24èmes Rencontres HelioSPIR - 13-15/06/2023 - Montpellier

Bénéfices et limites de la méthode de visualisation t-SNE pour la spectroscopie proche infrarouge

François STEVENS*, Vincent BAETEN & Juan Antonio FERNÁNDEZ PIERNA

Centre wallon de Recherches agronomiques
Unité Qualité et authentification des produits
5030 Gembloux – Belgique

Contact: f.stevens@cra.wallonie.be

Origines de la méthode t-SNE

- 2003: Développement de la méthode “Stochastic Neighbor Embedding” (SNE)
  Hinton, G., & Roweis, S. (2003). Stochastic neighbor embedding. *Advances in Neural Information Processing Systems 15*.
- 2008: Mise au point de la variante “t-SNE”
  van der Maaten, L., & Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research 9*, *9*, 2579–2605.

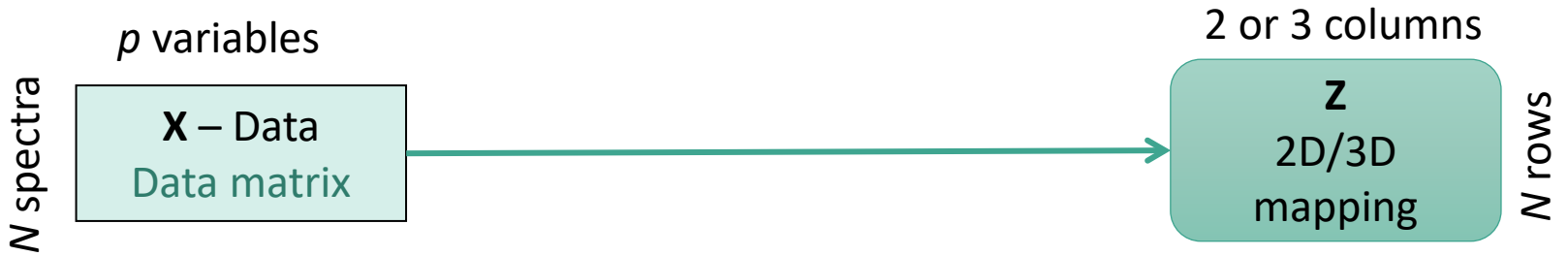
Publications, implémentations, exemples, conseils, …
→ Page “t-SNE” de Laurens van der Maaten: https://lvdmaaten.github.io/tsne/

t-SNE: t-distributed Stochastic Neighbor Embedding

Méthode

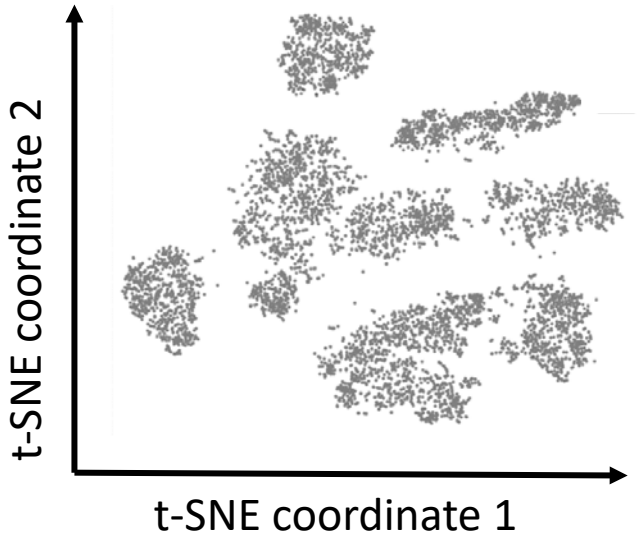
- de réduction de dimension
- non supervisée (utilise seulement la matrice spectrale, pas les valeurs de références)
- non paramétrique
- non linéaire
- utilisée principalement pour la visualization et l'exploration de données

Fonctionnement basique

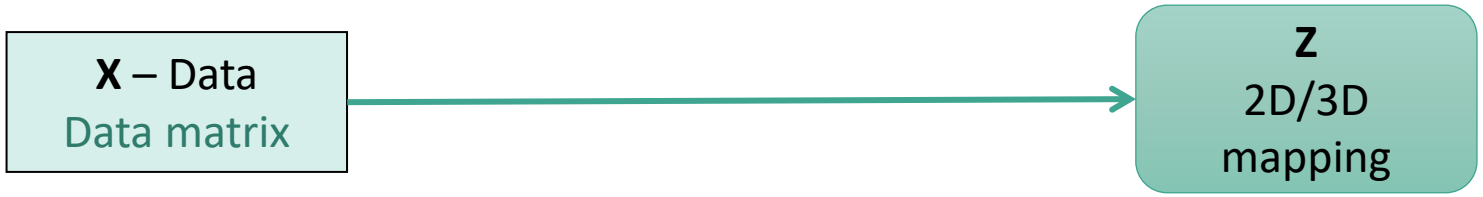


Deux point qui sont *proches* dans l'espace d'origine (spectres *similaires*) on une *grande* probabilité d'être proches dans l'espace t-SNE

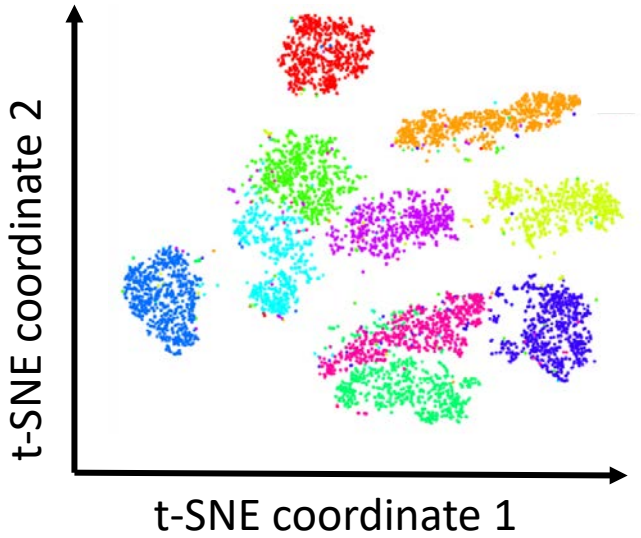
Deux point qui sont *éloignés* dans l'espace d'origine (spectres *différents*) on une *faible* probabilité d'être proches dans l'espace t-SNE



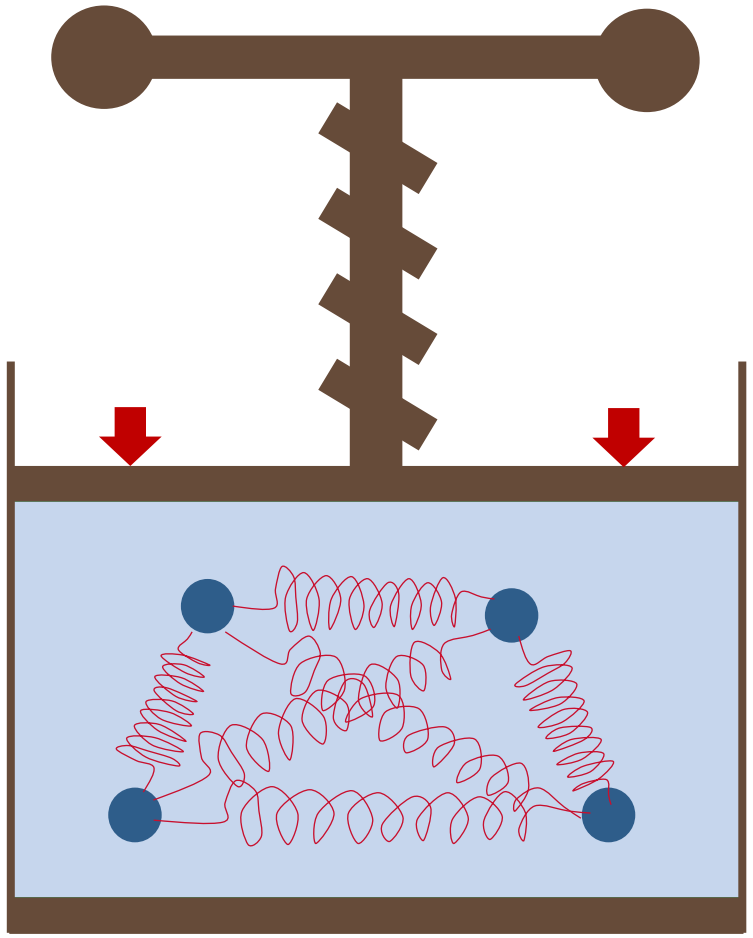
Fonctionnement basique



Si on colore **A POSTERIORI** les points en fonction d'une valeur de référence ou d'une variable catégorique, on peut vérifier si cette variable a un effet sur la matrice X (les spectres)



Fonctionnement intuitif



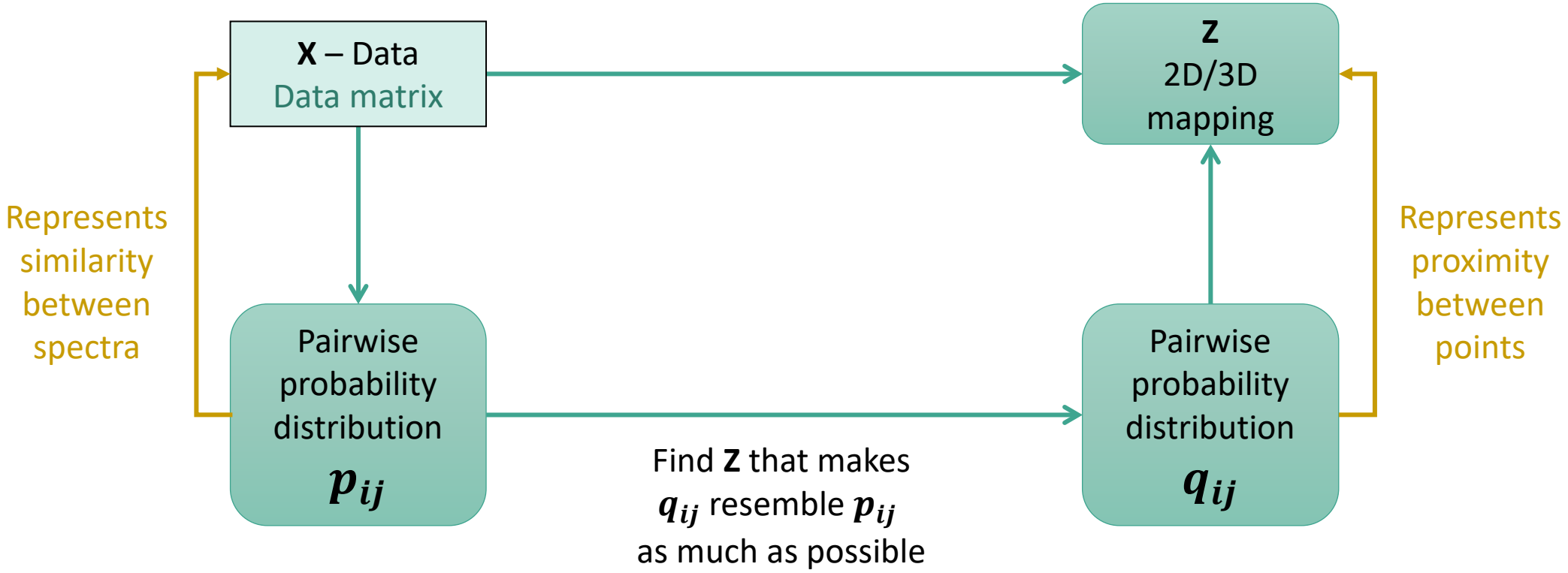
Analogie de la presse et des ressorts

Imaginons

- N boules
- dans un espace multidimensionel
- se déplaçant librement (eau)
- mais subissant des forces de répulsions ou d'attraction d'intensité variable (ressorts)
- tout en étant progressivement comprimés au moyen d'une presse

→ il en résultera une disposition des boules en 2 dimensions tenant compte au mieux de leur attirances ou répulsions respectives

Fonctionnement théorique



Fonctionnement théorique

Normal distribution

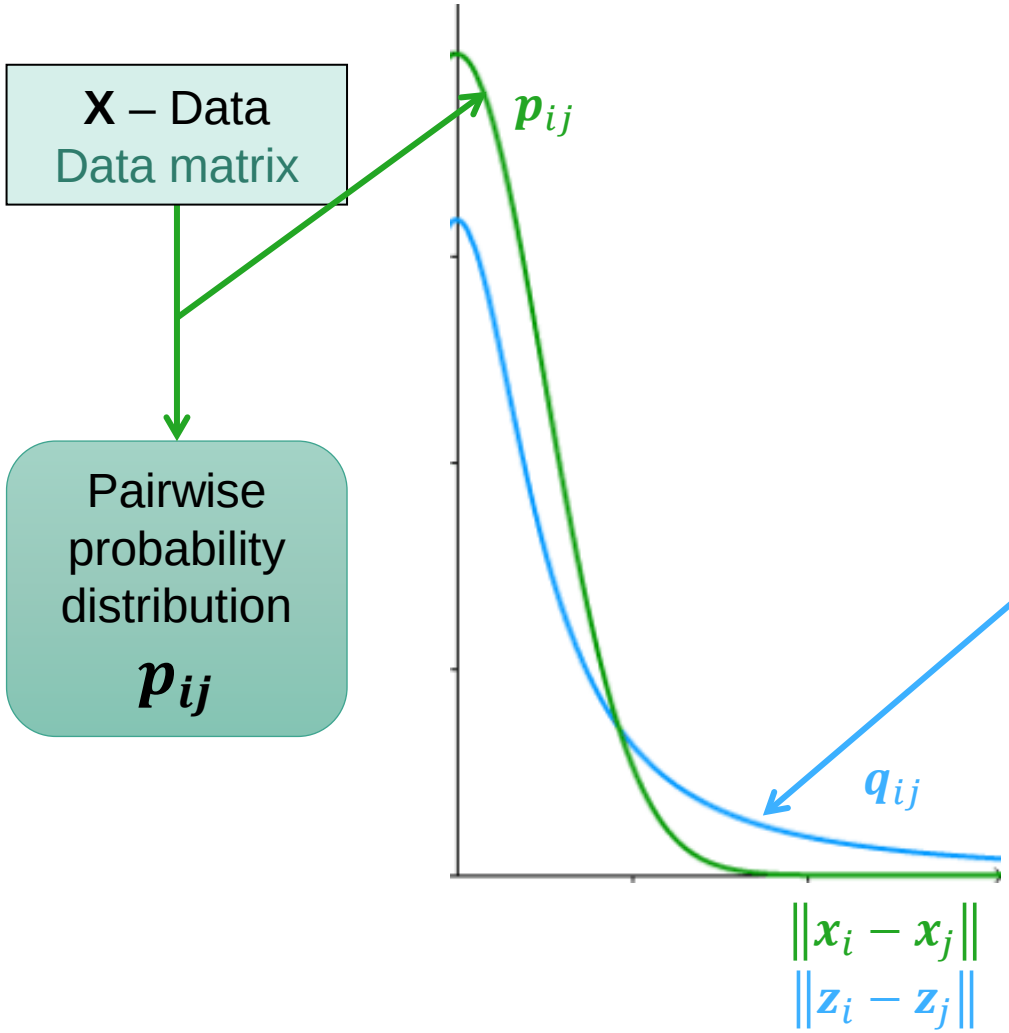
$$p_{j|i} = \frac{e^{-\|x_i - x_j\|^2 / 2\sigma_i^2}}{\sum_{k \neq i} e^{-\|x_i - x_k\|^2 / 2\sigma_k^2}}$$

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N}$$

with σ_i fixed such that

$$\forall i: \; 2^{-\sum_j p_{j|i} \log_2 p_{j|i}} = \mathrm{P}$$

with P the *perplexity*

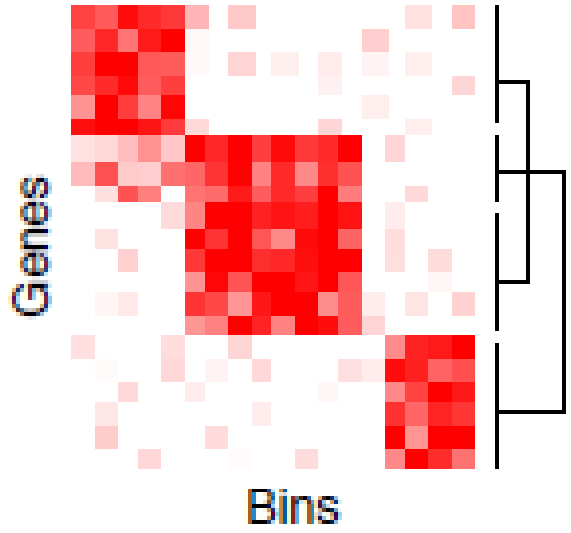


Student t distribution

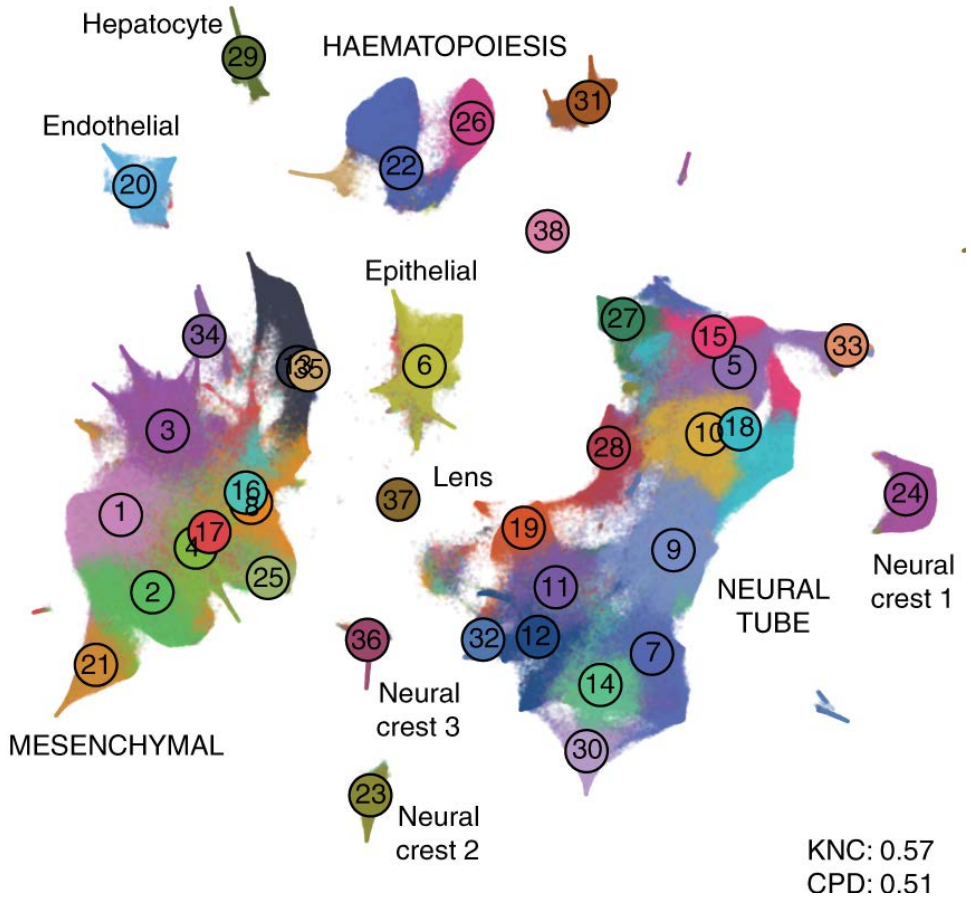
$$q_{ij} = \frac{\left(1 + \|z_i - z_j\|^2\right)^{-1}}{\sum_{k \neq l} (1 + \|z_k - z_l\|^2)^{-1}}$$

Applications actuelles: biologie et « omics »

Standard pour l'exploration visuelle de données de séquençage ARN sur cellules uniques



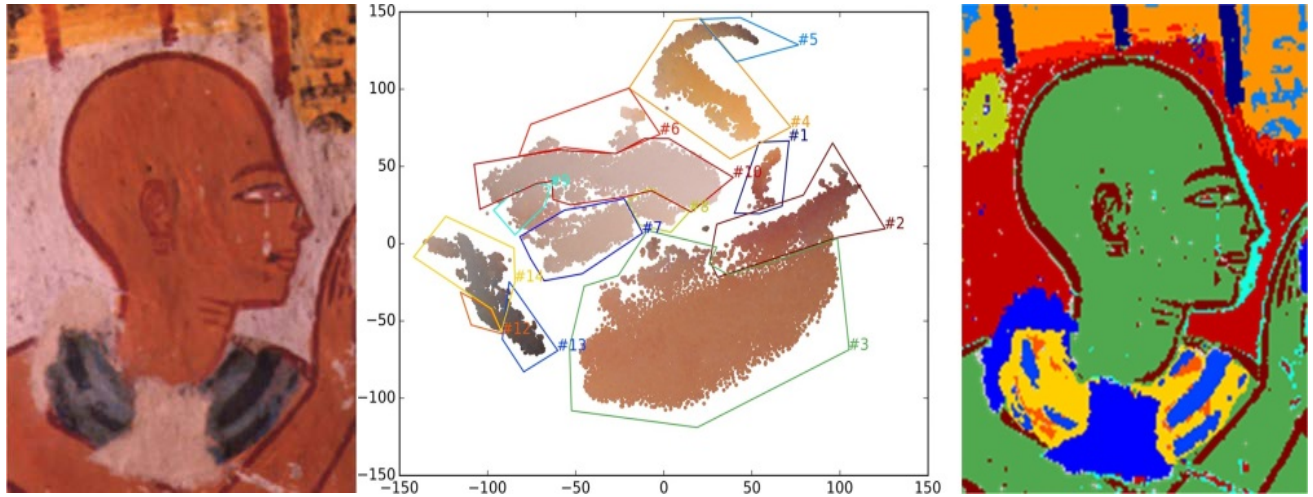
Linderman, G. C. et al. (2019) 'Fast interpolation-based t-SNE for improved visualization of single-cell RNA-seq data', Nature Methods, 16(3), pp. 243–245. doi: 10.1038/s41592-018-0308-4.



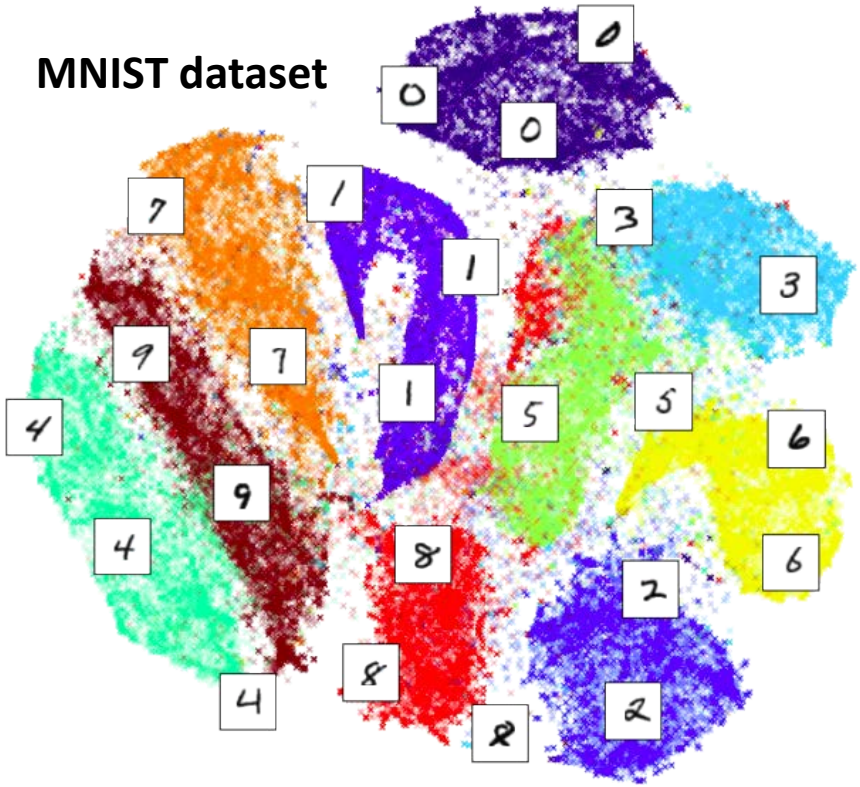
Kobak, D. and Berens, P. (2019) 'The art of using t-SNE for single-cell transcriptomics', Nature Communications, 10(1). doi: 10.1038/s41467-019-13056-x..

Applications actuelles : autres domaines

Utilisé dans plein de domaines utilisant de grandes bases de données: neurologie, analyse biomédicale, archéologie, sciences sociales, psychologie, sécurité informatique, traitement du langage naturel, reconnaissance optique de caractères, …



Alfeld, M. *et al.* (2018) 'Joint data treatment for Vis–NIR reflectance imaging spectroscopy and XRF imaging acquired in the Theban Necropolis in Egypt by data fusion and t-SNE', *Comptes Rendus Physique*, 19(7), pp. 625–635. doi: 10.1016/j.crhy.2018.08.004.



https://nlml.github.io/in-raw-numpy/in-raw-numpy-t-sne/

t-SNE en spectroscopie vibrationelle

Actuellement, t-SNE semble être connue et utilisée par un nombre grandissant de chercheurs en tant que méthode d'exploration complémentaire, mais reste mentionnée dans un nombre limité de publication

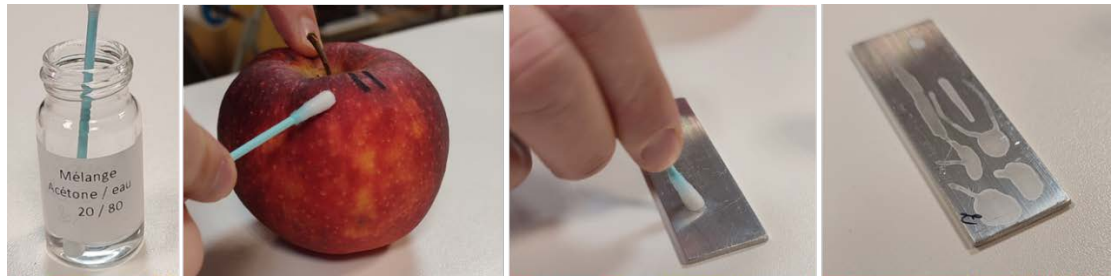
Nous suggérons ici quelques utilisations potentiellement intéressantes de t-SNE

1. contrôle rapide des données
2. aide au choix du prétraitement spectral
3. exploration approfondie en complément de la PCA

Quelques références

- Alfeld, M. *et al.* (2018) 'Joint data treatment for Vis–NIR reflectance imaging spectroscopy and XRF imaging acquired in the Theban Necropolis in Egypt by data fusion and t-SNE', *Comptes Rendus Physique*, 19(7), pp. 625–635. doi: 10.1016/j.crhy.2018.08.004.
- Luo, N. *et al.* (2021) 'Visualization of vibrational spectroscopy for agro-food samples using t-Distributed Stochastic Neighbor Embedding', *Food Control*, 126, p. 107812. doi: 10.1016/j.foodcont.2020.107812.
- Mwanga, E. P. *et al.* (2023) 'Using transfer learning and dimensionality reduction techniques to improve generalisability of machine-learning predictions of mosquito ages from mid-infrared spectra', *BMC Bioinformatics*, 24(1), pp. 1–15. doi: 10.1186/s12859-022-05128-5.
- Xie, B. *et al.* (2022) 'Detection of lipid efflux from foam cell models using a label-free infrared method', *Analyst*, 407, pp. 5372–5385. doi: 10.1039/d2an01041k.

Jeu de données pesticides



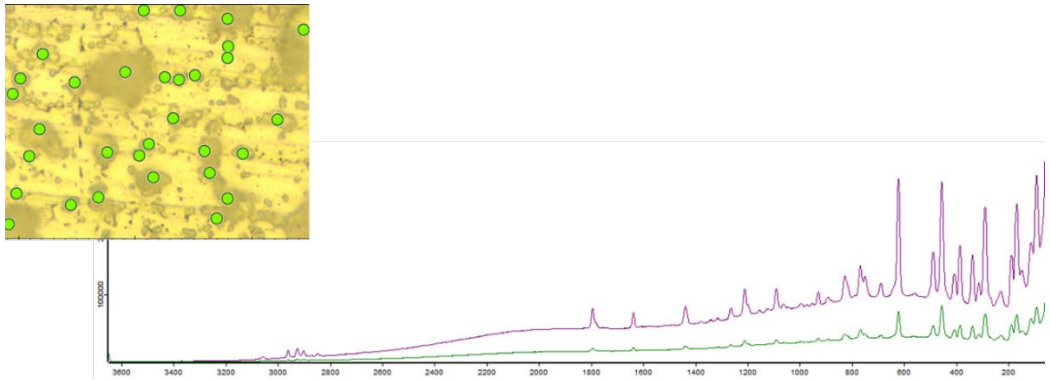
Pesticide spraying

Pesticide recovery

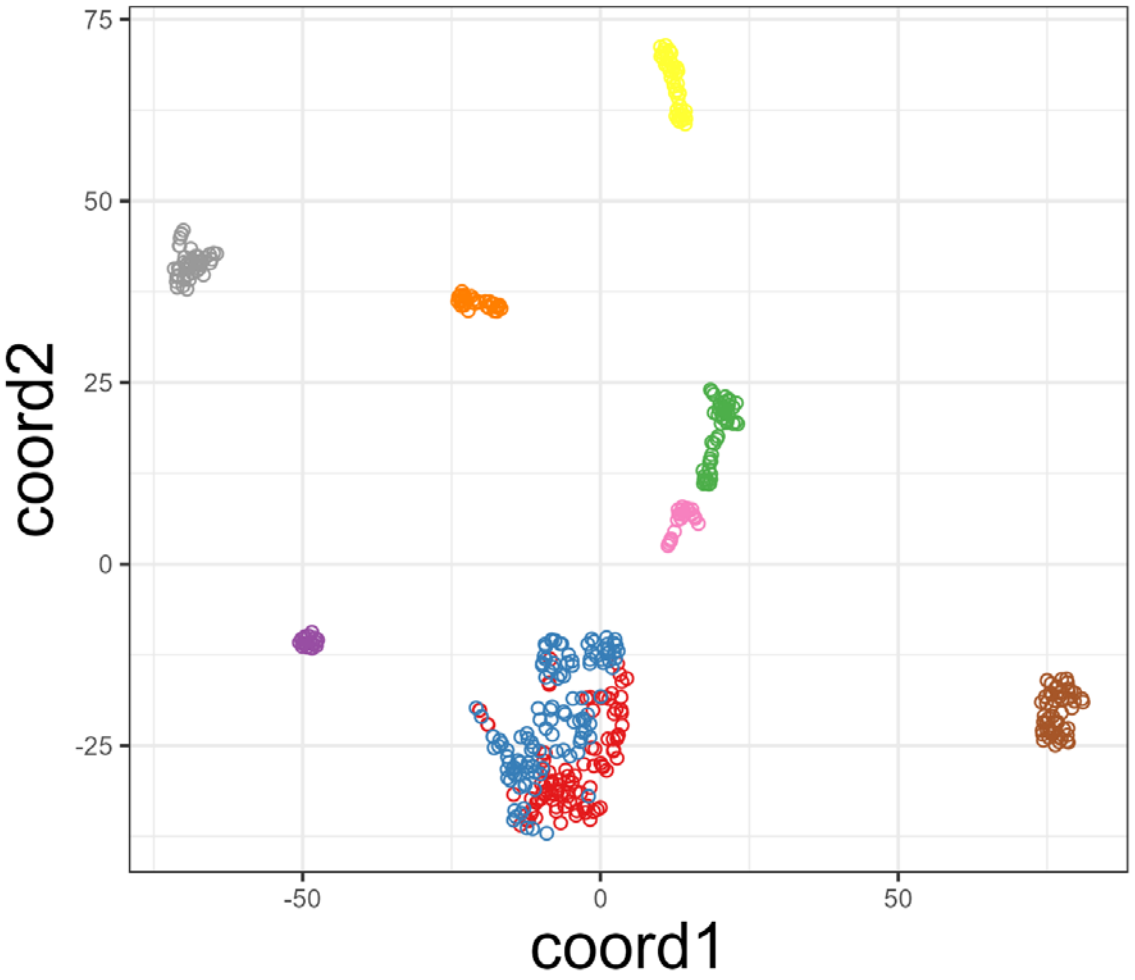
Raman microscopy

- 467 spectra
- 9 pesticides (or mixtures)

Source: Centre wallon de Recherches agronomiques



t-SNE pour le contrôle rapide des données



→ Une discrimination précise semble réalisable pour la plupart des pesticides

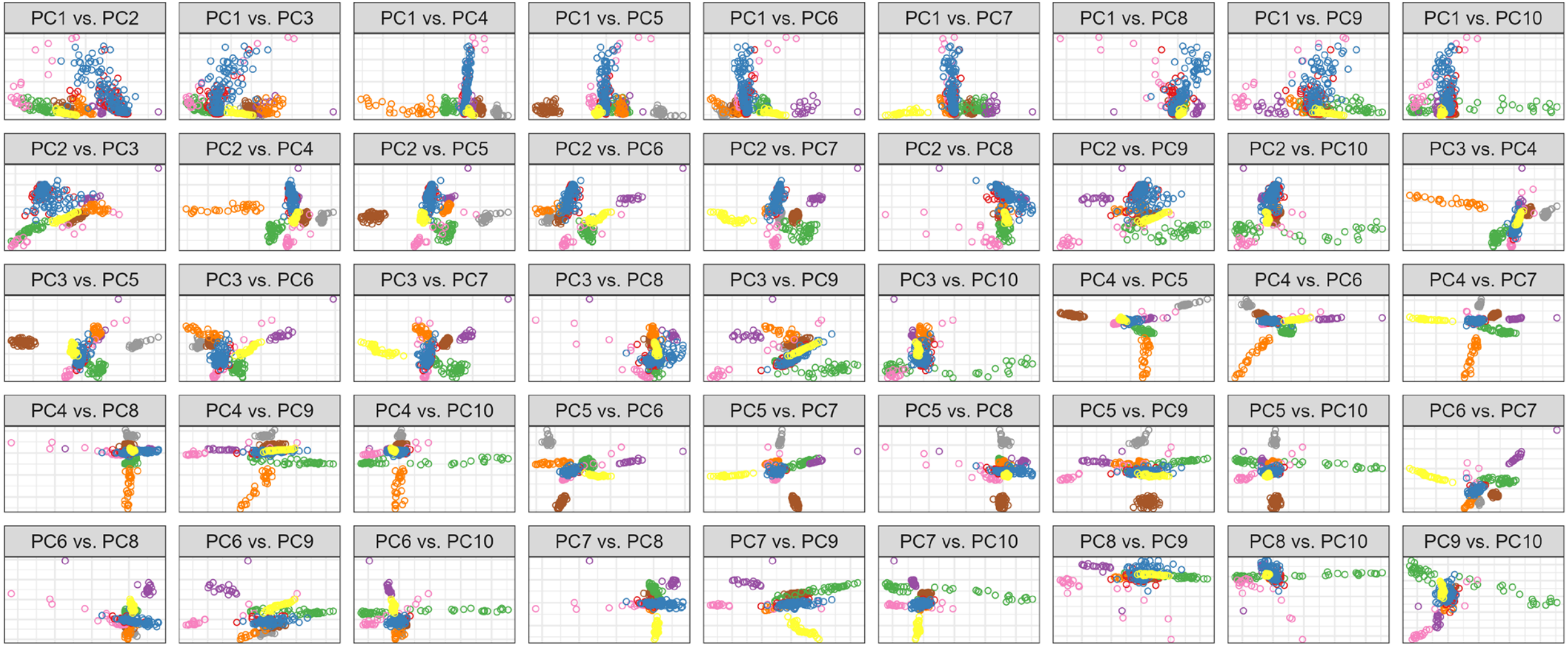
pesticide

- captan
- captan + trifloxystrobin
- difenoconazole
- folpet
- mancozeb
- pyrimethanil
- spirotetramat
- tebuconazole
- thiabendazole

Pre-processing: 1st derivative then SNV

Perplexity: $\sqrt{N}$ = 21.61

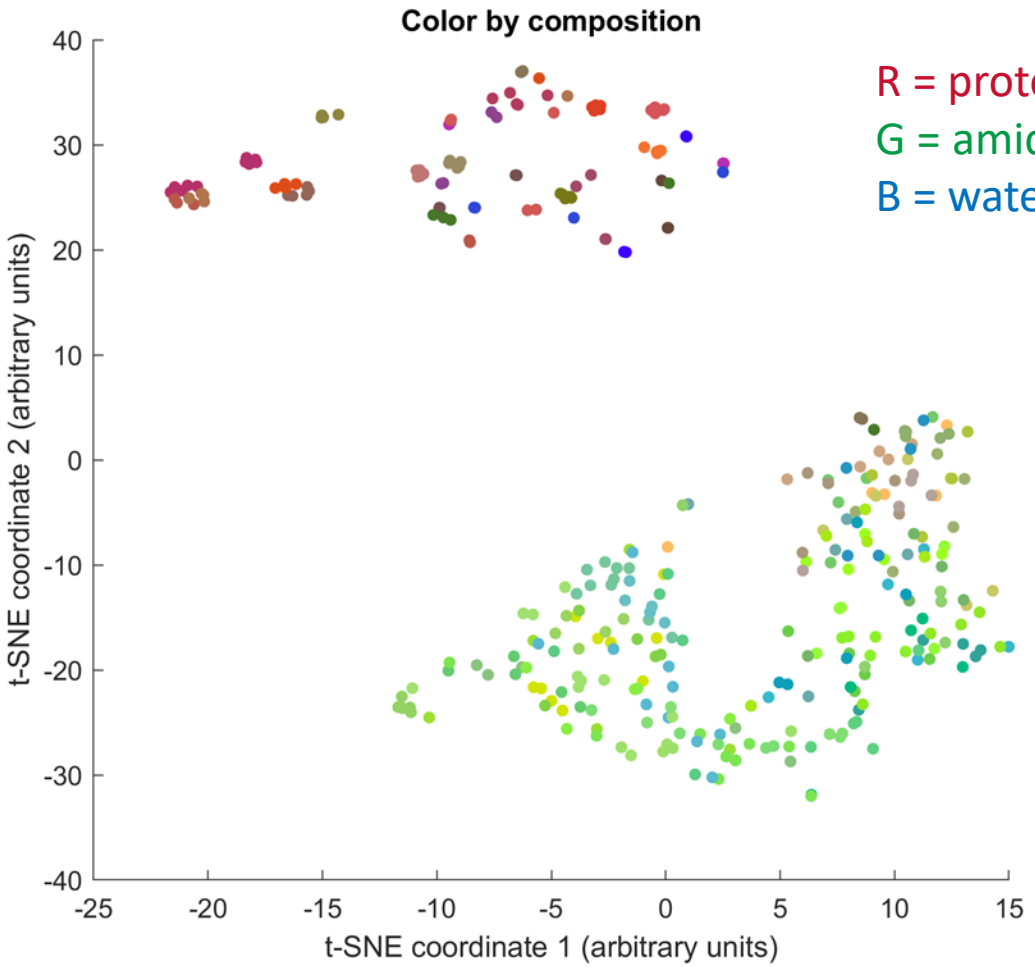
Et avec PCA …

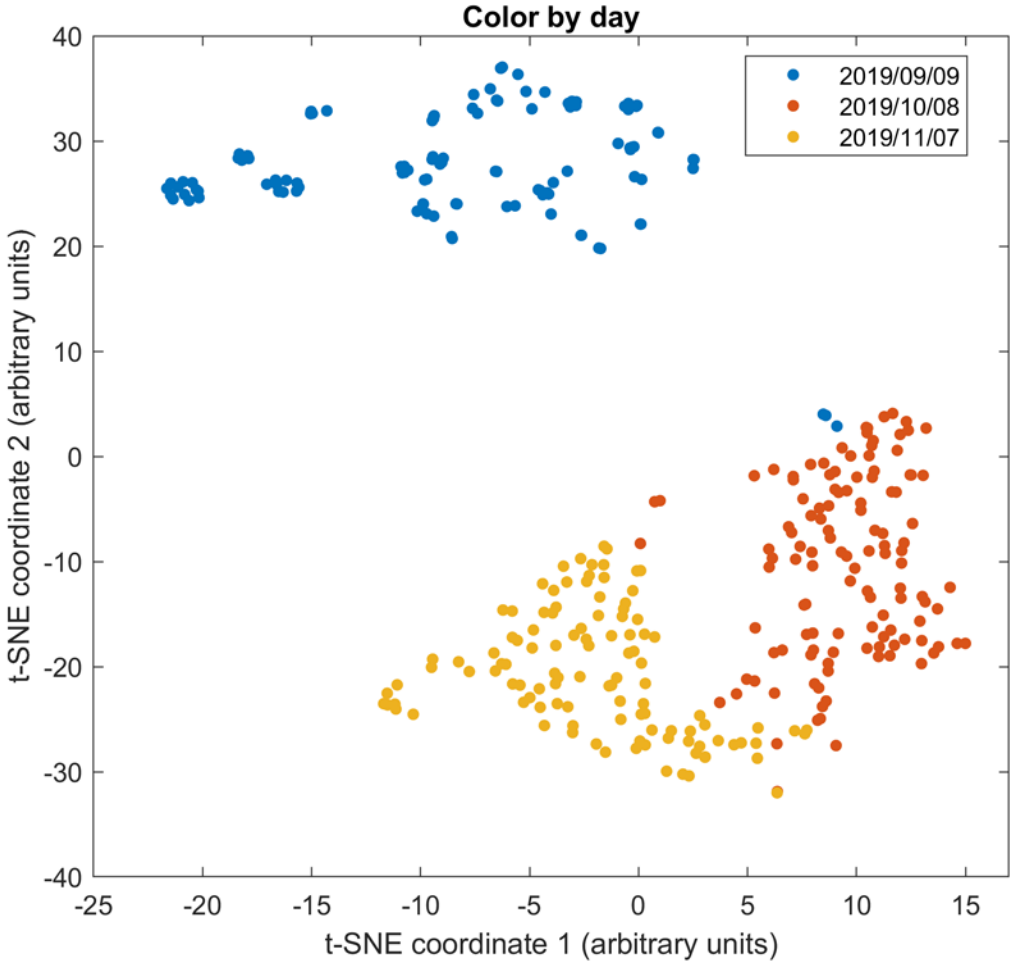

PC1 vs. PC2
PC1 vs. PC3
PC1 vs. PC4
PC1 vs. PC5
PC1 vs. PC6
PC1 vs. PC7
PC1 vs. PC8
PC1 vs. PC9
PC1 vs. PC10
PC2 vs. PC3
PC2 vs. PC4
PC2 vs. PC5
PC2 vs. PC6
PC2 vs. PC7
PC2 vs. PC8
PC2 vs. PC9
PC2 vs. PC10
PC3 vs. PC4
PC3 vs. PC5
PC3 vs. PC6
PC3 vs. PC7
PC3 vs. PC8
PC3 vs. PC9
PC3 vs. PC10
PC4 vs. PC5
PC4 vs. PC6
PC4 vs. PC7
PC4 vs. PC8
PC4 vs. PC9
PC4 vs. PC10
PC5 vs. PC6
PC5 vs. PC7
PC5 vs. PC8
PC5 vs. PC9
PC5 vs. PC10
PC6 vs. PC7
PC6 vs. PC8
PC6 vs. PC9
PC6 vs. PC10
PC7 vs. PC8
PC7 vs. PC9
PC7 vs. PC10
PC8 vs. PC9
PC8 vs. PC10
PC9 vs. PC10

PCA ou t-SNE pour l'exploration de données ?

	PCA	t-SNE
Tous les scores en un seul graphique	✗	✓
Informations supplémentaires liées à la nature des spectres	✓ (Loadings)	✗
Possibilité de projeter un nouveau point	✓	✗
Utiliser les scores pour détecter les outliers ou créer des clusters	✓	⚠

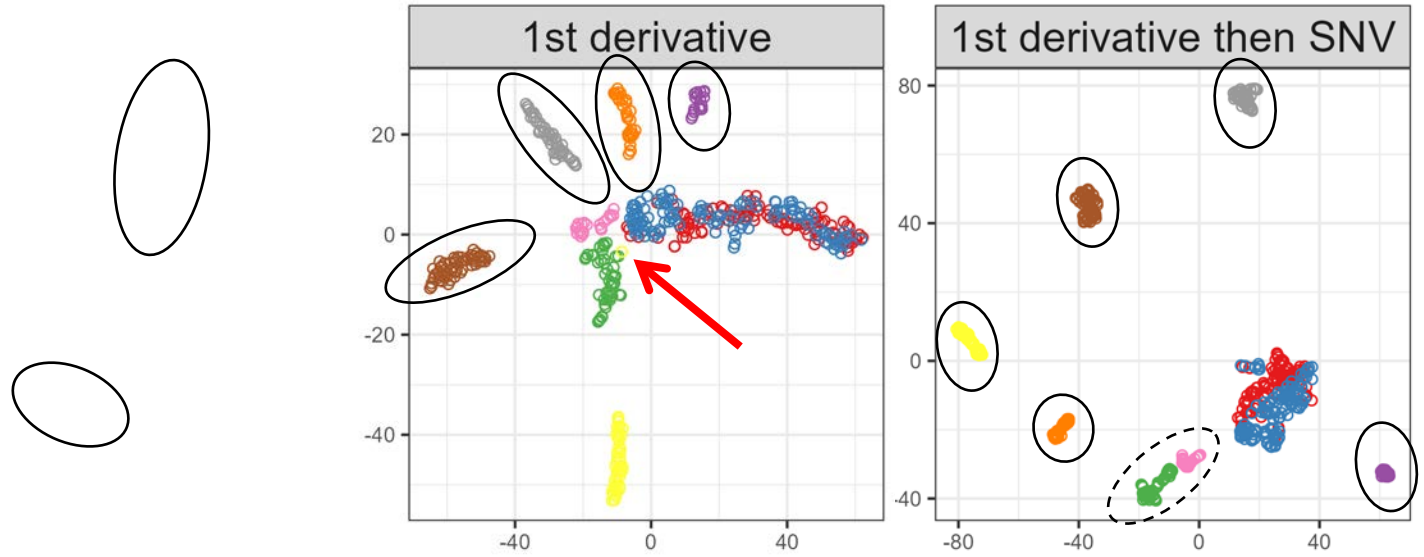
Observation de clusters, effet jour, effet batch, …





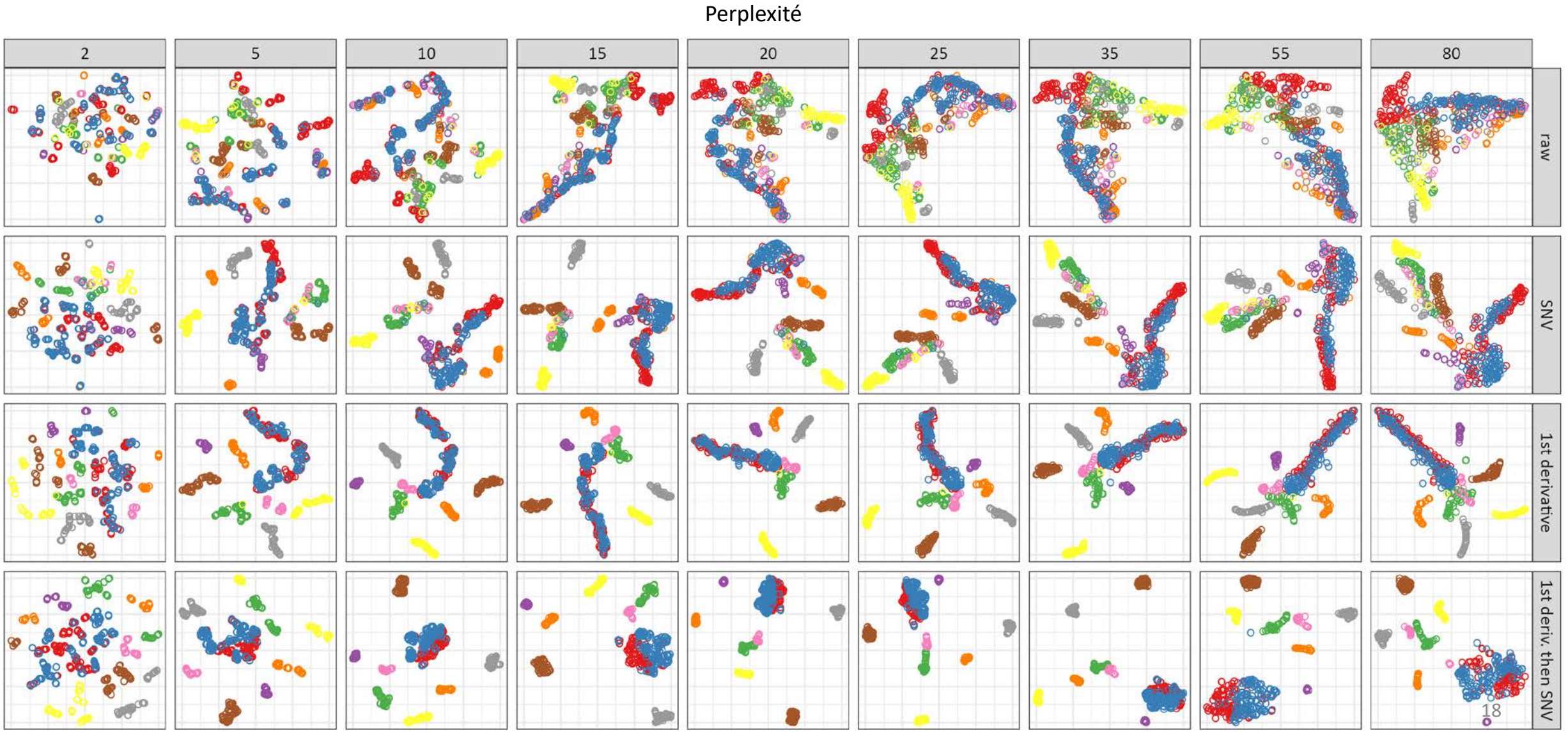
Données fourrage de maïs mesurées avec un XDS (FOSS, DK)
Source : Centre wallon de Recherches agronomiques

Est-ce que le prétraitement a de l'importance ?

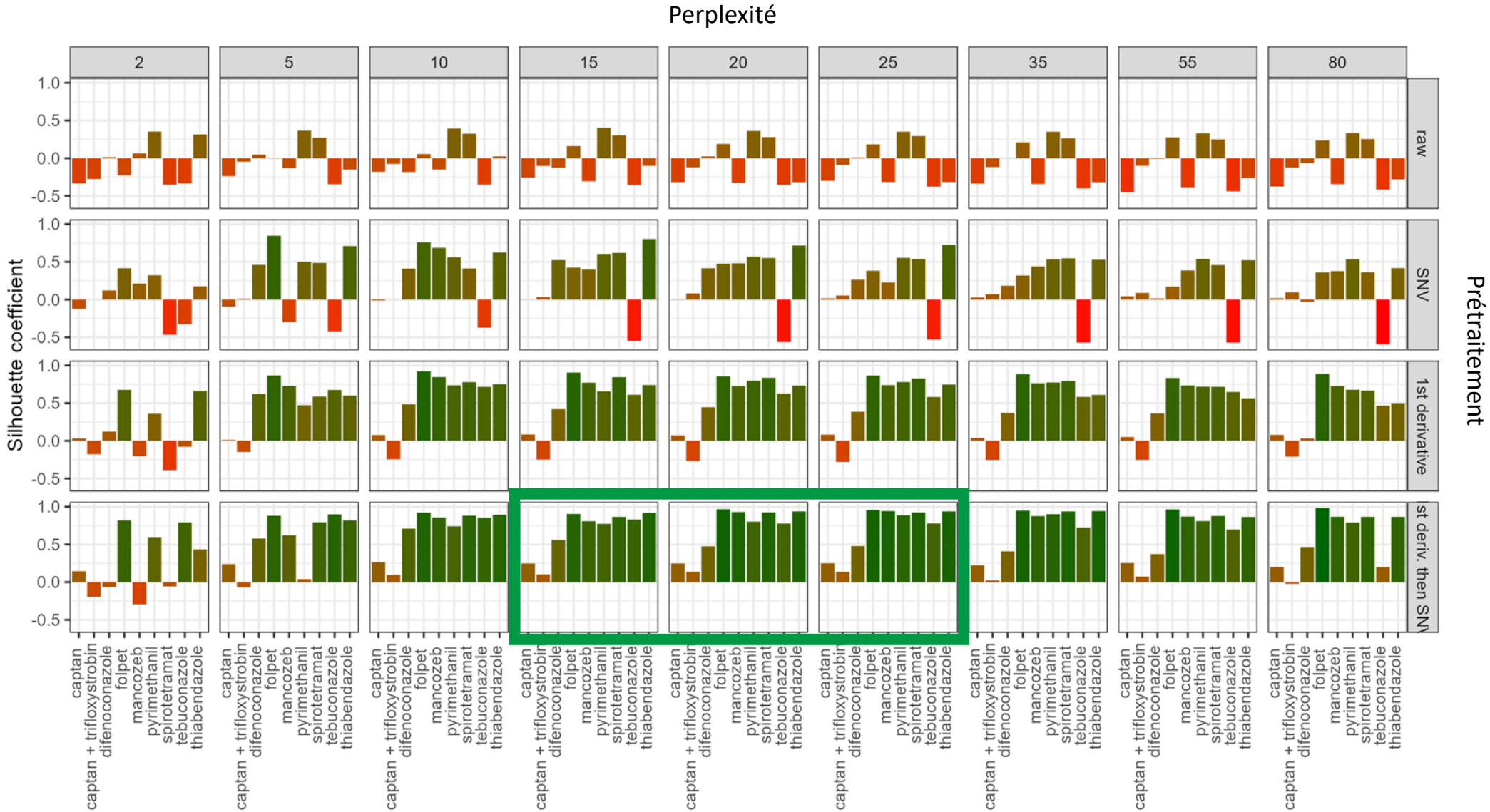


pesticide
captan
captan + trifloxystrobin
difenoconazole
folpet
mancozeb
pyrimethanil
spirotetramat
tebuconazole
thiabendazole

t-SNE pour aider au choix du prétraitement spectral



t-SNE pour aider au choix du prétraitement spectral



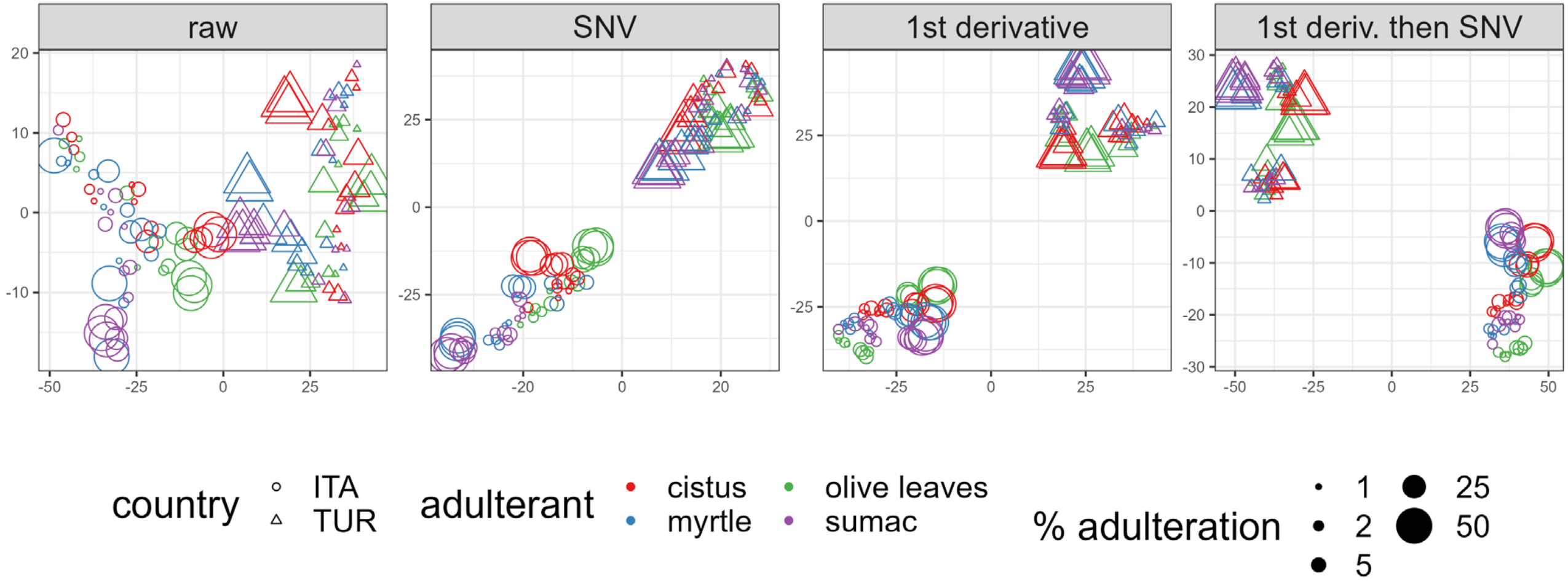
Jeu de données adulteration origan

- **4 adultérants**
  Olive, myrte, sumac, ciste
- **2 pays**
  Italie et Turquie
- **5 niveaux d'adulteration**
  1, 2, 5, 25 and 50%
- **3 repetitions**

→ **120 NIR spectra**
FOSS NIRS™ DS2500
400 to 2500 nm

Van De Steene, J. *et al.* (2022) 'Authenticity analysis of oregano: development, validation and fitness for use of several food fingerprinting techniques', *Food Research International*, 162, p. 111962.

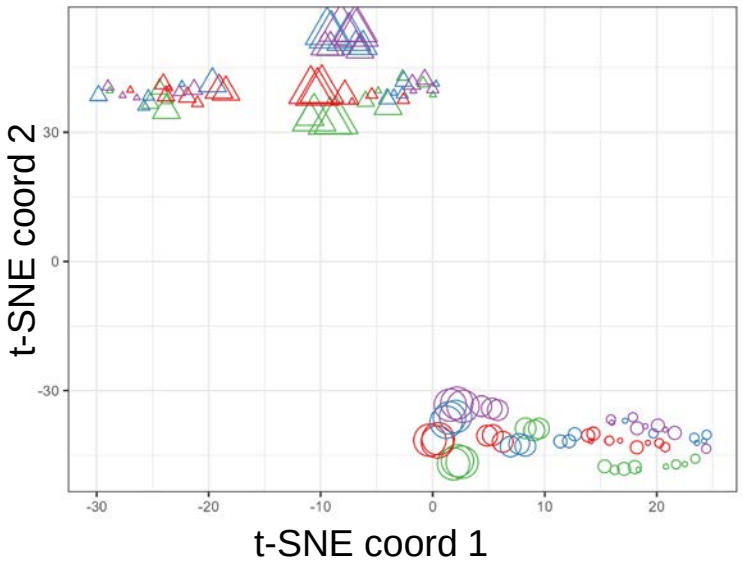
Choix du prétraitement



Combiner t-SNE et PCA ?

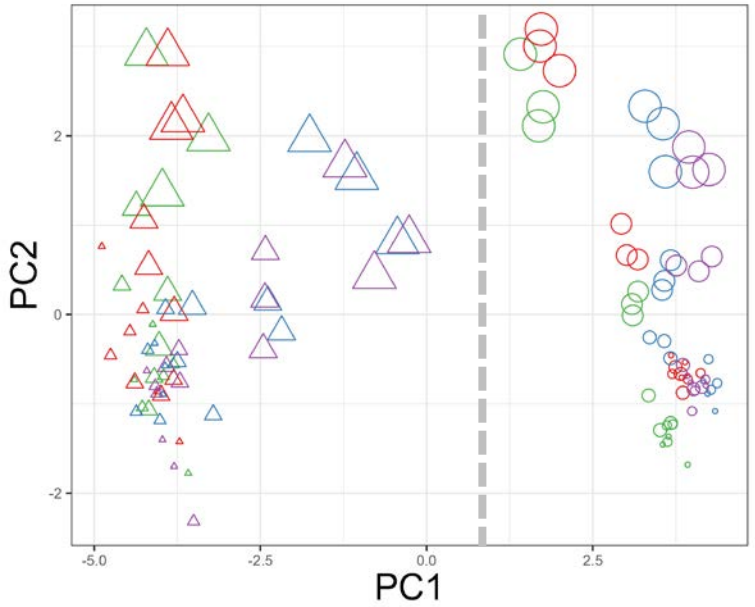
t-SNE on the 100 first PCs

Country dominates



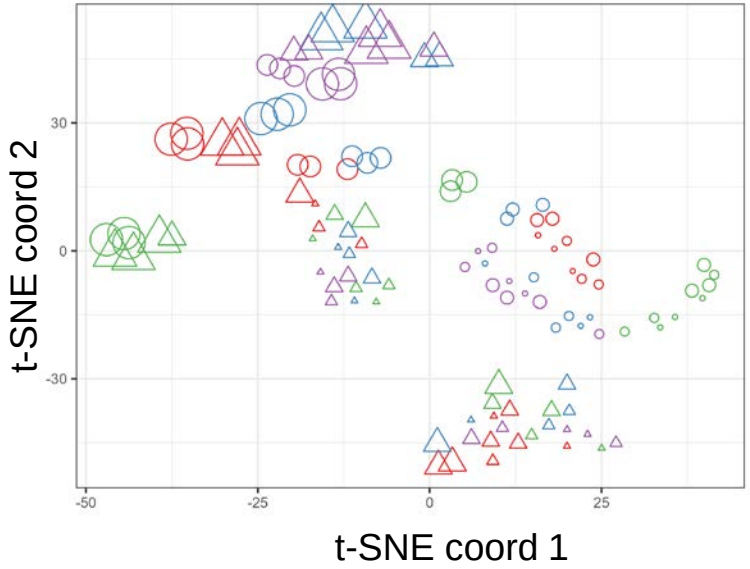
Scores of PC1 and PC2

PC1 largely explains country



t-SNE on PCs **2**-100

Still a small effect of country



country ○ ITA △ TUR

adulterant • cistus • myrtle • olive leaves • sumac

% adulteration • 1 • 2 • 5 ● 25 ● 50

Conclusion

Quel est l’intérêt de t-SNE en spectroscopie vibrationnelle ?

t-SNE permet de :

- Comprendre la structure générale des données et la hiérarchie des facteurs qui influencent les spectres
- Détecter des anomalies (erreur dans le jeu de données, etc) ou des tendances dominantes (effet batch)
- Donner une première estimation de l’efficacité de différents prétraitements
- Anticiper les possibilités en terme de modélisation predictive

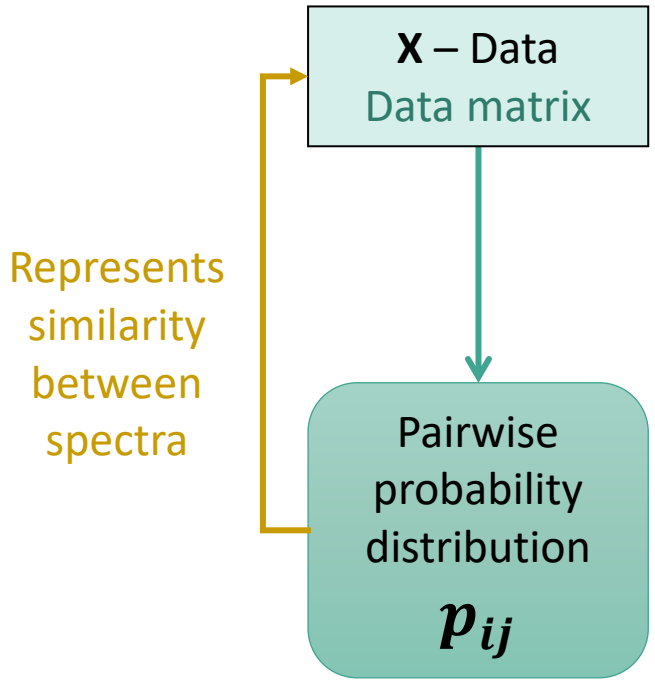
Merci pour votre attention !

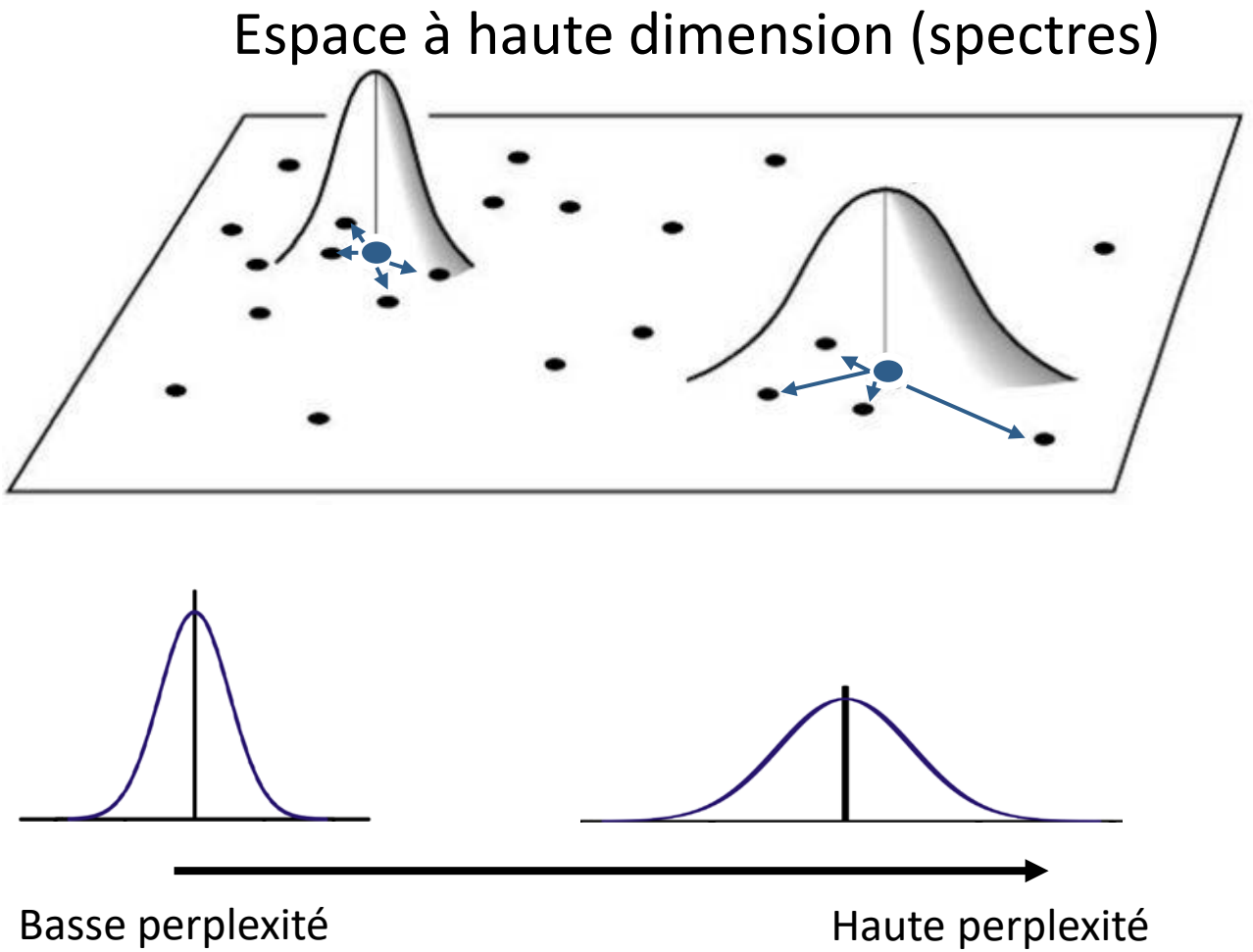
François STEVENS, Vincent BAETEN & Juan Antonio FERNÁNDEZ PIERNA

Centre wallon de Recherches agronomiques
Unité Qualité et authentification des produits
5030 Gembloux – Belgique

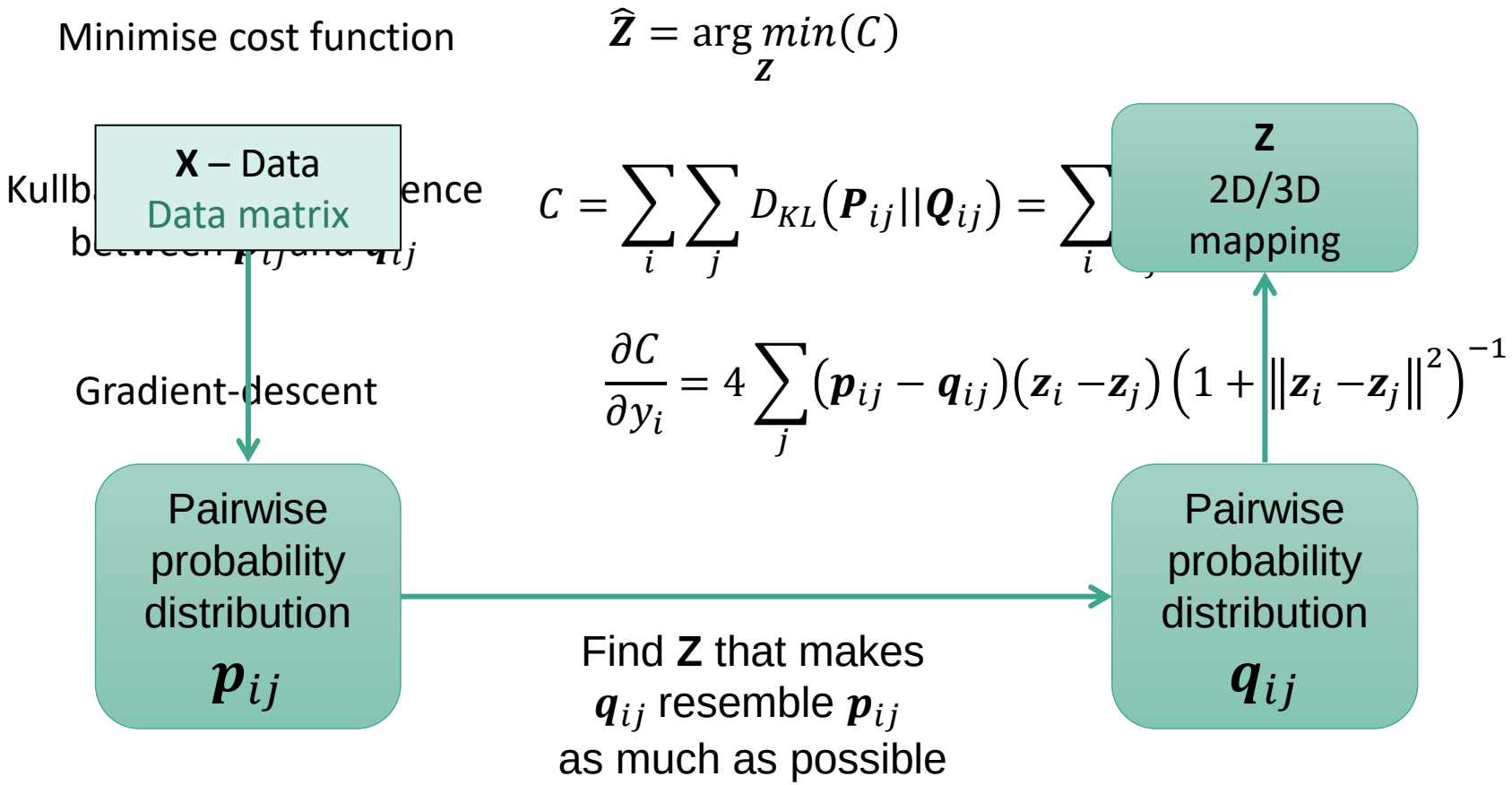
Vibrational Spectroscopy and Chemometric course
02-06 October 2023
Gembloux-Belgium

La perplexité, un paramètre pour ajuster l'échelle





Fonctionnement théorique



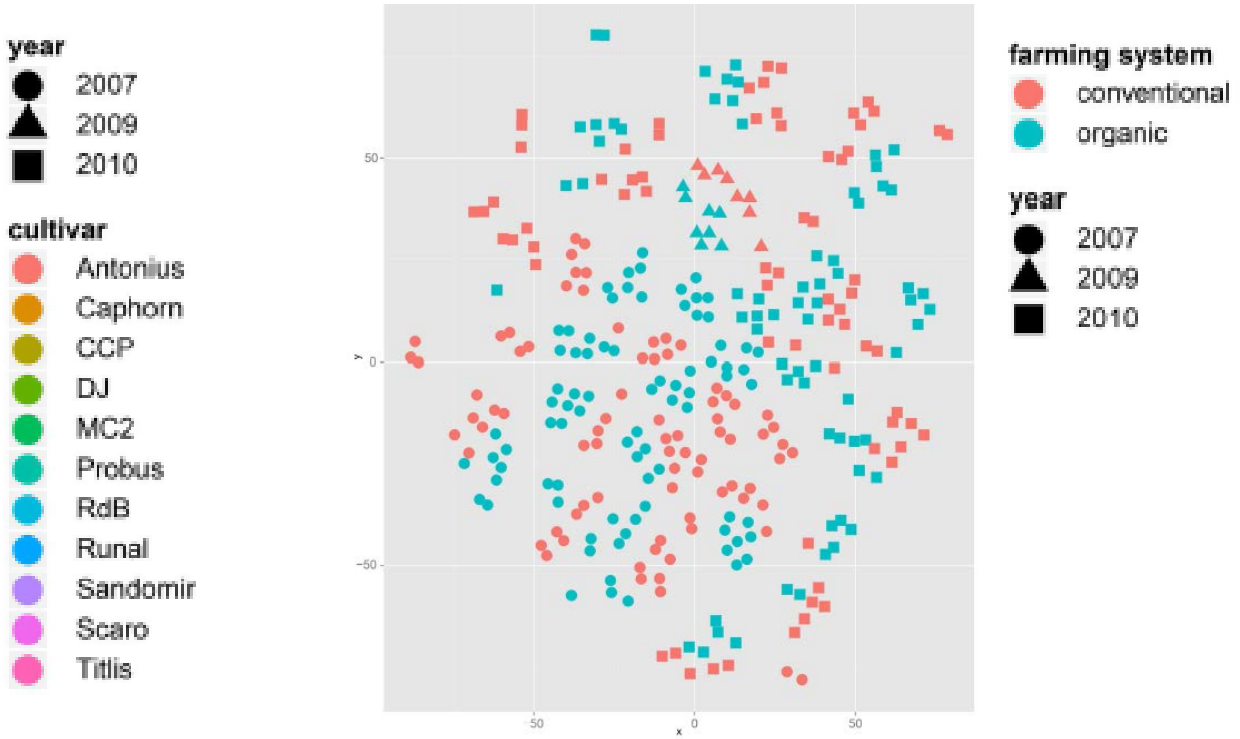
Détails pratiques

Différentes techniques sont utilisées afin d'améliorer la convergence et la recherche d'une solution optimale ou la plus satisfaisante possible

- **early compression** : empêche les points de trop s'éloigner durant les premières iterations afin de mieux explorer l'espace des solutions (ajout d'une pénalité L2 sur les distances dans l'espace t-SNE)
- **early exaggeration** : permet d'augmenter la tendance au clustering durant les premières phases de l'optimisation afin d'augmenter l'espace disponible et donc les solutions accessibles
- **Barnes-Hut** : algorithme basé sur decomposition de l'espace HD en une structure en arbre, permettant d'accélerer le calcul des similarités entre objets en utilisant des approximations
- **Fast interpolation-based t-SNE (FIt-SNE)** : autre implémentation utilisant des approximations et efficace sur les très gros jeux de données

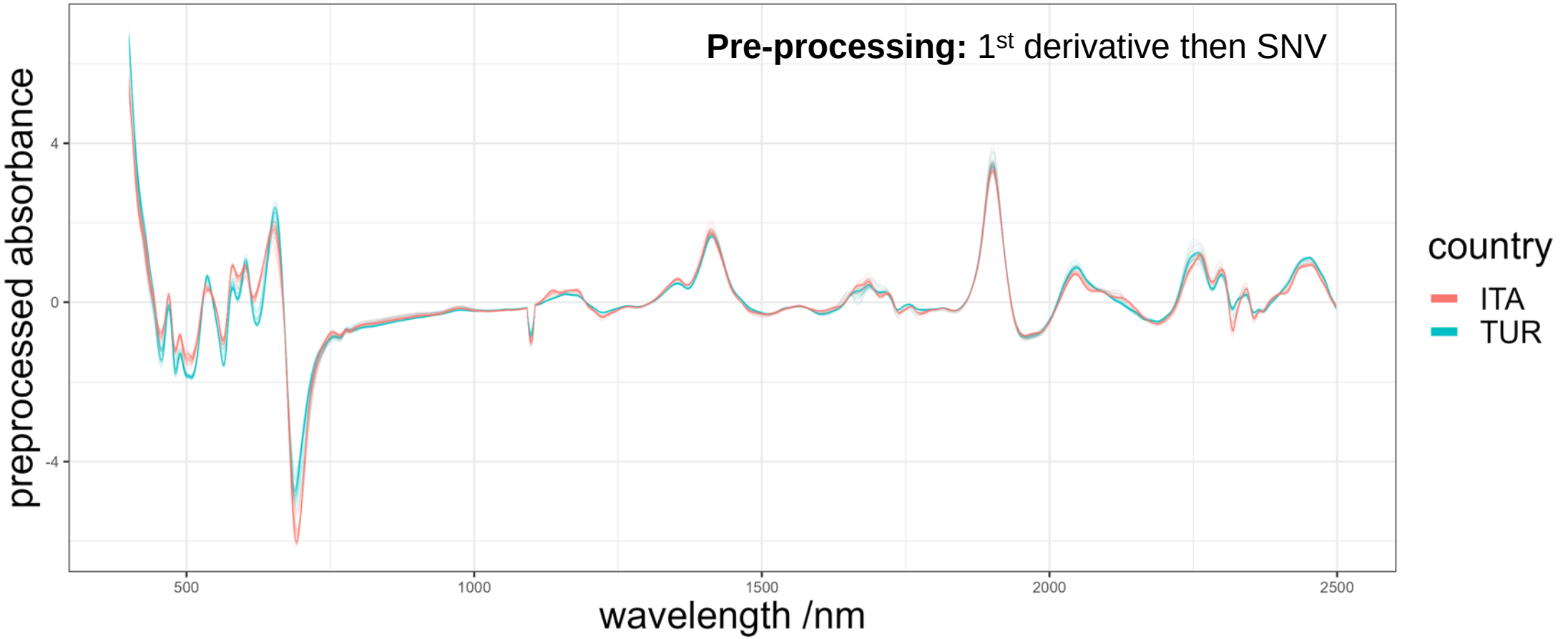
Applications actuelles: biologie et « omics »

Exemple dans le domaine de la métabolomique : exploration de donnée GC-MS de blé avec différents modes de culture (biologique et conventionnel)

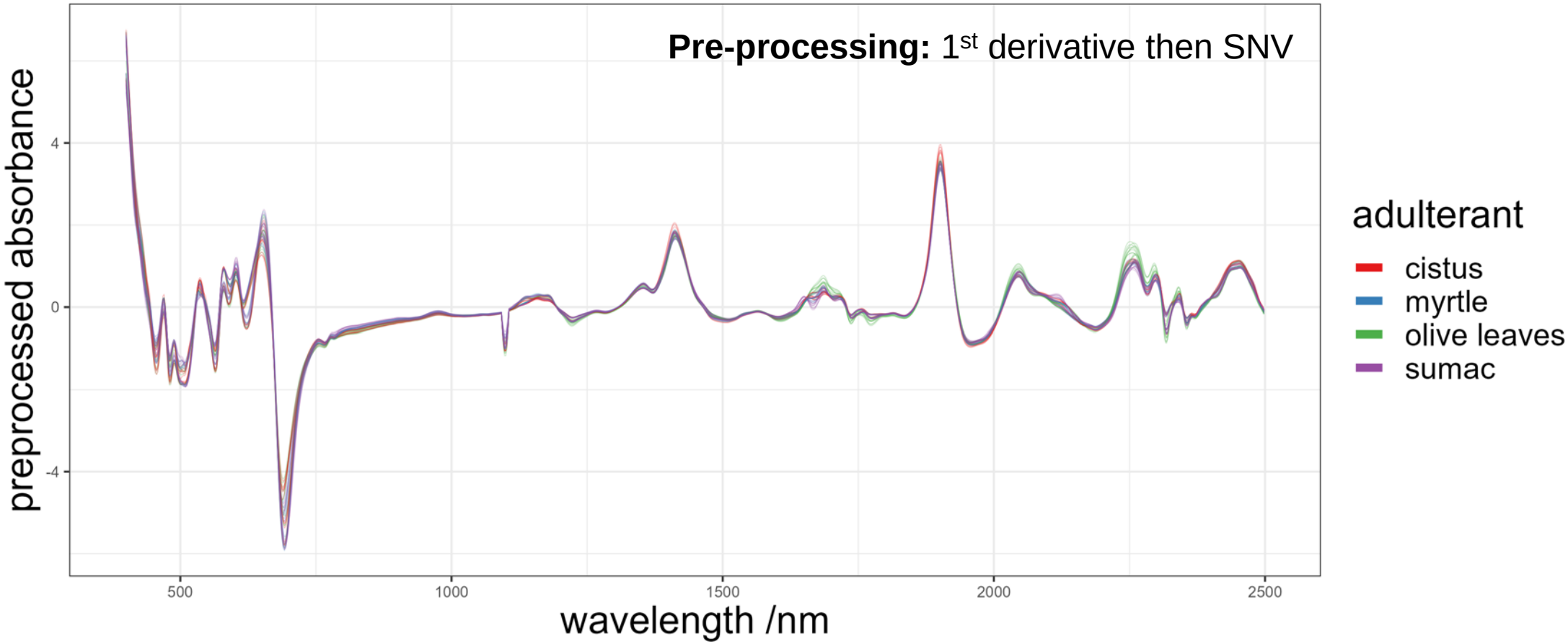


Kessler, N. *et al.* (2015) 'Learning to classify organic and conventional wheat - A machine learning driven approach using the MeltDB 2.0 metabolomics analysis platform', *Frontiers in Bioengineering and Biotechnology*, 3(MAR), pp. 0–10. doi: 10.3389/fbioe.2015.00035.

Tous les spectre, couleur par pays



Tous les spectre, couleur par adultérant



Dérivée 1ère + SNV