DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

ARTIFICIAL INTELLIGENCE AND LIFE SCIENCES:
Applications in Biomedicine and Life Sciences

DIFFERENT TYPES OF APPLICATIONS OF MACHINE LEARNING IN MEDICINE

- Medical diagnosis
- Medical prognosis

- Finding biomarkers for different diseases (e.g., brain tumors in children)
- Finding disease signatures and biomarkers for them

- Drug design/repurposing
- Advising when to change therapy
- Identifying the source of tumors metasthasized to the liver

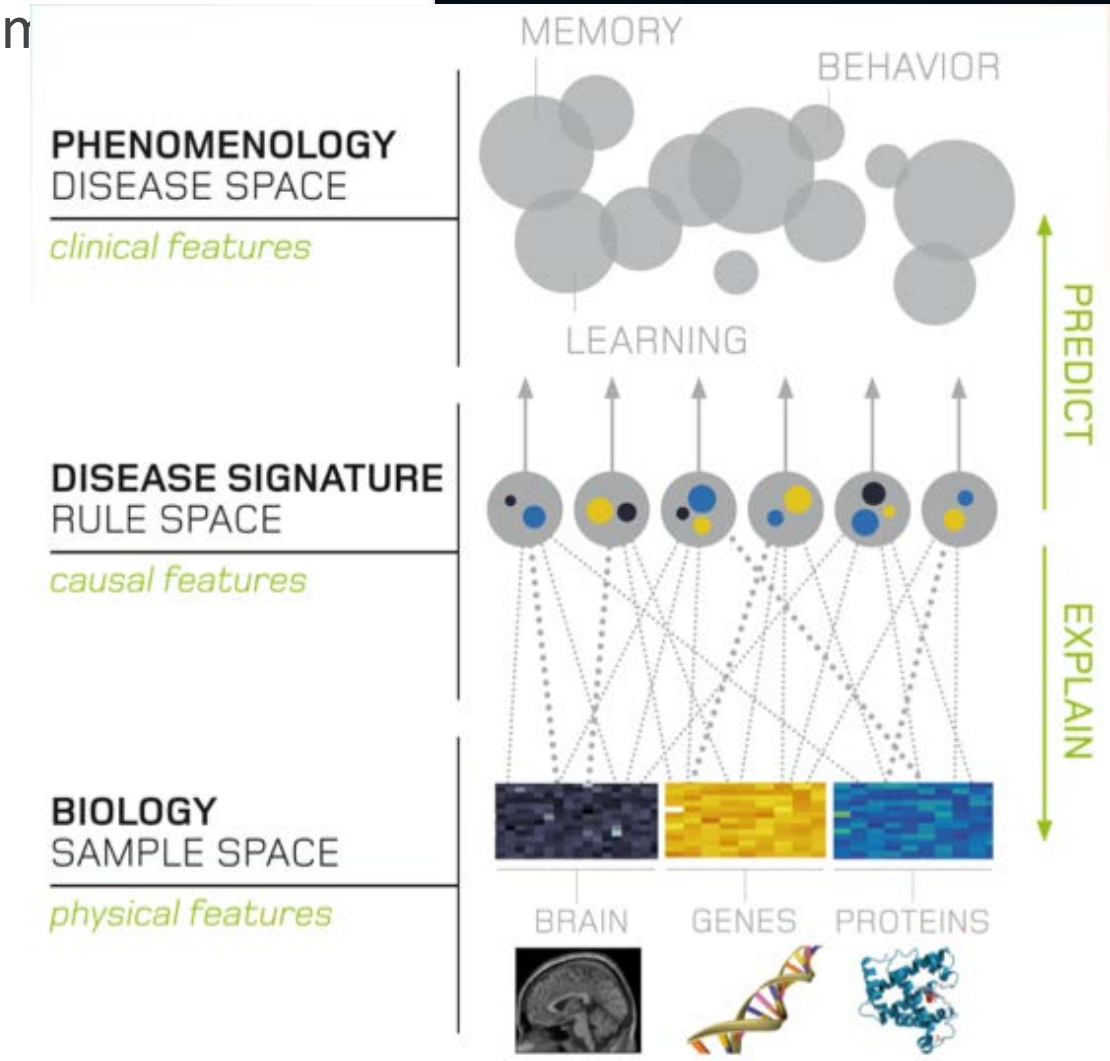
FINDING BIOLOGICAL SIGNATURES OF DISEASE

Relating biomarkers and combinations thereof to sym
diagnoses: B -> C [+ D] (multi-target prediction)

The data: D(iagnosis), C(linical), B(iomarkers)

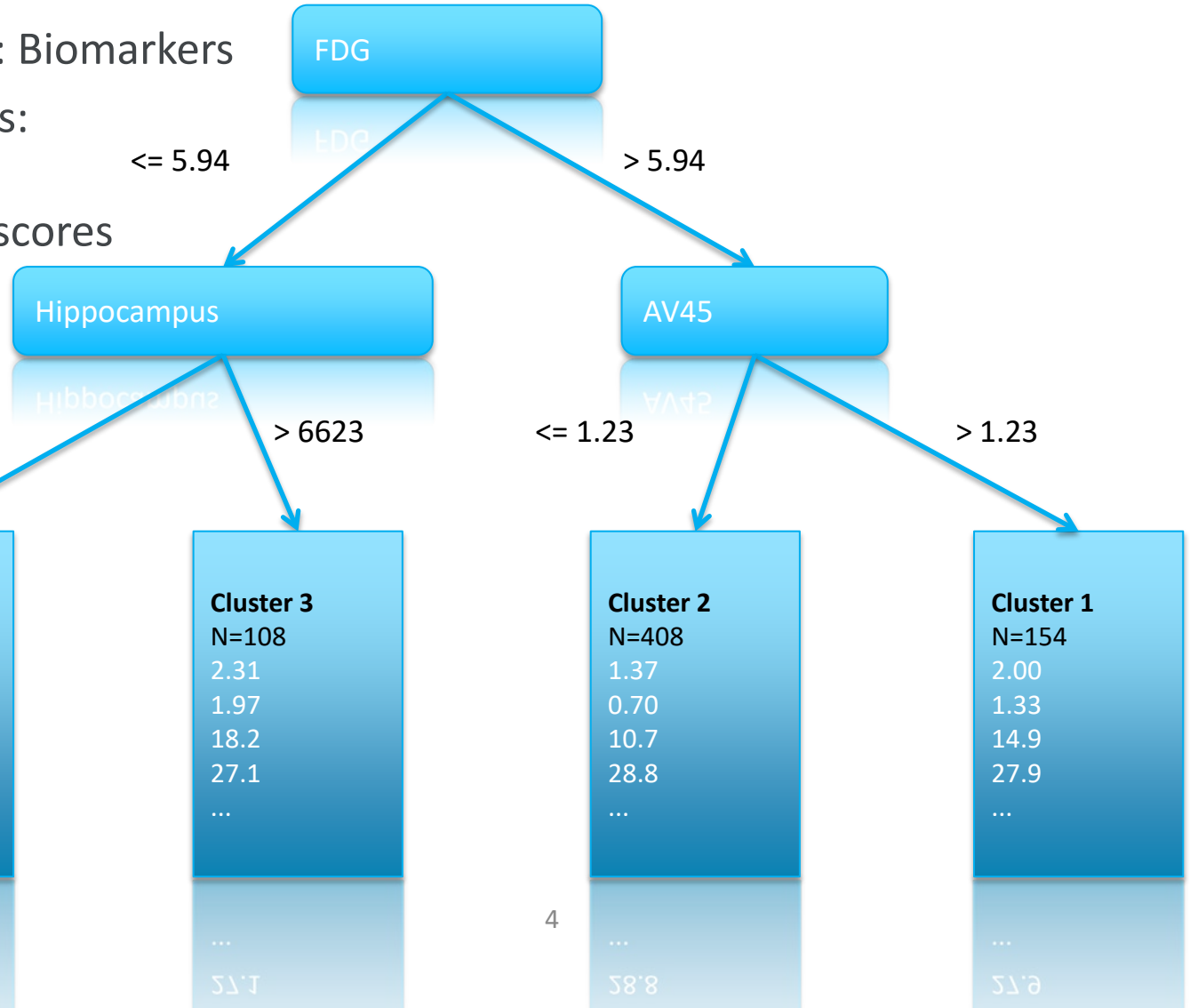
The data analysis tasks:

- Finding populations of patients with similar symptoms (clustering in C)
- Relating symptoms and combinations thereof to diagnoses: C -> D (predictive modeling)
- Finding biological signatures (clustering in B)
- Relating biomarkers and combinations thereof to diagnoses: B -> D (predictive modeling)



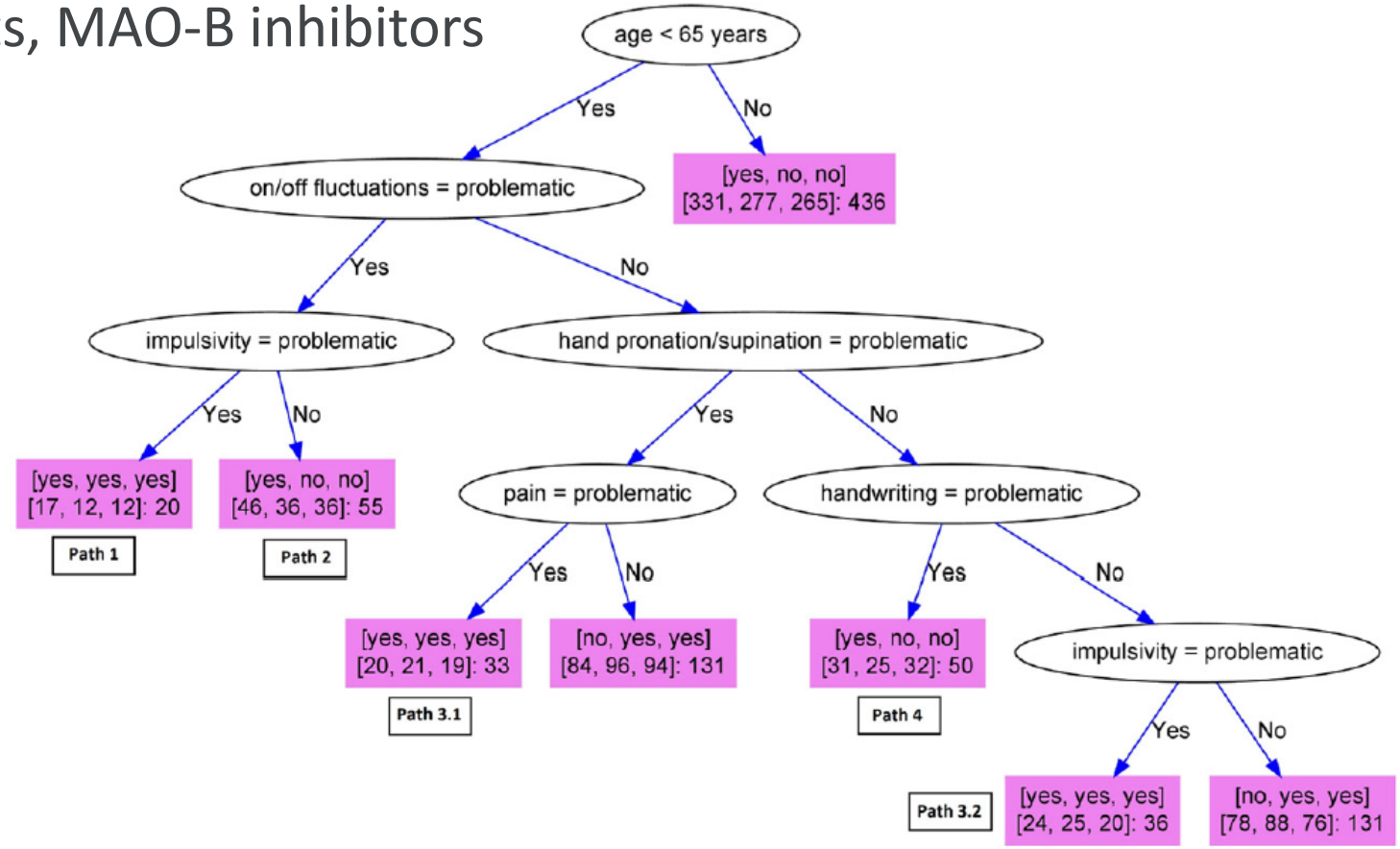
EXAMPLE OF FINDING DISEASE SIGNATURES

- Descr. attributes: Biomarkers
- Targets attributes: **diagnosis** (0-4)+ clinical msmnts/scores (23 of them)
  - DX
  - CDRSB
  - ADAS1
  - MMSE
  - …



WHEN TO CHANGE THERAPY FOR PARKINSONS

From patient's symptoms at visit, predict whether physician will change each of three groups of meds: Levodopa, dopamine agonists, MAO-B inhibitors



DETERMINING CANCER TYPE FROM METHYLATION DATA

Cancers metastasized to the liver

The task is to identify where they come from

Overall workflow

Determine a subset of useful probes

- Remove useless probes
- Select probes where normal and tumor differ

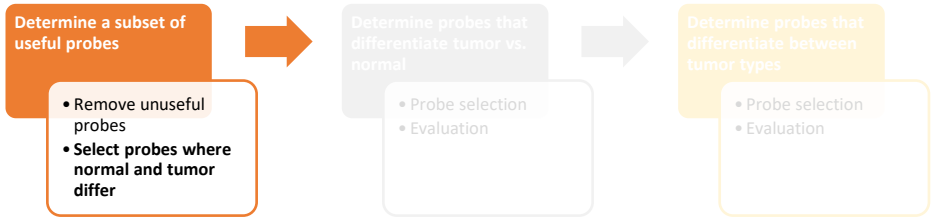
Determine probes that differentiate tumor vs. normal

- Probe selection
- Evaluation

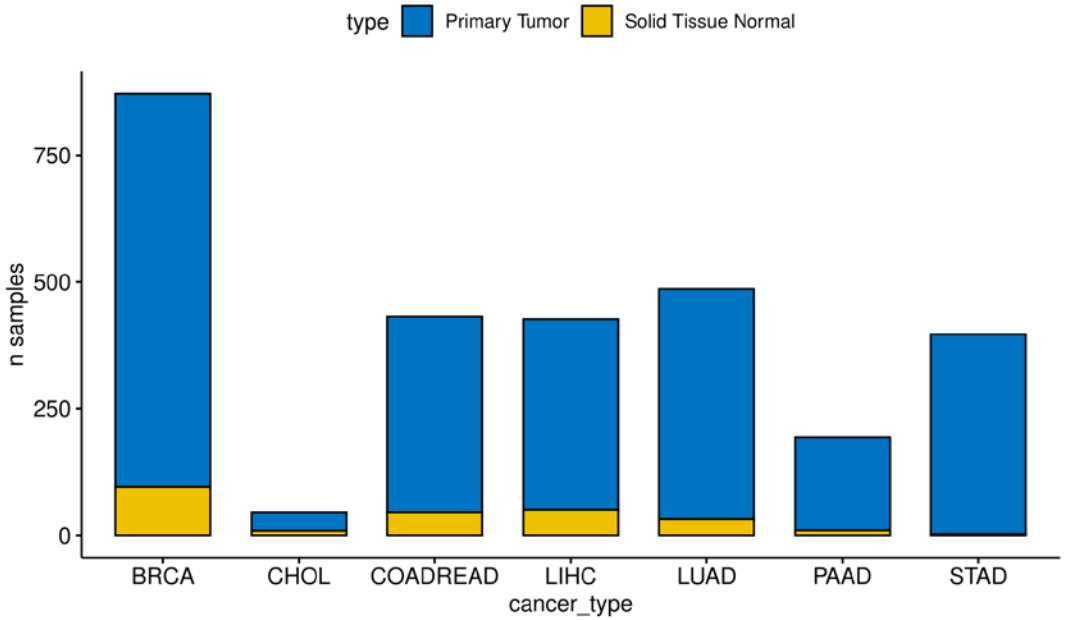
Determine probes that differentiate between tumor types

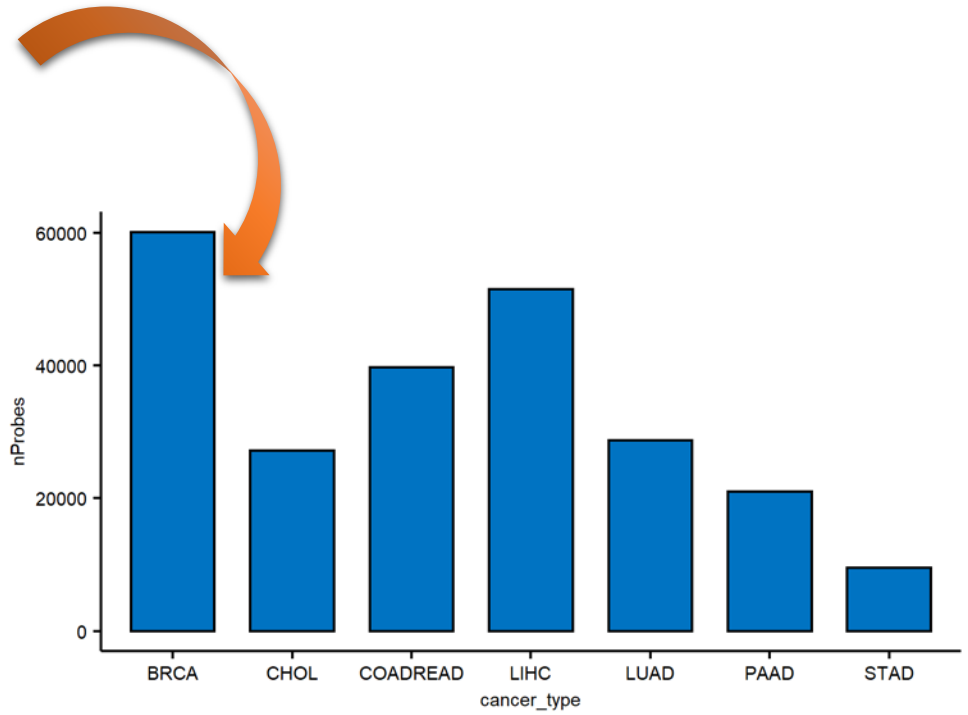
- Probe selection
- Evaluation

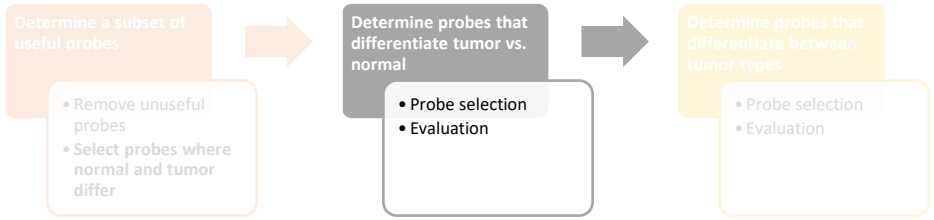
Determine useful probes



- Around **100k probes removed** for each cancer type (Probes with all missing values; Probes on chr X and Y; **Probes that have missing values;** …)
- We took probes with methylation: **tumor > 0.3 and normal < 0.1**
  - For each cancer type, an **union** of such probes was taken (across all tumor-normal pairs)
  - Between 60k (BRCA) and 10k (STAD) probes left



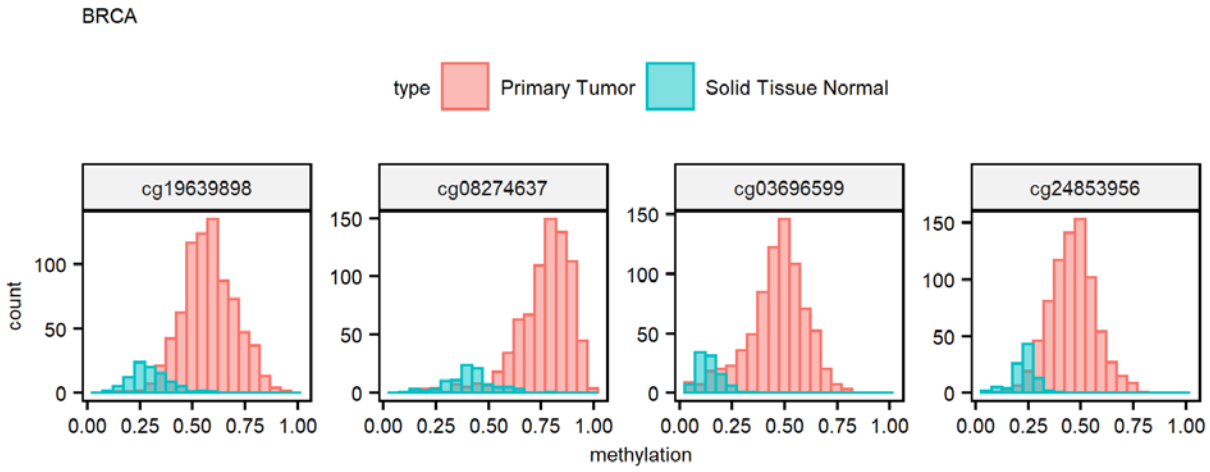


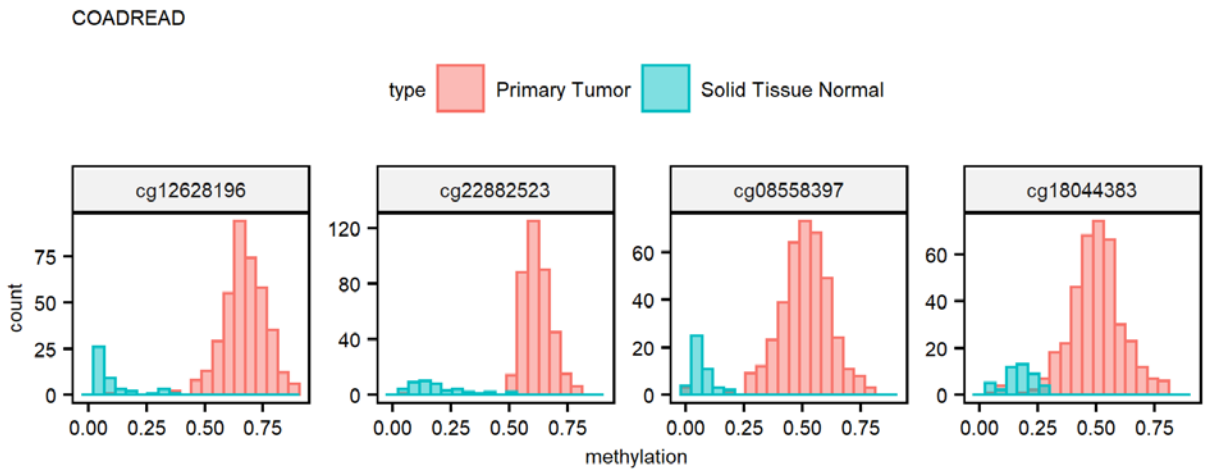


Probes that differentiate tumor vs. normal

- For all cancer types, **few probes** (2-15) were sufficient to differentiate tumor vs. normal with **high accuracy** (>95%)
  - STAD has only 2 normal samples so it was not evaluated
  - PAAD had lower accuracy (85%)

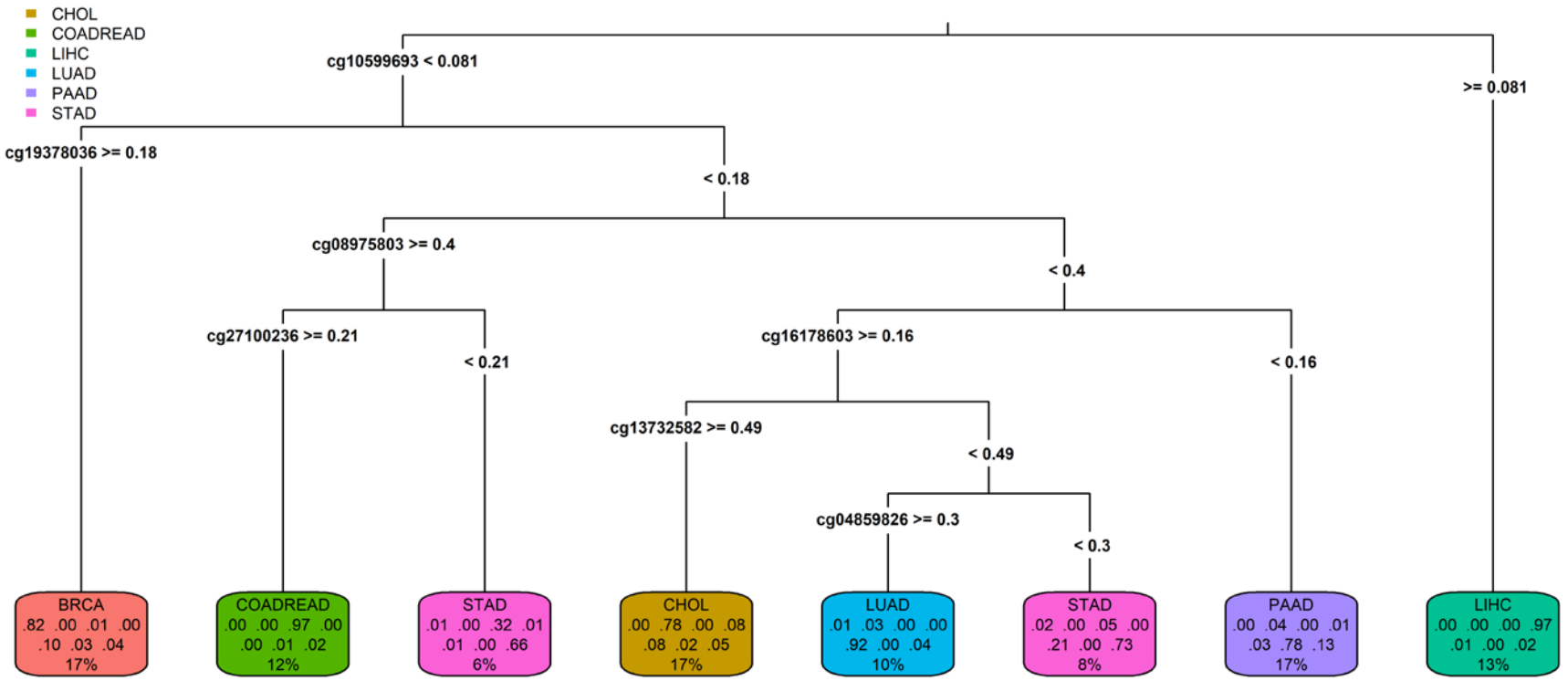
Examples of **histograms of methylation** of **most important probes**:





Differentiating among tumor types
1 model for all cancer types simultaneously

- Tree with methylation threshold = 0.15
  - Note, splitting point of tree is in the middle, hence 0.081 for the first split.

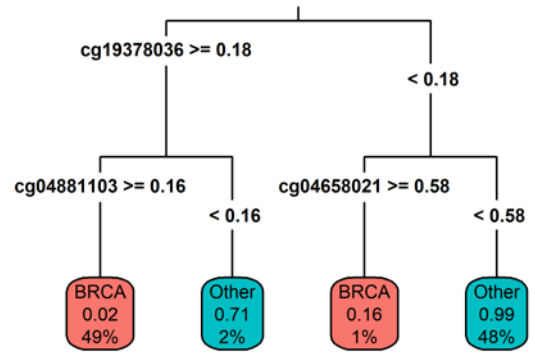


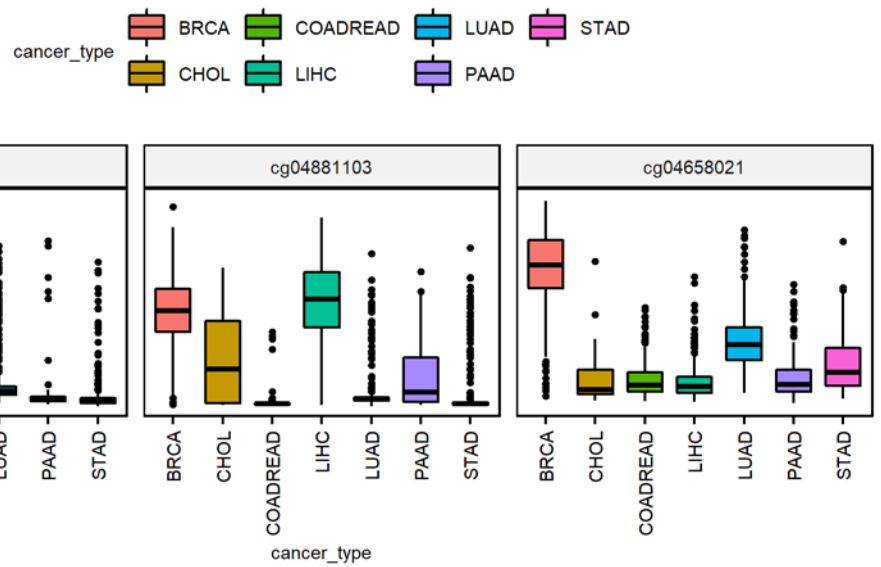
Differentiating among tumor types

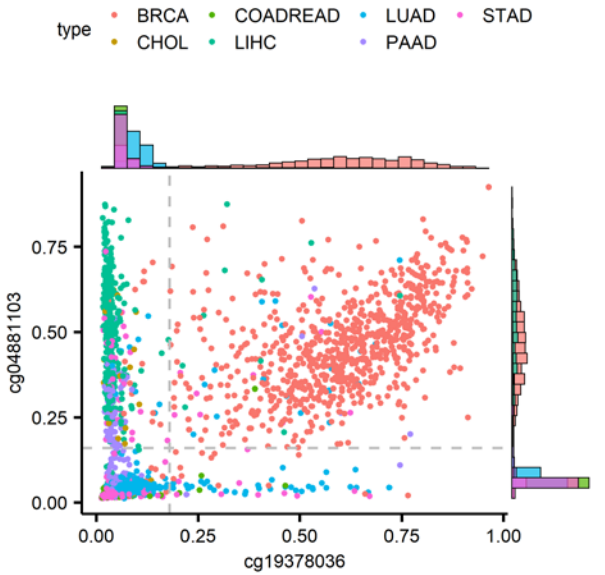
1 vs All models for each cancer type individually

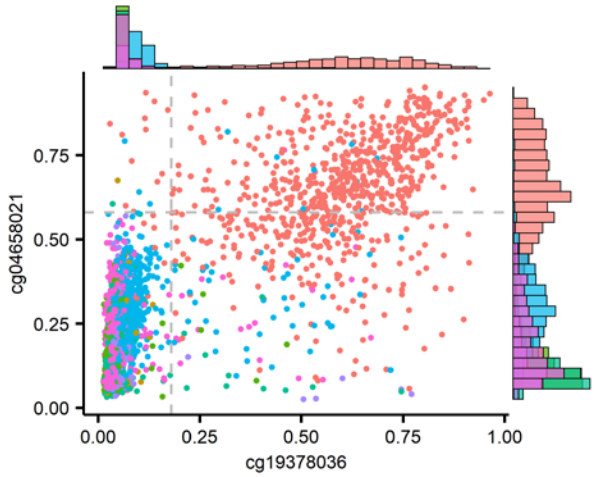
BRCA

Sensitivity	0.938
Specificity	0.9895
Pos Pred Value	0.9755
Neg Pred Value	0.9727
Precision	0.9755
Recall	0.938
F1	0.9564
Prevalence	0.3093
Detection Rate	0.2901
Detection Prevalence	0.2974
Balanced Accuracy	0.9637
cancer_type	BRCA
Feature threshold	0.15
treeDepth	2









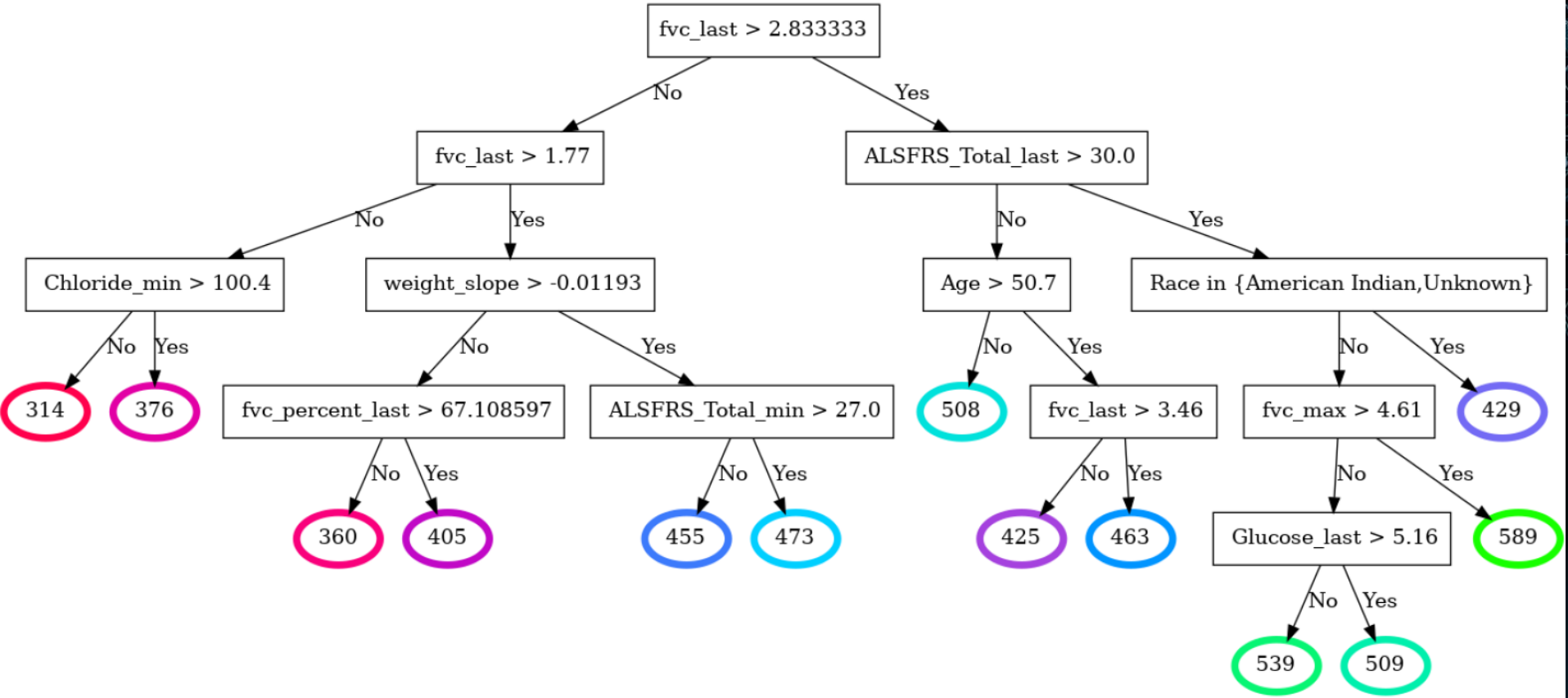
SURVIVAL ANALYSIS

- Also called time-to-event analysis
- A branch of statistics for analyzing the expected duration of time until one event occurs
  - Death in biological organisms
  - Failure in mechanical systems
- What is the proportion of a population which will survive past a certain time?
- Survival curves: Survival probability at time t

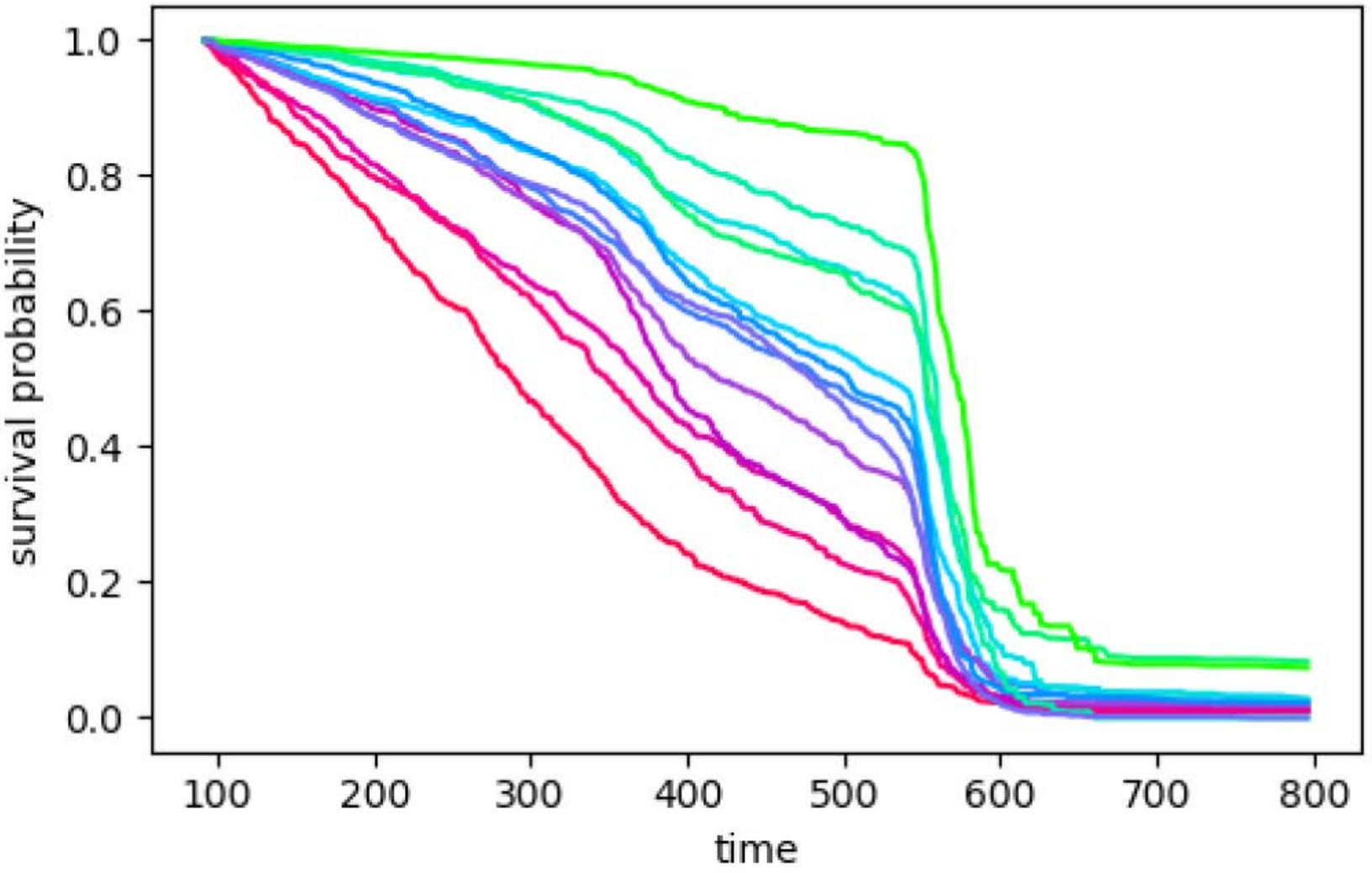
$$S_t = \frac{\text{Number of subjects living at the start} - \text{Number of subjects died}}{\text{Number of subjects living at the start}}$$

SURVIVAL TREES

- Leaves correspond to subpopulations with distinct survival curves (as opposed to single curve for whole population in classical survival analysis)



SURVIVAL CURVES IN LEAVES OF TREE



APPLICATIONS IN THE LIFE SCIENCES

- ***Predicting gene functions (HMLC=Hierarchical MLC)***

  In model organisms and bacterial genomes (genome-wide)

- ***Virtual compound screening for drug repurposing/ design***

  Tuberculosis & Salmonella; Fibrosis in myocardial infarction

- ***Relating environmental factors to biota structure***

  Microbiota in chicken/human gut; Funghi in beach sand

- ***Using machine learning to help select peptide insertion sites for regulation of protein function***

Plaper et al. *Cell Discovery* (2024)10:8
https://doi.org/10.1038/s41421-023-00635-y

Cell Discovery
www.nature.com/celldisc

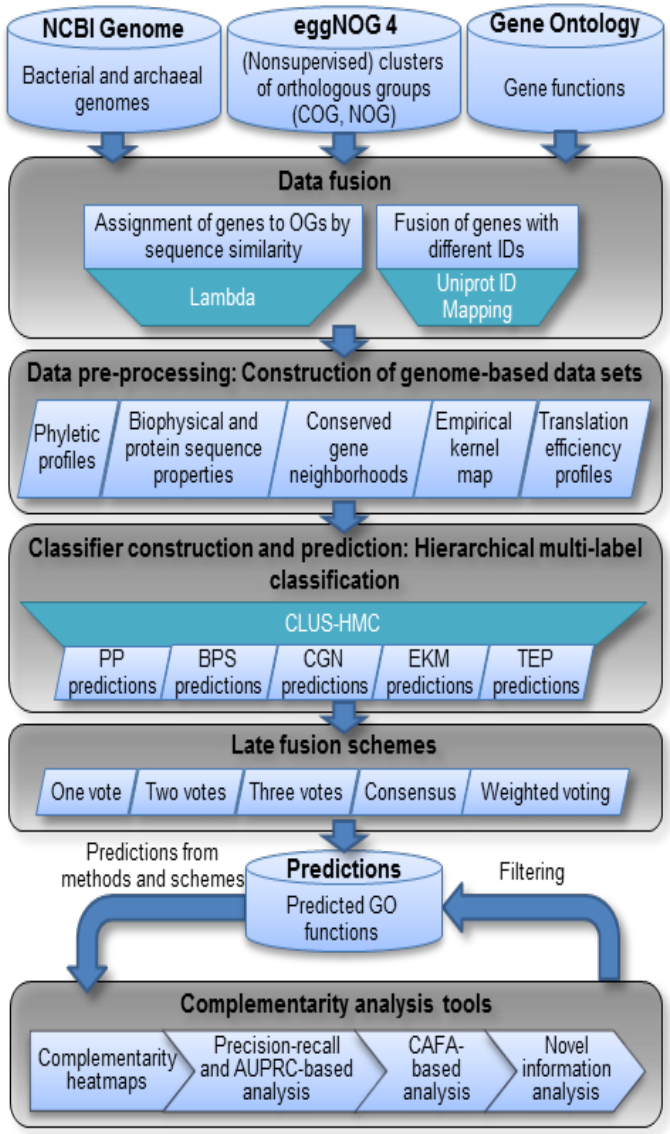
ARTICLE Open Access

Designed allosteric protein logic

Tjaša Plaper[1], Estera Merljak[1], Tina Fink[1], Tadej Satler[1,2], Ajasja Ljubetič[1], Duško Lainšček[1], Vid Jazbec[1,2], Mojca Benčina[1,3], Sintija Stevanoska[4], Sašo Džeroski[4] and Roman Jerala[1,3]✉

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

GENOME-WIDE GENE FUNCTION PREDICTION IN BACTERIA WITH PCT ENSEMBLES FOR HMLC ON DIFFERENT FEATURES



Instances : 21,626 eggNOG 4 OGs

(a)	g_1	g_2	g_3	g_4	GO
OG_1	1	0	0	1	
OG_2	1	1	0	1	?
OG_3	0	1	0	1	
OG_4	1	0	1	1	?

Phyletic Profiles (PP) — Features: 2,071 microbial genomes — Class: 4,145 Gene Ontology (GO) functions

(b)	A Entropy	Cysteine Spaced Motif	hp2: AAAAC	GO
OG_1	-0.27	-0.21	0.28	
OG_2	-0.12	-0.19	0.09	?
OG_3	-0.15	-0.18	0.08	
OG_4	-0.77	-0.24	-1.11	?

Features: 1,170 biophysical and sequence derived attributes — Biophysical and Protein Sequence Properties (BPS)

(c)	g_1	g_2	OG_1	OG_2	GO
OG_1	0.71	0.53	1	0.71	
OG_2	0.48	0.25	0.71	1	?
OG_3	1.22	0.56	-0.27	0.44	
OG_4	0.66	0.56	0.34	-0.59	?

Features: 2,071 microbial genomes + 5,891 OGs that appear in at least 100 genomes — Translation Efficiency Profiles (TEP)

Instances : 21,626 eggNOG 4 OGs

(d)	OG_1	OG_2	OG_3	OG_4	GO
OG_1	-33.22	19.96	13.88	11.21	
OG_2	19.96	-33.22	23.81	23.81	?
OG_3	13.88	23.81	-33.22	20.38	
OG_4	11.21	23.81	20.38	-33.22	?

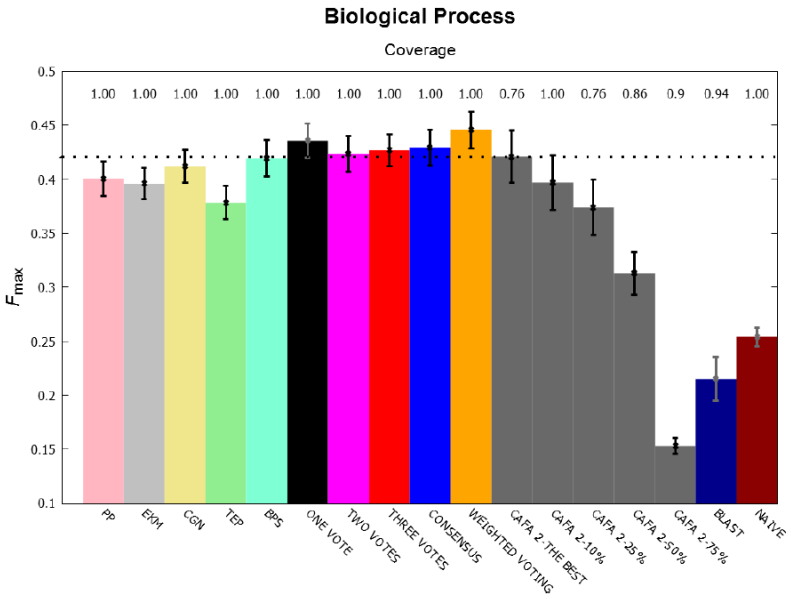
Features: 5,891 OGs that appear in at least 100 genomes — Conserved Gene Neighborhoods (CGN)

(e)	OG_1	OG_2	OG_3	OG_4	GO
OG_1	0	0.24	6.64	6.64	
OG_2	0.24	0	-9.87	1.32	?
OG_3	6.64	-9.87	0	6.64	
OG_4	6.64	1.32	6.64	0	?

Features: 8,447 OGs from 6 model organism genomes — Empirical Kernel Map (EKM)

- 5000 bacterial genomes
- 5 different feature sets
- Predictive models learned from each FS: Tree ensembles for

GENOME-WIDE GENE FUNCTION PREDICTION IN BACTERIA USINNG META-GENOME PHYLETIC PROFILES AS FEATURES



- Excellent results, comparable to
- the best in the CAFA 2 challenge

Metagenome Phyletic Profiles (MPP)

Features: metagenomes

	m_1	m_2	m_3	m_4	GO
OG_1	0	10E-5	0	0	
OG_2	10E-6	0	10E-7	10E-9	?
OG_3	0.008	0.02	0	0.01	
OG_4	0	0	0.003	0	?

Feature values: sum of OG member genes abundances in metagenomes

Phyletic Profiles (PP)

Features: microbial genomes

	g_1	g_2	g_3	g_4	GO
OG_1	1	0	0	1	
OG_2	1	1	0	1	?
OG_3	0	1	0	1	
OG_4	1	0	1	1	?

Feature values: presence/absence of genes in genomes

MPP can predict hundreds of gene functions that would not be predicted using only PP

VIRTUAL COMPOUND SCREENING

- Descriptive variables refer to compound structure
  - Functional groups
  - Fingerprints
  - Bulk properties
- May also describe the compound in terms of the proteins it targets (e.g. from PubChem)
  - Their functional annotations
  - Pathways they are involved in
  - Proteins that the targets interact with (and/or their functional annotations, pathways they are involved in)
- Target variables describe compound activity and toxicity

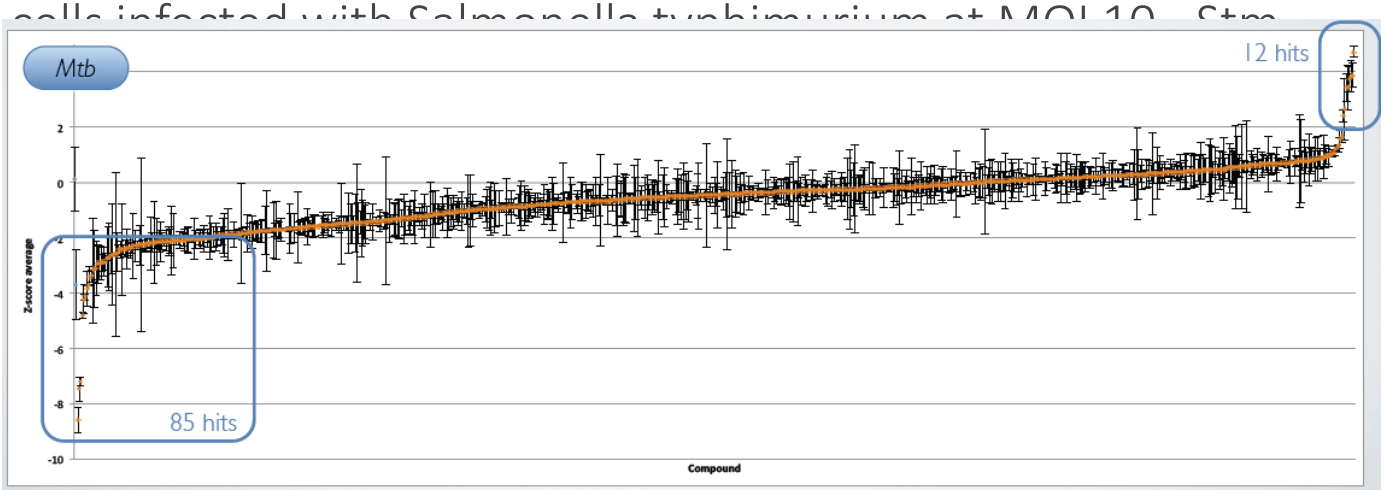
HOST-TARGETED DRUGS FOR MTB (TUBERCULOSIS) AND STM (SALMONELLA)

Library of compounds

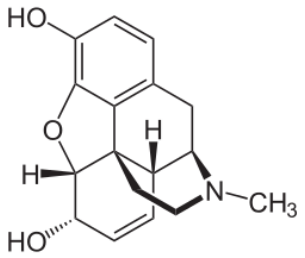
- LOPAC library - Library Of Pharmacologically Active Compounds
  - 1260 compounds
- Well-characterized compounds, many already applied in clinical practice for a range of conditions

Flow cytometry (FACS) - measured reduction in bacterial load

- MelJuSo cells infected with Mycobacterium tuberculosis at MOI 10 – Mtb
- HeLa cells infected with Salmonella typhimurium at MOI 10 – Stm



MTB&STM: HOST-TARGETED DRUGS



Morphine

Given SDF files, find PubChemID

PubChem repository

- Retrieve the proteins that were found to be active in bio-assays with human cells

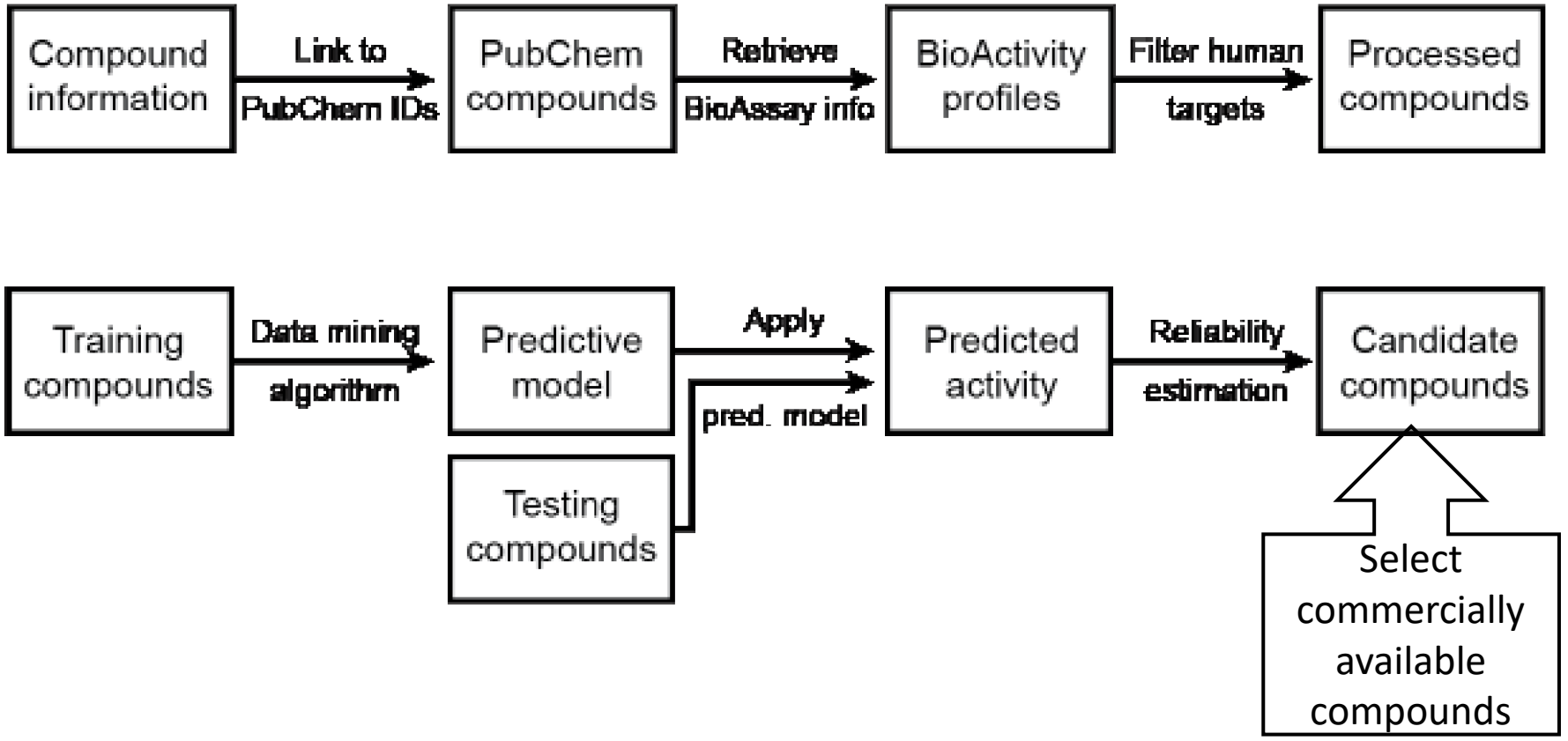
Dataset

- 964 compounds were found active on human protein targets
- 711 distinct protein targets were identified

Each compound is described with

- the respective protein targets
- functional annotations of the respective protein targets
- functional annotations of both the respective protein targets and the proteins they interact with

MTB&STM: HOST-TARGETED DRUGS THE DATA ANALYSIS WORKFLOW



MTB&STM: HOST-TARGETED DRUGS RESULTS

Greatly increased proportions of hit compounds

- 5 out of 9 (55.6%) for Mtb and
- LOPAC primary screen (90 out of 1260 (7.1%) for *Mtb*

The *in silico* predictive model successfully identified active compounds *de novo*

Abbr.	*Compound name*	*Alternative name(s)*	*Primary sceen z-score*	*Rescreen z-score*	*Activity*
Mycobacterium tuberculosis					
SU	SU 6656	2,3-Dihydro-N,N-dimethyl-2-oxo-3-[(4,5,6,7-tetrahydro-1H-indol-2-yl)methylene]-1H-indole-5-sulfonamide	**-5.79**	**-10.51**	Src family kinase inhibitor
Q	Quinacrine dihydrochloride		**-5.25**	**-9.90**	MAO inhibitor
SB	SB 216763	3-(2,4-Dichlorophenyl)-4-(1-methyl-1H-indol-3-yl)-1H-pyrrole-2,5-dione	**-6.02**	**-8.29**	GSK-3 kinase inhibitor
G	GW5074	3-(3, 5-Dibromo-4-hydroxybenzylidine-5-iodo-1,3-dihydro-indol-2-one)	**-4.86**	**-6.98**	Raf1 kinase inhibitor
T494	Tyrphostin AG 494	N-Phenyl-3,4-dihydroxybenzylidenecyanoacetamide	**-3.83**	**-6.93**	EGFR kinase inhibitor
L	3',4'-Dichlorobenzamil hydrochloride	L-594,881	**-3.87**	-5.13	$Na^+/Ca2^+$ exchanger inhibitor
H	Haloperidol		**-3.77**	-2.96	D2/D1 dopamine

ANALYZING DATA FROM HIGH-CONTENTS SCREENS

- Compounds described by fingerprints
- Generated by open-source chemoinformatics SW library RDkit
- The FCFP2 fingerprits were used (1024 features)
- Also considered profiles of targeted proteins
- These are the attributes

- Assays photographed under the microscope
- Features extracted from images
- These are then the targets

HTS: MODULATING FIBROBLAST TO MYOFIBROBLAST TRANSITION

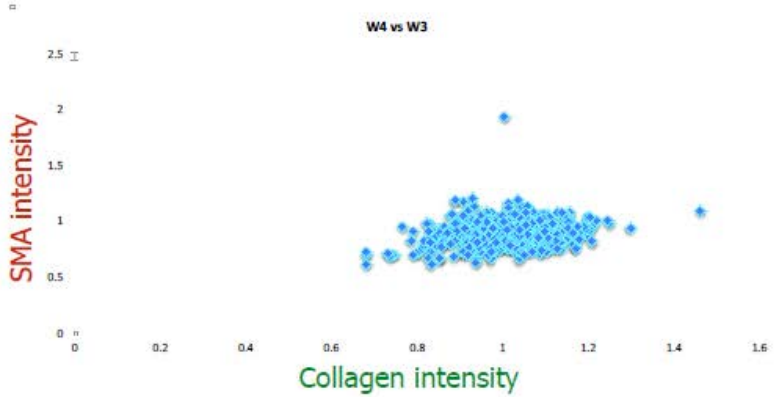
cardiac fibroblasts from a-SMA-RFP/
Coll α1(I)-EGFP mice

0h

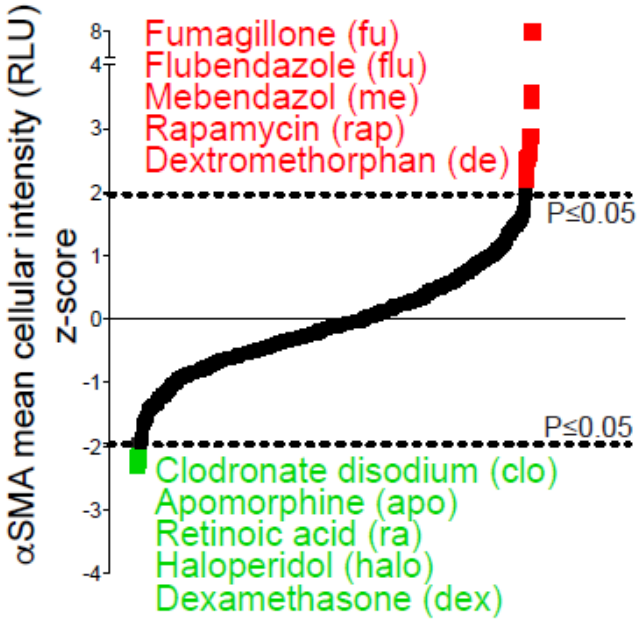
FDA-approved drugs (640 compounds)

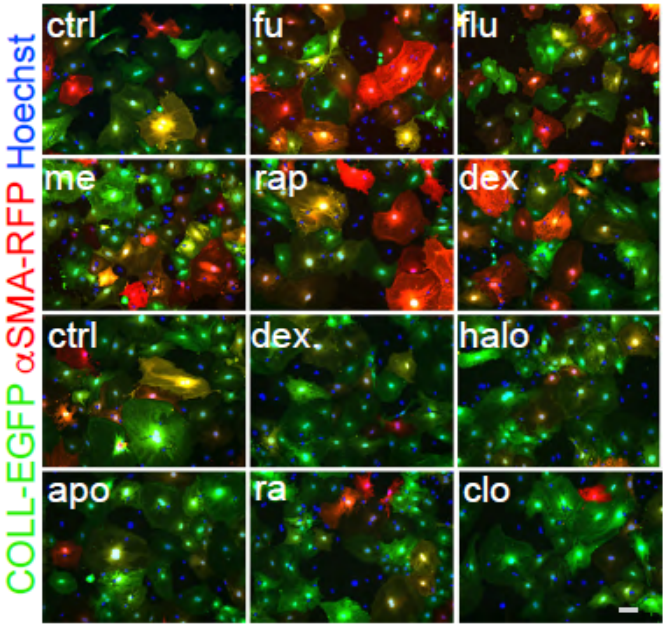
72h

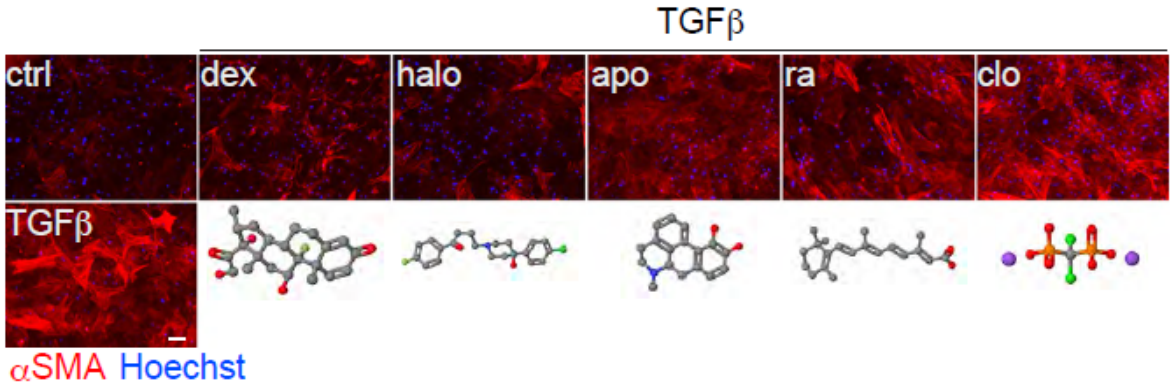
cell fixation, image acquisition and elaboration

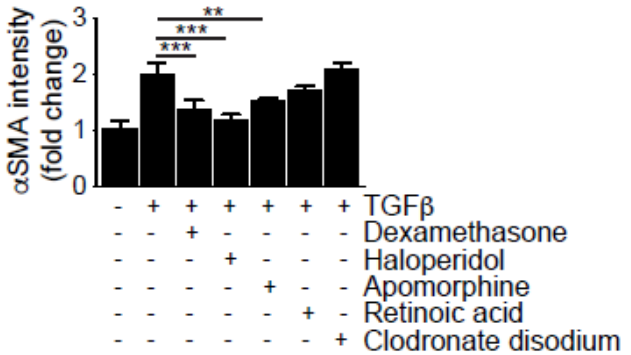


HITS IN THE HTS SCREEN



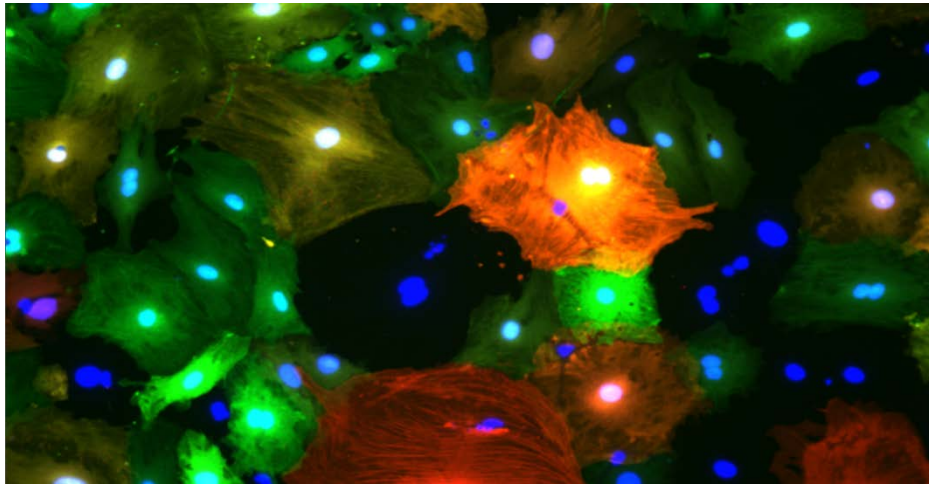






REDUCING FIBROSIS IN MYOCARDIAL INFARCTION

- High content screen using a library of 640
- FDA approved drugs (ENZO)
- Identify drugs to reduce fibrosis in myocardial infarction
- Screen used murine cardiac fibroblasts which differentiate into myofibroblasts in culture, expressing increased alpha SMA-RFP and collagen-alpha1-EGFP
- Targets: Intensity of
  - alphaSMA
  - Collagen
- Attributes
  - Fingerprints



TESTING THE PREDICTIONS

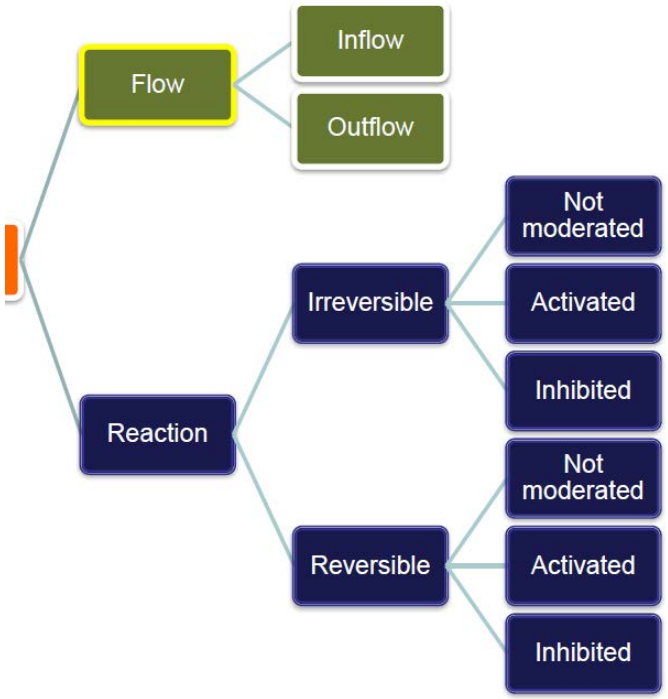
- Some domain-specific knowledge / constraints applied: Predicted compounds filtered for FDA approved drugs that are not corticosteroids
- SMILE strings used in Chemmine to identify substances with structural similarity to non commercial compounds with high predicted values
- Three related compounds identified which are described in literature to have an anti-fibrotic effect
- Four related compounds identified which were not previously described to have an anti-fibrotic effect
- Tested in the wet-lab and one works really well

APPLICATIONS OF PROCESS-BASED MOD.

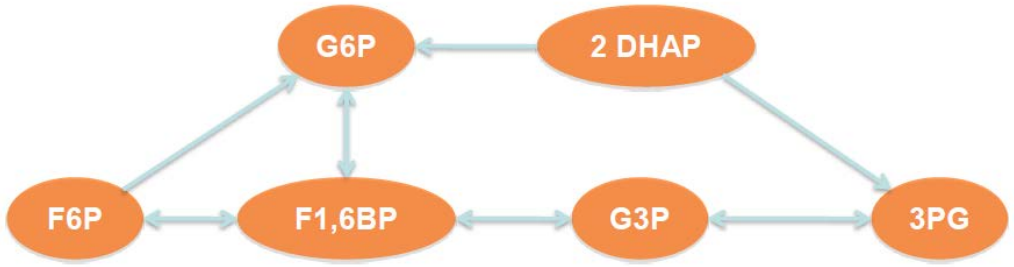
Application areas

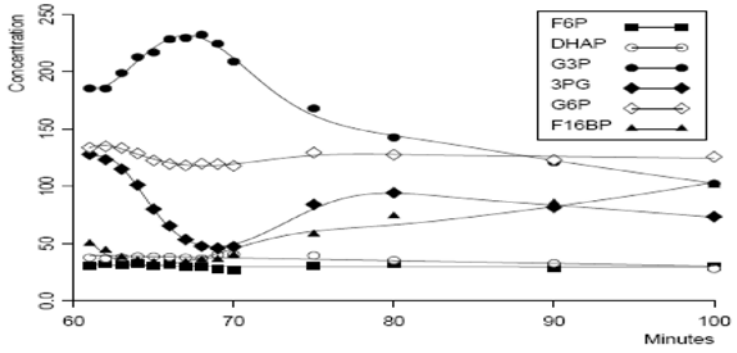
- **Systems Biology**
  - Metabolic networks
  - Immune response
- **Synthetic Biology**
  - Designing biological circuits
- **Neuroscience**
  - Modelling individual neurons
  - Modelling populations of neurons

MODELING METABOLIC NETWORKS



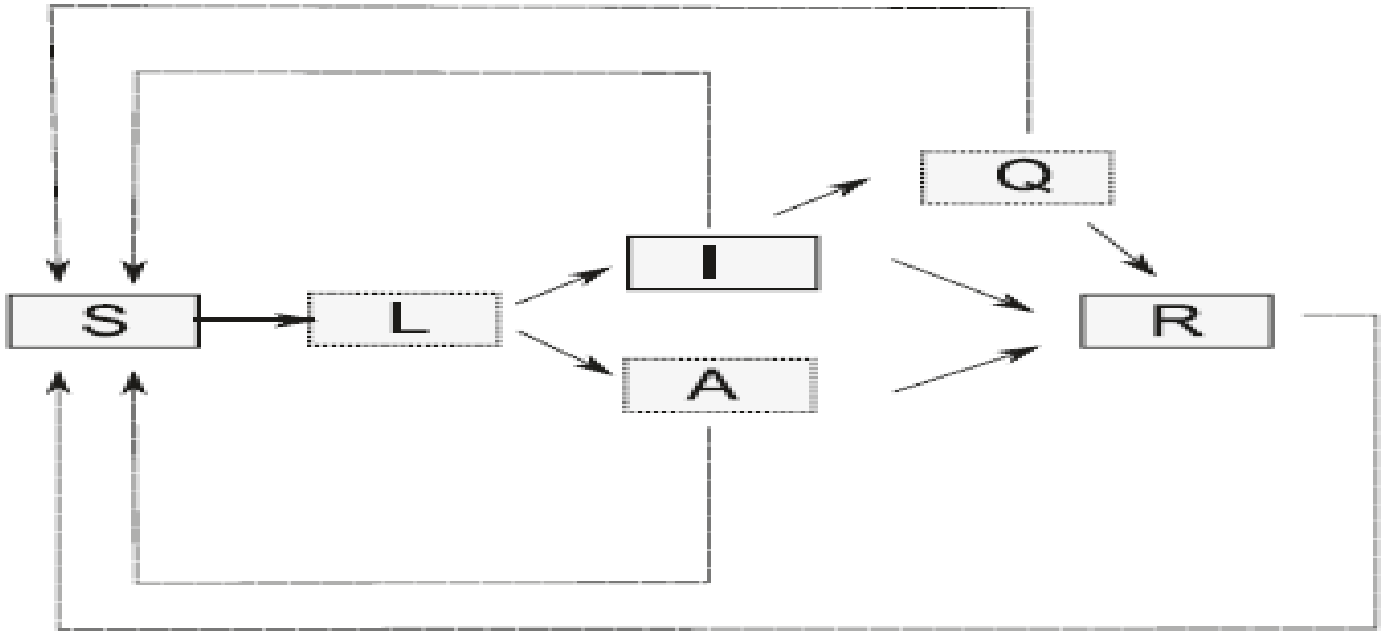
- Inducing (partial) chemical network of glycolisys
  - Data: temporal responses of species to pulse changes (14 time points)
  - From: Torralba et al. (2003) PNAS 100(4): 1494-1498
- Responses of six chemical compounds:
  - G6P (glucose 6-phosphate)
  - F6P (fructose 6-phosphate)
  - F1,6BP (fructose 1,6-bisphosphate)
  - G3P (glycerol 3-phosphate)
  - 3PG (3-phosphoglycerate)
  - DHAP (dihydroxyacetone 3-phosphate)





LEARNING EPIDEMIOLOGICAL MODELS

- Eyam plague outbreak (SIR, SLIR)
- Tristan da Cunha influenza outbreak (SLIR, SIR)
- COVID-19 in different countries (SLIAQRS)



LEARNING MODELS FOR COVID-19

- SLIAQRS model for Slovenia
- First wave in 2020
- Before first vaccinations