Evaluating Meta-Feature Fidelity: Detecting Information Leakage in Interpretable Latent Spaces

Charlotte Job[1,2], Majd Saleh[2], Thierry Dorval[2], Farida Zehraoui[1,3], Blaise Hanczar[1]

[1]IBISC, University Paris-Saclay (Univ. Evry), Evry, France
[2]Institut Servier d'Innovation Thérapeutique, Gif−sur−Yvette, France
[3]LIFAT, Université de Tours, Tours, France

Outline

- **Context**
  - Research context
  - Concept Learning
  - Motivation
  - Meta-Feature definition
  - Problem and scientific question
- **Method**
  - Knock Out Simulation (KOS) metric definition
  - Distance Conservation (DC) metric definition
- **Results**
  - Experimental tests on simulated neurons using synthetic data
  - Use case on biologically constrained neurons using biological data
- **Conclusion / Perspectives**

ibisc université evry val-d'essonne université PARIS-SACLAY SERVIER moved by you

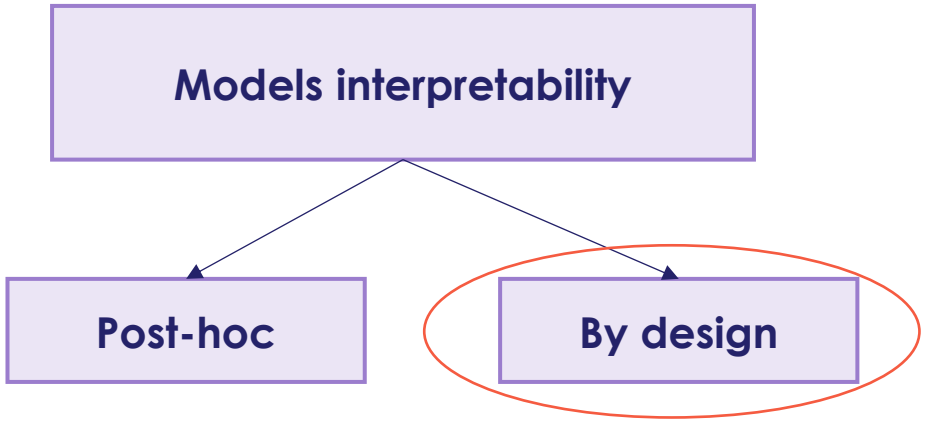
Research context

Deep Neural Networks are powerfull tools for predictive tasks in various domains

Deep Neural Networks are black-boxes: lack of trust and reliability

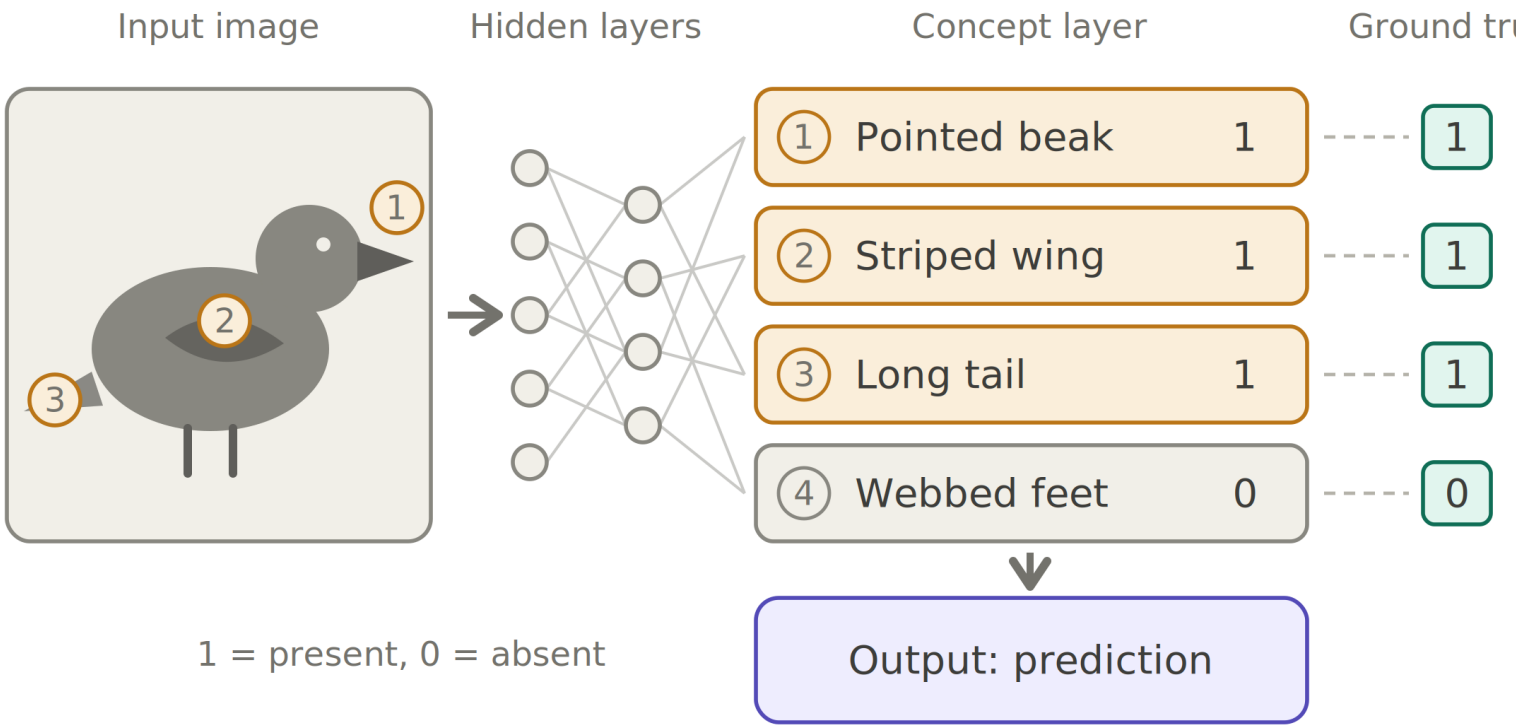
Model's interpretablility is necessary

New metrics to evaluate interpretability are needed



Concept Learning (CL)

Concept Learning (CL): mapping neurons to human-understandable entities (concepts)

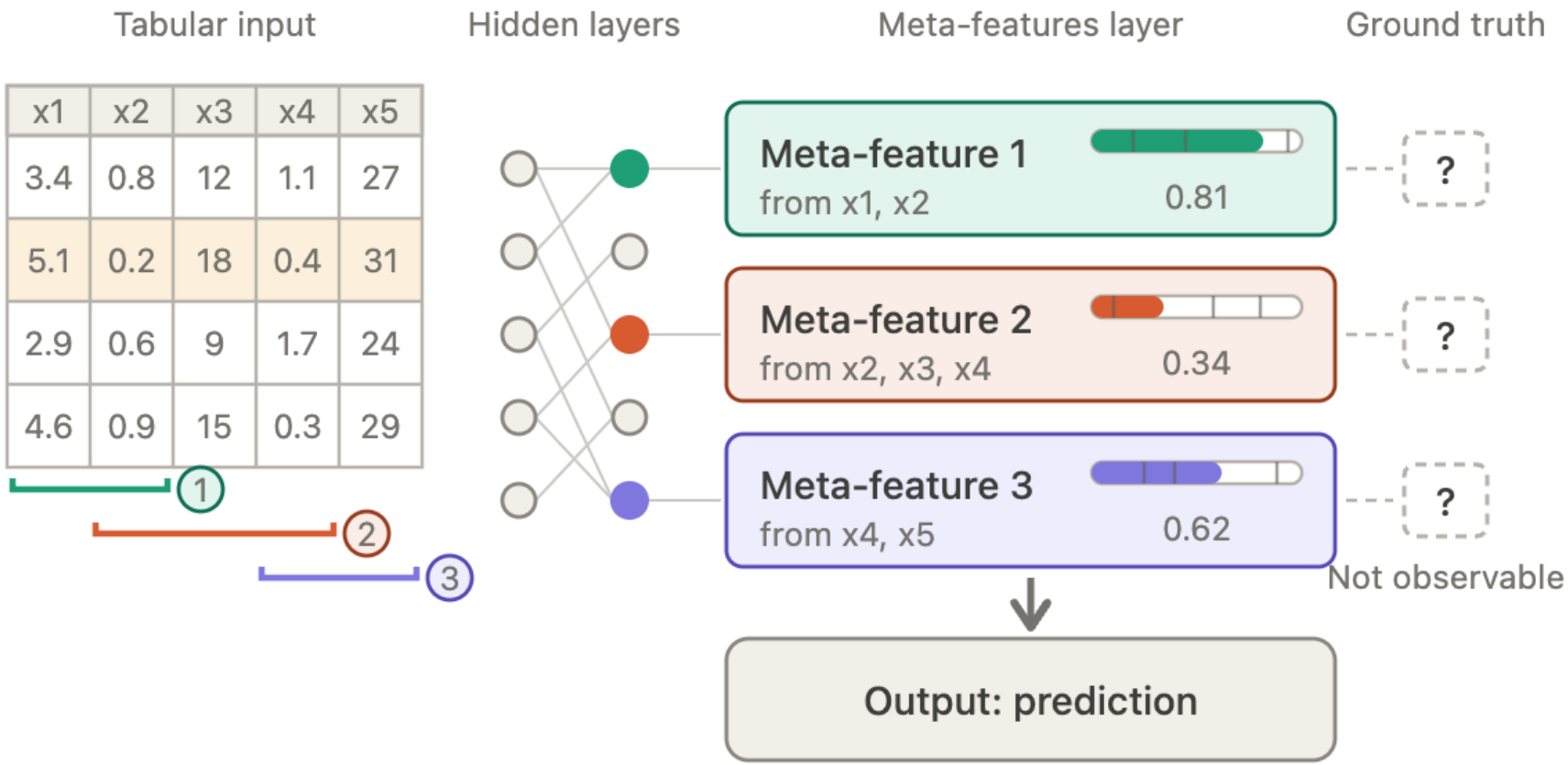


Concepts are:

- Top-down semantic characteristics of individual examples.
- Discrete attributes (presence or absence).
- Have a ground truth.

Meta-Feature Definition

Meta-feature (MF): specific type of concept.

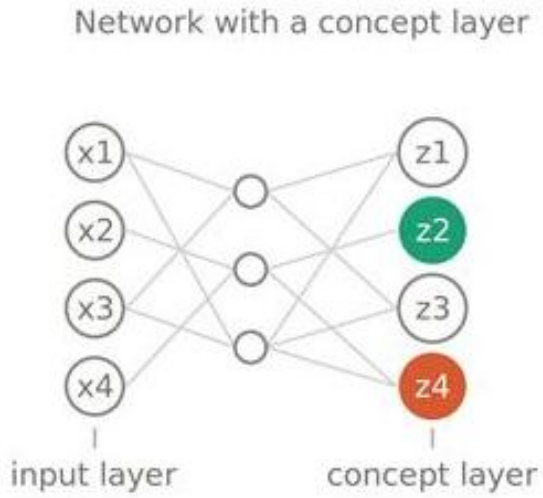


MFs are:

- Bottom-up concepts.
- Specific subsets of input variables.
- Continuous activation values present across all instances.
- Don't have any ground truth in real-world applications.

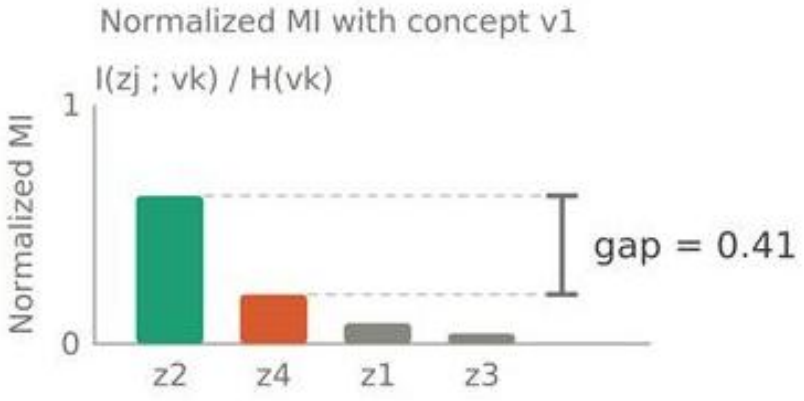
Motivation – Example of existing metric for CL

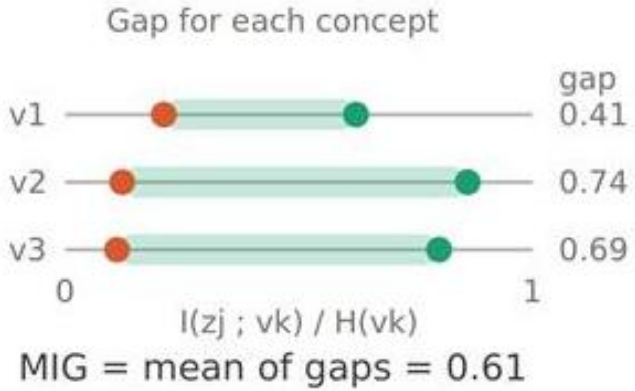
Mutual Information Gap (MIG) metric: computes the difference between the highest and second-highest mutual information values between neurons and true concept values.



Ground-truth concepts

k	v1 (round)	v2 (large)	v3 (red)
1	1	0	1
2	0	1	1
3	1	1	0
4	0	0	1





Chen, R. T., Li, X., Grosse, R. B., Duvenaud, D. K. (2018). Isolating sources of disentanglement in variational autoencoders. Advances in neural information pro-cessing systems, 31

Problems and Scientific Question

Interpretable models are evaluated for prediction or clustering tasks over the latent space.
Only some models have their interpretability evaluated with dedicated metrics.

↓

Current evaluation metrics that aim to evaluate if one neuron corresponds to one concept are based on ground truth factors like the metric MIG.

↓

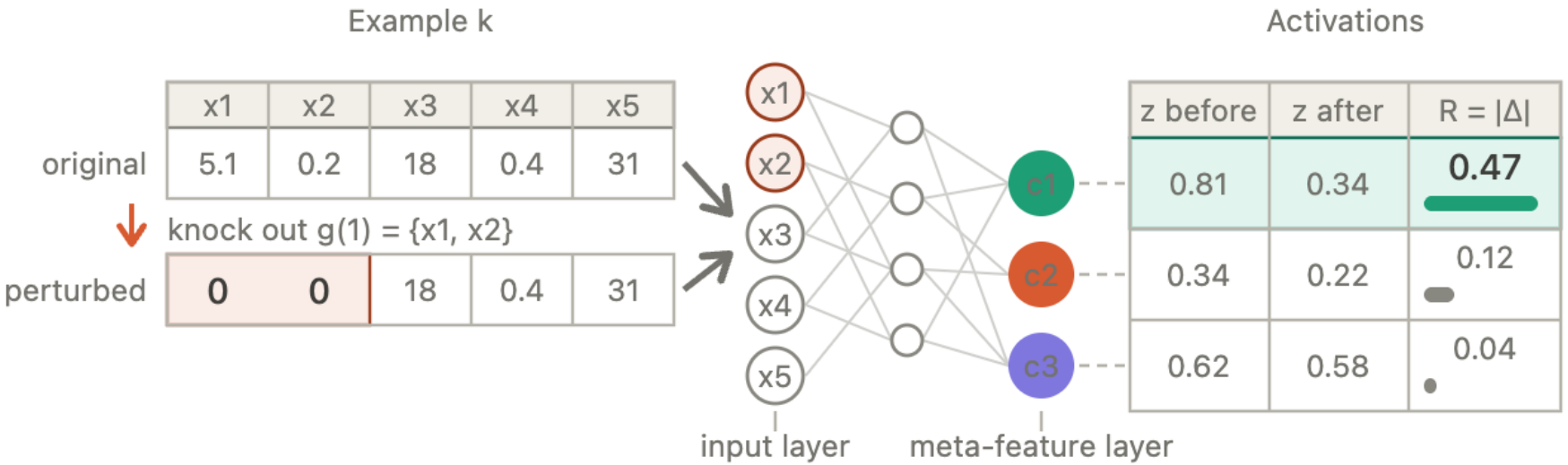
There is no specific metric defined for concepts like MFs.
Existing metrics for general concepts are not adapted to MFs.

↓

How to evaluate quantitatively to what extent one hidden neuron represents one specific MF ?

Method – Knock-Out Simulation Metric

Objective: evaluate if one neuron represents one known human-understandable MF.



Reduction score:

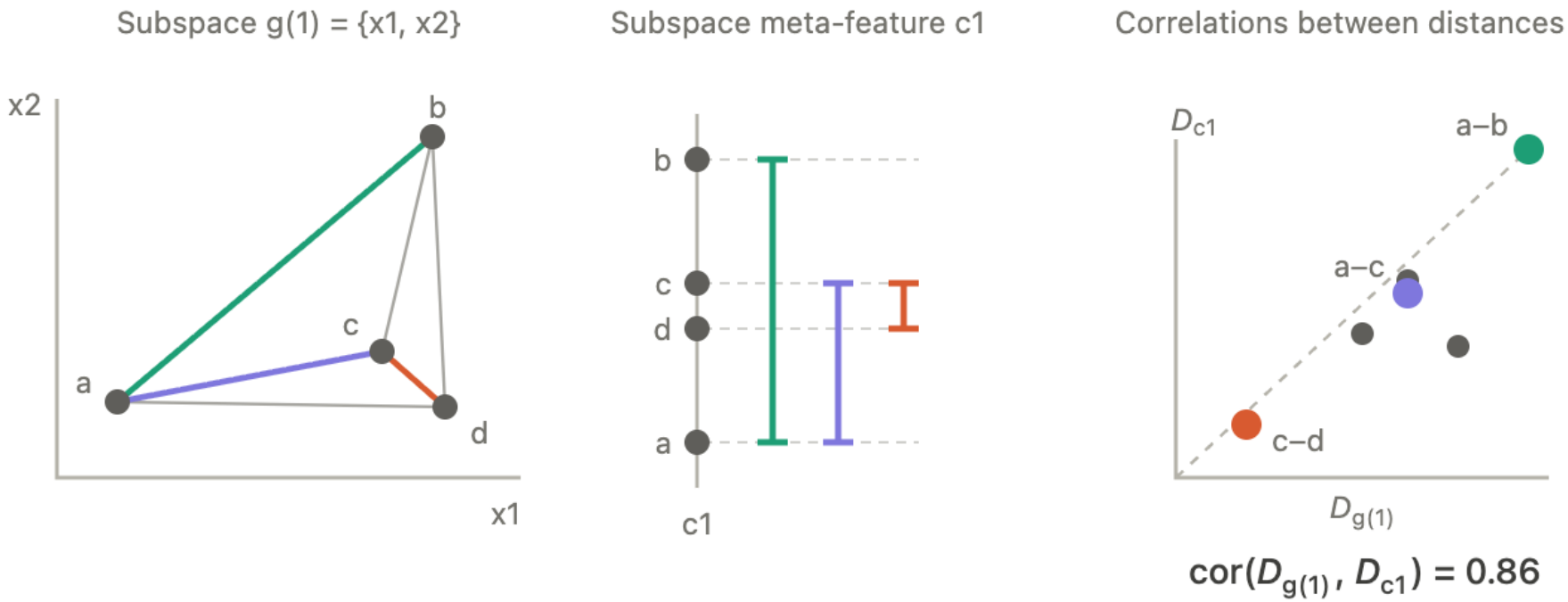
$$R_{c_i,k}^{(c_i)} = z_{c_i,k} - z_{c_i,k}^{(c_i)}$$

$$\boldsymbol{KOS_{c_i}} = \frac{\sum_{k=1}^{N}\sum_{j=1}^{K} \boldsymbol{I}[(\boldsymbol{R}_{c_i,k}^{(c_i)} > \boldsymbol{R}_{c_j \neq c_i,k}^{(c_i)})\text{^}(\boldsymbol{O}_{c_j,c_i} < \boldsymbol{t_0})]}{\sum_{k=1}^{N}\sum_{j=1}^{K} \boldsymbol{I}[\boldsymbol{O}_{c_j,c_i} < \boldsymbol{t_0}]}$$

- K: number of concepts.
- N: number of examples.
- Overlap between two concepts: O_{c_j,c_i}.

Method – Distance Conservation Metric

Objective: evaluate if relative distances between examples are preserved from the original space to the latent space.



$$DC = \frac{1}{K}\sum_{i=0}^{K} cor(D_{g(i)}, D_{c_i})$$

- K: number of concepts.
- $D_{g(i)}$: pairwise distances between examples in the input space restricted to variables $g(i)$.
- D_{c_i}: pairwise distances between examples in the dimension of concept c_i of the latent space.

Results – Experimental Tests on Synthetic Data

Neuron Simulation

Level of **activation of neuron z_{c_i}**:

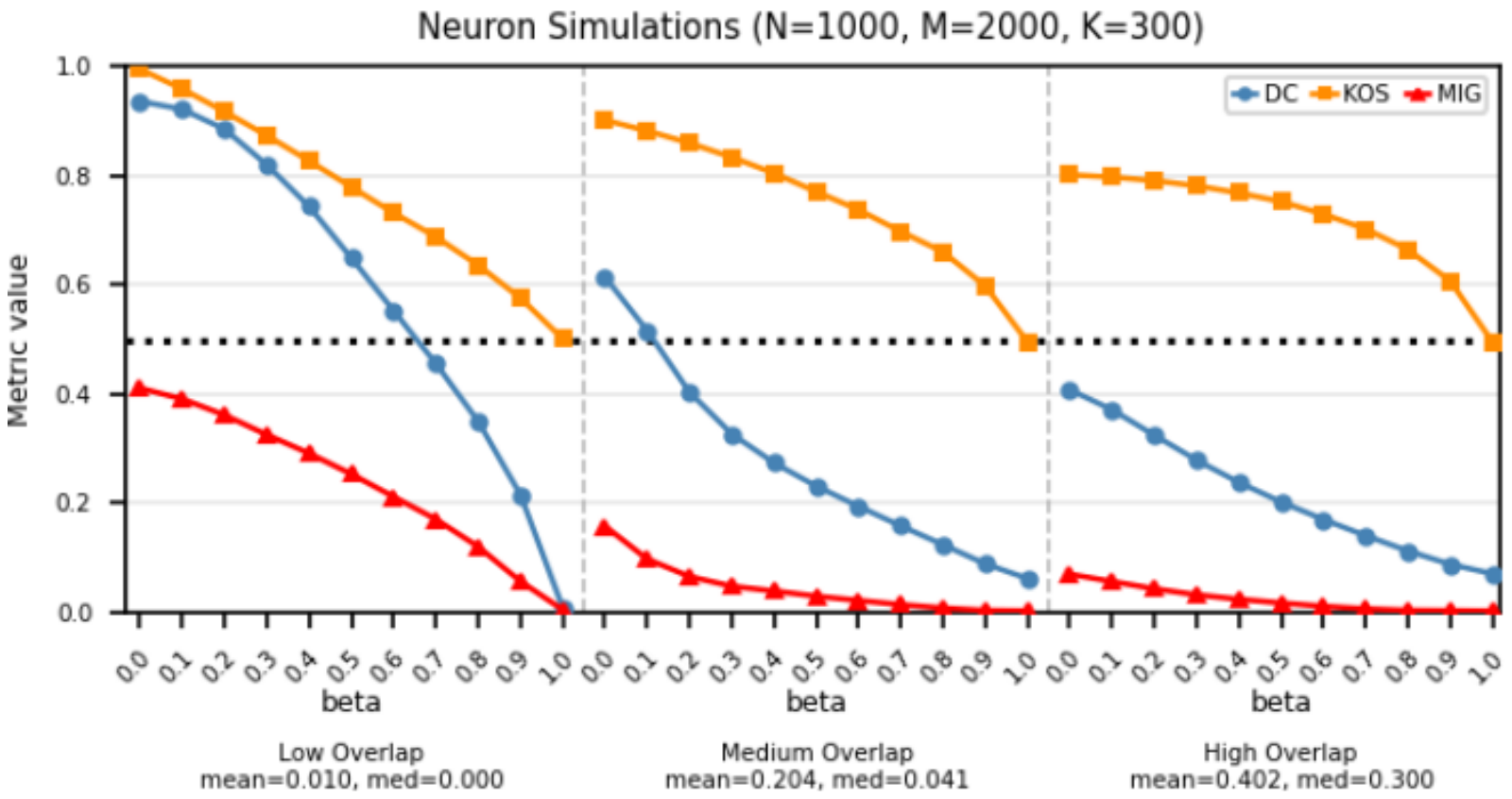
$$z_{c_i} = (1 - \beta) \sum_{j \in g(i)} x_j + \beta \sum_{k=1}^{K} \alpha_k \Big(\sum_{\substack{j \in g(k) \\ j != k}} x_j \Big)$$

- β: MF entanglement coefficient
- $\alpha_k = \frac{\alpha_{k'}}{\sum_{l=1}^{K} \alpha_{l'}} < here\ \alpha'_k \sim Exp(\alpha)$
- K: number of concepts

→ **Ideal case:** z_{c_i} depends only on variables belonging to MF c_i $(\beta = 0)$.

Results – Experimental Tests on Synthetic Data

Evolution of metrics KOS, DC and MIG with β evolution



- KOS: varies between 0.5 and 1.
- DC: Varies between 0 and 1.
- MIG: varies between 0 and 1.

→ **KOS and DC evaluate the interpretability of neurons representing MFs more effectively than MIG.**

Results – Use case on Biological Data

Single-cell RNA-seq data

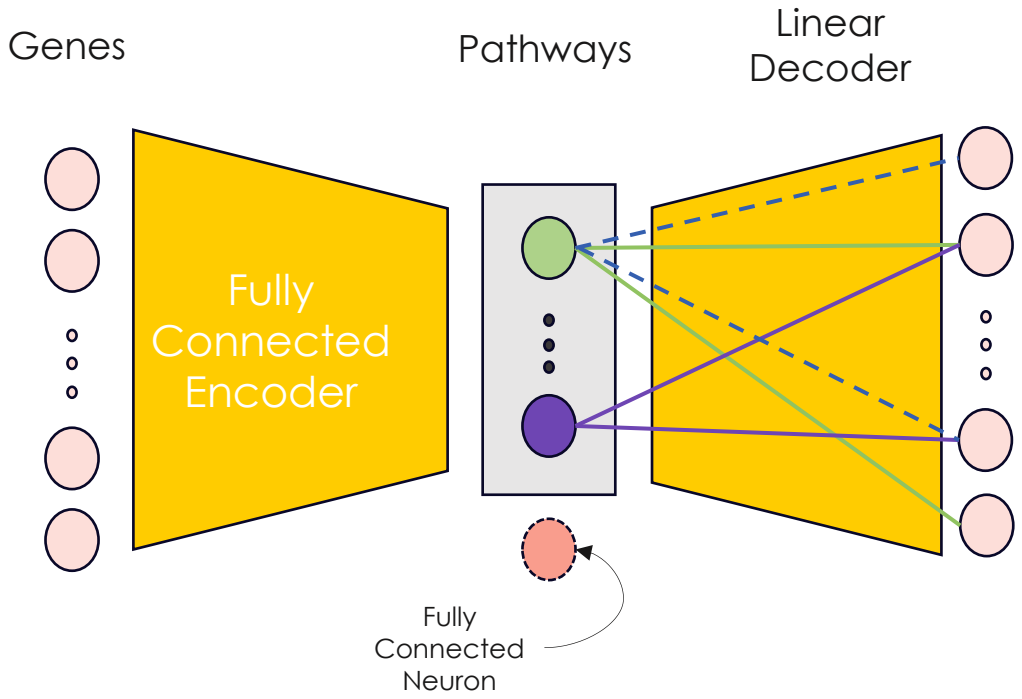
	$Gene_1$	$Gene_2$	…	$Gene_N$
$Cell_1$	0	5	2	…
$Cell_2$	10	0	0	…
…	15	0	0	…
$Cell_n$	…	…	…	…

$Pathway_1$: $Gene_4$, $Gene_{10}$, $Gene_{500}$, $Gene_{1200}$

…

$Pathway_1$: $Gene_{18}$, $Gene_{259}$

VEGA

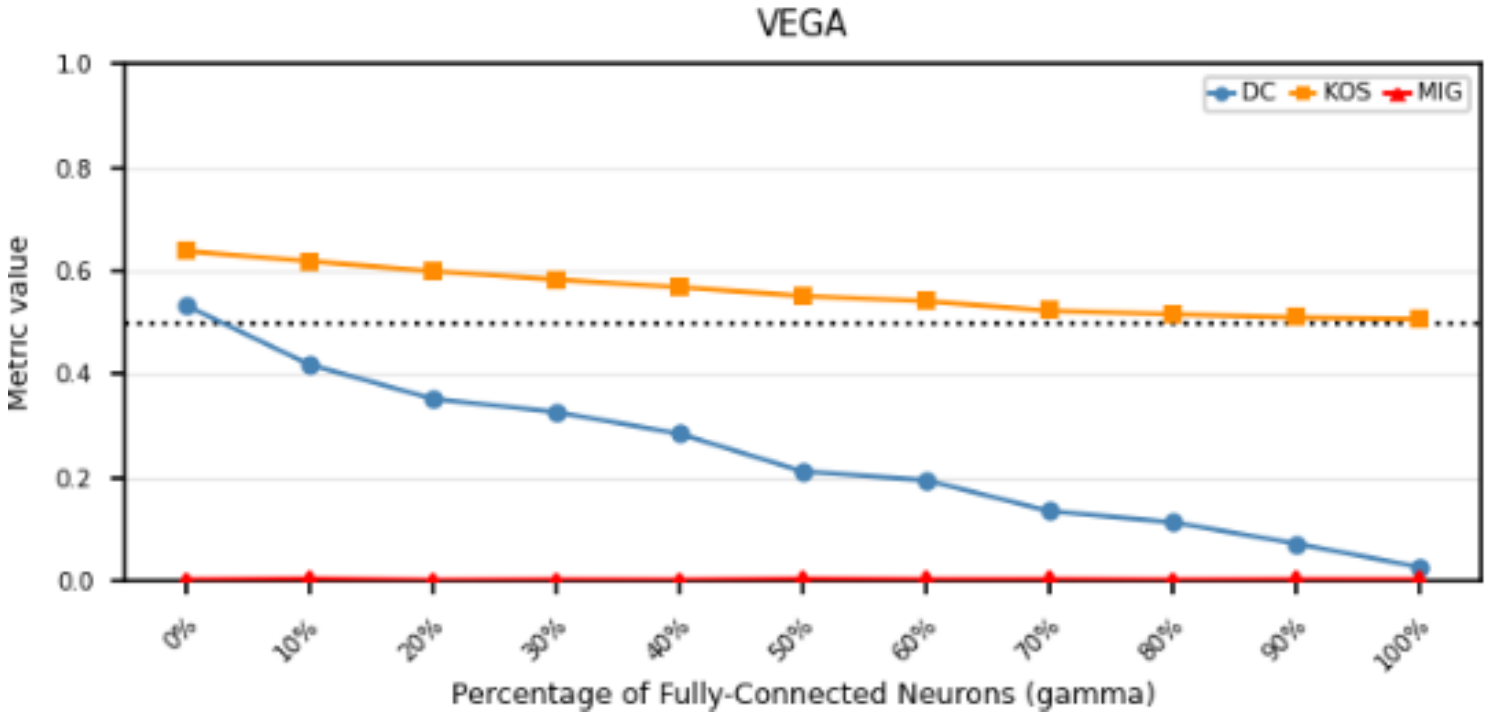


- - - Progressive addition of fully-connected (FC) neurons

Seninge, L., Anastopoulos, I., Ding, H., & Stuart, J. (2021). VEGA is an interpretable generative model for inferring biological network activity in single-cell transcriptomics. *Nature communications*

Results – Use case on Biological Data

Evolution of metrics KOS, DC and MIG with γ evolution



- γ controls the proportion of FC neurons.
- At $\gamma = 0$: sparse architecture.
- At $\gamma = 1$: FC architecture.
- Ground truth for MFs using GSEA.

→ KOS and DC successfully capture interpretability while MIG is not adapted to the specific scenario of MFs.

Conclusion and perspectives

Conclusion:

- MFs are dependent on variables and not on examples, and discrete ground truth factors can not describe continuous MFs.
- Existing metrics such as MIG are not adapted to the specific case where neurons are constrained to represent MFs.
- KOS and DC demonstrated their effectiveness and robusteness in assessing neuron's level of interpretability.

Perspectives:

- Employing KOS and DC as regularization terms during model training to enforce MFs disentanglement.

Questions & Answers

ibisc

université evry val-d'essonne

université PARIS-SACLAY

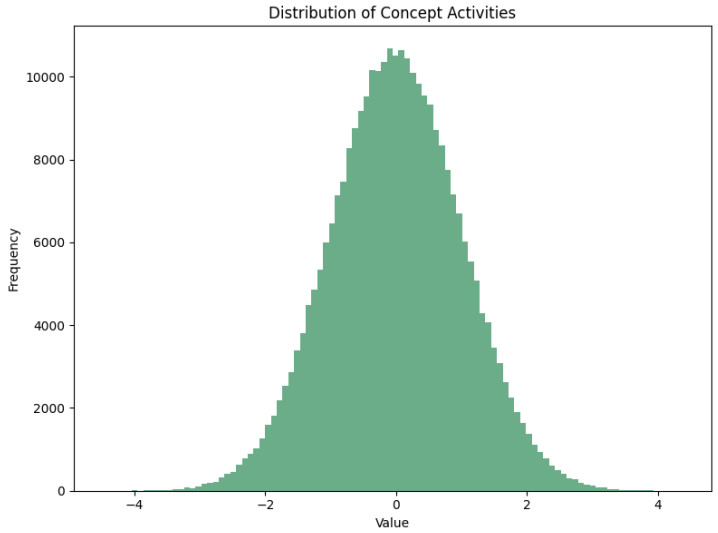
SERVIER
moved by you

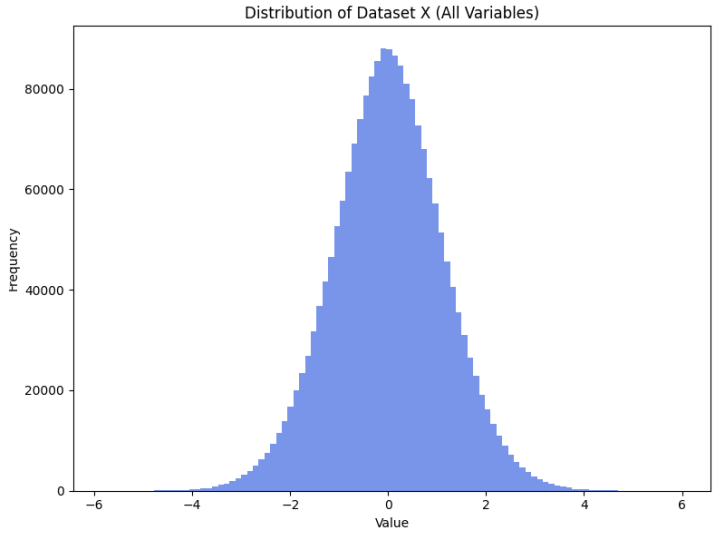
Results – Experimental Tests on Synthetic Data

Data simulation

MFs activity c_i: $\mu_{c_i} \sim N(0,1)$

Variables: x_i: $x_i \sim N(\mu, \sigma^2 I)$ $with\ \mu_i = \frac{1}{|g(i)|}\sum_{j \in g(i)} \mu_{c_j}$





Results – Experimental Tests on Synthetic Data

Data simulation

MFs sizes distribution