Forecasting Chronic Diseases using Data Fusion

Evrim Acar

University of Copenhagen

Joint work with

Gözde Gürdeniz, Rasmus Bro, Lars Ove Dragsted, University of Copenhagen

Francesco Savorani, Polytechnic University of Turin

Louise Hansen, Anja Olsen, Anne Tjønneland, Danish Cancer Society Research Center
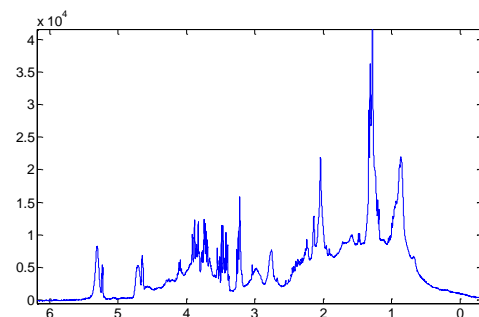
NuGO Week 2016

Motivation: Joint analysis of measurements from multiple platforms has the potential to enhance biomarker discovery!
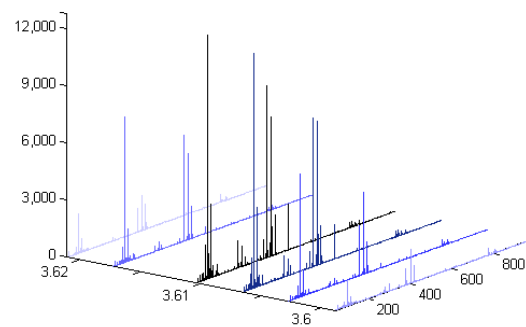
Metabolomics: The goal is to detect a wide range of chemical substances in biological fluids, e.g., blood, and to identify the chemicals related to certain conditions such as food intake and various diseases, e.g., cancer.

NMR (Nuclear Magnetic Resonance) Spectroscopy

Possible to measure using different analytical methods

**LC-MS
(Liquid Chromatography-
Mass Spectrometry)**

Motivation: Joint analysis of measurements from multiple platforms has the potential to enhance biomarker discovery!

Metabolomics: The goal is to detect a wide range of chemical substances in biological fluids, e.g., blood, and to identify the chemicals related to certain conditions such as food intake and various diseases, e.g., cancer.

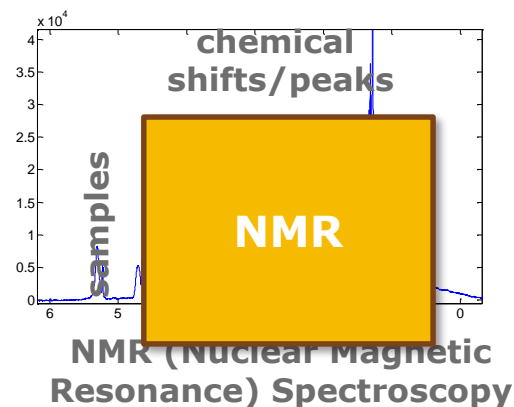chemical shifts/peaks

samples

NMR

NMR (Nuclear Magnetic Resonance) Spectroscopy

peaks

samples

LC-MS

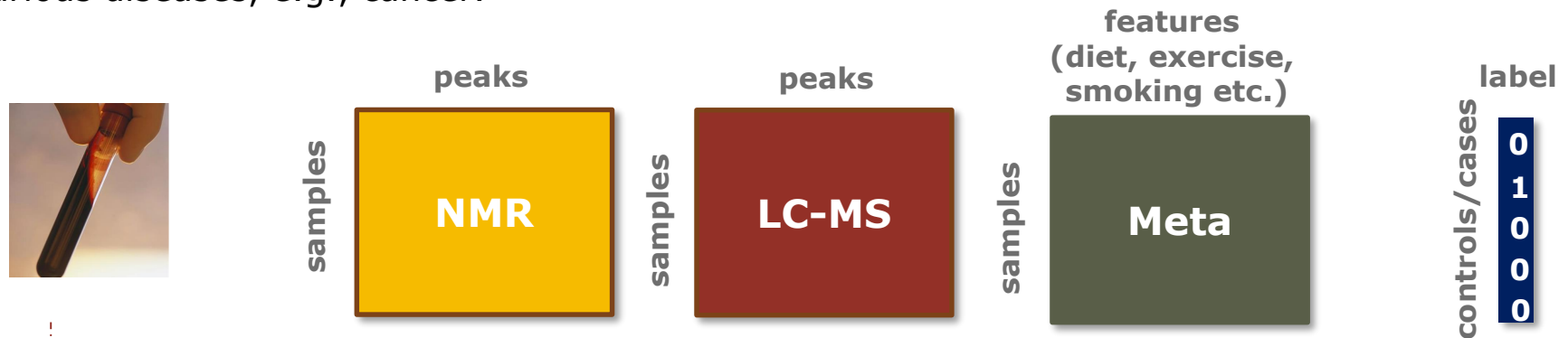LC-MS (Liquid Chromatography-Mass Spectrometry)

Motivation: Joint analysis of measurements from multiple platforms has the potential to enhance biomarker discovery!

Metabolomics: The goal is to detect a wide range of chemical substances in biological fluids, e.g., blood, and to identify the chemicals related to certain conditions such as food intake and various diseases, e.g., cancer.

1. *Given these data sets, can we forecast whether people will have certain diseases in near future?*

2. *Does data fusion improve our forecasting performance?*

3. *Can we capture biomarkers? Are we confident with those biomarkers?*

Our Data: Samples from subjects enrolled in the Danish Diet, Cancer and Health (DCH) Cohort [Tjønneland et al., 2007]

- Plasma samples (non-fasting) from subjects free of cancer at the time of sample collection (1993-1997). Some of those subjects develop the disease over time (time span is approximately up to 10 years).
- We are, in particular, interested in Acute Coronary Syndrome, Breast Cancer and Colorectal Cancer.
- Measurements: NMR and LC-MS
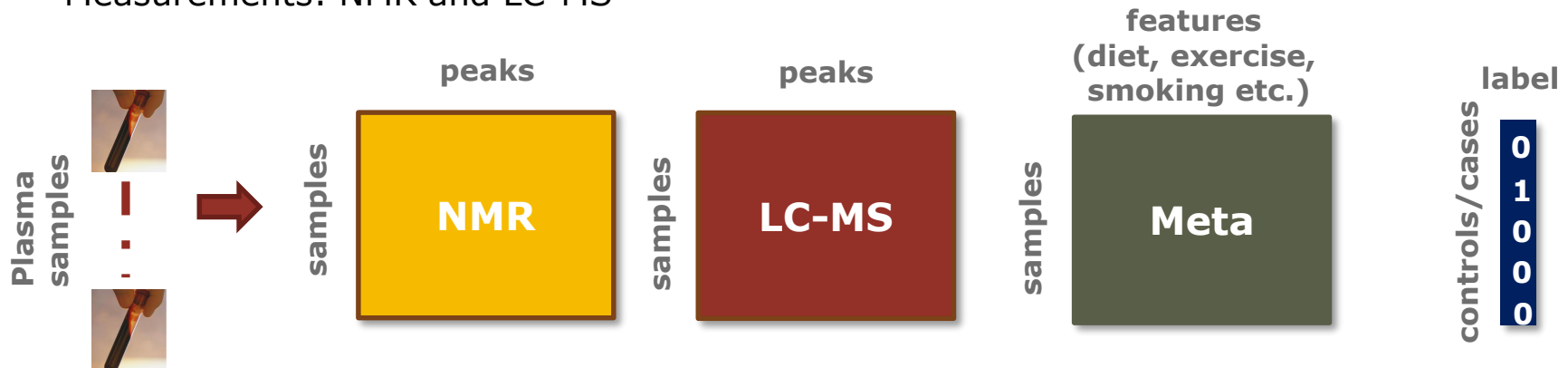
1. *Given these data sets, can we forecast whether subjects will develop the following diseases?*
   - *Acute Coronary Syndrome (ACS)*
   - *Breast Cancer*
   - *Colorectal Cancer*
2. *Does data fusion improve our forecasting performance?*
3. *Can we capture biomarkers? Are we confident with those biomarkers?*
   - *We will look into smoking and coffee to validate our approach.*

Supervised Data Fusion (a.k.a. multi-view learning)

peaks   peaks   features

samples   **NMR**   samples   **LC-MS**   samples   **Meta**   0 1 0 0 0   label
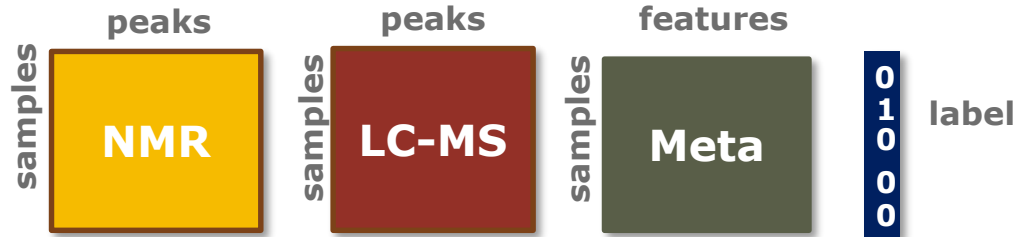
- Why not simply concatenate all views and represent each sample as one long vector of features? Due to many potential problems such as:
  - ➢ Increased risk of overfitting
  - ➢ Each data set may need different preprocessing
  - ➢ …
- Various multi-view learning approaches **[Xu et al., *arXiv:1304,* 2013]**:
  - *(i)* *high-level approaches* such as assigning a label based on predictions of multiple classifiers
  - *(ii)* *subspace learning-based methods* (i.e., finding latent subspaces and using those for classification)
  - *(iii)* *multiple kernel learning:* combining kernels corresponding to different views
- Many omics studies are interested in supervised data fusion:
  - *Subspace-based approaches:* Analysis of LC-MS and GC-MS (Gas Chromatography-Mass Spectrometry) data using multi-block PLS (Partial Least Squares) **[Smilde et al., 2005];** data integration in plant biology by jointly analyzing microarray and GC-MS data **[Bylesjö et al., 2007];** Joint analysis of LC-MS and NMR measurements of cerebrospinal fluid samples (CSF) **[Blanchet et al., 2011]**
  - Multiple kernel learning: fusion of GC-MS and NMR measurements of CSF samples to study multiple sclerosis **[Smolinska et al., 2012**]; consensus orthogonal PLS discriminant analysis for fusion of omics data **[Boccard et al., 2013]**

We use multiple kernel learning for supervised data fusion

Multiple Kernel Learning: Different kernels can be used as a measure of similarity for different views. Given multiple kernels, combining kernels is one possible way of combining information from multiple sources. See **[Gonen and Alpaydin, 2011]** for a nice survey.

**Form Kernel Matrices
(Linear/nonlinear kernels ??): LINEAR**

**WEIGHTED MEAN
(mkl-wmean)
Or
AVERAGE (mkl-mean)**
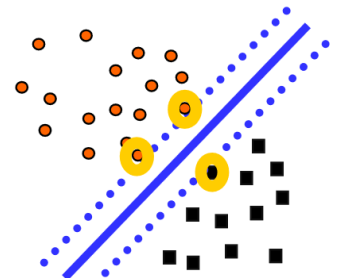
Combine Kernels
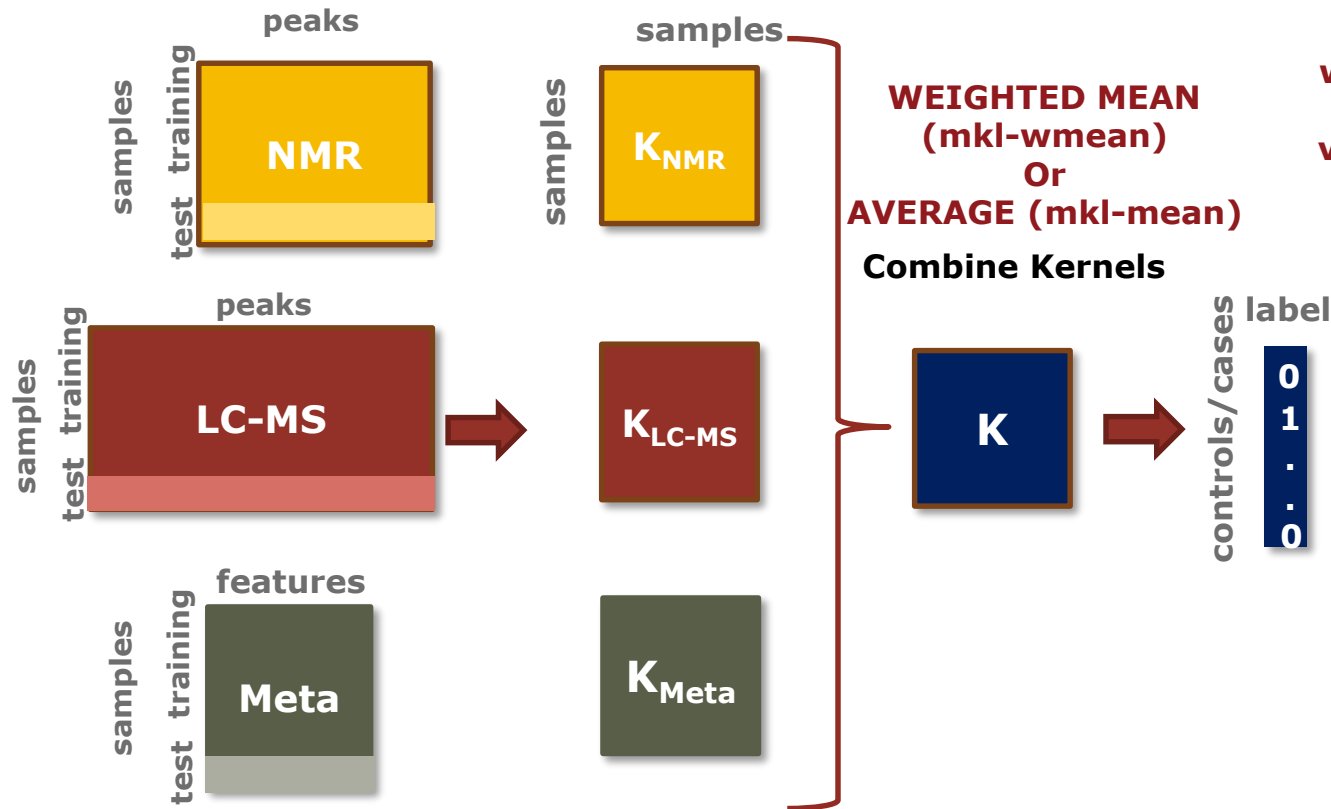
Parameters for SVM and weights in the weighted mean are learned using cross-validation in the training data.

Build a classifier using Support Vector Machines

$$D(x) = \mathbf{w}.\mathbf{x} + b$$

$$\mathbf{s}.\mathbf{t}.\mathbf{w} = \sum_{n=1}^{N} \alpha_n y_n \mathbf{x_n}$$

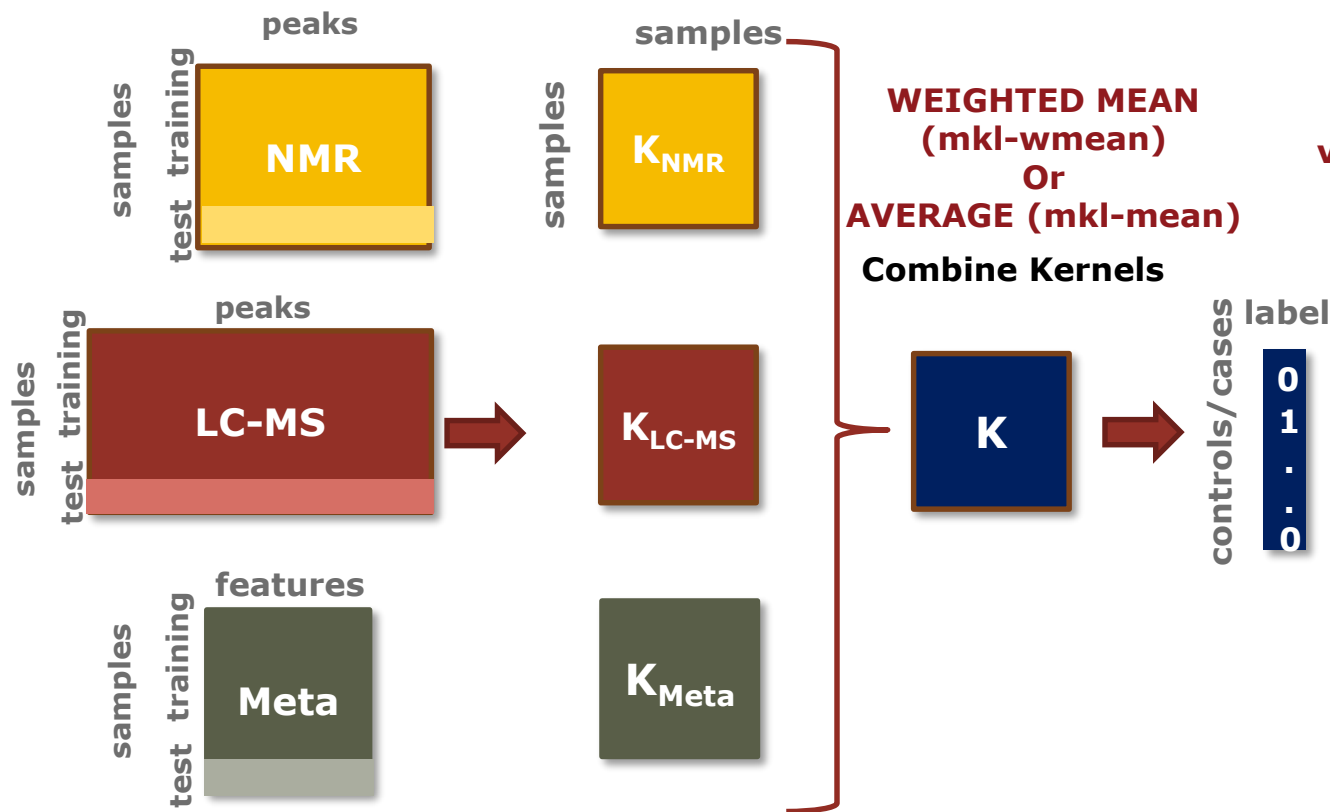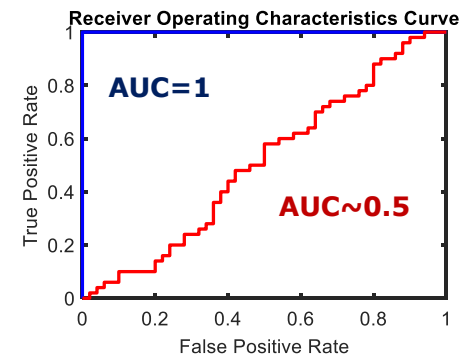$$b = <y_n - \mathbf{w}.\mathbf{x_n}>$$

[Bennett & Campbell, 2000]

Performance Evaluation in terms of ROC (Receiver Operating Characteristics) Curves

Multiple Kernel Learning: Different kernels can be used as a measure of similarity for different views. Given multiple kernels, combining kernels is one possible way of combining information from multiple sources. See **[Gonen and Alpaydin, 2011]** for a nice survey.

Form Kernel Matrices
(Linear/nonlinear kernels ??): LINEAR

peaks

samples
test training

NMR

samples

K$_{NMR}$

WEIGHTED MEAN (mkl-wmean) Or AVERAGE (mkl-mean)

Parameters for SVM and weights in the weighted sum are learned using cross-validation in the training data.

peaks

samples
test training

LC-MS

K$_{LC-MS}$

Combine Kernels

K

controls/cases

label

0
1
.
.
0

Build a classifier using Support Vector Machines

$$f(x) = \mathbf{w}.\mathbf{x}_{\text{test}} + b$$

$$\mathbf{v.s.}$$

$$y_{test}$$

features

samples
test training

Meta

K$_{Meta}$

Receiver Operating Characteristics Curve
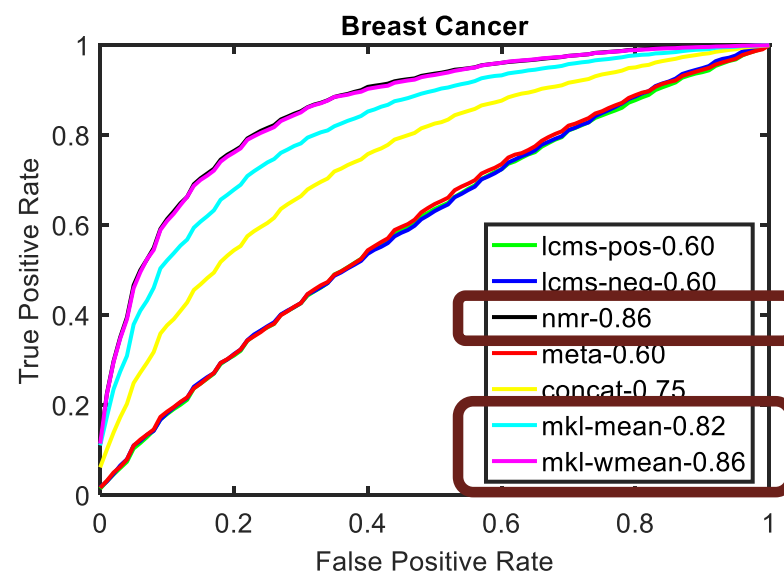
AUC=1

AUC~0.5

True Positive Rate

False Positive Rate
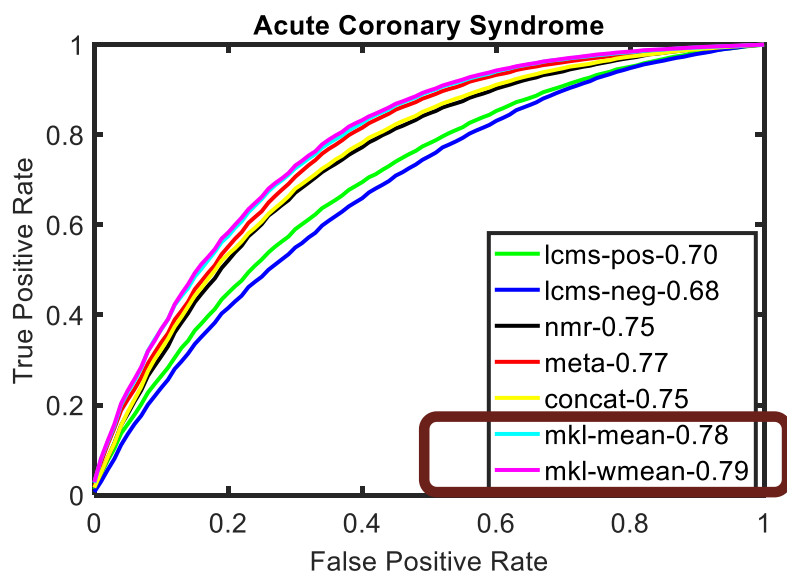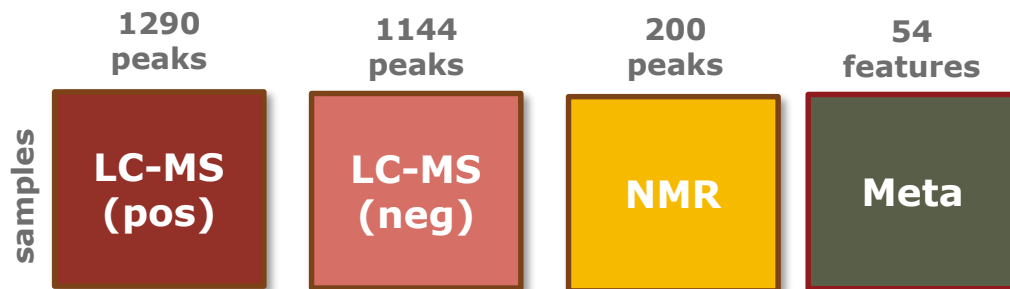
Results Summary
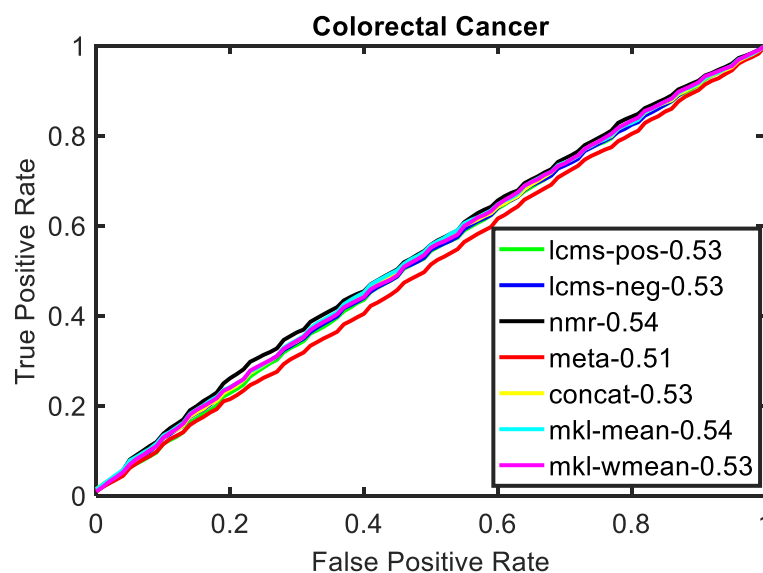
Acute Coronary Syndrome(ACS): Data fusion improves the forecasting performance for ACS!

1290 peaks

1144 peaks

200 peaks

54 features

samples

| LC-MS (pos) | LC-MS (neg) | NMR | Meta |

Acute Coronary Syndrome

True Positive Rate

False Positive Rate

lcms-pos-0.70
lcms-neg-0.68
nmr-0.75
meta-0.77
concat-0.75
mkl-mean-0.78
mkl-wmean-0.79

- In total, **3376** samples (**1092** cases/ **2284** controls).

- While forming training and test sets, we take equal number of cases and controls (randomly) and use 70% of the samples as the training set and 30% of the samples as the test set, i.e.,

  - Training set: **1530** samples

  - Test set: **654** samples

- Results are based on 100 such random training – test splits.
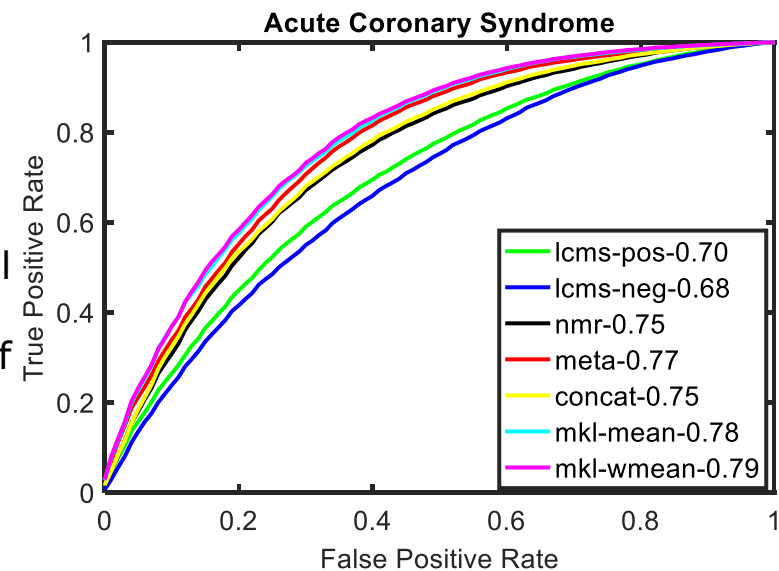
RESULTS:

- Data fusion (using both mkl-mean & mkl-wmean) performs better than individual analysis of data sets. Note that concatenation ("concat") is not the solution!

- In mkl-wmean, weights selected by cross-validation are higher for NMR in 44% of the runs, higher for meta data in 49% of the runs, and equal for all data sets in 7% of the runs.

ACS: Significant features in a fusion model with a performance close to the average performance (AUC=0.78)

$$D(x) = \mathbf{w}.\mathbf{x} + b$$

$$\mathbf{s.t.} \, \mathbf{w} = \sum_{n=1}^{N} \alpha_n y_n \mathbf{x_n}$$

$$b = \langle y_n - \mathbf{w}.\mathbf{x_n} \rangle$$
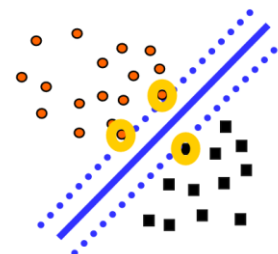
LC-MS (positive)	LC-MS (negative)	NMR	Meta
Cotinine *(smoking)*	Formate cluster of unknown plasma lipids	Cholesterol	Male **Gender effect**
Fragment of LPC(17:0) *(dairy)*	DG(40:8)	Choline	Female
5.7719 427.2842	Tetrahydrocorticosterone or Tetrahydrodeoxycortisol	?	Years of smoking
PE(40:6)	0.49796 972.2975	?	Low-level school **Smoking**
4.557 417.3468	1.0378 260.0448	?	Current smoker
Caffeine *(coffee)*	0.70387 111.9631	?	Never smoker
4.909 328. 2305	3.6106 725.5184	Valine	High-level school
4.723 594.3582	4.9296 457.225	Mainly glucose and protein	1 if either >35y at first birth or no children at all
0.69012 147.0294	3.0921 380.0944	?	Blood pressure
5.2717 786.507	4.7867 101.934	?	Years since quitting smoking

Acute Coronary Syndrome(ACS): Removing gender effect by looking into females and males separately!

ACS- Females Only

- lcms-pos-0.63
- lcms-neg-0.62
- nmr-0.72
- meta-0.74
- concat-0.70
- mkl-mean-0.77
- mkl-wmean-0.76

ACS- Males Only

- lcms-pos-0.62
- lcms-neg-0.59
- nmr-0.67
- meta-0.67
- concat-0.65
- mkl-mean-0.69
- mkl-wmean-0.70
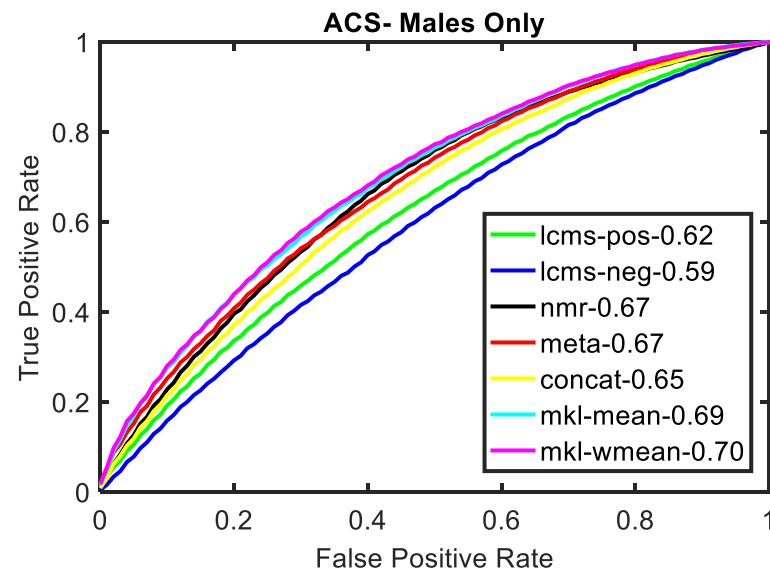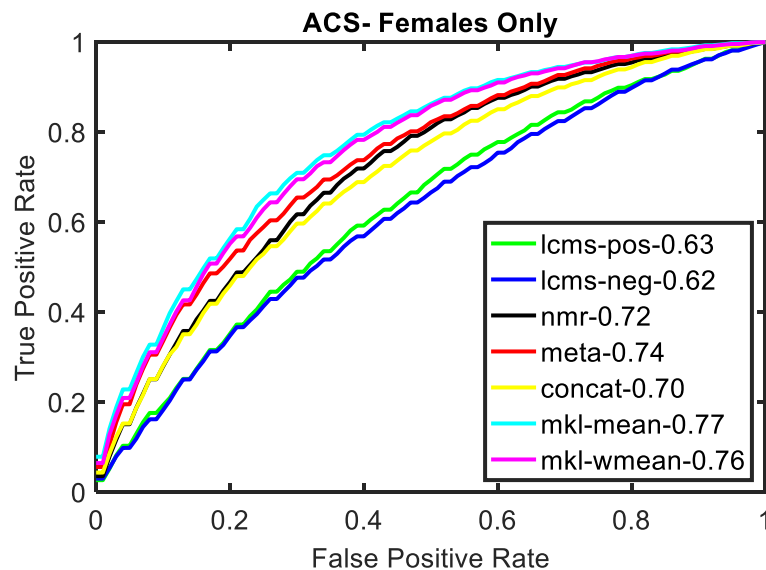
RESULTS:

- We see a drop in performance (in females from 0.79 to 0.77, and in males from 0.79 to 0.70) so gender does indeed play a significant role.
  - Females: Training set: 362 samples, Test set: 134 samples
  - Males: Training set: 1168 samples, Test set: 500 samples

  (We can expect a decrease in the performance due to the decrease in the number of samples but still we have many samples to build the models).

- Data fusion still improves the performance in both males and females.

ACS – Females Only: Significant features in a fusion model with a performance close to the average performance (AUC=0.77)

$$D(x) = \mathbf{w}.\mathbf{x} + b$$

$$\mathbf{s.t.\,w} = \sum_{n=1}^{N} \alpha_n y_n \mathbf{x_n}$$

$$b = <y_n - \mathbf{w}.\mathbf{x_n}>$$
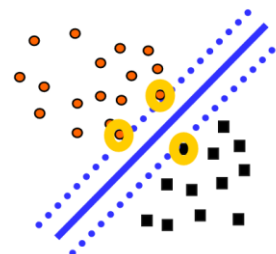
LC-MS (positive)	LC-MS (negative)	NMR	Meta
Cotinine	4.6767 142.9247	?	Current smoker
4.6161 277.2185	0.70361 193.1955	?	Years of smoking
5.3291 145.9565	4.2568 579.2974	?	Former smoker
3.668 472.3038	0.7048 192.3631	**Cholesterol**	Never smoker
4.4242 233.0796	4.8341 367.2635	**Choline**	High-level school
4.8958 137.06	4.8288 489.2595	?	Low-level school
Testosterone	0.70457 192.2054	?	Total amount of tobacco consumption (daily)
5.1136 178.9475	3.9639 445.1124	?	Years since quitting smoking
4.7973 287.1309	0.70662 192.1081	**Pyruvate**	Blood pressure, systolic
4.7477 548.3712	0.70048 111.9481	**Histidine**	Coffee (g/d)

Smoking effect

ACS - Females & Non-Smokers Only: Removing the effect of smoking by looking into only former smokers and the ones who have never smoked!
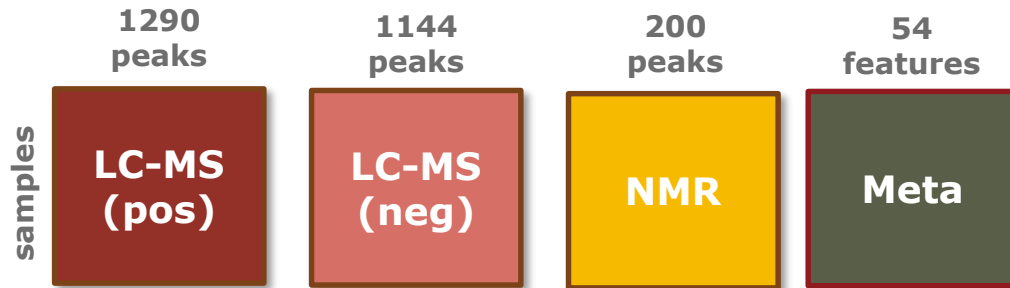

ACS- Females (Non-Smokers) Only
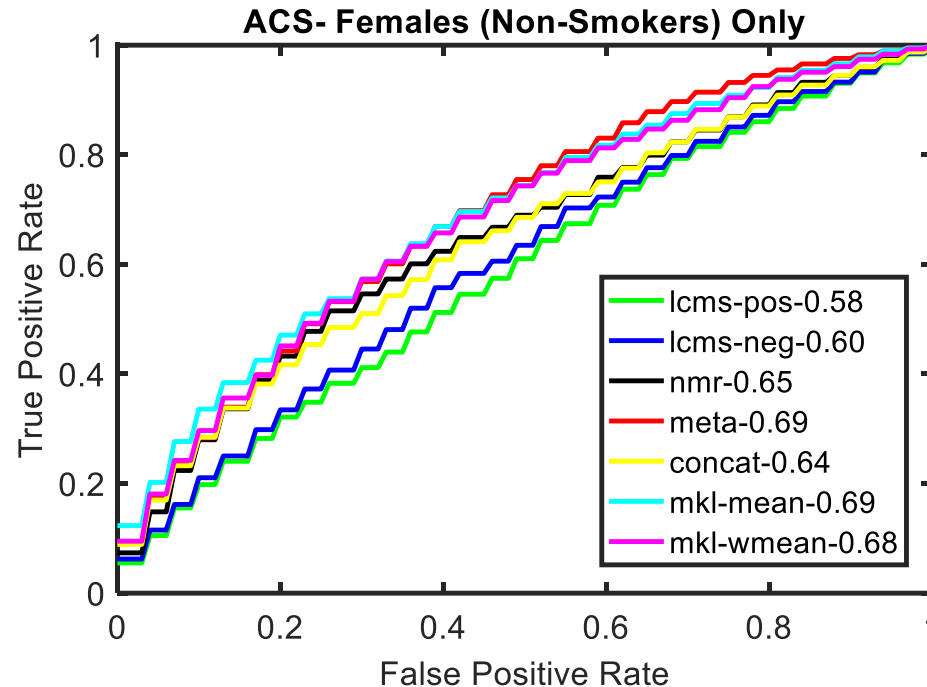
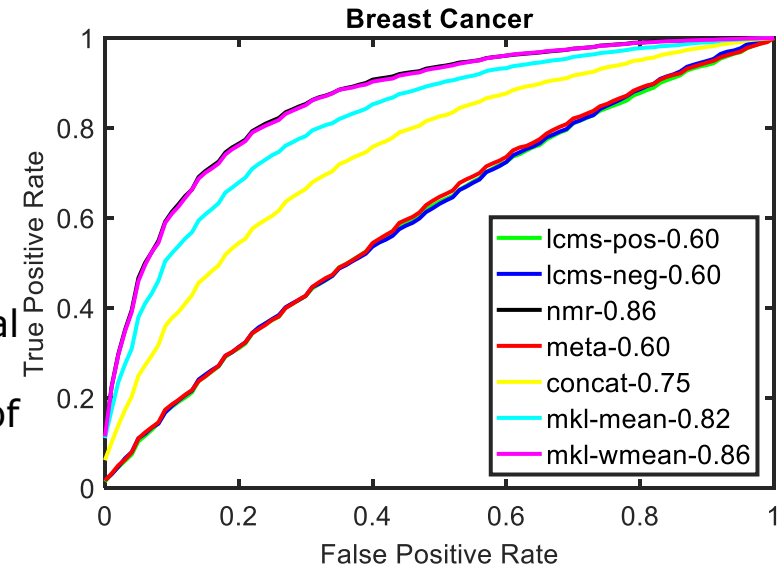*Better than random but low performance!*

Summary (ACS):

- Improved forecasting performance using supervised fusion
- Major effects are gender, smoking and cholesterol!

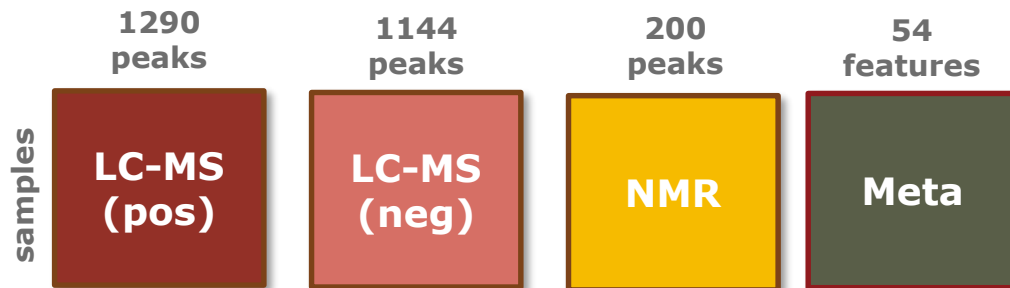# Breast Cancer: Data fusion is not always a good idea!

1290 peaks — **LC-MS (pos)**

1144 peaks — **LC-MS (neg)**

200 peaks — **NMR**

54 features — **Meta**

samples

- In total, **1589** samples (**412** cases/ **1177** controls) from *females*.
- While forming training and test sets, we take equal number of cases and controls (randomly) and use 70% of the samples as the training set and 30% of the samples as the test set, i.e.,
  - Training set: **578** samples
  - Test set: **246** samples
- Results are based on 100 such random training – test splits.

Breast Cancer

True Positive Rate / False Positive Rate

- lcms-pos-0.60
- lcms-neg-0.60
- nmr-0.86
- meta-0.60
- concat-0.75
- mkl-mean-0.82
- mkl-wmean-0.86

NMR performs the best!

Bro et al., *Metabolomics,* 2015 achieves a similar performance using NMR and meta data*.*

Colorectal Cancer (CRC): Nothing works if the goal is to forecast CRC cases!

samples

1290 peaks

1144 peaks

200 peaks

54 features

LC-MS (pos) **LC-MS (neg)** **NMR** **Meta**

Colorectal Cancer

- True Positive Rate vs False Positive Rate
  - lcms-pos-0.53
  - lcms-neg-0.53
  - nmr-0.54
  - meta-0.51
  - concat-0.53
  - mkl-mean-0.54
  - mkl-wmean-0.53

- In total, **3376** samples (**408** cases/ **2968** controls).

- While forming training and test sets, we take equal number of cases and controls (randomly) and use 70% of the samples as the training set and 30% of the samples as the test set, i.e.,

  - Training set: **572** samples

  - Test set: **244** samples

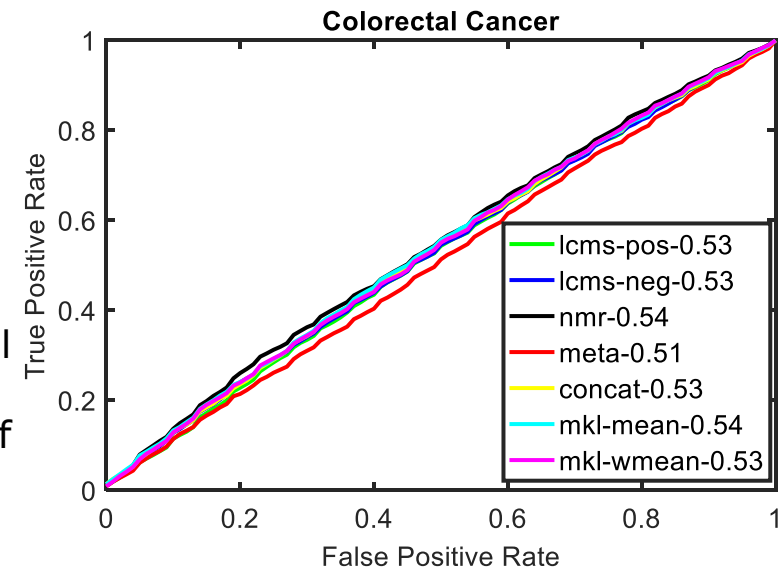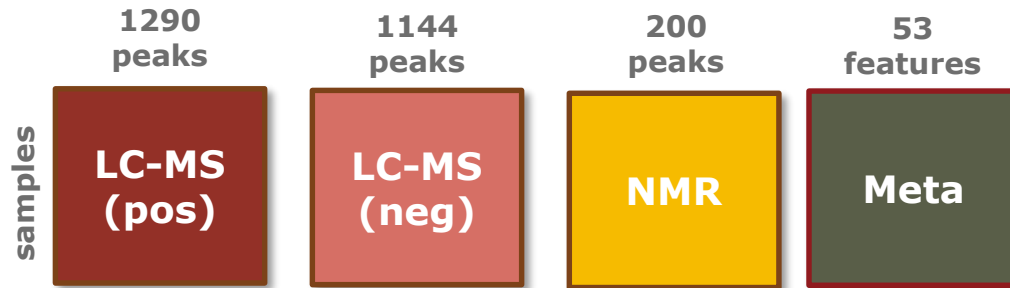- Results are based on 100 such random training – test splits.

*Is it the biology, time resolution or the modeling approach???*

The models reveal meaningful biomarkers!

Case Study 1: Coffee Markers

1290 peaks	1144 peaks	200 peaks	53 features

samples

LC-MS (pos)	LC-MS (neg)	NMR	Meta

Coffee Drinking (Non-Smokers)

Legend:
- lcms-pos-0.84
- lcms-neg-0.92
- nmr-0.74
- meta-0.75
- concat-0.92
- mkl-mean-0.91
- mkl-wmean-0.93

- **Non-smokers**
- In total, **641** samples (**370** cases(coffee-drinkers)/ **271** controls (not drinking coffee)).
- While forming training and test sets, we take equal number of cases and controls (randomly) and use 70% of the samples as the training set and 30% of the samples as the test set, i.e.,
  - Training set: **380** samples
  - Test set: **162** samples
- Results are based on 100 such random training – test splits.
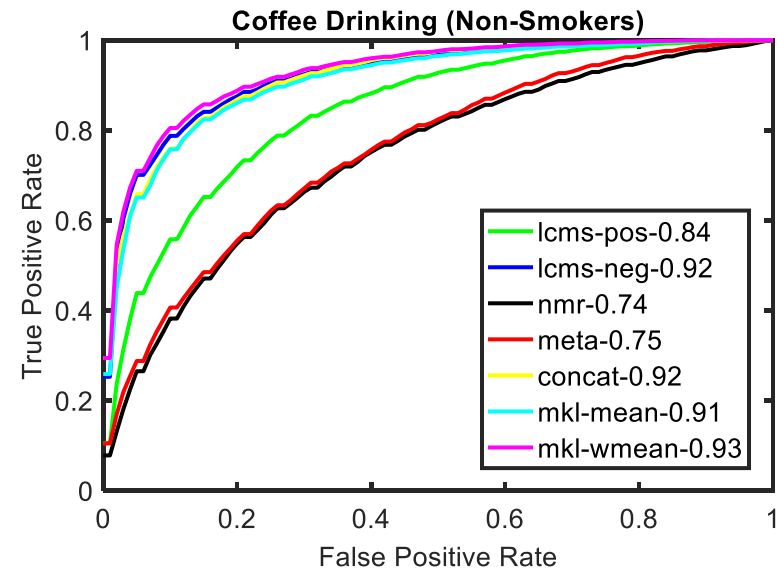
- *Fusion using mkl-wmean improves the performance!*
- *In 86% of the runs, mkl-wmean gives the following weights to the data sets: 0.7 (LC-MS negative), 0.1 (LC-MS positive), 0.1 (NMR), 0.1 (Meta). 11% of the runs uses only LC-MS negative mode.*

Coffee drinking: Significant features in a fusion model with a performance close to the average performance (AUC=0.93)

$$D(x) = \mathbf{w}.\mathbf{x} + b$$

$$\mathbf{s.t.w} = \sum_{n=1}^{N} \alpha_n y_n \mathbf{x_n}$$

$$b = < y_n - \mathbf{w}.\mathbf{x_n} >$$
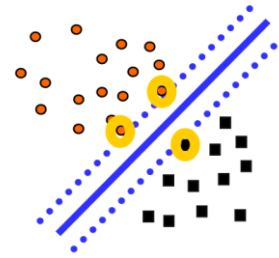
LC-MS (positive)	LC-MS (negative)	NMR	Meta
0.66518 181.0732	**Quinic acid** *(coffee)*	?	Rye bread[g/d]
Caffeine *(coffee)*	3.6983 495.2231 *(coffee)*	?	Low-level school
0.52274 138.055	3.6969 517.2054 *(coffee)*	**Creatinine**	High-level school
1,7-dimethylxanthine Fragment *(coffee)*	3.4666 429.1583	**Glucose**	Female
4.9097 382.2082	3.6386 413.1632	**Mainly glucose and protein**	Male
3.1213 138.0668	**Cafestol adduct 1** *(coffee)*	**Glucose**	Fatty dairy products (g/d)
1,7-dimethylxanthine *(coffee)*	**Cafestol adduct 2** *(coffee)*	?	Current user of NSAIDS
4.6052 98.98434	2.3967 195.0512	?	Never Smoker
2.3447 185.1292	**1 or 3 Methyluric acid** *(coffee)*	?	Former Smoker
3-methylxanthine OR 7-methylxanthine *(coffee)*	0.51941 179.0552	?	Sugars total (g/d)

The models will also pick up the confounding variables!

Case Study 2: Smoking Markers

1290 peaks

1144 peaks

200 peaks

48 features

samples

LC-MS (pos)	LC-MS (neg)	NMR	Meta
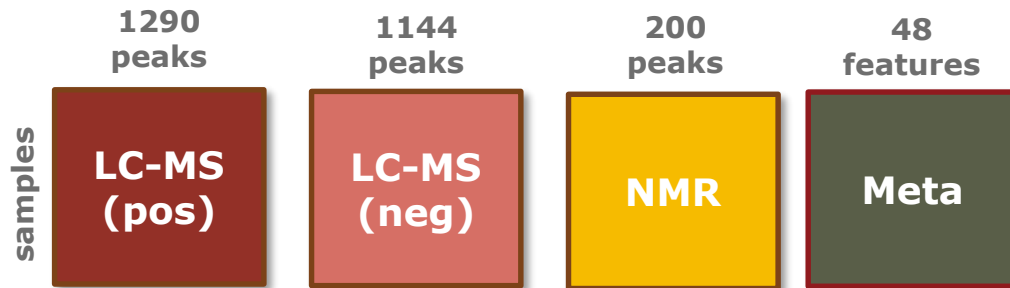
- **Current smokers & never-a-smokers (former smokers excluded)**
- In total, **2466** samples (**1471** smokers/ **995** never-a-smokers).
- While forming training and test sets, we take equal number of cases and controls (randomly) and use 70% of the samples as the training set and 30% of the samples as the test set, i.e.,
    - Training set: **1394** samples
    - Test set: **596** samples
- Results are based on 100 such random training – test splits.

Smoking

True Positive Rate vs False Positive Rate

- lcms-pos-0.90
- lcms-neg-0.86
- nmr-0.80
- meta-0.80
- concat-0.91
- mkl-mean-0.90
- mkl-wmean-0.92
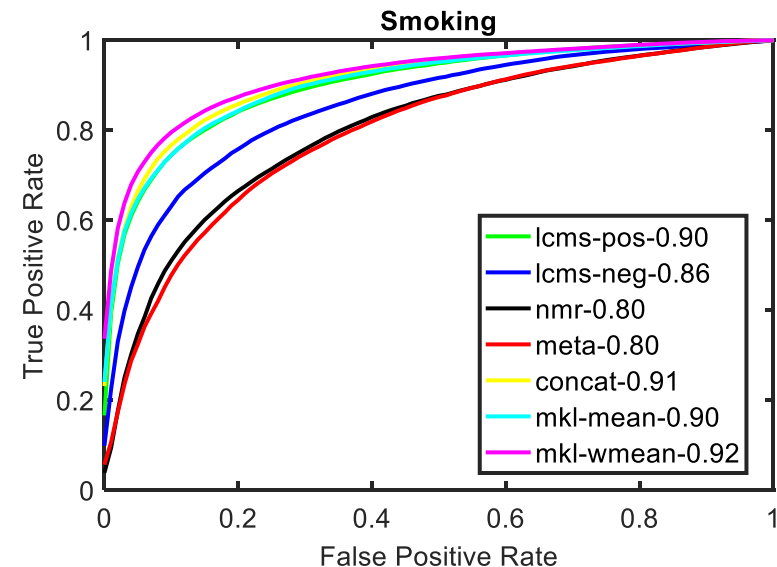
- *Data fusion using mkl-wmean performs the best!*
- *In all runs, mkl-wmean uses the following weights for the data sets: 0.7 for LC-MS positive, 0.1 for LC-MS negative mode, 0.1 for NMR and 0.1 for Meta data.*

Smoking: Significant features in a fusion model with a performance close to the average performance (AUC=0.92)

$$D(x) = \mathbf{w}.\mathbf{x} + b$$

$$\mathbf{s.t.w} = \sum_{n=1}^{N} \alpha_n y_n \mathbf{x_n}$$

$$b = <y_n - \mathbf{w}.\mathbf{x_n}>$$
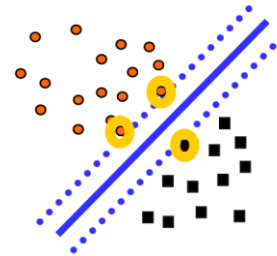
| LC-MS (positive)	LC-MS (negative)	NMR	Meta
Cotinine *(smoking)*	3.5214 651.3422	?	Coffee (g/d)
Caffeine *(coffee)*	3.5254 199.0062	?	Low-level school
0.57949 160.1341	3.5912 201.0218	Leucine and Isoleucine	Medium-level school
3.1213 138.0668	0.70663 192.3785	?	Dietary fibers (g/d)
0.4542 313.0382	0.70675 391.032	?	Male
3-methylxanthine OR 7-methylxanthine *(coffee)*	3.9759 435.1458	?	Female
4.7541 538.3133	4.2615 737.2553	Lipid	1 if either >35 y at first birth …
0.70023 791.0996	0.70223 192.9709	Choline	Marine fats (n-3) in diet (g/d)
Theobromine *(coffee)*	5.0276 809.0491	Lipid, CH3	Fruits (ex. juice) (g/d)
Hesperetin 7-o-|a-d-glucuronide	0.70775 192.9034	?	Carbohydrate (g/d)

Smoking: Smokers drink more coffee!
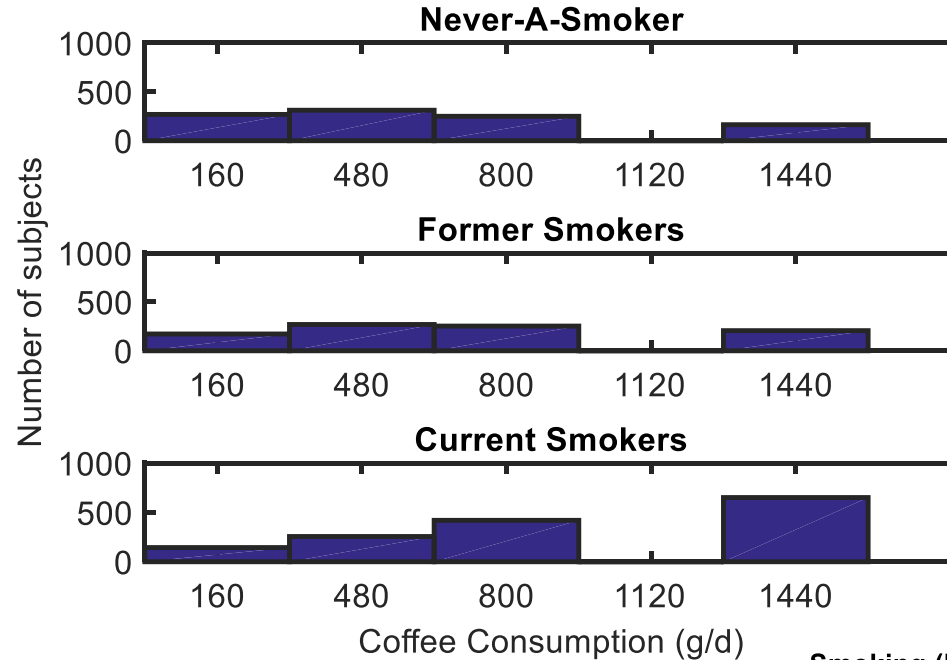
In order to eliminate the coffee effect, if we only take the subjects who do not drink coffee, we are left with very few samples and cannot build good models (64 never-a-smoker & 35 smokers).
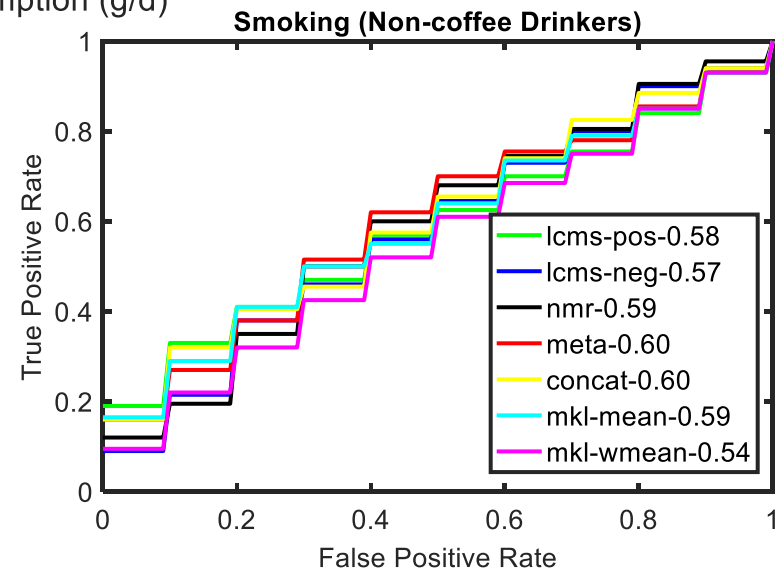
Revisiting: Joint analysis of measurements from multiple platforms has the potential to enhance biomarker discovery!

Metabolomics: The goal is to detect a wide range of chemical substances in biological fluids, e.g., blood, and to identify the chemicals related to certain conditions such as food intake and various diseases, e.g., cancer.
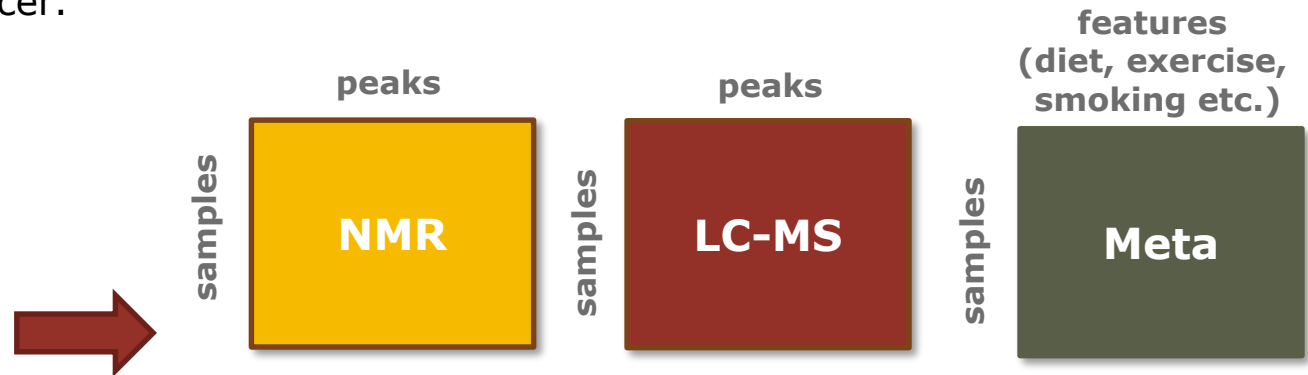
1. *Given these data sets, can we forecast whether people will have certain diseases in near future?*
Yes!

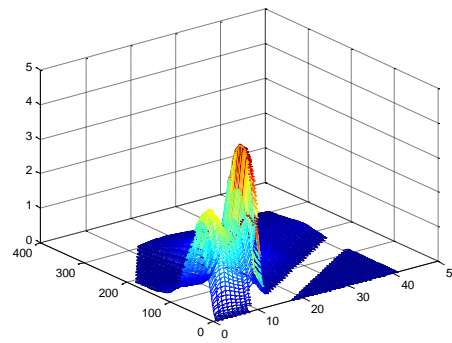2. *Does data fusion improve our forecasting performance?*
Depends on the disease!

3. *Can we capture biomarkers? Are we confident with the biomarkers?*
We can capture biomarkers but we should be aware of the confounding effects!

Motivation: Joint analysis of measurements from multiple platforms has the potential to enhance biomarker discovery!

EEM (Fluorescence Spectroscopy)

NMR (Nuclear Magnetic Resonance)

Possible to measure using different analytical methods

LC-MS (Liquid Chromatography-Mass Spectrometry)

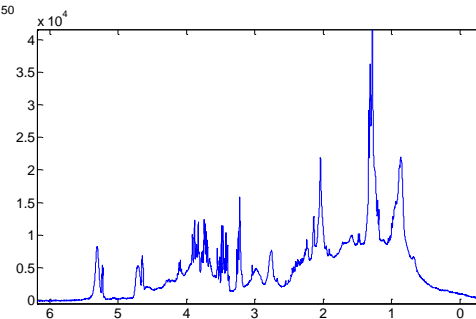# Motivation: Joint analysis of measurements from multiple platforms has the potential to enhance biomarker discovery!

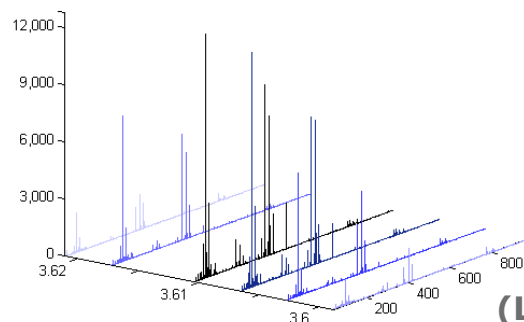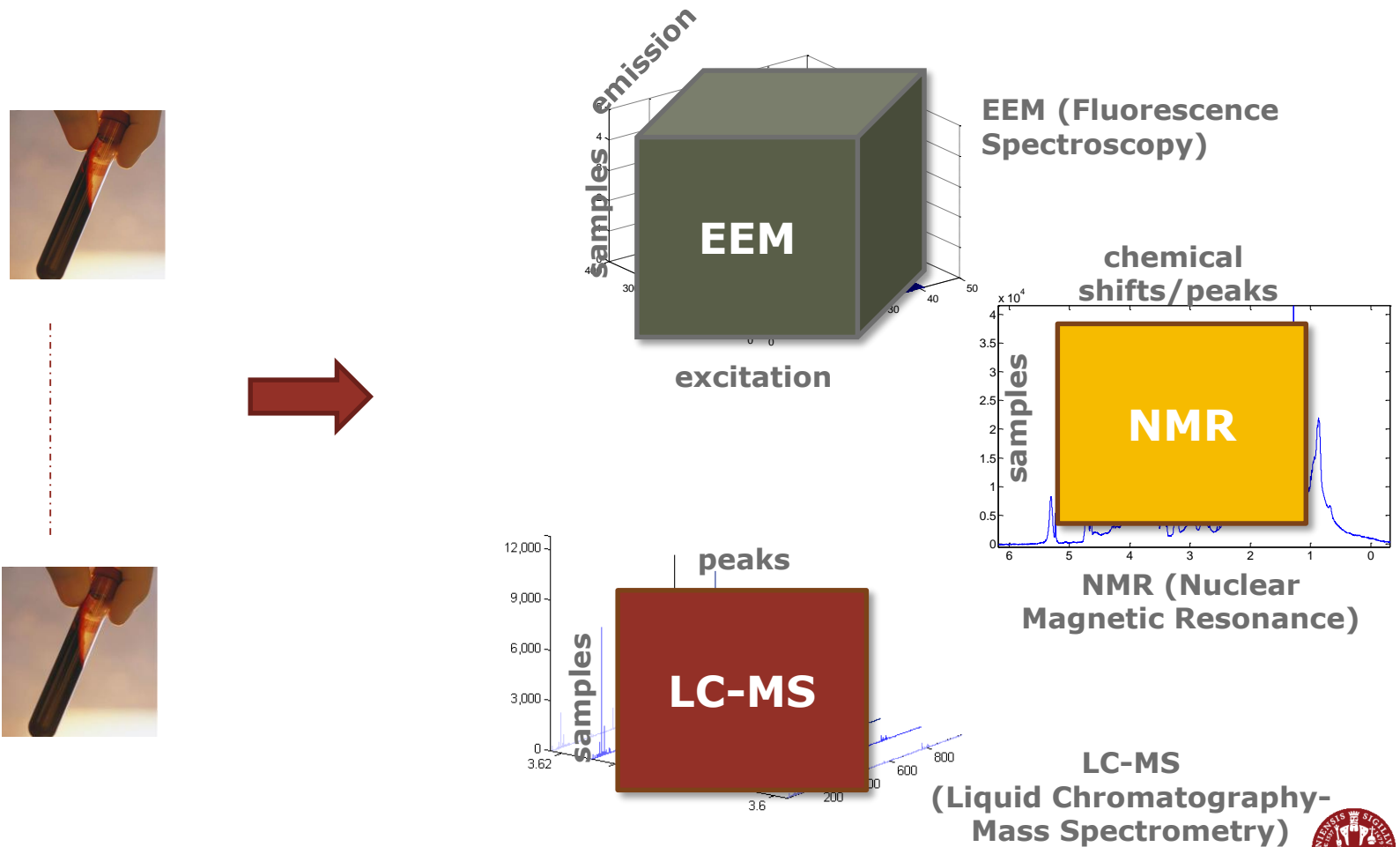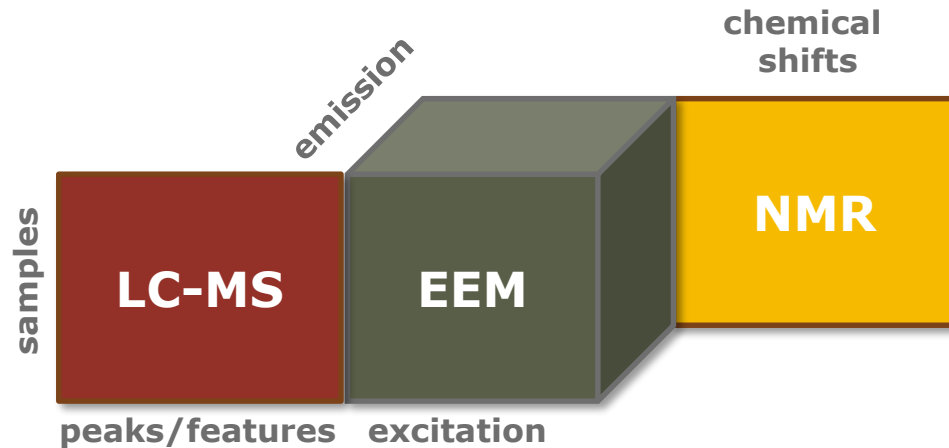# Motivation: Joint analysis of measurements from multiple platforms has the potential to enhance biomarker discovery!

How can we jointly analyze these coupled matrices and higher-order tensors to enhance knowledge discovery?

Unsupervised Data Fusion using Coupled Matrix and Tensor Factorizations (CMTF)

Joint analysis of heterogeneous data from multiple sources can be formulated as a coupled matrix and tensor factorization problem. In CMTF, higher-order tensors and matrices are simultaneously factorized by fitting a CP model to higher-order tensors and factorizing matrices in a coupled manner.

Tensor Factorization: CANDECOMP/PARAFAC(CP)

[Hitchcock, 1927; Harshman, 1970; Carroll & Chang, 1970]

Matrix Factorization:

$$\mathbf{Y} \approx \sum_{r=1}^{R} \mathbf{a}_r \mathbf{d}_r^{\mathsf{T}}$$

$$\approx \mathbf{A}\mathbf{D}^{\mathsf{T}}$$

$$\mathbf{A} \in \mathbb{R}^{I \times R} = \begin{bmatrix} \mathbf{a}_1 & \cdots & \mathbf{a}_R \end{bmatrix}$$

$$\mathbf{D} \in \mathbb{R}^{M \times R} = \begin{bmatrix} \mathbf{d}_1 & \cdots & \mathbf{d}_R \end{bmatrix}$$

$$\mathbf{Y} \approx \mathbf{A}\mathbf{D}^{\mathsf{T}} \quad \mathcal{X} \approx [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!]$$

$$\mathcal{X} \approx \sum_{r=1}^{R} \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r$$

$$\approx [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!]$$

$$\mathbf{A} \in \mathbb{R}^{I \times R} = \begin{bmatrix} \mathbf{a}_1 & \cdots & \mathbf{a}_R \end{bmatrix}$$

$$\mathbf{B} \in \mathbb{R}^{J \times R} = \begin{bmatrix} \mathbf{b}_1 & \cdots & \mathbf{b}_R \end{bmatrix}$$

$$\mathbf{C} \in \mathbb{R}^{K \times R} = \begin{bmatrix} \mathbf{c}_1 & \cdots & \mathbf{c}_R \end{bmatrix}$$

The problem can be formulated as:

$$\min_{\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}} \|\mathcal{X} - [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!]\|^2 + \|\mathbf{Y} - \mathbf{A}\mathbf{D}^{\mathsf{T}}\|^2$$

Data Fusion based on Coupled Tensor Factorizations
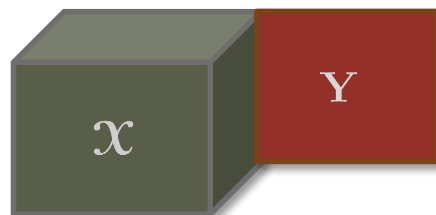
$$\min_{A,B,C,D} \| \mathcal{X} - [\![A, B, C]\!] \|^2 + \| Y - AD^\mathsf{T} \|^2$$

- **Psychometrics:** Linked-mode PARAFAC **[Harshman and Lundy, 1984]**
- **Chemometrics:** Multi-way Multi-block component models **[Smilde et al., 2000]**
- **Bioinformatics:** Coupled analysis of in vitro and histology tissue samples **[Acar et al., 2012]**
- **Signal Processing:** Joint analysis of a covariance matrix and a cumulant tensor **[De Lathauwer and Vandewalle, 2004; Comon, 2004]**; Generalized Coupled Tensor Factorizations **[Yilmaz et al., 2011]**; Structured Data Fusion **[Sorber et al., 2015]**
- **Data Mining:** Multi-way Clustering **[Banerjee et al., 2007]**; Community detection **[Lin et al., 2009]**; Missing value estimation **[Zheng et al., 2010]**; All-at-once optimization for CMTF **[Acar et al., 2011]**; Link prediction **[Ermis et al., 2012]**; Scalable CMTF approaches (sampling–based **[Papalexakis et al., 2014]**, distributed stochastic gradient running on MapReduce **[Beutel et al., 2014]**, distributed ALS running on MapReduce **[Jeon et al., 2016]**)

Models identifying shared/unshared factors:
- Structure-revealing data fusion models **[Acar et al., 2013; Acar et al., 2014; Acar et al., 2015]**
- Mining labelled tensors by discovering common and discriminative subspaces **[Lie et al., 2013]**
- Joint decompositions with flexible couplings **[Farias et al., 2015; Rivet et al., 2015]**

Our Approach: Structure-Revealing CMTF

[Acar et al., *BMC Bioinformatics*, 2014]

We reformulate the coupled matrix and tensor factorization problem by having factor matrices with unit norm columns and explicitly representing the weights of rank-one components in the formulation. Through modeling constraints/penalties, we let the model identify shared/unshared components.

$$\mathbf{Y} \approx \mathbf{A\Sigma D}^\mathsf{T} \qquad \mathfrak{X} \approx [\![\lambda; \mathbf{A}, \mathbf{B}, \mathbf{C}]\!]$$
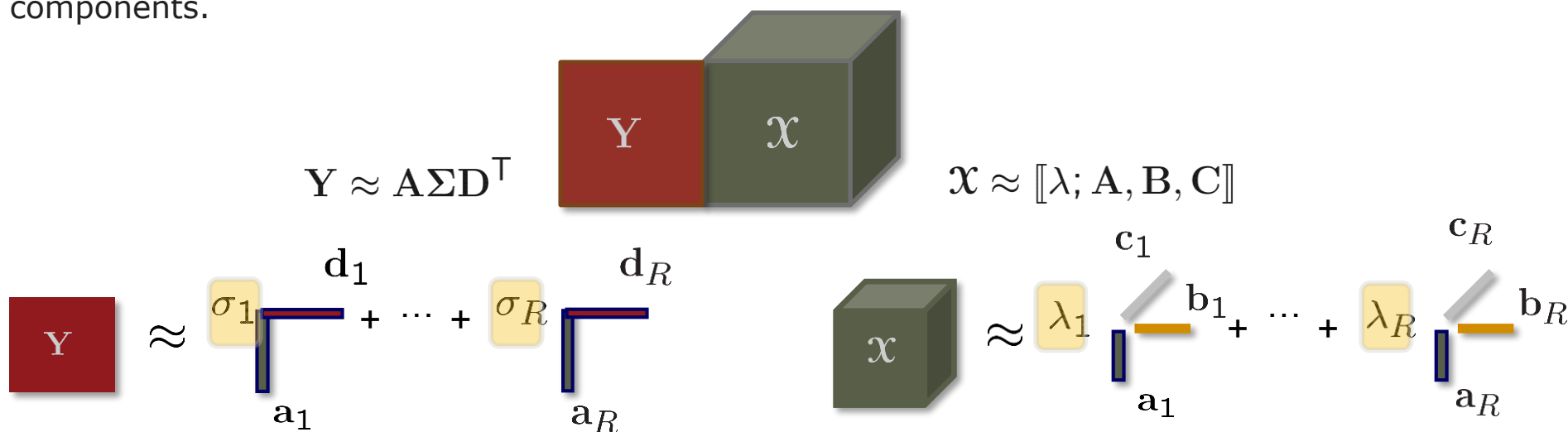
Structure-revealing model:

l1-norm
$\|\mathbf{x}\|_1 = \sum |x_i|$

$$\min_{\mathbf{A,B,C,D,\Sigma,\lambda}} \|\mathfrak{X} - [\![\lambda; \mathbf{A}, \mathbf{B}, \mathbf{C}]\!]\|^2 + \|\mathbf{Y} - \mathbf{A\Sigma D}^\mathsf{T}\|^2 + \beta\|\lambda\|_1 + \beta\|\sigma\|_1$$

$$\text{s.t.} \|\mathbf{a}_r\|_2 = \|\mathbf{b}_r\|_2 = \|\mathbf{c}_r\|_2 = \|\mathbf{d}_r\|_2 = 1, \text{for } r = 1, ..., R$$

Original CMTF
$\min\limits_{\mathbf{A,B,C,D}} \|\mathfrak{X} - [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!]\|^2 + \|\mathbf{Y} - \mathbf{AD}^\mathsf{T}\|^2$

Application: Cancer Metabolomics

[Acar et al., *IEEE EMBC*, 2013]

- Plasma samples measured using fluorescence spectroscopy and ^{1}H NMR.

- 119 samples (61 Female/58 Male)
  - **cancer:** 55 samples from verified colorectal cancer (CRC)
  - **control:** 64 samples with other nonmalignant findings

- **NMR** data is converted to a set of peaks.

- **Fluorescence** measurements: Samples measured with excitation wavelengths from 250 to 450 with 5 nm increment, and emission wavelengths from 300 to 600 with 1 nm increment.
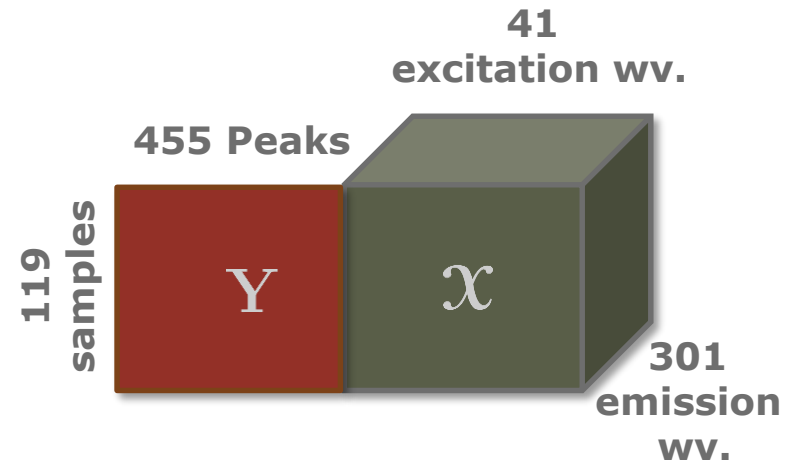
Goal: We want to jointly analyze these data sets and identify shared and unshared components, and if there are any components related to CRC.

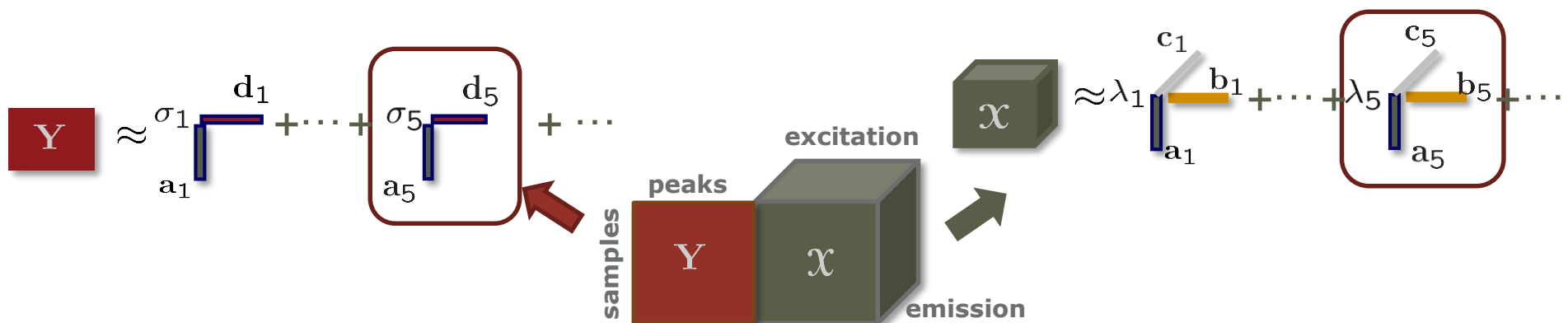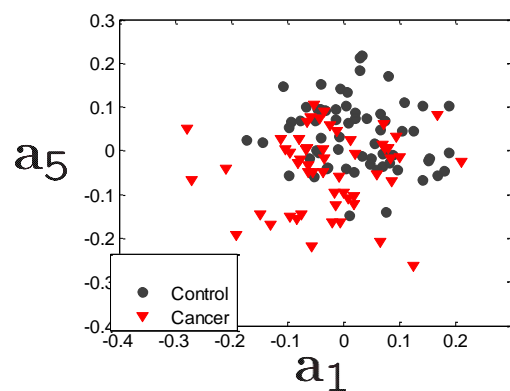# Structure-Revealing CMTF captures shared/unshared components!
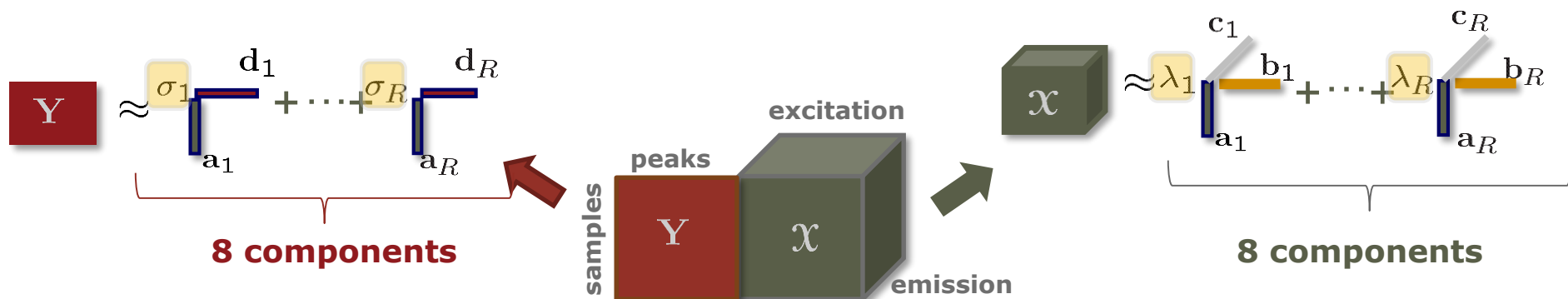
σ_r	λ_r
0.14	0.82
0.20	0.72
0.35	0.30
0.14	0.39
0.42	0.10
0.59	0.06
0.46	0.09
0.39	0.00

Only in NMR

Cancer-related component:

$$Y \approx {}^{\sigma_1} \Big\lfloor \mathbf{d}_1 \ + \cdots + \ {}^{\sigma_5} \Big\lfloor \mathbf{d}_5 \ + \cdots$$

$$\mathcal{X} \approx \lambda_1 \ \mathbf{c}_1 \ \mathbf{b}_1 \ + \cdots + \lambda_5 \ \mathbf{c}_5 \ \mathbf{b}_5 \ + \cdots$$

σ_r	λ_r
0.14	0.82
0.20	0.72
0.35	0.30
0.14	0.39
0.42	0.10
0.59	0.06
0.46	0.09
0.39	0.00

Accuracy: 71%
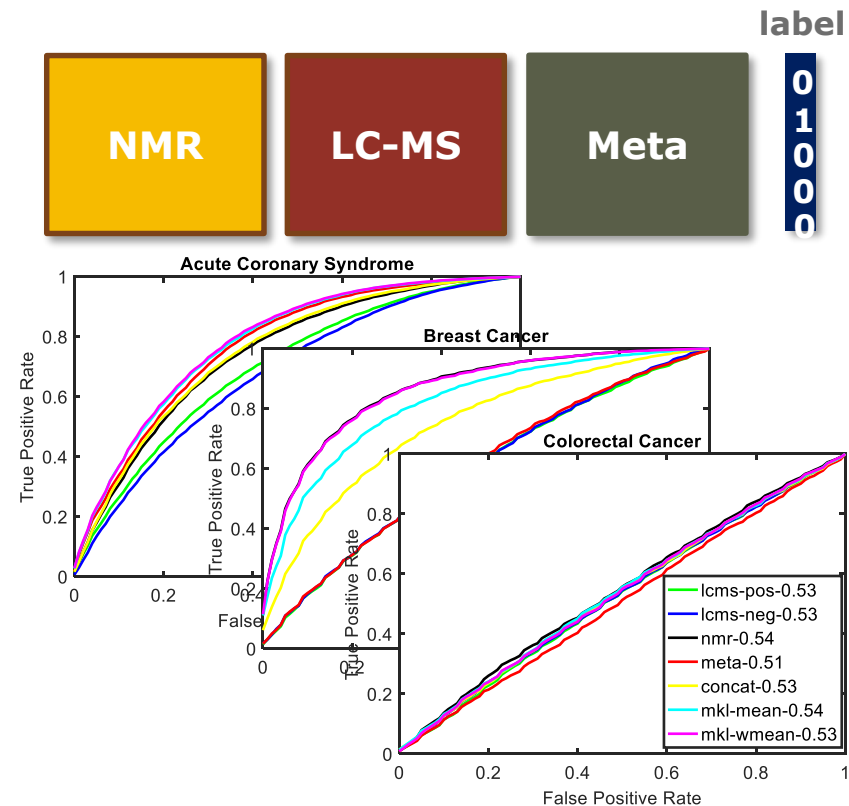**Sensitivity: 64%**
Specificity: 78%
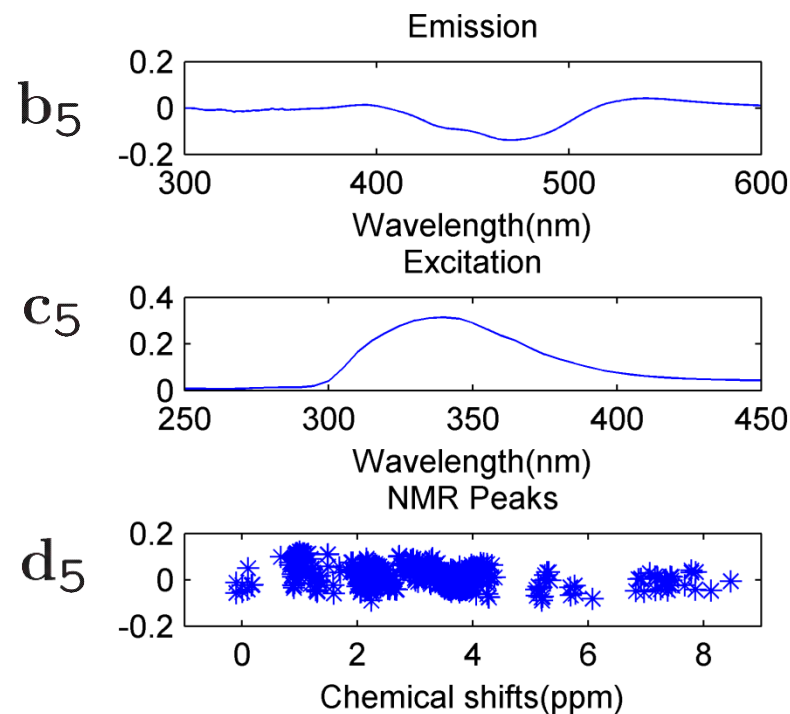
Summary

Goal: To forecast Acute Coronary Syndrome, Breast Cancer and Colorectal Cancer using measurements from multiple platforms and the meta data.

**Approach:** Supervised data fusion using multiple kernel learning

- Using linear kernels (no improvements in performance yet using nonlinear kernels)
- Identifying significant features

Results:

- ACS: Data fusion improves the forecasting performance.
- Breast Cancer: Fusion degrades the forecasting performance.
- CRC: Neither the individual data sets nor their fusion can forecast CRC cases.
- Validation of the models by capturing known biomarkers (e.g., for coffee)
- Limitations of the models due to the confounding effects (e.g., for smoking)

label

NMR LC-MS Meta

0 1 0 0 0 0

Acute Coronary Syndrome

Breast Cancer

Colorectal Cancer

lcms-pos-0.53
lcms-neg-0.53
nmr-0.54
meta-0.51
concat-0.53
mkl-mean-0.54
mkl-wmean-0.53

Joint analysis of data sets in the form of matrices and higher-order tensors in cancer metabolomics

LC-MS FFM NMR
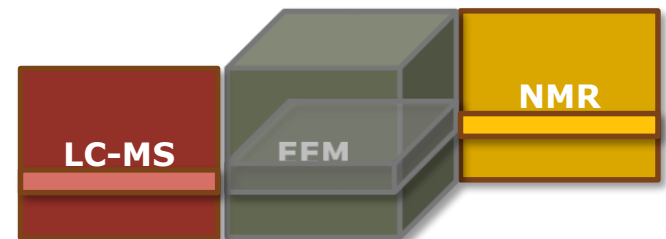
Thank you!

Structure-revealing data fusion model

E. Acar, R. Bro, and A. K. Smilde, Data Fusion in Metabolomics using Coupled Matrix and Tensor Factorizations, *Proceedings of the IEEE*, 103: 1602-1620, 2015.

E. Acar, E. E. Papalexakis, G. Gurdeniz, M. A. Rasmussen, A. J. Lawaetz, M. Nilsson, and R. Bro, Structure Revealing Data Fusion, *BMC Bioinformatics*, 15: 239, 2014.

E. Acar, M. Nilsson, and M. Saunders, A Flexible Modeling Framework for Coupled Matrix and Tensor Factorizations, *EUSIPCO,* pp. 111-115, 2014.

E. Acar, A. J. Lawaetz, M. A. Rasmussen, and R. Bro, Structure Revealing Data Fusion Model with Applications in Metabolomics, *IEEE EMBC,* pp. 6023 - 6026, 2013.

Evrim Acar
http://www.models.life.ku.dk/~acare/

JODA: Joint Data Analysis for Enhanced Knowledge Discovery
http://www.models.life.ku.dk/joda

Danish Agency for Science
Technology and Innovation
Ministry of Science, Innovation
and Higher Education

Sapere Aude