Kernel-based perturbation response analysis in single-cell omics

Polina Arsenteva[1,2] & *Anthony Ozier-Lafontaine*[2] & *Franck Picard*[1] & *Bertrand Michel*[2]

[1]*polina.arsenteva@ens-lyon.fr*

[1] ***Laboratoire de biologie et modélisation de la cellule***
Université de Lyon, ENS de Lyon, Lyon, France
[2] ***Laboratoire de Mathématiques Jean Leray***
Nantes Université - École Centrale de Nantes, Nantes, France

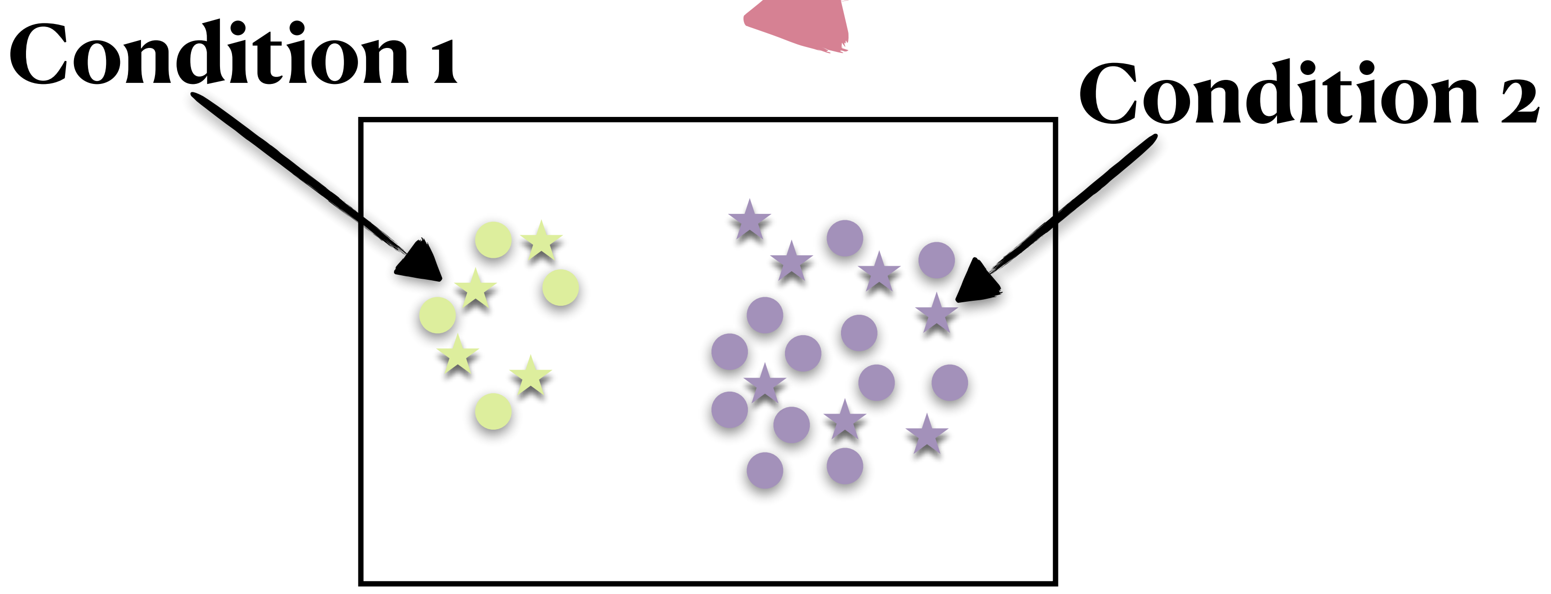
Perturbation experiments on cells

Change conditions and modify the basal cell state through a controlled external event

→

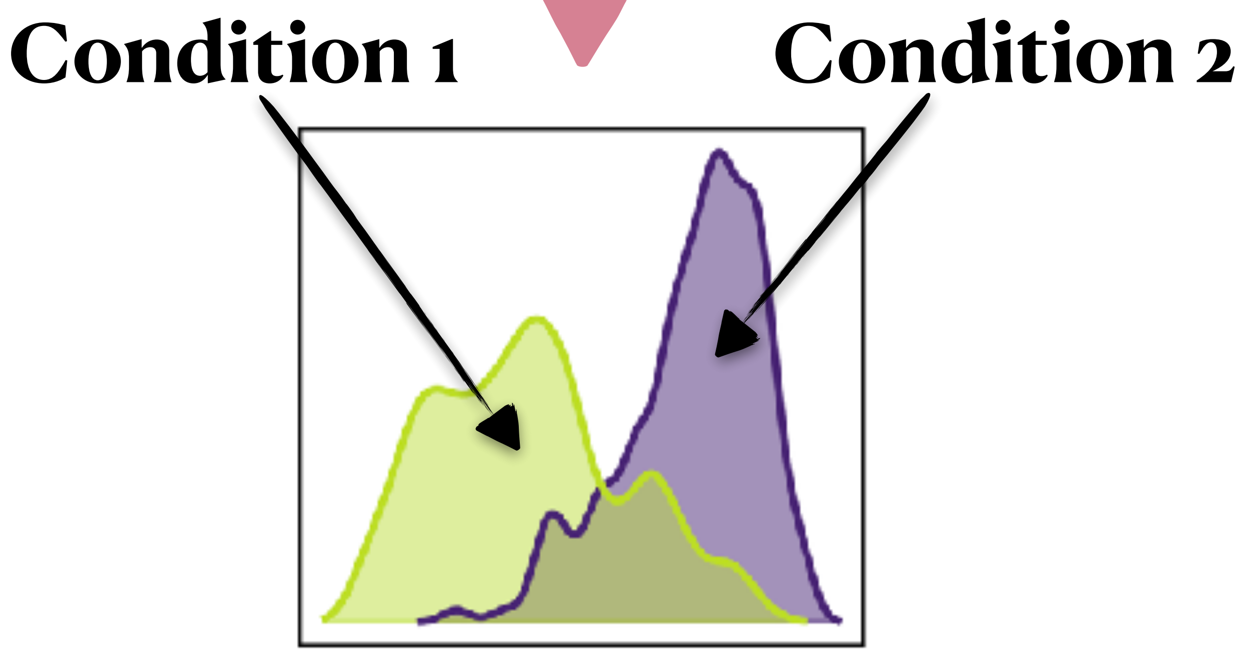
? Impact of the perturbation on gene expression/cell type?

Three types of approaches:



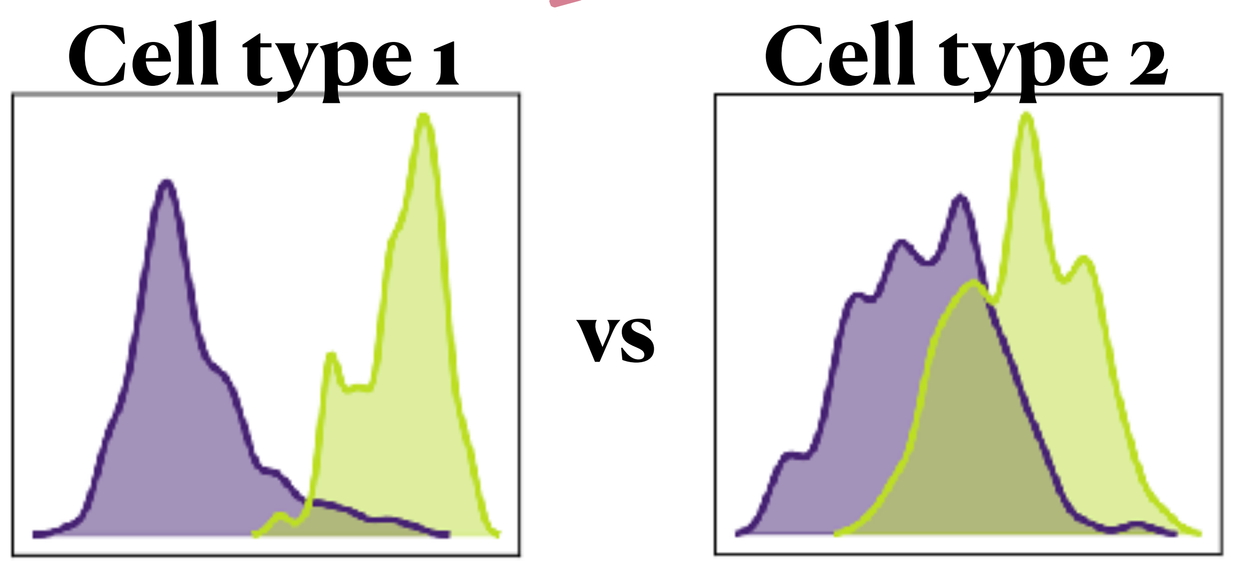
Differential Abundance

Detect differences in the cell type **composition**



Differential Expression

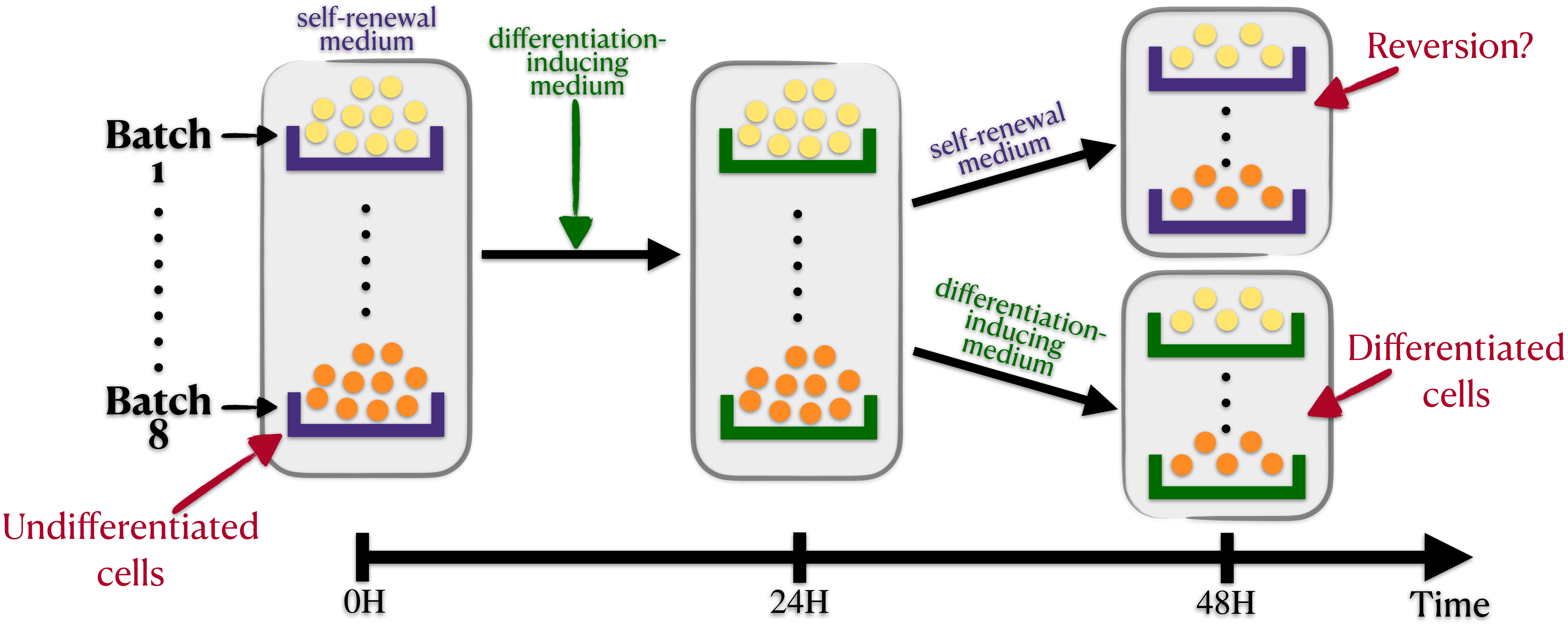
Detect differences in **gene expression**



Differential State

Detect differences in **cell types'** responses

Application to experimental data



? What is the transcriptomic impact of the medium perturbation on cell differentiation?

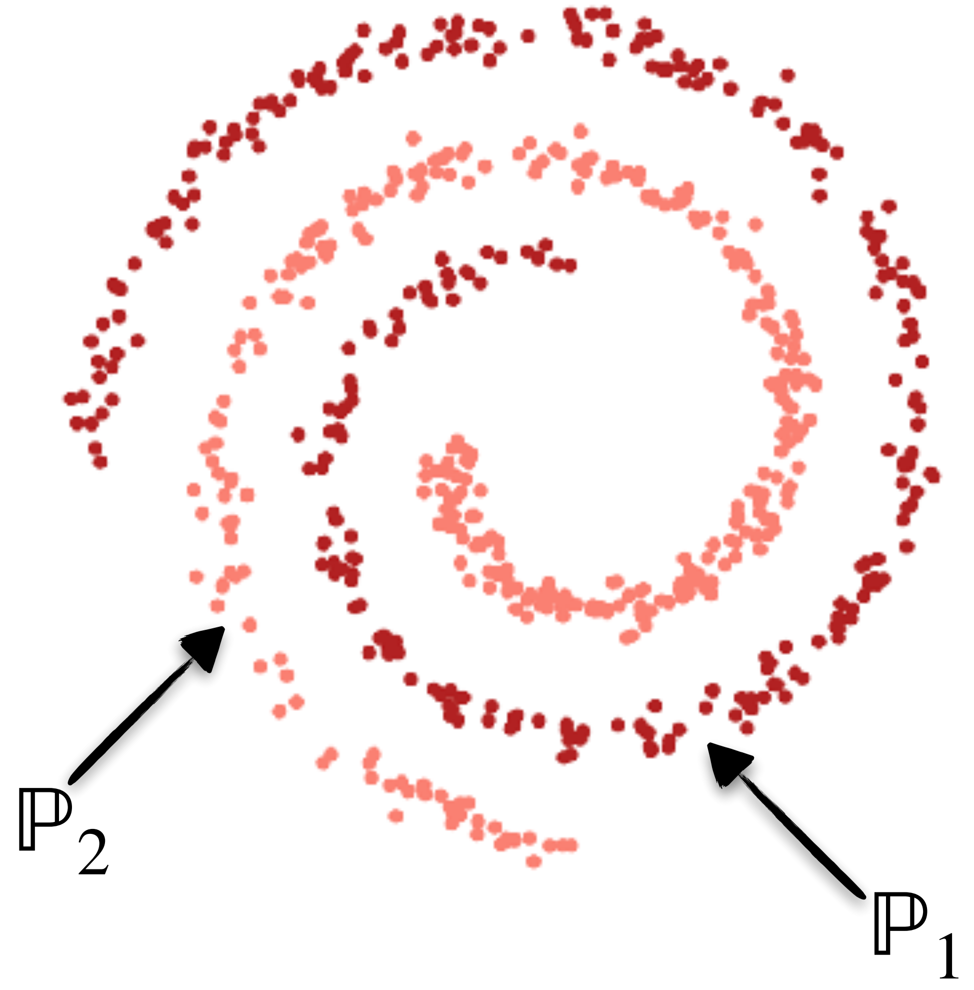
? Are the reverted cells close to undifferentiated?

Kernel testing for perturbation analysis

Complex nature of the gene-expression distribution ➔ **non-linear approach is required**

Original data

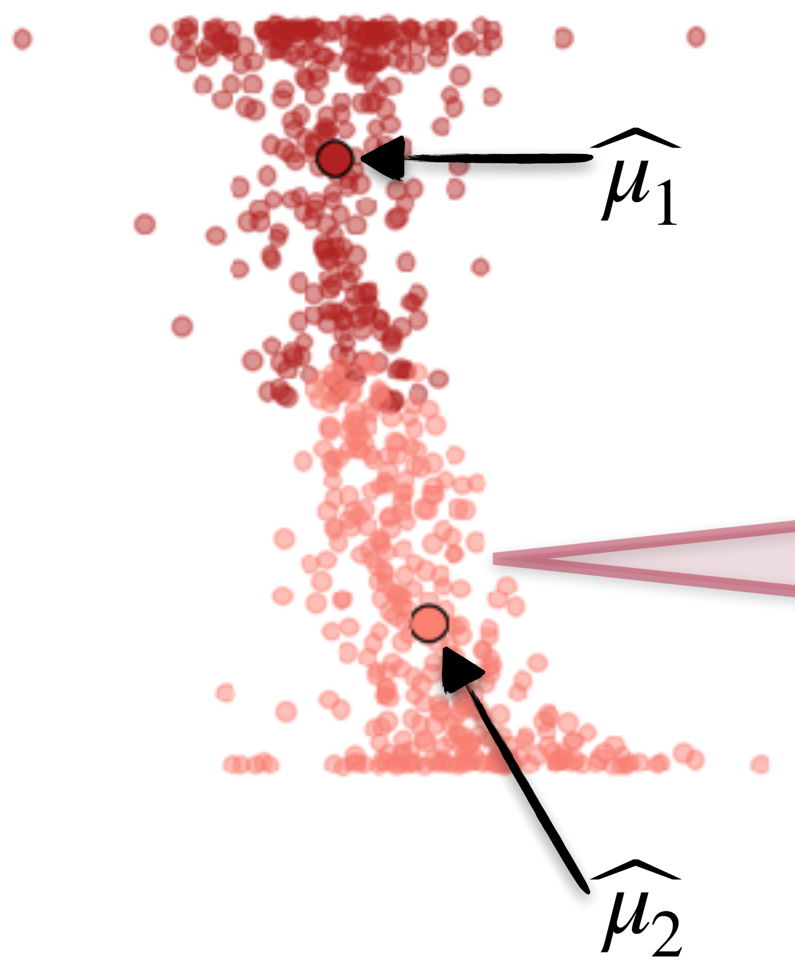
$Y_1, \ldots, Y_n \in \mathcal{Y}^{n \times d}$



kernel $k(\cdot, \cdot)$
such that
$k(Y_i, \cdot) = \phi(Y_i) \in \mathcal{H}$

Embedded data

$\phi(Y_1), \ldots, \phi(Y_n) \in \mathcal{H}^n$ ← **Reproducing Kernel Hilbert Space (RKHS)**



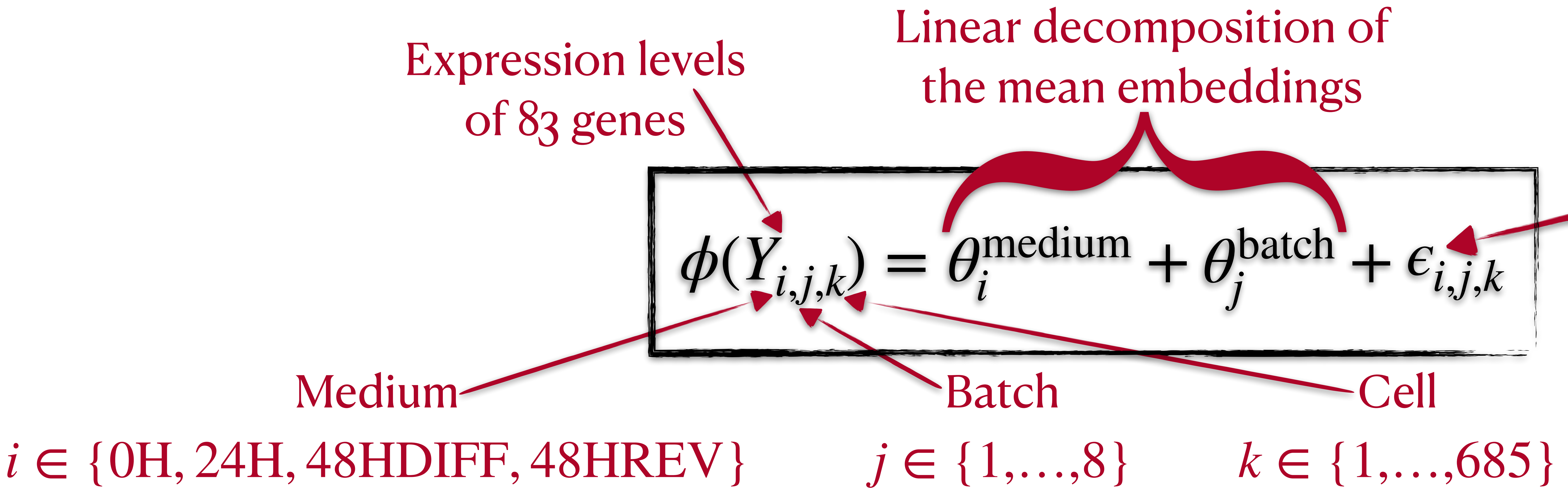
Non-linear for the original data, but linear in the feature space

Simple two-sample test:

$$H_0 : \{\mathbb{P}_1 = \mathbb{P}_2\} \iff \{\mu_1 = \mu_2\}$$

Generalizing kernel testing to any design

- Embed the expression data into the feature space
- Regress the embeddings against:
  - ✓ the medium effect (i.e. **perturbation**)
  - ✓ the batch effect



Expression levels of 83 genes

Random error

$$\boxed{\phi(Y_{i,j,k}) = \overbrace{\theta_i^{\text{medium}} + \theta_j^{\text{batch}}}^{\text{Linear decomposition of the mean embeddings}} + \epsilon_{i,j,k}}$$

Medium

$i \in \{0\text{H}, 24\text{H}, 48\text{HDIFF}, 48\text{HREV}\}$

Batch

$j \in \{1,\ldots,8\}$

Cell

$k \in \{1,\ldots,685\}$

- **Matrix formulation** in the feature space: $\Phi(\mathbf{Y}) = \mathbf{X\Theta} + \mathbf{E}$

Perturbation analysis via statistical tests

- **Differential Expression Analysis via pairwise test:**

$$\boxed{H_0 : \theta_i^{\text{medium}} = \theta_{i'}^{\text{medium}}}$$

$(i, i') \in \{0\text{H}, 24\text{H}, 48\text{HDIFF}, 48\text{HREV}\}^2,$
$i \neq i'$

- Kernel T-truncated Hotelling-Lawley test statistic: $\mathscr{F}_T = \mathrm{Tr}(\widehat{\Sigma}^{-}_{T}\widehat{H}_{\mathbf{L}})$

$\approx \mathrm{Cov}(\widehat{\mathbf{E}})^{-1}$

A test operator encoding
the test hypothesis

- $\mathscr{F}_T$ is asymptotically χ^2_{dT} ➔ **asymptotic p-values**

Perturbation analysis via statistical tests

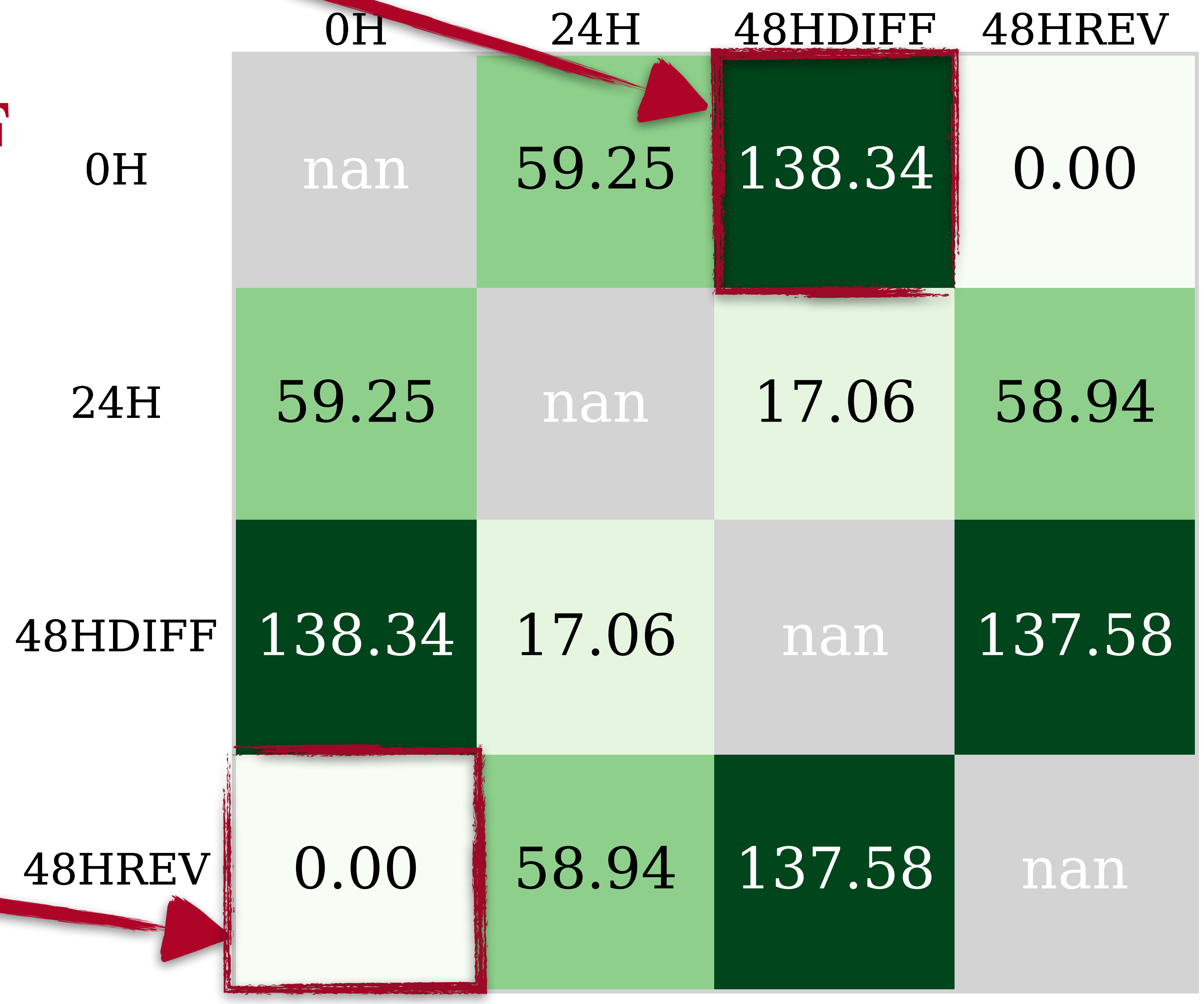
- **Differential Expression Analysis via pairwise test:**

$$\boxed{H_0 : \theta_i^{\text{medium}} = \theta_{i'}^{\text{medium}}}$$

$$(i, i') \in \{\text{0H}, \text{24H}, \text{48HDIFF}, \text{48HREV}\}^2,$$
$$i \neq i'$$

The most distinct medium pair:
0H and 48HDIFF

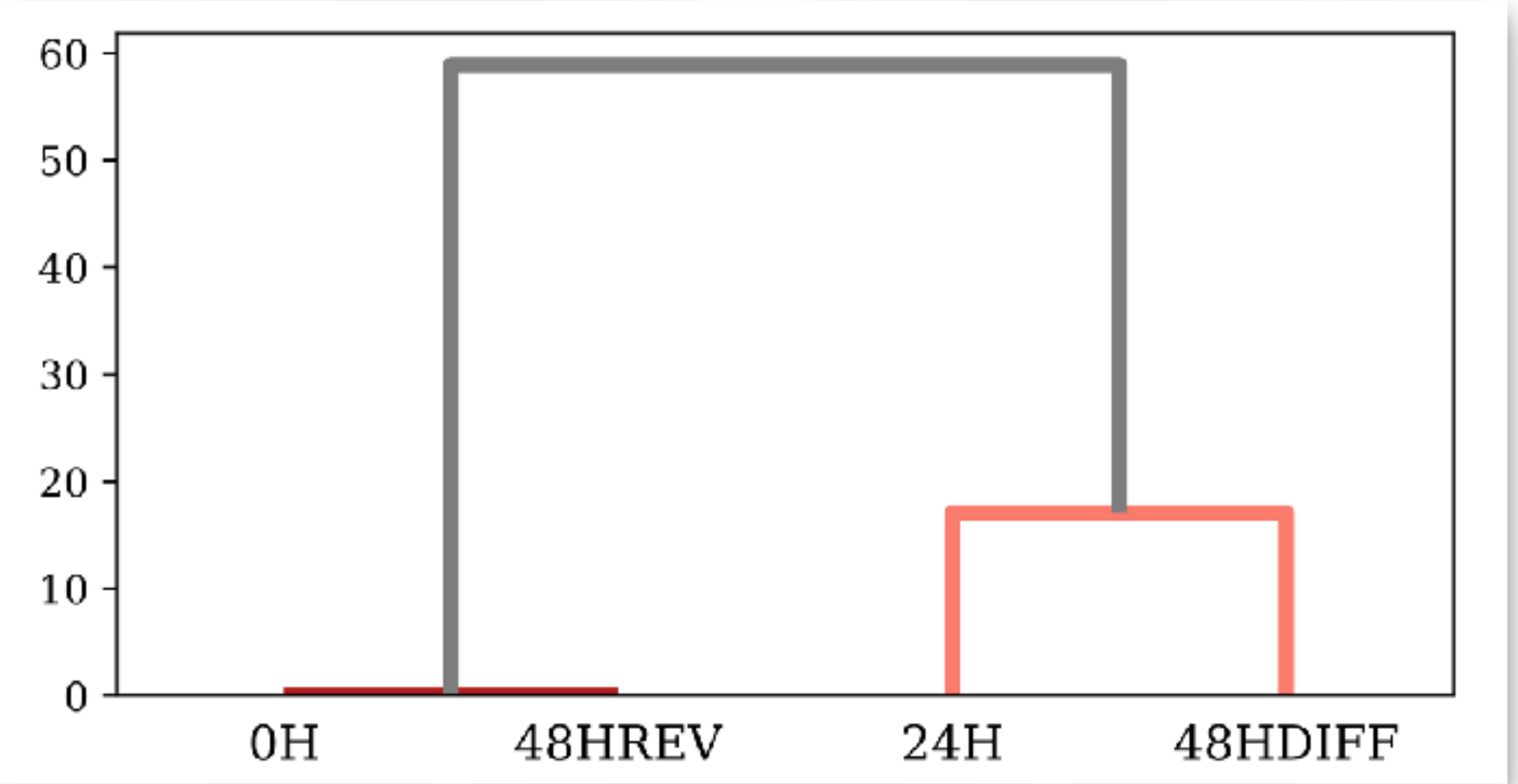
$\mathscr{F}_T =$



	0H	24H	48HDIFF	48HREV
0H	nan	59.25	138.34	0.00
24H	59.25	nan	17.06	58.94
48HDIFF	138.34	17.06	nan	137.58
48HREV	0.00	58.94	137.58	nan

The closest medium pair:
0H and 48HREV

Pairwise Hotelling-Lawley test statistic values



Hierarchical clustering based on the pairwise test statistic

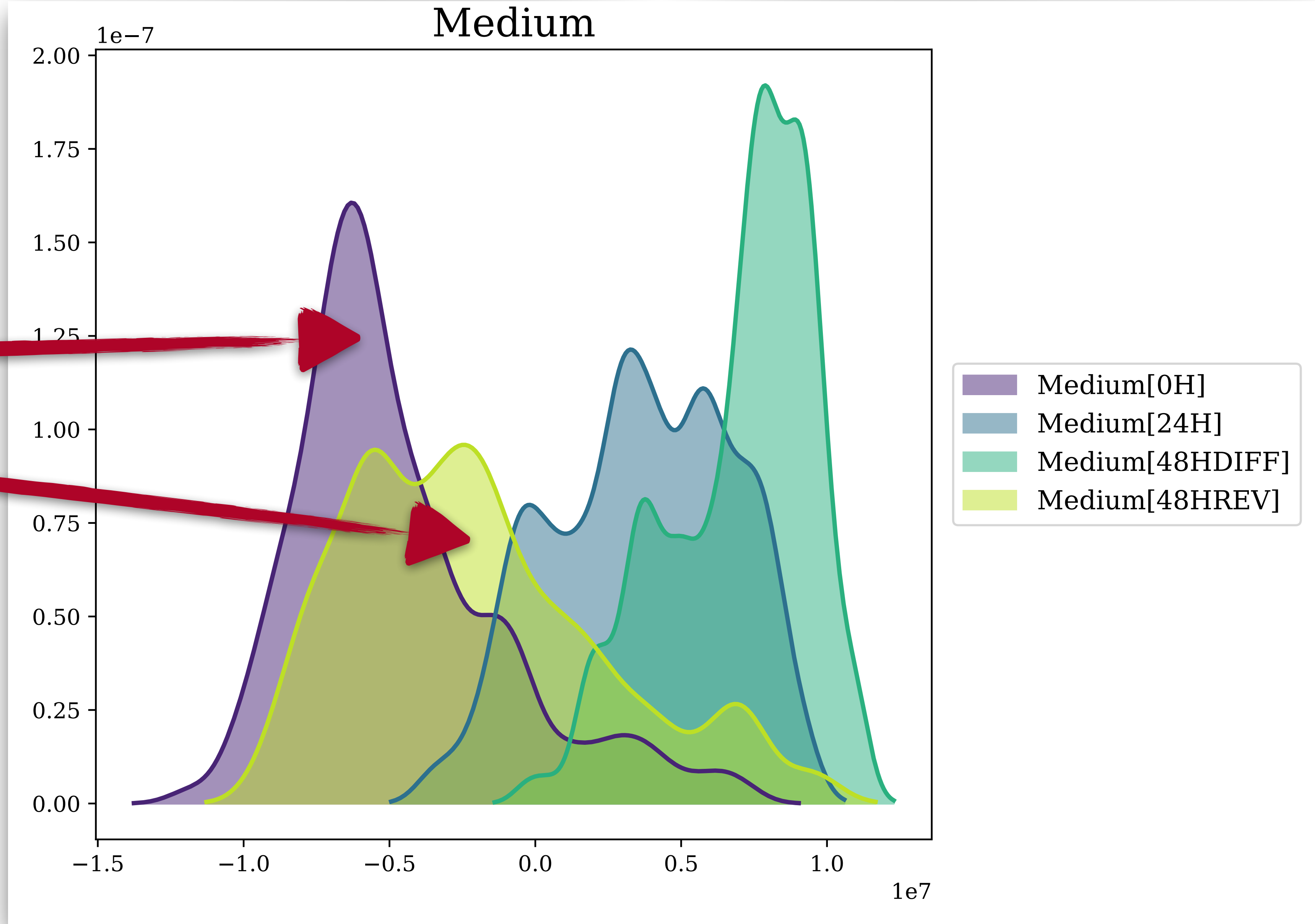
Projections on discriminant axes

- Project the cell embeddings onto the eigenfunctions of $\widehat{\Sigma}_T^{-1}\widehat{H}_{\mathbf{L}}$

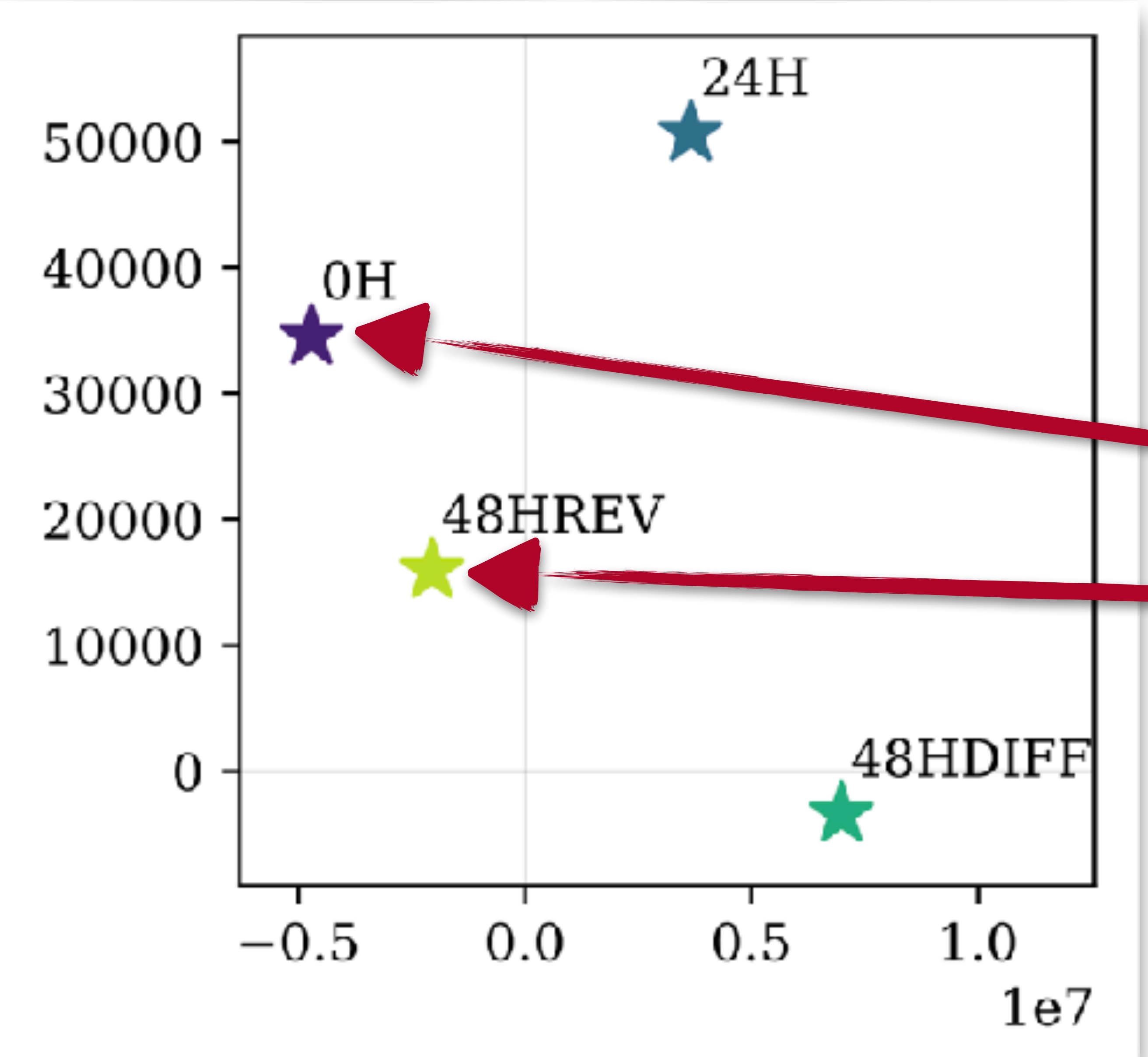
“Discrimination” operator associated with the test

- Test for the **medium effect**: $H_0 : \theta^{\text{medium}}_{\text{0H}} = \theta^{\text{medium}}_{\text{24H}} = \theta^{\text{medium}}_{\text{48HDIFF}} = \theta^{\text{medium}}_{\text{48HREV}}$

Proximity between 0H and 48HREV



Densities of the embeddings’ projections



Projections of the mean embeddings

Proximity between 0H and 48HREV

From DEA to DSA

Differences in **gene expression**

Differences in **cell types'** responses

- Differential **expression** analysis (DEA):

$$\phi(Y_{i,k}) = \theta_i^{\text{condition}} + \epsilon_{i,k}$$

- Differential **state** analysis (DSA):

Cell type fixed effect

$\theta_l^{\text{celltype}} \in \mathscr{H}$

Cell type fixed effect for a given condition

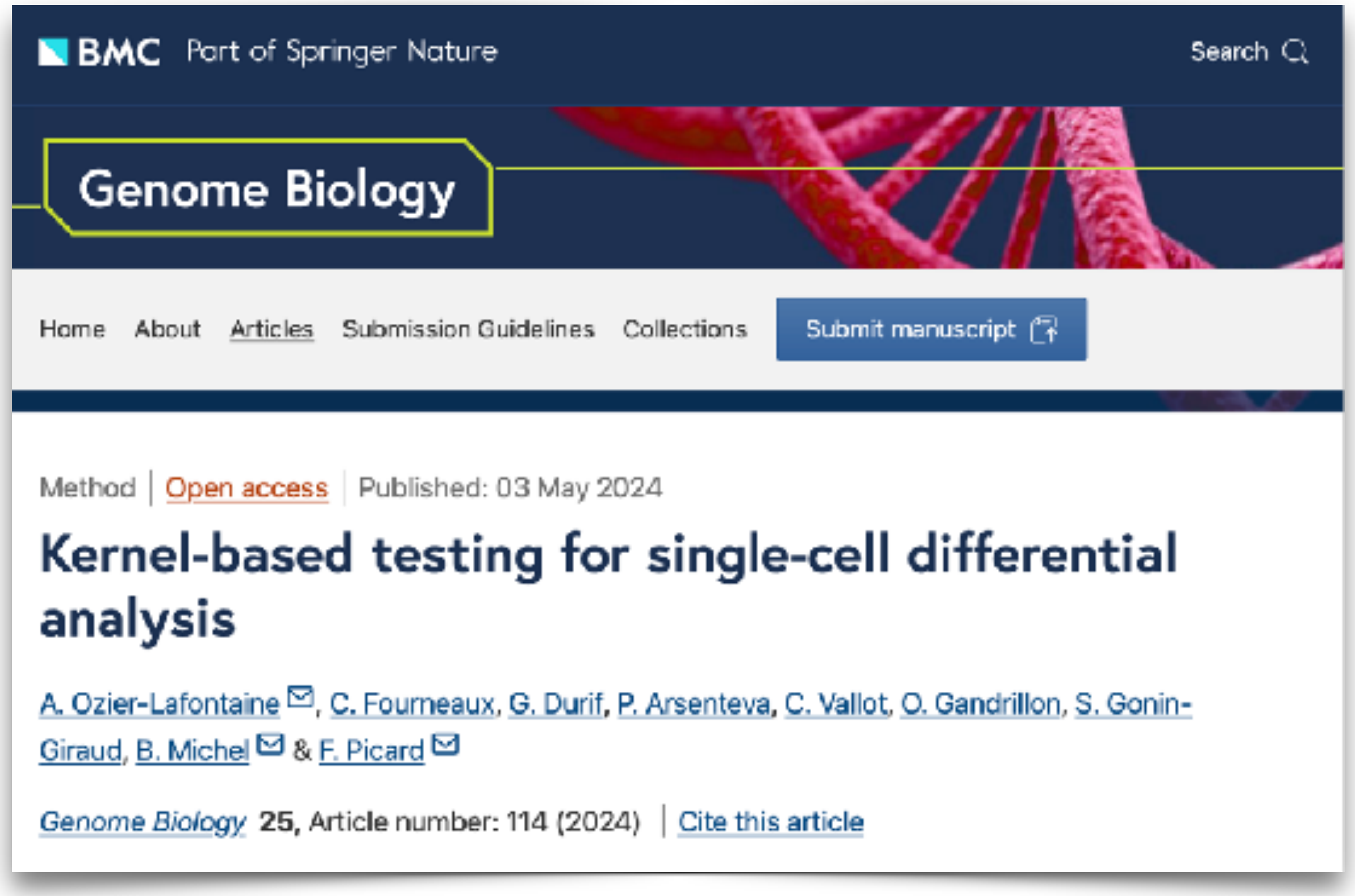
$\theta_{i,l}^{\text{condition\&celltype}} \in \mathscr{H}$

$$\phi(Y_{i,k,l}) = \theta_i^{\text{condition}} + \theta_l^{\text{celltype}} + \theta_{i,l}^{\text{condition\&celltype}} + \epsilon_{i,k,l}$$

- **Test (pairwise):** $\cancel{H_0 : \theta_i^{\text{condition}} = \theta_{i'}^{\text{condition}}}$ $\boxed{H_0 : \theta_{i,l}^{\text{condition\&celltype}} = \theta_{i',l'}^{\text{condition\&celltype}}}$

Two-sample approach

Paper



Method | Open access | Published: 03 May 2024

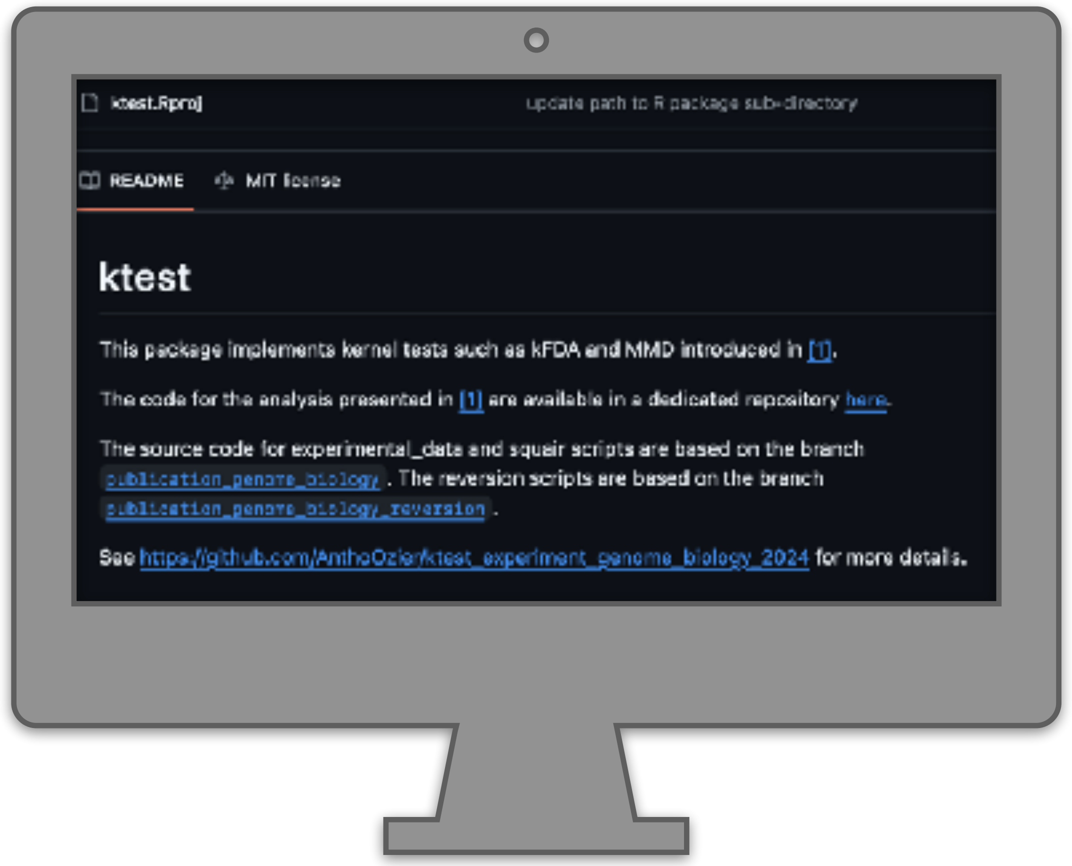
Kernel-based testing for single-cell differential analysis

A. Ozier-Lafontaine, C. Fourneaux, G. Durif, P. Arsenteva, C. Vallot, O. Gandrillon, S. Gonin-Giraud, B. Michel & F. Picard

Genome Biology 25, Article number: 114 (2024) | Cite this article

+

Python + R library **ktest**



ktest

This package implements kernel tests such as kFDA and MMD introduced in [1].

The code for the analysis presented in [1] are available in a dedicated repository here.

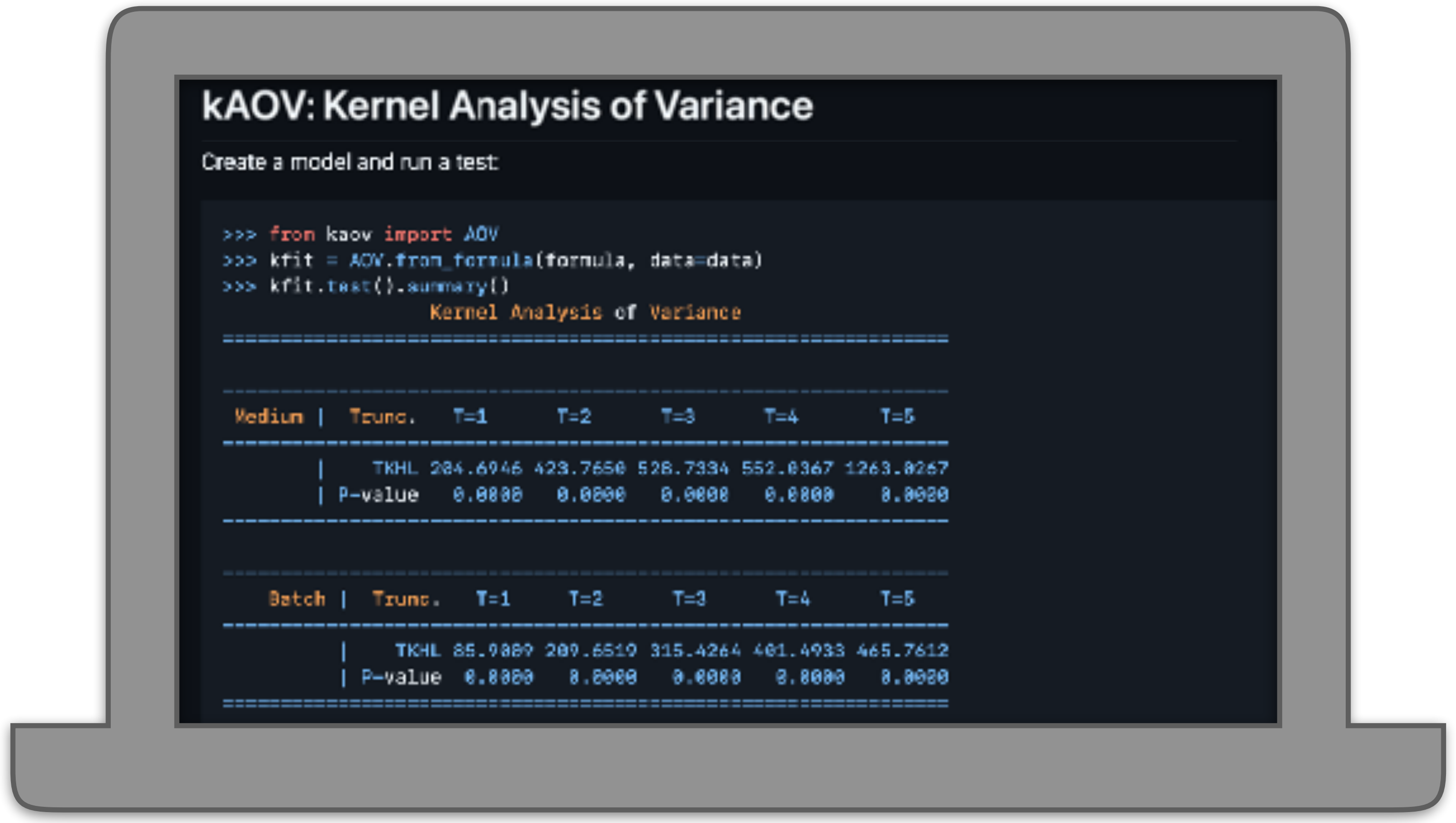
The source code for experimental_data and squair scripts are based on the branch `publication_genome_biology`. The reversion scripts are based on the branch `publication_genome_biology_reversion`.

See https://github.com/AnthoOzier/ktest_experiment_genome_biology_2024 for more details.

Thank you for attention!

General hypothesis testing

Python library **kAOV**



kAOV: Kernel Analysis of Variance

Create a model and run a test

```
>>> from kaov import AOV
>>> kfit = AOV.from_formula(formula, data=data)
>>> kfit.test().summary()
                    Kernel Analysis of Variance
================================================================


----------------------------------------------------------------
 Medium |  Trunc.   T=1      T=2      T=3      T=4       T=5
----------------------------------------------------------------
        |    TKHL 204.6946 423.7650 528.7334 552.0367 1263.0267
        | P-value   0.0000   0.0000   0.0000   0.0000    0.0000
----------------------------------------------------------------


----------------------------------------------------------------
   Batch |  Trunc.   T=1     T=2      T=3      T=4      T=5
----------------------------------------------------------------
         |    TKHL 85.9889 209.6519 315.4264 401.4933 465.7612
         | P-value  0.0000   0.0000   0.0000   0.0000   0.0000
================================================================
```

+

Preprint

[Submitted on 22 May 2024]

Extending Kernel Testing To General Designs

Anthony Ozier-Lafontaine, Franck Picard, Bertrand Michel