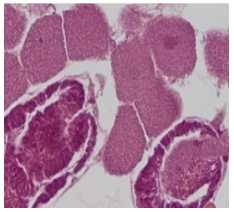
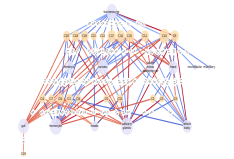
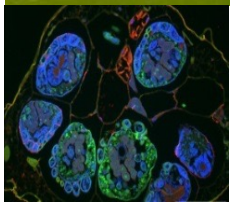
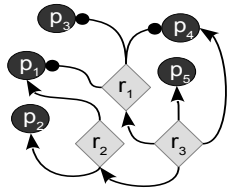
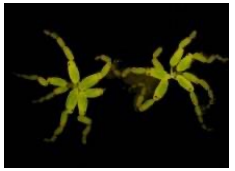










Generalizable Gene Self-Expressive Network

Sergio Peignier et Federica Calevro

Laboratory of Functional Biology, Insects and Interactions (BF2i)
INRAE/INSA-Lyon, University of Lyon

1ère Journée SPE et Numérique, INSA Lyon, 31 mai 2022

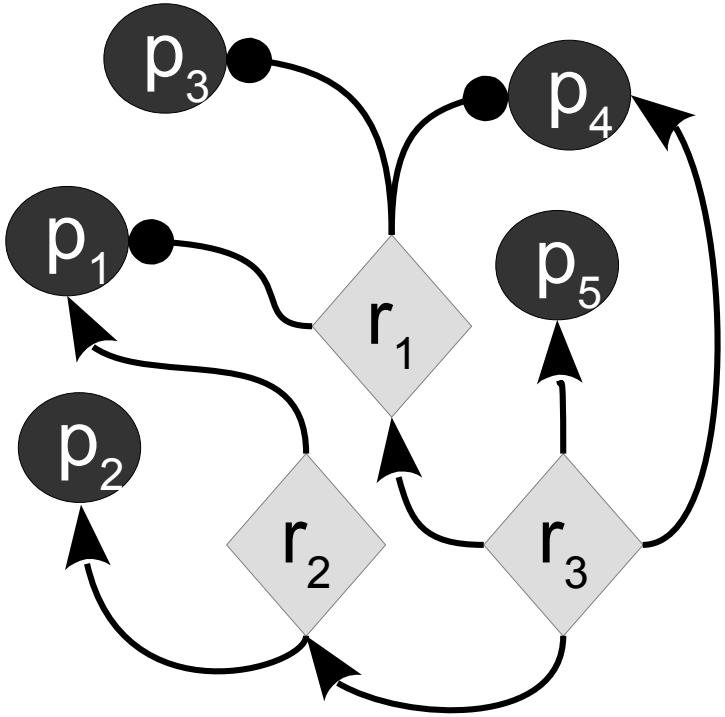
Gene Regulatory Networks (GRNs)

Definition

Interacting molecular regulators (e.g. transcription factors) **controlling** the **gene expression**

G = ⟨V, E⟩: **oriented graph** with nodes V and edges E s.t.

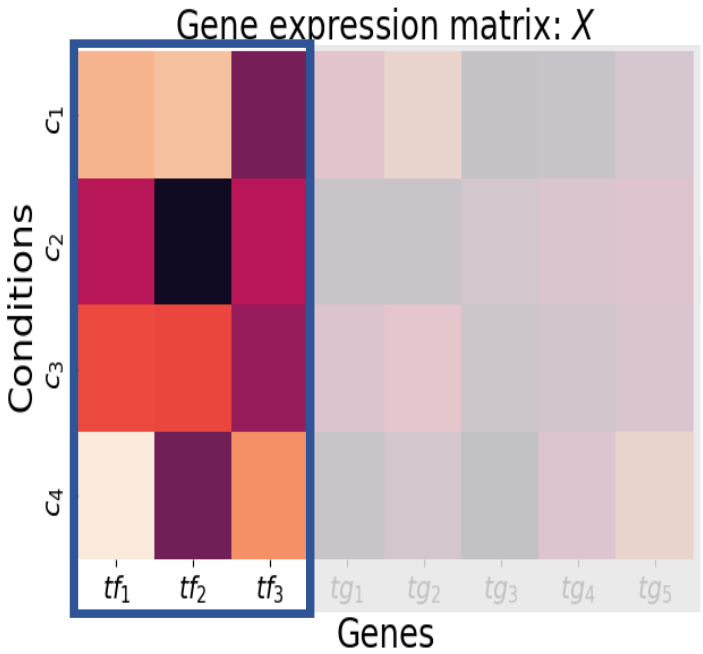
- Let V is the set of **genes**
- Let R ⊂ V be a set of **regulators**
- ∃e = (r, p) ∈ E if r ∈ R **controls the gene expression** of p



- **Wide range of mechanisms:**
(e.g., epigenetic, transcriptional ...)

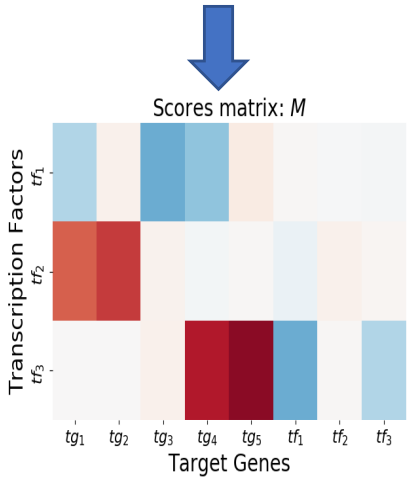
- **Important biological role:**
  - Adaptation
  - Versatility
  - Differentiation
  - Morphogenesis ...

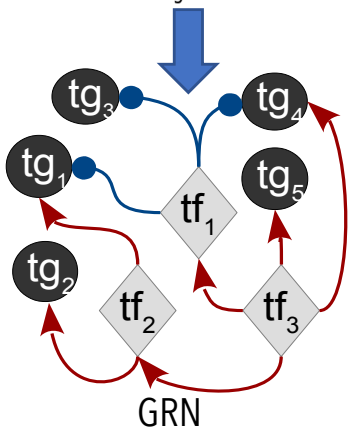
Data-driven GRN inference methodology

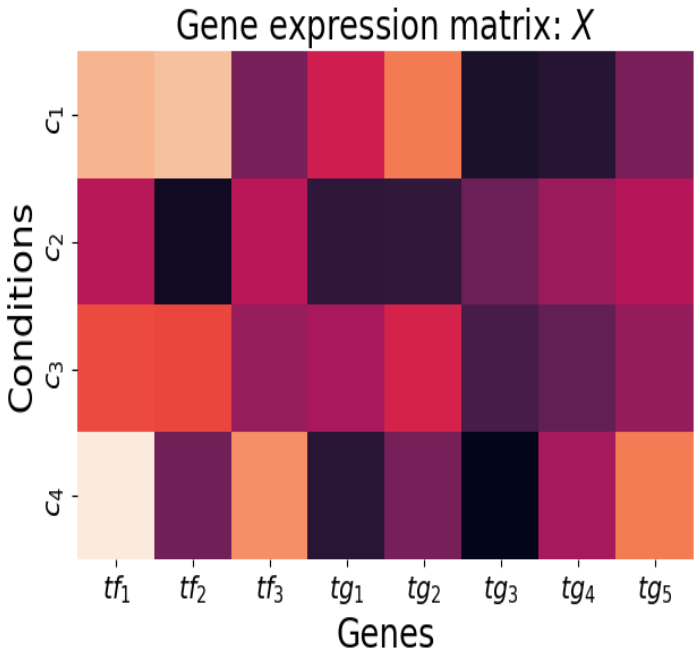


Machine Learning Algorithm
(Regression/Classification)

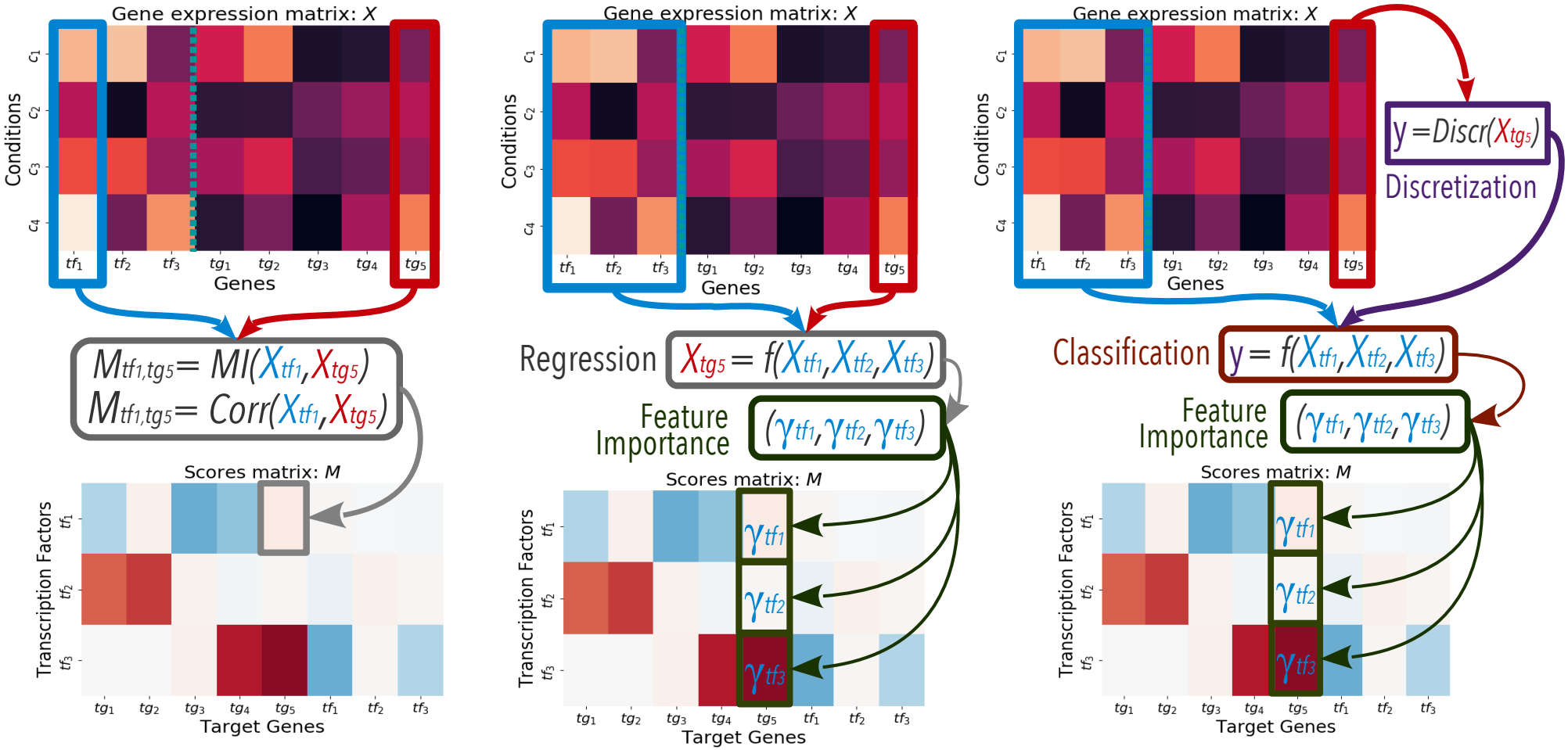
(e.g., Irrthum et. al 2010, Peignier et al. 2019)





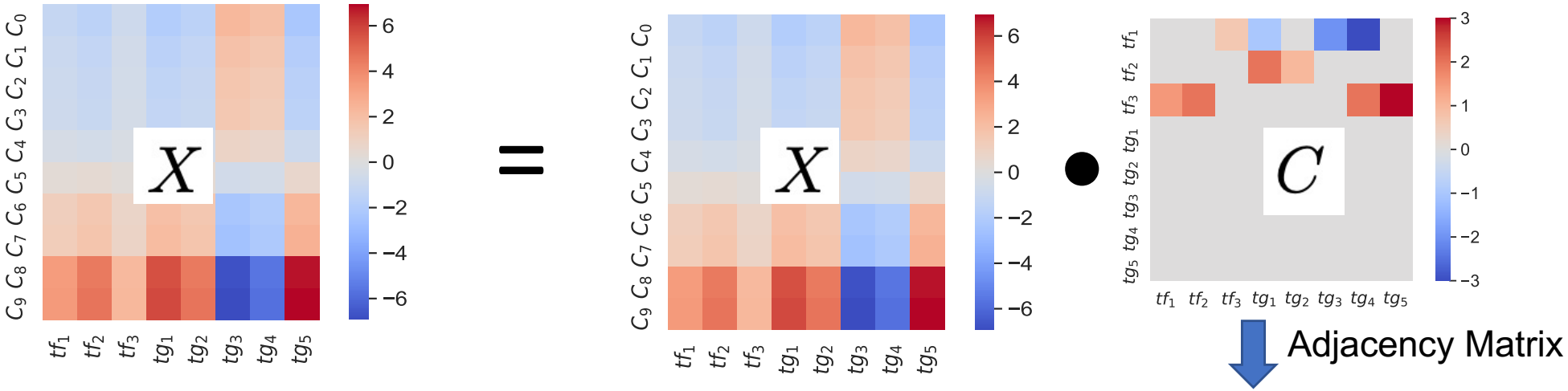


Data-driven GRN Inference families



Self-expressiveness

- Characterize the **relationship between object in a dataset**:
Express each object as a **linear combination of other objects**
- Used in computer vision, time series analysis, inference of Causal Graphs
- Self-expressiveness and GRN Inference are **strongly related**



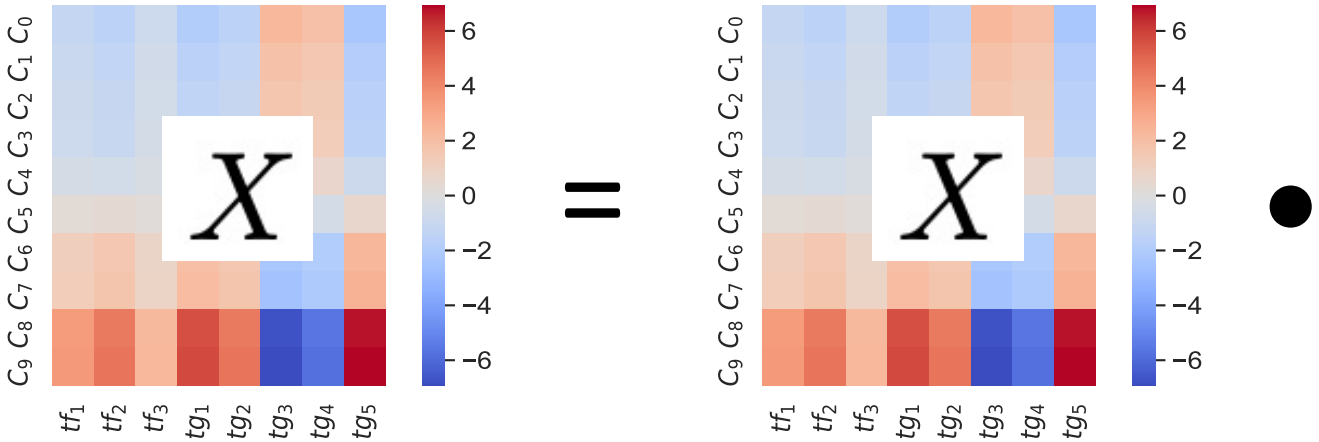
$$X = X \cdot C \quad \text{s.t. } C_{j,j} = 0 \quad \forall\, j$$

Many solutions exist for matrix *C* …
which one should be choosen?

Subspace preserving self-expressiveness

- Express **each data point** as **linear combinations** of **other points** laying in the **same subspace**
- **Sparsity-inducing norms** -> **remove connections** between points from **different subspaces:**

  - **L0 norm** (e.g., Orthogonal Matching Poursuit - OMP)
  - **L1 norm** (e.g., LASSO)
  - **L1 & L2 norm** (e.g., ElasticNet)

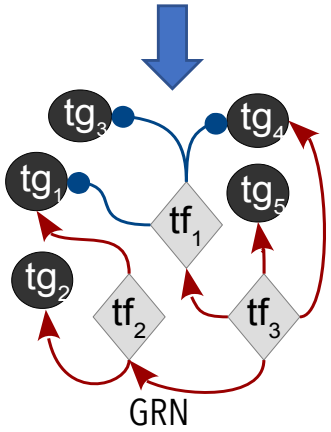


Sparse
Adjacency Matrix

$$C^*_{\star,j} = \underset{C_{\star,j}}{\operatorname{argmin}} \ \|C_{\star,j}\|$$

$$\text{with } X_{\star,j} = X \cdot C_{\star,j} \quad \text{and } C_{j,j} = 0$$

Possible to add other constrains (e.g., motifs presence)



Generalizable Gene Self-eXpression Networks - GXN

- D : Number of conditions
- Γ : Set of *N* genes
- Ψ : Set of regulatory genes ($\Psi \subset \Gamma$)

GXN•OMP : Orthogonal Matching Poursuit – L0 norm

$$C^*_{\star,g} = \underset{C_{\star,g}}{\operatorname{argmin}} \ \|X_{\star,g} - X \cdot C_{\star,g}\|_2^2$$

$$\text{with } \|C_{\star,g}\|_0 \leq d_0, \ C_{g,g} = 0 \ \ \forall g \in \{1, \ldots, N\} \text{ and } C_{j,g} = 0 \ \ \forall j \notin \Psi$$

- d_0 : Maximal number of regulators per Target Gene (parameter)

GXN•EN : ElasticNet – L1 & L2 norm

$$C^*_{\star,g} = \underset{C_{\star,g}}{\operatorname{argmin}} \ \frac{\|X_{\star,g} - X \cdot C_{\star,g}\|_2^2}{2D} + \alpha\rho\|C_{\star,g}\|_1 + \frac{\alpha(1-\rho)}{2}\|C_{\star,g}\|_2^2$$

$$\text{with} \quad C_{g,g} = 0 \quad \forall\, g \in \{1, \ldots, N\} \ \text{ and } \quad C_{j,g} = 0 \quad \forall\, j \notin \Psi$$

- $\rho \in [0, 1]$: Mixing parameter (0 -> L2 regulatiration, 1 -> L1 regularization)
- $\alpha \in \mathbb{R}_+$: Regularization strength (0 -> No regularization)

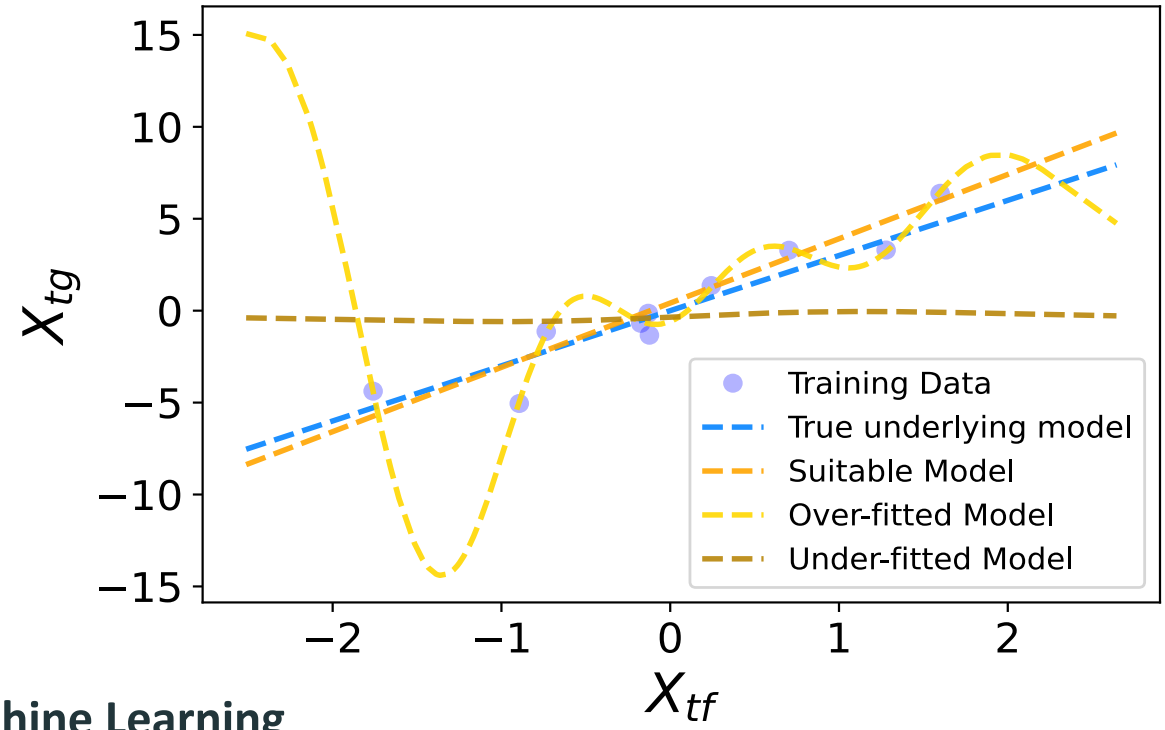
Parameters calibration?

Data-driven GRN Inference limitations

Parameter calibration -> Under/over-fitting tradeoff

- **Underfitted models:**
Lack of expressiveness, lead to over-simplistic predictions

- **Overfitted models:**
More parameters than needed. Poor and biased predictions



Generalization: a major goal in Machine Learning

Predict accurately the outputs of examples that were not used during the training phase.

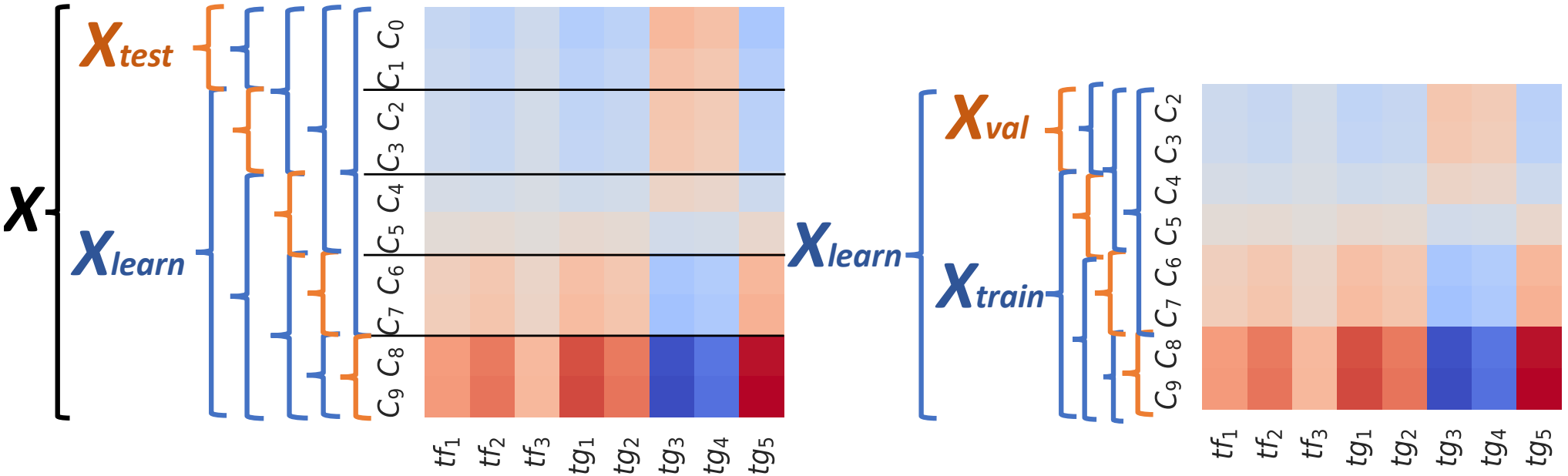
Generalization and GRN inference

GRN Inference methods do not measure model's generalization during parameters tunning or evaluation

Nested K-fold cross-validation

Outer cross-validation:

- Xtrain : Train the model with a given parameter setting
- Xval : Evaluate the Generalization capabilities to choose the best parameters
- Xtest : Evaluate the Generalization capabilities of the best model



Internal evaluation metric: R^2 determination coefficient ($-\infty$: worse quality -> 1: best quality)

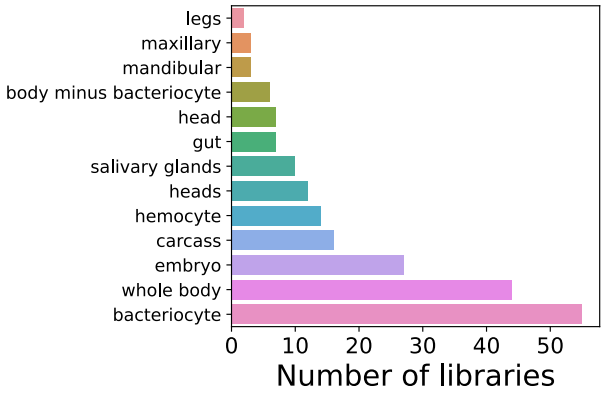
Datasets

DREAM5 Benchmark and RNAseq Eukaryote datasets

DREAM5	Data	D	$\|\Gamma\|$	$\|\Psi\|$	$\|E_{gold}\|$	$\frac{\|E_{gold}\|}{\|E^{full}_{gold}\|}$	$\|E^{full}\|$
In silico	Simulated	805	1,643	195	4,012	0.014	320,190
S. aureus	Microarray	160	2,810	99	515	0.028	278,091
E. coli	Microarray	805	4,511	334	2,066	0.013	1,506,340
S. cerevisiae	Microarray	536	5,950	333	3,940	0.017	1,981,017
Eukaryotes	Data	D	# Tissues	$\|\Psi\| = \|\Gamma\|$	$\|E^{full}\|$		
C. familiaris	RNAseq	75	6	2,286	5,223,510		
R. norvegicus	RNAseq	80	11	2,358	5,557,806		
H. sapiens	RNAseq	657	3	2,454	6,019,662		

D. melanogaster *: D* = 72, 1 tissue (eyes), $|\Gamma|$ = 15345 , $|\Psi|$ = 841

A. pisum *: D* = 206, 10 tissues + Whole aphids, $|\Gamma|$ = 17594 , $|\Psi|$ = 1046



Evaluation : Evaluation Metrics

- **Internal regression evaluation metric (on test set):**

  R^2 determination coefficient ($-\infty$: worse quality -> 1: best quality)

- **Network topology assessment:**

  Set of edges: $E = \{(\psi, g) \in \Psi \times \Gamma \mid C_{g,\psi} \neq 0\}$

  Set of possible edges: $E^{full} = \{(\psi, g) \in \Psi \times \Gamma \mid \psi \neq g\}$

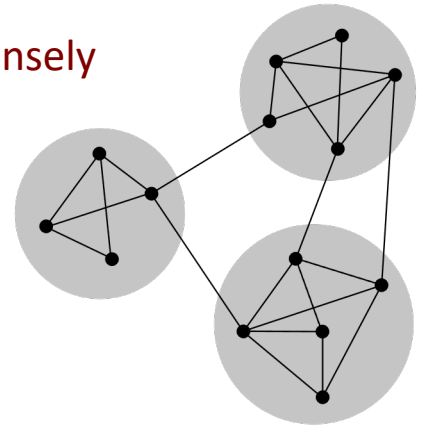
  - $sparsity = \frac{|E^{full} \setminus E|}{|E^{full}|} \times 100.$
  - $deg^+(\psi) = |\{g \in \Gamma \mid (\psi, g) \in E\}|$
  - $deg^-(g) = |\{\psi \in \Psi \mid (\psi, g) \in E\}|$

- **Gold Standard Comparison (DREAM5):**

  - **AUPR** (0: worse quality -> 1: best quality)
  - **AUROC** (0: worse quality -> 1: best quality)

Evaluation: Community Detection

Do inferred GRNs structure in communities (i.e., sub-networks of genes densely intra-connected by regulatory links) sharing common functional roles?



- Set of Edges: $E = \{(\psi, g) \in \Psi \times \Gamma \mid C_{g,\psi} \neq 0\}$
- In-degree: $deg^-(g) = |\{\psi \in \Psi \mid (\psi, g) \in E\}|$
- Out degree: $deg^+(\psi) = |\{g \in \Gamma \mid (\psi, g) \in E\}|$

Communities detection method: Clauset-Newman-Moore greedy **modularity maximization**
High modularity Q -> 1, Low modularity Q -> 0

$$Q = \frac{1}{|E|} \sum_{g \in \Gamma, \psi \in \Psi} \left(C_{\psi,g} - r \cdot \frac{deg^+(\psi) \cdot deg^-(g)}{|E|} \right) \cdot \zeta(\psi, g)$$

$$\zeta : \Psi, \Gamma \rightarrow \{0, 1\} \quad \begin{cases} \zeta(\psi, g) = 1 & \text{Gene and regulator are in the same community} \\ \zeta(\psi, g) = 0 & \text{Gene and regulator are not in the same community} \end{cases}$$

$r \in \mathbb{R}_+^*$ **Resolution** (High-> many small communities, low -> few large communities)

Evaluation: Communities GSEA and GO enrichment

Gene Set Enrichment Analysis - GSEA:

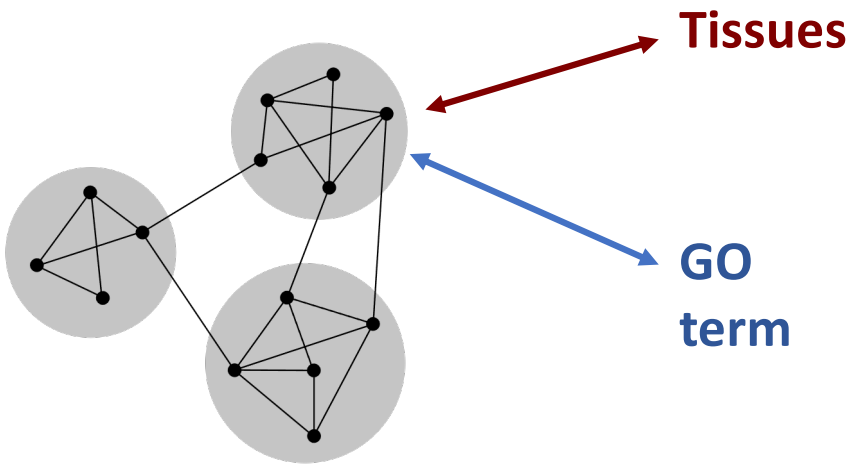
Are **communities** of genes that are **collectively over-expressed** or **under-expressed**, with statistical significance, in a **particular tissue or cell-type**.

- Ranking method: difference-of-classes
- 10,000 gene-set permutations
- Retain relationships with a False-Discovery-Rate (FDR) < 0.05

Gene Ontologies – GO (GO:0048856 anatomical structure development):

Do **communities** of genes exhibit significantly **over-represented GO terms**?

- GOATOOLS
- Benjamini Hochberg multiple test FDR correction
- Retain relationships with a False-Discovery-Rate (FDR) < 0.05



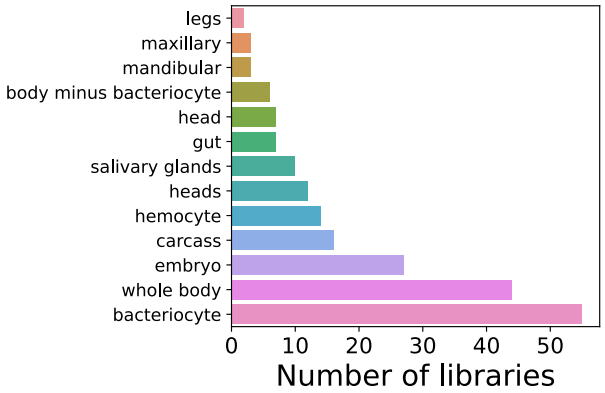
Datasets

DREAM5 Benchmark and RNAseq Eukaryote datasets

DREAM5	Data	D	$\lvert\Gamma\rvert$	$\lvert\Psi\rvert$	$\lvert E_{gold}\rvert$	$\frac{\lvert E_{gold}\rvert}{\lvert E^{full}_{gold}\rvert}$	$\lvert E^{full}\rvert$
In silico	Simulated	805	1,643	195	4,012	0.014	320,190
S. aureus	Microarray	160	2,810	99	515	0.028	278,091
E. coli	Microarray	805	4,511	334	2,066	0.013	1,506,340
S. cerevisiae	Microarray	536	5,950	333	3,940	0.017	1,981,017
Eukaryotes	Data	D	# Tissues	$\lvert\Psi\rvert = \lvert\Gamma\rvert$	$\lvert E^{full}\rvert$		
C. familiaris	RNAseq	75	6	2,286	5,223,510		
R. norvegicus	RNAseq	80	11	2,358	5,557,806		
H. sapiens	RNAseq	657	3	2,454	6,019,662		

D. melanogaster *: D* = 72, 1 tissue (eyes), $|\Gamma|$ = 15345 , $|\Psi|$ = 671

A. pisum *: D* = 206, 10 tissues + Whole aphids, $|\Gamma|$ = 17594 , $|\Psi|$ = 1046



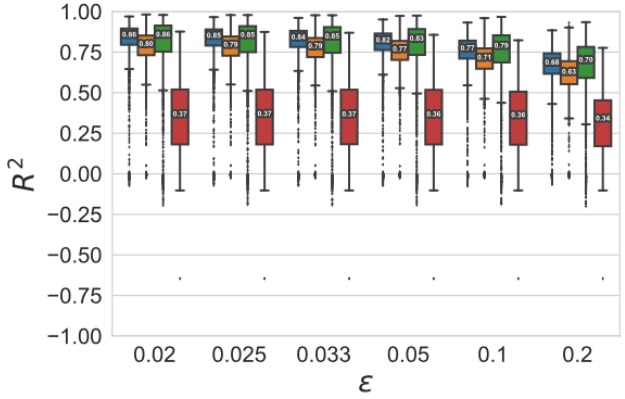
Results: DREAM5 - Sparsity

State-of-the-art methods

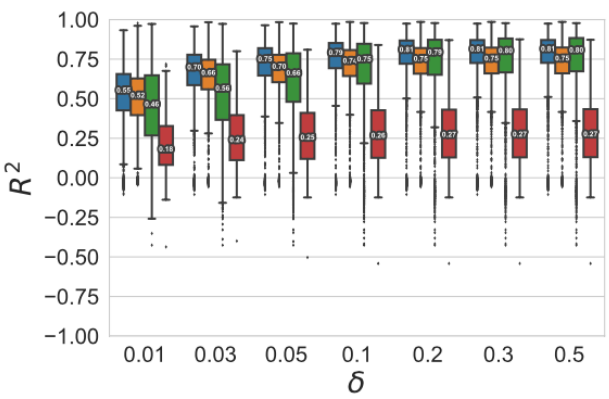
	S. cerevisiae	*E. coli*	*S. aureus*	*In silico*
SVR	0.025%	0.027%	0.021%	0.021%
RF	0.036%	0.002%	0.002%	0.019%
GXN•OMP (Min)	84.474%	83.530%	75.760%	97.037%
GXN•OMP (Max)	99.258%	99.181%	98.990%	99.487%
GXN•EN (Min)	79.119%	69.949%	67.240%	86.456%
GXN•EN (Max)	94.929%	94.454%	90.238%	93.429%

GXN•OMP GXN•EN exhibit high sparsity structure

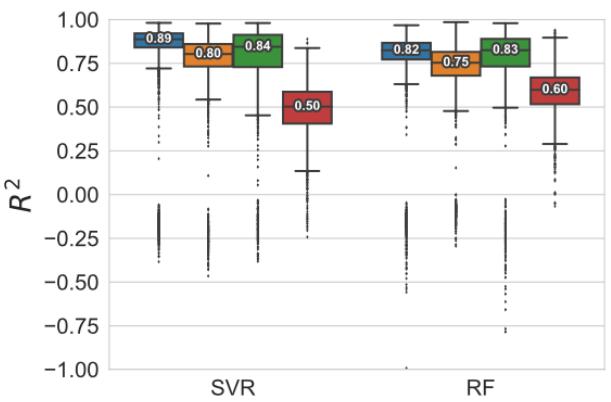
Results: DREAM5 – Inner evaluation



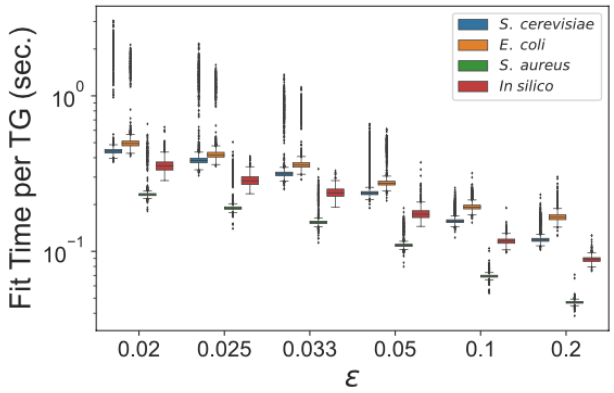
(a) GXN•EN - R^2



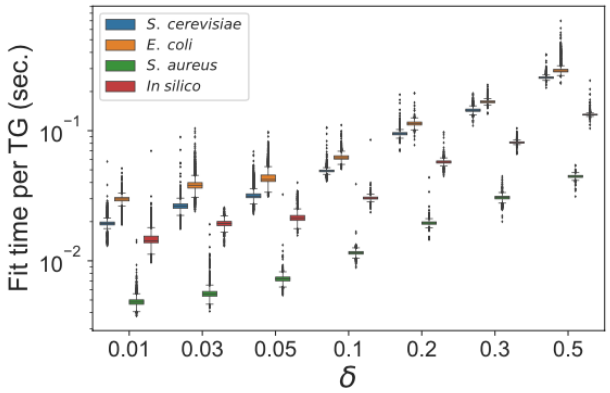
(b) GXN•OMP - R^2



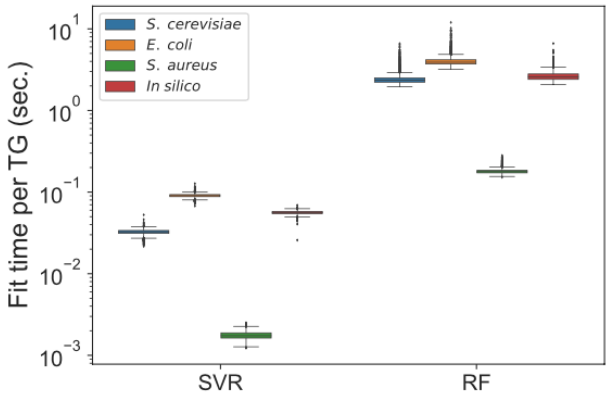
(c) SVR and RF - R^2



(d) GXN•EN - Runtime



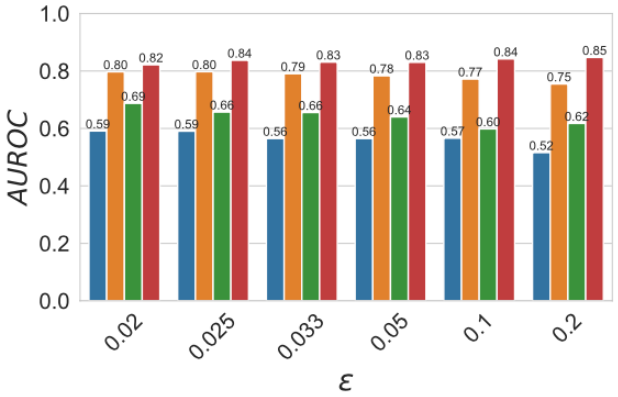
(e) GXN•OMP - Runtime



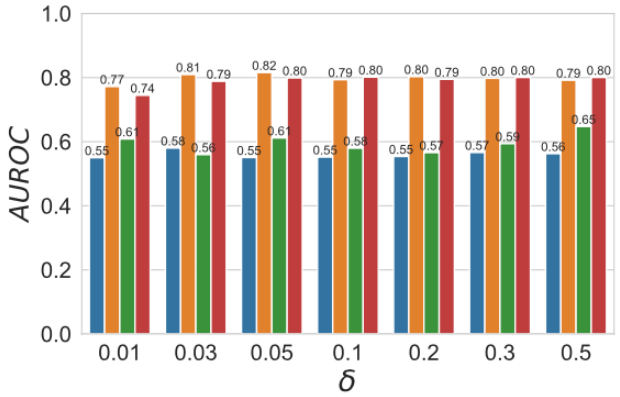
(f) SVR and RF - Runtime

GXN•OMP and **GXN•EN** exhibit comparable R2 scores wrt SVR and RF
GXN•OMP and **GXN•EN** training runtime are lower than RF and comparable to SVR

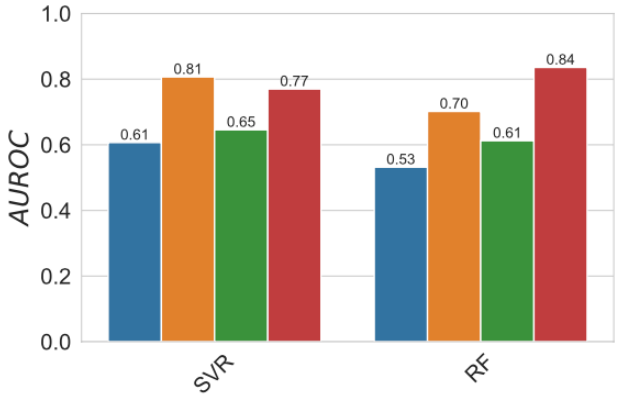
Results: DREAM5 – External evaluation



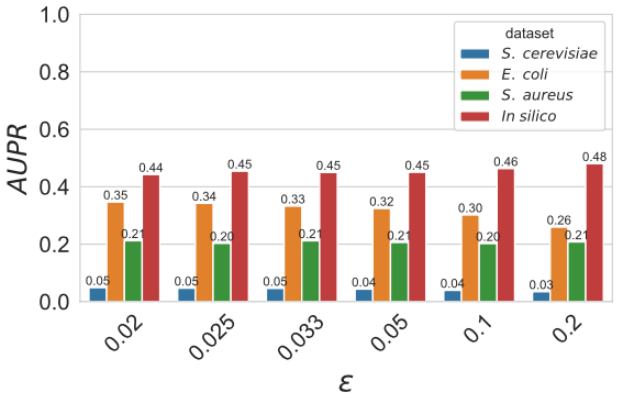
(a) GXN•EN - $AUROC$



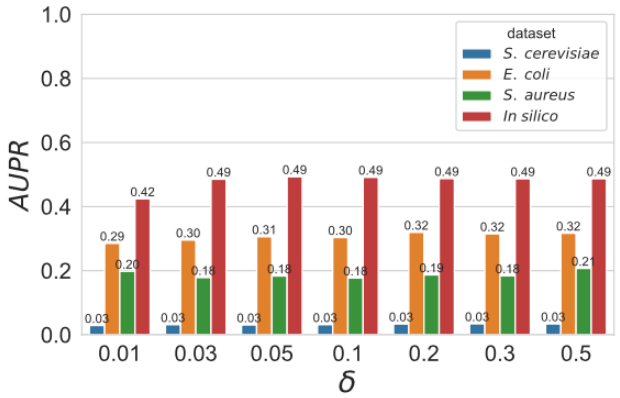
(b) GXN•OMP - $AUROC$



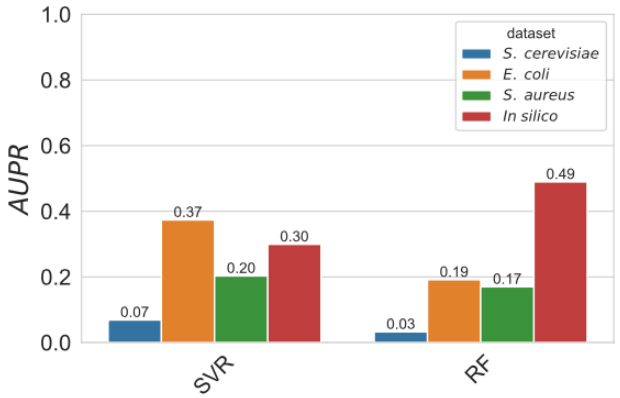
(c) SVR and RF - $AUROC$



(d) GXN•EN - $AUPR$



(e) GXN•OMP - $AUPR$

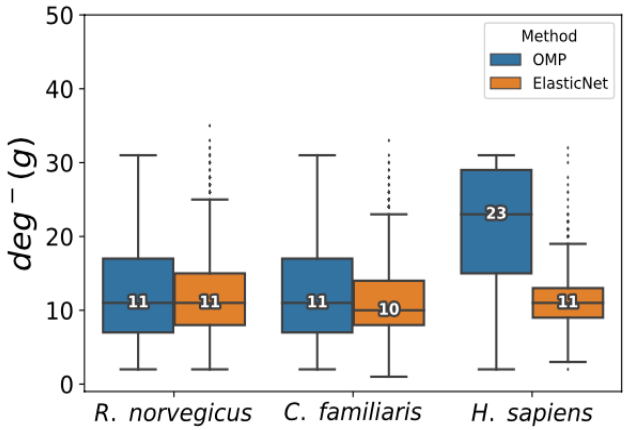


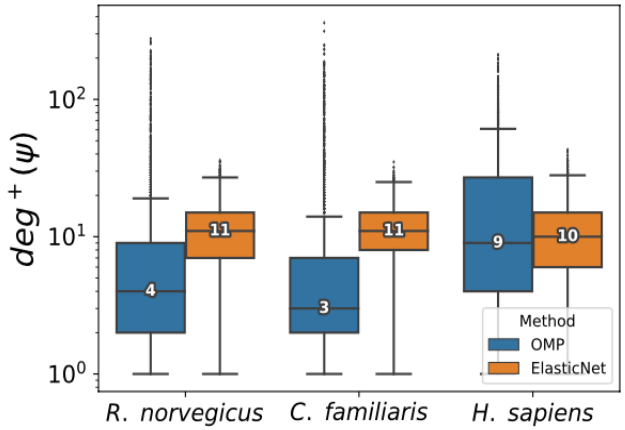
(f) SVR and RF - $AUPR$

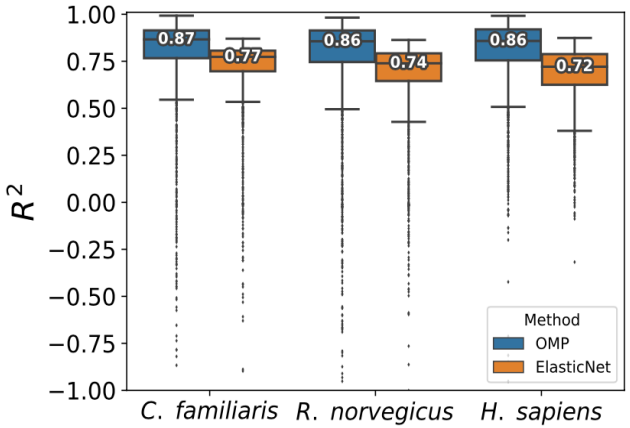
GXN•OMP and **GXN•EN** exhibit comparable AUROC and AUPR wrt SVR and RF

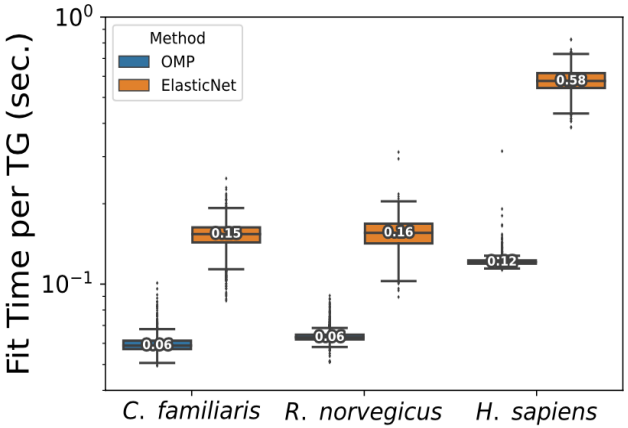
Results: Mammal datasets – evaluation metrics

	Sparsity (%)		Modularity	
	GXN•EN	GXN•OMP	GXN•EN	GXN•OMP
R. norvegicus	99.495	99.39	0.575	0.829
C. familiaris	99.491	99.357	0.627	0.835
H. sapiens	99.538	99.087	0.658	0.573







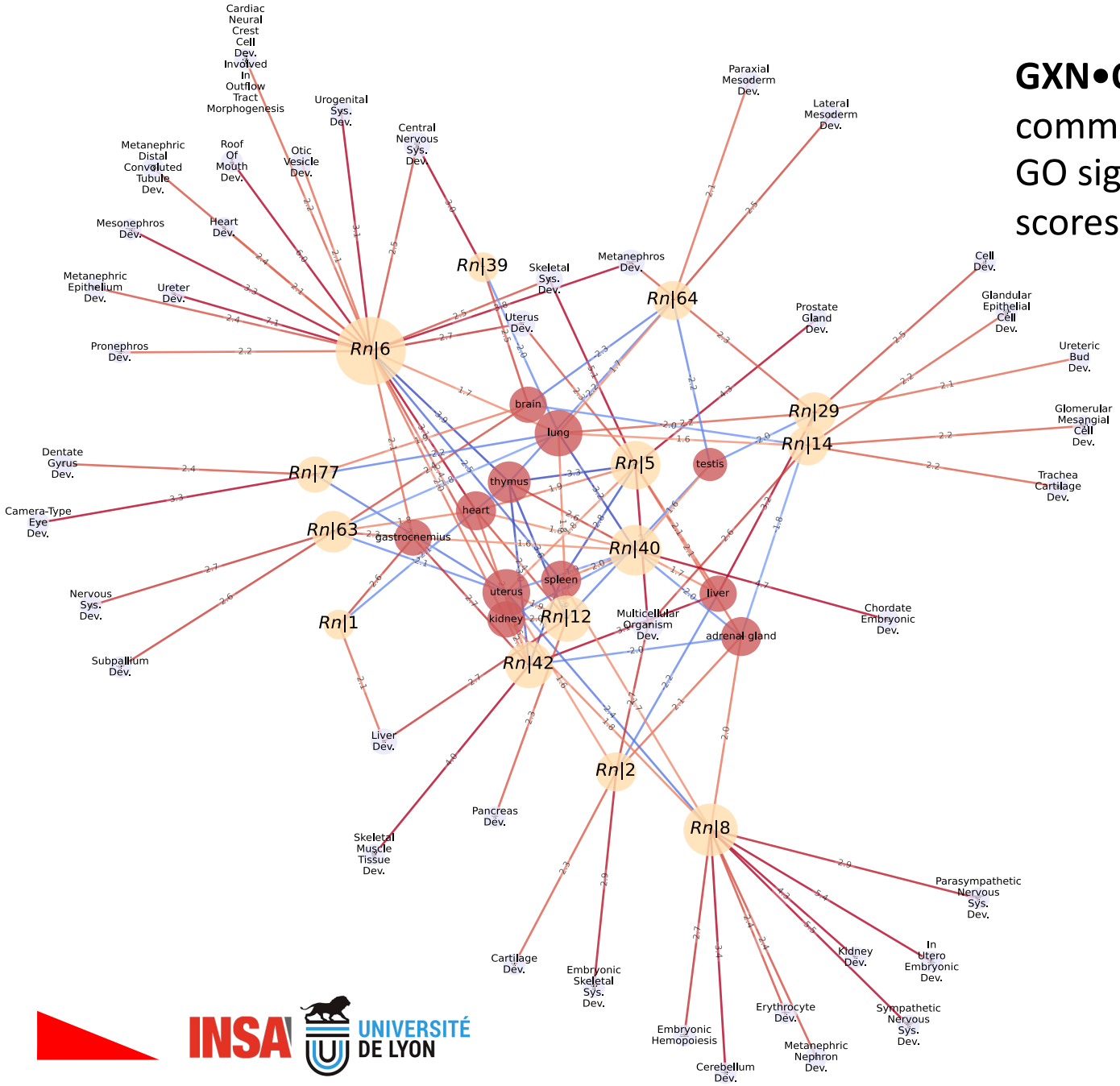


GXN•OMP GXN•EN exhibit **high sparsity** and **modularity**

GXN•OMP GXN•EN exhibit **simple** and **interpretable** regulatory models

GXN•OMP GXN•EN exhibit **high R2 scores** and are **fast to train**

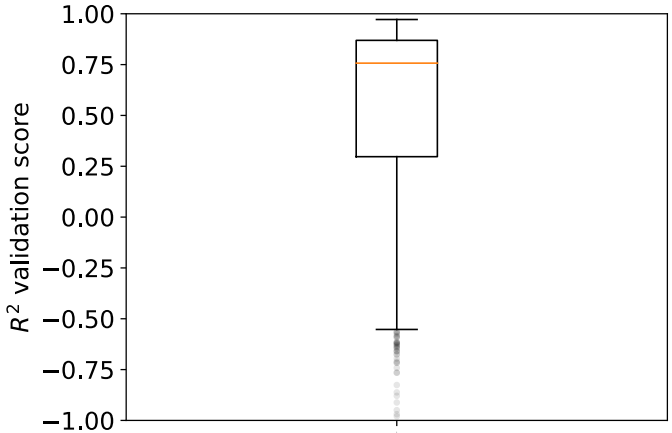
Results: Mammal datasets – Community structure



GXN•OMP on *R. norvegicus:* communities exhibit GSEA and GO significant and coherent scores

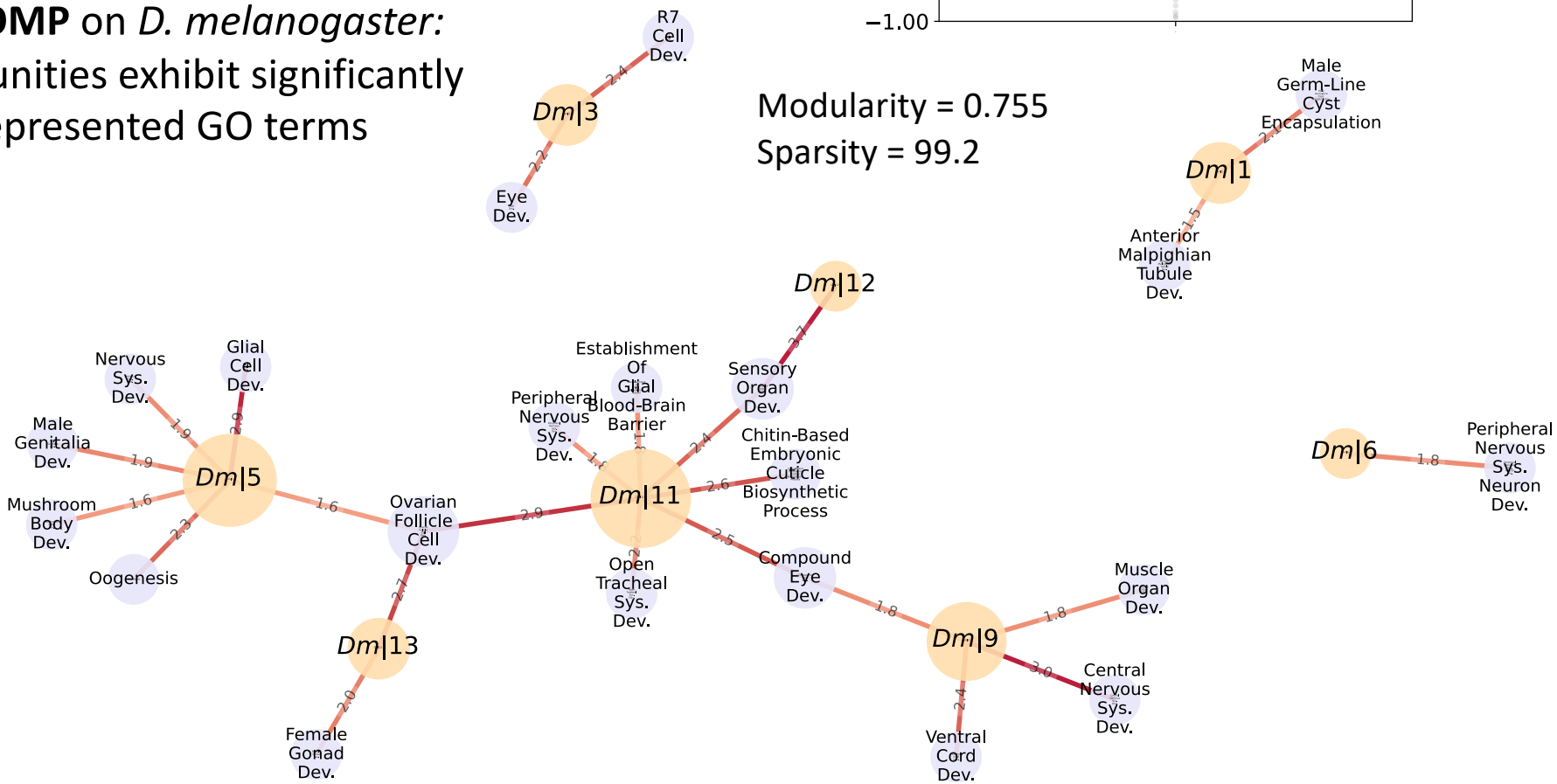
Results: *D. melanogaster* dataset

Motif constrained inference (motifs for 372 TFs)

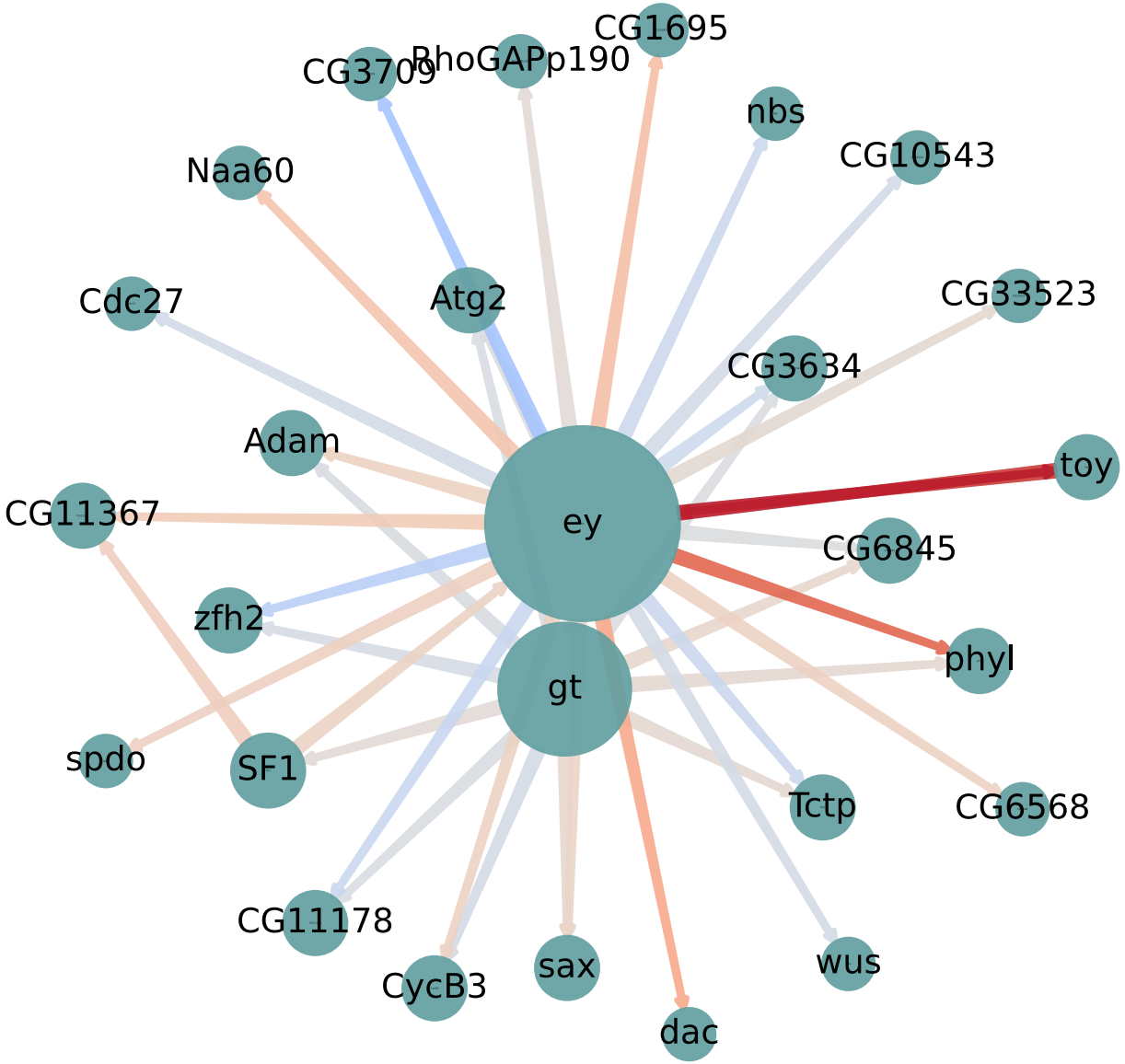


GXN•OMP on *D. melanogaster:* communities exhibit significantly over-represented GO terms

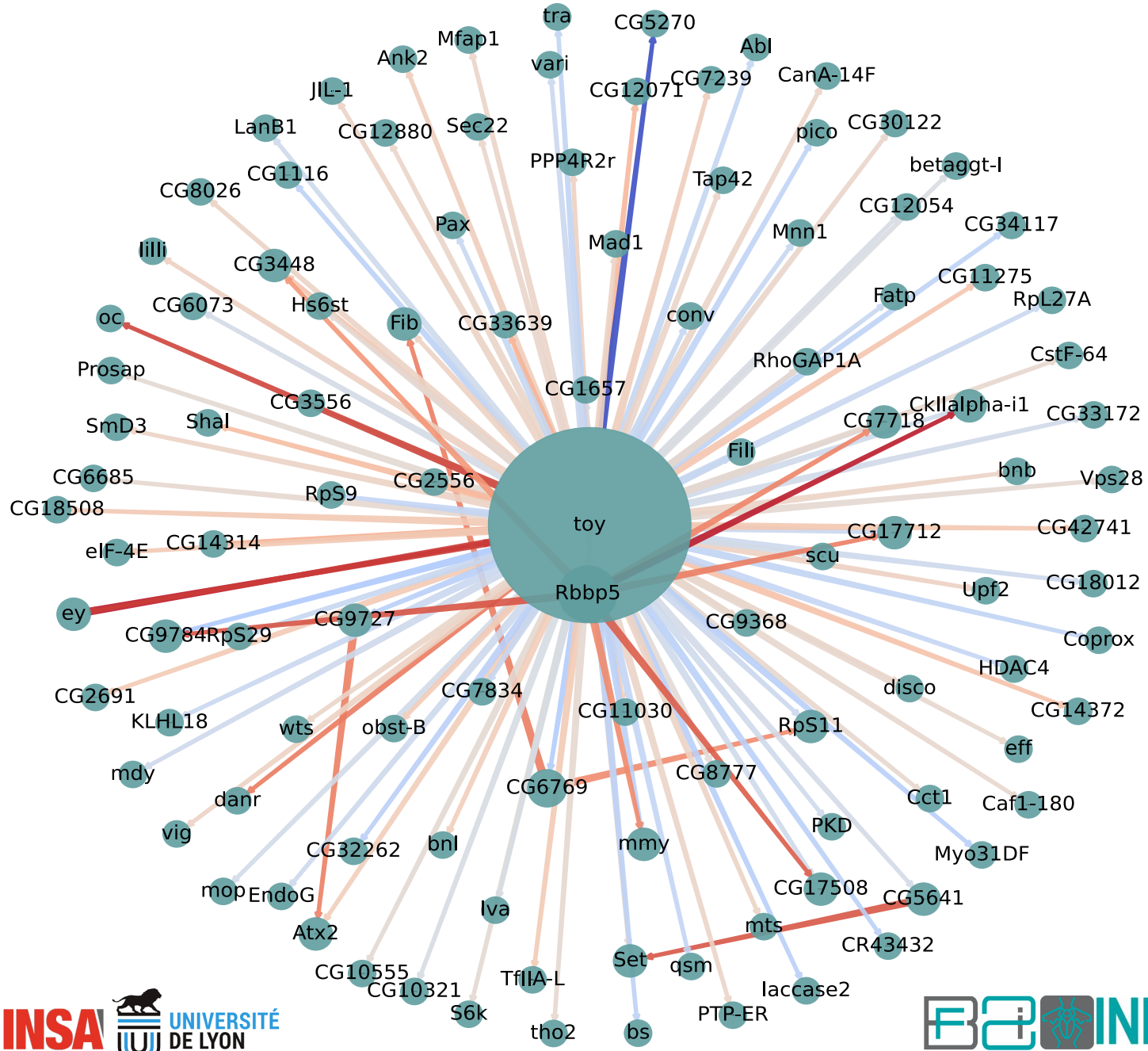
Modularity = 0.755
Sparsity = 99.2



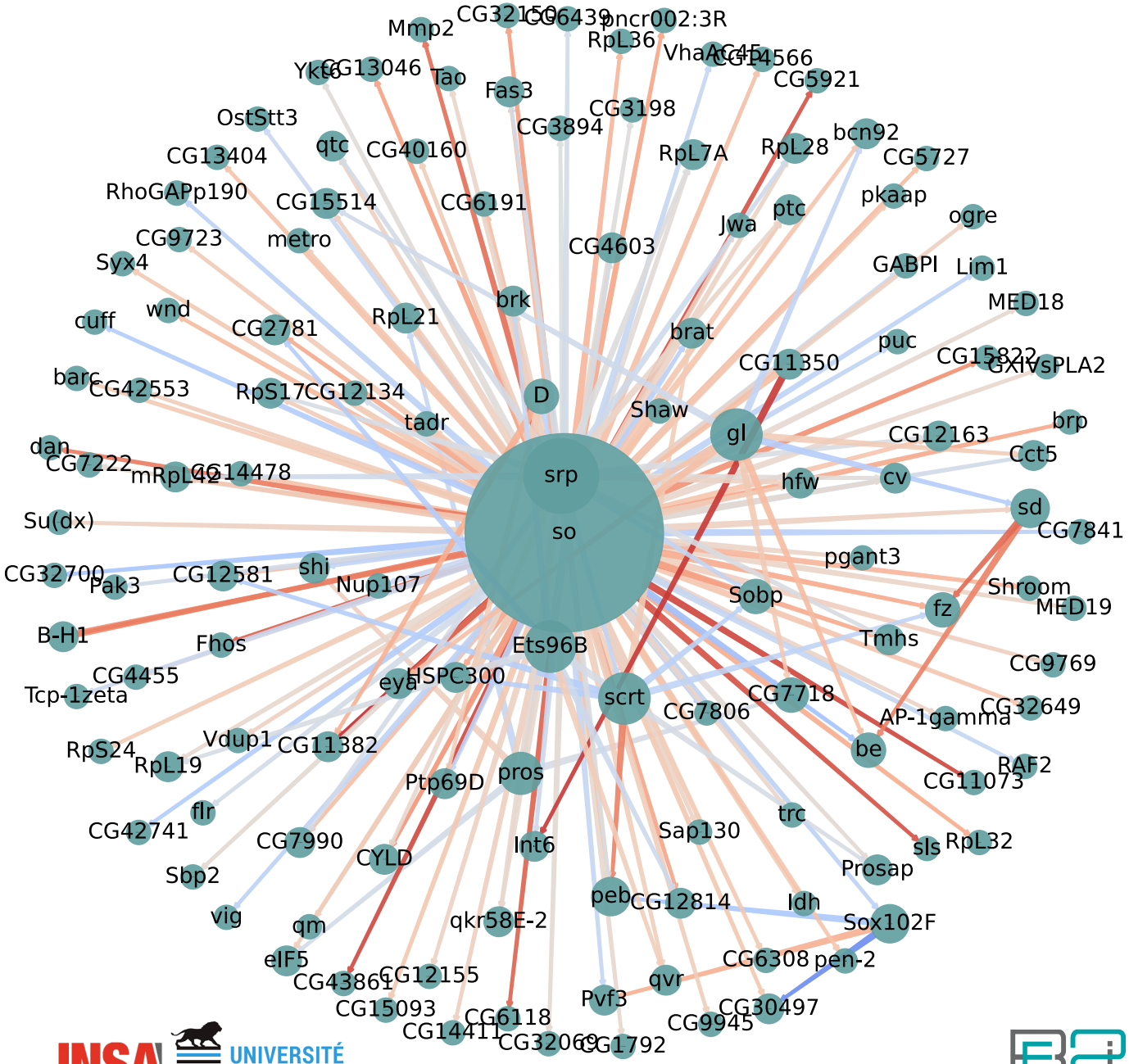
Results: *D. melanogaster* dataset - ey gene



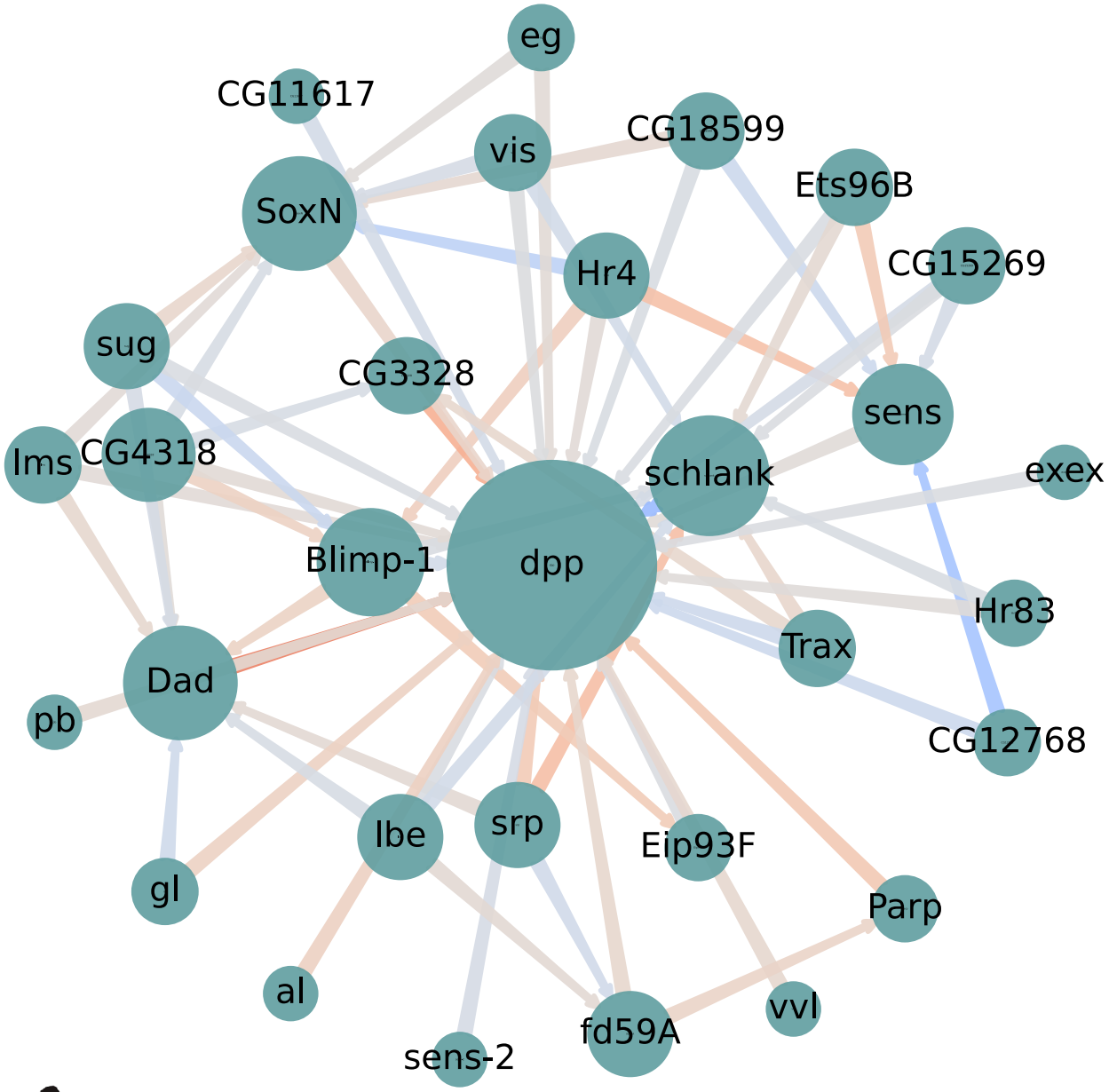
Results: *D. melanogaster* dataset - toy gene



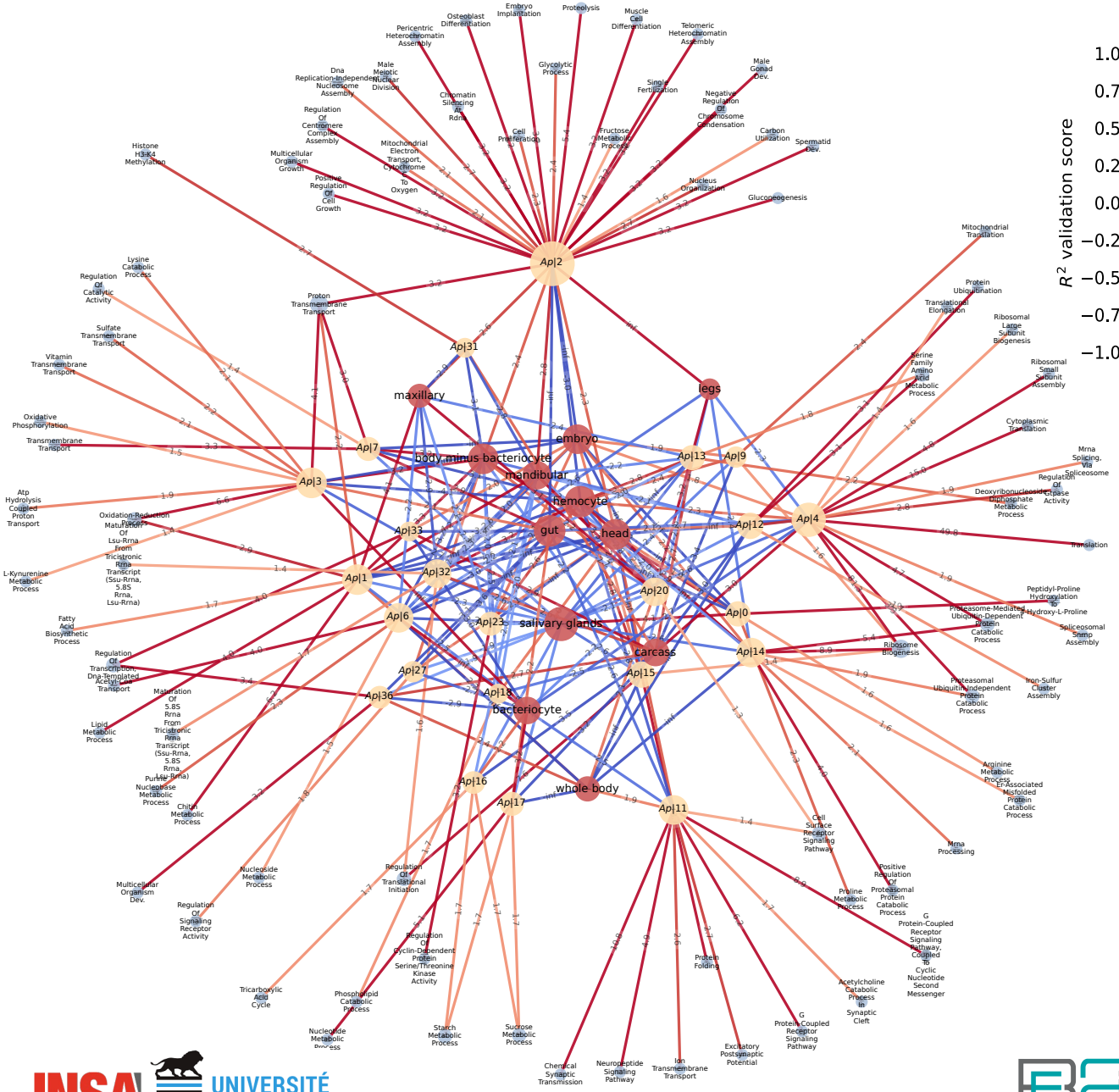
Results: *D. melanogaster* dataset - so gene

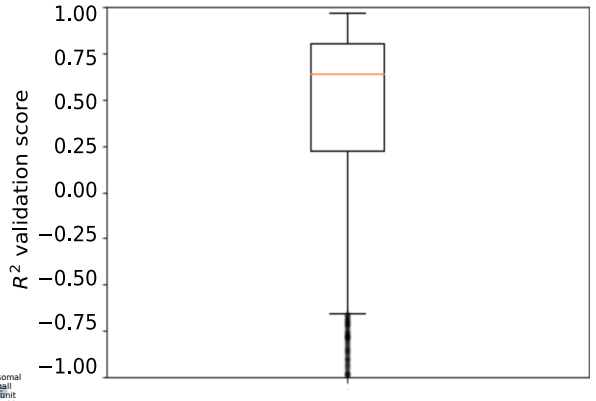


Results: *D. melanogaster* dataset - dpp gene



Results: *A. pisum* dataset – highly over-expressed only

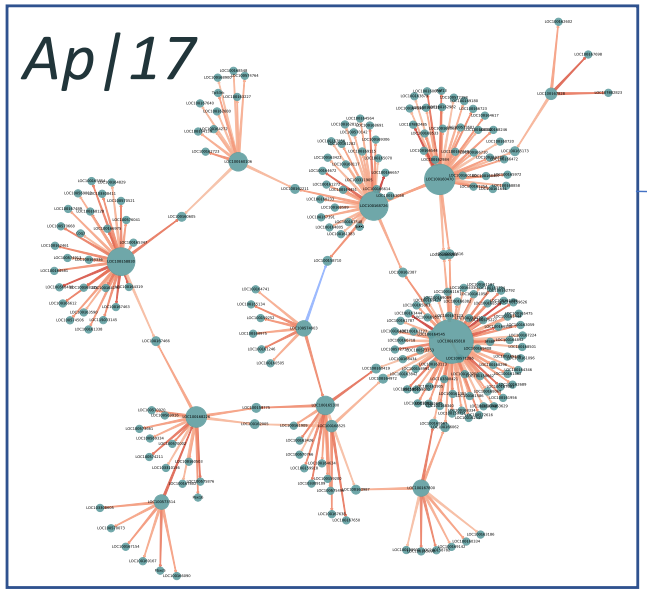


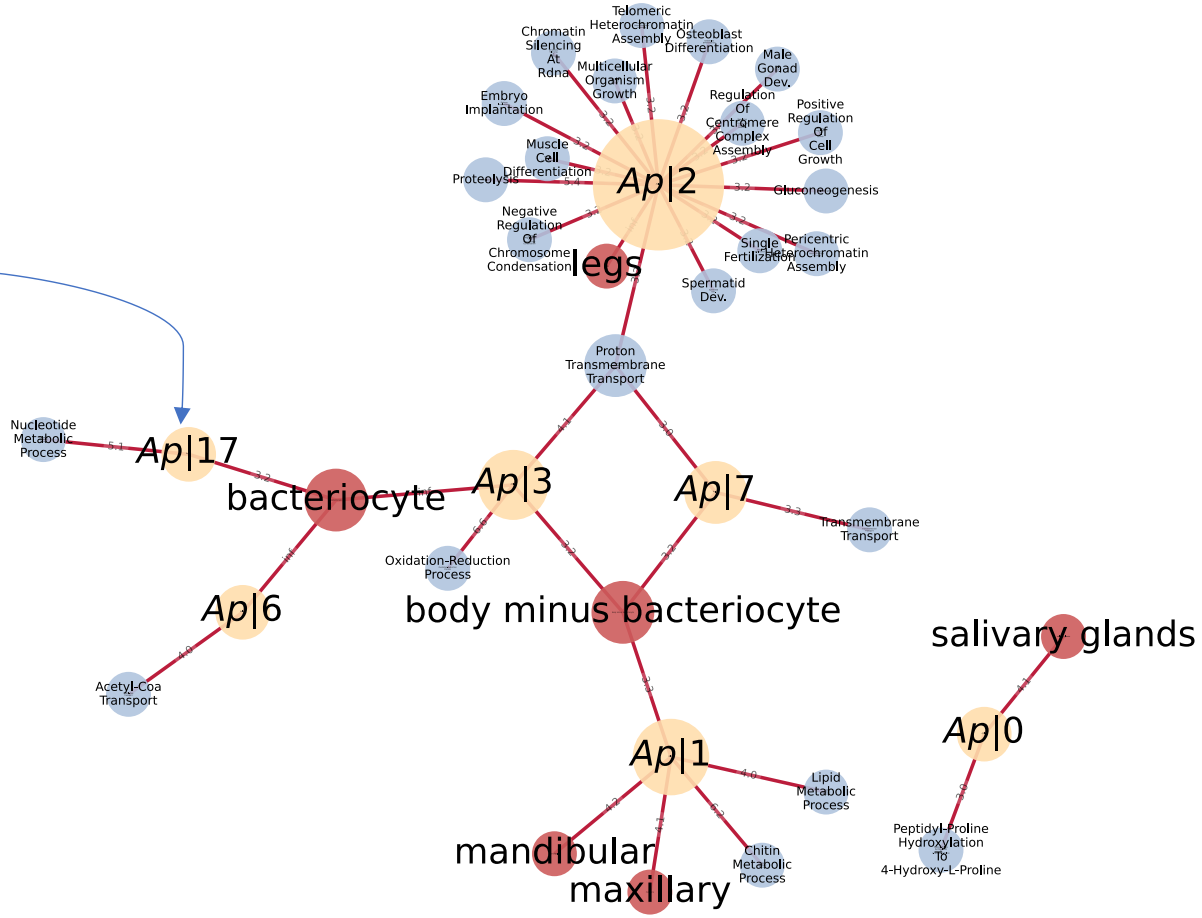


Modularity = 0.692

Sparsity = 98.4

Results: *A. pisum* dataset – highly over-expressed only





- **First representation of the GRN communities for different aphid tissues**
- **Discovery of bacteriocyte-specific GRN communities**

Statistically significant GO terms and positive GSEA scores between **GRN genes communities** and *A. pisum* **tissue-types**

Thanks to …

Nicolas Parisot

Patrick Callaerts

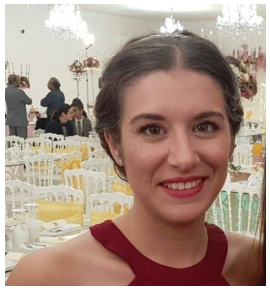
The students:

Pauline Schmitt (M1 INSA Lyon)

Baptiste Sorin (M1 INSA Lyon)

Timothée Frouté (M2 Lyon1)

Patrice Baa-Puyoulet



Mélanie Ribeiro Lopes

Gabrielle Duport

Karen Gaget

Hubert Charles

Financial support: BQR INSA Lyon, AAP FR BIOEEnViS, ANR HMicMac, Europe/FNRS IGut