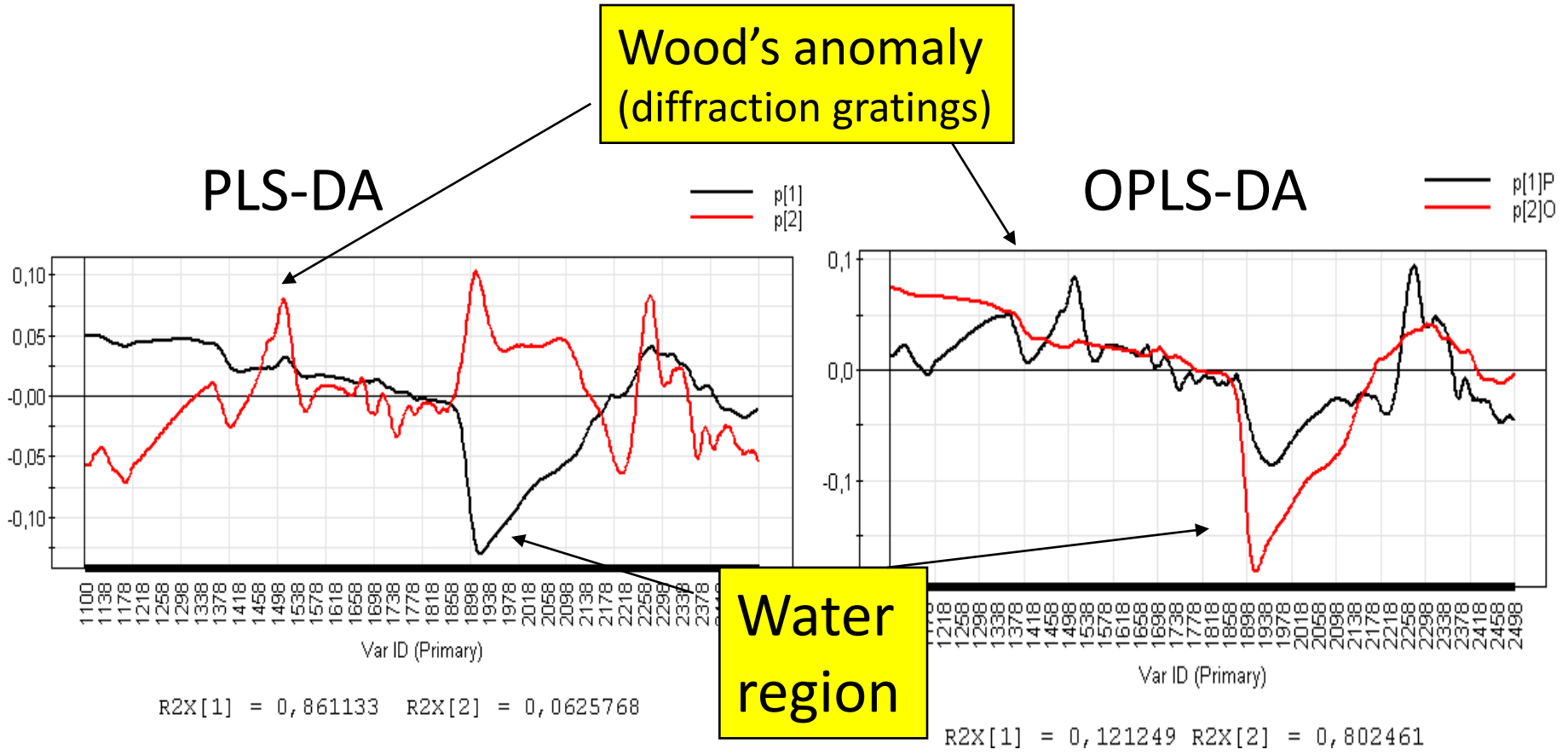
- Instrument A, B used to measure NIR spectra of an active pharma compound
- 15 batches specially selected to cover a variation of the water content
- A reference spectrum measured every second
- Water content varied from 1.38 to 4.47 wt./wt.% (Karl–Fischer titration)
- **Y =class (-1,1) [Instrument A vs Instrument B]**



Sjöblom, J.; Svensson, O.; Josefson, M.; Kullberg, H.; Wold, S.**; An evaluation of orthogonal signal correction applied to calibration transfer of near infrared spectra,** Chemometrics and Intelligent Laboratory Systems, **1998**, 44, 229–244.

OPLS as a filter

Example: Calibration transfer of near infrared spectra



Example PAT: Binary powder

- Diffuse reflectance NIR spectroscopy
- Mixture of two powders with markedly different particle size
- 11 batches of powders, 0% to 100% in steps of 10%.

- X = NIR spectra (SNV) in the range 1080-2025 nm
- Y = % binary mix of powders

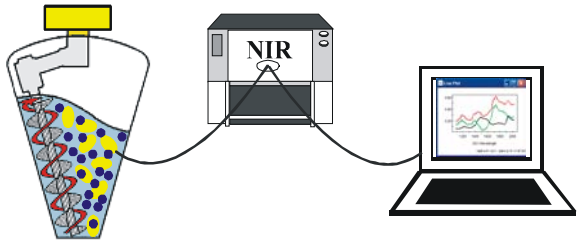
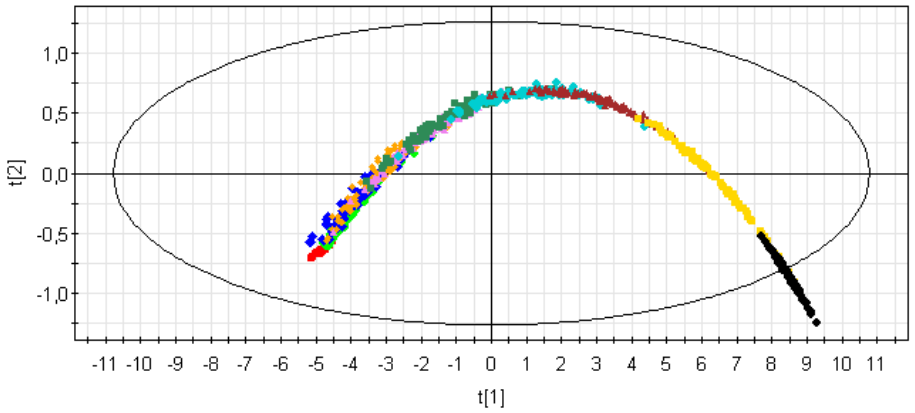


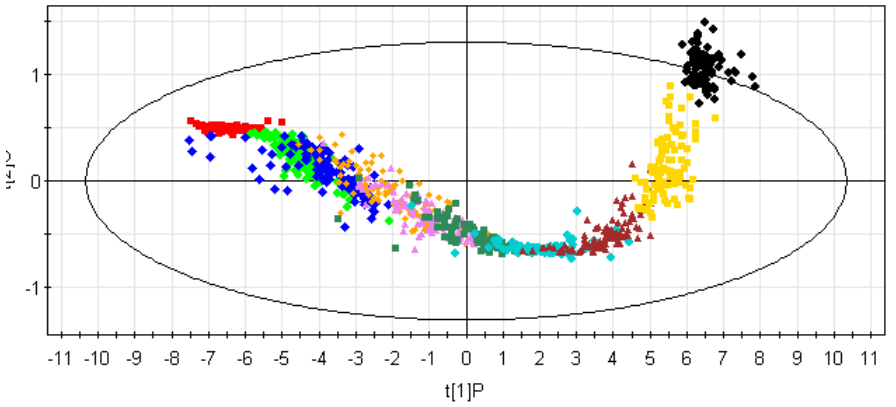
Figure: Schematic overview of the vertical cone mixer and the fibre-optic probe set-up.

PLS model scores



R2X[1] = 0,983807 R2X[2] = 0,0137599
Ellipse: Hotelling T2 (0,95)

OPLS model scores

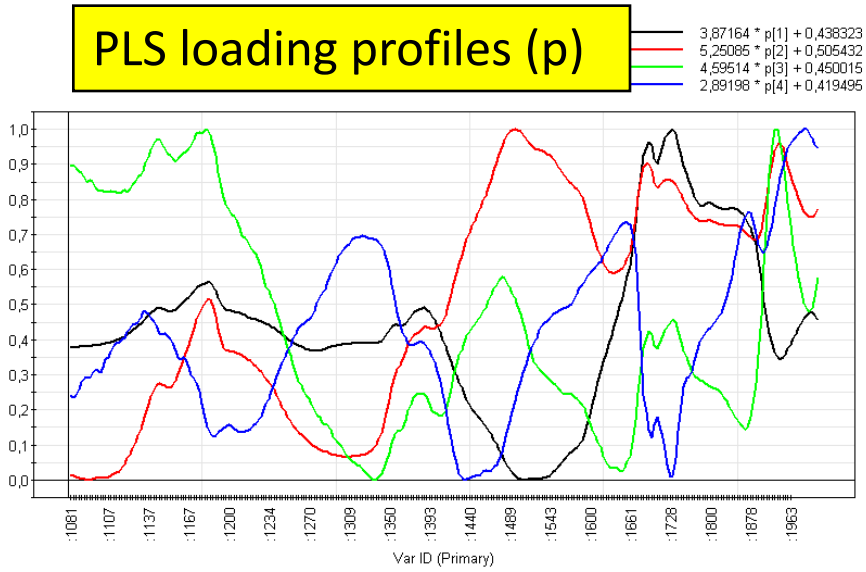


R2X[1] = 0,901427 R2X[2] = 0,0682814
Ellipse: Hotelling T2 (0,95)

Example PAT: Binary powder

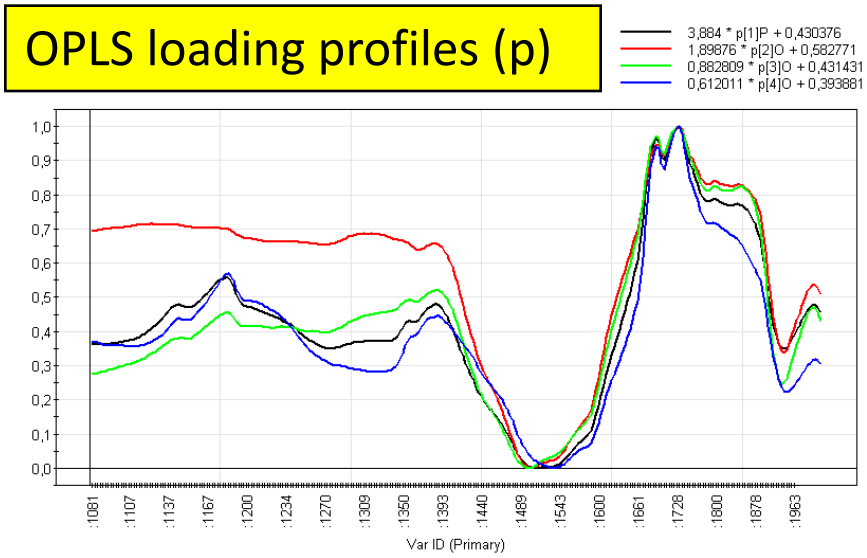
Non-linearities transparent in OPLS loading profiles

PLS loading profiles (p)



2X[1] = 0,983807 R2X[2] = 0,0137599 R2X[3] = 0,00107454 R2X[4] = 0,0005891

OPLS loading profiles (p)



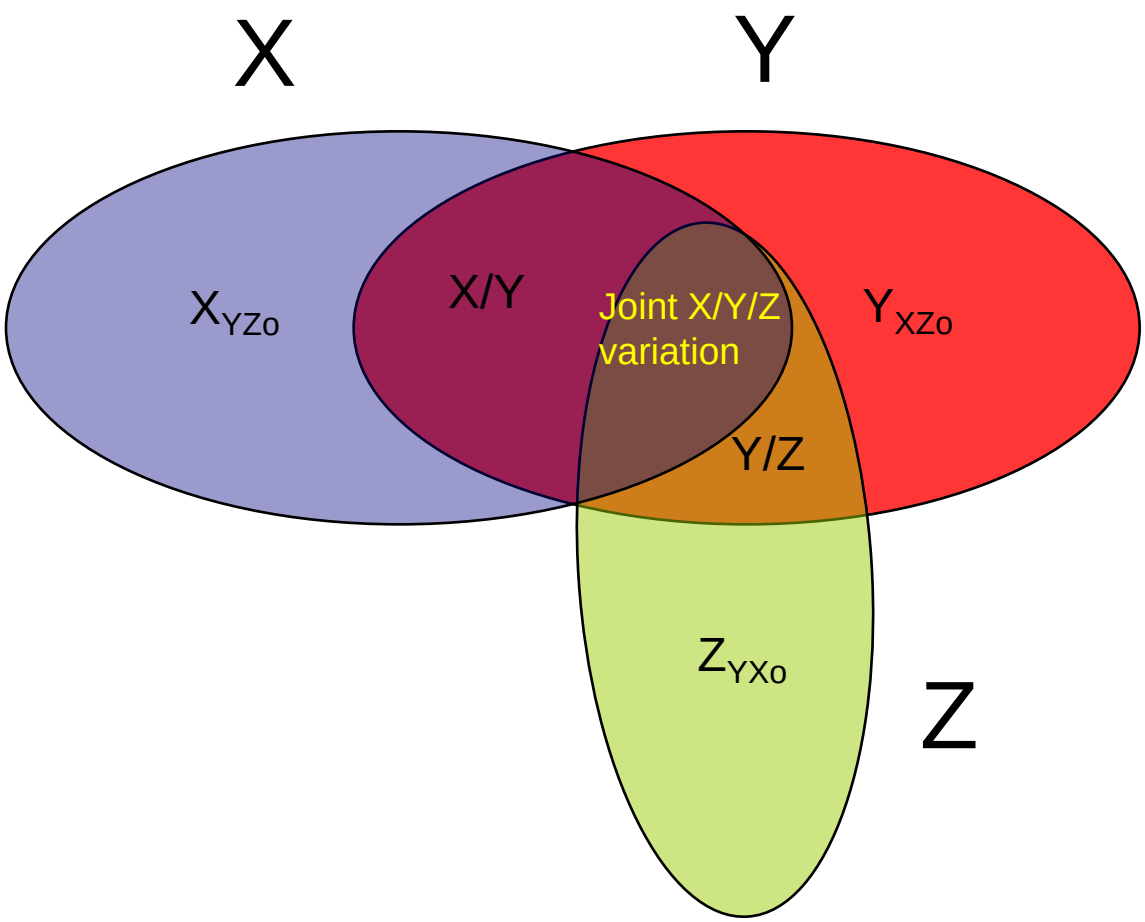
R2X[1] = 0,901427 R2X[2] = 0,0682814 R2X[3] = 0,0185834 R2X[4] = 0,0109389

OPLS-derived methods

- Bifocal OPLS (BIF-OPLS)
- Kernel OPLS
- Multi-block modeling OPLS

.....

Three blocks of data (X/Y/Z)
BIF-OPLS
(near publication)

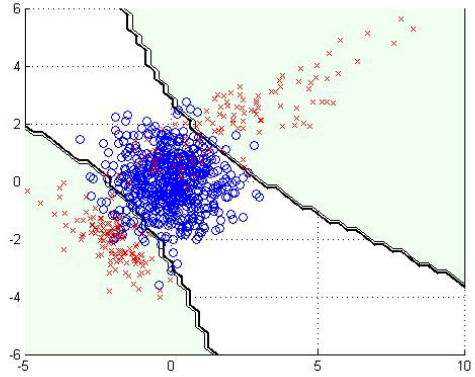


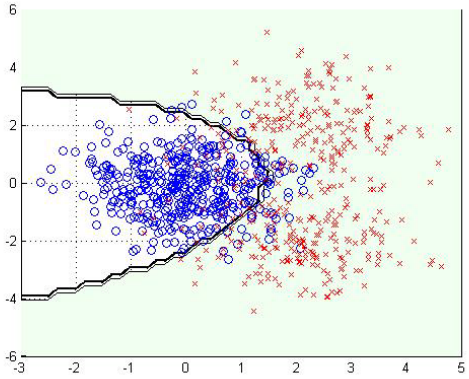
Non-linear modeling techniques
Kernel-OPLS

- There are situations where linear modeling techniques are insufficient
  - Biological and chemical systems, image analysis, *etc*.

- Many alternatives exist for prediction and classification
  - Artificial neural networks (ANNs)
  - Bayesian networks
  - Support Vector Machines (SVMs)
  - Kernel-based Partial Least Squares (KPLS)

- **K-OPLS**
  - Benefits are related to the interpretation of **Y**-predictive and **Y**-orthogonal scores
  - Not possible with KPLS or SVMs



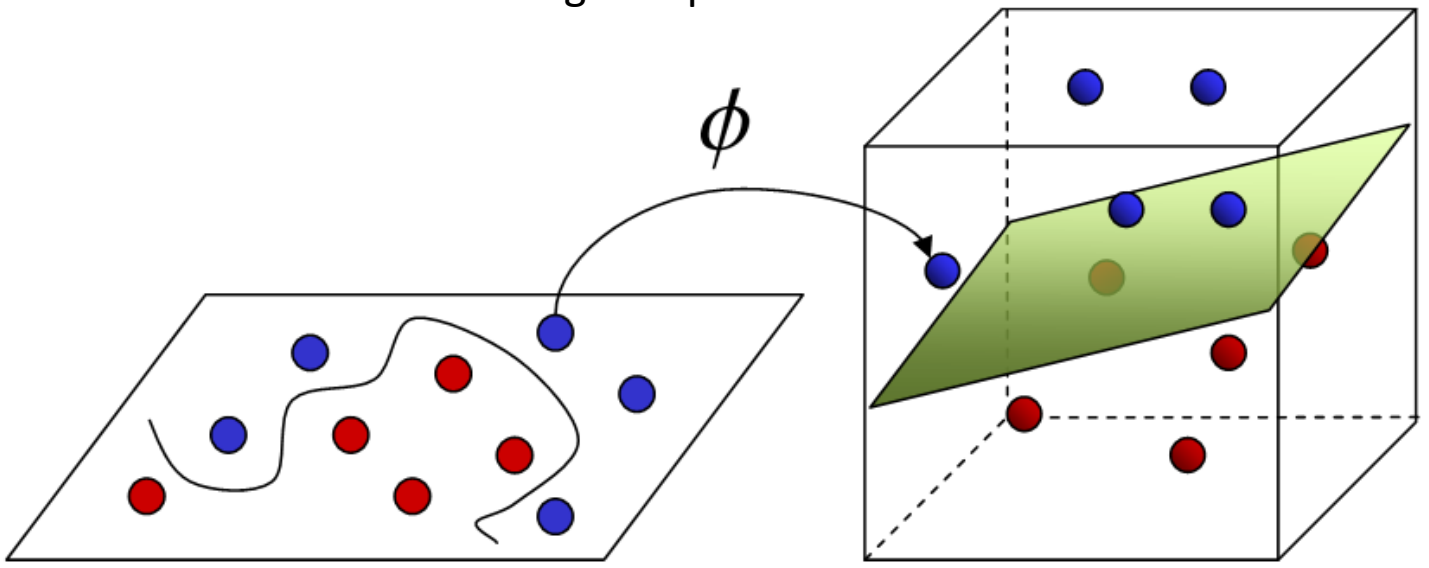


Rantalainen, M.; Bylesjo, M.; Cloarec, O.; Nicholson, J.K.; Holmes, E.; Trygg, J**.; Kernel-based orthogonal projections to latent structures (K-OPLS)**, Journal of chemometrics, **2007**, 21, 376-385.

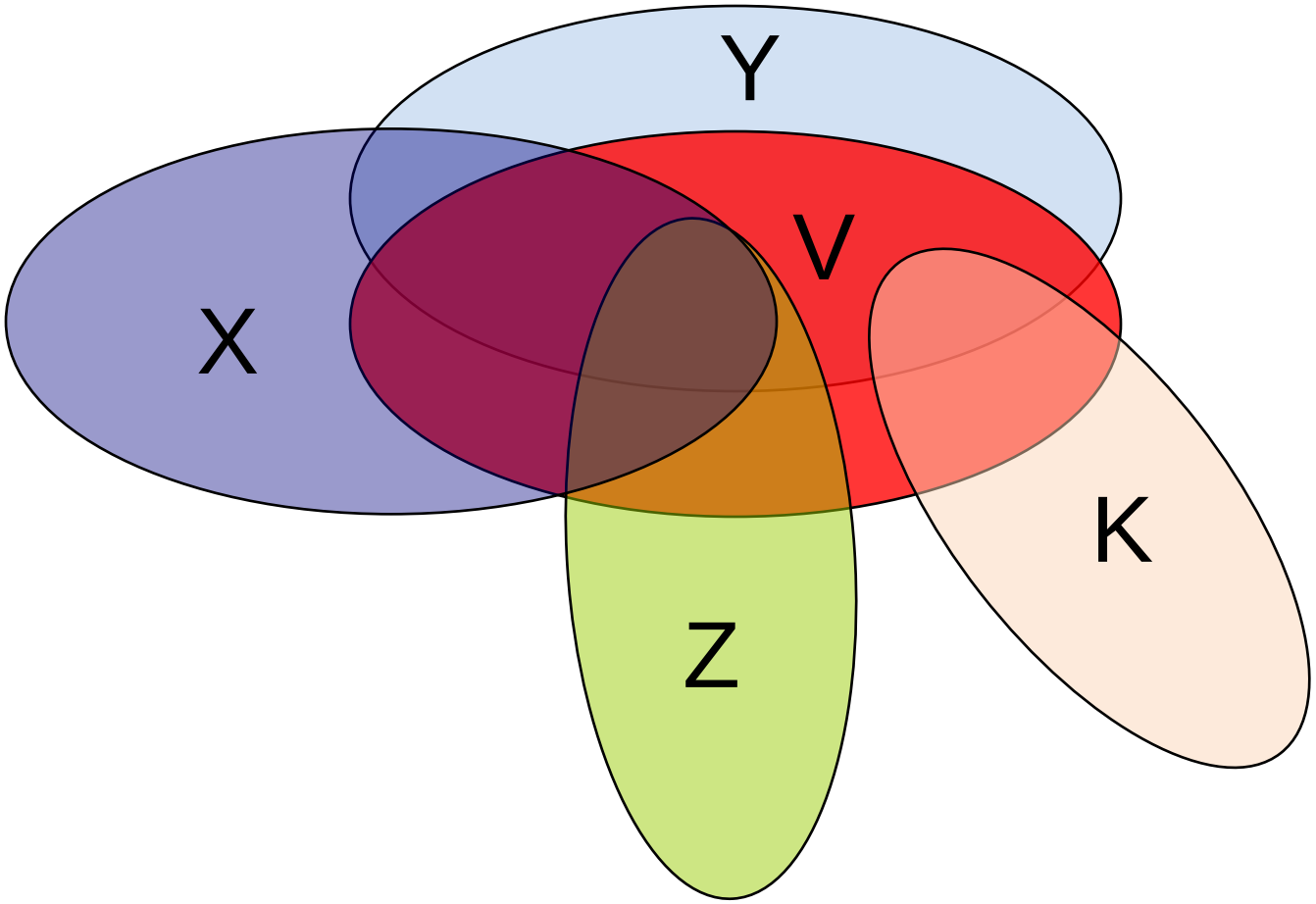
Kernel-based methods

Image from http://www-kairo.csce.kyushu-u.ac.jp/~norikazu/research.en.html

- Kernel-based methods utilize Φ(**X**) instead of **X** to predict **Y**
- The function Φ(·) extends **X** into a high-dimensional space (*feature space*)
- In this higher-dimensional space, a linear model is used for regression or classification
- The model is non-linear in the original space



Multi-block modeling OPLS
(in development)

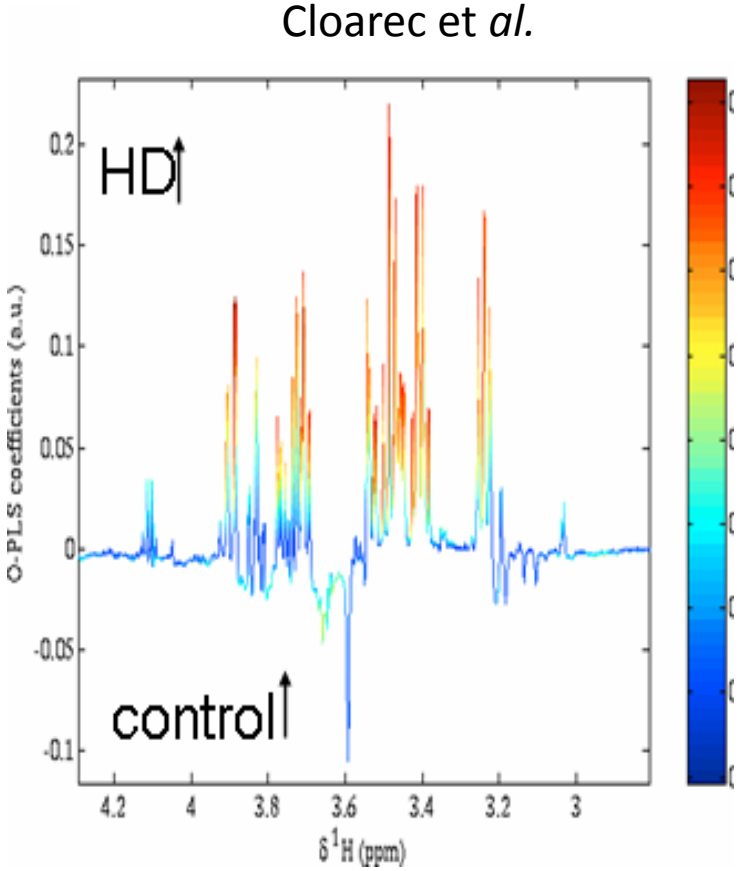


Visualisation of OPLS model

STOCSY & S-plot: correlation and covariation combined into one plot

STOCSY (NMR)

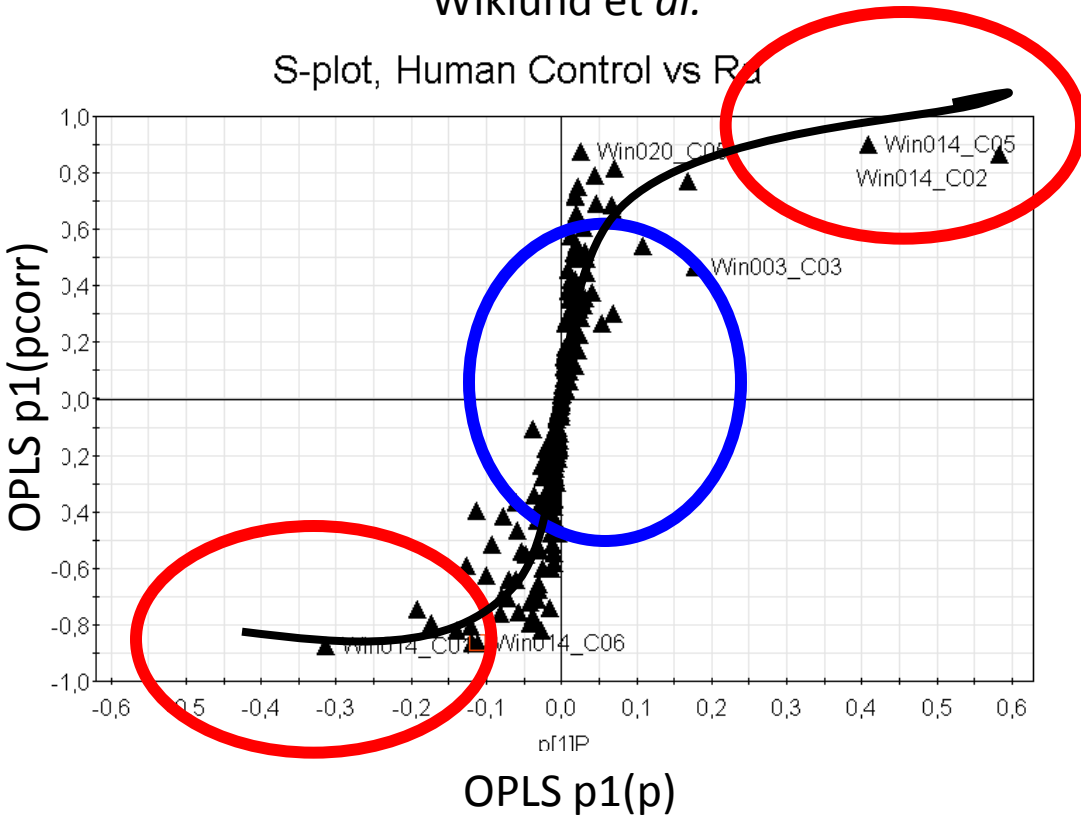
Cloarec et *al.*



Line plot

S-plot (NMR, MS,etc...)

Wiklund et *al.*



Scatter plot

Covariation and correlation

- **Covariation** is the measure of how much two variables vary together (strength)
  - Covariation is scale dependant (i.e. dependant upon the size of variability of the two variables)
  - Can hold positive, 0, and negative values

$$\mathrm{Cov}\,(\mathbf{t}, \mathbf{y}) = [(\mathbf{t})^{\mathrm{T}}(\mathbf{y})] / (\mathrm{N}\text{-}1)$$

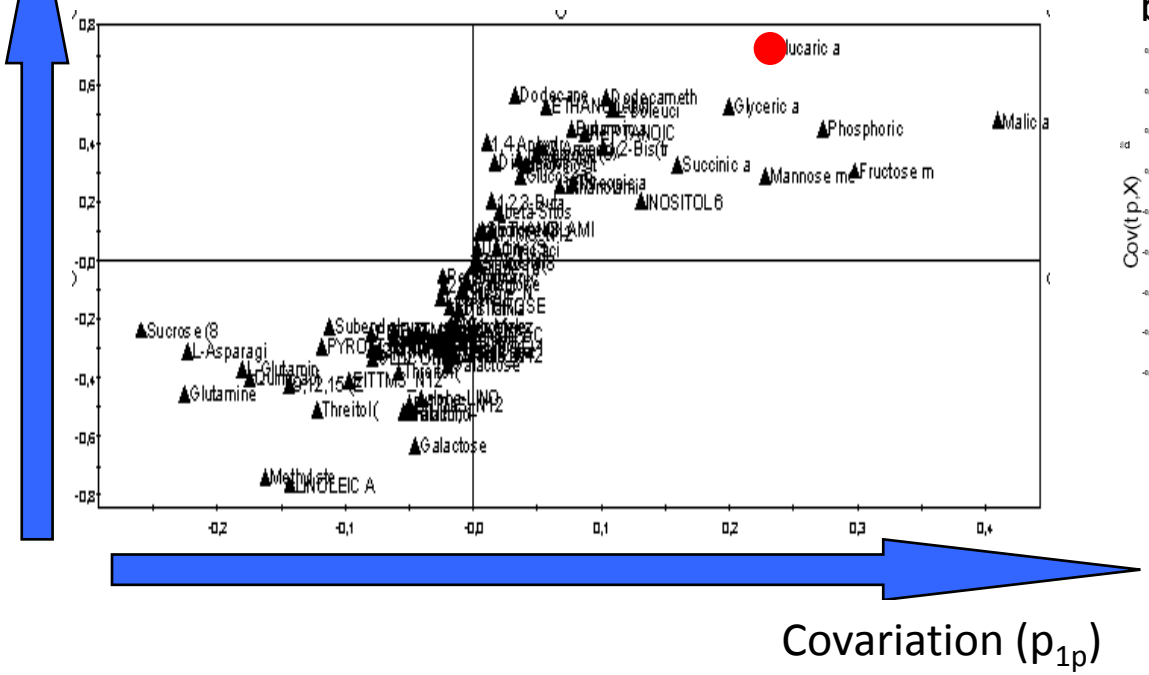
- **Correlation** = Fit is a dimensionless measure of covariation
  - Correlation is scale invariant (i.e. not dependant upon the size of variability of the two variables)
  - Can hold values between -1 to +1

$$\mathrm{Corr}\,(\mathbf{t}, \mathbf{y}) = [\mathrm{Cov}\,(\mathbf{t}, \mathbf{y}) / (\|\mathbf{t}\|\,\|\mathbf{y}\|)]\,(\mathrm{N}\text{-}1)$$

Understand the most influential metabolites
related to class separation
→ S-plot of the OPLS predictive component

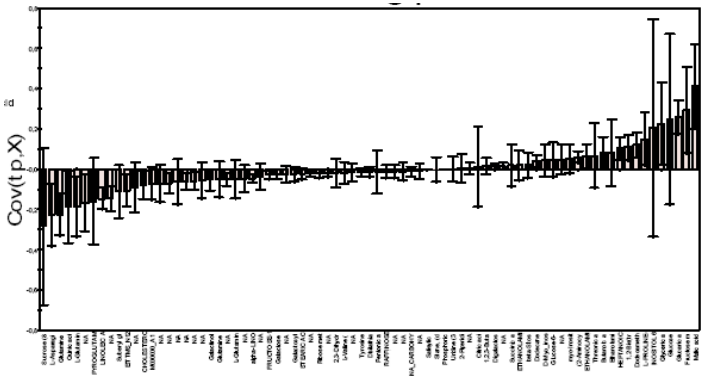
Correlation
($p_{1p\ cprr}$)

S-plot



Covariation (p_{1p})

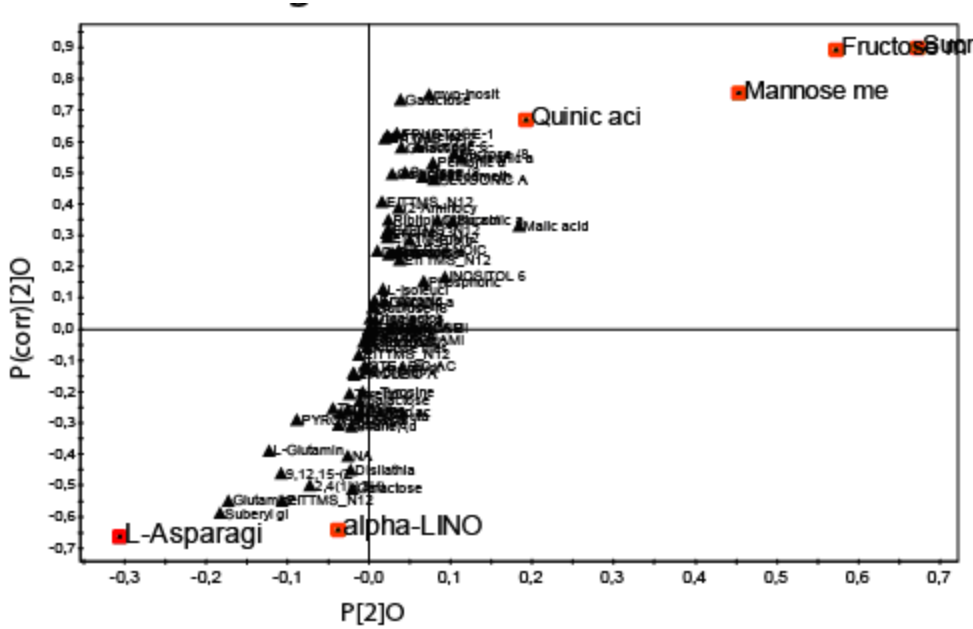
Confidence interval of variables
based on Jack-knifing estimation

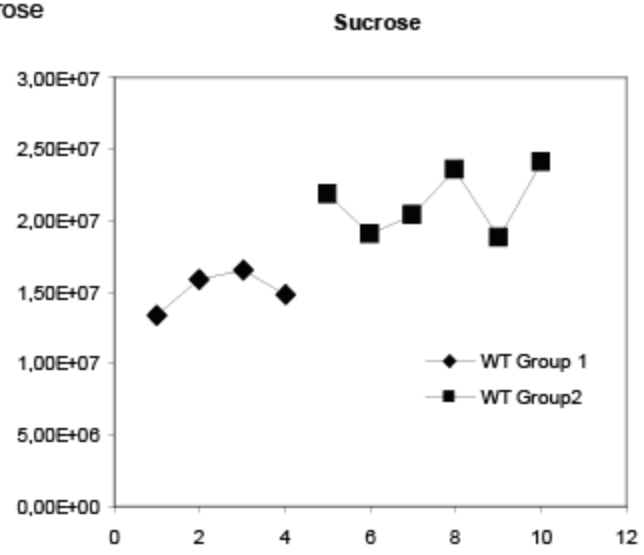


Understand the most influential metabolites (putative)

NOT CORRELATED to class separation

→ S-plot of the OPLS orthogonal component





Examples

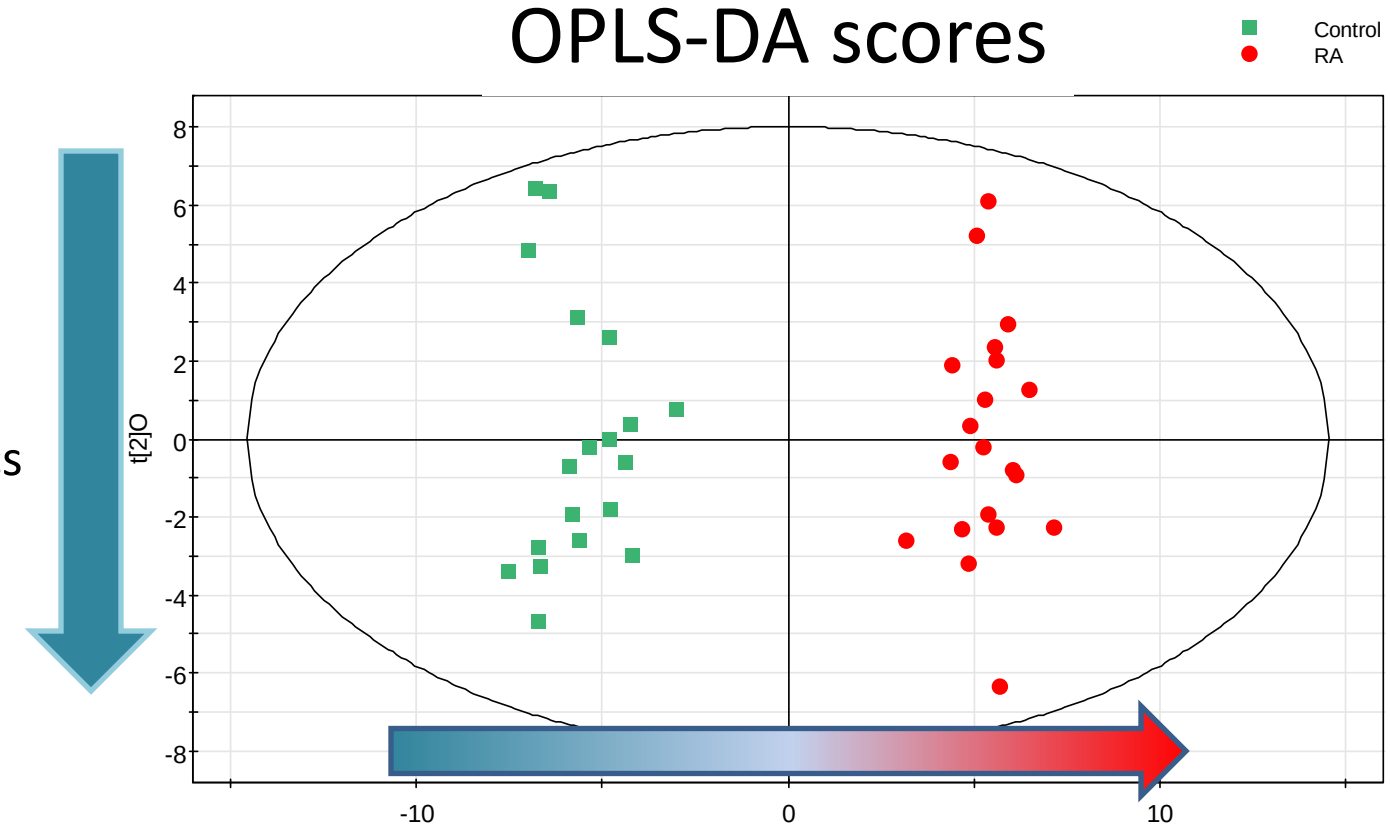
2-class separation OPLS

Disease diagnosis:
Rheumatoid Arthritis – brief background

- Worldwide prevalence of approximately 1%
- Autoimmune disease, the body attacks itself, aetiology largely unknown
- Treatment; irreversible disease, no known cure, medication to maintain mobility and ease pain
- Early diagnosis critical
  - More successful treatment with early medication
- Diagnosis for rheumatoid arthritis
  - Physical examination, antibodies (today not specific for RA), X-ray, MRI
- New diagnostic tools are needed…

Two class separation - Rheumatoid arthritis

Blood serum samples from 40 individuals (20 RA/20 Control)



Within group dynamics

Group separating direction

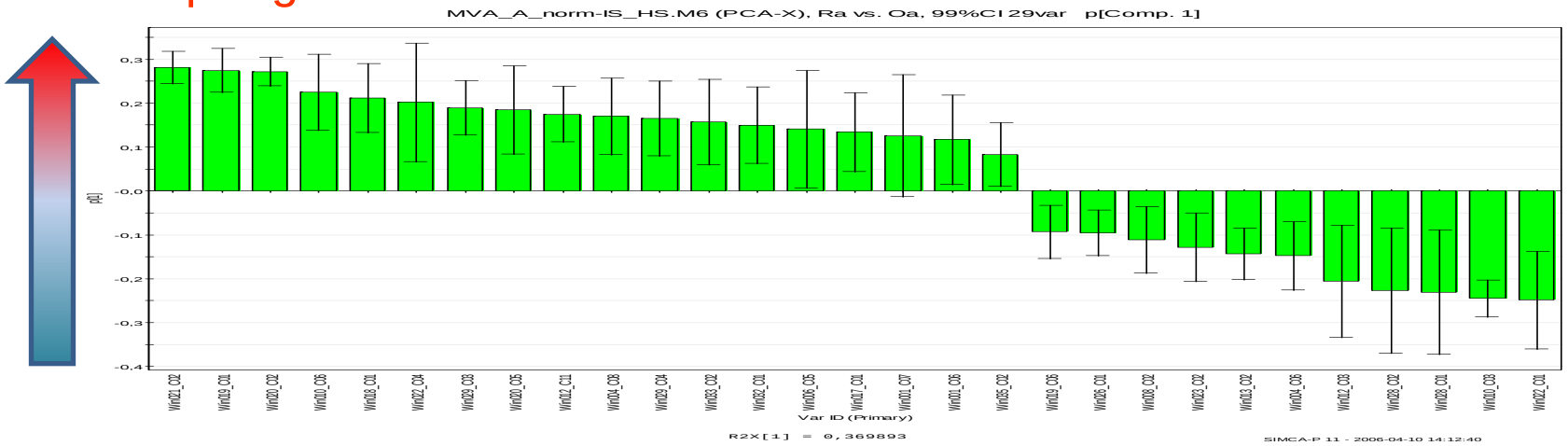
Specific metabolites for healthy and diseased

Rheumatoid arthritis: Control vs. RA

Understand biochemical differences

- Significant (subset) metabolites for separation of RA samples from healthy controls.
  - Variables represent endogenous metabolites

Up regulated in Ra



Down regulated in RA

RA: Comparison of the human case and animal models

- Great overlap of metabolites between humans and animals
  - Different metabolites show overlap in different animal models
  - Allows for identification of relevant animal models
  - Selection of model system for treatment studies

BM	Human Rheumatoid Arthritis	Mouse Collagen Induced Arthritis	Rat Adjuvant Induced Arthritis
EC001	↑	na	Na
EC002	↑	?	?
EC003	↑	↓	↓
EC004	↑	0/↓	↓
EC005	↓	na	na
EC006	↓	↓	↓
EC007	↓	↓	↓
EC008	↓	↓	↑
EC009	↓	↓	↓
EC010	↓	↑	↑
EC011	↓	0/↓	↓
EC012	↓	na	na
EC013	↓	↓	↓
EC014	↓	↓	?
EC015	↓	↓	↓
EC016	↓	?	↓
EC017	0	↓	↓
EC018	↑	↑	↓/?
EC019	↓	↓	↓
EC020	↓	↓/?	↓
EC021	?	↑/?	↑
EC022	↓	↓	↓
EC023	0	↓	↓
EC024	↑	↓	0/↑
EC025	↑	↓	↓
EC026	0/↑	↓	↑

RA: Comparison of therapies in animal model

- Metabolites levels are affected by administered therapeutics
  - New drug (X) restore levels in more metabolites compared to MTX*
  - Useful in development of novel drugs
  - Tool in clinical studies to verify therapeutic effect in clinical studies
  - Concomitant development of novel drug and diagnostic test, theranostics?

	Vehicle	MTX	X 1mg	X 3mg	X 10mg
EC004	0/↑	↓	↓	0/↓	↓
EC006	0/↑/?	0/?	0	↑	↑
EC007	↓	0/↑	0/↑	0/↓	↑
EC009	0	↑	↓	↑	↑
EC010	↑	↑	↑	↑	↑
EC011	0	0/↓	↓	0/↓	↑
EC012	0/↓	↑	0/↓	↑	↑
EC013	↑	0/↑	0/↓	↑	0/↑
EC014	↑	0/?	↑	↑	↑
EC015	0/↑	↑	0/↓	↑	↑
EC016	0	↓	↑	↑	↓
EC017	↓	↓	↓	↓	↓
EC018	↓	↓	↓	0/↑	0/↑
EC019	↓	↓	↓	↓	↓
EC022	↑	↑	0/↑	↑	↑
EC023	↓	0/↓	↓	↓	↓
EC024	↓	↓	↓	↓	↓
EC025	↓	↓	↓	↓	↓
EC026	↑	↑	↑	↑	↑

*MTX, methotrexate

Multi-class separation OPLS

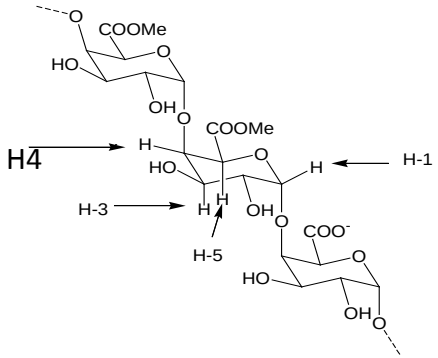
OPLS in multi class metabolomics
Example: Plant metabolomics on Poplar

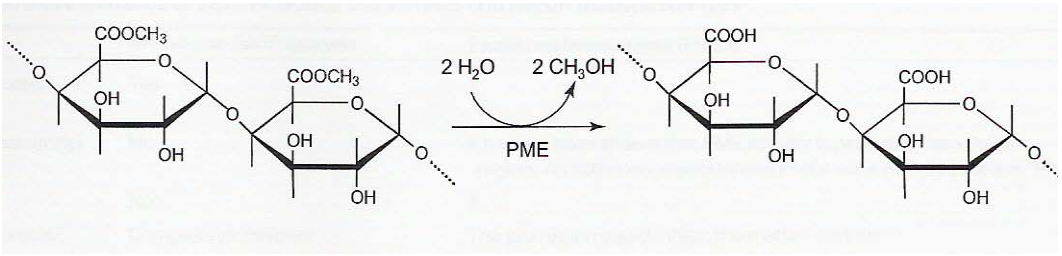
***PttPME1* expression was up and down regulated in transgenic aspen trees**
PME enzyme activity in wood forming tissues was correspondingly altered

<u>Lines in this study</u>
WT poplar
2B – up regulated *PttPME1* gene
5- down regulated *PttPME1* gene

Metabolomics study of xylem and phloem, here only the xylem results are presented.

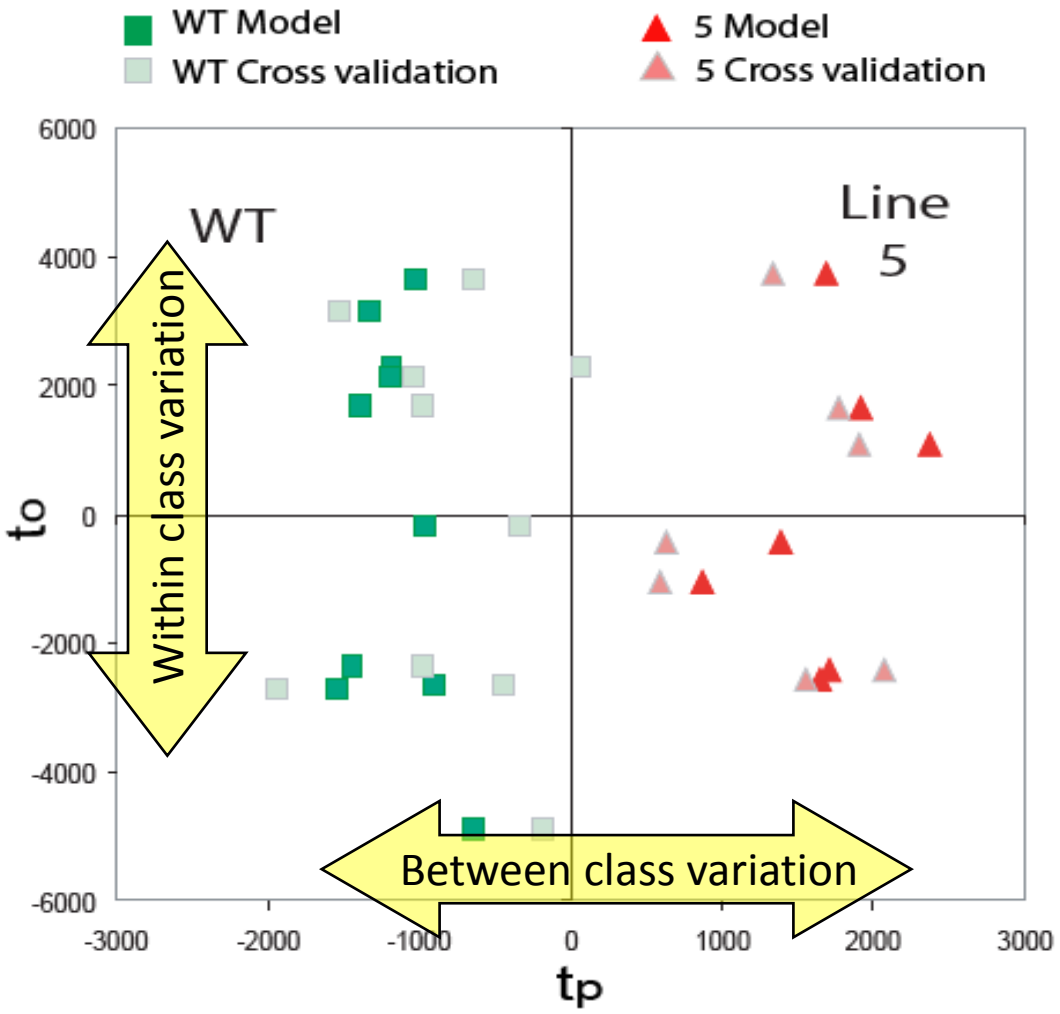
pectin





OPLS-DA model of Line 5 vs Wildtype Score plot

OPLS model
1 predictive component
3 orthogonal components
$R^2X(p)=12\%$
$R^2X(o)=20\%$
$Q^2Y=80\%$
$R^2Y=96\%$



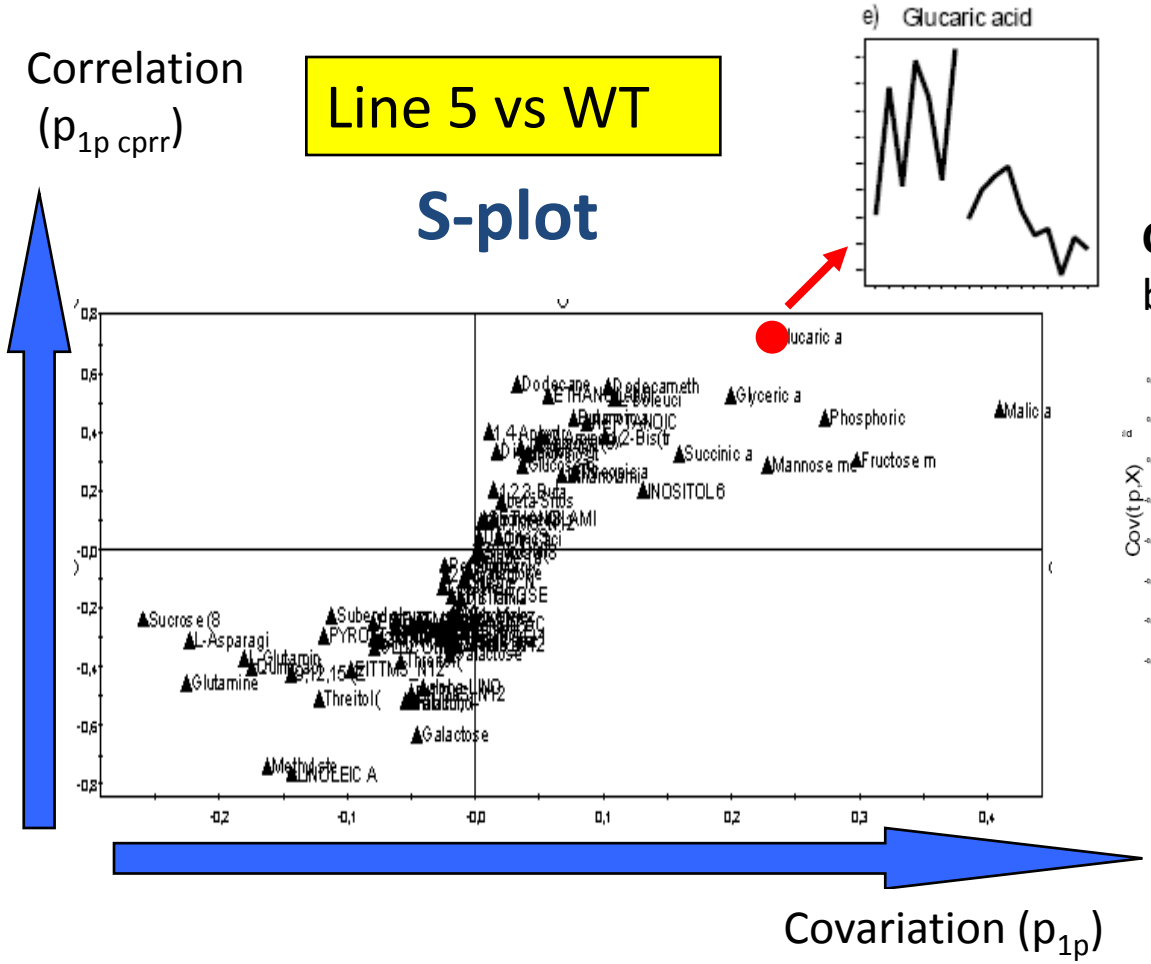
Understand the most influential metabolites (putative) related to class separation (transgene vs wildtype)

→ **S-plot** of the OPLS predictive component

Correlation ($p_{1p\ cprr}$)

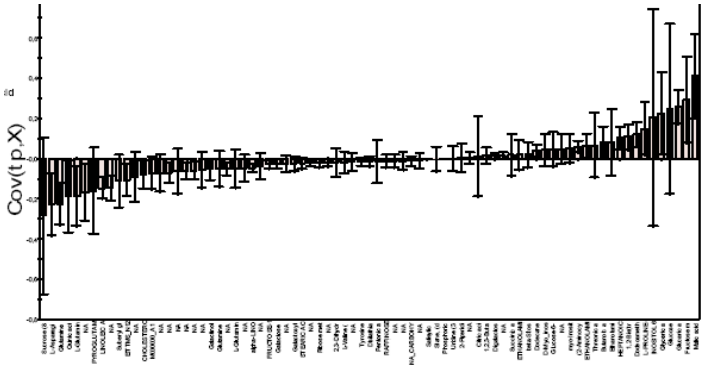
Line 5 vs WT

S-plot



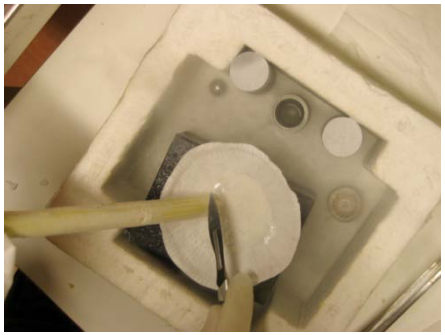
Covariation (p_{1p})

Confidence interval of variables
based on Jack-knifing estimation

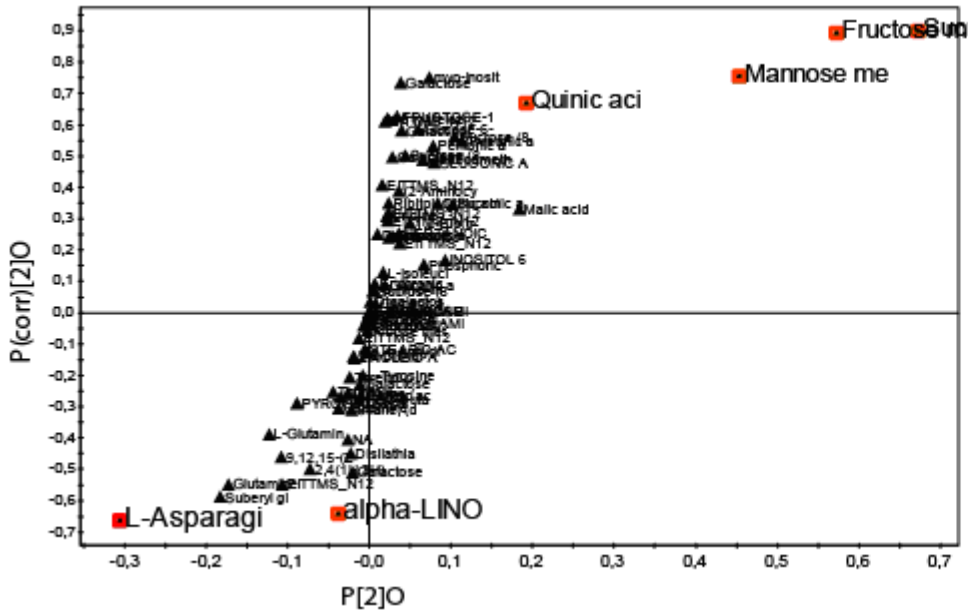


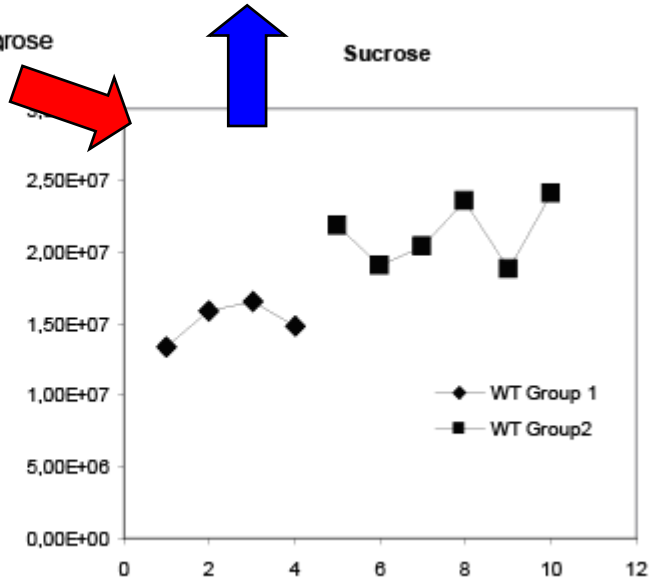
Understand the most influential metabolites (putative)
NOT CORRELATED to class separation

Orthogonal S-plot



Line 5 vs WT
Orthogonal S-plot





Multiblock modeling - O2PLS

Combined profiling projects at UPSC

Transcriptomics

sample preparation
RNA extraction

↓

labelling, hybridization

↓

scanning,
(Scanarray)

↓

data extraction
(GenePix)

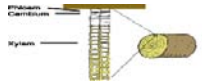
↓

normalization,
statistical treatment
(UPSC-BASE)

↓

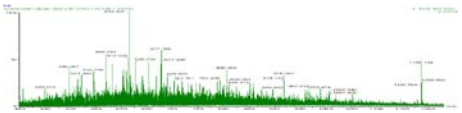
visualization and
interpretation of data
(MAPMAN, GeneSpring)

Proteomics



Preparation and fractionation
of samples

↓

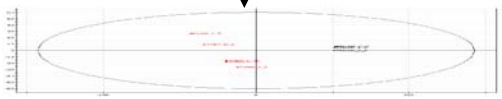


LC/MS analysis and data processing

↓

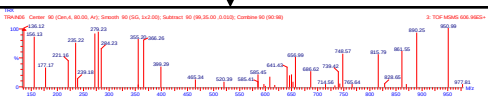
Combined Profiling

↓



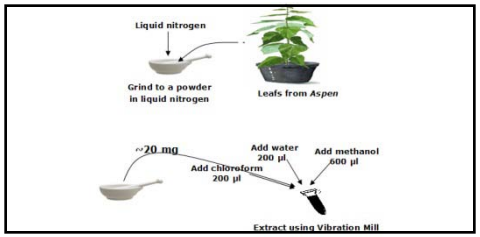
Selection of differential expressed peptides
derived from multivariate analysis

↓

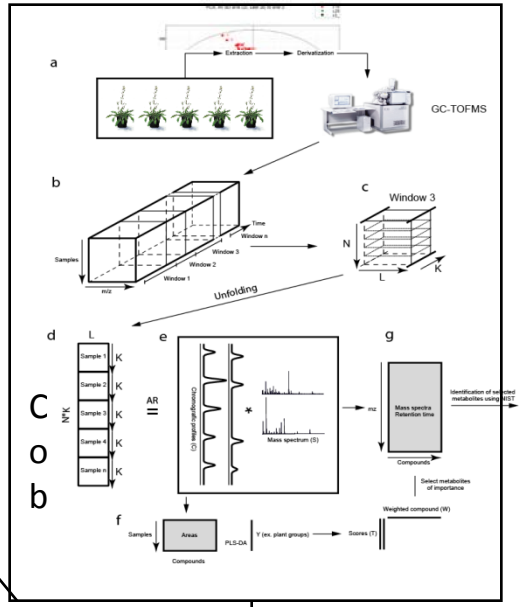


LC/MS/MS analysis and identification
of proteins by database search

Metabolomics



↓

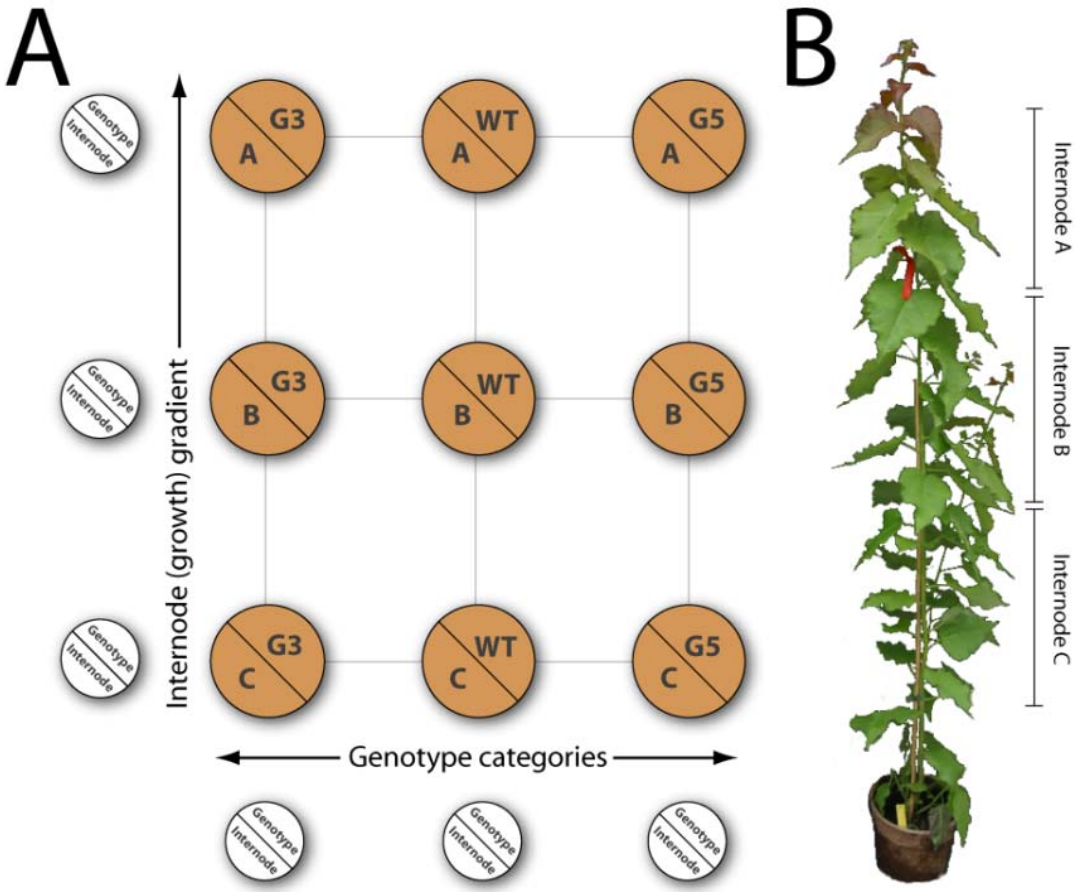


↓

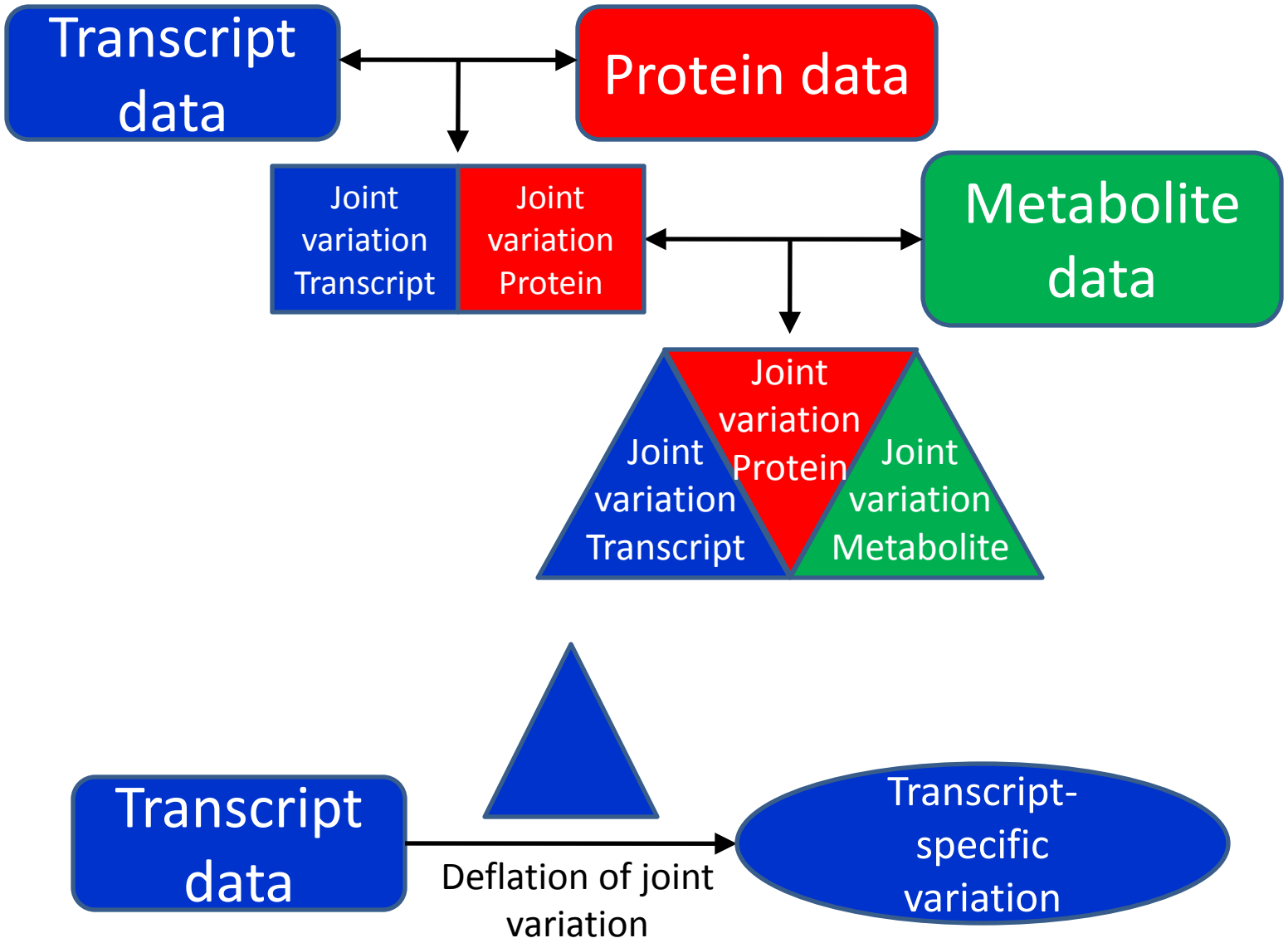
visualization and
interpretation of data

Combined profiling of transgenic Poplar

DOE (Genotype/Internode)

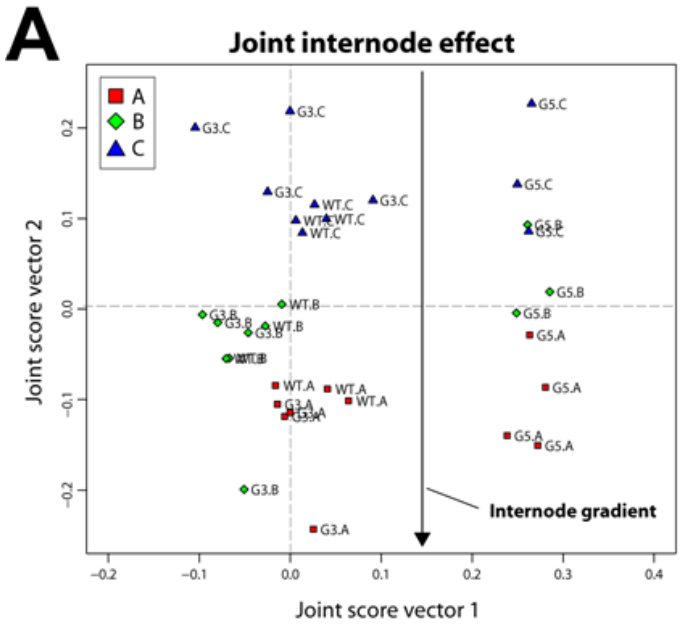


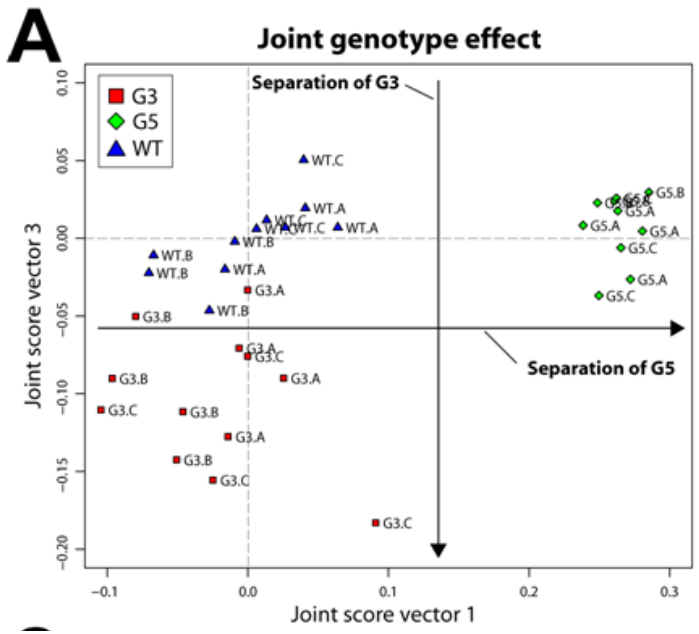
Combined profiling of transgenic Poplar

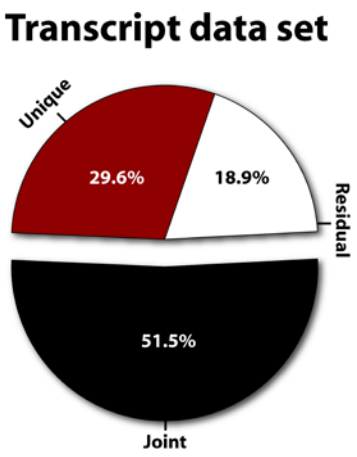


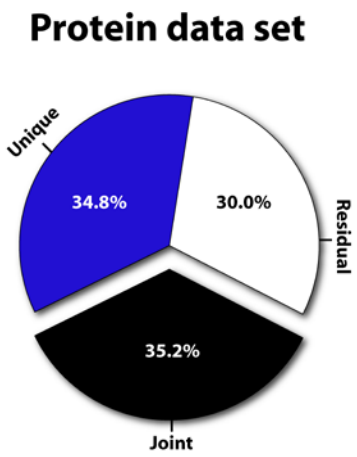
Same for Protein and Metabolite data

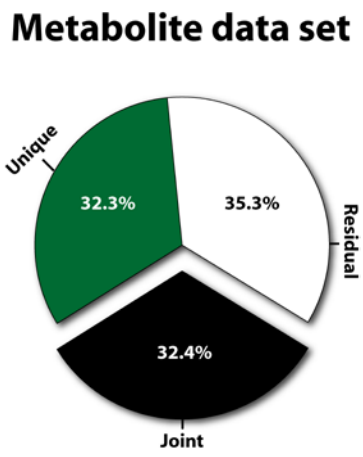
Combined profiling of transgenic Poplar







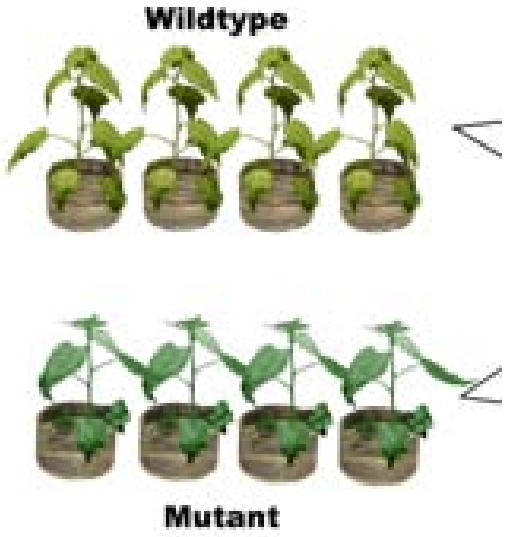


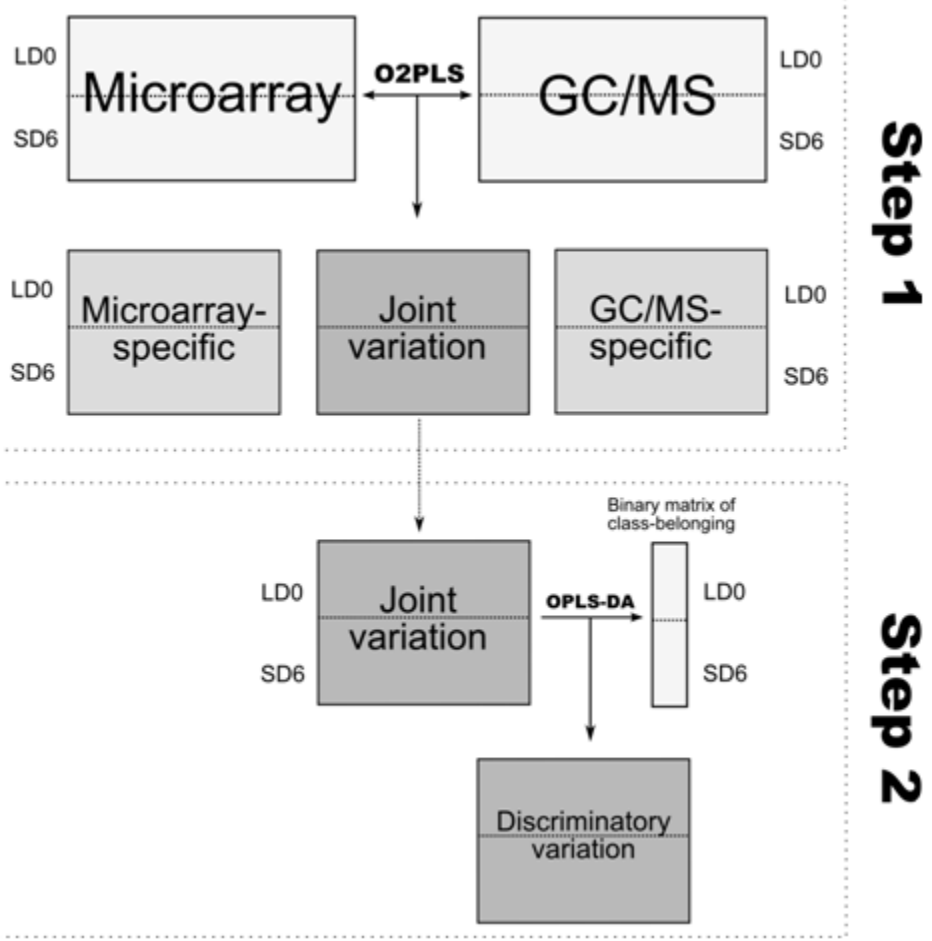


A combined profiling study of *Populus tremula* × *P. tremuloides*, investigating **short-day induced** effects at transcript and metabolite levels

-24 hybrid aspen (*Populus tremula* × *P. tremuloides*) trees
- growth chamber under long day conditions (12 h of PAR light (400 µEin m-2 s-1) and a 6-h daylength extension with low light (30 µEin m-2 s-1).

LD0 - Long day samples
SD6 – Short day

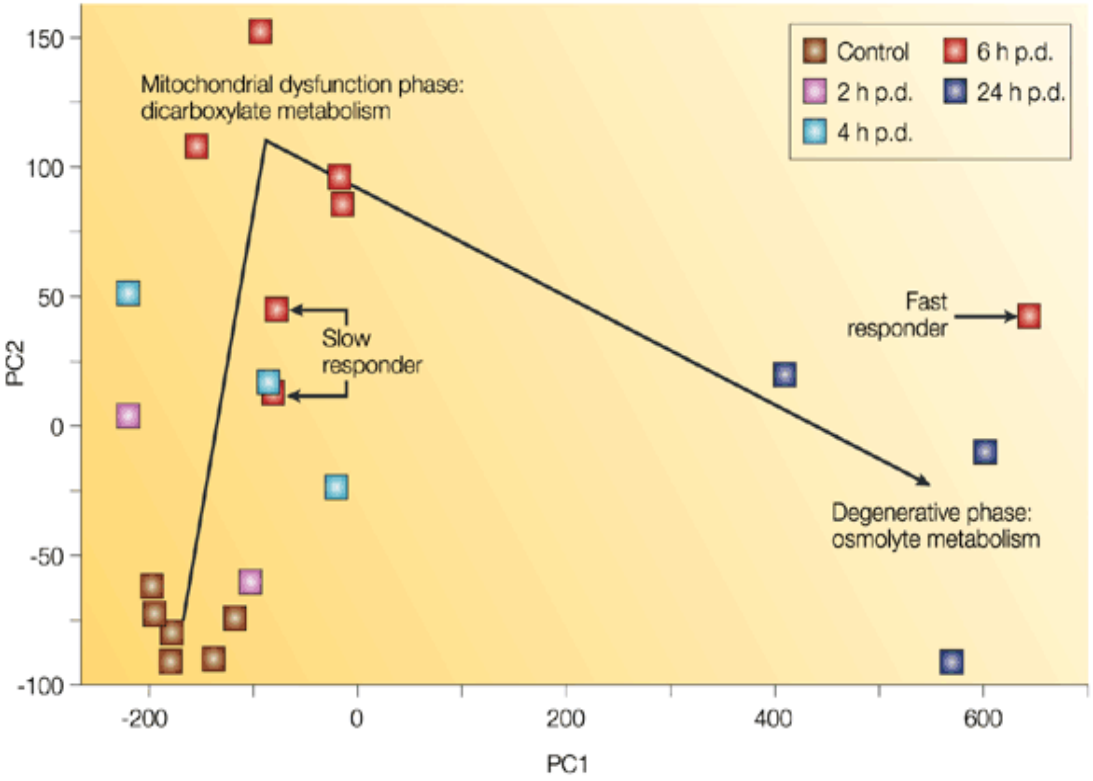




Dynamic modeling

Dynamic modeling

- Biological systems are dynamic processes that react to changes in their environment at both the cellular and organism levels.
- Modeling the time-related behavior of biological systems is essential for understanding the biology and underlying dynamics.

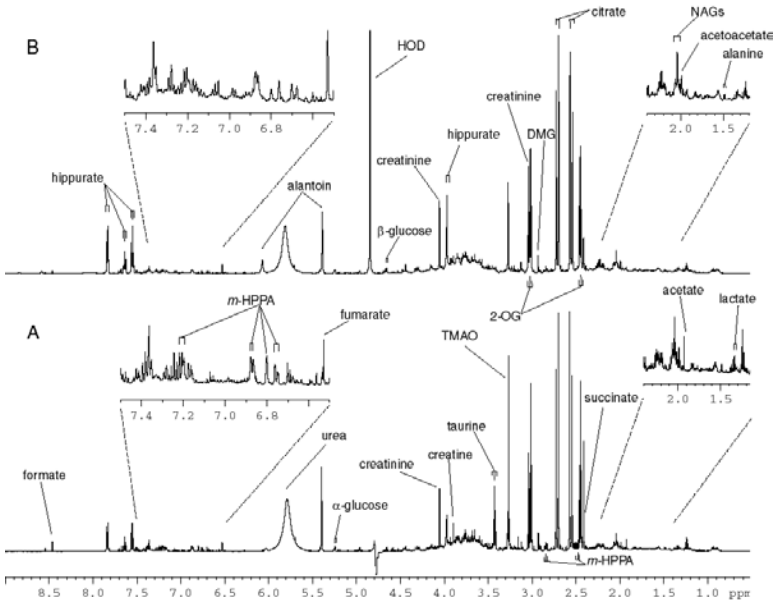


• PCA scores showing the trajectory of biochemical changes in the kidney after the administration of 2-bromoethanamine.
• Some animals respond to the intoxication faster than others, even though they are of uniform age and sex and were raised under the same conditions.
• This is a typical type of response, with 'slow' and 'fast' responders being characteristic of many drugs and toxins.

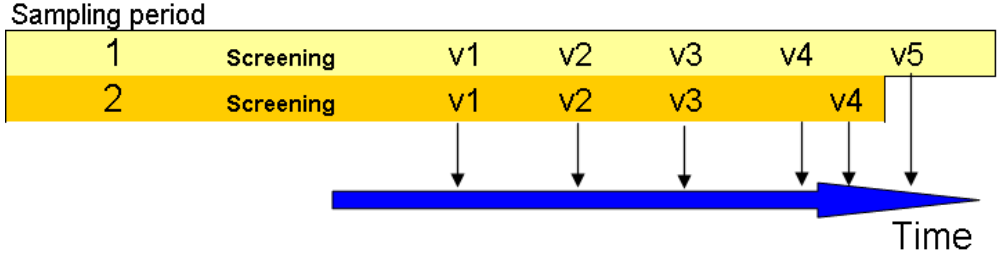
Nicholson, J.; Connelly, J.; Lindon, J.C.; Holmes, E.; **Metabonomics: a platform for studying drug toxicity and gene function,** Nature Reviews Drug Discovery , **2002**, 1, 153-161.

Example: Functional foods study

- Functional foods: Foodstuffs with a documented health-promoting effect – besides energy addition
- Centre for Human Studies of Foodstuffs, Sweden
  - Inclusion/exclusion criteria
  - 9 individuals given prepared foodstuff
  - Multiple visits – document effect over time

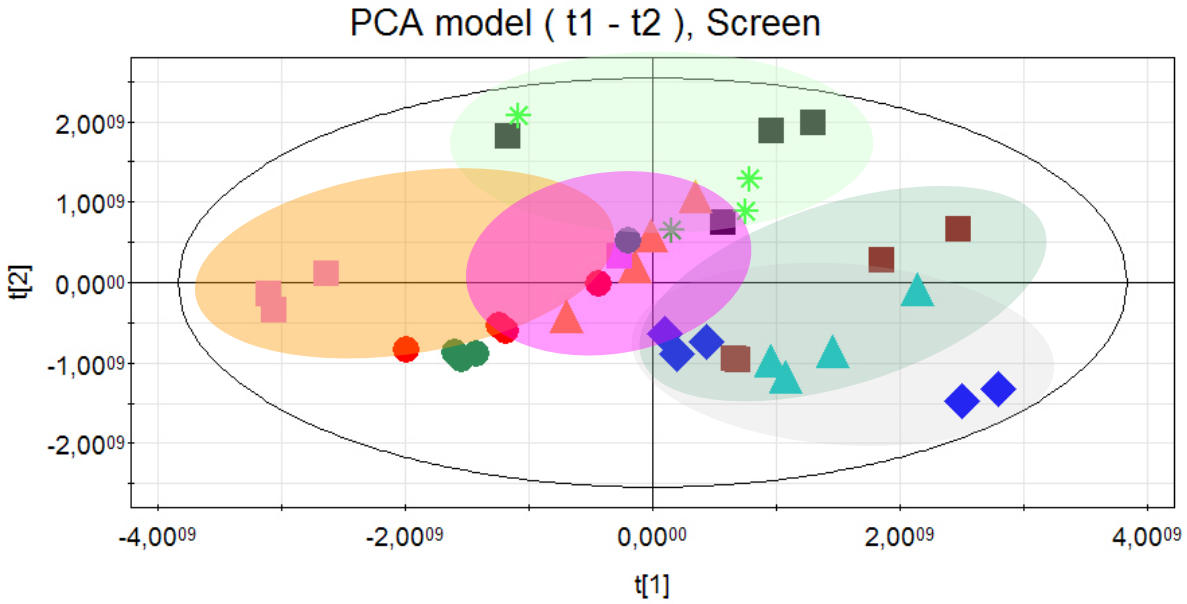


Study design



Functional foods study:

Individual metabolism vs metabolic response to food intake



Individuals metabolism baseline greater than the effect of foodstuffs

But… we are interested in the effect of foodstuffs

Dynamic (time-series) modeling

- In ‘omics (e.g. metabolic profiling) studies
  - the sampling rate and number of time points are often restricted (experimental, cost and biological constraints (< 4-15 time points).
  - Chemometrics:
    - MSPC batch modeling (Antti et al)
    - ANOVA based modeling, e.g. ASCA (Smilde et al), ANOVA-PCA (Harrington et al)
    - Dynamic Bayesian networks (Kim et al)
    - Auto-regressive moving average (ARMA, Box et al)
    - SMART analysis (Keun et al)
    - Independent component analysis (Morgenthal et al)
    - PARAFAC (Forshed et al)

Existing strategies for modeling dynamic data rests on two major assumptions:

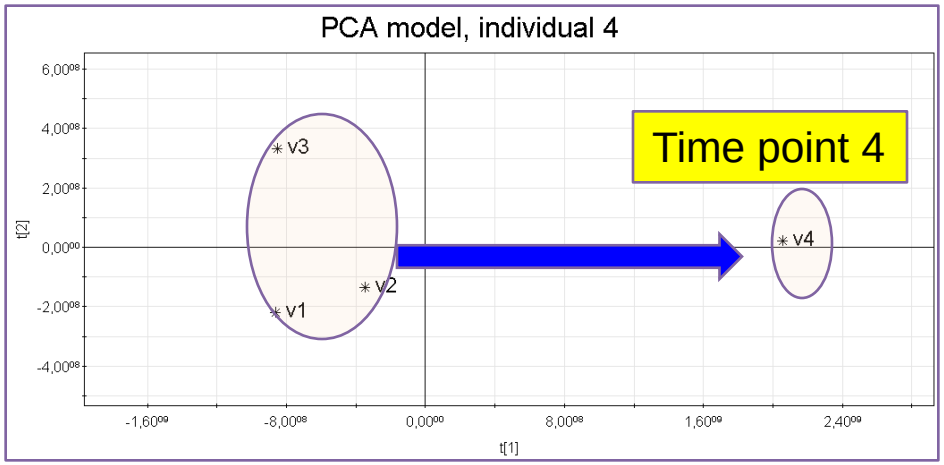
(1) The multivariate profile or fingerprint is comparable over all individuals.

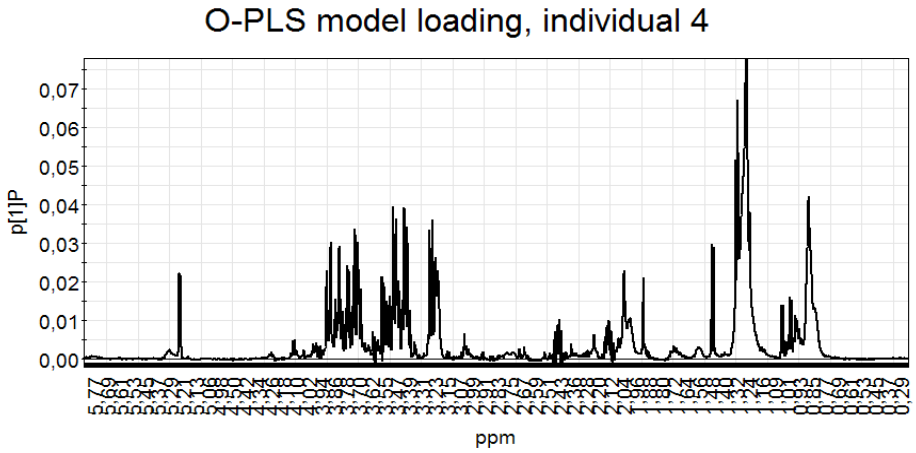
(2) The global temporal behavior is aligned between all individuals.

Dynamic modeling

- **Two alternative approaches using the OPLS model**
  - **Use OPLS property of single predictive components (+ Orthogonal components)**

1. Piece-wise dynamic modeling (Rantalainen et al)
2. **Dynamic modeling of individual effect profiles (Trygg et al)**





Understanding biochemistry

<u>Food stuff INCREASED</u>

Myo-inositol

Glutamine

Histidine

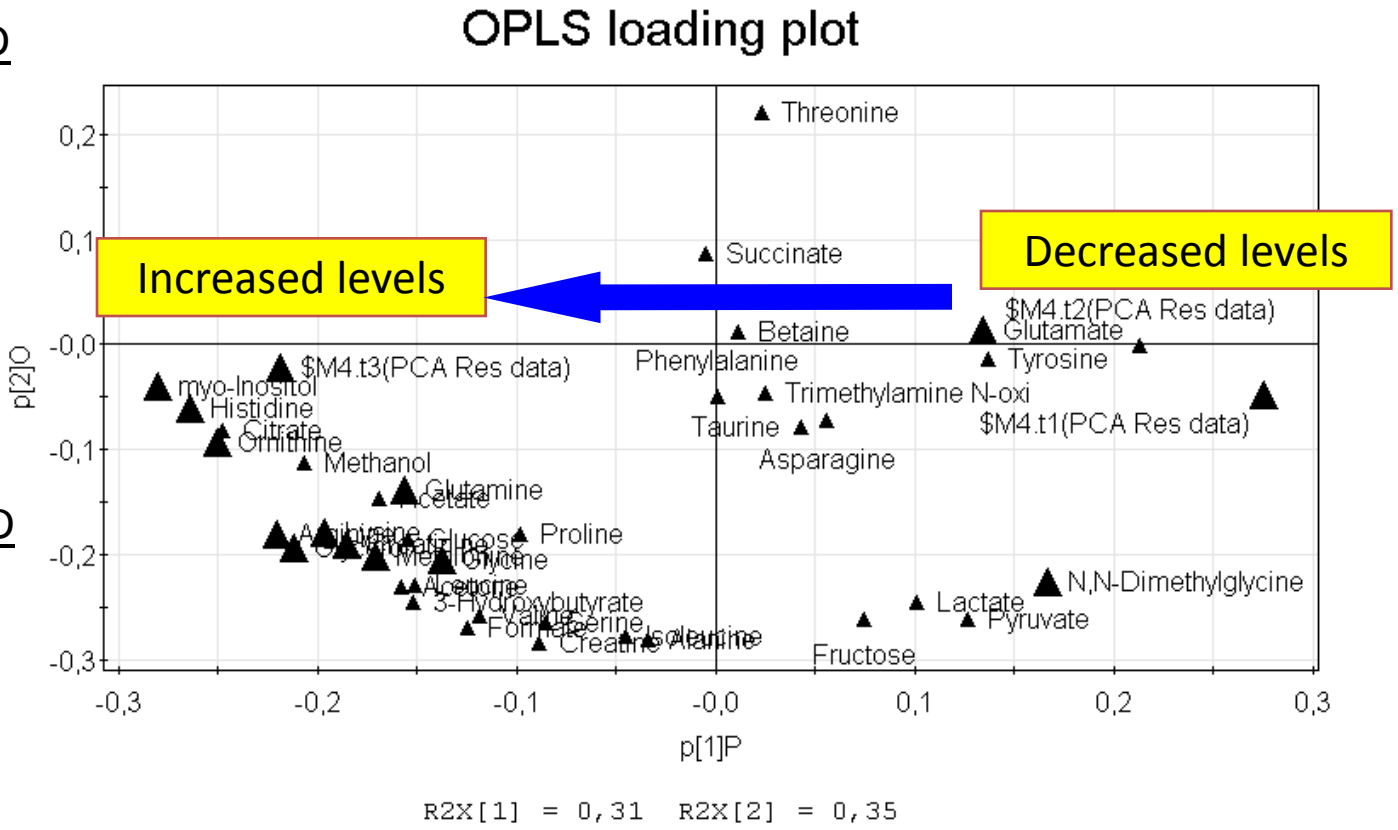
Citrate

Ornithine

<u>Food stuff DECREASED</u>

Glutamate

Triglycerides



Myo-inositol can have an effect on aminotransferase, supported by increase of ornithine, citrate and acetate. Myo-inositol has shown to have protective effect on cardiac dysfunction in diabetic rats.

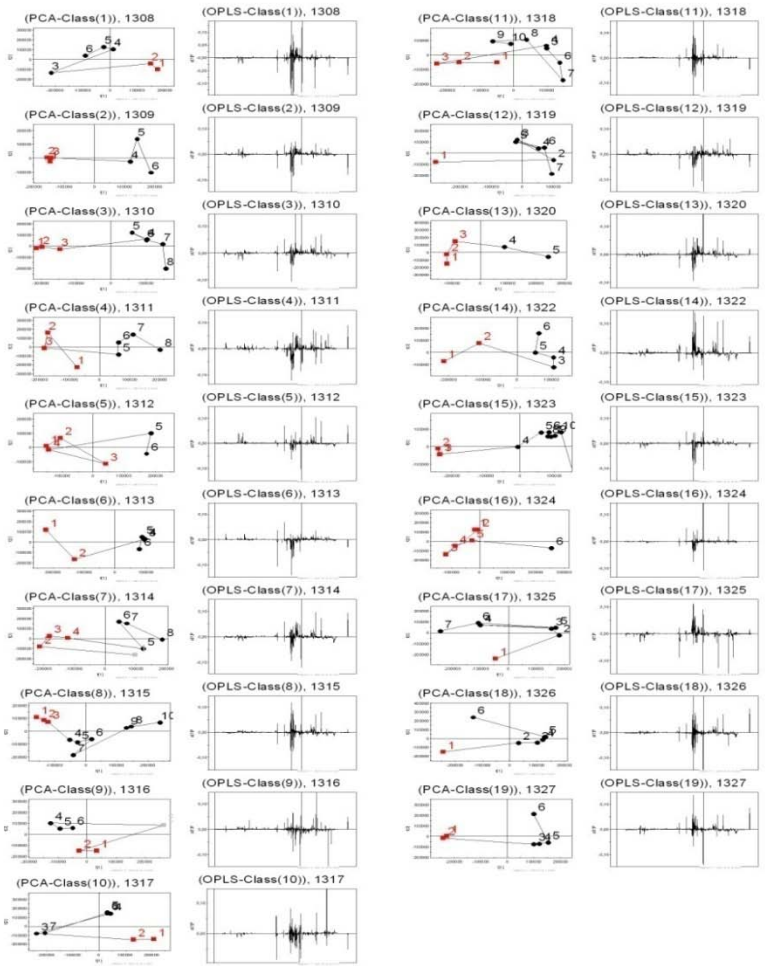
Example: Dynamic modeling

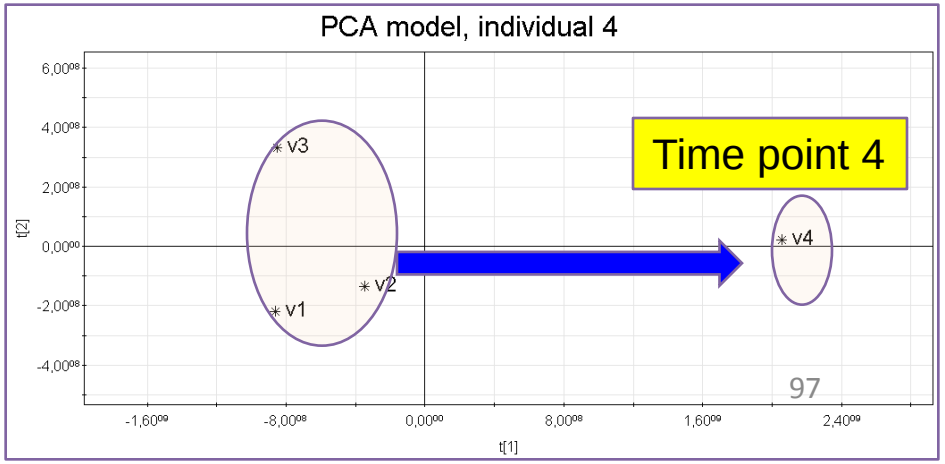
Kidney transplant study

NMR profiles of human urine samples after surgery

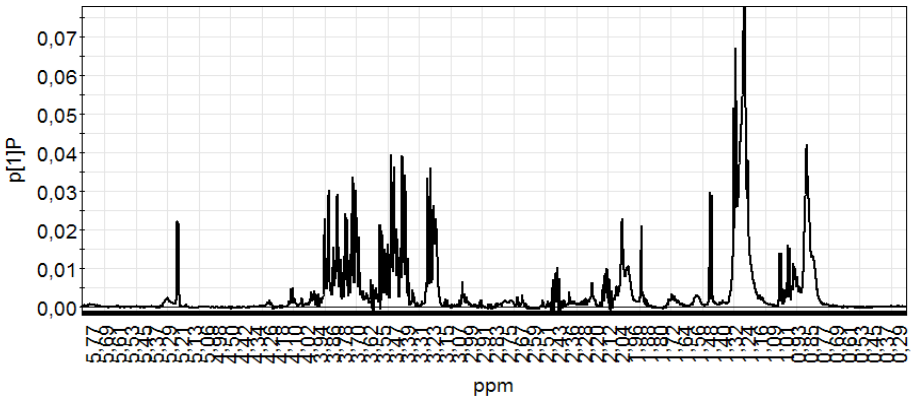
1.) Principal component analysis (PCA) t1/t2 score

Post-operative time trajectory





O-PLS model loading, individual 4

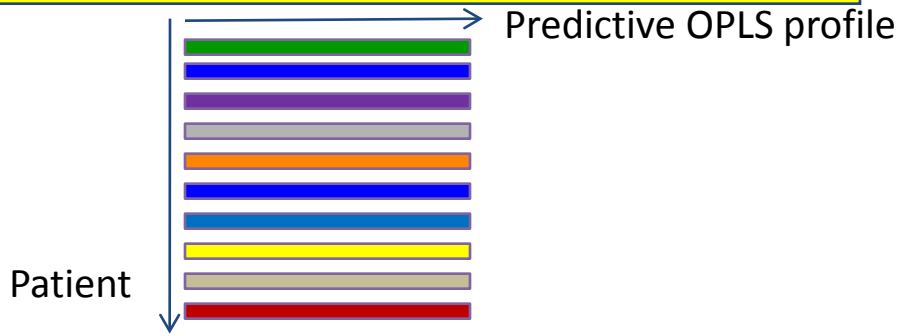


Example: Dynamic modeling

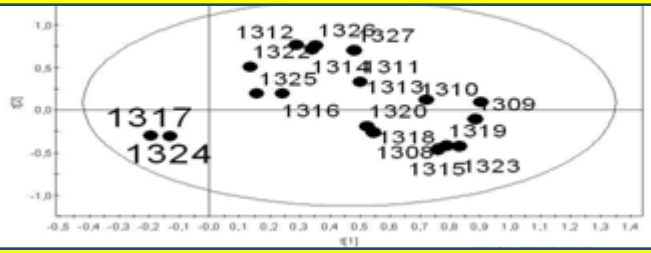
Kidney transplant study

NMR profiles of human urine samples after surgery

2.) Collect ALL patient ”recovery” profiles (Predictive OPLS component)



3.) PCA/SIMCA analysis of ALL patient ”recovery” profiles (Predictive OPLS component)



4.) Interpretation

Concluding remarks

- Metabolomics has a promising future in different areas in the post-genomic era
- Chemometrics shall be used in all steps of the metabolomics pipelines
- New methods in chemometrics are needed to understand huge loads of information
- Multi-block modeling strategies needed
- OPLS approach is appropriate to model data from metabolomics
- O-PLS is a multivariate prediction method, similar to PLS,
  - separates two different types of variations in the modelled data
  - TpPp = X-Y related variation
  - ToPo = Y-Orthogonal variation in X (unique variation in X)

- Regression coefficient profile b should not be used for interpretation
- OPLS allows model diagnostics, prediction and interpretation
- Different strategies using OPLS/O2PLS are useful for different purposes

Acknowledgements

Chemistry dep, Umeå University
M. Bylesjö
H. Stenlund
R. Madsen
M. Hedenström
S. Wiklund
P. Jonsson
H. Antti

Uppsala University
T. Lundstedt
J. Olsson

Umeå University Hospital
SB. Rantapää,
GM Alenius

Lund University & MAS
Å. Lernmark
L. Åkesson

Imperial College
J. Nicholson
E. Holmes
M. Rantalainen
O. Cloarec

Umeå Plant Science Center, Umeå Univ, Sweden
T. Moritz
D. Eriksson
A. Johansson
A. Sjödin
R. Nilsson
A Grönlund
S Jansson
B Sundberg
G. Wingsle
J. Karlsson
V. Srivastava
R. Bahlerao
G. Sandberg

Italy (Univ. Siena and more)
M. Calderisi
A. Vivi
M. Tassini
M Valensin,
M. Carmellini,
M. Cocchi

Riken University
M. Kusano

Acure Pharma
P. Lek
E. Seifert

AcureOmics
J. Gabrielsson

Anamar Medical
G Ekström

SweTreeTechnologies
M Hertzberg
K Johansson
A Karlsson

Umetrics AB,
E. Johansson
L. Eriksson
M. Earll
S. Wold

Chenomx
J. Newton
A. Weljie

Swedish Foundation for Strategic Research (SSF)
Swedish Research Council
FORMAS FuncFibre
Knut & Alice Wallenberg Foundation
GlaxoSmithKline
AstraZeneca
MKS Umetrics