# SIMPATI
# NETWORK-BASED SYSTEM FOR PATIENTS' CLASSIFICATION REVEALS DISEASE SPECIFIC PATHWAYS DRIVEN BY COHESIVE COMMUNITIES

Luca Giudice, Claudia Mengoni, Rosalba Giugno

Presentation by Rosalba Giugno
Department of Informatics, University of Verona
Verona, Italy, rosalba.giugno@univr.it

UNIVERSITÀ di VERONA | Dipartimento di INFORMATICA



- Rosalba Giugno is Associate Professor in Computer Science at the Department of Computer Science, University of Verona since May 2016. She received the bachelor degree in Computer Science at the University of Catania in 1998 cum Laude. She received the PhD in Computer Science from the University of Catania in the 2003. For her PhD program, she spent three years of research abroad, at the University of Maryland, at the New York University, and at the Cornell University (NY) working with prof. Dennis Shasha, prof. VS Subrahmanian and prof. Ronitt Rubinfeld.

- Since she moved in Verona in 2017, she leads a research group, InfOmics (https://infomics.github.io/InfOmics/), of 1 post-docs, 4 PhD students, 3 young researchers and 10 master thesis students. She is director of the Master Degree in Medical Bioinformatics at University of Verona. She is also Director of the Italian National Laboratory of Bioinformatics, coordinating 35 Italian academic groups of Bioinformatics researchers. She leads the bioinformatics work packages of several European Union's Horizon 2020 research and innovation programme and JPcofuND2 Personalised Medicine for Neurodegenerative Diseases projects.

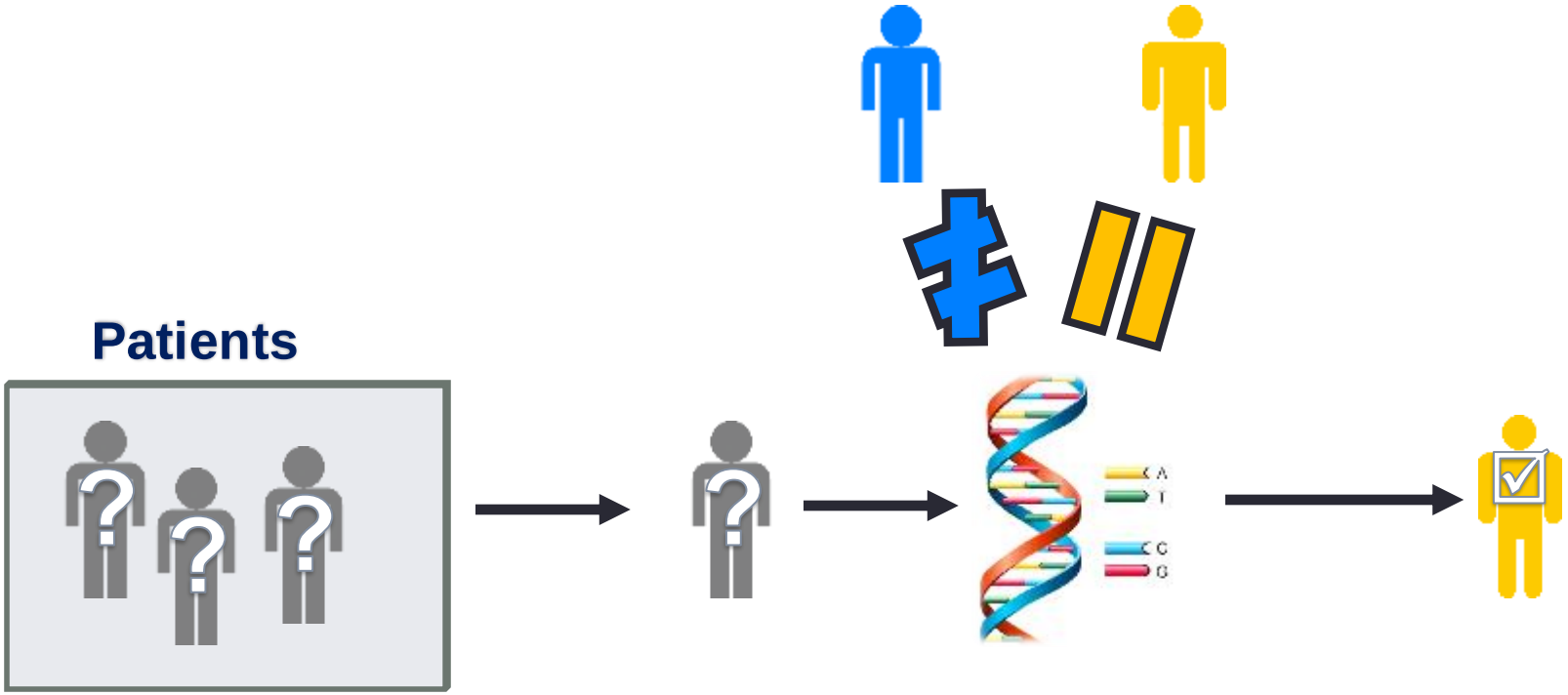
- She is editor for Nature Elsevier Information Systems, Frontiers in Genetics, and Frontiers in Bioinformatics.

- Her scientific interests concern algorithms on graphs and biological networks, machine learning for classification of subjects, for the reuse of drugs, for the analysis of single cell multi-omics data and the modeling of biological systems.

# Aim of the work

Build a patient classifier

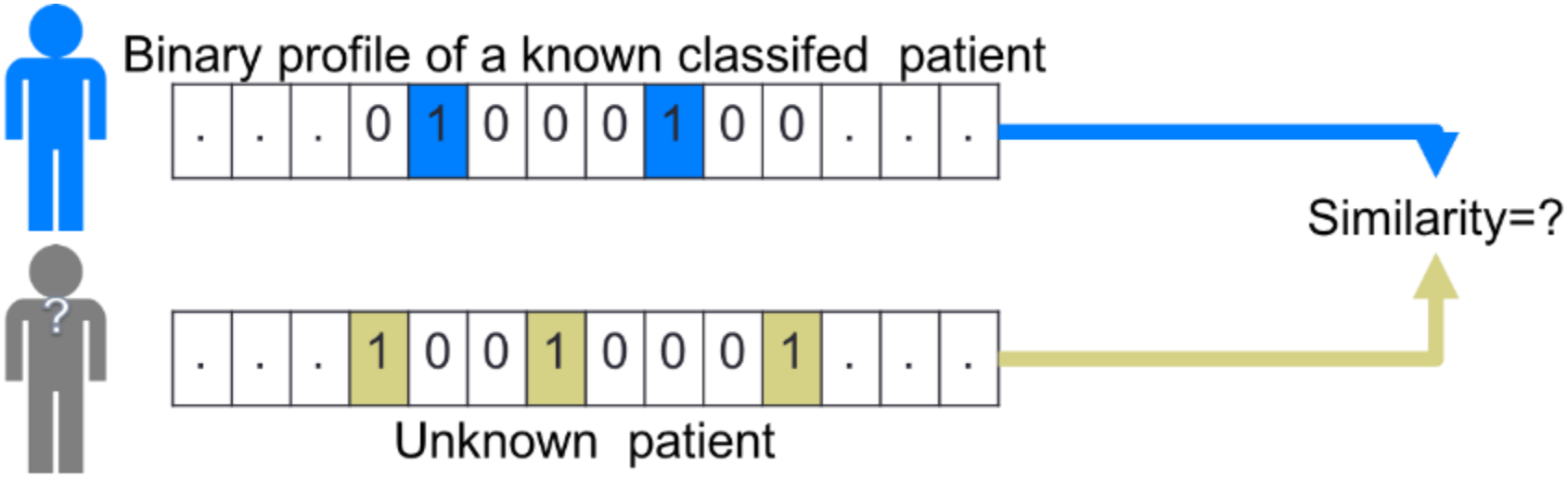
1. Efficiently dealing with sparse data
2. Exploiting pathway-level information
3. Identify the input features predictive of patient classes
4. Accurately assign new patients to the correct class.



# Mutations form binary profiles

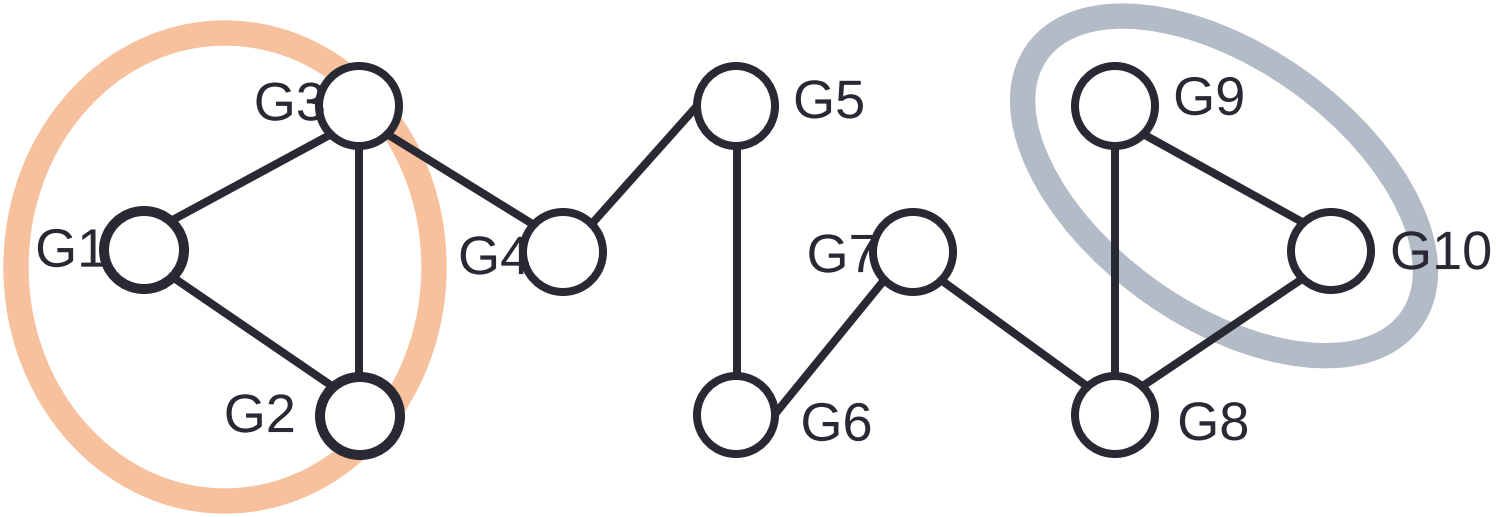
Binary somatic mutation profiles shows few mutated genes

- Profiles are difficult to compare, a lot of zero's, few 1's, curse of dimensionality



# Mutated genes tend to fall into recurrent pathways

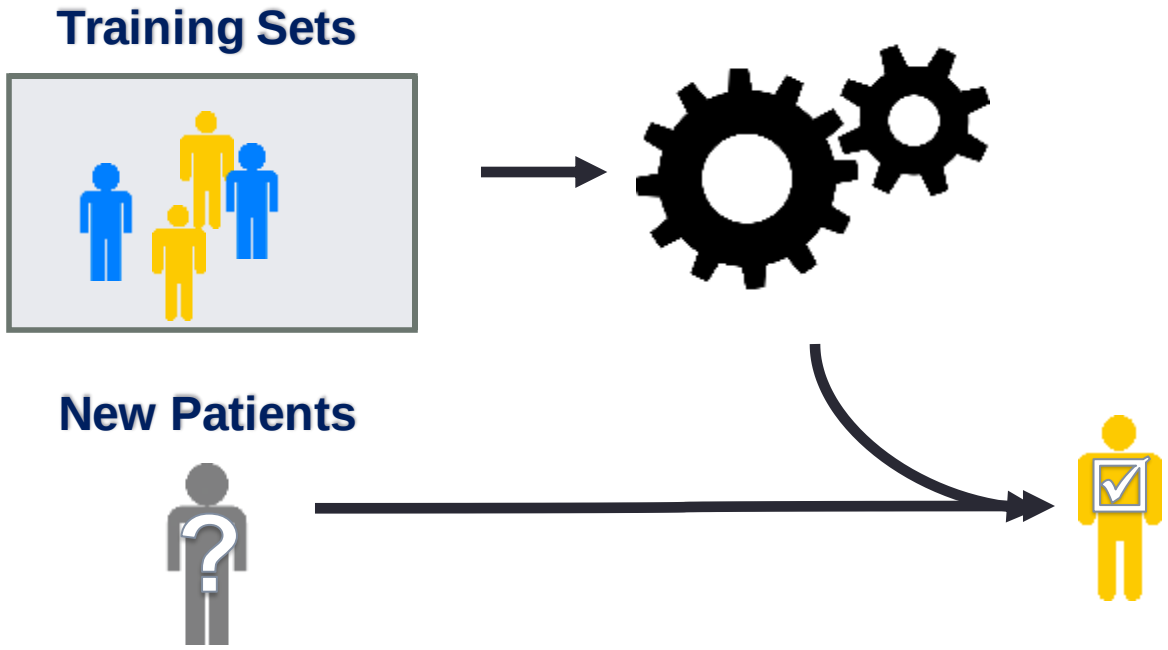
| | GLUCONEOGENESIS | | | | | | | | CELL CYCLE PATHWAY | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **G1** | **G2** | **G3** | G4 | G5 | G6 | G7 | G8 | **G9** | **G10** |
| **P1** | **1** | **1** | 0 | 0 | 0 | 0 | 0 | 0 | **1** | **0** |
| **P2** | 0 | 0 | **1** | 1 | 0 | 0 | 0 | 0 | 0 | **1** |




As for the case of renal clear cell carcinoma and metabolic pathways like fatty acid synthesis.
The Cancer Genome Atlas Research Network *Nature*, 499, pages 43–49, 2013

# Let’s construct a classifier

Given known cancer patients in two classes, **early stage** and **late stage**:
Build a classifier s.t. to predict the status of unknown patient comparing his properties to the knowing's ones



# Simpati: a classifier based on networks

For each pathway build a patient similarity network (PSN)

## Patient data

| | G1 | G2 | G3 | G4 | … |
|---|---|---|---|---|---|
| P1 | 1 | 1 | 1 | 0 | 0 |
| P2 | 0 | 1 | 0 | 1 | 0 |
| P3 | 0 | 1 | 0 | 1 | 0 |
| P4 | 1 | 0 | 0 | 0 | 0 |

Somatic mutation profiles

## Pathway grouping

**APOPTOSIS**

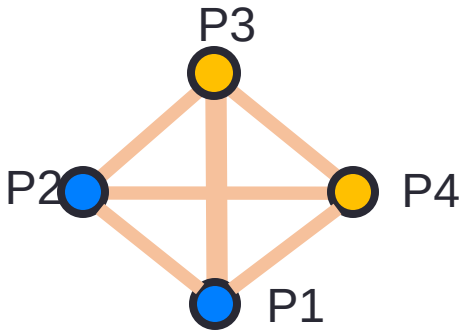
| | G1 | G2 |
|---|---|---|
| P1 | 1 | 1 |
| P2 | 0 | 1 |
| P3 | 0 | 1 |
| P4 | 1 | 0 |

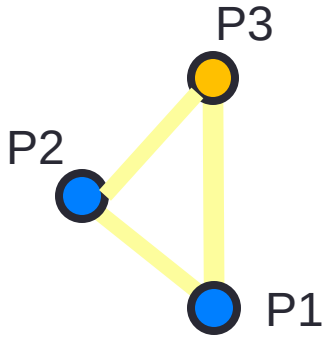
**GLYCOLYSIS**

| | G3 | G4 |
|---|---|---|
| P1 | 1 | 0 |
| P2 | 0 | 1 |
| P3 | 0 | 1 |

…..

## Pathway-level PSN

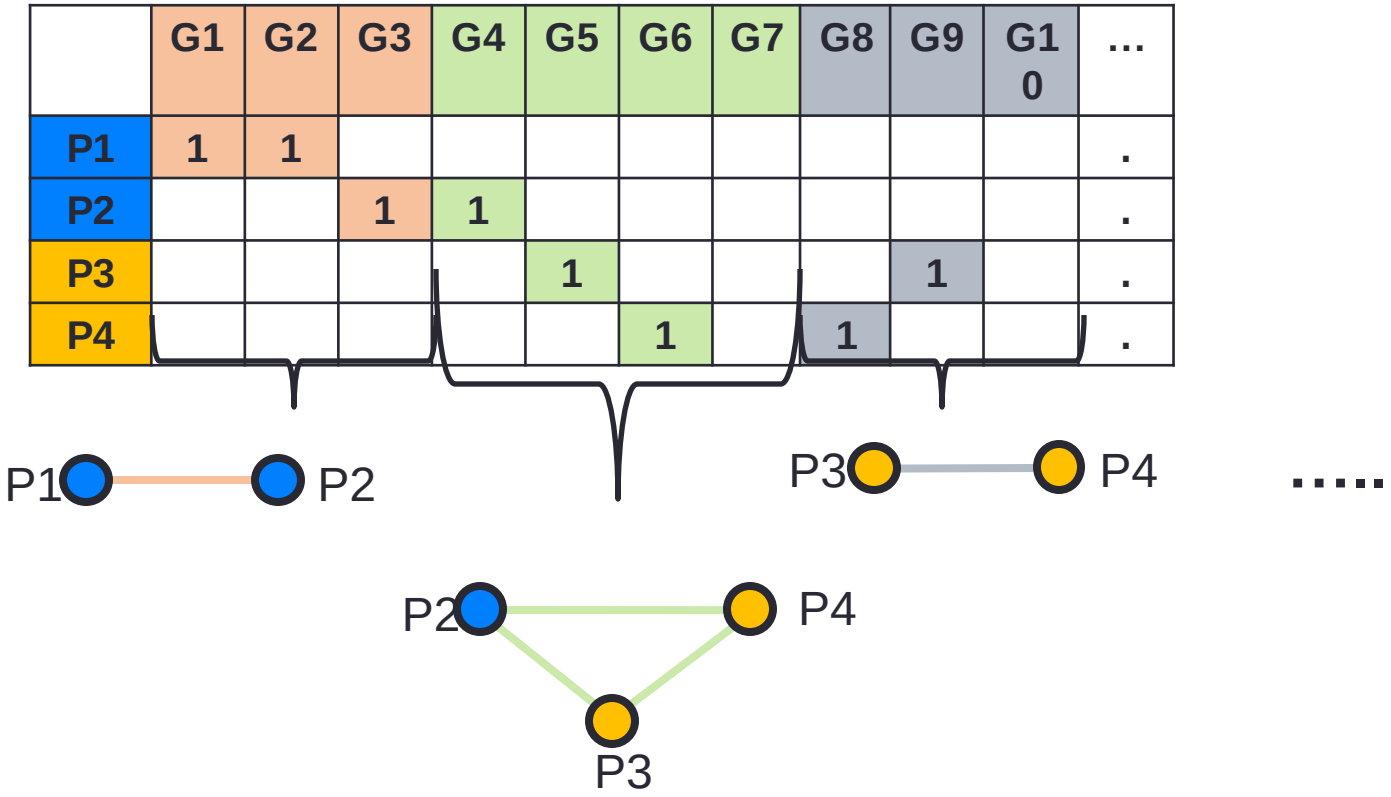



…..

# Classifying by using a somatic mutation profile

- Build PSNs with only the patients with mutated genes in pathways
- PSNs are complete networks
- Patients in the same PSN are considered all similar



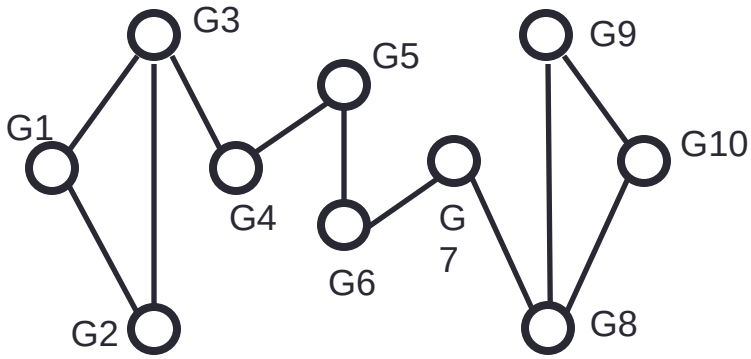
# Gene-gene networks help smooth sparse data

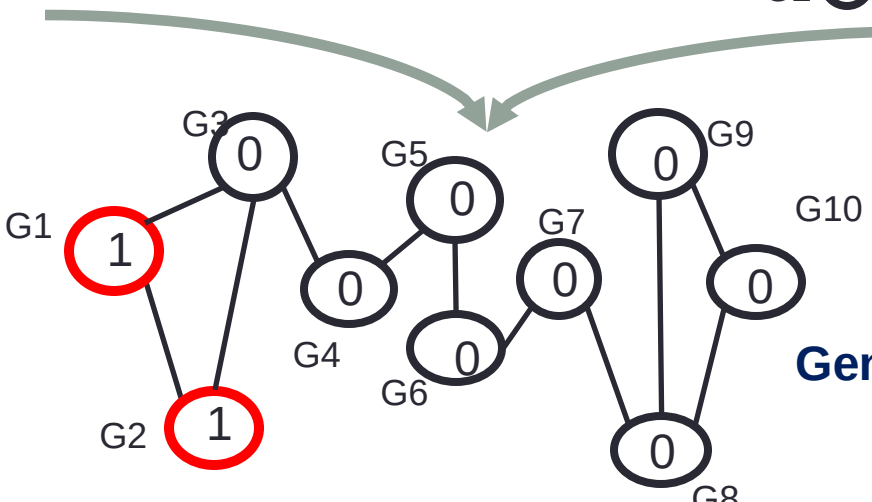
1. Genes of the profiles represented as nodes in a gene interaction network
2. Mutations are represented by the weights of the nodes

**Patient data**

| | G1 | G2 | G3 | G4 | G5 | G6 | G7 | G8 | G9 | G10 |
|---|---|---|---|---|---|---|---|---|---|---|
| P | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

**From literature**

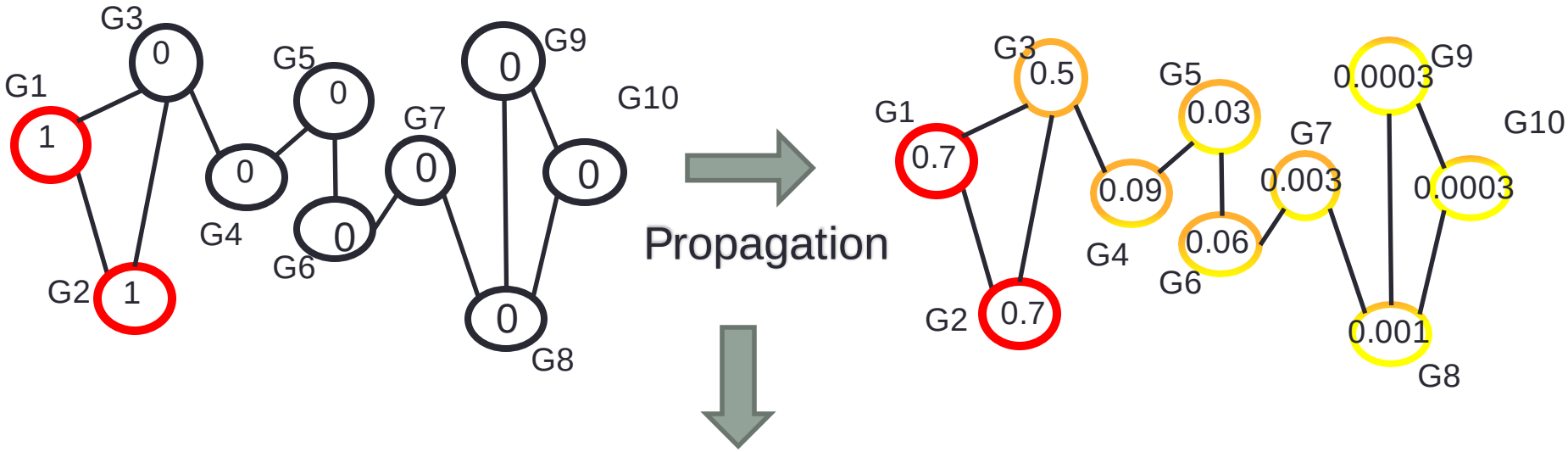



**Gene network for patient P**

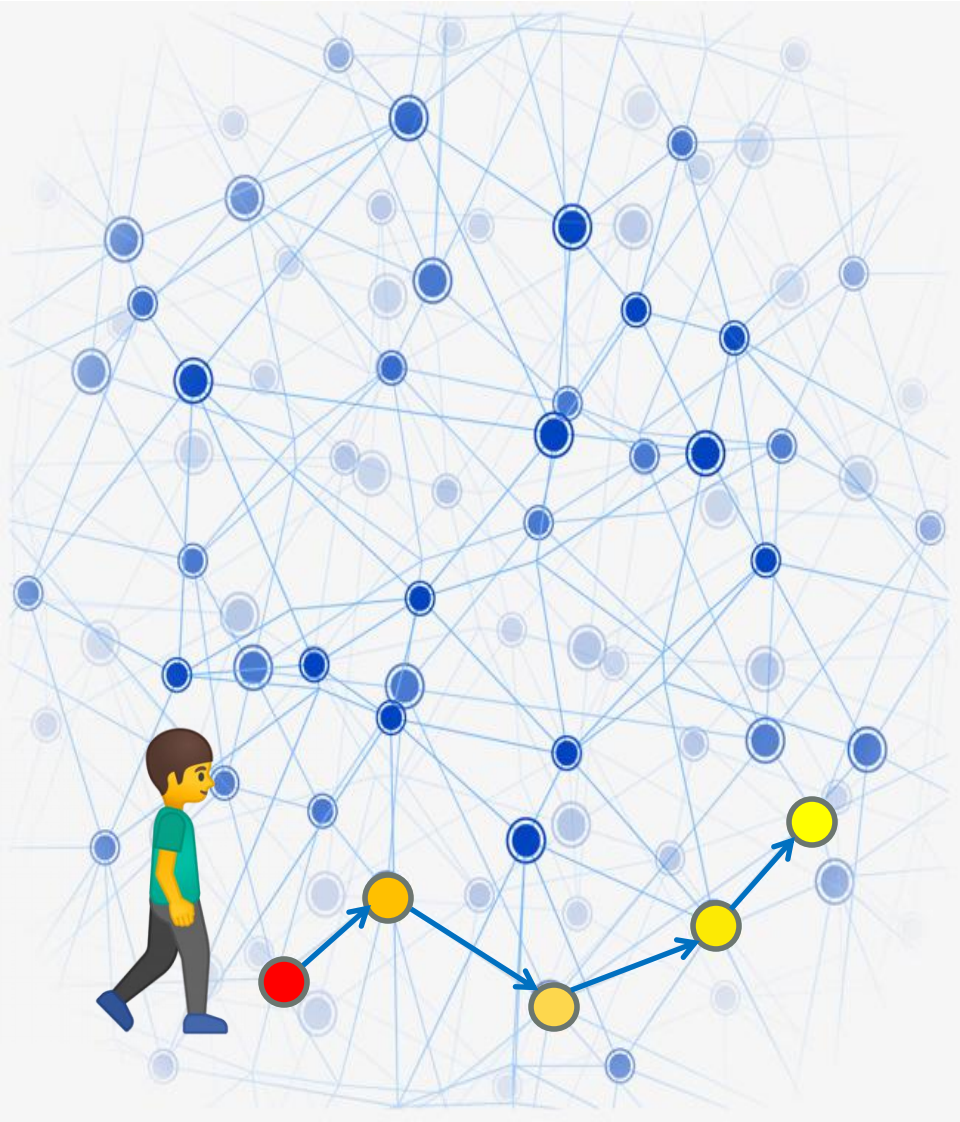
# Network propagation to deal with sparse data

3. Gene information is propagated following the guilt by association principle.
   - A diffusion function computes a score for each node




| | G1 | G2 | G3 | G4 | G5 | G6 | G7 | G8 | G9 | G10 |
|---|---|---|---|---|---|---|---|---|---|---|
| P1 | 0.69999 | 0.6999 | 0.4499 | 0.0999 | 0.0333 | 0.0110 | 0.0035 | 0.0012 | 0.0003 | 0.0003 |

# Network propagation



The process simulates a random walk on a network with restarts, that is mathematically defined by:

$$F_{t+1} = \alpha F_t A + (1-\alpha) F_0$$

restart rate

α controls how far a node's influence can diffuse: the higher α is, further is the distance reached by the smoothing.

# PSNs with smoothed weights

- For each pathway, for example GLUCONEOGENESIS,

1. Build a complete network representing patients as nodes
2. Edges weights are computed by Trending Matching similarity

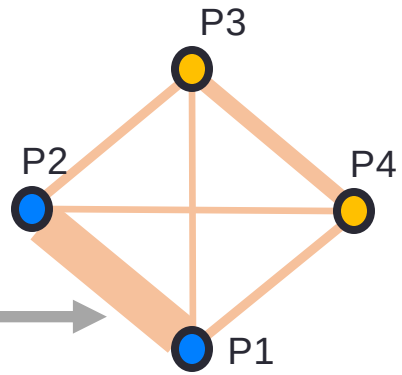
| | G1 | G2 | G3 | G4 | G5 | G6 | G7 | G8 | G9 | G10 | ….. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| P1 | 1 | 1 | 0.9 | 0.8 | 0.7 | 0.6 | 0.5 | 0.4 | 0.3 | 0.3 | . |
| P2 | 0.9 | 0.9 | 1 | 1 | 0.9 | 0.8 | 0.7 | 0.6 | 0.3 | 0.3 | . |
| P3 | 0.5 | 0.5 | 0.7 | 0.9 | 1 | 0.9 | 0.8 | 0.9 | 1 | 0.9 | . |
| P4 | 0.3 | 0.3 | 0.5 | 0.7 | 0.9 | 1 | 0.9 | 1 | 0.9 | 0.9 | . |

GLUCONEOGENESIS

| | G1 | G2 | G3 |
|---|---|---|---|
| P1 | 1 | 1 | 0.9 |
| P2 | 0.9 | 0.9 | 1 |
| P3 | 0.5 | 0.5 | 0.7 |
| P4 | 0.3 | 0.3 | 0.5 |

| Corr | P1 | P2 | P3 | P4 |
|---|---|---|---|---|
| P1 | | | | |
| P2 | | | | |
| P3 | | | | |
| P4 | | | | |

TM(P2,P3)




PSN

# Overview of classification process

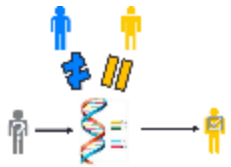
Classify Patients

Data

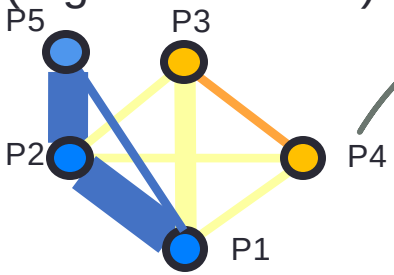
Prediction design

Early Stage vs Late Stage

Smoothed Somatic Mutation Profile



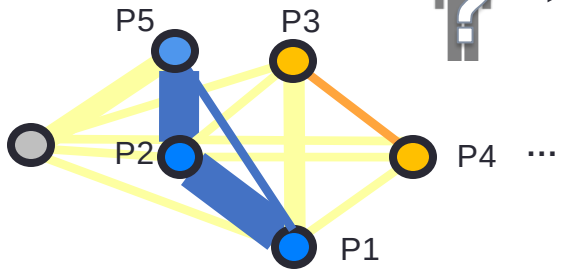
Feature selection (Signature PSNs)




EARLY ☑

Repeat over each leave-one-out CV split

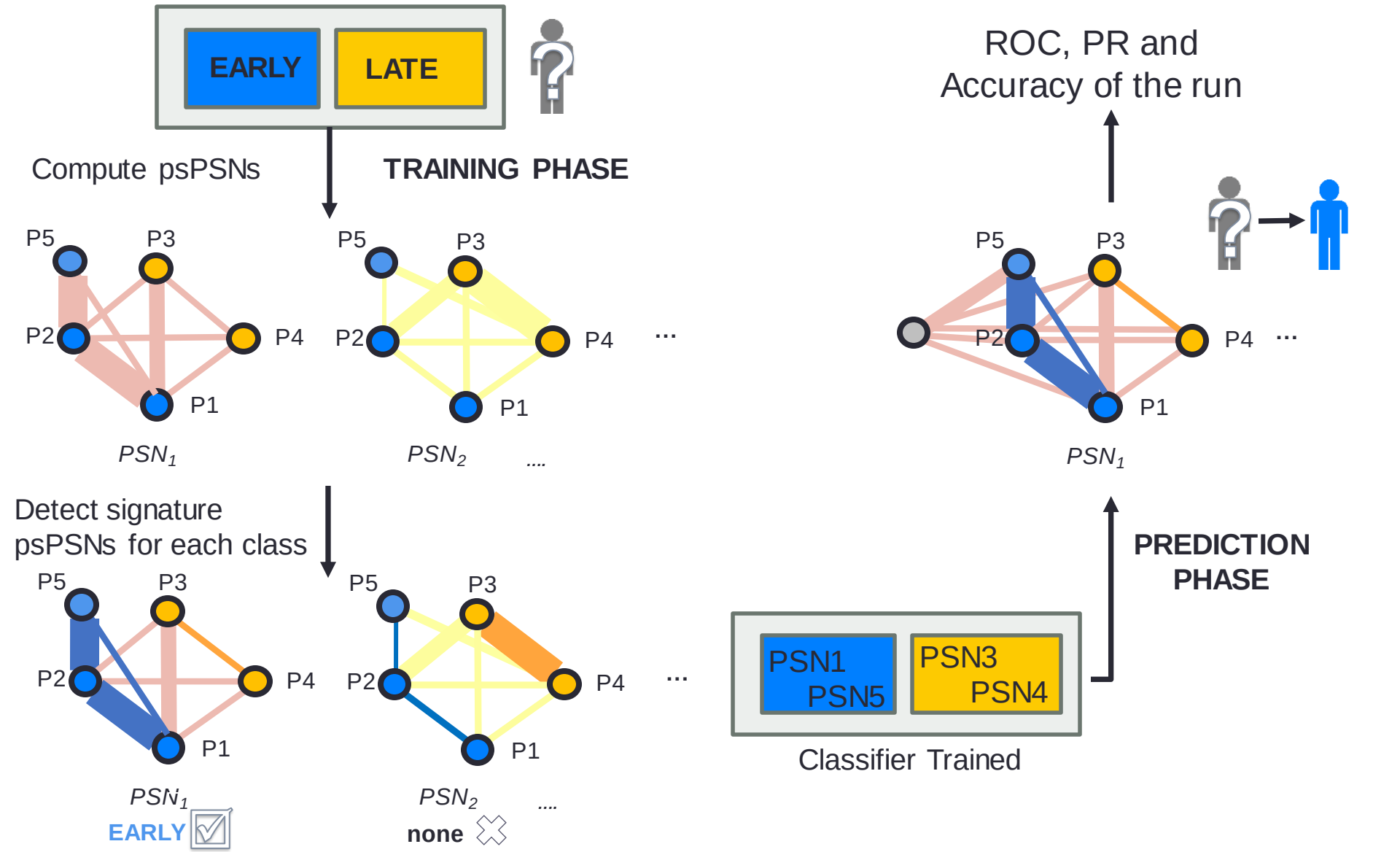
Classification




Evaluate perfomance over splits

Obtain a set of signature features (PSN) per class

# Overview of the classification process



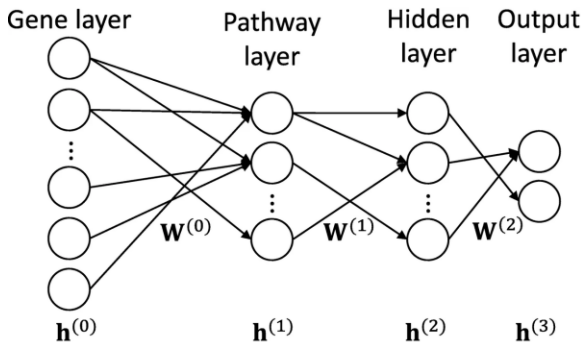
# Tested datasets

- RNA dataset
- Somatic mutation dataset

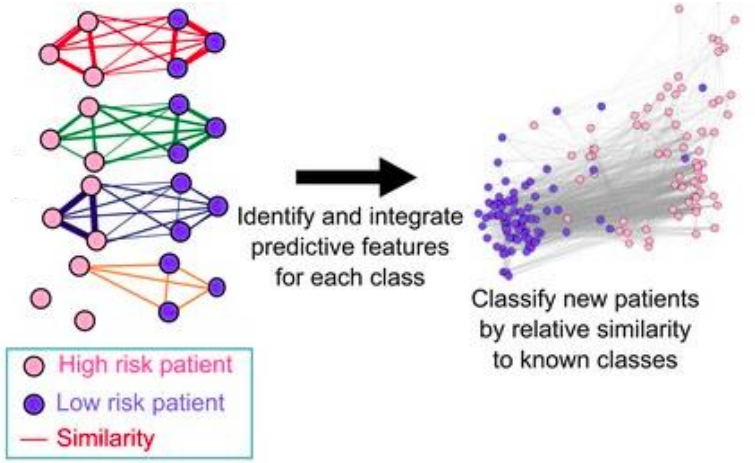
| Cancer type | Dimension |
|---|---|
| LIHC | 7 early, 7 late |
| STAD | 8 early, 13 late |
| KIRC | 24 early, 13 late |
| BLCA | 8 early, 37 late |
| LUSC | 60 early, 13 late |
| ESCA | 91 early, 61 late |

# Competitors

• PASNet[1]: incorporates biological pathways in a Deep Neural Network.

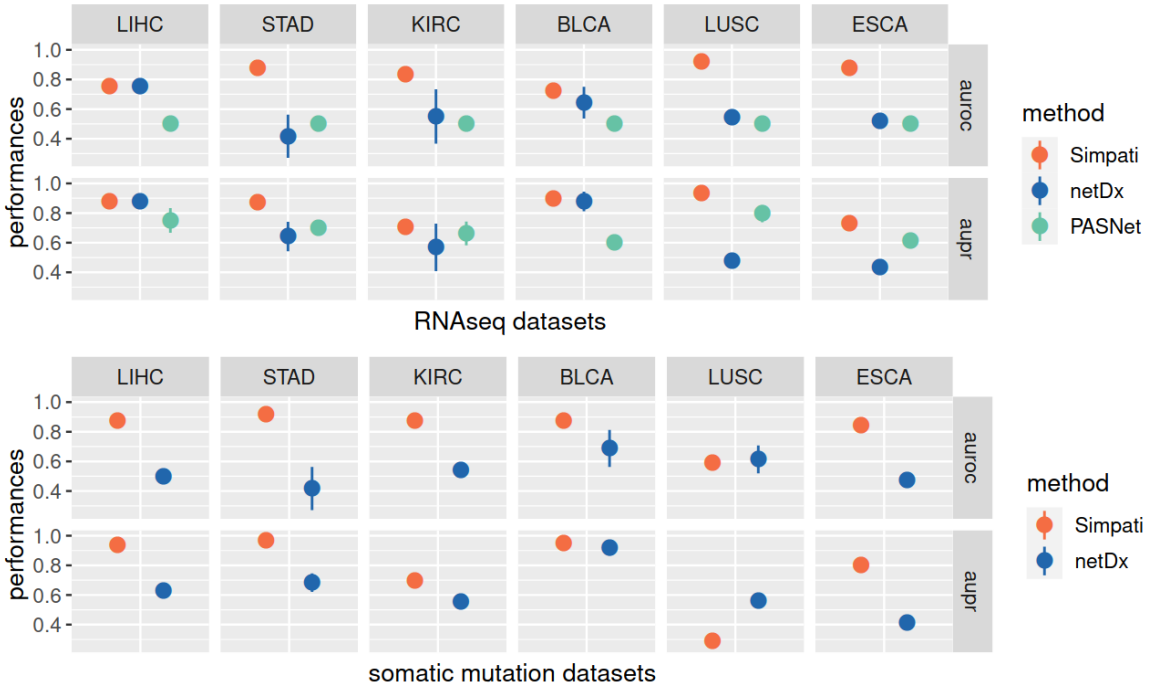



• NetDx[2]: PSN-based tool using external software for feature selection and integration

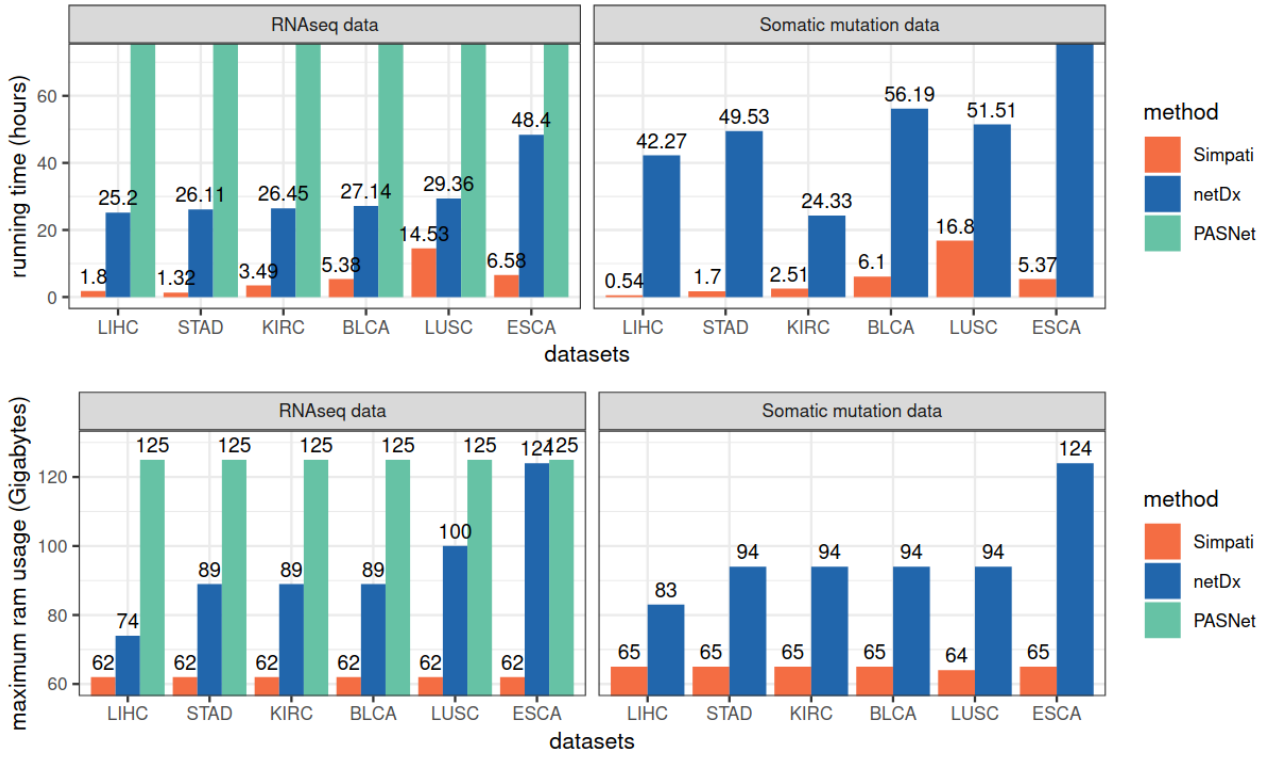



[1] J. Hao, Y. Kim, T. K. Kim, and M. Kang, "PASNet: pathway-associated sparse deep neural network for prognosis prediction from high-throughput data", BMC Bioinformatics,no.19 pp.510, 2018.
[2] S. Pai, et al., "netDx: interpretable patient classification using integrated patient similarity networks", Mol Syst Biol no.15, 2019.

# Classification performance



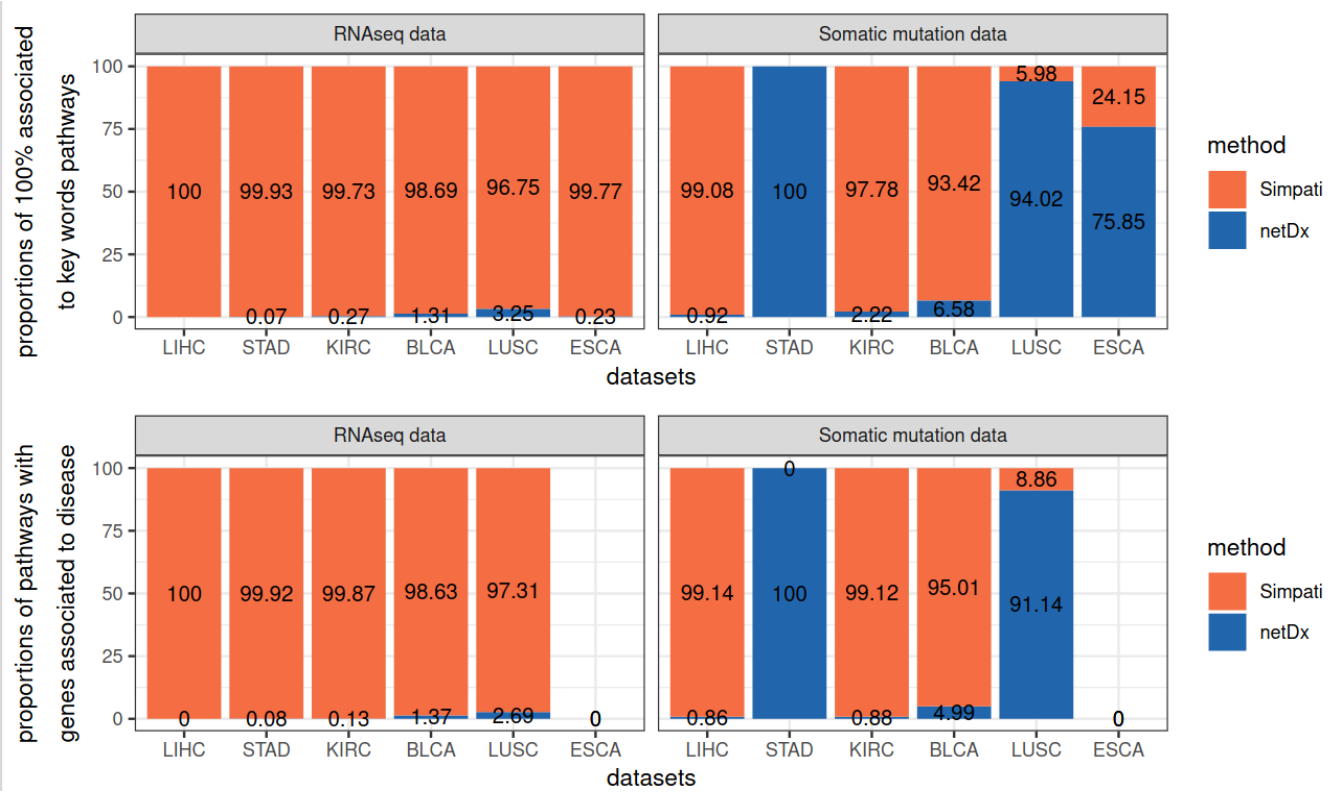
# Computational performance



# Biological validation

Pathways used to classify are functional for the patient class. We compute:

- pathways having occurrences of disease-specific key words in the Human Protein Atlas
- pathways which are associated with the cancer type in DisGeNet.



# Future directions

- Test on larger datasets
- Integrate multiple omics

# Software Availability

- R package https://github.com/InfOmics/Simpati
- Results reproduction https://github.com/LucaGiudice/supplementary-Simpati
- Graphical interface https://github.com/LucaGiudice/propaGUIation

# Thanks!

## Let's meet in Verona at the end of June for BITS2022

Bioinformatics Italian Society

### Welcome to BITS 2022

18th Annual Meeting of the Bioinformatics Italian Society

The Bioinformatics Italian Society is glad to invite you to its 18th Annual Conference.

The conference highlights keynote talks by excellent scientists in bioinformatics and its applications, presentations of state-of-the-art research in bioinformatics and computational biology, and poster sessions on the latest research progress.

BITS 2022
Verona
June 27 - 29

The Congress will take place on 27th-29th of June 2022 in Verona at Polo Santa Marta ( Via Cantarane, 24, 37129 Verona VR ) .

https://bioinformatics.it/bits2022

## International Master Degree in Medical Bioinformatics in the Computer Science Class : http://www.di.univr.it/

lab website: https://infomics.github.io/InfOmics

Contact us (rosalba.giugno@univr.it) to apply for the following scholarships

- Visiting research
- Short teaching at the Master Degree
- International PhD program
- Master Degree international students