# Utilisation de la NMF supervisée intégrative pour l'étude d'altérations de la peau

Aurélie Mercadié[1,2], Éléonore Gravier[2], Gwendal Josse[2], Nathalie Vialaneix[1] & Céline Brouard[1]

[1] *Université de Toulouse, INRAE, UR MIAT, F-31320, Castanet-Tolosan, France,* {*aurelie.mercadie, nathalie.vialaneix, celine.brouard*}*@inrae.fr*
[2] *Pierre Fabre Dermo-Cosmétique, F-31300, Toulouse, France,* {*eleonore.gravier, gwendal.josse*}*@pierre-fabre.com*

**Résumé.** Cette communication est motivée par un problème fréquent en recherche clinique : des patients, stratifiés en groupes d'intérêt (typiquement sains / malades ou contrôles / traités) sont caractérisés par des mesures biologiques correspondant à plusieurs caractéristiques différentes (métabolomiques, protéomiques, etc). L'objectif est alors de découvrir des signatures moléculaires multi-omiques caractérisant les groupes.

Ici, nous proposons une approche de Factorisation Matricielle Non-négative (NMF) que nous étendons à ce cadre. De manière plus précise, notre proposition se fonde sur une variante du problème d'optimisation FR-lda [8], qui, par l'introduction d'un terme supervisé, permet de prendre en compte la structuration des individus en groupes d'intérêt. Notre proposition étend cette méthode à un cadre multi-tableaux, en assurant l'intégration des informations par le biais d'une composante de pondération commune à tous les tableaux. Nous proposons deux approches pour résoudre le problème d'optimisation induit, l'une classique, par approche multiplicative, et l'autre, nouvelle, par approche proximale et qui permet d'obtenir une parcimonie exacte dans les signatures moléculaires.

Nous illustrerons l'utilisation de cette extension sur des données de protéomique et de transcriptomique issus d'échantillons de peau prélevés sur une zone non-lésionnelle et montrerons comment elle nous a permis d'identifier une signature multi-omique de la Dermatite Atopique (DA), une maladie inflammatoire commune principalement caractérisée par une fonction barrière de la peau dysfonctionnelle.

**Mots-clés.** intégration multi-omiques, apprentissage supervisé, NMF, optimisation proximale.

**Abstract.** This communication is motivated by a frequent problem in clinical research: patients, stratified into groups of interest (typically healthy/sick or control/treated patients) are described by biological measurements corresponding to different omics (metabolomics, proteomics, etc.). The aim is then to discover molecular signatures characterizing the groups.

Here, we propose a Non-negative Matrix Factorization (NMF) approach that we extend to this framework. More specifically, our proposal is based on the FR-lda variant of NMF [8]. This method introduces a supervised term, aiming at explaining the two groups of individuals. Our proposal extends this method to a multi-table framework, by integrating information through a weighting matrix common to all omics. We also propose two approaches to solve

the induced optimization problem, the classical multiplicative approach (MU) and a novel proximal approach that achieves exact sparsity in molecular signatures.

The use of this extension is illustrated on proteomic and transcriptomic data from skin samples taken in a non-lesional area of subjects with Atopic Dermatitis (DA). First results show that our NMF variant identifies a multi-omic signature of the disease.

**Keywords.** multi-omics integration, supervised learning, NMF, proximal optimization.

# 1 Introduction

Cette communication est motivée par un problème fréquent en recherche clinique : des patients, stratifiés en groupes d'intérêt (typiquement sains / malades ou contrôles / traités) sont caractérisés par des mesures biologiques correspondant à plusieurs caractéristiques différentes (métabolomiques, protéomiques, etc). L'objectif est alors de découvrir des signatures moléculaires multi-omiques caractérisant les groupes. En particulier, les Laboratoires Pierre Fabre Dermo Cosmétique sont engagés dans de multiples projets de ce type dans lesquels des données omiques multiples ont été acquises dans le but de mieux comprendre des altérations de la peau. Ici, nous nous focalisons en particulier sur un projet lié à la dermatite atopique (DA) qui est une maladie inflammatoire commune, principalement caractérisée par une fonction barrière de la peau dysfonctionnelle. Plusieurs études multi-omiques portant sur des échantillons de peau humaine relevés sur zones lésionnelles ont pu caractériser la DA sur le plan protéomique et transcriptomique, or les analyses séparées de ces différentes omiques ne permettent pas de comprendre les relations gènes-protéines potentiellement instrumentales dans le développement de cette maladie.

Or, si les approches permettant l'intégration de données multiples, en particulier de données omiques multiples, se sont développées de manière importante ces dernières années (voir notamment [10, 9, 3] pour des revues sur ce sujet), un faible nombre sont destinées à une analyse exploratoire tenant compte de cette structure. Ici, nous abordons donc la question de l'intégration multi-omiques sous un angle mixte, celui de l'analyse exploratoire d'omiques multiples dans laquelle une information complémentaire caractérisant ces individus (attribut clinique ou expérimental par exemple) est d'intérêt pour la compréhension du phénomène biologique.

Dans cette communication, nous présentons une extension de la Factorisation Matricielle Non-négative (NMF) [6] et plus particulièrement de sa version supervisée [7] pour extraire un profil multi-omique de la DA. En effet, cette méthode de réduction de dimension offre un cadre bien adapté au problème d'intégration de données omiques pour des individus structurés en groupe. Également, elle est spécifiquement conçue pour l'analyse de données à valeurs positives, ce qui est le cadre naturel de nombreuses données omiques (données de comptages comme le RNA-seq ou les données métagénomiques, données compositionnelles comme en métabolomique ou protéomique, etc). En particulier, son interprétation est elle-même facilitée par la contrainte de positivité de la solution, la décomposition retenue s'expliquant aisément en termes de profils types et d'appartenance à ces profils de chacun des individus.

Dans la suite, nous introduisons le cadre général de la NMF, certaines variantes ainsi que l'approche que nous proposons, une version intégrative et supervisée de la NMF, dans la section 2. Enfin, dans la section 3, nous présentons les premiers résultats des expériences obtenues sur données simulées et sur les données du projet étudiant la dermatite atopique.

# 2 NMF supervisée intégrative

## 2.1 La NMF et ses variantes

Soit $\mathbf{X} \in \mathbb{R}_+^{n \times p}$ une matrice de données à entrées positives de grande dimension (typiquement $n \ll p$). Le but de la NMF est de fournir une approximation de faible rang de $\mathbf{X}$ telle que :

$$\mathbf{X} \simeq \mathbf{W} \times \mathbf{H},$$

avec $\mathbf{W} \in \mathbb{R}_+^{n \times K}$ et $\mathbf{H} \in \mathbb{R}_+^{K \times p}$, deux matrices à entrées positives respectivement appelées matrice des poids et matrice des signatures (ou composantes latentes). $K$ est le nombre de signatures (ou le rang de la décomposition) et est choisi par l'utilisateur.

[6] décrivent deux variantes de la NMF qui se fondent sur deux fonctions de coût distinctes dont le but est de minimiser l'erreur de l'approximation : la divergence de Kullback-Leibler et la norme de Fröbenius. Le choix de la fonction de coût dépend de la distribution statistique que l'on attribue au terme d'erreur et pour des distributions gaussiennes, la norme de Fröbenius est généralement indiquée :

$$\underset{\mathbf{W},\mathbf{H} \geq 0}{\arg\min} \frac{1}{2} \|\mathbf{X} - \mathbf{W}\mathbf{H}\|_F^2 .$$

Ainsi, la NMF est initialement une méthode non supervisée, conçue pour les analyses exploratoires. Toutefois, [8] ont développé plusieurs variantes de la NMF adaptées à la classification d'images MALDI[1], dont la NMF « FR-lda ». Cette variante intègre un terme supervisé qui assure que la décomposition obtenue (en particulier les signatures) est prédictive d'une structuration des individus en deux groupes, $\mathbf{y} \in \{0,1\}^n$ :

$$\underset{\mathbf{W},\mathbf{H},\boldsymbol{\beta} \geq 0}{\arg\min} \underbrace{\frac{1}{2} \|\mathbf{X} - \mathbf{W}\mathbf{H}\|_F^2 + \frac{\mu}{2} \|\mathbf{W}\|_F^2 + \lambda \|\mathbf{H}\|_1 + \frac{\nu}{2} \|\mathbf{H}\|_F^2 + \frac{\gamma}{2} \|\mathbf{y} - \mathbf{X}\mathbf{H}^\top \boldsymbol{\beta}\|_2^2}_{= \mathcal{F}_0(\mathbf{W},\mathbf{H},\boldsymbol{\beta})}$$

avec :

- $\mathbf{W} \in \mathbb{R}_+^{n \times K}$, la matrice des poids (contribution des individus aux composantes latentes) ;
- $\mathbf{H} \in \mathbb{R}_+^{K \times p}$, les signatures (composantes latentes) ;
- $\boldsymbol{\beta} \in \mathbb{R}_+^K$, les coefficients de régression ;

[1] Matrix Assisted Laser Desorption/Ionization

- $\mu$, $\lambda$, $\nu$, $\gamma > 0$, les paramètres de régularisation, fixés.

Outre le terme classique de perte de l'approximation et le terme supervisé correspondant à un critère de moindres carrés, $\|\mathbf{y} - \mathbf{X}\mathbf{H}^\top \boldsymbol{\beta}\|_2^2$, les termes de perte $\ell_2$ assurent la régularisation de la solution (ainsi que la définition d'une perte trivialement non identifiable) et le terme de perte $\ell_1$ assure la parcimonie des signatures obtenues.

## 2.2 Extension de la NMF FR-lda pour l'intégration de données

Dans le cadre multi-omique décrit dans l'introduction, nous considérons maintenant $\mathbf{X}^{(j)} \in \mathbb{R}_+^{n \times p_j}$ ($j \in \{1, 2\}$), deux matrices à entrées positives contenant les mesures de deux types d'omiques sur les mêmes $n$ individus ($p_j$ étant le nombre de variables mesurées dans chaque omique). On notera également $\mathbf{y} \in \{0, 1\}^n$ le vecteur d'appartenance des individus à deux groupes d'intérêt biologique.

Notre proposition consiste à étendre l'approche de [5] pour l'intégration de données en minimisant le critère :

$$\begin{aligned} \underset{\mathbf{W}, \mathbf{H}^{(1)}, \mathbf{H}^{(2)}, \boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}}{\arg\min} & \frac{1}{2} \left( \sum_{j=1}^{2} \|\mathbf{X}^{(j)} - \mathbf{W}\mathbf{H}^{(j)}\|_F^2 \right) + \frac{\gamma}{2} \left( \sum_{j=1}^{2} \|\mathbf{y} - \mathbf{X}^{(j)}\mathbf{H}^{(j)\top}\boldsymbol{\beta}^{(j)}\|_2^2 \right) \\ & + \sum_{j=1}^{2} \lambda \|\mathbf{H}^{(j)}\|_1 + \frac{\mu}{2} \|\mathbf{W}\|_F^2 \end{aligned} \tag{1}$$

avec :

- $\mathbf{W} \in \mathbb{R}_+^{n \times K}$, la matrice des poids, commune aux deux composantes latentes ;
- $\forall j \in \{1, 2\}$, $\mathbf{H}^{(j)} \in \mathbb{R}_+^{K \times p_j}$, les signatures, spécifiques de chaque omique ;
- $\forall j \in \{1, 2\}$, $\boldsymbol{\beta}^{(j)} \in \mathbb{R}_+^K$, les coefficients de régression ;
- $\gamma$, $\lambda$, $\mu > 0$, les paramètres de régularisation, fixés.

Les problèmes d'optimisation qui apparaissent dans la NMF sont des problèmes non-convexes et non-linéaires [5] car la fonction de perte n'est pas convexe en $\mathbf{W}$, $\mathbf{H}^{(1)}$, $\mathbf{H}^{(2)}$, $\boldsymbol{\beta}^{(1)}$ et $\boldsymbol{\beta}^{(2)}$ simultanément. Toutefois, la marginalisation en chacune de ces variables conduit à des problèmes d'optimisation convexes, deux de ces problèmes incluant une contrainte non lisse (la pénalité $\ell_1$). Les problèmes d'optimisation de type NMF sont donc généralement résolus par des approches itératives résolvant successivement chacun des problèmes marginaux en utilisant des méthodes de Majoration-Minimisation (MM). En outre, dans le cas particulier de la NMF, ce principe permet d'obtenir des étapes de mises à jour multiplicatives (MU) qui assurent la positivité des matrices obtenues à chaque itération. Toutefois, ces méthodes ne permettent qu'une parcimonie approchée. Alternativement, nous avons proposé une méthode permettant la parcimonie exacte des signatures en remplaçant l'étape MU par une étape d'optimisation proximale. De manière générale, le problème de l'équation (1) est résolu par l'algorithme 1.

**Algorithme 1** Vue d'ensemble de l'algorithme utilisé pour la résolution de l'équation (1)

1: Initialiser les matrices $\mathbf{W}^{(0)}$, $\mathbf{H}^{(j,0)}$ et vecteurs $\boldsymbol{\beta}^{(j,0)}$ avec des valeurs strictement positives ($\forall j \in \{1,2\}$).
2: **Pour tout** $t = 1, \ldots, T$ **Faire**
3: Mise à jour MU : $\mathbf{W}^{(t+1)} \leftarrow \mathbf{W}^{(t)} \odot \mathbf{A}(\mathbf{W}^{(t)})$
4: Mise à jour MU ou Prox : $\forall\, j = 1, 2$, (Prox)

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

OU (MU) :

$$\mathbf{H}^{(j,t+1)} \leftarrow \mathbf{H}^{(j,t)} \odot \mathbf{S}(\mathbf{H}^{(j,t)})$$

5: Mise à jour MU : $\forall\, j = 1, 2$, $\boldsymbol{\beta}^{(j,t+1)} \leftarrow \boldsymbol{\beta}^{(j,t)} \odot \mathbf{u}(\boldsymbol{\beta}^{(j,t)})$
6: **Fin pour**
7: **renvoyer** $\mathbf{W} := \mathbf{W}^{(T+1)}$, $\mathbf{H}^{(j)} := \mathbf{H}^{(j,T+1)}$ et $\boldsymbol{\beta}^{(j)} := \boldsymbol{\beta}^{(j,T+1)}$ $(j = 1, 2)$

$\odot$ est l'opérateur de multiplication terme à terme et les valeurs spécifiques de $\mathbf{A}(.)$, $\text{prox}_{\tilde{g}_j}(.)$, $f_j(.)$ $\mathbf{S}(.)$, et $\mathbf{u}(.)$ ont des formes explicites omises ici pour la clarté du propos.

# 3 Applications

## 3.1 Données simulées

L'approche que nous proposons ici a également été évaluée sur des données simulées. Nous avons utilisé le même processus de génération de données que celui décrit dans [12][2] et qui a été utilisé pour évaluer une approche de NMF intégrative non supervisée (iNMF) (voir aussi l'article de comparaison [2] qui utilise ces mêmes données).

En bref, le processus de génération fonctionne en deux temps, le premier correspondant à la génération de données $\mathbf{W}$ et $\mathbf{H}^{(j)}, \forall j \in \{1,2\}$ (ici $n = 50$, $p_1 = 2500$, $p_2 = 400$, $K = 2$) pour lesquelles des signatures typiques des deux groupes sont générées selon une loi $\mathcal{B}eta(2,2) \times 2$ (les autres variables, non informatives, ou les valeurs des variables de signatures pour le groupe complémentaire étant initialisées à la valeur 0). Enfin, dans un deuxième temps, les matrices $\mathbf{X}^{(j)}$ reconstruites à partir de $\mathbf{W}$ et $\mathbf{H}^{(j)}$ sont bruitées.

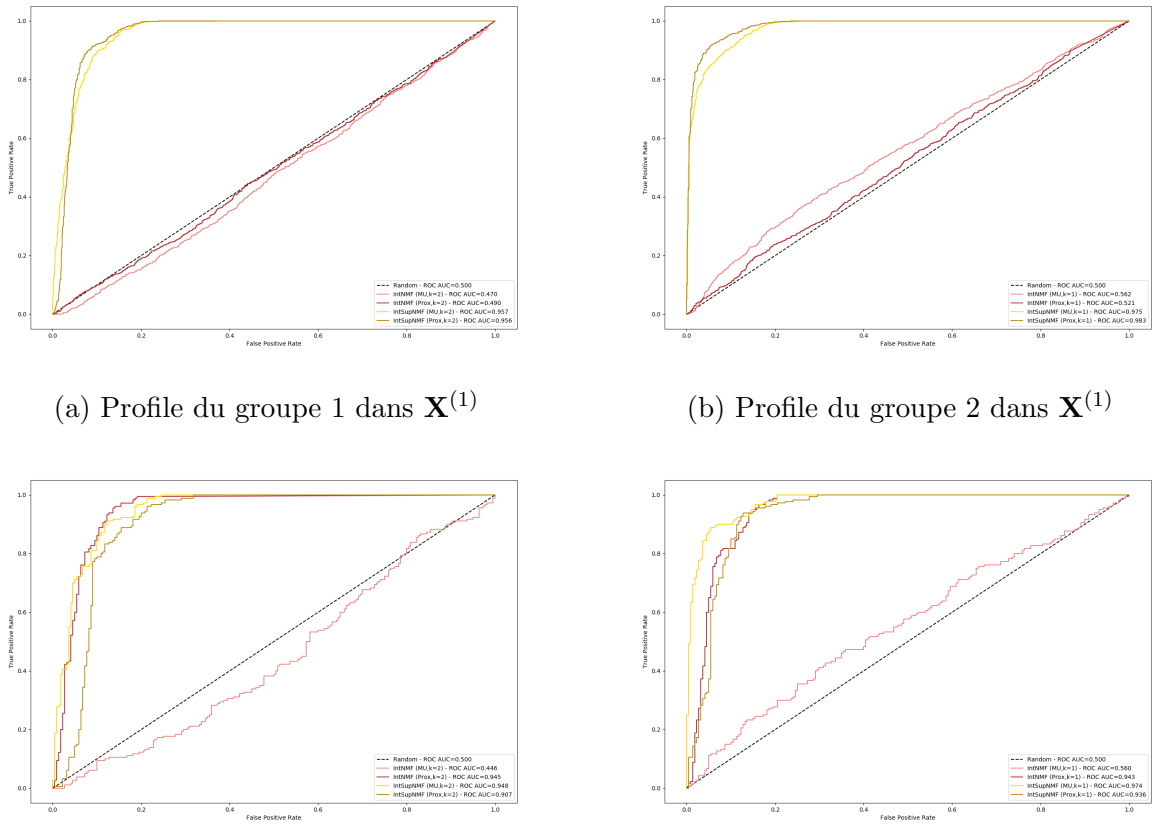
La flexibilité de ce modèle de génération de données nous a permis de tester la NMF intégrative supervisée sur différents aspects, comme la sensibilité au bruit dans les données ou au déséquilibre dans les groupes d'intérêt ou encore dans le nombre de variables caractérisant chacun des groupes. En particulier, nous avons constaté que :

- d'une manière générale, lorsque le niveau de bruit est modéré, la solution fournie par l'approche de résolution proximale extrait des signatures moléculaires directement parcimonieuses, discriminant les individus selon leur groupe d'appartenance. En revanche, lorsque le niveau de bruit dans la génération des données est plus fort, l'approche MU,

[2]Les scripts associés sont disponibles à l'adresse `https://github.com/yangzi4/iNMF/tree/master`.

qui ne fournit pas de parcimonie exacte dans les signatures extraites, est plus robuste et sélectionne correctement les variables expliquant les groupes ;

- la partie supervisée dans le terme de reconstruction permet d'améliorer considérablement la qualité des signatures retrouvées, comme illustré sur la Figure 1, et ce particulièrement lorsque le bruit augmente ou que le déséquilibre dans la taille des deux groupes devient important.




(a) Profile du groupe 1 dans $\mathbf{X}^{(1)}$

(b) Profile du groupe 2 dans $\mathbf{X}^{(1)}$

(c) Profile du groupe 1 in $\mathbf{X}^{(2)}$

(d) Profile du groupe 2 in $\mathbf{X}^{(2)}$

FIGURE 1 : Courbes ROC et scores AUC selon la version de la NMF et de l'approche d'optimisation utilisée

Par ailleurs, nous avons comparé notre approche à deux approches d'intégration de données très utilisées, DIABLO [11] (qui est une approche supervisée parcimonieuse basée sur l'analyse canonique des corrélations régularisée) et MOFA [1] qui est une approche non supervisée assez similaire à une Analyse Factorielle Multiple (MFA [4]). Les expériences sont encore en cours mais les premiers résultats semblent montrer que :

- comme la NMF supervisée, DIABLO extrait des signatures moléculaires parcimonieuses et permet bien d'identifier quelques-unes des variables explicatives des groupes dans celles-ci. Toutefois, les signatures obtenues sont généralement trop conservatrices ;

- MOFA n'extrait pas de signatures moléculaires parcimonieuses mais affecte correctement un poids plus important aux variables expliquant les deux groupes, donnant, de ce point de vue, des résultats assez similaires à notre approche.

### 3.2 Etude de la Dermatite Atopique (DA) aux niveaux protéomique et transcriptomique

La NMF supervisée intégrative a également été utilisée pour analyser des données transcriptomiques et protéomiques issues d'une étude sur la Dermatite Atopique (DA) sur peau non lésionnelle. Cette étude a été menée sur $n = 12$ sujets, divisés en deux groupes. Cinq d'entre eux étaient des sujets atteints de DA et les sept autres des sujets sains. Les données finales correspondent ainsi à des données transcriptomiques (issues de la technologie biopuces) contenant l'expression de $p_1 = 22\ 557$ gènes et des données protéomiques contenant la quantification de $p_2 = 281$ protéines.

Sur ces données, la méthode de résolution MU permet de bien discriminer les deux groupes. Les signatures moléculaires sont actuellement à l'étude pour savoir si des éléments connus pour être spécifiques de la DA sont retrouvés.

## 4 Conclusion

Nous avons décrit une approche d'intégration de données adaptée à un problème classique dans les études cliniques dans lesquelles les patients sont souvent classés en groupes d'intérêt biologique. Sur données réelles et simulées, la méthode permet de correctement extraire les signatures biologiques spécifiques des groupes. Les résultats sont en cours d'approfondissement, notamment pour affiner la comparaison avec d'autres approches d'intégration de données mais aussi pour approfondir l'interprétation biologique des signatures extraites pour la DA.

Enfin, notons que la méthode propose un cadre permettant une extension flexible de son utilisation : l'utilisation d'une divergence de Kullback-Leibler à la place de la norme de Fröbenius permettrait de l'adapter à des données fortement non gaussiennes et la modification du terme supervisé permettrait de l'exprimer comme un problème de régression logistique (plutôt que linéaire) ou de régression logistique multiple (pouvant s'adapter à plus de 2 groupes).

## Bibliographie

[1] Ricard Argelaguet, Britta Velten, Damien Arnol, Sascha Dietrich, Thorsten Zenz, John C. Marioni, Florian Buettner, Wolfgang Huber, and Oliver Stegle. Multi-Omics Factor Analysis-a framework for unsupervised integration of multi-omics data sets. *Molecular Systems Biology*, 14(6):e8124, 2018.

[2] Cécile Chauvel, Alexei Novoloaca, Pierre Veyre, Frédéric Reynier, and Jérémie Becker. Evaluation of integrative clustering methods for the analysis of multi-omics data. *Briefings in Bioinformatics*, 21(2):541–552, 2020.

[3] Tara Eicher, Garrett Kinnebrew, Andrew Patt, Kyle Spencer, Kevin Ying, Qin Ma, Raghu Machiraju, and Ewy A. Mathé. Metabolomics and multi-omics integration: a survey of computational methods and resources. *Metabolites*, 10(5):202, 2020.

[4] B. Escofier and J. Pagès. Multiple factor analysis (AFMULT package). *Computational Statistics and Data Analysis*, 18(1):121–140, 1994.

[5] Pascal Fernsel and Peter Maass. A survey on surrogate approaches to non-negative matrix factorization. *Vietnam Journal of Mathematics*, 46:987–1021, 2018.

[6] Daniel Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. *Advances in Neural Information Processing Systems (NIPS 2000)*, 13:556–562, 2001.

[7] Hyekyoung Lee and Seungjin Choi. Group nonnegative matrix factorization for EEG classification. In David van Dyk and Max Welling, editors, *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pages 320–327, Clearwater Beach, Florida, USA, 2009. PMLR.

[8] Johannes Leuschner, Maximilian Schmidt, Pascal Fernsel, Delf Lachmund, Tobias Boskamp, and Peter Maass. Supervised non-negative matrix factorization methods for MALDI imaging applications. *Bioinformatics*, 35:1940–1947, 2019.

[9] Chen Meng, Oana A. Zeleznik, Gerhard G. Thallinger, Bernhard Kuster, Amin M. Gholami, and Aedín C. Culhane. Dimension reduction techniques for the integrative analysis of multi-omics data. *Briefings in Bioinformatics*, 17(4):628–641, 2016.

[10] Marylyn D. Ritchie, Emily R. Holzinger, Ruowang Li, Sarah A. Pendergrass, and Dokyoon Kim. Methods of integrating data to uncover genotype-phenotype interactions. *Nature Reviews Genetics*, 16:85–97, 2015.

[11] Amrit Singh, Casey P. Shannon, Benoît Gautier, Florian Rohart, Michaël Vacher, Scott J. Tebbutt, and Kim-Anh Lê Cao. DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. *Bioinformatics (Oxford, England)*, 35:3055–3062, 2019.

[12] Zi Yang and George Michailidis. A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data. *Bioinformatics*, 32(1):1–8, 2016.