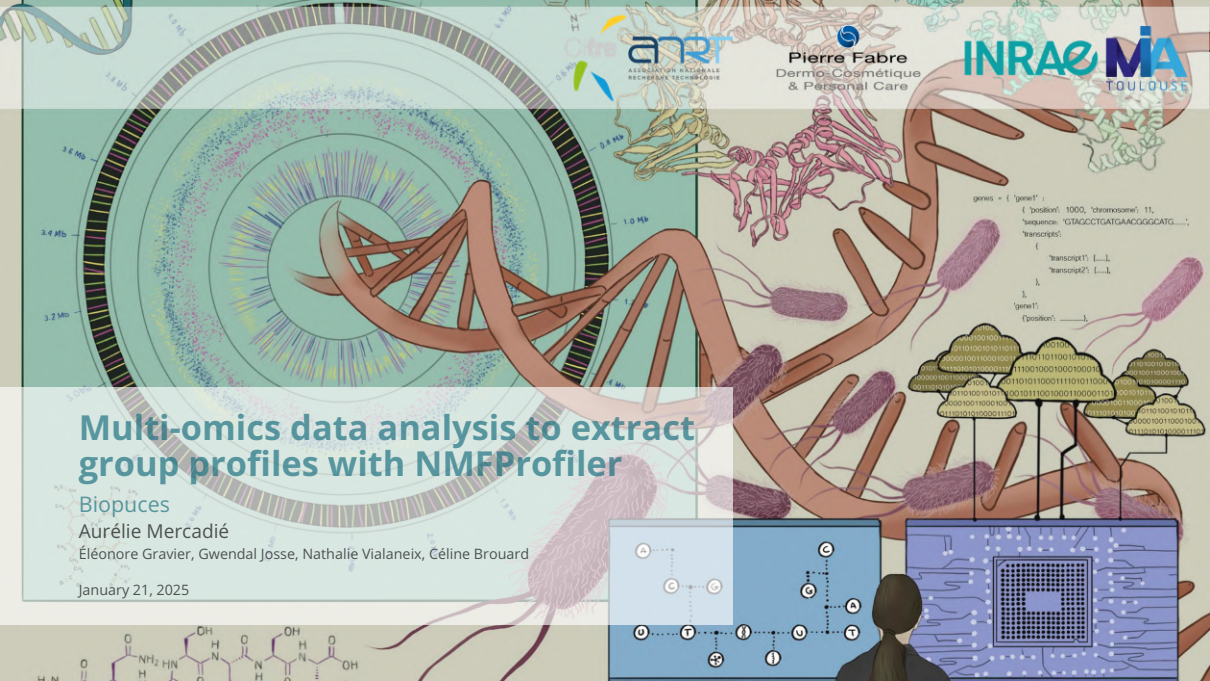


Multi-omics data analysis to extract group profiles with NMFProfiler

Biopuces

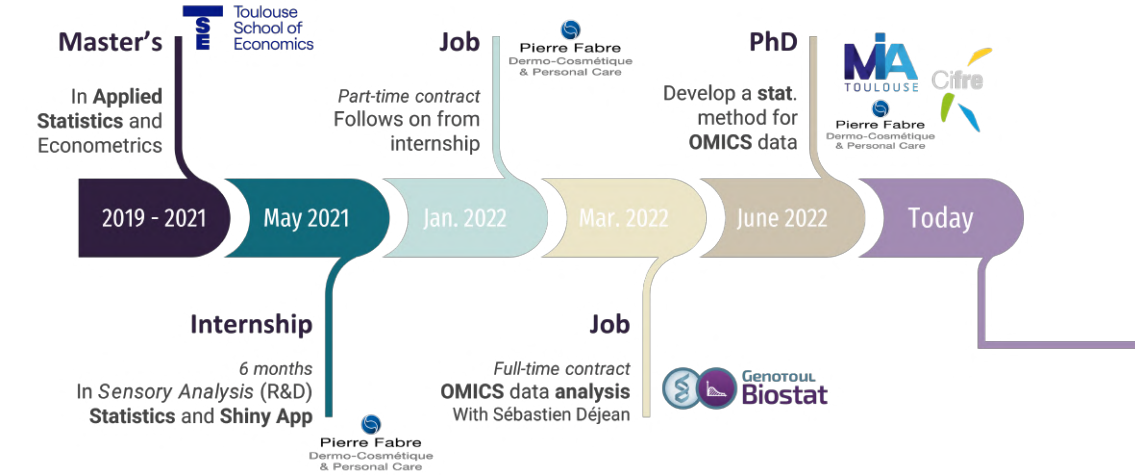
Aurélie Mercadié

Éléonore Gravier, Gwendal Josse, Nathalie Vialaneix, Céline Brouard

January 21, 2025

About me

A brief overview of the past years



Skin research

Pierre-Fabre Laboratories

- **Prevent**, **soothe** and treat **skin disorders** (e.g. acne, alopecia, eczema / atopic dermatitis, seborrheic dermatitis, rosacea, skin cancer…) or **changes** (e.g. skin aging…).
- Develop products *taking care of* healthy as well as damaged skin, providing **homeostasis**.

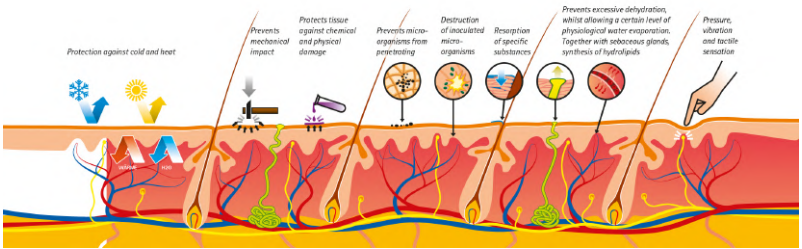
Skin research

Pierre-Fabre Laboratories

- **Prevent**, **soothe** and treat **skin disorders** (e.g. acne, alopecia, eczema / atopic dermatitis, seborrheic dermatitis, rosacea, skin cancer…) or **changes** (e.g. skin aging…).
- Develop products *taking care of* healthy as well as damaged skin, providing **homeostasis**.

Skin - A short definition

Skin = "*[…] the largest organ in the body [covering] the body's entire external surface*". Hani et al., 2022, NCBI website.




How can we study this ecosystem?

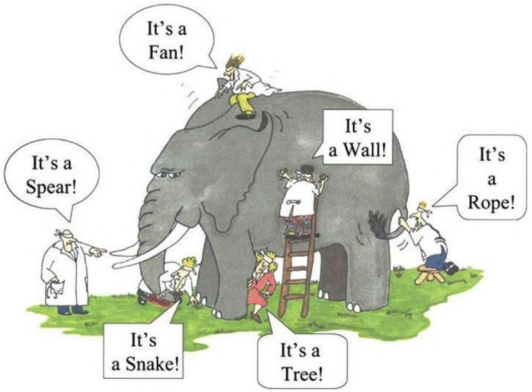
Multi-omics integration

What is it?
Why (and how) should we combine heterogeneous omics datasets?

Why combining them?

Why combining them?

To get an **overview** of possible interactions between different molecular levels happening on a given biological system of interest, *e.g.* the cutaneous ecosystem.



The (famous) elephant parable.

Linear multi-table approaches

Constraints: samples as basis, intermediate integration [Picard et al., 2021]

Linear multi-table approaches

Constraints: samples as basis, intermediate integration [Picard et al., 2021]

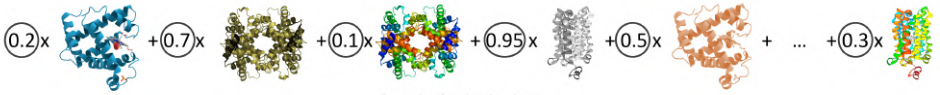
Generally, they produce...

Linear multi-table approaches

Constraints: samples as basis, intermediate integration [Picard et al., 2021]

Generally, they produce...

- **signatures**;



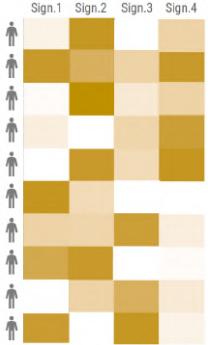
Example of molecular signature

Linear multi-table approaches

Constraints: samples as basis, intermediate integration [Picard et al., 2021]

Generally, they produce...

- **signatures**;
- individual **scores** associated to signatures.



Linear multi-table approaches

Constraints: samples as basis, intermediate integration [Picard et al., 2021]

Generally, they produce...

- **signatures**;
- individual **scores** associated to signatures.

Linear multi-table approaches

Constraints: samples as basis, intermediate integration [Picard et al., 2021]

Generally, they produce...

- **signatures**;
- individual **scores** associated to signatures.

They differ on...

1. optimisation problem;

Linear multi-table approaches

Constraints: samples as basis, intermediate integration [Picard et al., 2021]

Generally, they produce...

- **signatures**;
- individual **scores** associated to signatures.

They differ on...

1. optimisation problem;
2. sparsity (or not);

Linear multi-table approaches

Constraints: samples as basis, intermediate integration [Picard et al., 2021]

Generally, they produce...

- **signatures**;
- individual **scores** associated to signatures.

They differ on...

1. optimisation problem;
2. sparsity (or not);
3. way to combine signatures;

Linear multi-table approaches

Constraints: samples as basis, intermediate integration [Picard et al., 2021]

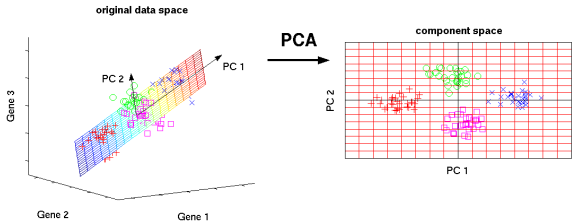
Generally, they produce...

- **signatures**;
- individual **scores** associated to signatures.

They differ on...

1. optimisation problem;
2. sparsity (or not);
3. way to combine signatures;
4. signature constraint(s).

MOFA [Argelaguet et al., 2018] [1]



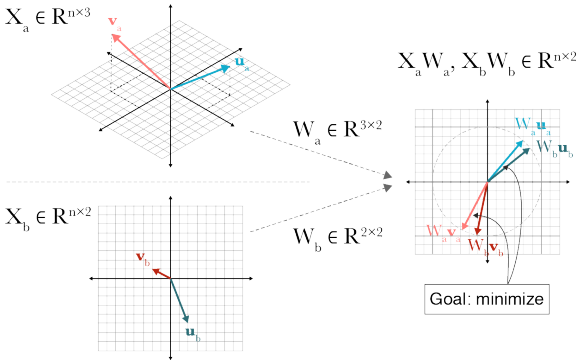
Source: nlpca website

1. **MFA** (*weighted PCA*) variant with Bayesian framework
2. no (direct) sparsity
3. common scores

⇒ **unsupervised**

[1] **MOFA2** R package - MOFA website

DIABLO [Singh et al., 2019] [2]



Source: Gundersen G. (2018)

1. sG**CCA** variant
   [Tenenhaus et al., 2014]
2. sparsity
3. scores for each omic

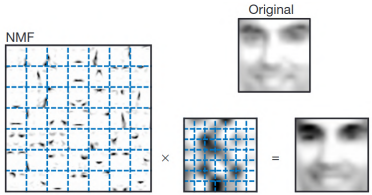
⇒ **mixed**[4]

→ groups described in relation to each other

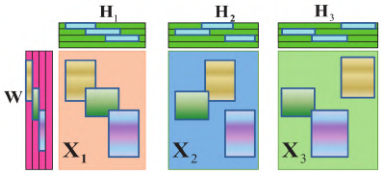
[2] **mixOmics** R package - mixOmics website

[4] performing both supervised and unsupervised tasks

jNMF [Zhang et al., 2012]



NMF on a picture [Lee and Seung, 1999]

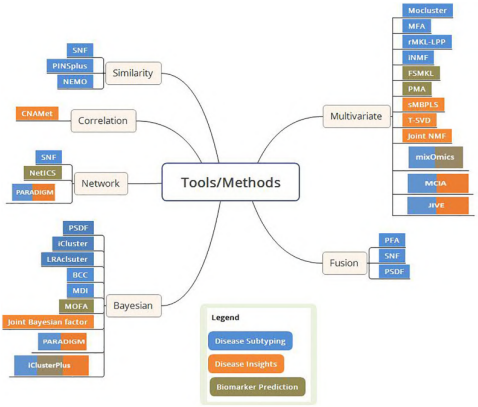


joint NMF on omics [Zhang et al., 2012]

1. NMF variant
2. no sparsity
3. common scores
4. positive signatures

⇒ **unsupervised**

The choice will depend on the question!



Source: [Subramanian et al., 2020]

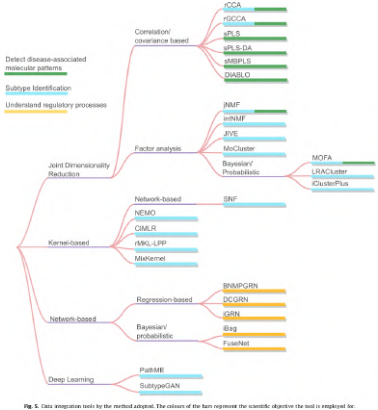


Fig. 5. Data integration tools by the method adopted. The colours of the bars represent the scientific objective the tool is employed for.

Source: [Athieniti and Spyrou, 2022]

Question(s) of interest

Biomarker research

Which are the describing elements of a dermatological state measured through multiple OMICS and clinical data?

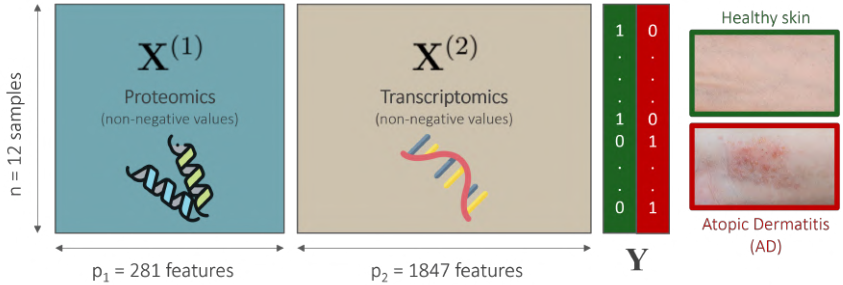
Association question

What are the molecular elements coming from 2 different OMIC types associated in all (or specific) samples?

Question(s) of interest

An example

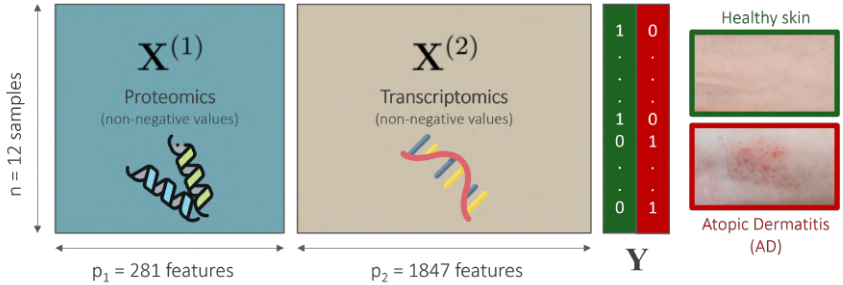
What group(s) of proteins and genes are associated together and explain the presence (or absence) of *Atopic Dermatitis* (AD) on samples?



Question(s) of interest

An example

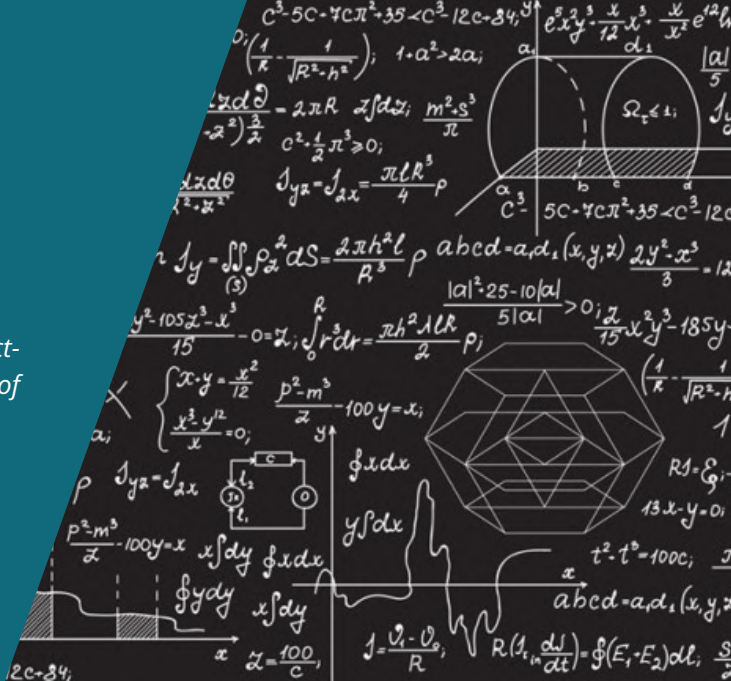
What group(s) of proteins and genes are associated together and explain the presence (or absence) of *Atopic Dermatitis* (AD) on samples?



Existing *mixed* approaches? Only a few.

NMFProfiler

A mixed integrative NMF extracting typical profiles of groups of interest.



Non-negative Matrix Factorization (NMF), [Lee and Seung, 1999]

A short introduction

Non-negative Matrix Factorization (NMF), [Lee and Seung, 1999]

A short introduction

Aim? Extract **typical profiles** of individuals with latent components (here, molecular **signatures**).

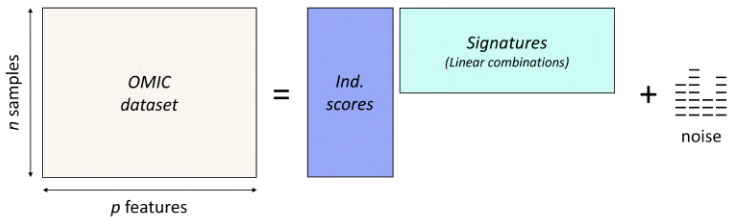
Non-negative Matrix Factorization (NMF), [Lee and Seung, 1999]

A short introduction

Aim? Extract **typical profiles** of individuals with latent components (here, molecular **signatures**).

How? Decompose the data matrix $\mathbf{X} \in \mathbb{R}_+^{n \times p}$ into **two non-negative** matrices $\mathbf{W} \in \mathbb{R}_+^{n \times K}$ and $\mathbf{H} \in \mathbb{R}_+^{K \times p}$:

$$\mathbf{X} \simeq \mathbf{WH}$$



- $\mathbf{W}$: "**contribution**" matrix of scores for n samples wrt each signature $k \in \{1, \cdots, K\}$;
- $\mathbf{H}$: "**signature**" (or "dictionary") matrix for K signatures.

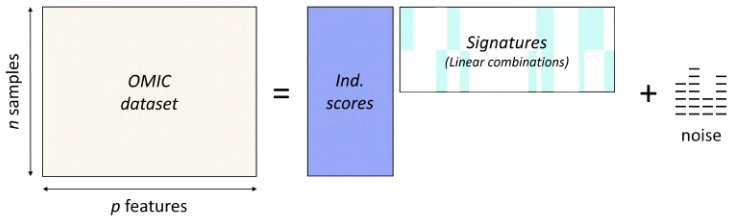
Non-negative Matrix Factorization (NMF), [Lee and Seung, 1999]

A short introduction

Aim? Extract **typical profiles** of individuals with latent components (here, molecular **signatures**).

How? Decompose the data matrix $\mathbf{X} \in \mathbb{R}_+^{n \times p}$ into **two non-negative** matrices $\mathbf{W} \in \mathbb{R}_+^{n \times K}$ and $\mathbf{H} \in \mathbb{R}_+^{K \times p}$:

$$\mathbf{X} \simeq \mathbf{WH}$$



- $\mathbf{W}$: "**contribution**" matrix of scores for n samples wrt each signature $k \in \{1, \cdots, K\}$;
- $\mathbf{H}$: "**signature**" (or "dictionary") matrix for K signatures.

Non-negative Matrix Factorization (NMF), [Lee and Seung, 1999]

A short introduction

Aim? Extract **typical profiles** of individuals with latent components (here, molecular **signatures**).

How? Decompose the data matrix $\mathbf{X} \in \mathbb{R}_+^{n \times p}$ into **two non-negative** matrices $\mathbf{W} \in \mathbb{R}_+^{n \times K}$ and $\mathbf{H} \in \mathbb{R}_+^{K \times p}$:

$$\mathbf{X} \simeq \mathbf{WH}$$

→ Decompose **jointly several** datasets? **[Zhang et al., 2012]**

→ Extract typical **group profiles**? **[Leuschner et al., 2019]**

Non-negative Matrix Factorization (NMF), [Lee and Seung, 1999]

A short introduction

Aim? Extract **typical profiles** of individuals with latent components (here, molecular **signatures**).

How? Decompose the data matrix $\mathbf{X} \in \mathbb{R}_+^{n \times p}$ into **two non-negative** matrices $\mathbf{W} \in \mathbb{R}_+^{n \times K}$ and $\mathbf{H} \in \mathbb{R}_+^{K \times p}$:

$$\mathbf{X} \simeq \mathbf{WH}$$

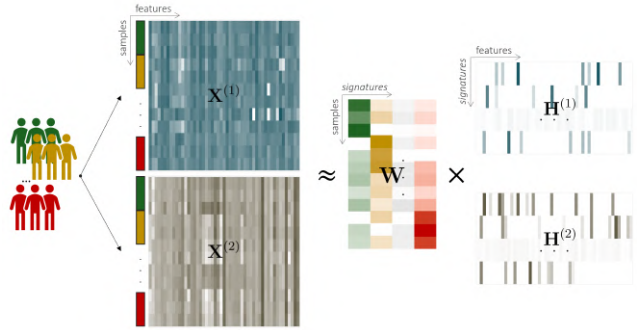
→ Decompose **jointly several** datasets? [Zhang et al., 2012]

→ Extract typical **group profiles**? [Leuschner et al., 2019]

NMFProfiler, a mixed integrative NMF

The math behind

Framework? $\mathbf{X}^{(j)} \in \mathbb{R}_+^{n \times p_j}$ the J OMICS datasets and $\mathbf{Y} \in \{0,1\}^{n \times U}$ the one-hot encoded groups.

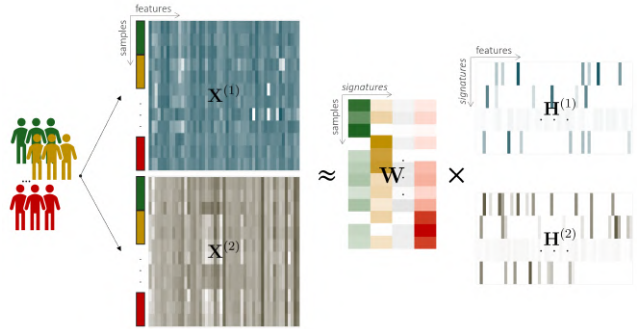


$$\frac{1}{2}\sum_{j=1}^{J} \underbrace{\left\|\mathbf{X}^{(j)} - \mathbf{W}\mathbf{H}^{(j)}\right\|_F^2}_{\text{goodness-of-fit}} + \lambda \sum_{j=1}^{J} \underbrace{\left\|\mathbf{H}^{(j)}\right\|_1}_{\text{sparsity}} + \frac{\mu}{2} \underbrace{\|\mathbf{W}\|_F^2}_{\text{regularization}} + \frac{\gamma}{2}\sum_{j=1}^{J} \underbrace{\left\|\mathbf{Y} - \mathbf{X}^{(j)}\mathbf{H}^{(j)\top}\mathrm{Diag}\left(\boldsymbol{\beta}^{(j)}\right)\right\|_F^2}_{U \text{ independent linear regressions}} \quad (1)$$

NMFProfiler, a mixed integrative NMF

The math behind

Framework? $\mathbf{X}^{(j)} \in \mathbb{R}_+^{n \times p_j}$ the J OMICS datasets and $\mathbf{Y} \in \{0,1\}^{n \times U}$ the one-hot encoded groups.

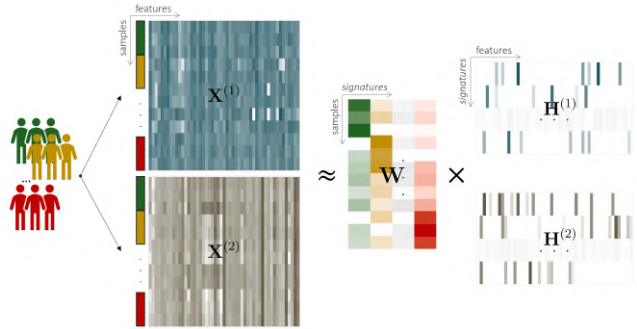


$$\frac{1}{2}\sum_{j=1}^{J}\underbrace{\left\|\mathbf{X}^{(j)}-\mathbf{W}\mathbf{H}^{(j)}\right\|_F^2}_{\text{goodness-of-fit}} + \lambda\sum_{j=1}^{J}\underbrace{\left\|\mathbf{H}^{(j)}\right\|_1}_{\text{sparsity}} + \frac{\mu}{2}\underbrace{\|\mathbf{W}\|_F^2}_{\text{regularization}} + \frac{\gamma}{2}\sum_{j=1}^{J}\underbrace{\left\|\mathbf{Y}-\mathbf{X}^{(j)}\mathbf{H}^{(j)\top}\mathrm{Diag}\left(\boldsymbol{\beta}^{(j)}\right)\right\|_F^2}_{\text{U independent linear regressions}}$$

NMFProfiler, a mixed integrative NMF

The math behind

Framework? $\mathbf{X}^{(j)} \in \mathbb{R}_+^{n \times p_j}$ the J OMICS datasets and $\mathbf{Y} \in \{0,1\}^{n \times U}$ the one-hot encoded groups.

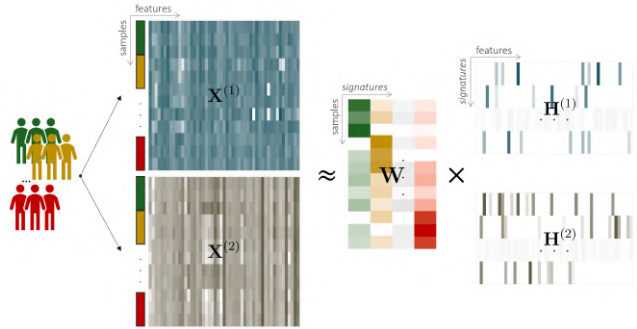


$$\frac{1}{2}\sum_{j=1}^{J}\underbrace{\left\|\mathbf{X}^{(j)}-\mathbf{W}\mathbf{H}^{(j)}\right\|_F^2}_{\text{goodness-of-fit}} + \lambda\sum_{j=1}^{J}\underbrace{\left\|\mathbf{H}^{(j)}\right\|_1}_{\text{sparsity}} + \frac{\mu}{2}\underbrace{\|\mathbf{W}\|_F^2}_{\text{regularization}} + \frac{\gamma}{2}\sum_{j=1}^{J}\underbrace{\left\|\mathbf{Y}-\mathbf{X}^{(j)}\mathbf{H}^{(j)\top}\mathrm{Diag}\left(\boldsymbol{\beta}^{(j)}\right)\right\|_F^2}_{\text{U independent linear regressions}}$$

NMFProfiler, a mixed integrative NMF

The math behind

Framework? $\mathbf{X}^{(j)} \in \mathbb{R}_+^{n \times p_j}$ the J OMICS datasets and $\mathbf{Y} \in \{0,1\}^{n \times U}$ the one-hot encoded groups.



$$\frac{1}{2}\sum_{j=1}^{J}\underbrace{\left\|\mathbf{X}^{(j)}-\mathbf{W}\mathbf{H}^{(j)}\right\|_F^2}_{\text{goodness-of-fit}}+\lambda\sum_{j=1}^{J}\underbrace{\left\|\mathbf{H}^{(j)}\right\|_1}_{\text{sparsity}}+\frac{\mu}{2}\underbrace{\|\mathbf{W}\|_F^2}_{\text{regularization}}+\frac{\gamma}{2}\sum_{j=1}^{J}\underbrace{\left\|\mathbf{Y}-\mathbf{X}^{(j)}\mathbf{H}^{(j)\top}\mathrm{Diag}\left(\boldsymbol{\beta}^{(j)}\right)\right\|_F^2}_{\text{U independent linear regressions}}$$

NMFProfiler

A new optimization algorithm Appendix 3

- All terms **alternatively updated** thanks to an **iterative approach** (gradient descent).

NMFProfiler

A new optimization algorithm Appendix 3

- All terms **alternatively updated** thanks to an **iterative approach** (gradient descent).
- Generally, updates under the shape of **Multiplicative Updates** (MU).

NMFProfiler

A new optimization algorithm Appendix 3

- All terms **alternatively updated** thanks to an **iterative approach** (gradient descent).
- Generally, updates under the shape of **Multiplicative Updates** (MU).
- But $\mathbf{H}^{(j)}$ obtained not directly sparse.
  Thus, use the **Proximal approach** to update these matrices.

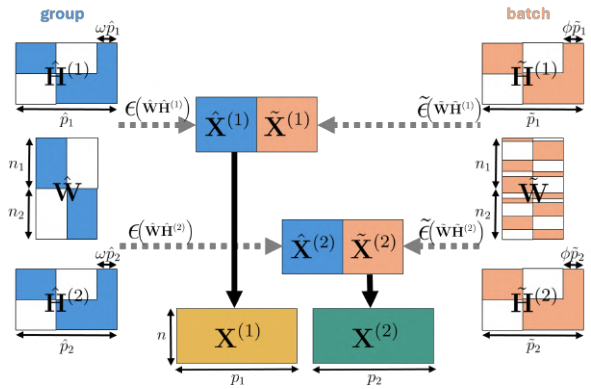
Results from simulations

Test NMFProfiler on simulated datasets. Compare with state-of-the-art methods.

Simulated data see data

Simulated data

see data

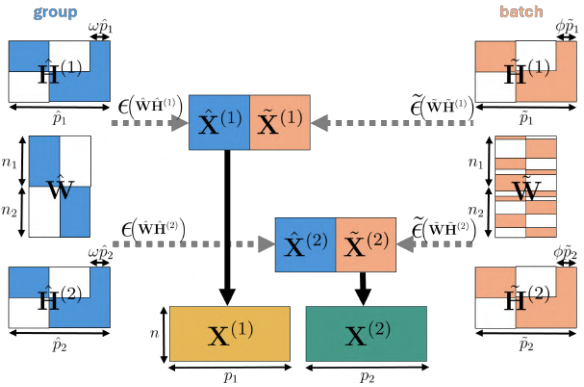


Data generation process (based on [Yang and Michailidis, 2016])

- 2 OMICS datasets with $n = 50$, $p_1 = 2500$ and $p_2 = 400$;
- $K = 2$ signatures (since 2 groups);

Simulated data

see data



Data generation process (based on [Yang and Michailidis, 2016])

- 2 OMICS datasets with $n = 50$, $p_1 = 2500$ and $p_2 = 400$;
- $K = 2$ signatures (since 2 groups);

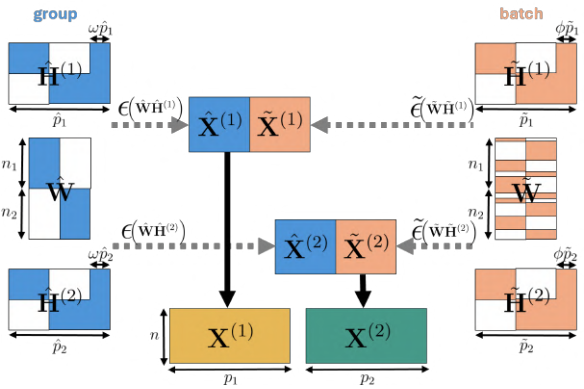
→ Groups perfectly balanced;

→ Group and batch patterns of the same size and no noisy features;

→ More noise in group patterns.

Simulated data

see data



Data generation process (based on [Yang and Michailidis, 2016])

- 2 OMICS datasets with $n = 50$, $p_1 = 2500$ and $p_2 = 400$;
- $K = 2$ signatures (since 2 groups);

→ Groups perfectly balanced;

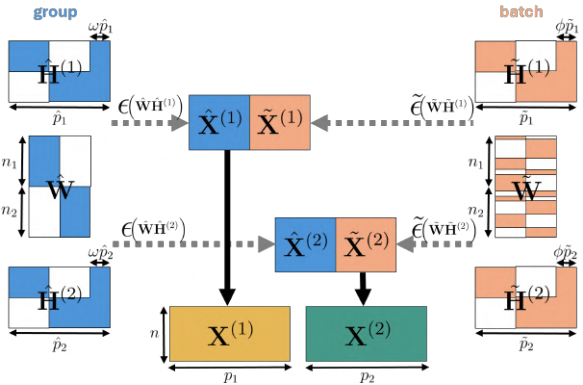
→ Group and batch patterns of the same size and no noisy features;

→ More noise in group patterns.

- **Aim?** Identify precisely biomarkers of a given group (e.g. healthy / DA) only.

Simulated data

see data



Data generation process (based on [Yang and Michailidis, 2016])

- 2 OMICS datasets with $n = 50$, $p_1 = 2500$ and $p_2 = 400$;
- $K = 2$ signatures (since 2 groups);

→ Groups perfectly balanced;

→ Group and batch patterns of the same size and no noisy features;

→ More noise in group patterns.

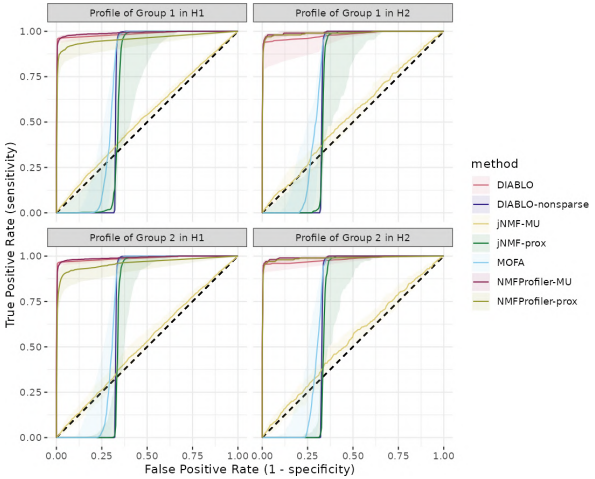
- **Aim?** Identify precisely biomarkers of a given group (e.g. healthy / DA) only.

- Compared **NMProfiler**[c] to state-of-the-art methods in multi-omics analysis ([Argelaguet et al., 2018, Singh et al., 2019]).

[c] : both MU and proximal solvers

Results

Median ROCs on 50 simulations for each dataset j and group k



Results

General conclusions on simulated data Appendix 5

- Best methods: **NMFProfiler** *(both solvers)* and DIABLO.

[3] Up to a certain threshold.

[4] Unadapted hyperparameters. See Cohen and Leplat (2024).

Results

General conclusions on simulated data Appendix 5

- Best methods: **NMFProfiler** *(both solvers)* and DIABLO.
- **NMFProfiler-prox** succeeds in **selecting features** depending on the phenotype, **classifies** well **samples** and runs fast.

[3] Up to a certain threshold.

[4] Unadapted hyperparameters. See Cohen and Leplat (2024).

Results

General conclusions on simulated data Appendix 5

- Best methods: **NMFProfiler** *(both solvers)* and DIABLO.
- **NMFProfiler-prox** succeeds in **selecting features** depending on the phenotype, **classifies** well **samples** and runs fast.
- **Robust to group desequilibrium** and **batch effect**.[3]

[3] Up to a certain threshold.

[4] Unadapted hyperparameters. See Cohen and Leplat (2024).

Results

General conclusions on simulated data Appendix 5

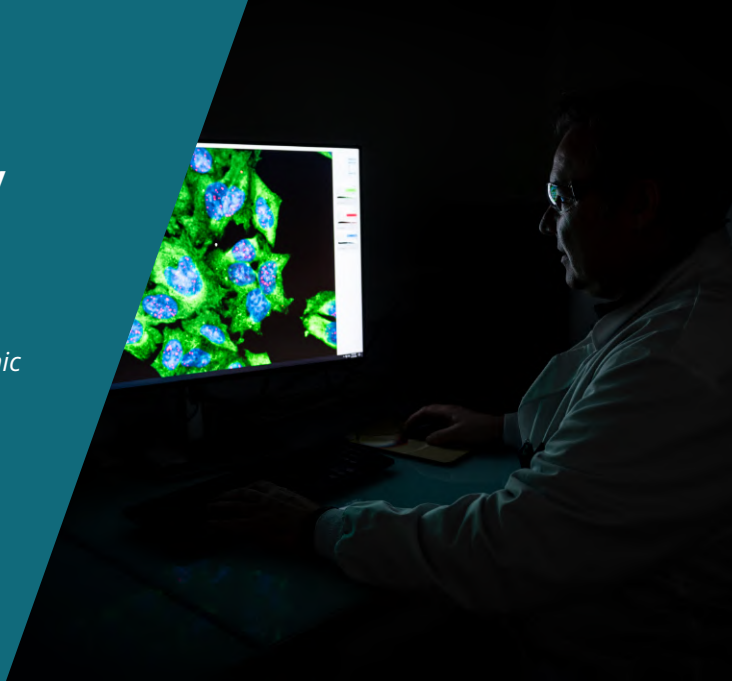
- Best methods: **NMFProfiler** *(both solvers)* and DIABLO.
- **NMFProfiler-prox** succeeds in **selecting features** depending on the phenotype, **classifies** well **samples** and runs fast.
- **Robust to group desequilibrium** and **batch effect**.[3]
- **NMFProfiler-MU** more robust to noise than NMFProfiler-prox.[4]

[3] Up to a certain threshold.

[4] Unadapted hyperparameters. See Cohen and Leplat (2024).

Atopic Dermatitis study

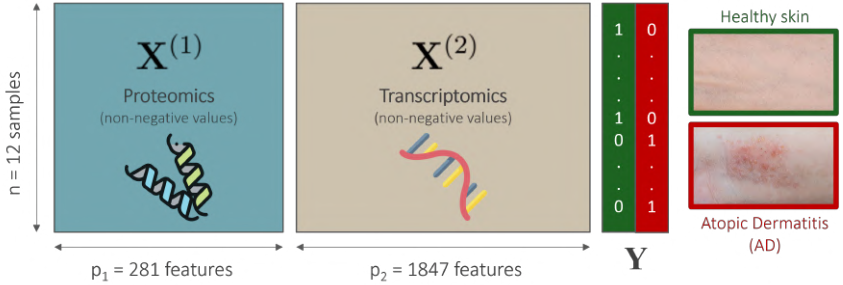
Profile 2 groups based on 2 omic datasets



Profile Atopic Dermatitis

Question and datasets

What group(s) of proteins and genes are associated together and explain the presence (or absence) of *Atopic Dermatitis* (AD) on samples?

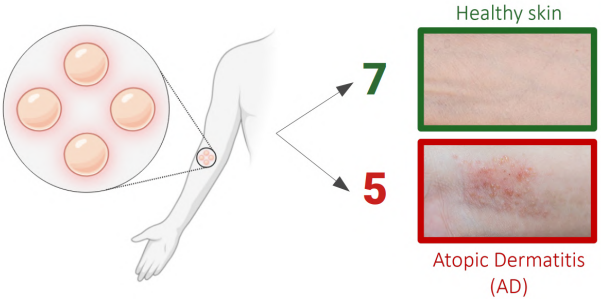


→ Analyzed with **NMFProfiler-prox** and DIABLO

Profile Atopic Dermatitis

Question and datasets

What group(s) of proteins and genes are associated together and explain the presence (or absence) of *Atopic Dermatitis* (AD) on samples?

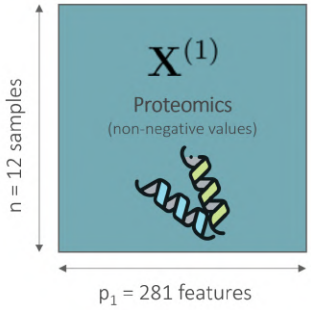


- common inflammatory skin disease
- $n = 12$ volunteers
- suction blister in non-lesional area

Profile Atopic Dermatitis

Question and datasets

What group(s) of proteins and genes are associated together and explain the presence (or absence) of *Atopic Dermatitis* (AD) on samples?

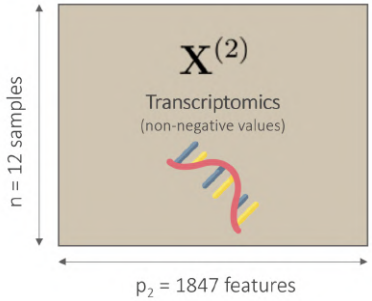


- LC/MS
  $\rightarrow p_1 = 1303$
- $\log 2$-transformed
- quantile normalization
- low count / variance proteins filtered out
- batch correction with `ComBat`
  [Johnson et al., 2007]

Profile Atopic Dermatitis

Question and datasets

What group(s) of proteins and genes are associated together and explain the presence (or absence) of *Atopic Dermatitis* (AD) on samples?



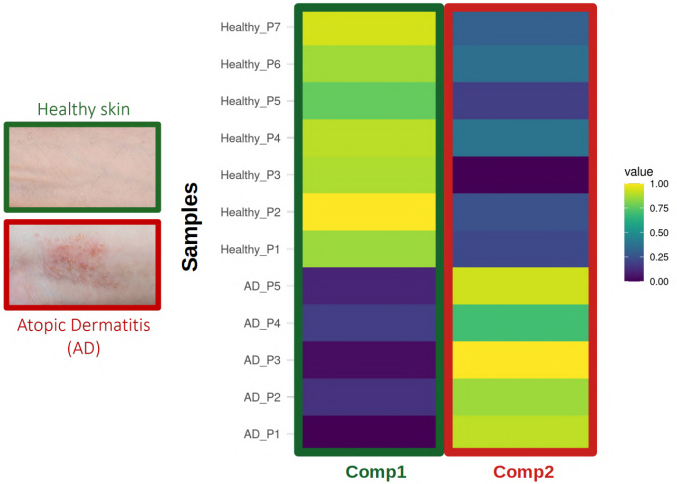
- Human Gene Array Plates (Affymetrix)
  $\rightarrow p_2 = 53617$
- RMA **[Irizarry et al., 2003]**
- probes expressed below background filtered out
- batch correction (as before)

Profile Atopic Dermatitis

Heatmap of contribution matrix $\mathbf{W}$

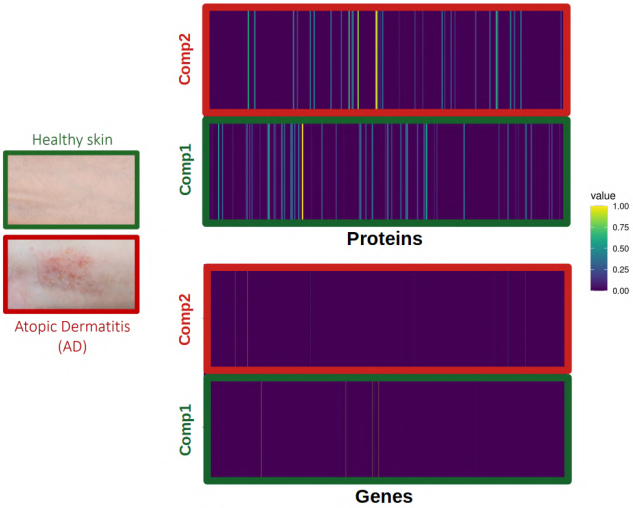
Profile Atopic Dermatitis

Heatmap of contribution matrix $\mathbf{W}$



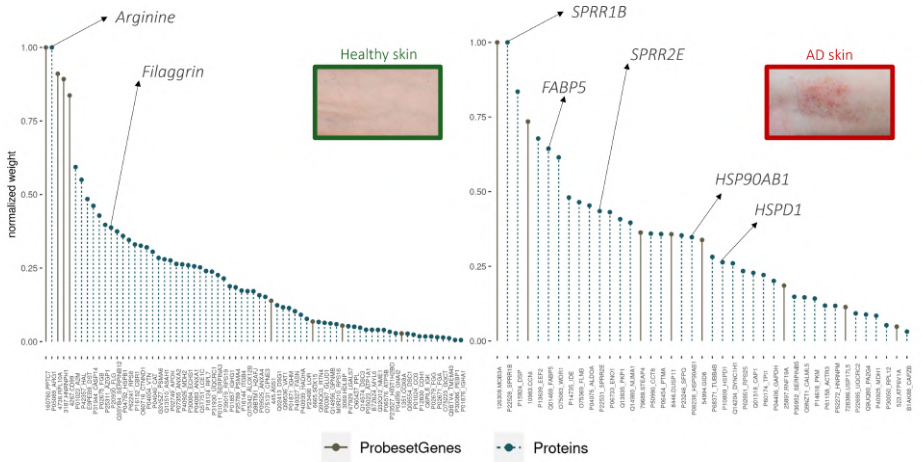
Profile Atopic Dermatitis

Heatmaps of dictionary matrices $\mathbf{H}^{(j)}$



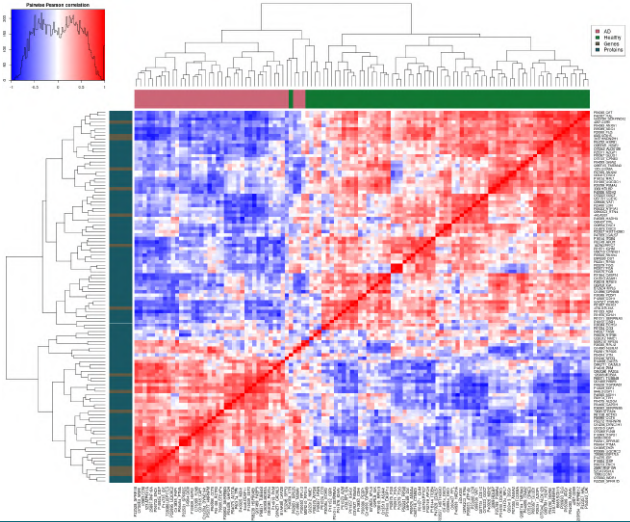
Profile Atopic Dermatitis

Features characterizing each profile



Profile Atopic Dermatitis

Pairwise correlation matrix



Profile Atopic Dermatitis

Conclusions

- **Sparse** signatures containing **known biomarkers** of (non-lesional) **AD** skin;
- Features selected for describing a given group are **associated** together;
- Possibly *new biomarkers* uncovered;

Profile Atopic Dermatitis

Conclusions

- **Sparse** signatures containing **known biomarkers** of (non-lesional) **AD** skin;
- Features selected for describing a given group are **associated** together;
- Possibly *new biomarkers* uncovered;

- Similar results with DIABLO;
- NMFProfiler's **signatures less redundant**.

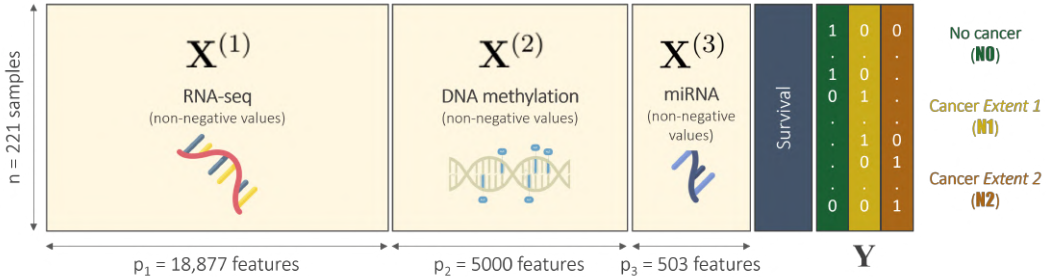
Colon adenocarcinoma study (TCGA)

Profile 3 groups based on 3 omic datasets

TCGA: Colon adenocarcinoma study (COAD)

Question and datasets (**downloaded here**)

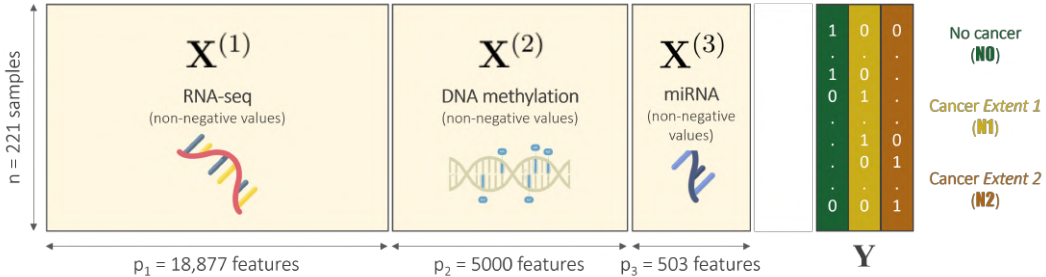
What group(s) of genes, methylated DNA sites and miRNA are associated together and explain the different stages of regional lymph nodes involvement?



TCGA: Colon adenocarcinoma study (COAD)

Question and datasets (**downloaded here**)

What group(s) of genes, methylated DNA sites and miRNA are associated together and explain the different stages of regional lymph nodes involvement?



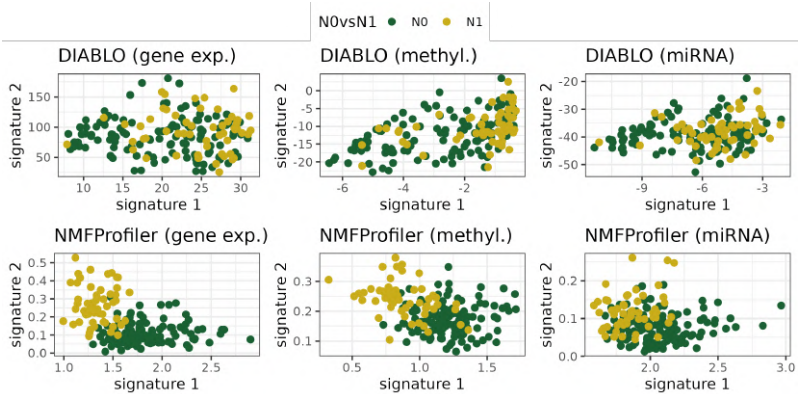
→ Analyzed with **NMFProfiler-MU** (pairs' subsample, full categorical variable) and DIABLO (only pairs' subsample).

Profile lymph nodes extent stages

N0 versus N1

Profile lymph nodes extent stages

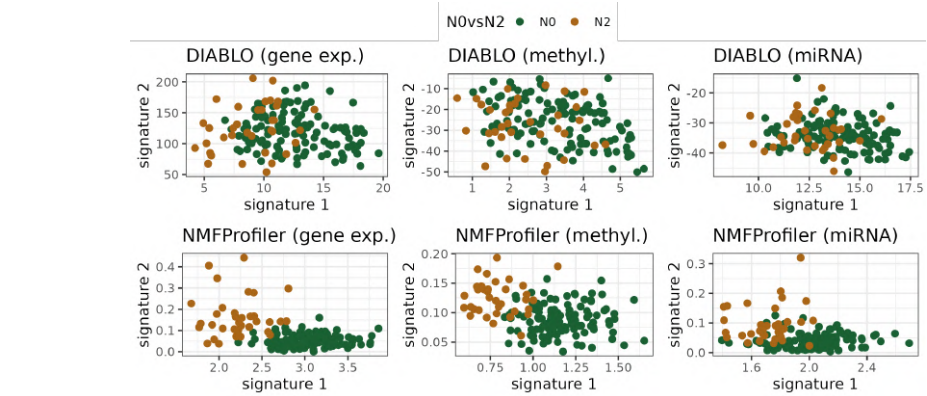
N0 versus N1



Projection of samples onto signatures obtained for `N0vsN1` *for each omic and method. For DIABLO, only the x-axis (first signature) is relevant (split based on sign).*

Profile lymph nodes extent stages

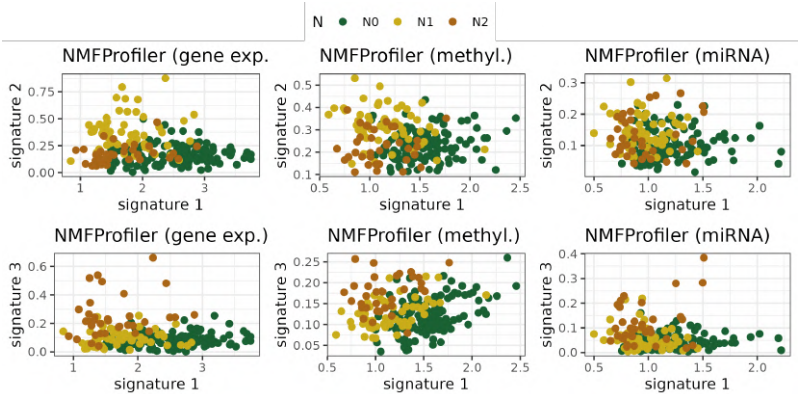
N0 versus N2



Projections of samples onto signatures of `N0vsN2` *for each omic and method. For DIABLO, only the x-axis (first signature) is relevant (split based on sign).*

Profile lymph nodes extent stages

N0, N1 and N2



Projections of samples onto signatures of `N` *obtained by NMFProfiler for each omic. The first signature corresponds to group* `N0`*, the second to group* `N1` *and the third to group* `N2`*.*

Profile lymph nodes extent stages

Conclusions Appendix 6

NMFProfiler:

- shows a better ability than DIABLO to separate the two groups;
- produces **a specific profile for each group**, easing the interpretation;

Profile lymph nodes extent stages

Conclusions Appendix 6

NMFProfiler:

- shows a better ability than DIABLO to separate the two groups;
- produces **a specific profile for each group**, easing the interpretation;
- finds some signatures to be **predictive of survival** (hard to find on COAD **[Rappoport and Shamir, 2018, Cantini et al., 2021]**).

Conclusion

Conclusion and perspectives

- Developped **NMFProfiler** for multi-omics group profile extraction.
- Flexible, interpretable and competitive.
- Able to draw a **specific profile by group** from $J \geq 2$ omics and for $U \geq 2$ distinct groups.
- **Used** on real data to uncover biomarkers of AD.

[5] like γ, λ

Conclusion and perspectives

- Developped **NMFProfiler** for multi-omics group profile extraction.
- Flexible, interpretable and competitive.
- Able to draw a **specific profile by group** from $J \geq 2$ omics and for $U \geq 2$ distinct groups.
- **Used** on real data to uncover biomarkers of AD.

→ **nmfprofiler** implemented in a Python package: check out GitLab or PyPI.

[5] like γ, λ

Conclusion and perspectives

- Developped **NMFProfiler** for multi-omics group profile extraction.
- Flexible, interpretable and competitive.
- Able to draw a **specific profile by group** from $J \geq 2$ omics and for $U \geq 2$ distinct groups.
- **Used** on real data to uncover biomarkers of AD.

→ **nmfprofiler** implemented in a Python package: check out GitLab or PyPI.

- Keep on investigating results on real data (*e.g.* multi-omics enrichment analysis).
- Calibrate hyperparameters[5].

[5] like γ, λ

Thank you for your attention.

Feel free to ask questions.

Contact: aurelie.mercadie@inrae.fr | aurelie.mercadie@pierre-fabre.com

Credits:

Presentation template: Inspired by the IUC template by Usama Muneeb (GitHub: https://github.com/usamamuneeb/uic-beamer-template)

Pictures: Pierre Fabre pictures from *Communication ToolKit*, Krassowki et al. (2020) for the 2nd chapter.

References I

Figure 1: The Blind Men and the Elephant.

Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J. C., Buettner, F., Huber, W., and Stegle, O. (2018).
Multi-omics factor analysis—a framework for unsupervised integration of multi-omics data sets.
Molecular Systems Biology, 14(6):e8124.

Athieniti, E. and Spyrou, G. M. (2022).
A guide to multi-omics data collection and integration for translational medicine.
Computational and structural biotechnology journal, 21:134–149.

Cantini, L., Zakeri, P., Hernandez, C., Naldi, A., Thieffry, D., Remy, E., and Baudot, A. (2021).
Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer.
Nature Communications, 12(1):124.

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., and Speed, T. P. (2003).
Exploration, normalization, and summaries of high density oligonucleotide array probe level data.
Biostatistics (Oxford, England), 4(2):249–264.

Johnson, W. E., Li, C., and Rabinovic, A. (2007).
Adjusting batch effects in microarray expression data using empirical bayes methods.
Biostatistics (Oxford, England), 8(1):118–127.

Lee, D. D. and Seung, H. S. (1999).
Learning the parts of objects by non-negative matrix factorization.
Nature, 401:788–791.

References II

Leuschner, J., Schmidt, M., Fernsel, P., Lachmund, D., Boskamp, T., and Maass, P. (2019).
Supervised non-negative matrix factorization methods for maldi imaging applications.
Bioinformatics, 35:1940–1947.

Picard, M., Scott-Boyer, M.-P., Bodein, A., Périn, O., and Droit, A. (2021).
Integration strategies of multi-omics data for machine learning analysis.
Computational and Structural Biotechnology Journal, 19:3735–3746.

Rappoport, N. and Shamir, R. (2018).
Multi-omic and multi-view clustering algorithms: review and cancer benchmark.
Nucleic Acids Research, 46(20):10546–10562.

Singh, A., Shannon, C. P., Gautier, B., Rohart, F., Vacher, M., Tebbutt, S. J., and Lê Cao, K.-A. (2019).
DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays.
Bioinformatics (Oxford, England), 35:3055–3062.

Subramanian, I., Verma, S., Kumar, S., Jere, A., and Anamika, K. (2020).
Multi-omics data integration, interpretation, and its application.
Bioinformatics and biology insights, 14:1177932219899051.

Tenenhaus, A., Philippe, C., Guillemot, V., Le Cao, K.-A., Grill, J., and Frouin, V. (2014).
Variable selection for generalized canonical correlation analysis.
Biostatistics, 15(3):569–583.

References III

Yang, Z. and Michailidis, G. (2016).
A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data.
Bioinformatics, 32(1):1–8.

Zhang, S., Liu, C.-C., Li, W., Shen, H., Laird, P. W., and Zhou, X. J. (2012).
Discovery of multi-dimensional modules by integrative analysis of cancer genomic data.
Nucleic acids research, 40(19):9379–9391.

Appendix 1

Integrative NMF, [Zhang et al., 2012]

Framework? $\mathbf{X}^{(j)} \in \mathbb{R}_+^{n \times p_j}$ the J OMICS datasets.

jNMF:

$$\min_{\mathbf{W}, \mathbf{H}^{(1)}, \ldots, \mathbf{H}^{(J)} \geq 0} \sum_{j=1}^{J} \|\mathbf{X}^{(j)} - \mathbf{W}\mathbf{H}^{(j)}\|_F^2$$

- $\mathbf{W} \in \mathbb{R}_+^{n \times K}$ the **common** "contribution" matrix;
- $\mathbf{H}^{(j)} \in \mathbb{R}_+^{K \times p_j}$ the "dictionary" matrices;
- K the number of signatures to choose.

- **No supervision**.
- **No sparsity** in signatures.

Appendix 2

Supervised NMF, [Leuschner et al., 2019]

Framework? $\mathbf{X} \in \mathbb{R}_+^{n \times p}$ the OMIC dataset and $\mathbf{y} \in \{0,1\}^n$ the groups of interest.

FR-lda:

$$\min_{\mathbf{W},\mathbf{H},\boldsymbol{\beta} \geq 0} \frac{1}{2} \underbrace{\|\mathbf{X} - \mathbf{W}\mathbf{H}\|_F^2}_{\text{goodness-of-fit}} + \underbrace{\lambda \|\mathbf{H}\|_1}_{\text{sparsity}} + \underbrace{\frac{\mu}{2}\|\mathbf{W}\|_F^2 + \frac{\nu}{2}\|\mathbf{H}\|_F^2}_{\text{regularization}} + \frac{\gamma}{2} \underbrace{\|\mathbf{y} - \mathbf{X}\mathbf{H}^T\boldsymbol{\beta}\|_2^2}_{\text{LDA}}$$

- $\mathbf{W} \in \mathbb{R}_+^{n \times K}$ the "contribution" matrix;
- $\mathbf{H} \in \mathbb{R}_+^{K \times p}$ the "dictionary" matrix;
- $\lambda, \mu, \nu, \gamma > 0$ the regularization parameters (given);
- $\boldsymbol{\beta} \in \mathbb{R}_+^K$ the regression coefficients vector;
- K the number of signatures to choose.

- For **one OMIC** only.
- **Sparsity** obtained by thresholding (no *true sparsity*).
- Use of regularization terms?
- Efficiency of supervised term?

Appendix 3 back to slides

Algorithm Overview of the *Proximal* algorithm used to minimize Equation (1)

Initialize matrices $\mathbf{W}^{(0)}$, $\mathbf{H}^{(j,0)}$, vectors $\boldsymbol{\beta}^{(j,0)}$ with strictly positive values.
for all $t = 1, \ldots, T$ **do**
  MU update: $\mathbf{W}^{(t+1)} \leftarrow \mathbf{W}^{(t)} \odot \mathbf{A}(\mathbf{W}^{(t)})$ of interest
  Prox update: $\forall j = 1, \ldots, J$,

$$\mathbf{H}^{(j,t+1)} \leftarrow \text{prox}_{\tilde{g}_j}\left(\tilde{\mathbf{H}}^{(j)}\right), \ \tilde{\mathbf{H}}^{(j)} = \mathbf{H}^{(j,t)} - \frac{1}{\eta}\nabla f_j(\mathbf{H}^{(j,t)})$$

  OLS solution: $\forall j = 1, \ldots, J, \forall k = 1, \ldots, U$,

$$\boldsymbol{\beta}_k^{(j,t+1)} \leftarrow \frac{\mathbf{H}_{k.}^{(j,t+1)}\mathbf{X}^{(j)\top}\mathbf{Y}_{.k}}{\mathbf{H}_{k.}^{(j,t+1)}\mathbf{X}^{(j)\top}\mathbf{X}^{(j)}\mathbf{H}_{k.}^{(j,t+1)\top}}$$

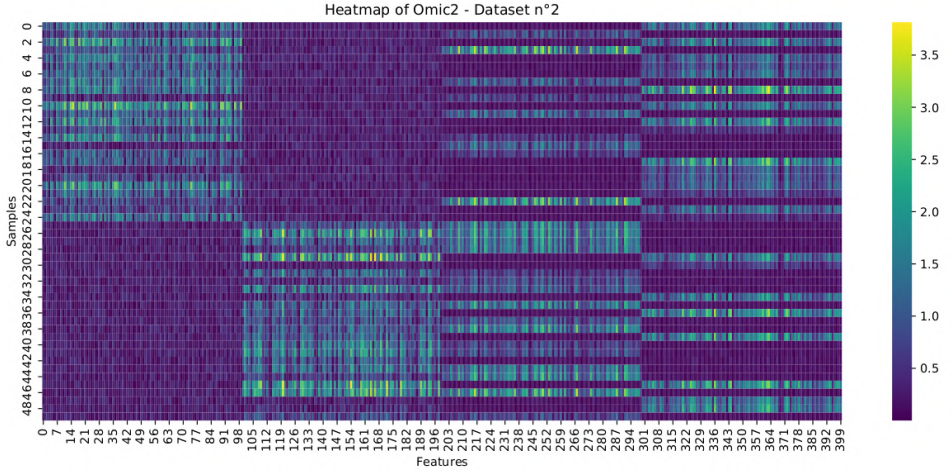
end for

return $\mathbf{W} := \mathbf{W}^{(T+1)}$, $\mathbf{H}^{(j)} := \mathbf{H}^{(j,T+1)}$ and $\boldsymbol{\beta}^{(j)} := \boldsymbol{\beta}^{(j,T+1)}$ $(j = 1, \ldots, J)$

where $\mathbf{A}(\mathbf{W}^{(t)})$ is a matrix with positive entries, prox is the proximal operator and f_j and $\tilde{g}_j$ are two functions.

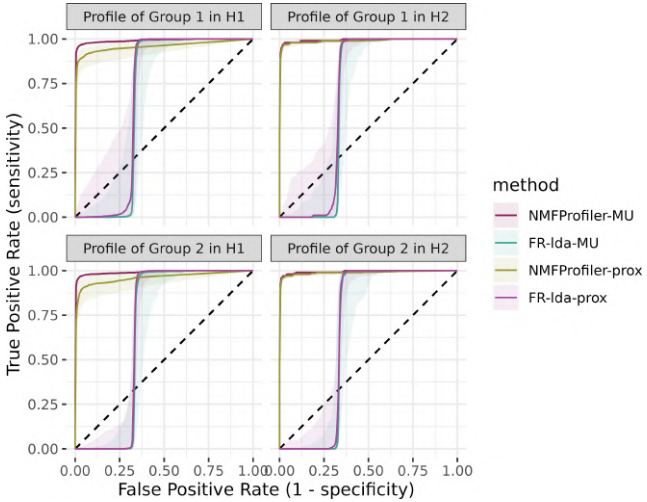
Appendix 4

Illustration of a simulated dataset back to slides



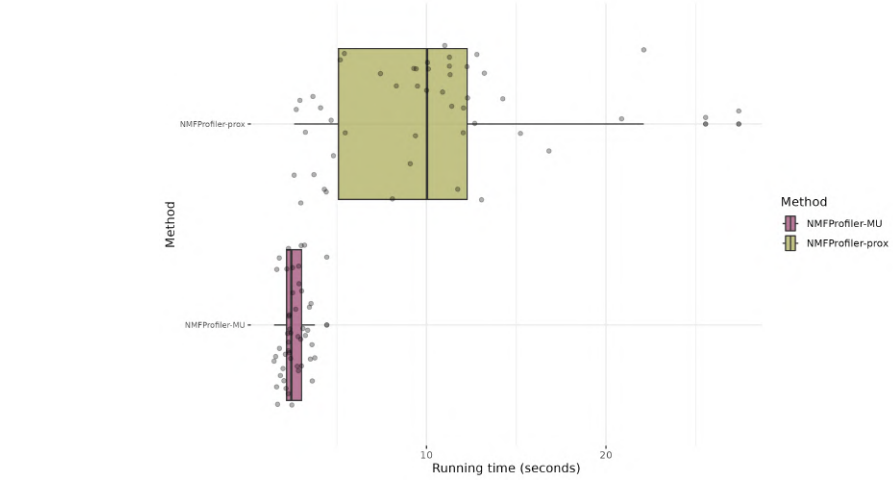
Appendix 5

Median ROCs on 50 simulations given both supervised terms and solvers back to slides



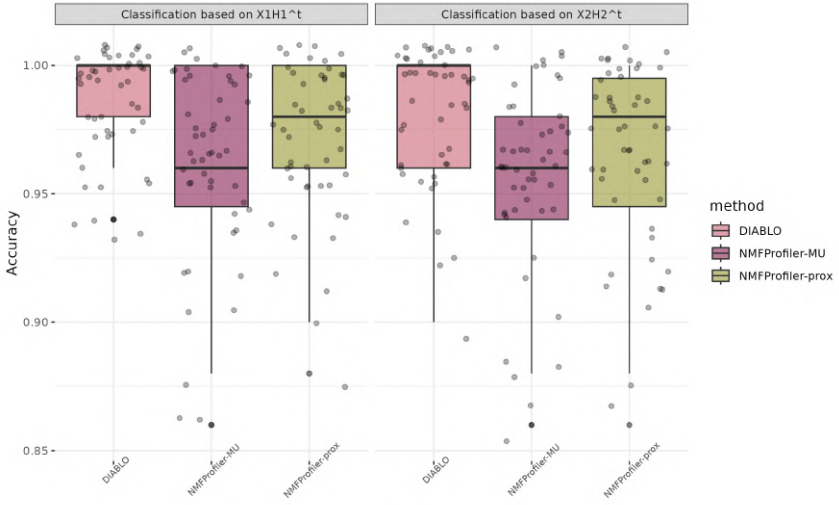
Appendix 5

Distribution of computational time across 50 simulations (focus on NMFProfiler) back to slides



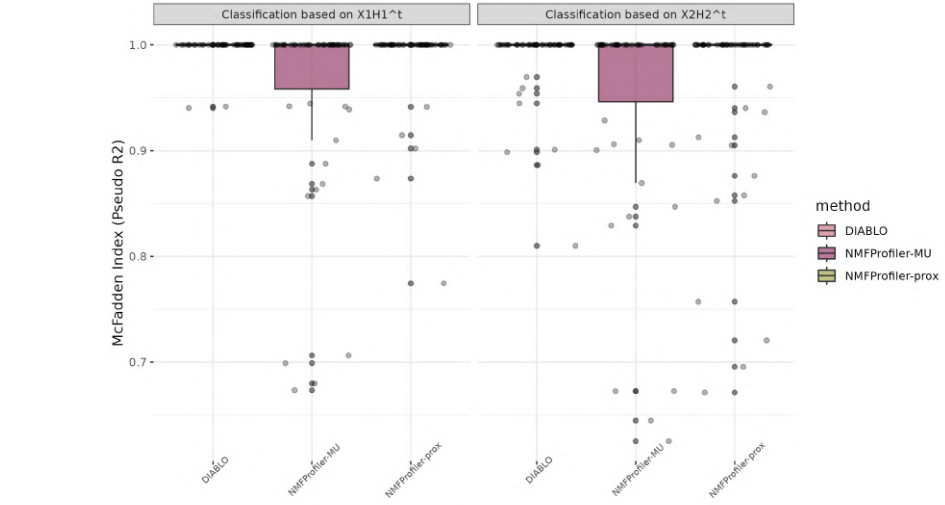
Appendix 5

Sample classification (accuracy measured on logistic reg.) back to slides



Appendix 5

Sample classification (McFadden index measured on logistic reg.) back to slides



Appendix 6

Are signature predictive of survival? back to slides

Cox proportional hazard model, based on [Rappoport and Shamir, 2018, Cantini et al., 2021] works:

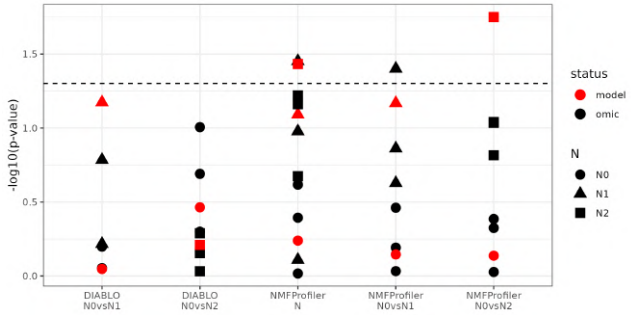
1. $\forall j \in \{1, 2, 3\}$, compute $\mathbf{X}^{(j)}\widehat{\mathbf{H}}^{(j)\top} \in \mathbb{R}^{n \times K}$, the projection of samples onto signatures;
2. For each group, extract the corresponding rows and signature in $\mathbf{X}^{(j)}\widehat{\mathbf{H}}^{(j)\top}, \forall j \in \{1, 2, 3\}$;
3. Concatenate the $J = 3$ submatrices of group $u \Rightarrow$ they become the 3 predictors in the CPH model (`coxph()` in R package **survival**);
4. Extract the 4 p-values: 3 omic-specific p-values[6] and a global model p-value.[7]

[6] Wald test

[7] Likelihood ratio test of the full model against the empty model

Appendix 6

Are signature predictive of survival? back to slides



$-\log_{10}$ *(p-values) obtained with Cox proportional hazard models for the association of survival to both* `N0vsN1` *and* `N0vsN2` *signatures obtained by DIABLO and NMFProfiler. The full (versus null) model p-value is displayed in red and the three p-values corresponding to an omic-specific signature are displayed in black. The dashed horizontal line corresponds to a p-value of* 0.05.