Contacts: nico.curti2@unibo.it
Github: https://github.com/Nico-Curti

Artificial Intelligence in Medicine

next AIM

DNetPRO: A network approach for low-dimensional signatures from high-throughput data

Nico Curti[1, 2], Giuseppe Levi[1, 2], Enrico Giampieri[2, 3], Gastone Castellani[4], Daniel Remondini[1, 2]

[1] Department of Physics and Astronomy, University of Bologna
[2] INFN, Bologna
[3] Department of Medical and Surgical Sciences, University of Bologna

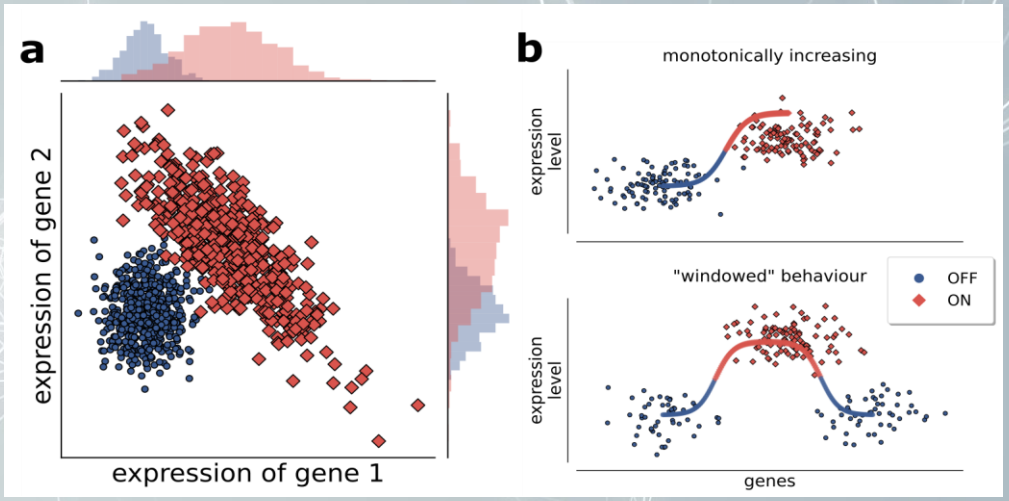
next AIM General Meeting
13/02/2023

Problem Statement

Artificial Intelligence in Medicine

next AIM

- **High-throughput** data (10^3 – 10^5 variables)
- Looking for **low**-dimensional set of observables
- **Gene** or **Protein** expression by an **up/down** regulation
- Features selection is a **critical** step
- Exploration of all feature space is an **NP-hard** problem
- **Few samples** available
- **Ill-posed** problem



Methods that select variables for multi-dimensional signatures based on single-variable performance can have limits in predicting higher-dimensional signature performance. As shown in the Fig. a, in which both variables taken singularly perform poorly, but their performance becomes optimal in a 2-dimensional combination, in terms of linear separation of the two classes.

Curti et al., *Scientific Report*, **2022**

Algorithm design

Artificial Intelligence in Medicine

next AIM

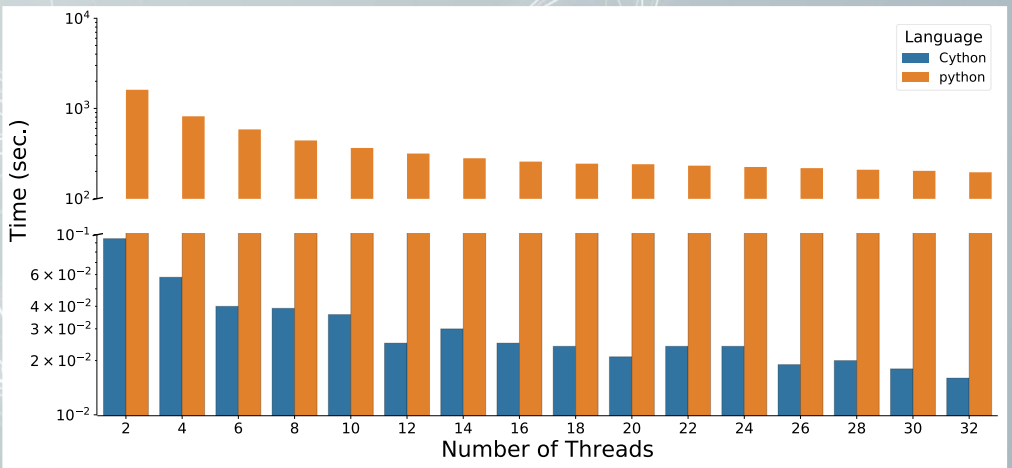
- Evaluation of **all possible couples** of features
- Discriminant classifier for the **couple scoring**
- Creation of the **fully connected network**
- **Thresholding** on network weights
- Extraction of **connected components** as putative signature
- Evaluation of the **signatures**

DNetPRO algorithm

The pseudo-code of the proposed DNetPRO algorithm could be sketched as:

Data: Data matrix (N, S)
Result: List of putative signatures
Divide the data into training and test by a Hold-Out method;
for *couple* ← (*feature_1*, *feature_2*) ∈ *Couples* **do**
  Score estimation using the DA Classifier through Leave-One-Out cross validation;
end
Sorting of the couples in ascending order according to their score;
Threshold over the couples score (*K*-best couples, e.g. based on the statistical distribution of couple performances) in order to obtain at least one connected component;
for *component* ∈ *connected_components* **do**
  if *reduction* **then**
    Iteratively pendant node removal;
  end
  Signature evaluation using the DA Classifier;
end

Algorithm 1: DNetPRO algorithm for Feature Selection.



Source code: https://github.com/Nico-Curti/DNetPRO
Implementation: C++ (backend), Python (frontend)
Algorithm complexity: O(N^2)
Parallelism: naïve parallel

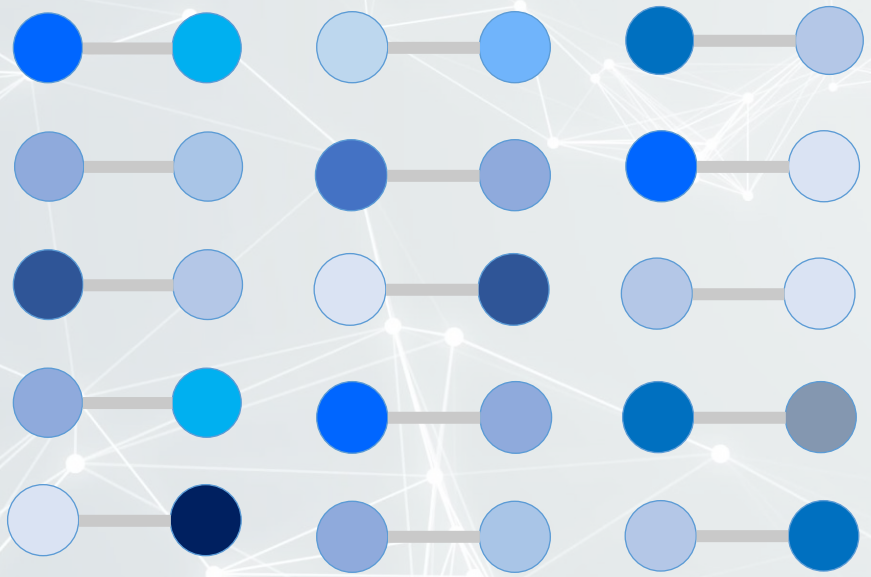
Curti et al., *Scientific Report*, **2022**

Algorithm representation

Artificial Intelligence in Medicine

next AIM

Step 1



Evaluation of **all the possible couples** of features (genes in this case)

Source code: https://github.com/Nico-Curti/DNetPRO
Implementation: C++ (backend), Python (frontend)
Algorithm complexity: $O(N^2)$
Parallelism: naïve parallel

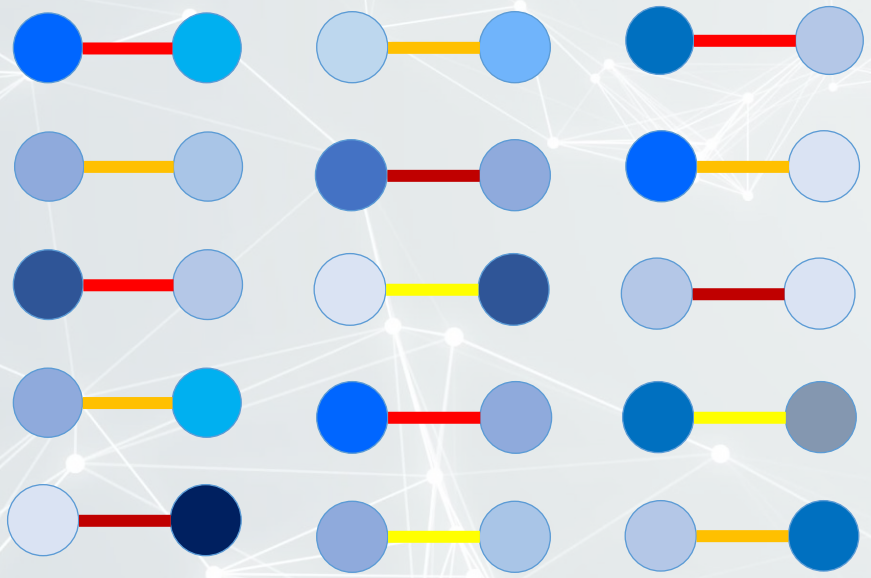
Curti et al., *Scientific Report*, **2022**

Algorithm representation

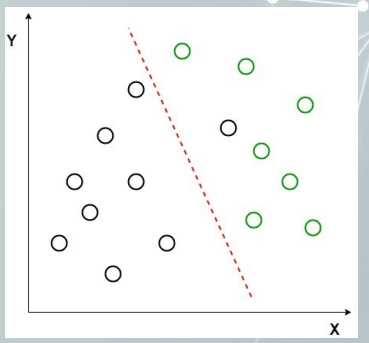
Artificial Intelligence in Medicine

next AIM

Step 2



Discriminant classifier for the **couple scoring**



Source code: https://github.com/Nico-Curti/DNetPRO
Implementation: C++ (backend), Python (frontend)
Algorithm complexity: $O(N^2)$
Parallelism: naïve parallel

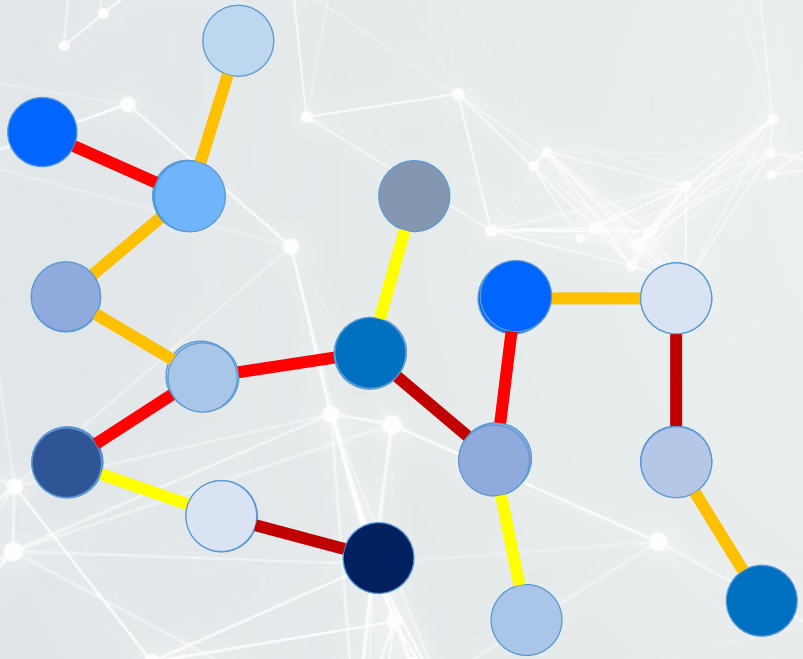
Curti et al., *Scientific Report*, **2022**

Algorithm representation

Artificial
Intelligence in
Medicine

next AIM

Step 3



Creation of the
fully connected network
weighted on couple
performances

Source code: https://github.com/Nico-Curti/DNetPRO
Implementation: C++ (backend), Python (frontend)
Algorithm complexity: $O(N^2)$
Parallelism: naïve parallel

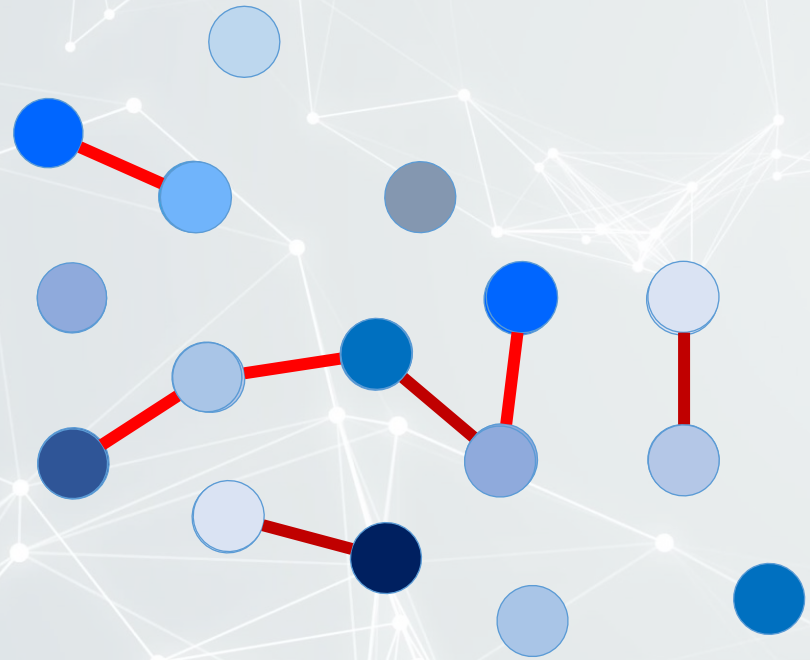
Curti et al., *Scientific Report*, **2022**

Algorithm representation

Artificial Intelligence in Medicine

next AIM

Step 4



Thresholding on network weights according to the performances, i.e., keep only the best couples

Source code: https://github.com/Nico-Curti/DNetPRO
Implementation: C++ (backend), Python (frontend)
Algorithm complexity: $O(N^2)$
Parallelism: naïve parallel

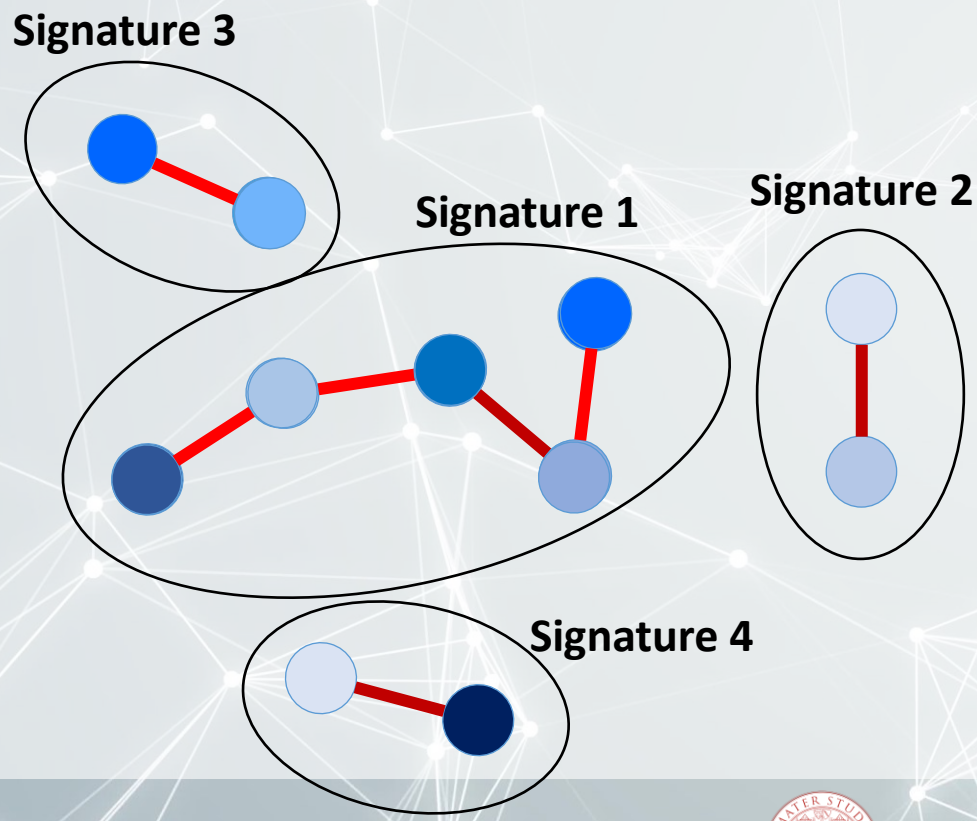
Curti et al., *Scientific Report*, **2022**

Algorithm representation

Artificial Intelligence in Medicine

next AIM

Step 5



Extraction of all the **connected components** as putative signature for the final classification task

Source code: https://github.com/Nico-Curti/DNetPRO
Implementation: C++ (backend), Python (frontend)
Algorithm complexity: $O(N^2)$
Parallelism: naïve parallel

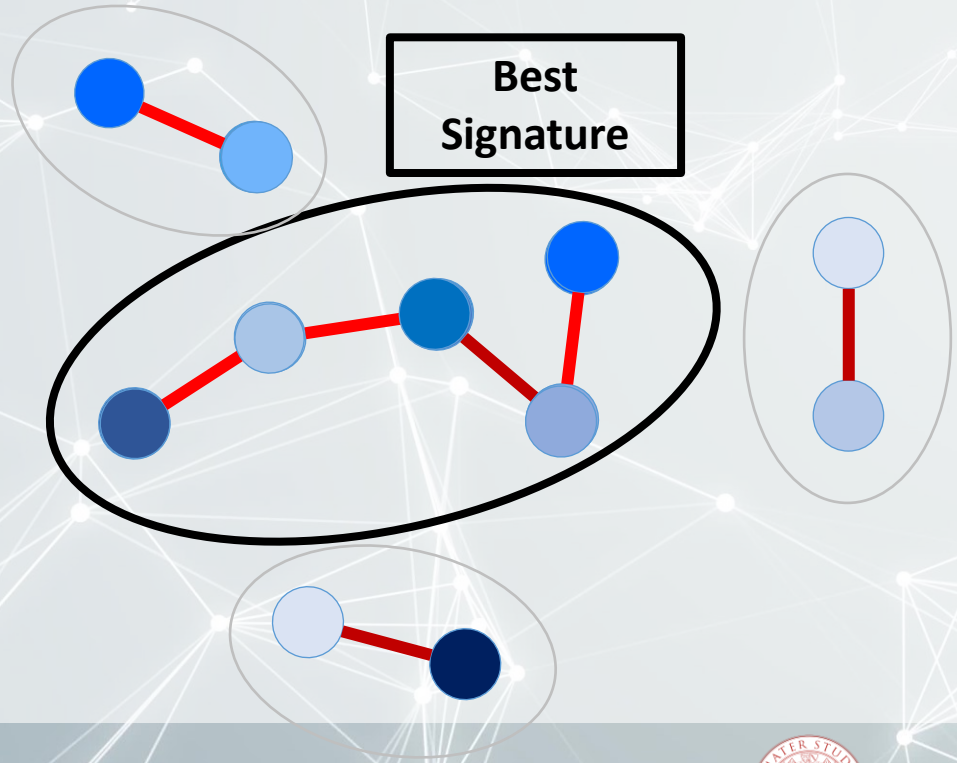
Curti et al., *Scientific Report*, **2022**

Algorithm representation

Artificial Intelligence in Medicine

next AIM

Step 6



Evaluation of the **signatures** in terms of classification efficiency and dimensionality

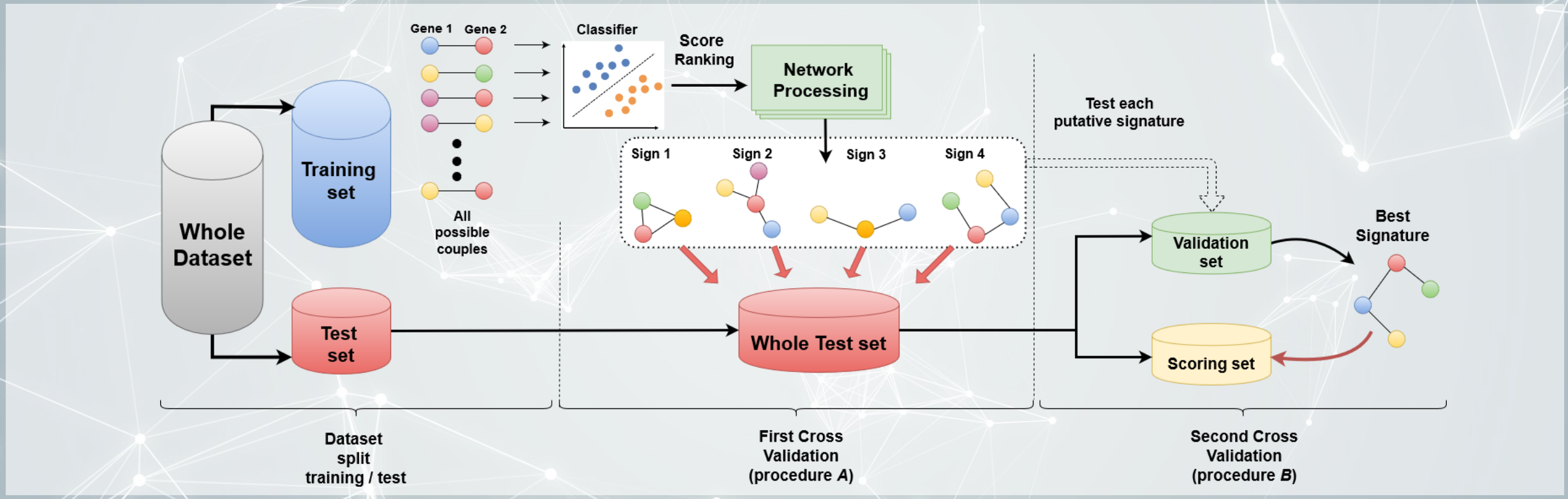
Source code: https://github.com/Nico-Curti/DNetPRO
Implementation: C++ (backend), Python (frontend)
Algorithm complexity: $O(N^2)$
Parallelism: naïve parallel

Curti et al., *Scientific Report*, **2022**

Algorithm representation

Artificial Intelligence in Medicine

next AIM



Source code: https://github.com/Nico-Curti/DNetPRO
Implementation: C++ (backend), Python (frontend)
Algorithm complexity: $O(N^2)$
Parallelism: naïve parallel

Curti et al., *Scientific Report*, **2022**

Application on real data

Artificial Intelligence in Medicine

next AIM

Synapse Dataset

- Application of **TCGA** dataset
- 4 cancer dataset (GBM, KIRC, OV, LUSC)
- 3 omics for each dataset (mRNA, miRNA, RPPA)

Cancer	mRNA	miRNA	Protein	Number of samples
GBM	AgilentG4502A	H-miRNA_8x15k	RPPA	
	17 814	533	[a]	210
KIRC	HiseV2	GA+Hiseq	RPPA	
	20 530	1 045	166	243
OV	AgilentG4502A	H-miRNA_8x15k	RPPA	
	17 814	798	165	379
LUSC	HiseqV2	GA+Hiseq	RPPA	
	20 530	1 045	174	121

Curti et al., *Scientific Report*, **2022**

Cytokine Dataset

- 289 patients
- CTL vs MCI vs AD

Boccardi et al., *JAD*, **2019**

Bovine Dataset

- 12k genes
- 15 samples (PP vs NP vs NN)

Malvisi et al., *Animals*, **2020**

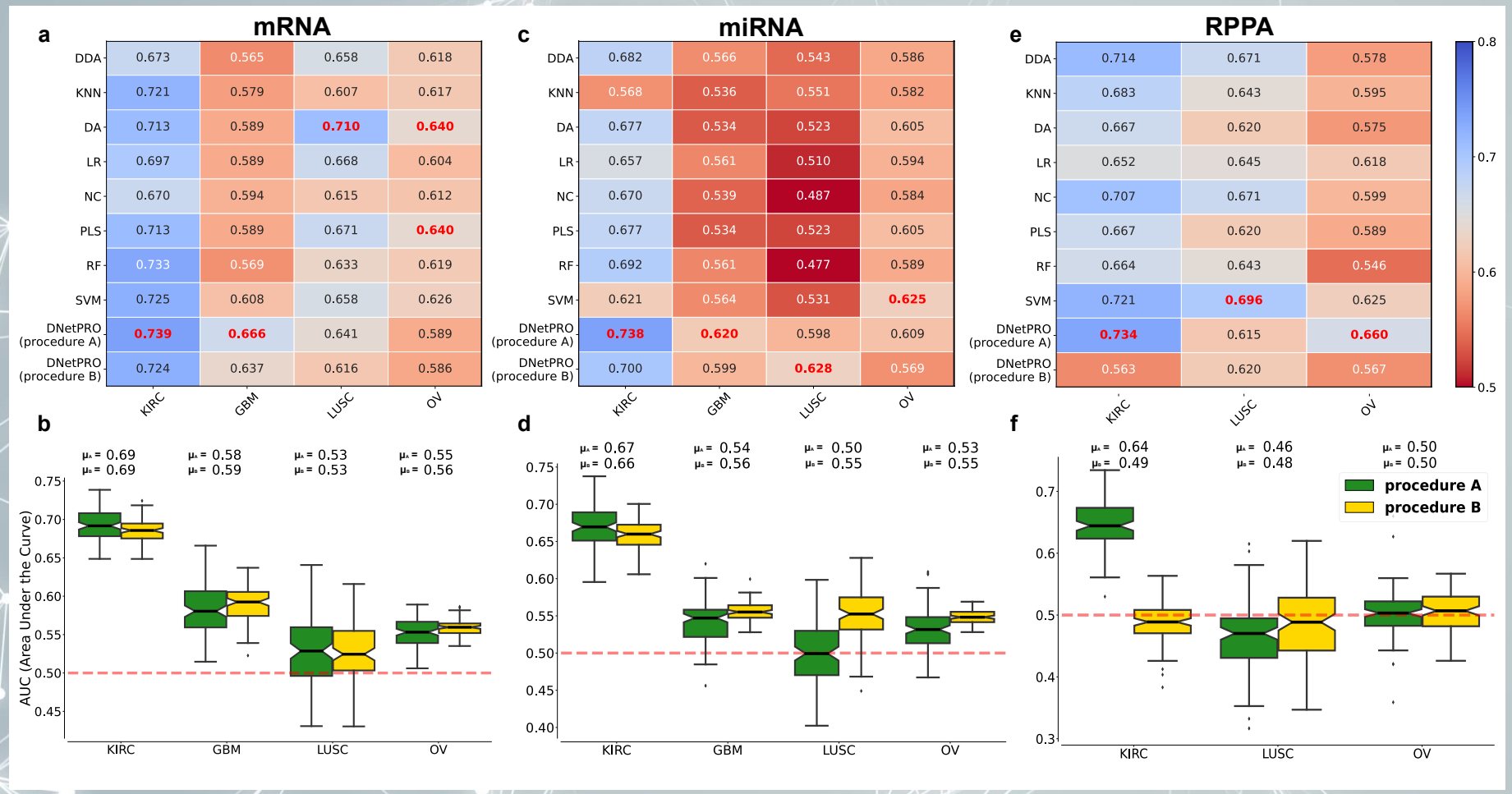
INFN Istituto Nazionale di Fisica Nucleare

Application on real data

Synapse Dataset

Artificial Intelligence in Medicine

next AIM



Benchmark with a set of state-of-the-art classifiers.

Yuan et al., *Nature Biotechnology*, **2014**

Results obtained by the DNetPRO on the mRNA, miRNA, and RPPA samples related to the four cancer types in the Synapse dataset.
Comparison of the DNetPRO results with the methods used in the work of Yuan et al., *Nature Biotechnology* (**2014**), in terms of the maximum AUC value obtained on a 10-fold cross-validation procedure.
Distributions of the AUC values related to each analyzed dataset.

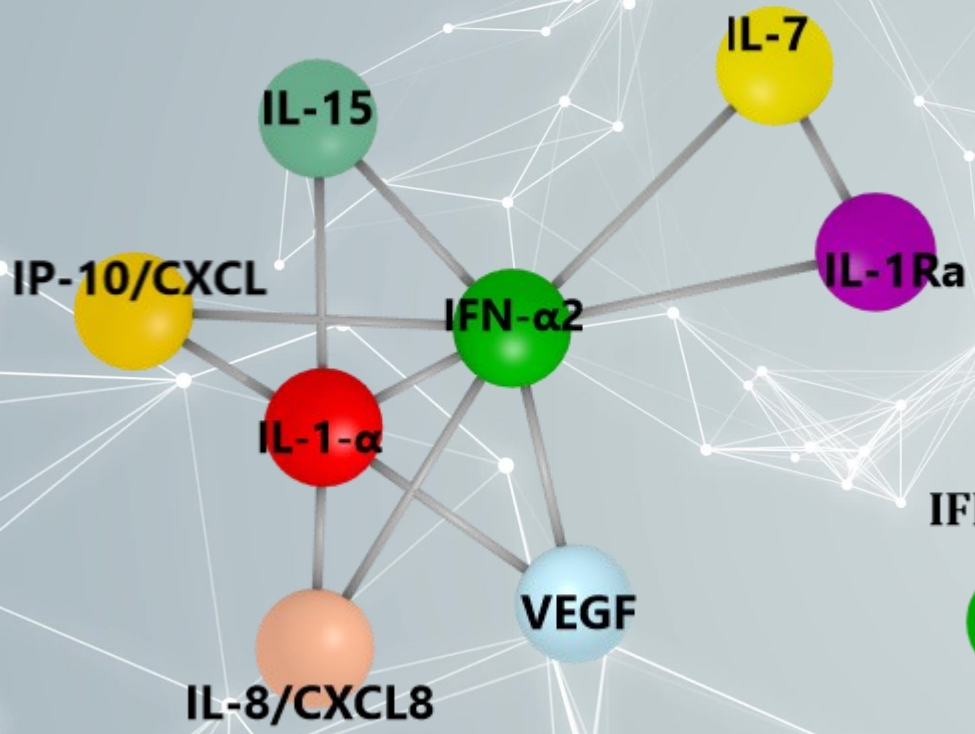
Curti et al., *Scientific Report*, **2022**

Application on real data

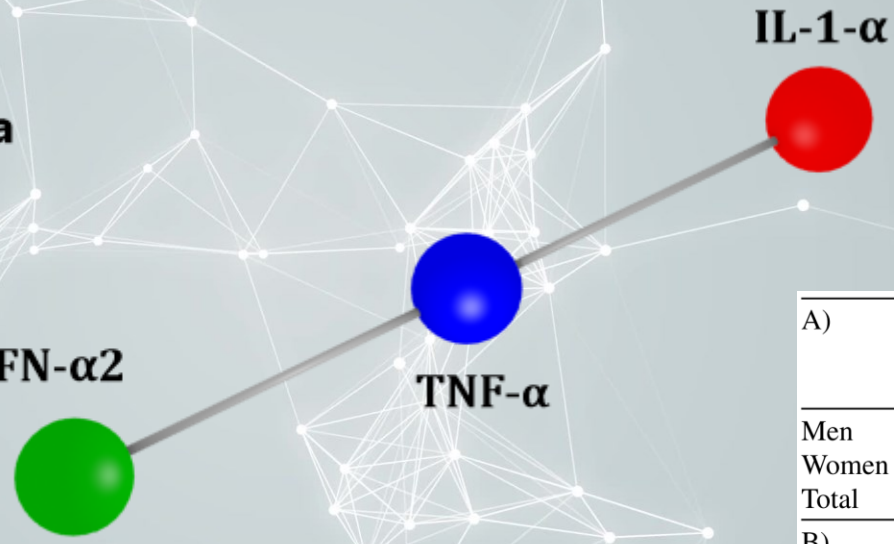
Cytokine Dataset

Artificial Intelligence in Medicine

next AIM



Signature CTL vs MCI



Signature CTL vs AD

A)

	Accuracy	Sensitivity	Specificity
	AD versus HC	*AD*	*HC*
Men	16/25 (64.00%)	8/12 (66.67%)	8/13 (61.54%)
Women	33/48 (68.75%)	27/38 (71.05%)	6/10 (60.00%)
Total	47/72 (65.24%)	36/54 (66.66%)	11/18 (61.11%)

B)

	Prediction	Sensitivity	Specificity
	MCI as non-HC	*MCI*	*HC*
Men	15/26 (57.69%)	15/26 (57.69%)	24/36 (66.67%)
Women	41/47 (87.23%)	41/47 (87.23%)	23/51 (45.09%)
Total	62/73 (84.93%)	62/73 (84.93%)	36/87 (41.38%)

Network rendering using:
https://github.com/Nico-Curti/BlendNet

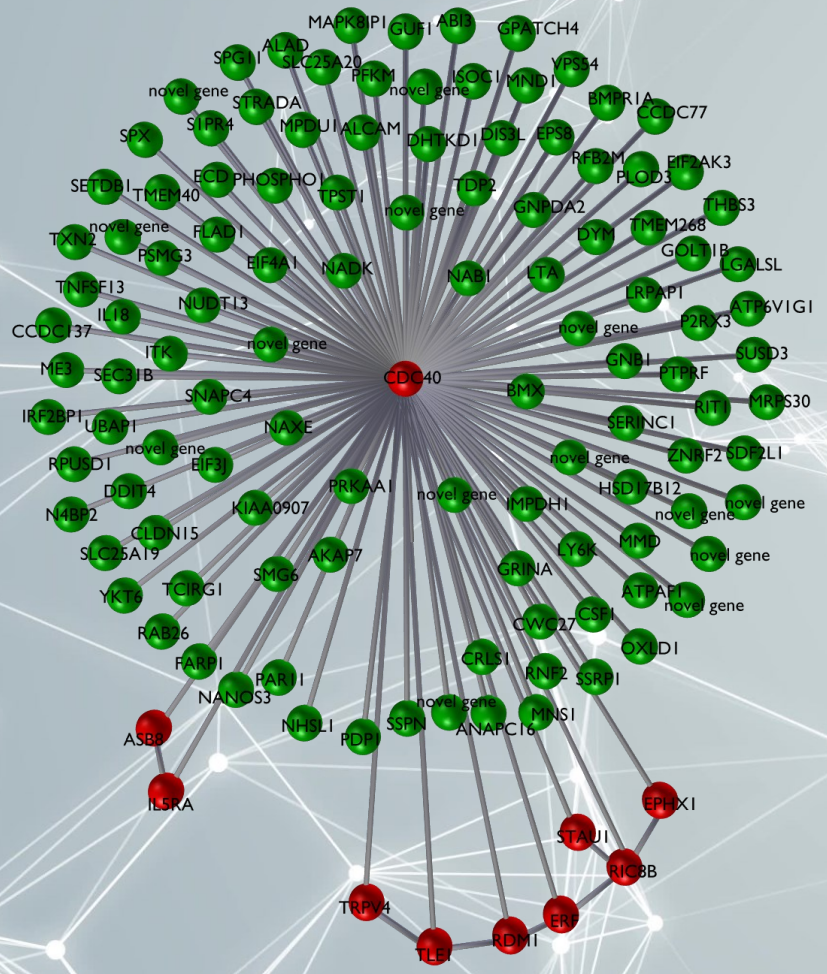
Boccardi et al., *JAD*, **2019**

Application on real data

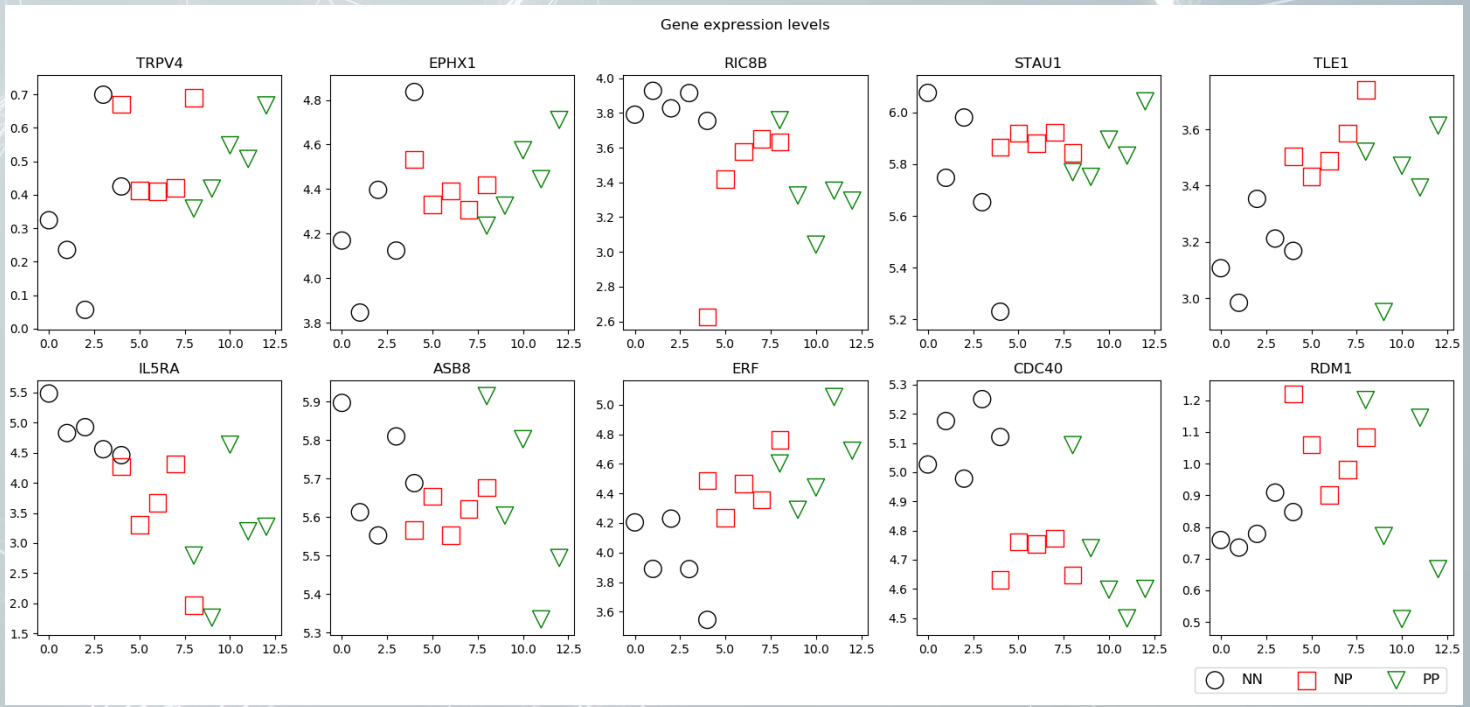
Bovine Dataset

Artificial Intelligence in Medicine

next AIM



Best signature of 123 probes with 100% accuracy



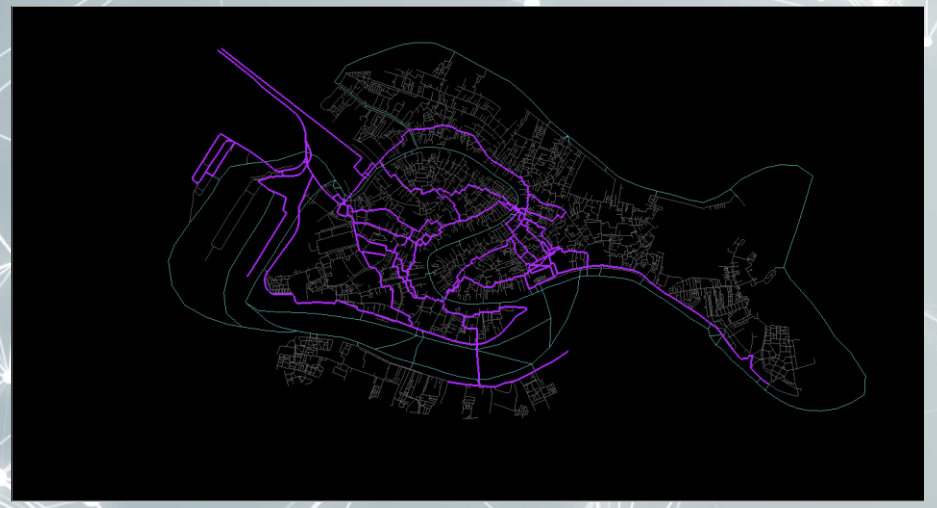
Network rendering using:
https://github.com/Nico-Curti/BlendNet

Malvisi et al., *Animals*, **2020**

Application on no-Bio data

Artificial Intelligence in Medicine

next AIM



Network pedestrian mobility

Unraveling pedestrian mobility on a road network using ICTs data during great tourist events.

- Reconstruction of the pedestrian mobility network
- Same network analysis of DNetPRO algorithm
- Roads play the role of genes
- Couples weighted according to the mobility score

Mizzi et al., *EPJ Data Science*, **2018**

Conclusion

Artificial
Intelligence in
Medicine

next AIM

- Combinatorial approach to explore the entire feature space
- Easy interpretation of the results
- Fast computation in parallel architectures
- Large scalability on high-throughput data
- Algorithm tailored to omics data but applied also to no-biological data

Artificial
Intelligence in
Medicine

next AIM

Thank you for the attention