Interpretable classification of molecular measurements via pathway-induced multiple kernel learning

—

Joris Cadow
Data Scientist

NetBio COSI @ ISMB ECCB 2019, Basel / July 23rd, 2019

IBM **Research**

Roadmap

Molecular data classification

Pathway-Induced Multiple Kernel Learning (PIMKL)

PIMKL benchmarking

PIMKL application

Roadmap

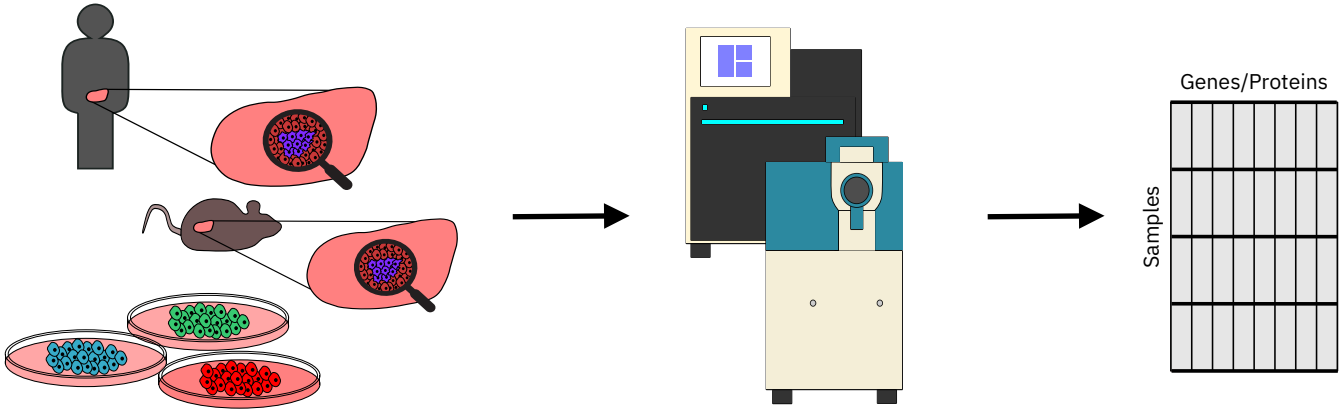
Molecular data classification

Pathway-Induced Multiple Kernel Learning (PIMKL)

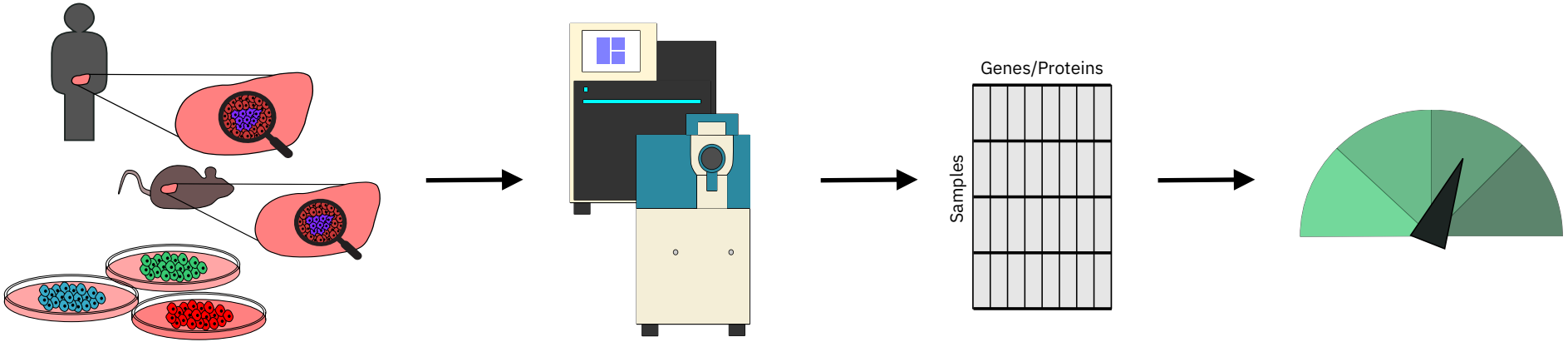
PIMKL benchmarking

PIMKL application

Molecular data classification



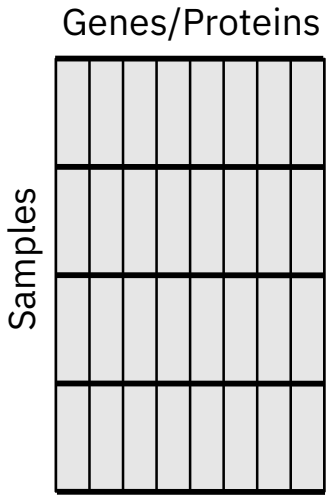
Molecular data classification



Molecular data classification - challenges

Experiments costs and noise

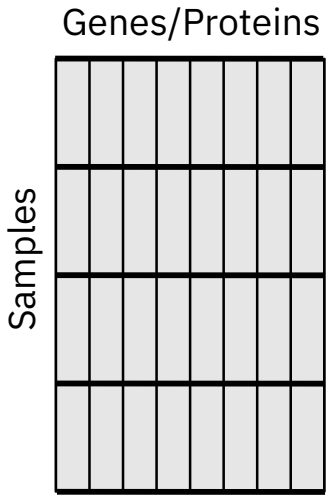
Scarce sample availability in high throughput experiments



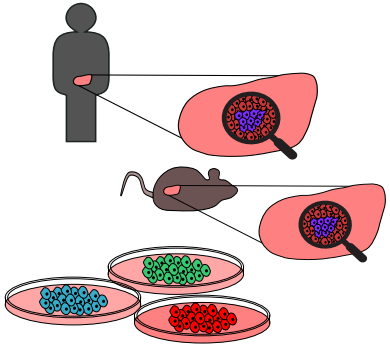
Molecular data classification - challenges

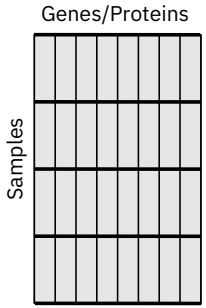
Experiments costs and noise

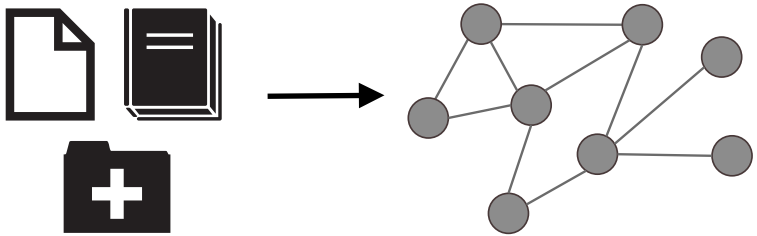
Scarce sample availability in high throughput experiments



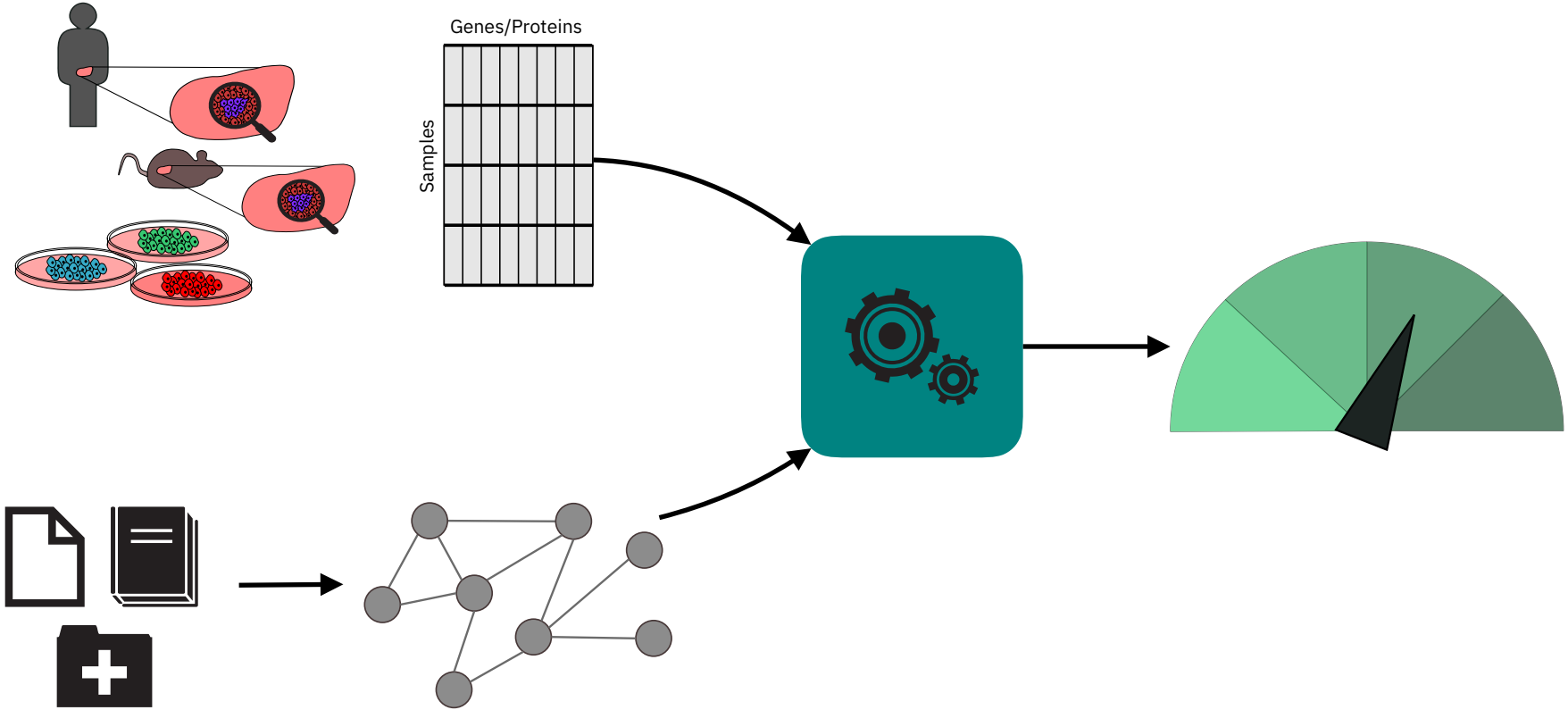
Molecular data classification - exploit prior knowledge



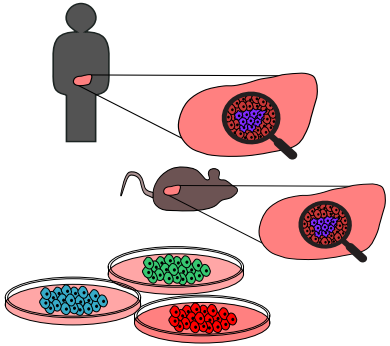


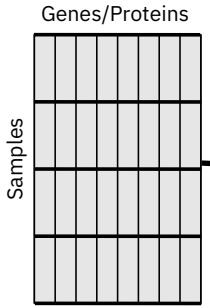


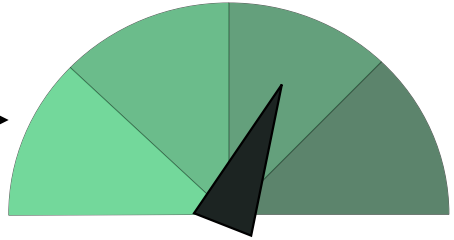
Molecular data classification - exploit prior knowledge

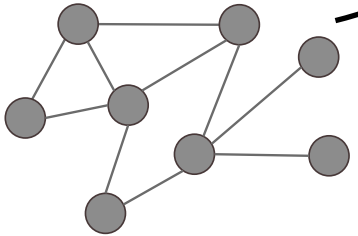


Molecular data classification - exploit prior knowledge









Use network topology as a regulariser

Define meta-features using pathway information

L. Chen et al., *BMC Systems Biology*, 2011

Z. Guo, *BMC Bioinformatics*, 2005

Roadmap

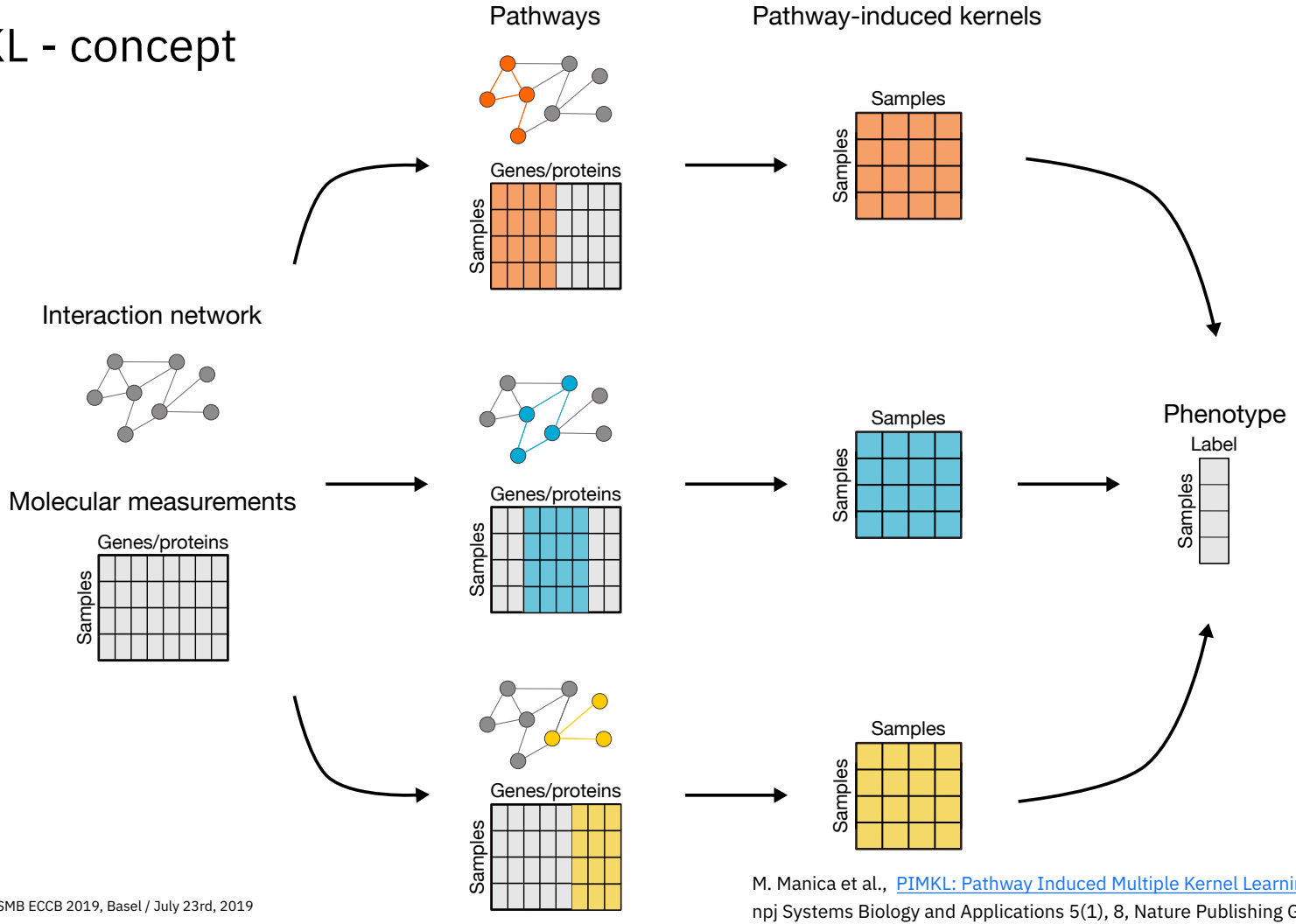
Molecular data classification

Pathway-Induced Multiple Kernel Learning (PIMKL)

PIMKL benchmarking

PIMKL application

PIMKL - concept

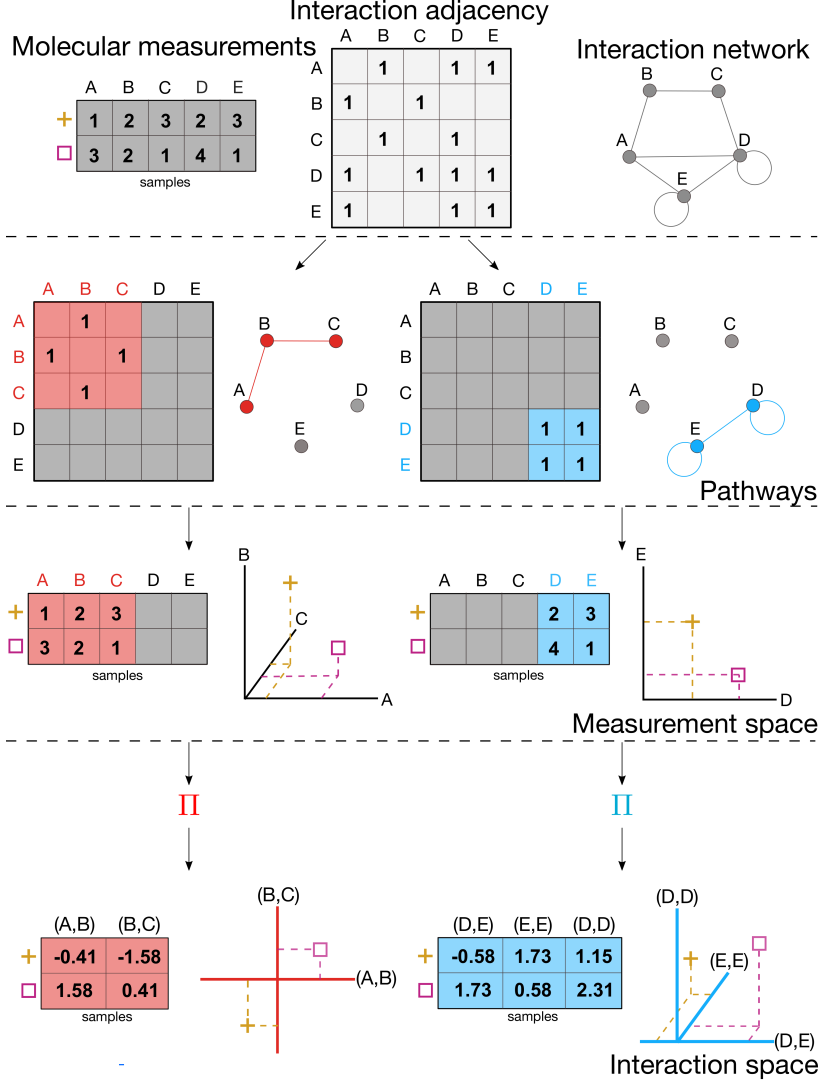


M. Manica et al., PIMKL: Pathway Induced Multiple Kernel Learning, npj Systems Biology and Applications 5(1), 8, Nature Publishing Group, 2019

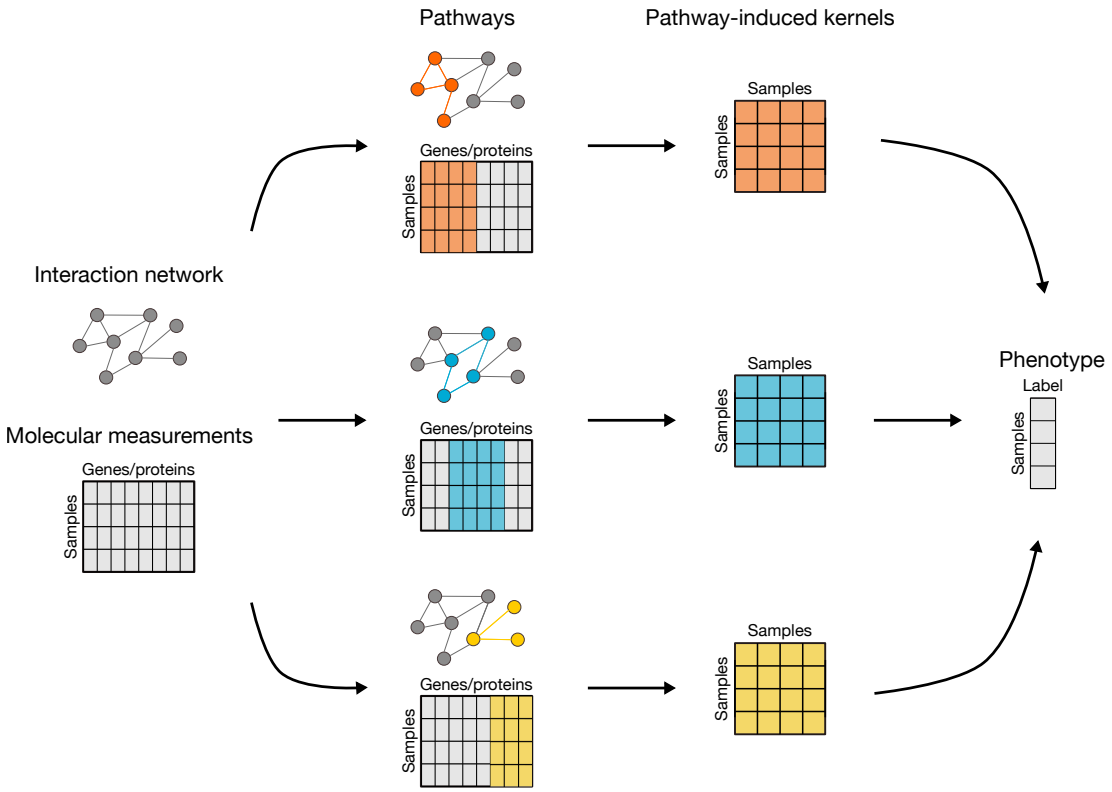
PIMKL - pathway induction

$$\begin{cases} k_{L_\omega}(x,y) = x^T L_\omega y = x^T S_\omega S_\omega^T y = \Pi(x)^T \Pi(y) \\ S_\omega = D^{-\frac{1}{2}} S W^{\frac{1}{2}} \end{cases}$$

Pathway induction map the data from the gene space to the interaction space



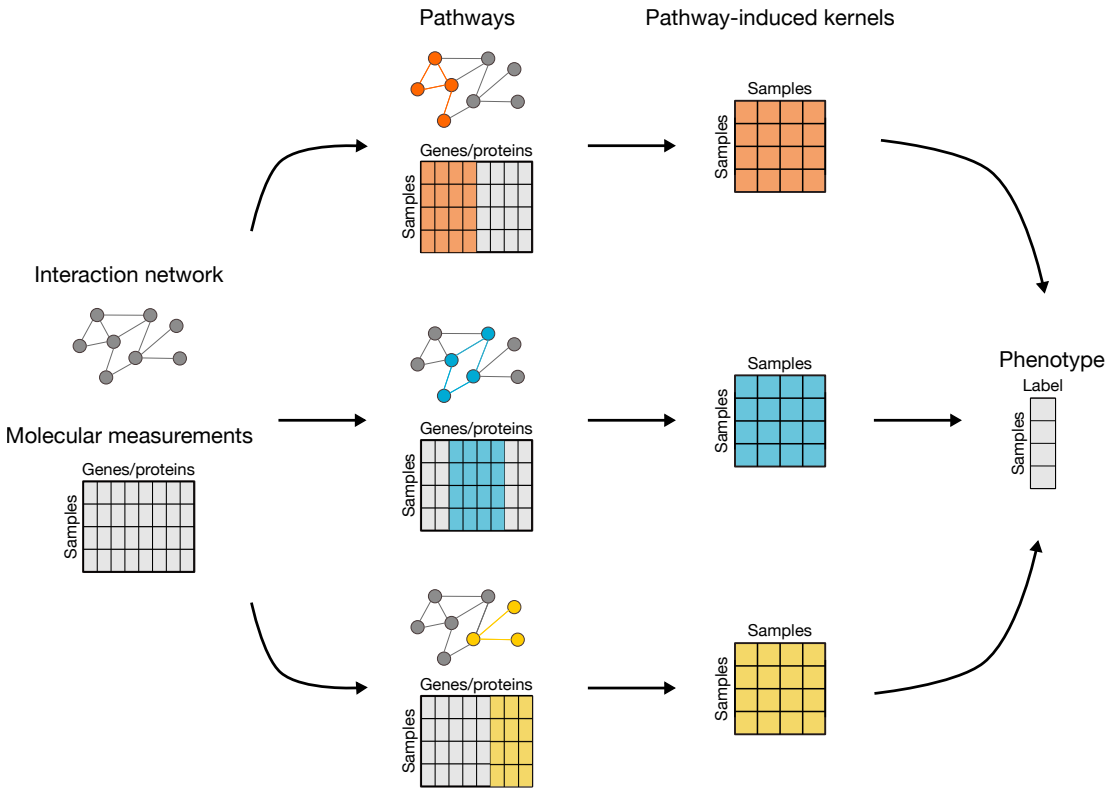
PIMKL - multiple kernel learning



Weighted combination of pathway-induced kernels to optimize phenotype prediction

$$\begin{cases} K^p_{ij} = k_{L^p_\omega}(x_i, x_j) \\ K = \sum_{p=1}^{P} w_p K^p, \quad w_p \geq 0 \end{cases}$$

PIMKL - multiple kernel learning



Weighted combination of pathway-induced kernels to optimize phenotype prediction

$$\begin{cases} K^p_{ij} = k_{L^p_\omega}(x_i, x_j) \\ K = \sum_{p=1}^{P} w_p K^p, \quad w_p \geq 0 \end{cases}$$

Included using the kernel trick in an iterative optimization procedure

$$\max_{\alpha_i \geq 0,\, w_p \geq 0} \quad \sum_i^N \alpha_i - \frac{1}{2} \sum_p^P \sum_i^N \sum_j^N \alpha_i \alpha_j y_i y_j w_p K^p_{ij}$$

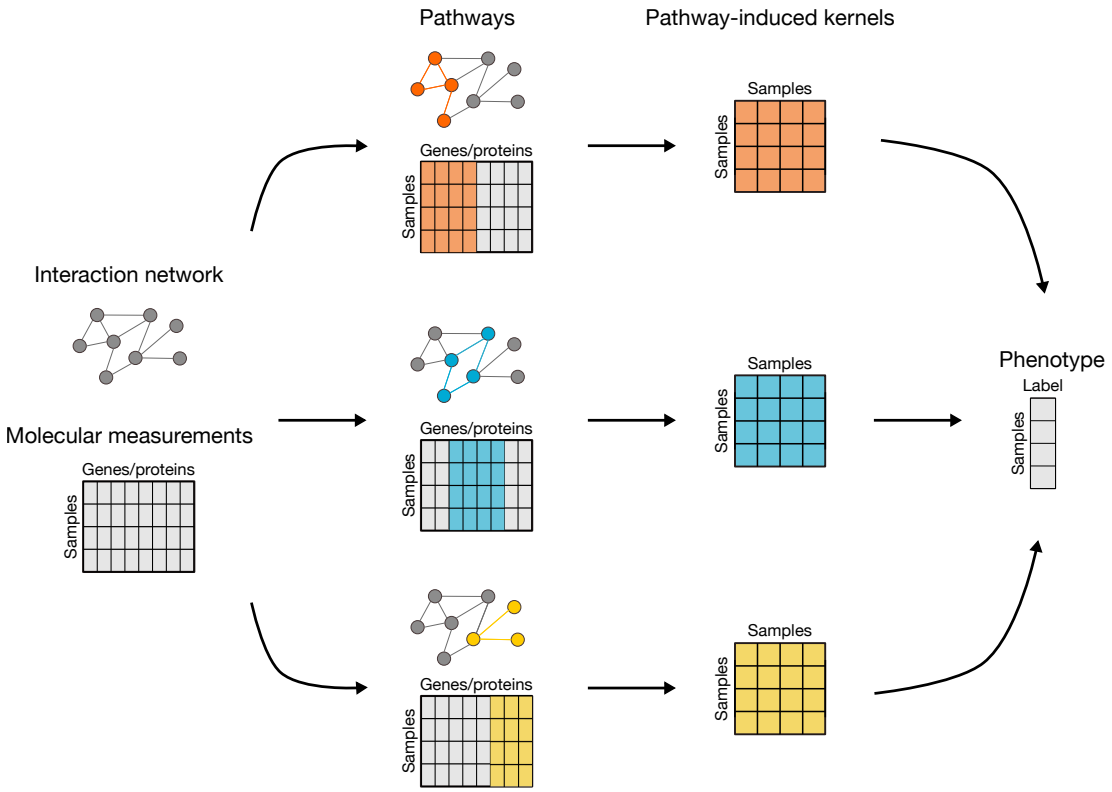
subject to

$$0 \leq \alpha_i \leq C$$

$$0 \leq w_p \leq C'$$

$$\sum_i^N \alpha_i y_i = 0$$

PIMKL - multiple kernel learning



Weighted combination of pathway-induced kernels to optimize phenotype prediction

$$\begin{cases} K^p_{ij} = k_{L^p_\omega}(x_i, x_j) \\ K = \sum_{p=1}^{P} w_p K^p, \quad w_p \geq 0 \end{cases}$$

By using EasyMKL, a scalable multiple kernel learning algorithm

F. Aiolli et al. *Neurocomputing*, 2015

Roadmap

Molecular data classification

Pathway-Induced Multiple Kernel Learning (PIMKL)

PIMKL benchmarking

PIMKL application

PIMKL benchmarking

Benchmark against other prior knowledge informed methods on multiple breast cancer cohorts

Breast cancer Affymetrix Human Genome U133A Array cohorts.

GEOid	Patients	dmfs/rfs <= 5 years	dmfs/rfs > 5 years
GSE2034	286	93	183
GSE1456	159	34	119
GSE2990	187	42	116
GSE4922	249	69	159
GSE7390	198	56	135
GSE11121	200	28	154

Y. Cun et al. *BMC bioinformatics*, 2012

A. Liberzon et al. *Cell systems*, 2015

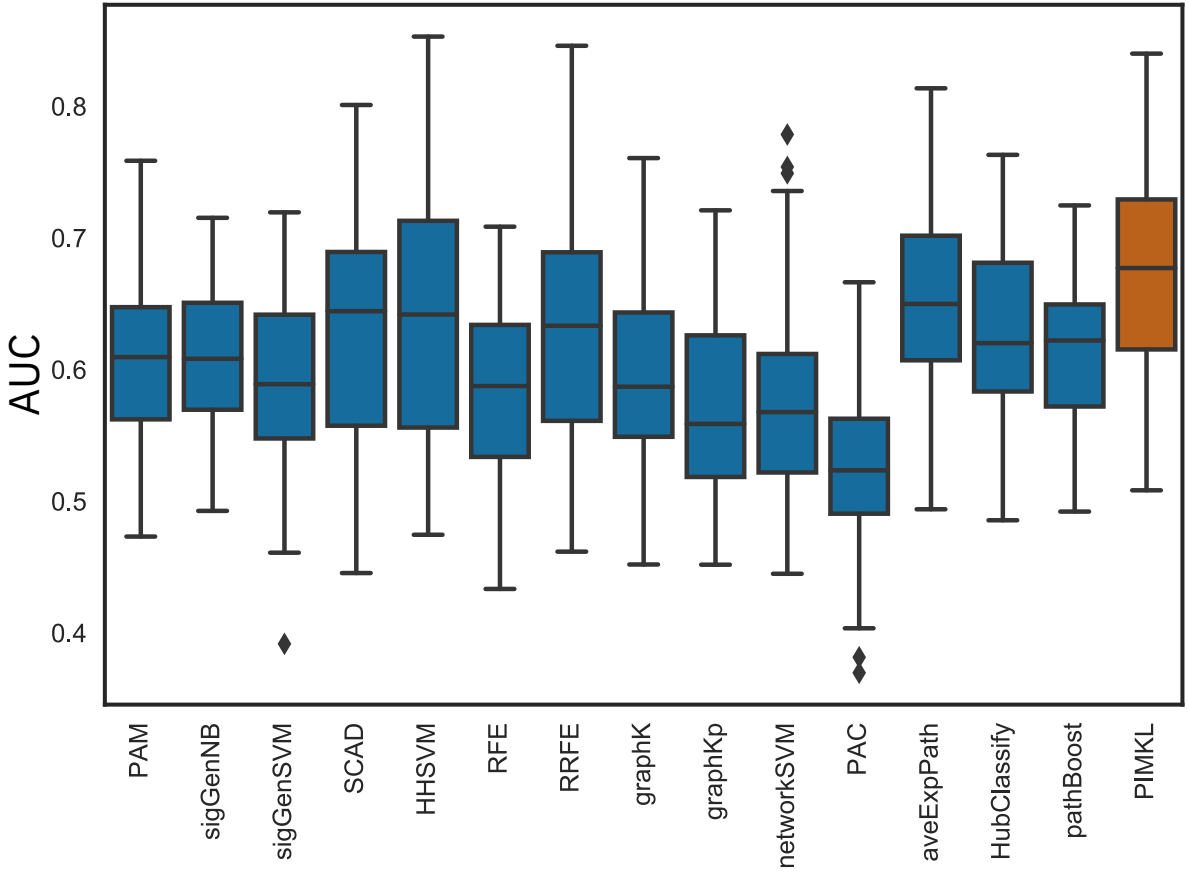
M. Kanehisa et al. *Nucleic Acids Research*, 2016

E. G. Cerami et al. *Nucleic Acids Research*, 2011

PIMKL benchmarking

Benchmark against other prior knowledge informed methods on multiple breast cancer cohorts

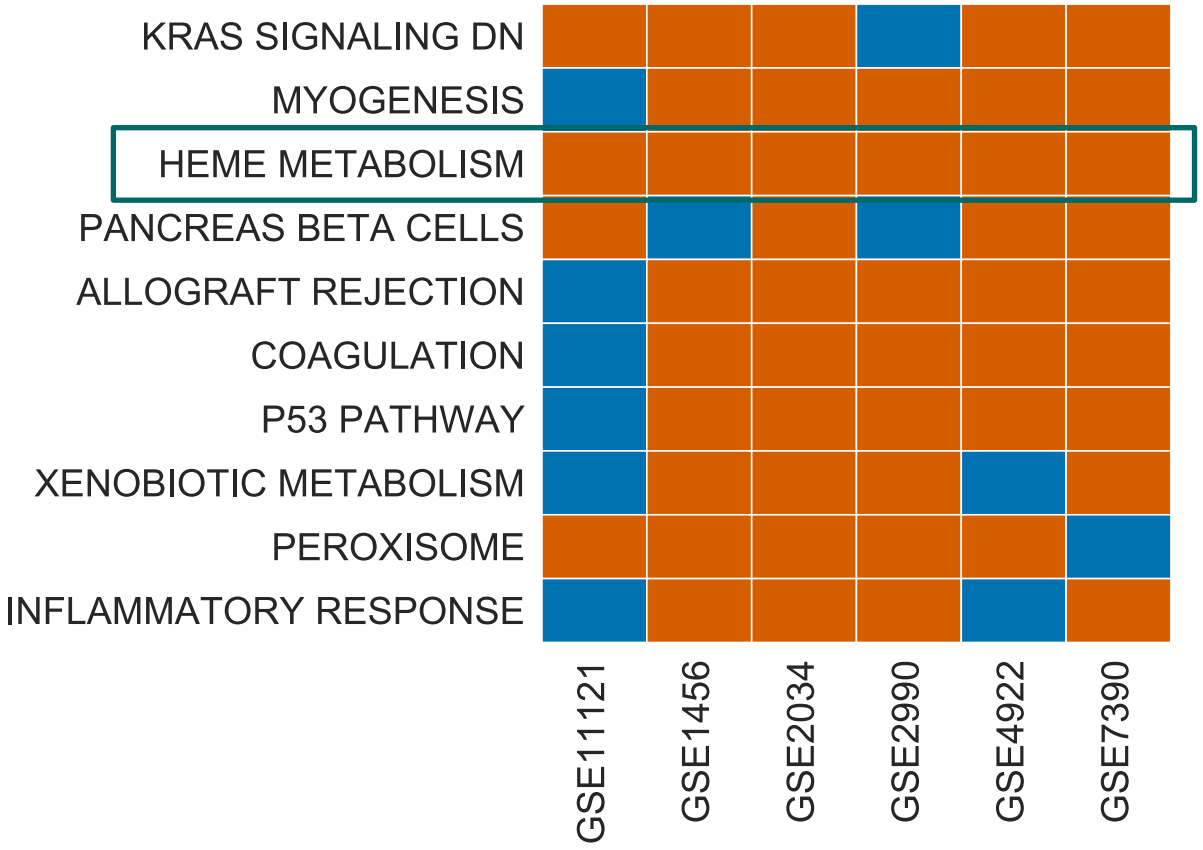
PIMKL significantly improves prediction of tumor relapse



Y. Cun et al. *BMC bioinformatics*, 2012
A. Liberzon et al. *Cell systems*, 2015
M. Kanehisa et al. *Nucleic Acids Research*, 2016
E. G. Cerami et al. *Nucleic Acids Research*, 2011

PIMKL benchmarking

PIMKL detects stable signatures across cohorts



Y. Cun et al. *BMC bioinformatics*, 2012
A. Liberzon et al. *Cell systems*, 2015
M. Kanehisa et al. *Nucleic Acids Research*, 2016
E. G. Cerami et al. *Nucleic Acids Research*, 2011

Roadmap

Molecular data classification

Pathway-Induced Multiple Kernel Learning (PIMKL)

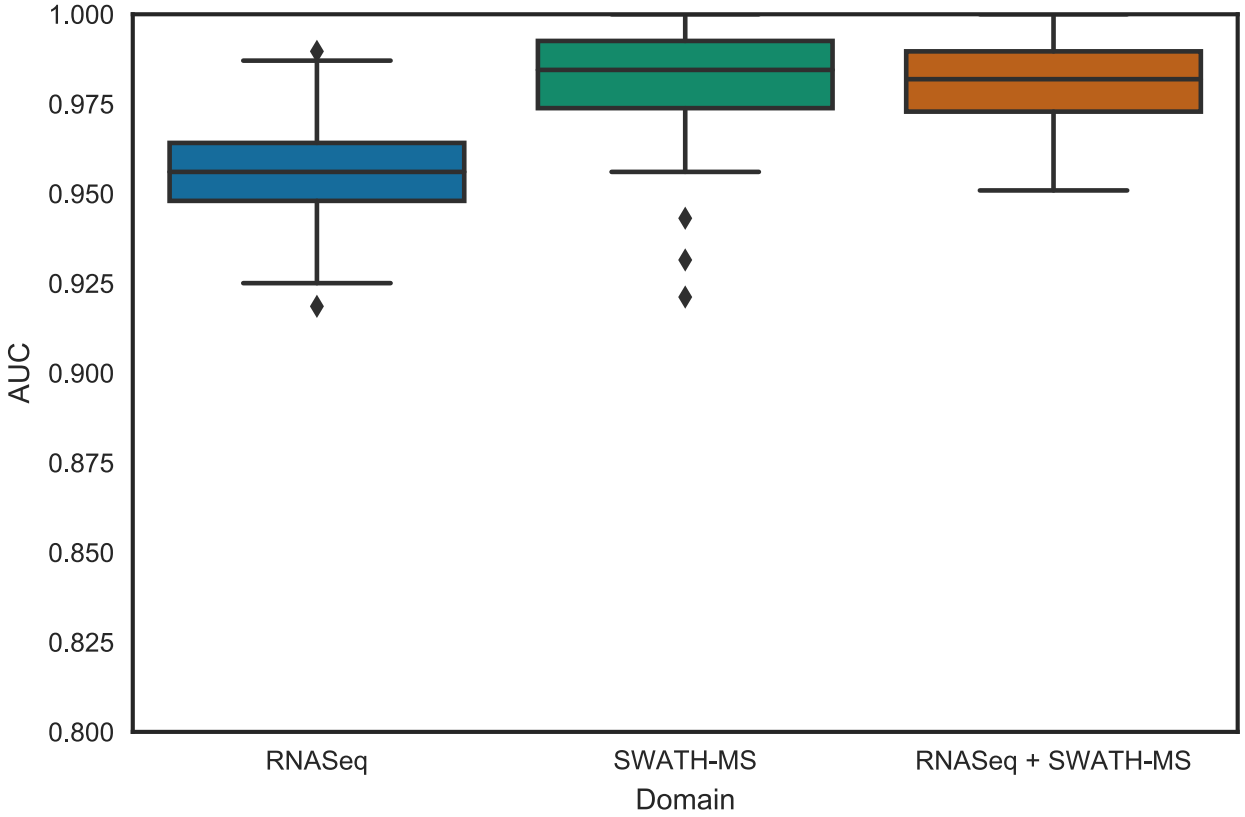
PIMKL benchmarking

PIMKL application

PIMKL application - prostate cancer

PIMKL detects tumor samples accurately

PIMKL integrates multiple omics seamlessly

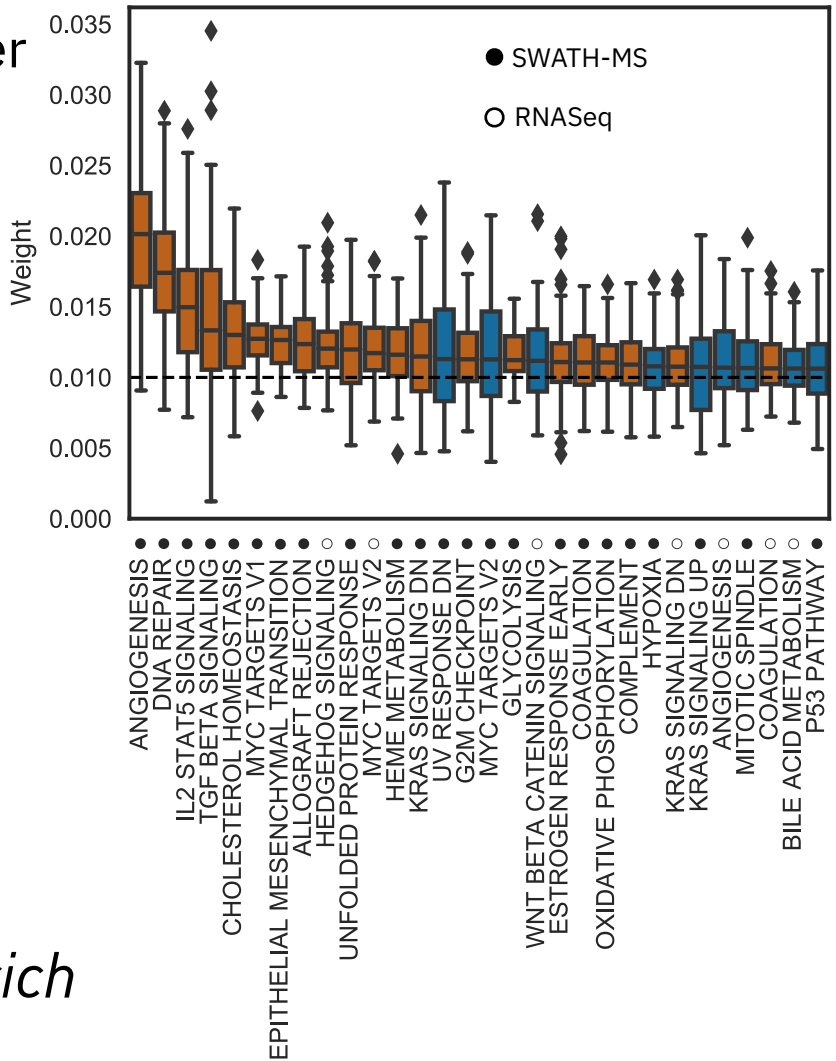


T. Guo et al., *in preparation*, 2019

UniversitätsSpital Zürich

ETH zürich

PIMKL application - prostate cancer



PIMKL identifies relevant pathways for each data type

T. Guo et al., *in preparation*, 2019

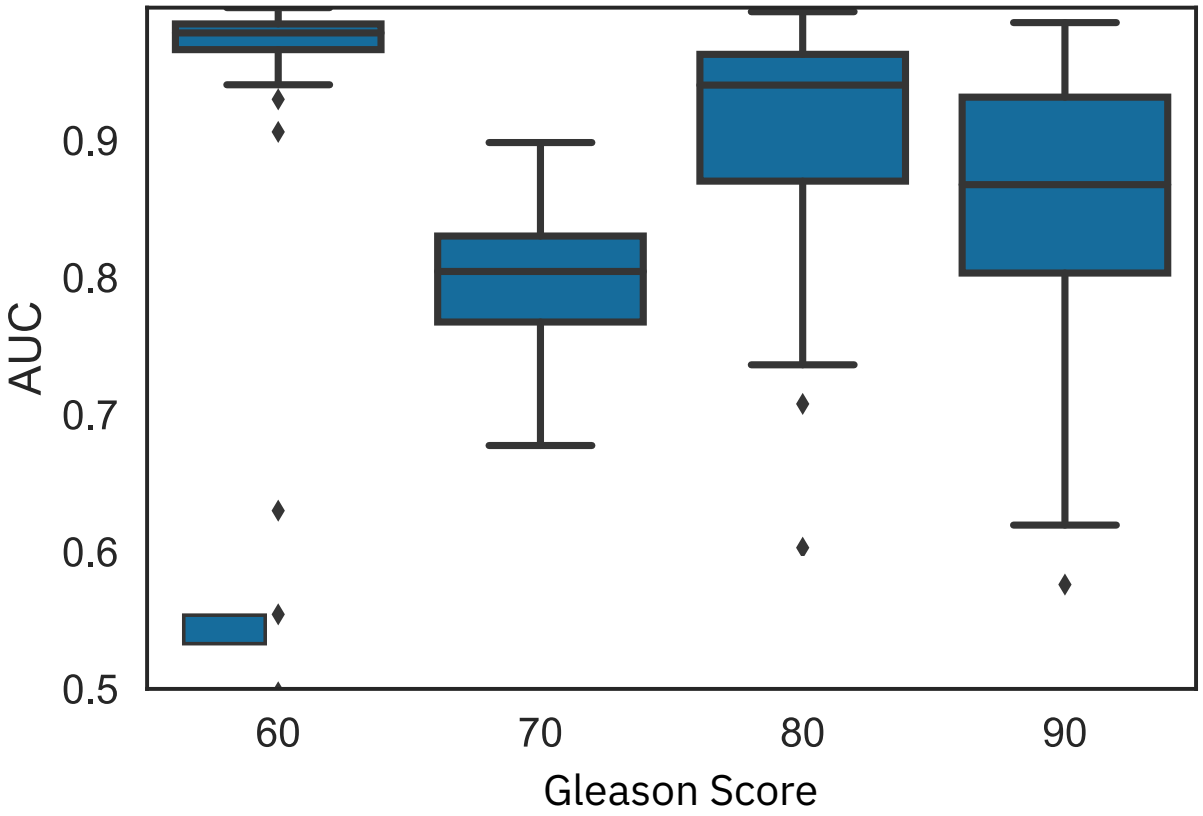
UniversitätsSpital Zürich

ETH zürich

PIMKL application - prostate cancer

PIMKL stratifies different disease grades

PIMKL reaches high performance by training on 30 samples per class

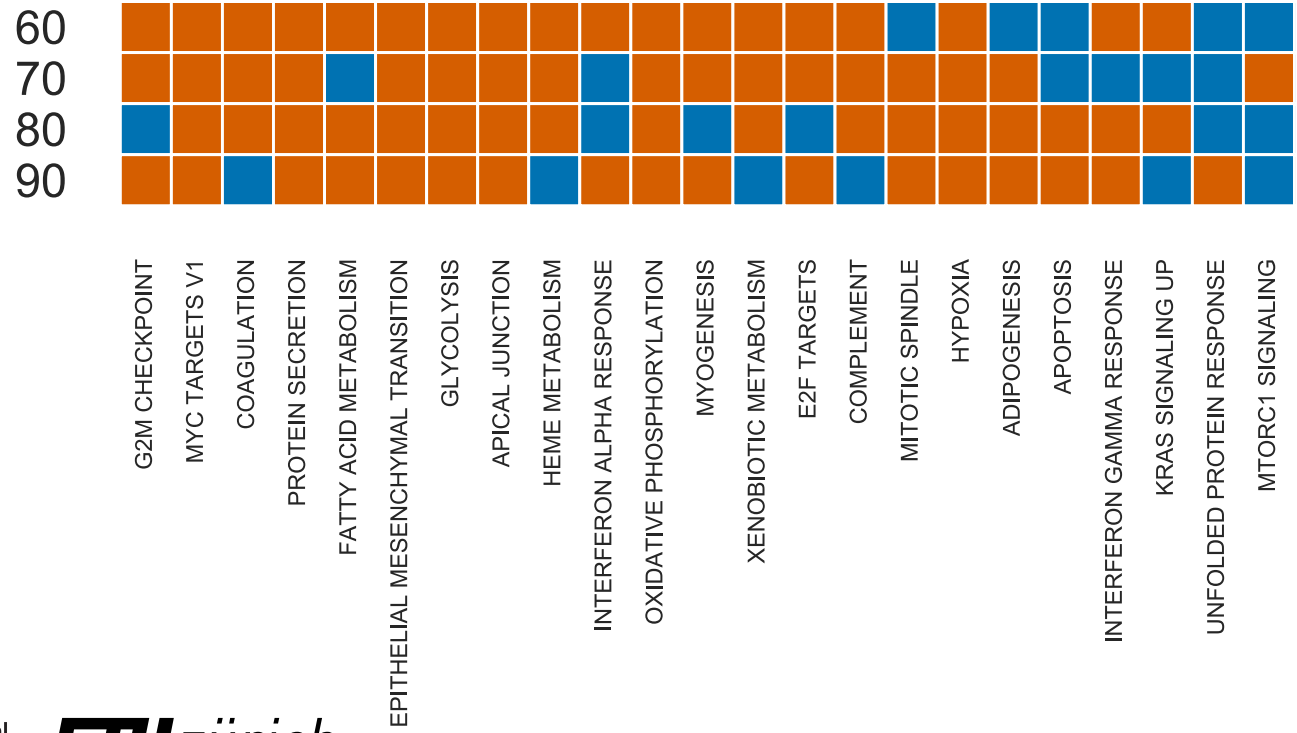


T. Guo et al., *in preparation*, 2019

UniversitätsSpital Zürich

ETH zürich

PIMKL application - prostate cancer



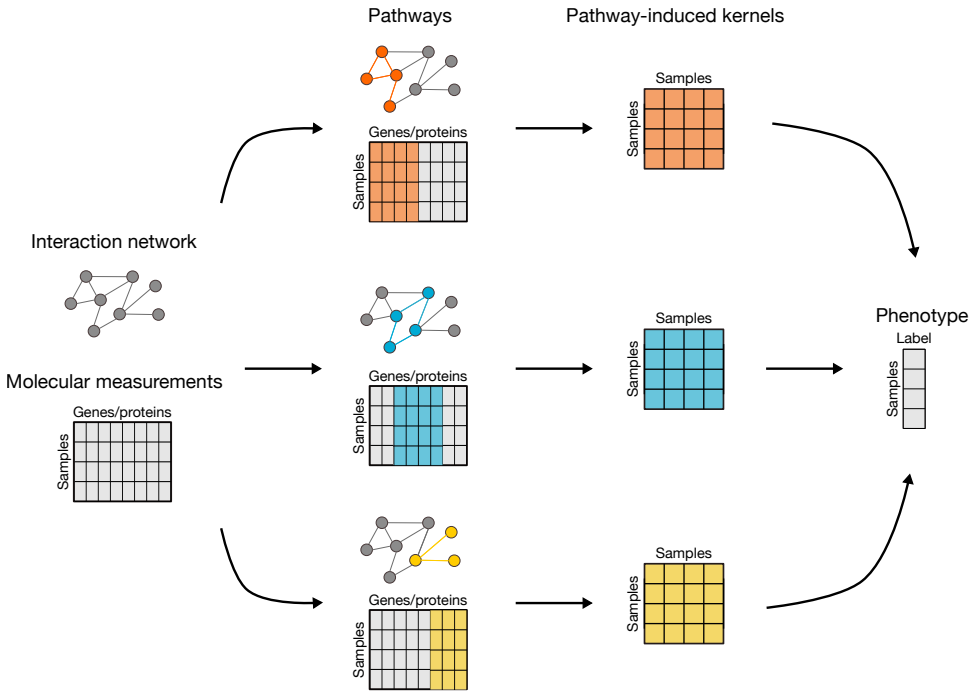
PIMKL helps to find the differences in the active pathways for each grade

T. Guo et al., *in preparation*, 2019

UniversitätsSpital Zürich

ETH zürich

PIMKL - pathway-induced multiple kernel learning



Available on IBM Cloud

Open source library

https://github.com/IBM/mimkl

https://ibm.biz/pimkl-aas

Thanks for your attention

Find me at the poster M-44
or at our booth

This work is part of the PrECISE project. PrECISE combines hypothesis-driven strategies with data-driven analysis in a novel mathematical and computational methodology for the integration of genomic, epigenetic, transcriptomic, proteomic, and clinical data with the goal of risk-stratifying patients and suggesting personalized therapeutic interventions. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 668858.
Project website: www.precise-project.eu

PrECISE

ibm.biz/ibmpimkl

IBM **Research**

1. Chen, L., Xuan, J., Riggins, R. B., Clarke, R. & Wang, Y. Identifying cancer bio-markers by network-constrained support vector machines.
BMC Syst. Biol. 5, 161–181 (2011). https://doi.org/10.1186/1752-0509-5-161.

2. Guo, Z. et al. Towards precise classification of cancers based on robust gene functional expression profiles.
BMC Bioinformatics 6, 58 (2005).

3. Aiolli, F. & Donini, M. EasyMKL: a scalable multiple kernel learning algorithm.
Neurocomputing 169, 215–224 (2015).

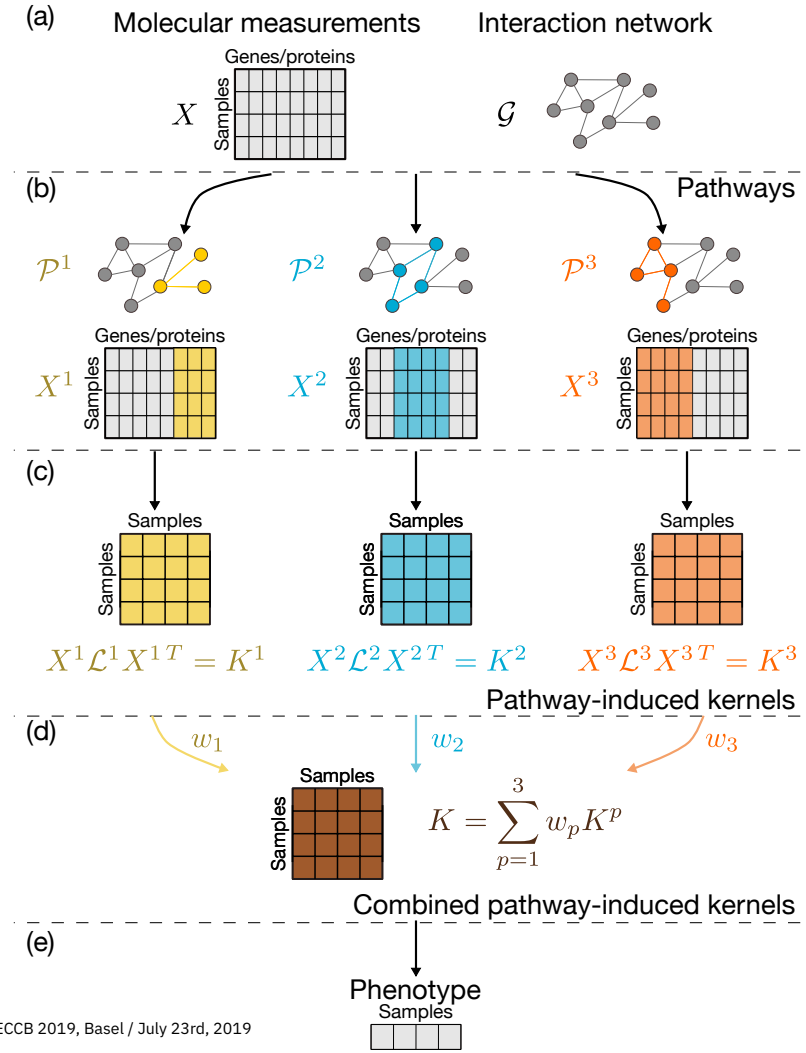
4. Cun, Y. & Fröhlich, H. Prognostic gene signatures for patient stratification in breast cancer-accuracy, stability and interpretability of gene selection approaches using prior knowledge.
BMC Bioinformatics 13, 69 (2012).

5. Liberzon, A. et al. The molecular signatures database hallmark gene set collection.
Cell Syst. 1, 417–425 (2015).

6. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes.
Nucleic Acids Res. 28, 27–30 (2000).

7. Cerami, E. G. et al. Pathway Commons, a web resource for biological pathway data.
Nucleic Acids Res. 39, D685–690 (2011).

8. Guo T. et al.
in preparation, 2019



Weighted combination of pathway-induced kernels to optimize phenotype prediction

$$\begin{cases} K^p_{ij} = k_{L^p_\omega}(x_i, x_j) \\ K = \sum_{p=1}^{P} w_p K^p, \quad w_p \geq 0 \end{cases}$$

By using EasyMKL, a scalable multiple kernel learning algorithm

F. Aiolli et al. *Neurocomputing*, 2015

PIMKL benchmarking

Benchmark against other prior knowledge informed methods on multiple breast cancer cohorts

Breast cancer Affymetrix Human Genome U133A Array cohorts.

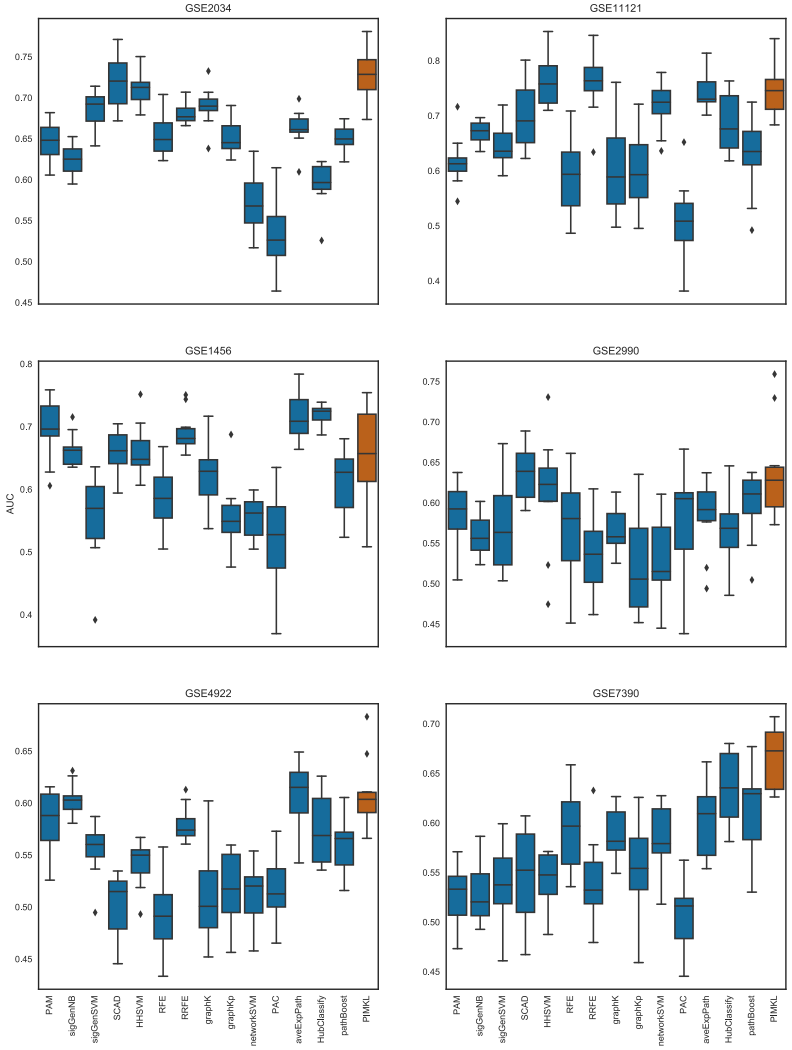
GEOid	Patients	dmfs/rfs <= 5 years	dmfs/rfs > 5 years
GSE2034	286	93	183
GSE1456	159	34	119
GSE2990	187	42	116
GSE4922	249	69	159
GSE7390	198	56	135
GSE11121	200	28	154

Y. Cun et al. *BMC bioinformatics*, 2012

A. Liberzon et al. *Cell systems*, 2015

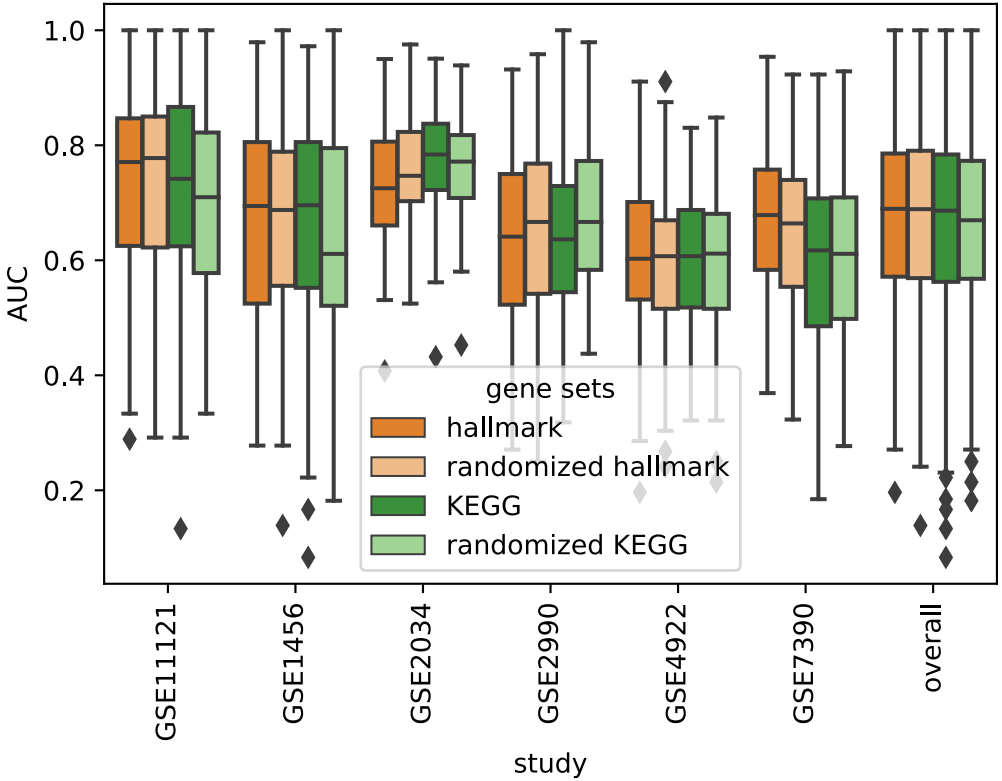
M. Kanehisa et al. *Nucleic Acids Research*, 2016

E. G. Cerami et al. *Nucleic Acids Research*, 2011



PIMKL randomised gene sets

PIMKL is stable in regard to gene set selection



Y. Cun et al. *BMC bioinformatics*, 2012
A. Liberzon et al. *Cell systems*, 2015
M. Kanehisa et al. *Nucleic Acids Research*, 2016
E. G. Cerami et al. *Nucleic Acids Research*, 2011