Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EXCELENCIA SEVERO OCHOA

Interpretability in Artificial Intelligence applications for rare diseases

Davide Cirillo

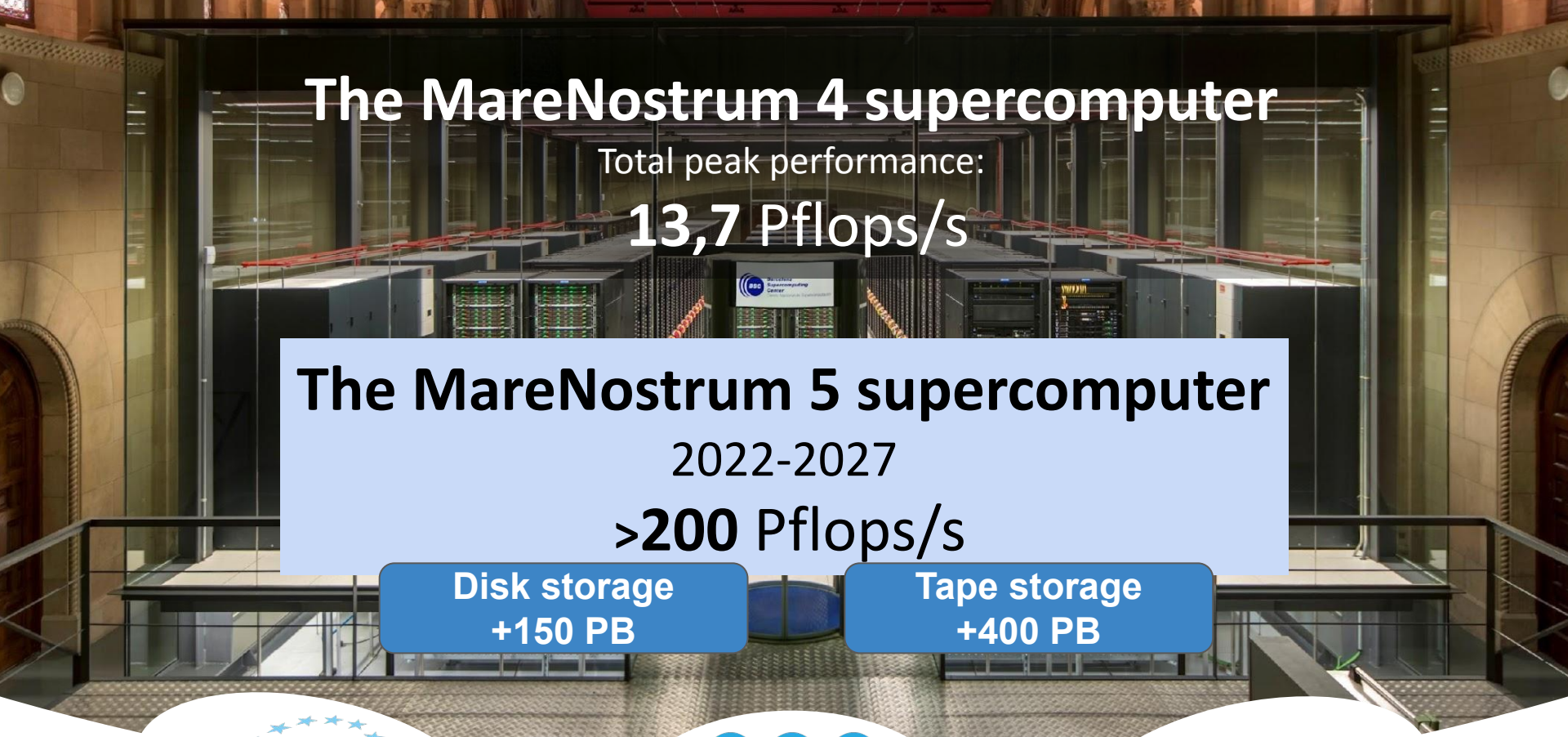
Machine Learning for Biomedical Research

Life Sciences Department

04/05/2022

Barcelona Supercomputing Center

Barcelona Supercomputing Center
Centro Nacional de Supercomputación



The MareNostrum 4 supercomputer

Total peak performance:

13,7 Pflops/s

The MareNostrum 5 supercomputer

2022-2027

>200 Pflops/s

Disk storage
+150 PB

Tape storage
+400 PB

PRACE

Access: prace-ri.eu/hpc-access

R E S

RED ESPAÑOLA DE SUPERCOMPUTACIÓN

Access: bsc.es/res-intranet

BSC Barcelona Supercomputing Center
Centro Nacional de Supercomputación

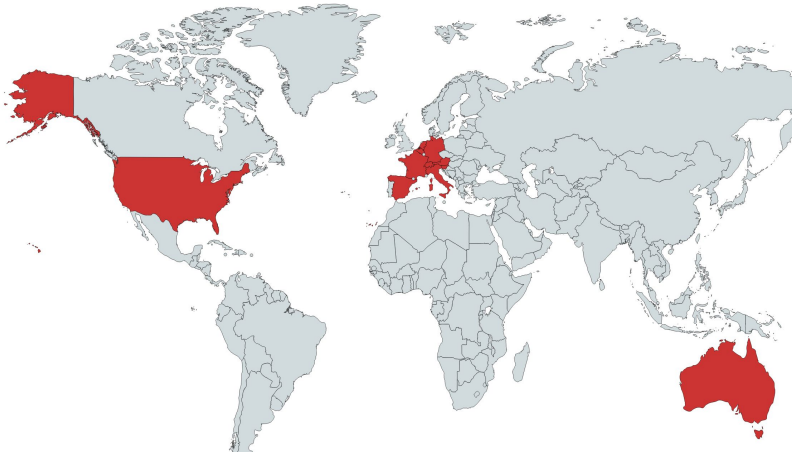
Individualized PaediatricCure

European Commission | Horizon 2020 European Union funding for Research & Innovation

Ref. 826121

TECHNIKON · IBM Research · Baylor College of Medicine · institut Curie · Technische Universität Darmstadt · Ghent University · Barcelona Supercomputing Center · XLAB · Princess Máxima Center pediatric oncology · MPIMG · Heidelberg University Hospital · IGTP · ALACRIS Theranostics GmbH · University of Zurich · LMU Ludwig-Maximilians-Universität München · National Research Council of Italy · Children's Medical Research Institute · Jeans for Genes · dkfz.

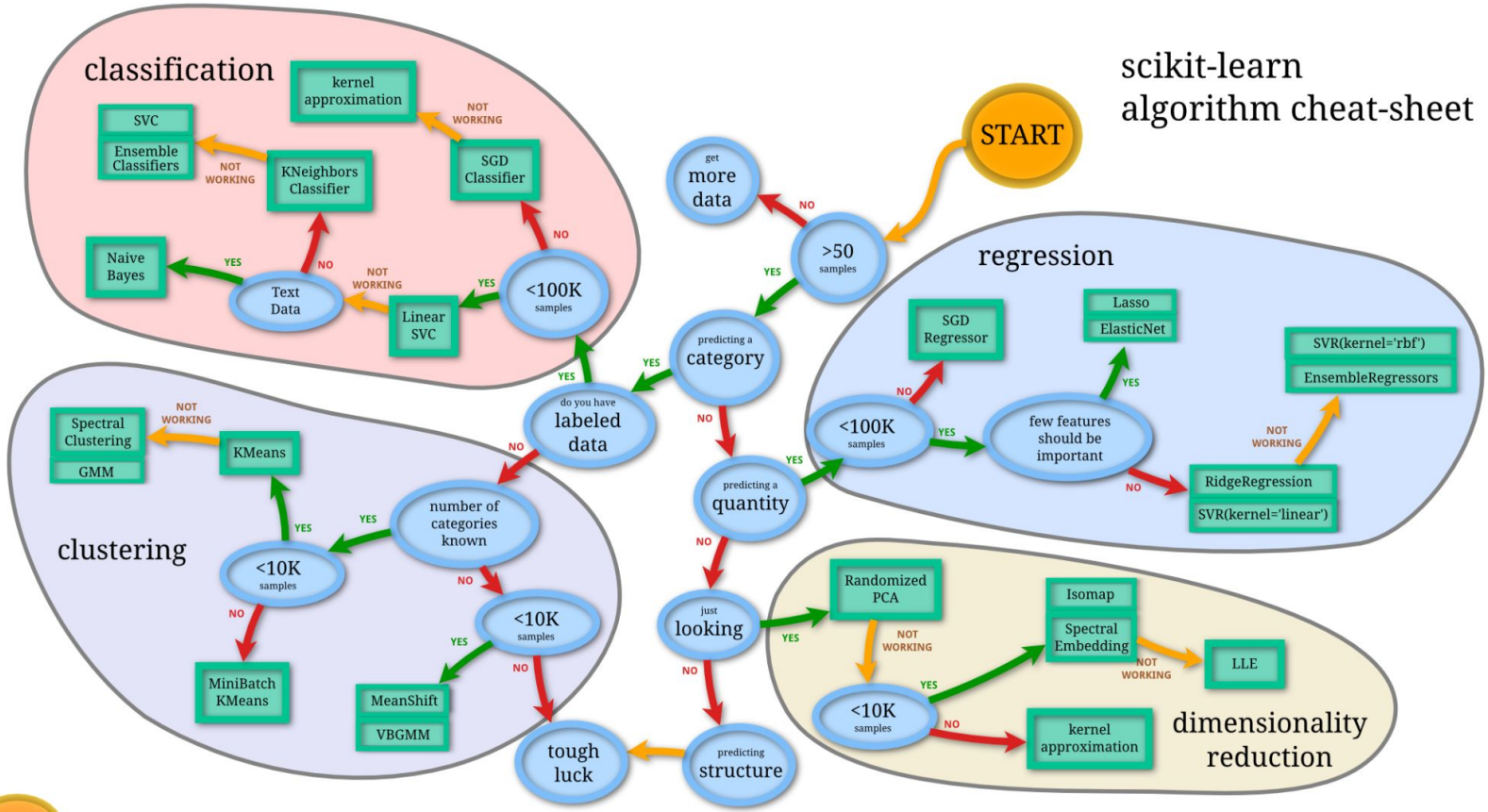
21 partners from 3 continents



- **Cloud-based platform** to share harmonized **data on paediatric cancers**.
- **Predictive models** verified through **clinical trials and preclinical models**.
- **Personalized treatment recommendations** for paediatric cancer patients.

BSC Barcelona Supercomputing Center
Centro Nacional de Supercomputación

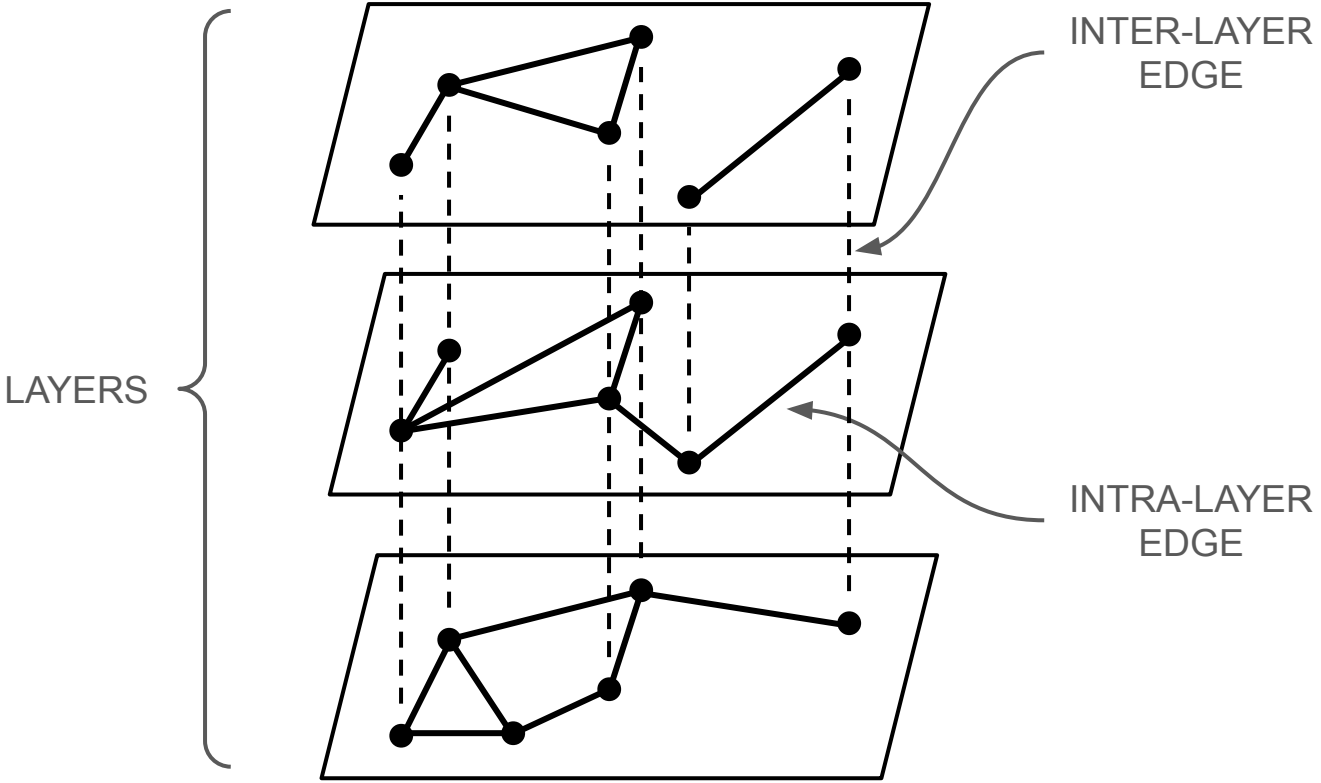
Rare diseases



source : https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

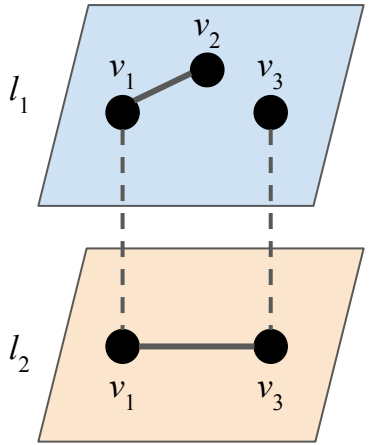
Multilayer networks



Kivelä et al. *J Complex Networks.* 2014

Multilayer network formalization

$$M = (V_M, E_M, V, L)$$



	v_1	v_2	v_3
l_1	1	1	1
l_2	1	0	1

	(l_1,v_1)	(l_1,v_2)	(l_1,v_3)	(l_2,v_1)	(l_2,v_2)	(l_2,v_3)
(l_1,v_1)	0					
(l_1,v_2)	1	0				
(l_1,v_3)	0	0	0			
(l_2,v_1)	1	0	0	0		
(l_2,v_2)	0	0	0	0	0	
(l_2,v_3)	0	0	1	1	0	0

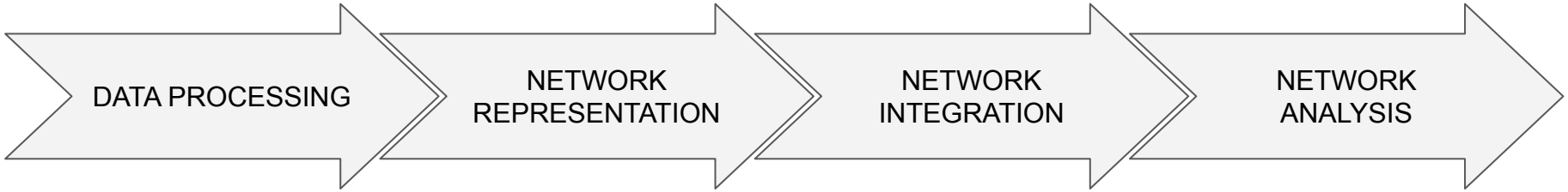
$$V = \{v_1, v_2, v_3\}$$
$$L = \{l_1, l_2\}$$

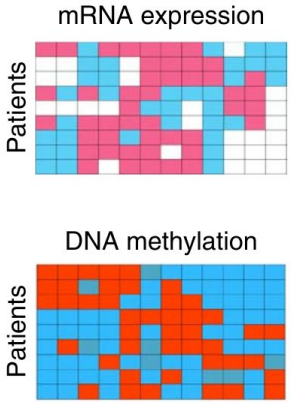
$$V_M \subseteq V \times L$$

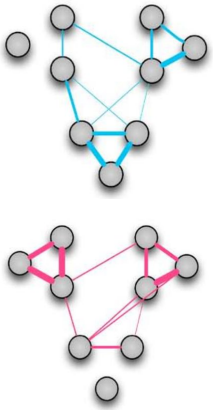
$$E_M \subseteq V_M \times V_M$$

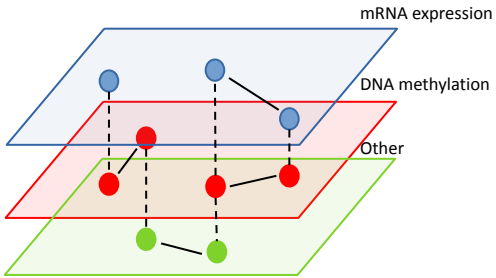
Núñez-Carpintero et al. *iScience*. 2021;24(4):102365.

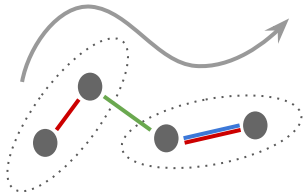
Multilayer network analysis of biomedical information









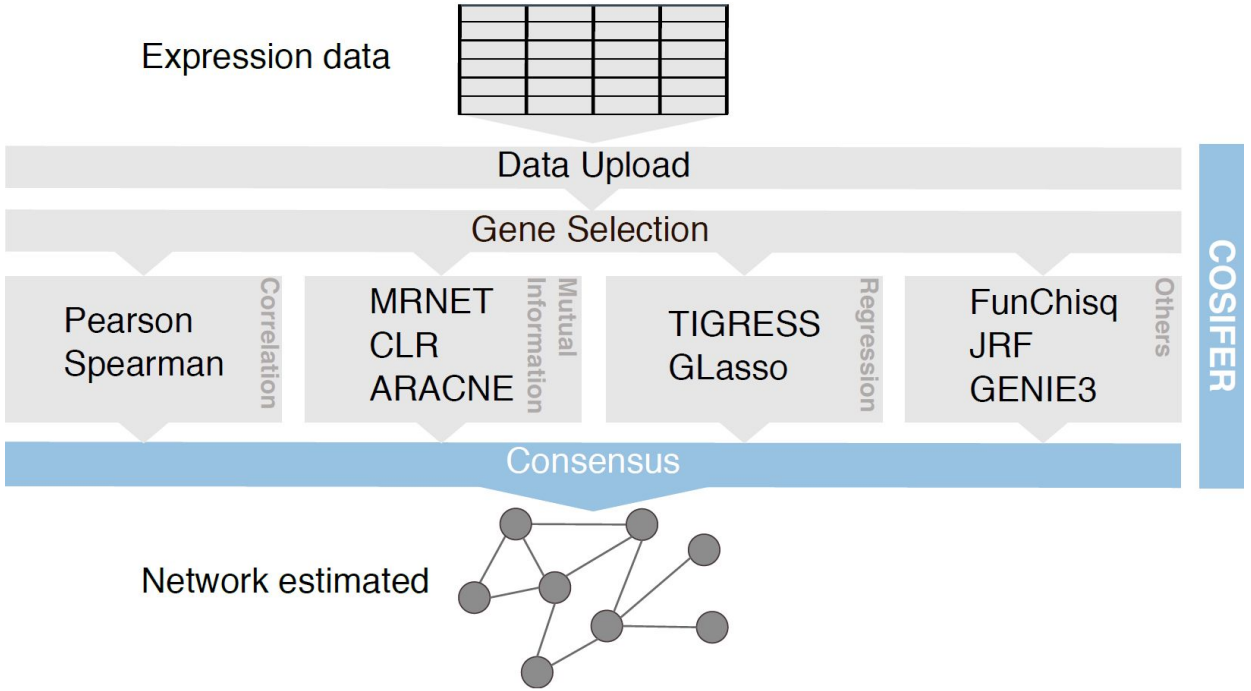


Network inference with COSIFER

COSIFER

Consensus inference of molecular networks

IBM

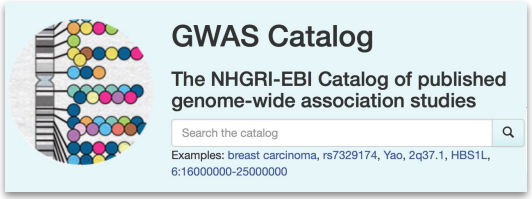


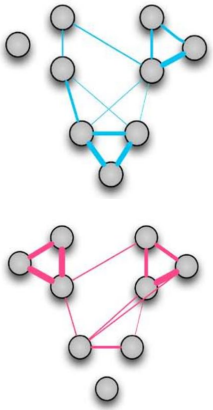
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

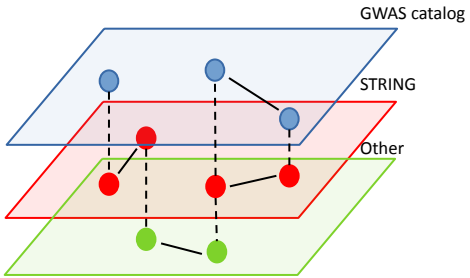
zurich.ibm.com/cosifer

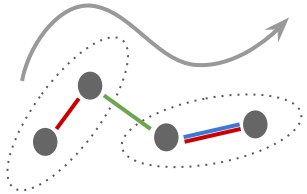
Multilayer network analysis of biomedical information

KNOWLEDGE BASES → NETWORK REPRESENTATION → NETWORK INTEGRATION → NETWORK ANALYSIS

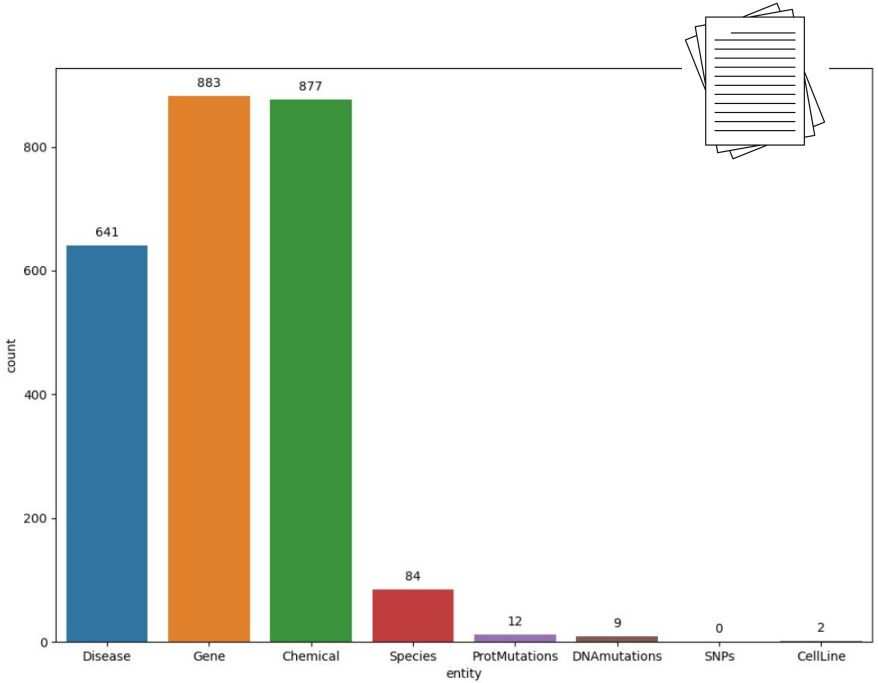


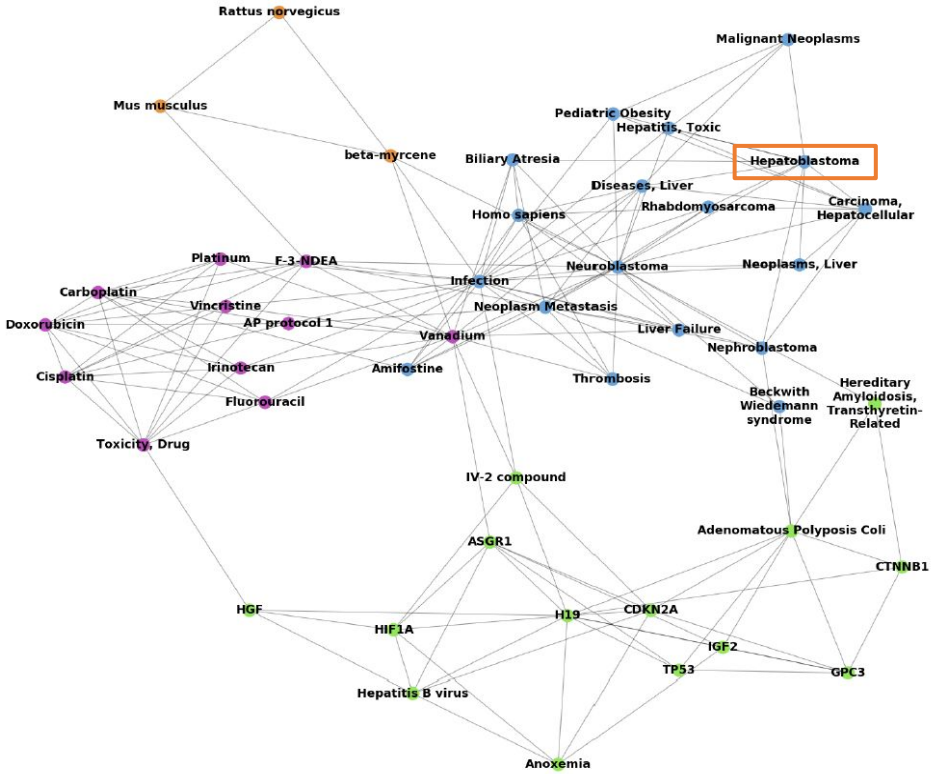






Hepatoblastoma





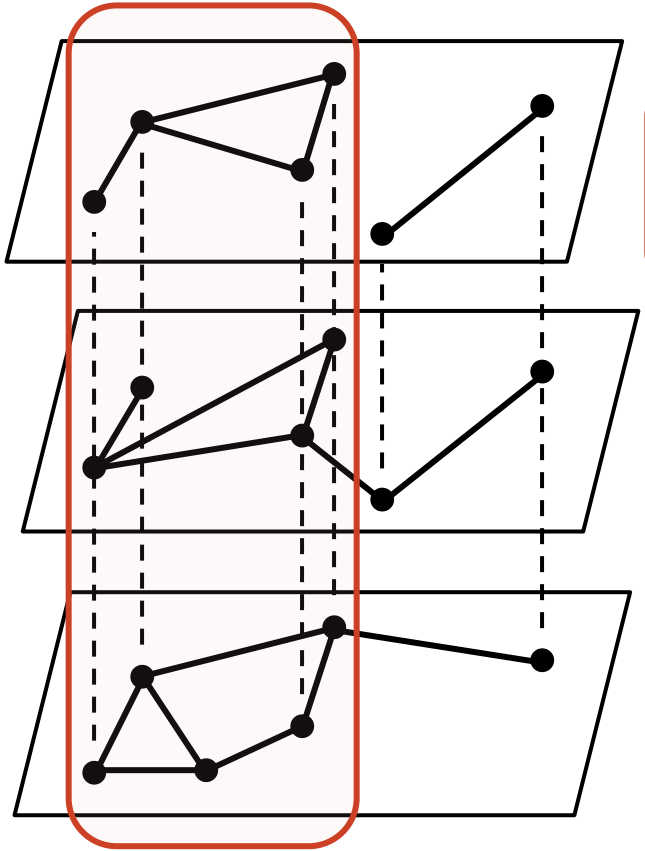
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

in collaboration with IBM

Multilayer community trajectories to study rare diseases

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Multilayer community detection

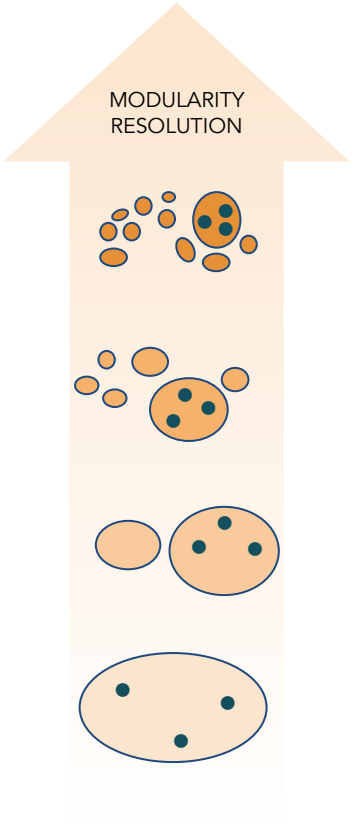


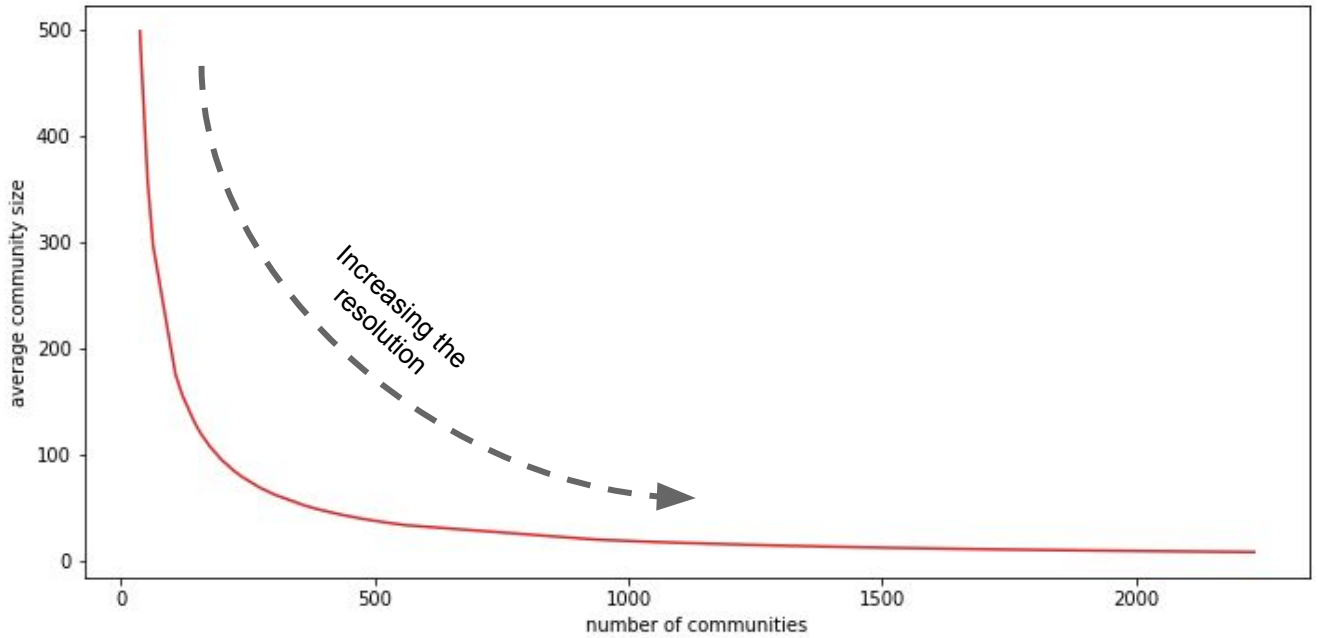
Multilayer community

$$\text{maximize} \sum_{g} \mathcal{Q}_{\gamma}\left(X^{(g)}, \mathbf{c}\right)$$

Newman & Girvan. *Phys Rev E.* 2004
Blondel et al. *J Stat Mech*. 2008
Didier et al. *PeerJ*. 2015
Didier et al. *F1000Res*. 2018

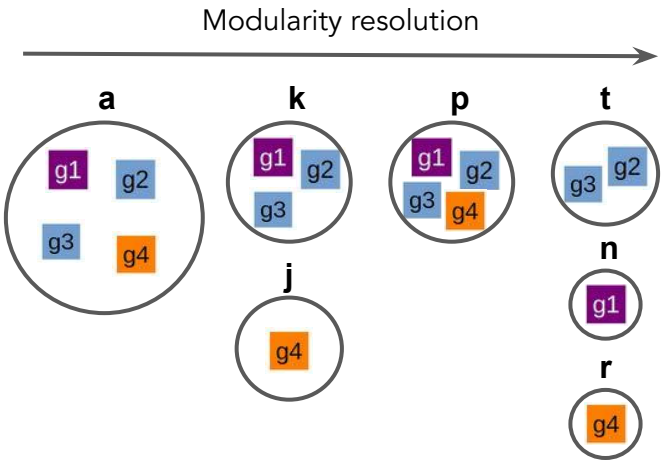
Modularity resolution

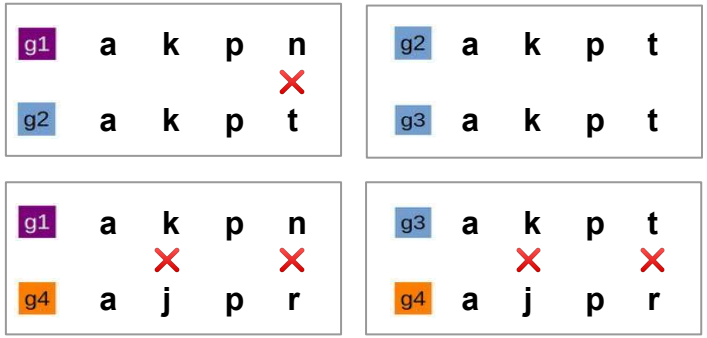




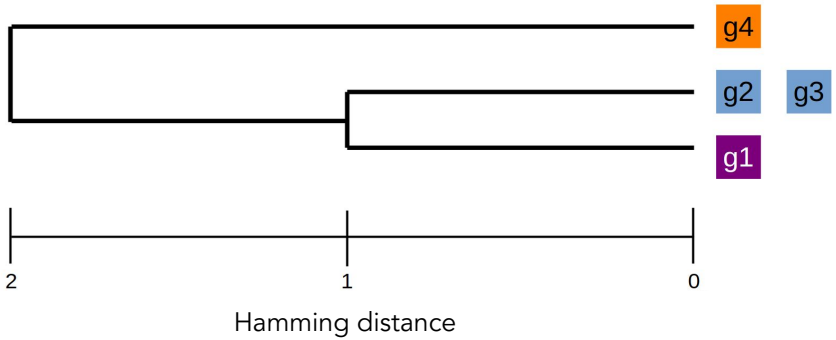
Fortunato & Barthélemy. *PNAS.* 2007
Arenas et al. *N J Phys*. 2008

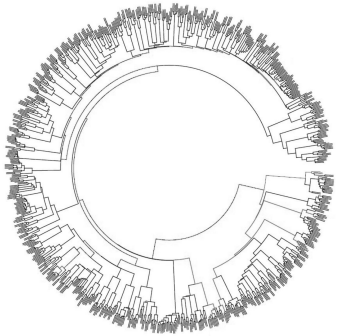
Multilayer community trajectories





	g1	g2	g3	g4
g1	0	1	1	2
g2	1	0	0	2
g3	1	0	0	2
g4	2	2	2	0



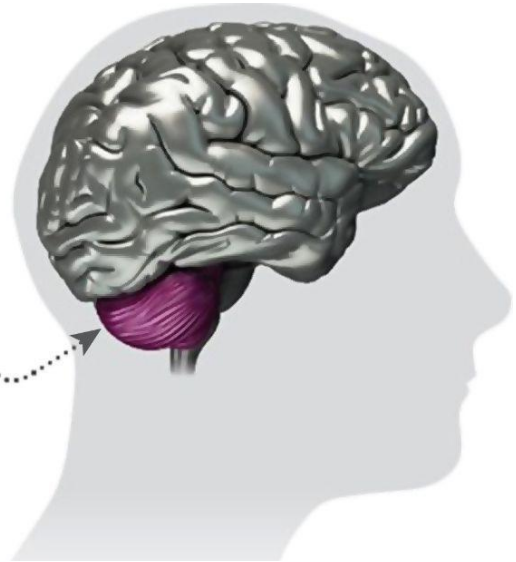


- distance in the tree
- community composition

Nuñez et al. *Cell iScience*. 2021

Medulloblastoma

Molecular subtype	Wnt group	Shh group	Group 3	Group 4
Prognosis	Very good	Good in infant and intermediate in others	Poor	Intermediate
Main signaling pathway	Wnt	Shh, PI3K	TGF-β, photoreceptor/ GABAergic	NF-κB
Metastasis	Rare	Uncommon	Very frequent	Frequent
Characteristic feature	*CTNNB1* mutation	*SMO/PTCH/SUFU* mutation	*MYC* amplification	*CDK6* amplification
***MYC* status**	MYC^{+}	$MYCN^{+}$	MYC^{+++}	Minimal *MYC/MYCN*

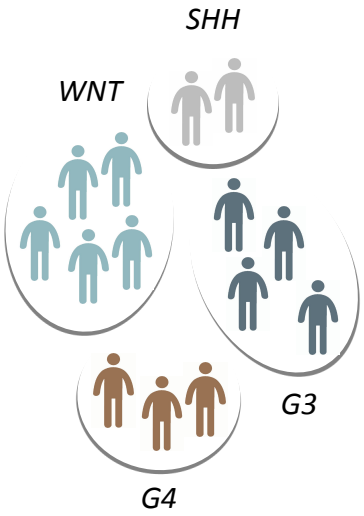


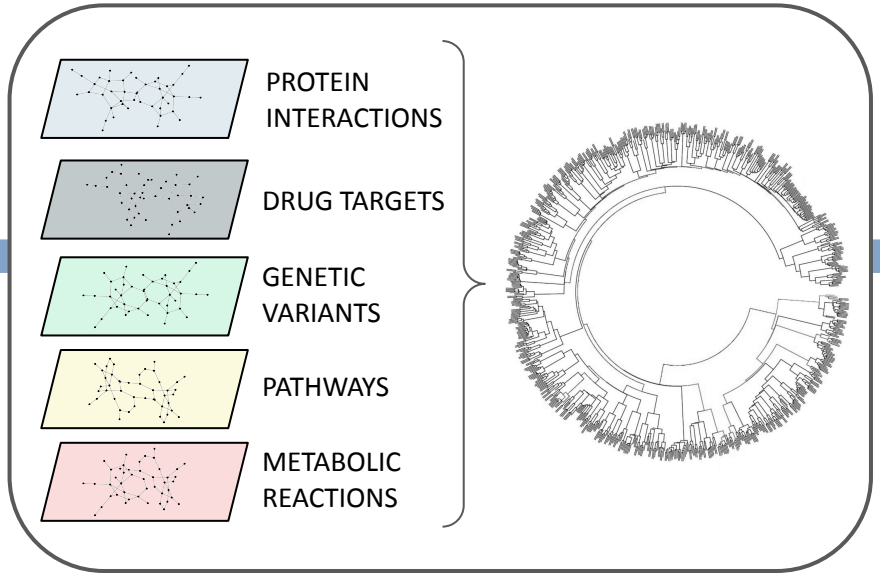
Kumar et al. *Trends in pharmacological sciences* vol. 38,12 (2017): 1061-1084. doi:10.1016/j.tips.2017.09.002

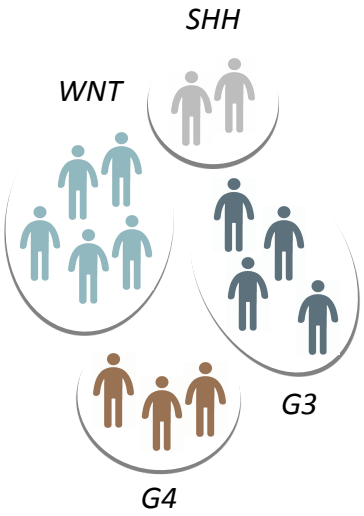
Patient stratification in medulloblastoma

PROTEOGENOMIC DATA
(~14,000 altered genes per patient)

MINIMAL SET OF GENES
(~1,800 altered genes per patient)





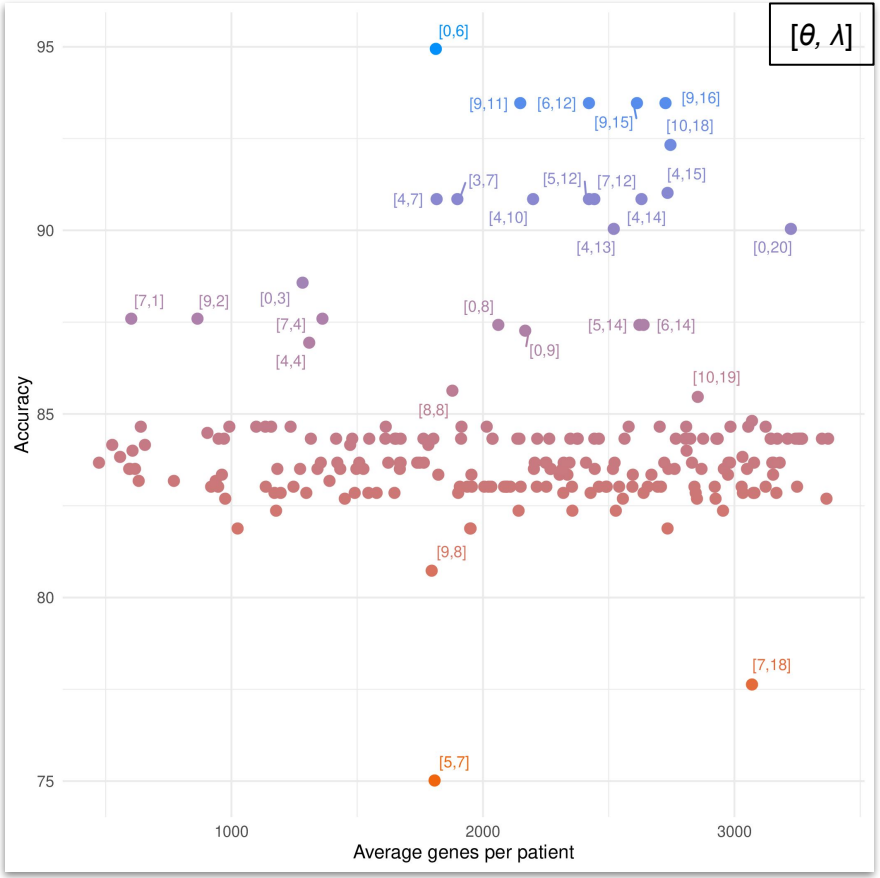


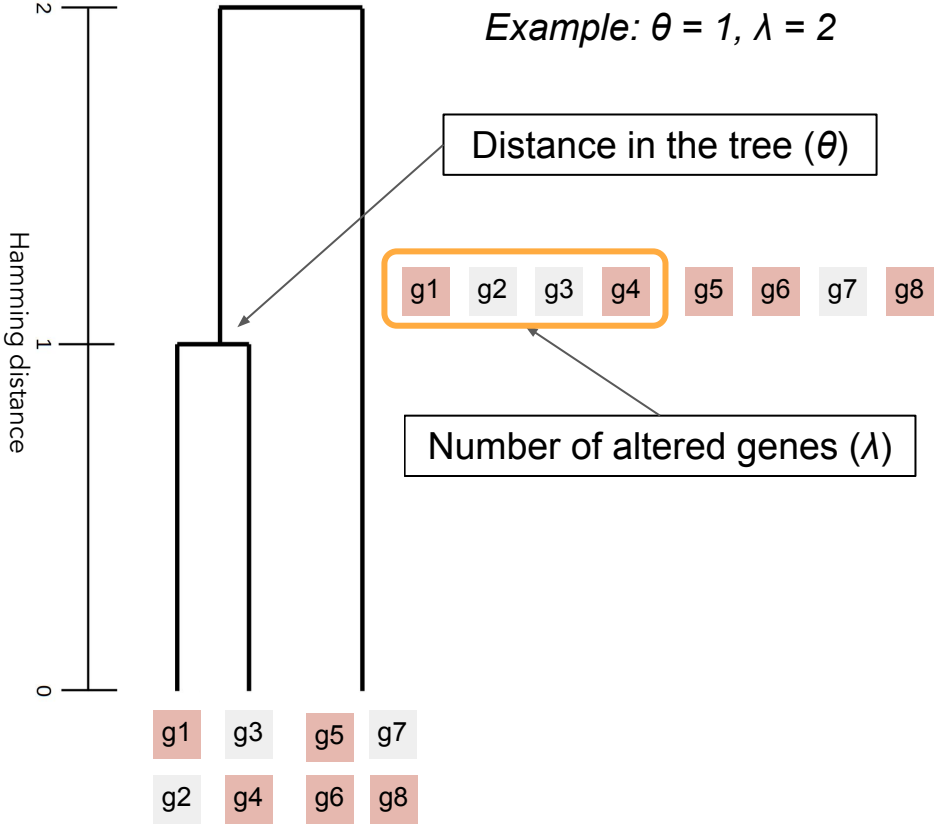
95% accuracy

87% dimensionality reduction

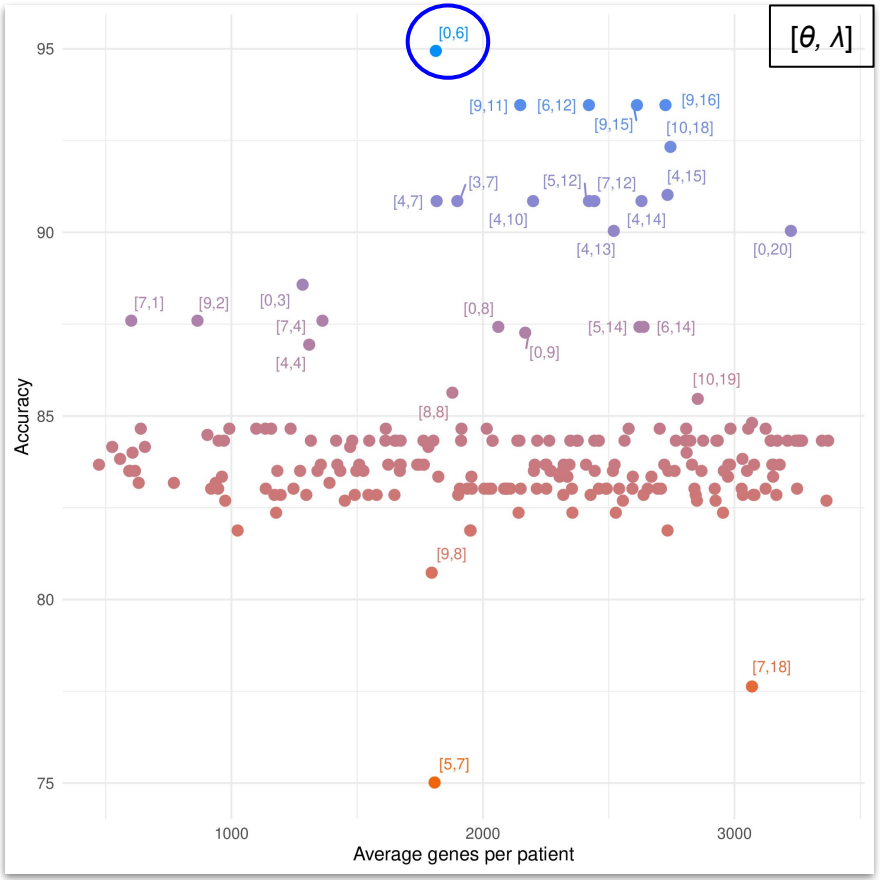
Forget et al. *Cancer Cell*. 2018
Nuñez et al. *Cell iScience*. 2021

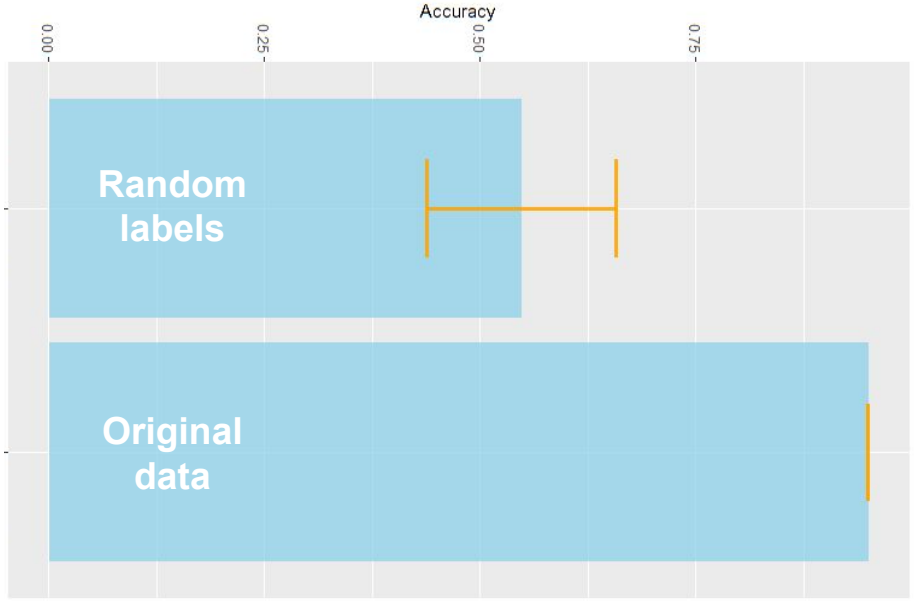
An optimization problem





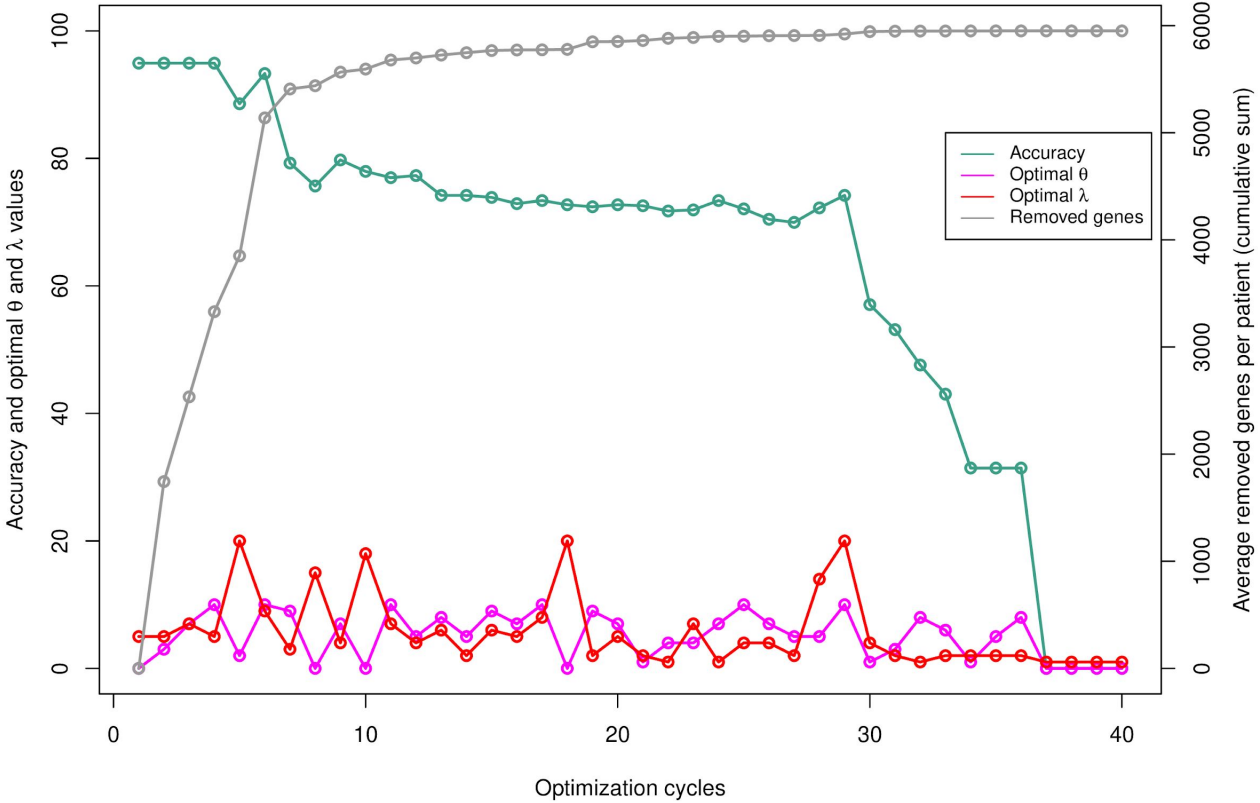
An optimization problem



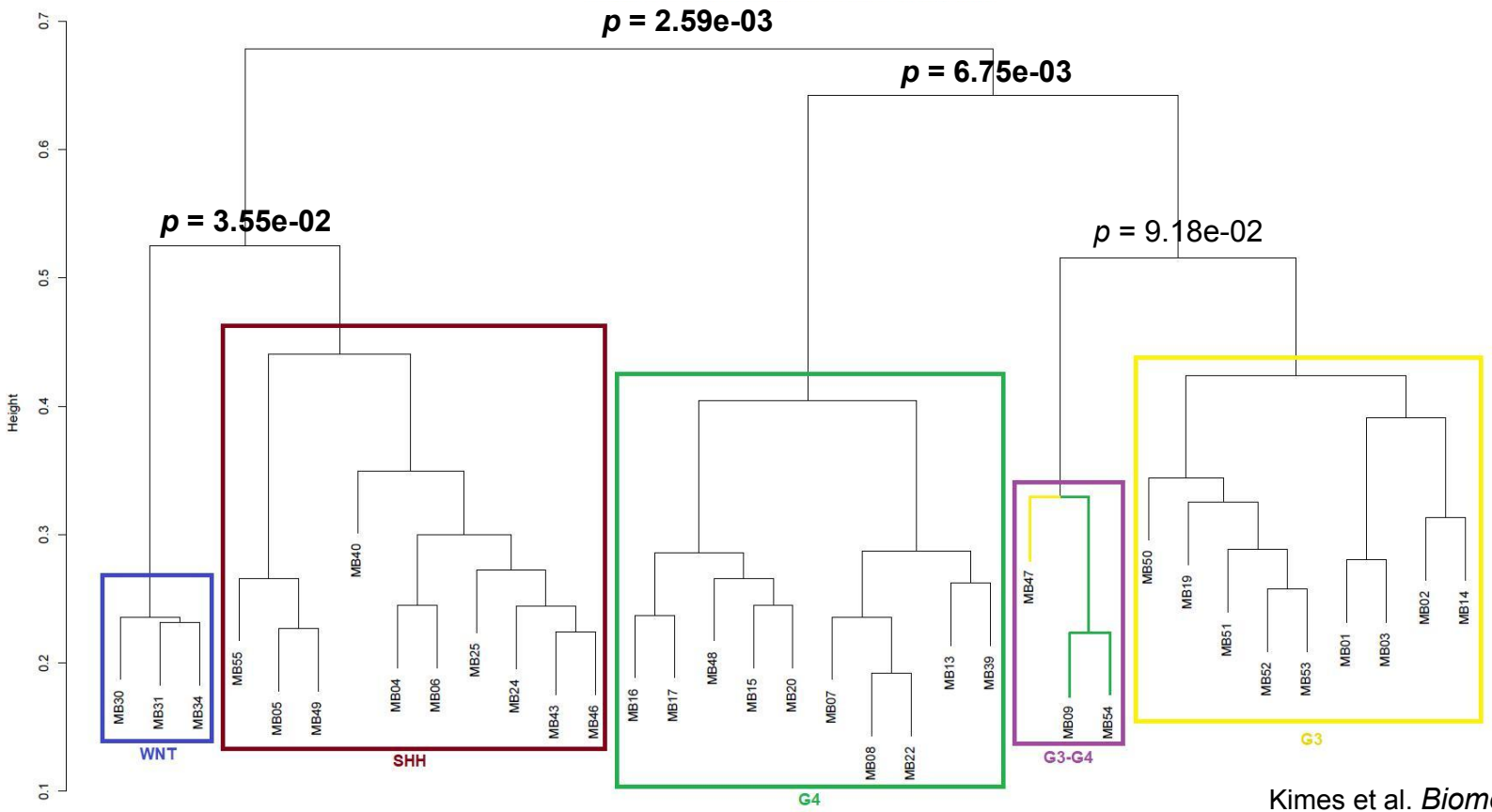


Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Sequential exclusion test



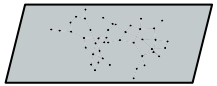
Patient stratification

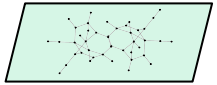


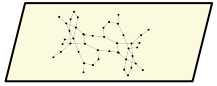
Kimes et al. *Biometrics*. 2017
Schwalbe et al. *Lancet Oncol.* 2017

Provenance of the associations


PROTEIN INTERACTIONS


DRUG TARGETS


GENETIC VARIANTS


PATHWAYS


METABOLIC REACTIONS

WNT	Nodal & Activin signaling	Rohr et al., 2001 Read et al., 2009
	Carbohydrates receptor activity	
SHH	Sonic Hegdehog signaling	Radeke et al., 1999 Briscoe and Thérond., 2003
	K^+ neuronal channels	
G3-G4	64 unique pathways	Guerreiro et al., 2008
	PI3K events in ERBB4 signaling	
G4	FGFR1	Egbivwie et al., 2019 Kessels and Malinow, 2009
	Synaptic transmission via glutamate receptors	
G3	57 unique pathways	Zhu et al., 2017
	PTEN stability	

Signorelli et al. *BMC Bioinformatics.* 2016

Severity in Congenital Myasthenic Syndromes

CHRNE c.1327delG

Severe | Non-severe

Omics analysis (WGS, RNAseq)

Curated databases

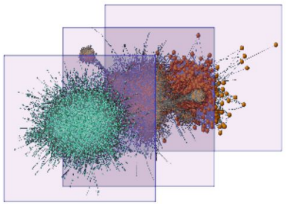
Compound heterozygous variants

CNVs

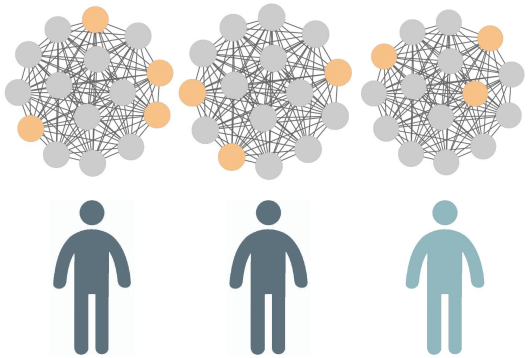
CMS causal genes

Multilayer community structure analysis

Personalized modules of damaged genes

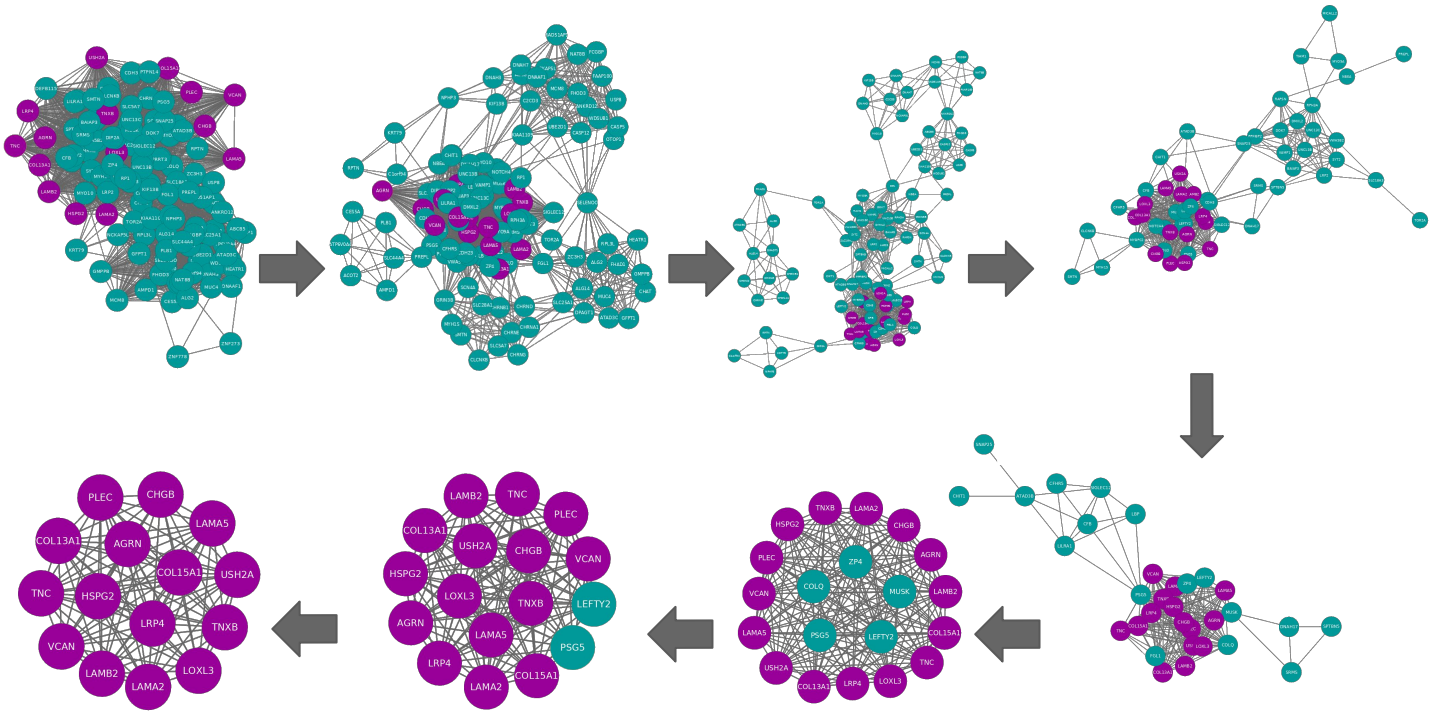


Interactome
Metabolome
Reactome



Nuñez et al. (in preparation)

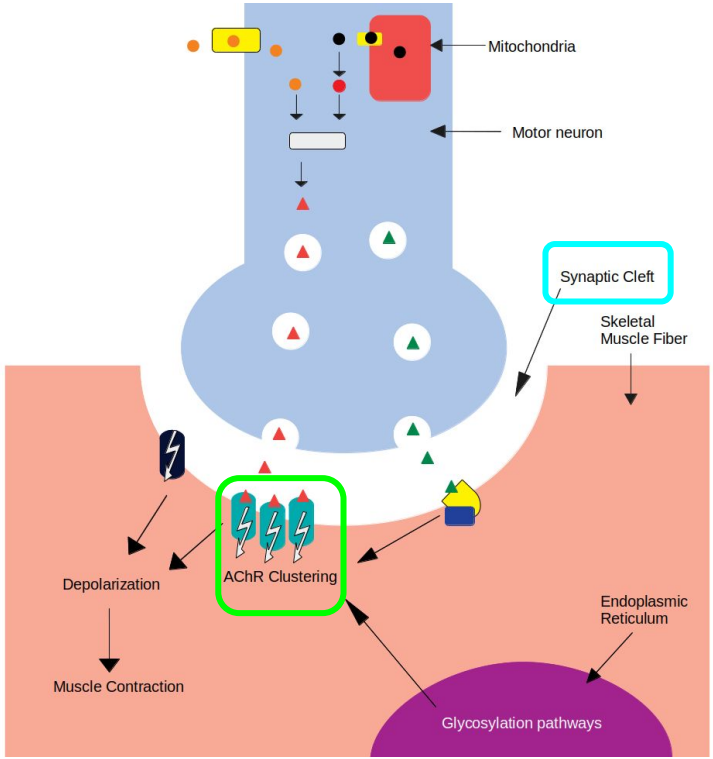
Severity in Congenital Myasthenic Syndromes



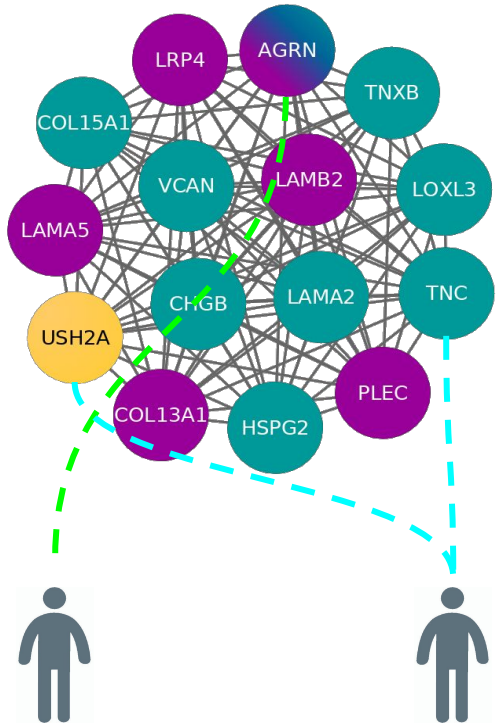
Nuñez-Carpintero et al. (in preparation)

Severity in Congenital Myasthenic Syndromes

Neuromuscular junction



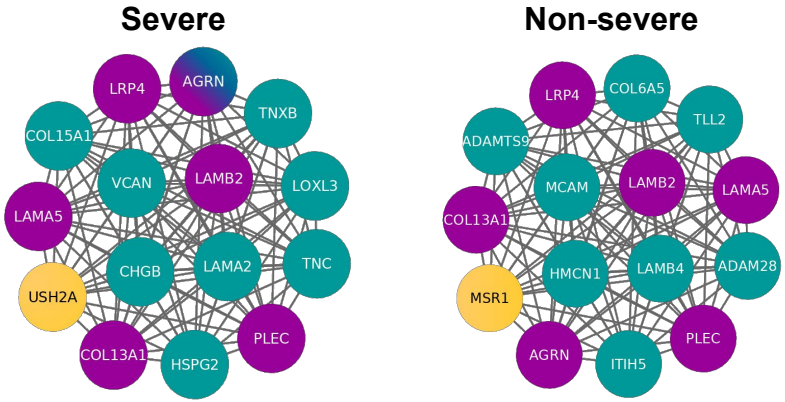
Gene module of severe cases



- Compound heterozygous variants
- Copy number variants
- Known CMS gene

Nuñez et al. (in preparation)

Severity in Congenital Myasthenic Syndromes

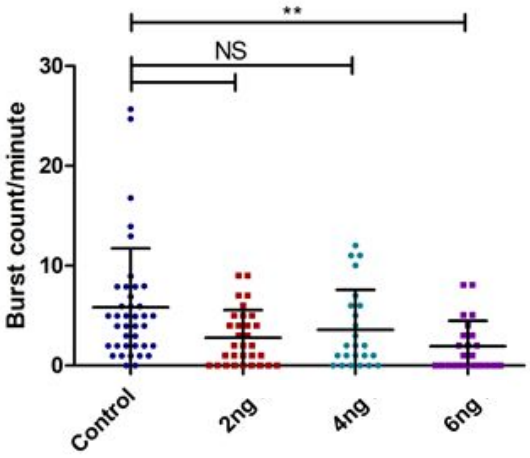


Activity localization	Class	CMS causal gene	Phenotype group		Function	Synaptic localization (Manual curation)	Localization (UniProt)
			Not-severe	Severe			
ECM (ECM)	Proteoglycans	AGRN	-	AGRN	Cell hydration and growth factor trapping	Pre- and post-synaptic (PMID:29462312)	Synaptic basal lamina / ECM
		-	-	HSPG2		Basement membrane (PMID:30453502)	Basement membrane / ECM
		-	-	VCAN		ECM (PMID: 29211034)	ECM
		-	-	COL15A1		Basement membrane (PMID:26937007)	ECM
	Collagens	COL13A1	-	-	Structural support	Basement membrane, post-synaptic (PMID:30768864)	Post-synaptic cell membrane
		-	COL6A5	-		Basement membrane (PMID:23869615)	Extracelullar matrix
	Laminins	LAMA5	-	-	Web-like structures	Pre-synaptic (PMID:28544784)	Basement membrane / ECM
		LAMB2	-	-		Basement mebrane (PMID:27614294)	Basement membrane / ECM / Synaptic cleft
		-	LAMB4	-		Myenteric plexus basement membrane (PMID:28595269)	Basement membrane / ECM
		-	-	LAMA2		Pre-synaptic (PMID: 9396756)	Basement membrane / ECM
		-	-	USH2A		Neuronal projection of stereocilia (PMID:19023448)	Stereocillium membrane / Secreted (Extracelullar region)
	Fibulins	-	HMCN1	-	Scaffolding	Glomerular Extracelullar matrix (PMID:29488390)	Basement membrane / ECM
	Tenascins	-	-	TNC	Anti-adhesion	Basement membrane (PMID:29466693)	ECM / Perisynaptic ECM (Ensembl)
				TNXB		Basement membrane (PMID:23768946)	ECM
	Enzymes	-	-	LOXL3	Collagen assembly	Basement membrane (PMID:26954549)	Secreted (extracellular region)
			ADAMTS9	-	Proteoglycan cleavage	Secreted to ECM (PMID:30626608)	ECM
			ADAM28			ECM (PMID:24613731)	Cell membrane / Secreted (extracellular region)
	Neuropeptides	-	-	CHGB	Regulatory peptides precursor	Pre- and post-synaptic (PMID:7526287)	Secreted (extracellular region)
	Others	-	ITIH5	-	Hyaluronic acid binding	ECM (PMID:27143355)	Secreted (extracellular region)
Cell surface	Receptors	-	MSR1	-	Proteoglycan and collagen binding	Macrophage surface Scavenger Receptor (PMID:12488451)	Plasma membrane
			MCAM			Plasma membrane (PMID:28923978)	Plasma membrane
		LRP4	-	-	Laminin binding	Post-synaptic (PMID:25319686)	Post-synaptic cell membrane
Cytoplasm	Cytoskeleton	PLEC	-	-	Structural support	Post-synaptic (PMID:20624679)	Post-synaptic cytoskeleton

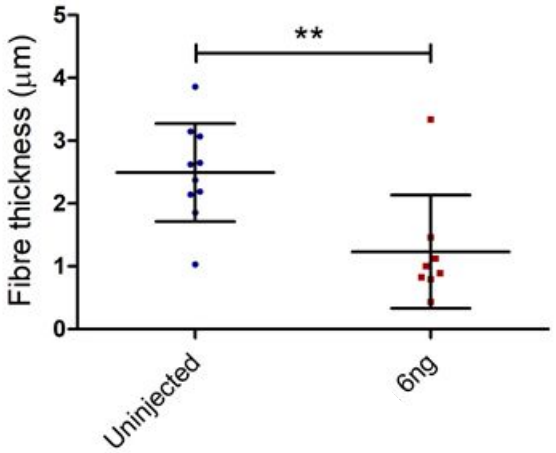
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Nuñez-Carpintero et al. (in preparation)

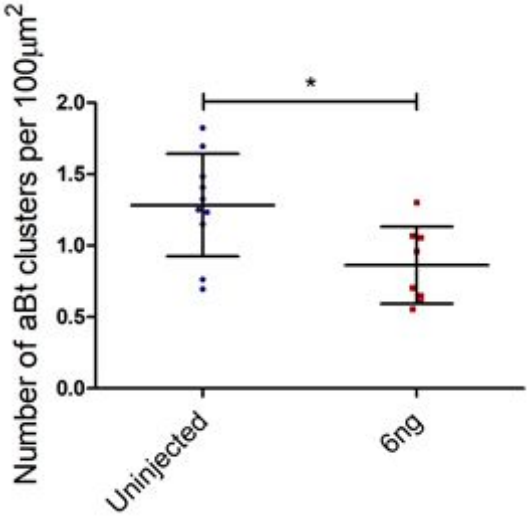
Decreased chorion movement
(1 day post fertilisation)



Decreased muscle fibre thickness
(5 days fish)

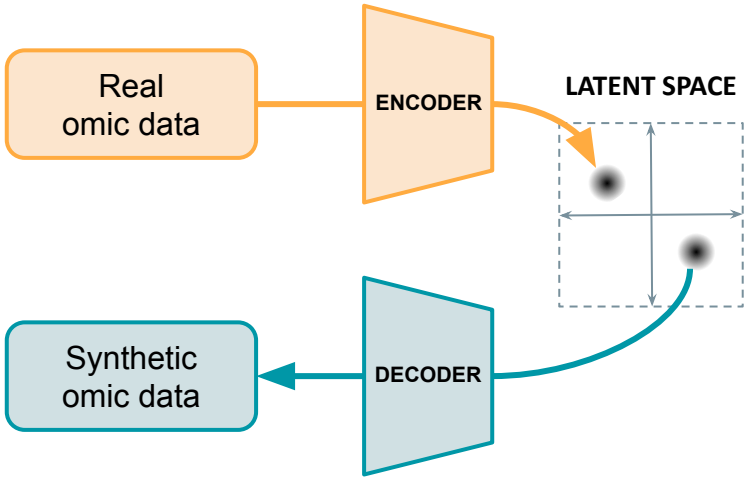


Decreased number of AChR clusters
(5 days fish)

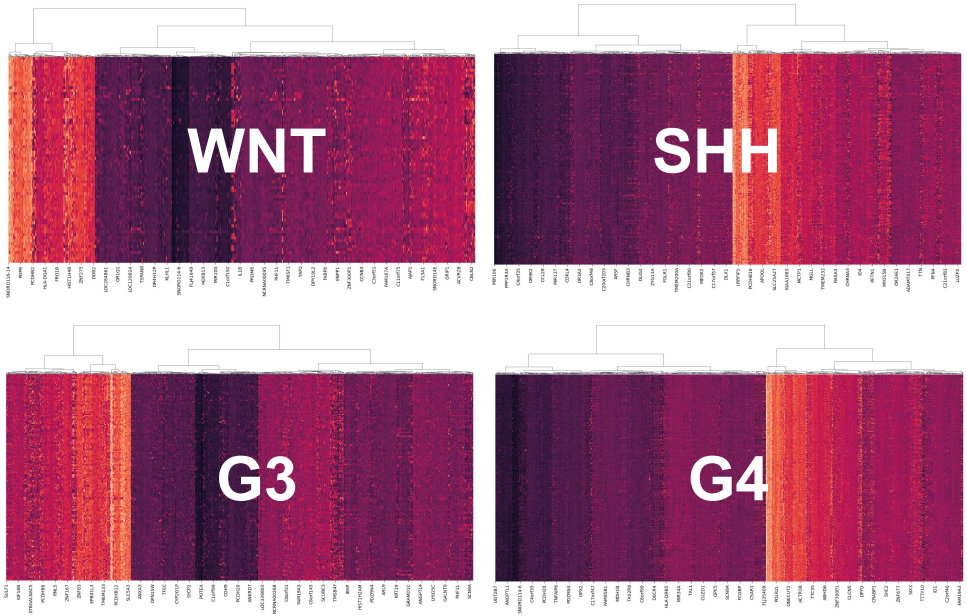


Explainable synthetic data generation for paediatric cancer

A **Variational Autoencoder** (VAE) can learn representations of **real data** of patients and therefore generate **synthetic data**.



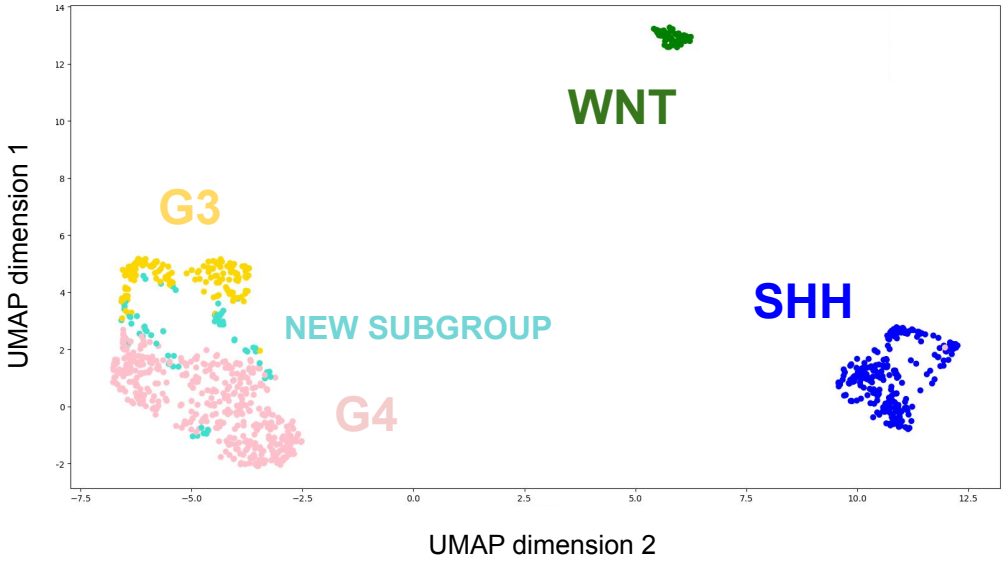
By studying **how** the VAE generates synthetic data we identified four distinct **omic signatures** of a **childhood brain tumor** (medulloblastoma).



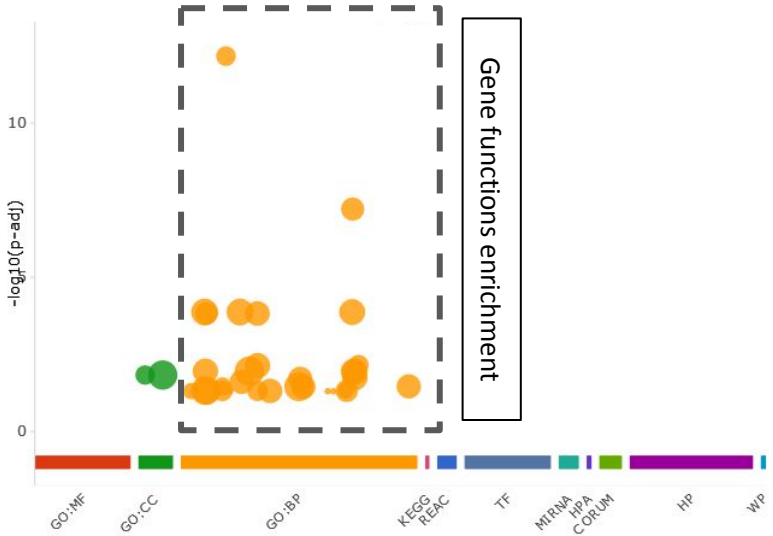
Tejada *et al.* 2022 [in preparation]

Explainable synthetic data generation for paediatric cancer

The VAE allowed us to discover **a new subgroup** of the childhood brain tumor characterized by **specific genes**.

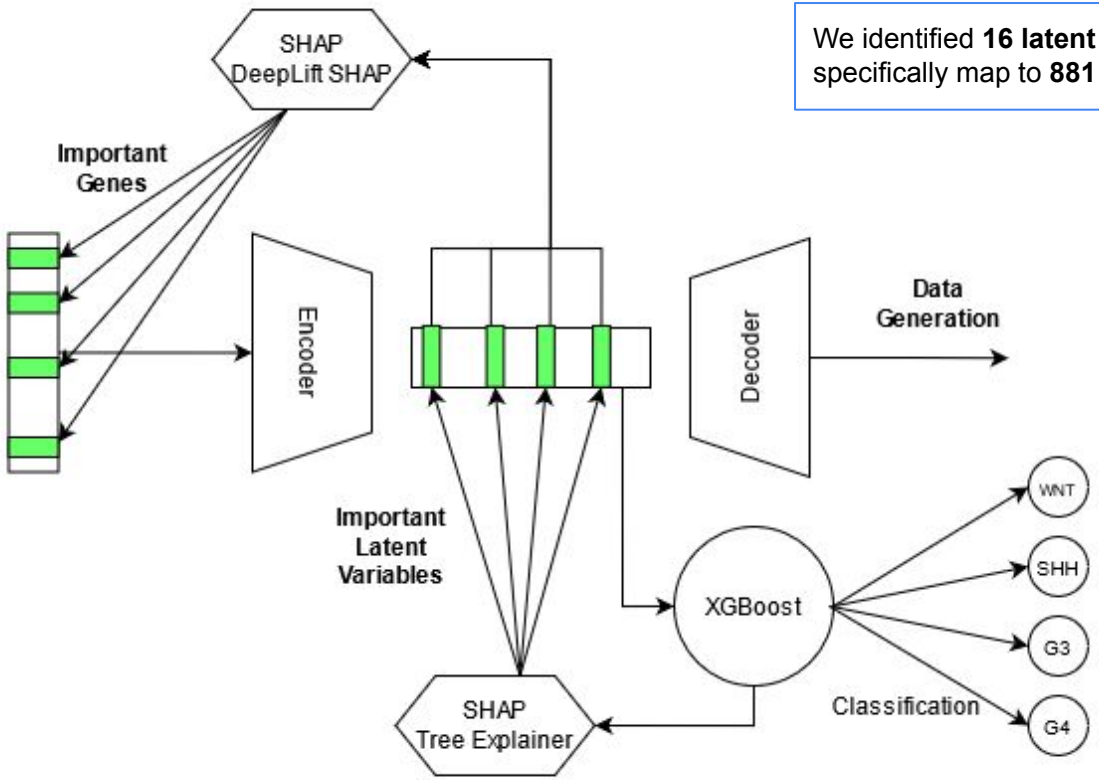


The functions of these genes are enriched in **synaptic signaling** and **nervous system development**.



Tejada *et al.* 2022 [in preparation]

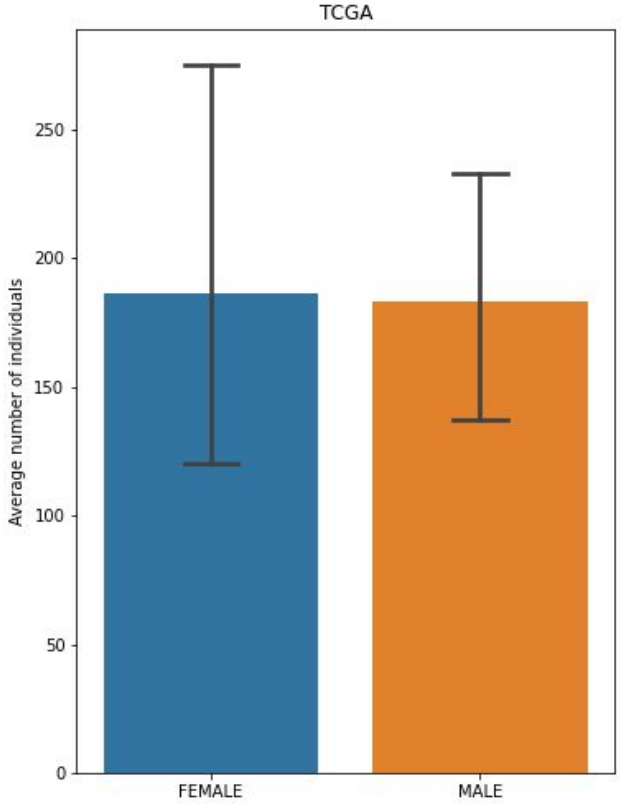
Explainable synthetic data generation for paediatric cancer

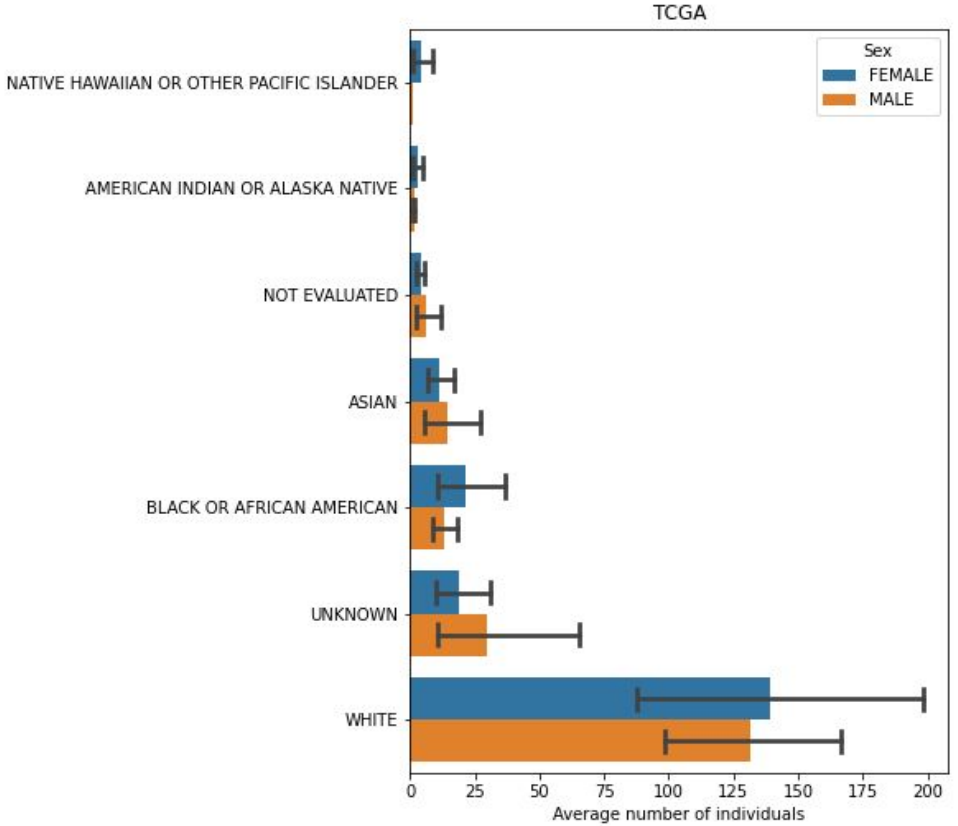


We identified **16 latent variables** that specifically map to **881 genes**.

Tejada *et al.* 2022 [in preparation]

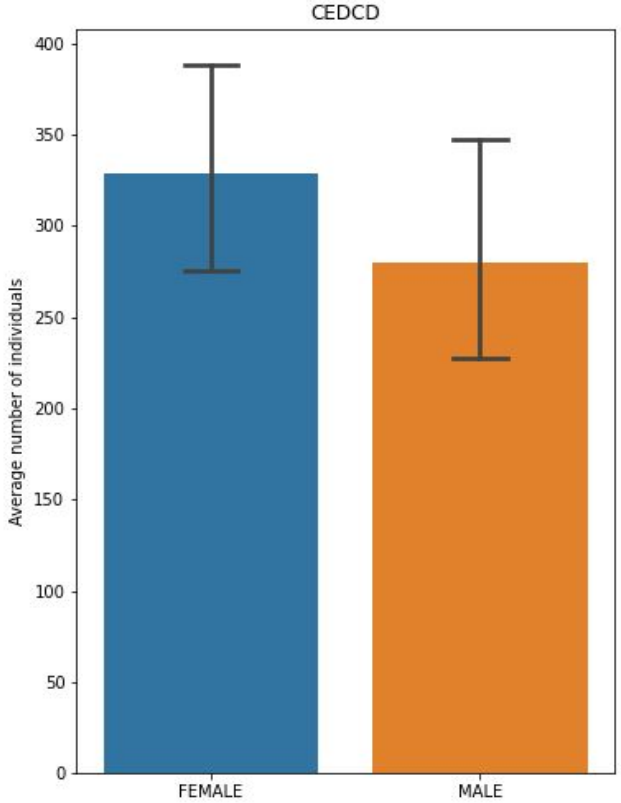
Sample size and label availability

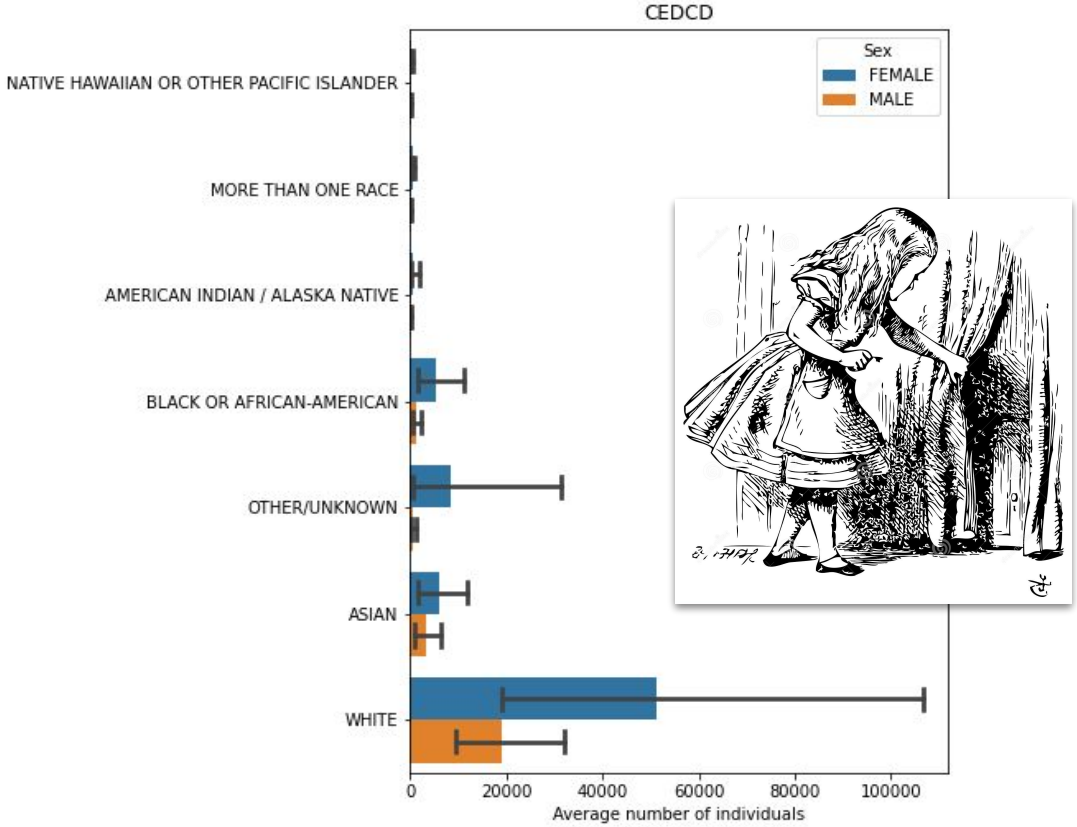




Cirillo et al. *Molecular oncology* vol. 15,4 (2021): 817-829. doi:10.1002/1878-0261.12920

Sample size and label availability





Cirillo et al. *Molecular oncology* vol. 15,4 (2021): 817-829. doi:10.1002/1878-0261.12920

Conclusions

- **Multilayer networks** represent a powerful tool for heterogeneous data integration in **rare diseases** such as medulloblastoma.
- The study of multilayer **community structure at different scales** enables to detect strong associations between bio-entities.
- The study of **multilayer community trajectories** allows to accurately performing tasks such as dimensionality reduction and molecular interpretation.
- Explainable **synthetic data generation** enables to both augment the data and to identify genes that are relevant to data synthesis.

Alfonso Valencia
Iker Núñez-Carpintero
Alejandro Tejada Lapuerta
Mar Batlle Pérez
Hanns Lochmüller
Maria Rigau
Mattia Bosio
Emily O'Connor
Salvador Capella
Steve Laurie
Sergi Beltran
Yoshiteru Azuma
Ana Topf
Rachel Thompson
Peter-Bram t'Hoen
Ivailo Tournev
Velina Guergueltcheva
Carolina Armengol
Marianyela Petrizzelli
Andrei Zinovyev
Anaïs Baudot
Alberto Valdeolivas
Léo Pio-Lopez

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EXCELENCIA SEVERO OCHOA

Thank you

Juan de la Cierva Programa

GOBIERNO DE ESPAÑA
MINISTERIO DE ECONOMÍA, INDUSTRIA Y COMPETITIVIDAD

davide.cirillo@bsc.es