Université de Rennes

UFR SVE
Sciences de la Vie et de l'Environnement

UNIVERSITÉ DE RENNES

MASTER 2 BIO-INFORMATIQUE
Parcours Informatique pour la Biologie et la Santé
2023–2024

Juliette Francis

Exploration des méthodes d'intégration de données dans l'objectif de prédiction ou de classification de catégories d'efficience alimentaire

---

Encadrants : **Yann Le Cunff, Mahendra Mariadassou, Quentin Le Graverand**

IRISA, équipe Dyliss
263 Av. Général Leclerc, 35042 Rennes

UMR IRISA

DyLiSS

Abréviations

RFI	Residual feed intake (Consommation résiduelle d'aliments)
RFI-	Lignée efficace (sélectionnée pour une consommation résiduelle négative)
RFI+	Lignée inefficace (sélectionnée pour une consommation résiduelle positive)

Sommaire

1 Introduction

1.1 Contexte méthodologique

1.1.1 Le Machine Learning appliqué à la biologie

Le Machine Learning (ML), ou apprentissage automatique en français, est une branche de l'intelligence artificielle. Il consiste à développer des modèles basés sur les mathématiques, utilisant les informations issues de données complexes pour générer du savoir. Les modèles apprennent automatiquement à extraire les informations utiles et à déceler des schémas de relations entre elles. Ils servent ensuite à faire des prédictions, à prendre des décisions ou à analyser les données [1]. Les méthodes de ML sont très utilisées dans le domaine de la biologie. Les progrès technologiques en biologie, comme les technologies de séquençage haut débit (HTS), permettent maintenant d'étudier des processus biologiques complexes. Ces progrès ont induit une explosion de la quantité de données biologiques. De plus, ces données sont hétérogènes et requièrent des méthodes d'analyse de données adaptées. Le ML offre une solution à ce besoin [2]. Il existe différentes grandes approches d'apprentissage dont les usages dépendent des objectifs d'analyse et des données utilisées [1, 2, 3]. Deux de ces approches sont utilisées dans les travaux de ce stage, l'approche supervisée et l'approche non supervisée.
L'approche supervisée consiste à trouver un lien entre plusieurs variables p décrivant par exemple des individus n. Ces individus appartiennent à ce qu'on appelle le jeu d'entraînement. L'objectif est d'établir une relation entre un ensemble de variables d'entrée et une variable de sortie discrète ou continue. Trouver cette relation consiste à construire une fonction f telle que $Y = f(X) + \epsilon$, où X sont les variables d'entrée, Y est la variable de sortie et ϵ est l'erreur entre la véritable valeur de Y et la valeur prédite par la fonction f. Cette fonction est construite en minimisant une fonction de perte $Loss(Y, \hat{Y})$ qui permet d'obtenir la meilleure fonction f, c'est-à-dire celle qui fournit un ϵ le plus petit possible. Après entraînement du modèle, il est utilisé pour prédire ou estimer les valeurs de la variable Y des individus d'un nouveau de jeu de données. En biologie, les variables d'entrée peuvent être hétérogènes, comme des variables démographiques ou des expressions de gènes. La variable de sortie peut être quantitative (ex : taille des individus) ou catégorielle (ex : malade / sain).
L'approche non supervisée tente d'identifier des structures parmi les individus et les variables étudiées. Un modèle non supervisé n'est pas entraîné avec une variable de sortie. L'approche non supervisée permet par exemple de regrouper les individus selon leurs caractéristiques communes sans connaissances a priori [1, 2, 3].

1.1.2 Nécessité d'une approche intégrative pour l'analyse des données biologiques

Pour bien étudier les processus biologiques complexes, il est important de prendre en compte les différentes couches d'informations biologiques mises à disposition par les technologies à haut débit. Ces technologies permettent de générer plusieurs types de données, dont certaines appelées données omiques. Un terme comportant le suffixe -omique désigne l'étude d'un sous-ensemble spécifique d'informations biologiques. Parmi les différentes données omiques disponibles se trouvent les données génomiques, transcriptomiques, protéomiques, métabolomiques,

etc [4, 5]. Ces différentes omiques sont liées entre elles et font partie d'une cascade d'événements qui commence avec le génome, et continue avec le transcriptome, le protéome et le métabolome. Le phénotype est le résultat final de cette cascade [5]. Le phénotype est l'ensemble des caractéristiques observables ou mesurables d'un organisme. Il résulte à la fois de l'influence des facteurs -omiques (en premier lieu le génotype de l'individu) et de celle des facteurs environnementaux [6]. Comprendre un système biologique complexe dans sa globalité pourrait être facilité par une analyse conjointe des différentes données omiques, c'est-à-dire une intégration de données hétérogènes. L'objectif de l'intégration de données est de combiner différentes données pour en extraire plus d'informations que ne pourrait être obtenue à partir d'un seul type de données [7]. Combiner différents types de données représentant les différentes couches d'informations biologiques en tant que variables prédictives pourrait alors améliorer la prédiction de phénotypes. Prédire un phénotype à partir de l'intégration de données hétérogènes, notamment omiques, et identifier les meilleurs prédicteurs de ce phénotype sont des enjeux actuels de la biologie et de la médecine mais présentent plusieurs difficultés méthodologiques [8].

1.1.3 Challenges de l'intégration de données omiques

Les omiques sont des données hétérogènes et bruitées. Les données manquantes peuvent créer du bruit dans n'importe quel type de données. Il est difficile de déterminer si l'absence de signal a une signification biologique ou si elle provient d'erreurs de mesure, d'une couverture de séquençage trop faible ou d'un signal en dessous de la limite de détection. Le bruit peut également être dû à des biais dans la collecte des données (effets batch). Sélectionner efficacement les données et les variables informatives présente également des difficultés. Les variables peuvent en effet être colinéaires : de la redondance peut être présente au sein de chaque donnée omique ou entre elles [5, 7]. Ainsi, lors d'une analyse de données omiques ou multi-omiques, il est nécessaire de procéder à une étape de prétraitement qui comprend habituellement le filtrage et la normalisation des données, l'élimination des effets batch et les contrôles de qualité. Ces étapes nécessitent un choix adéquat des méthodes utilisées, ce qui représente une difficulté supplémentaire au vu du nombre de méthodes disponibles dans la littérature et de l'absence de standard universel [4].
Les données omiques sont généralement des données de grande dimension, c'est-à-dire qu'elles contiennent un grand nombre de variables. Le nombre de variables dépasse souvent largement le nombre d'observations, ce qui peut provoquer "The curse of dimensionality", ou fléau de la dimension. Plus le nombre de variables augmente, plus la probabilité de trouver une relation par pur chance entre plusieurs variables d'entrée et la variable de sortie augmente également. Si un modèle de classification utilise ces relations fortuites pour discriminer les observations en fonction de la variable de sortie, il aura des difficultés à généraliser ces prédictions à nouveau jeu de données test. Il s'agit du problème d"overfitting" ou sur-ajustement du modèle, dont une solution est de réduire le nombre de dimensions en faisant de la sélection ou de l'extraction de variables. La sélection de variables conserve seulement les variables les plus informatives et l'extraction de variables créé un ensemble plus petit de nouvelles variables en transformant celles initiales tout en essayant de perdre le moins d'informations possibles [5, 2].
Enfin, la nature hétérogène des données soulève plusieurs problèmes. Les jeux de données hétérogènes ont des tailles, des formats, des distributions, des types de variables (continues ou discrète)

et des dimensions différentes. Utiliser un modèle pour analyser conjointement des données hétérogènes nécessite une stratégie d'intégration qui combine efficacement les différentes variables des jeux de données [4, 5, 7].

1.1.4 Méthodes d'intégration de données

On distingue trois schémas d'intégration de données dont les utilisations dépendent de la nature des données. La P-integration intègre des observations issues de différentes études concernant les mêmes P variables d'un type de données. La N-integration intègre les observations de variables de plusieurs types de données concernant les mêmes N échantillons ou individus. La NP-integration intègre des observations issues de différentes études et de variables de plusieurs types [5, 9].

Il existe plusieurs stratégies d'intégration utilisant diverses manières de combiner les variables des différents jeu de données et à des moments différents de l'analyse :

- Early integration : concaténation des variables pour créer un jeu de données augmenté, qui est ensuite fournie au modèle.
- Mixed : transformation des variables séparément afin de leur trouver des représentations simplifiées qui sont ensuite intégrées et utilisées par le modèle.
- Intermediate : création d'une représentation commune aux différentes variables, qui est par la suite fournie au modèle.
- Late : application d'un modèle sur chaque type de variables puis agrégation des résultats des différents modèles [7].

1.2 Contexte biologique

L'objectif de ce stage est de prédire un phénotype à partir de données multi-omiques dans un contexte agroécologique, plus précisément dans le contexte de la filière ovine. Le genre ovin regroupe des mammifères herbivores ruminants, tels que les moutons et les mouflons. L'espèce *Ovis aries* (appelé couramment mouton) a été domestiquée entre 11000 et 9000 ans avant JC. Depuis cette période et jusqu'à aujourd'hui, l'élevage des différentes races de moutons permet de produire de la viande, du lait, de la laine et de la peau [10]. En agriculture, la reproduction sélective est utilisée afin de produire des animaux ayant des phénotypes désirées. Elle consiste à accoupler des individus qui pourront potentiellement transmettre le phénotype d'intérêt à leur descendance [11]. Un phénotype intéressant à utiliser en sélection d'animaux est l'efficience alimentaire. L'efficience alimentaire mesure la capacité d'un animal à transformer les ressources alimentaires qui lui sont fournis en produits animaux. Améliorer l'efficience alimentaire pourrait aider le secteur de l'élevage à devenir plus durable. Sélectionner des animaux selon l'efficience alimentaire nécessite de phénotyper un grand nombre d'animaux individuellement [12]. Cependant, à cause du coût élevé et des contraintes du phénotypage, l'efficience alimentaire n'est pas un critère très utilisée pour sélectionner les animaux d'élevage. Une solution à ce problème serait de prédire l'efficience alimentaire à partir de données omiques [13].

Le stage s'inscrit dans la continuité de la thèse de Quentin Le Graverand, maître de conférences en alimentation des ruminants à l'ENVT et co-encadrant du stage [5]. Les travaux de la thèse présentent les performances de prédiction de l'efficience alimentaire d'agneaux de plusieurs modèles linéaires à partir de données hétérogènes, dont des données omiques. Les objectifs du stage sont :

- Prédire l'efficience alimentaire d'agneaux à l'aide des méthodes linéaires de prédiction et d'intégration de données utilisées dans la thèse de Quentin Le Graverand.
- Comparer les performances de ces méthodes à celle de méthodes non linéaires.

2 Matériel et méthodes

2.1 L'efficience alimentaire

2.1.1 Residual Feed Intake (RFI)

L'efficience alimentaire est définie selon plusieurs critères dans la littérature [14]. Je me suis intéressée au critère RFI, "Residual Feed Intake" ou consommation résiduelle d'aliments. Le critère RFI explique l'efficience alimentaire comme étant la différence entre l'apport alimentaire réel et l'apport alimentaire prévu selon les besoins de l'animal. Pour estimer ce critère pour un individu dans la population étudiée, la régression suivante est utilisée :

$$ADFI = \mu + \beta_1 ADG + \beta_2 MD + \beta_3 BFT + \beta_4 finalBW^{0.75} + \textbf{RFI}$$

Où l'ADFI (Average Daily Feed Intake) est la consommation journalière moyenne d'aliments en g/jour d'un individu (moyenne calculée sur six semaines), μ est la valeur moyenne des ADFI de tous les individus de la population étudiée, ADG (Average Daily Gain) est le gain en masse corporelle quotidien moyen en g/jour, MD (Muscle Depth) est l'épaisseur de muscle en cm, BFT (Back Fat Thickness) est l'épaisseur de la graisse dorsale en mm, BW (Body Weight) est la masse corporelle en kg et $BW^{0.75}$ la masse métabolique. La RFI s'exprime en g/jour. Les besoins alimentaires d'un animal regroupent les besoins de maintenance et les besoins de production. Les besoins de maintenance sont les besoins en énergie utilisée pour la thermorégulation, le métabolisme basal et l'activité physique. Les besoins de production sont les besoins en aliments utilisés pour la croissance des tissus et peuvent être estimés à partir du gain de masses. $\beta_1, \beta_2, \beta_3$ sont les effets associés respectivement à ADG, MD et BFT, et tiennent compte des besoins de production. β_4 est l'effet associé à la masse métabolique, et tient compte des besoins de maintenance. Les animaux efficients au niveau alimentaire ont des valeurs de RFI négatives et les animaux non efficients ont des valeurs de RFI positives [5].

2.1.2 Les lignées RFI

Une lignée est l'ensemble des descendants d'un individu. Une sélection des moutons basée sur la RFI a été réalisée et a permis de créer deux lignées de moutons divergentes, la lignée RFI- et la lignée RFI+. La lignée RFI- regroupe des animaux sélectionnés pour une RFI négative (et

donc à forte efficience alimentaire) et la lignée RFI+ regroupe des animaux sélectionnés pour une RFI positive (et donc à faible efficience alimentaire). Il est nécessaire d'attendre plusieurs générations avant d'obtenir deux lignées bien divergentes [5].

2.2 Les données

Les données utilisées dans les travaux du stage ont été fournies par Quentin Le Graverand. Les données ont été collectées chez des moutons de race Romane, aussi appelé race INRA 401. Elles font partie d'une étude menée de 2018 à 2022 sur des agneaux mâles, à l'unité expérimentale P3R de l'INRAE. Une partie de ces animaux appartiennent à la lignée RFI- et une autre à la lignée RFI+. Un ensemble de pré-traitements usuels en mentionner quelques uns pour montrer que tu as une idée de ce qui a été fait a été appliqué aux données lors de la thèse de Quentin Le Graverand [5]. Ces données traitées constituent le point de départ de mon stage, à partir duquel j'ai réalisé les travaux présentés dans les sections suivantes.

2.2.1 Génotype

Le génotypage des SNP (Single Nucleotide Polymorphism) a été réalisé à partir d'échantillons de sang des moutons. Un SNP est une différence d'un seul nucléotide à un endroit spécifique du génome. Le génotypage consiste à établir le génotype d'un individu, c'est-à-dire la composition allélique de tous ses gènes. Un allèle est une forme alternative d'un gène par rapport au gène établi comme référence [15]. Le génotypage des SNP consiste à déterminer quelle version de chaque SNP étudié possède un individu. Les données de génotype contiennent les dosages alléliques de 30 957 SNPs pour 649 moutons. Les dosages alléliques peuvent être 0 (homozygote pour l'allèle alternatif), 1 (hétérozygote) ou 2 (homozygote pour l'allèle de référence) À vérifier mais le codage généralement utilisé est l'inverse : 0 pour l'allèle de référence et 1 pour l'allèle alternatif. Le dosage 0 signifie que le SNP étudié de l'individu est différent de celui de référence pour les deux chromosomes. Le dosage 1 signifie que le SNP étudié de l'individu est différent de celui de référence sur un seul chromosome. Le dosage 2 signifie que le SNP étudié de l'individu est le même que celui de référence pour les deux chromosomes.

2.2.2 Microbiote

L'étude du microbiote des agneaux se fait grâce aux abondances résiduelles d'OTUs (Operational Taxonomic Unit). Un OTU est un groupe d'organismes formé selon leur similarité d'une séquence d'ADN d'un gène donné, qui sert de marqueur taxonomique. Le marqueur utilisé ici est le gène codant pour l'ARN ribosomal 16S. L'ADN a été extrait, amplifié et séquencé à partir d'échantillons de fluides du rumen (le premier compartiment du système digestif des agneaux). Un clustering des lectures de séquençage a ensuite été réalisé afin de former les différents OTUs. Les données de microbiote décrivent les abondances 625 OTU de bactéries ou archées chez 442 animaux.

2.2.3 Zootechnie

Les données zootechniques décrivent le phénotype de 439 agneaux. Plusieurs variables sont disponibles dans ces données. J'ai utilisé pour mes travaux, l'identifiant des agneaux, l'année de phénotypage, la lignée RFI, et la RFI des agneaux.

A partir des données de génotype, de microbiote et de phénotype, deux jeux de données ont été créés. Certains individus étudiés ne sont pas décrits par les trois fichiers. Seuls les individus communs aux données de génotype, de microbiote et de phénotype ont été conservés pour les analyses, c'est-à-dire 425 agneaux. Les deux jeux de données correspondent d'une part aux dosages alléliques des SNP et d'autre part aux abondances des OTUs, chacun d'entre eux étant augmentés des informations d'identifiant, d'année de phénotypage, de lignée RFI et de la RFI quantitative des animaux.

2.2.4 Sanctuariser un jeu de test

Les travaux du stage consistent à explorer des méthodes pour prédire la RFI (régression) et la lignée RFI (classification). Afin d'évaluer les performances des modèles utilisés, les jeux de données de génotype et de microbiote ont chacun été séparés en un jeu d'entraînement et un jeu de test. Les jeux d'entraînement et de test conservent les proportions d'animaux RFI- et RFI+, ainsi que les proportions des années de phénotypage des jeux de données initiaux. Ils regroupent respectivement 80% et 20% des individus. La figure 1a met en évidence la stratification par lignée RFI et par année de phénotypage des jeux de données. Les jeux d'entraînement et de test ont également une distribution de la RFI similaire à celle du jeu de données initial (figure 1b).

(a). **(b)**.

Figure 1. Proportion des agneaux dans les jeux de données d'entraînement, de test, et initiaux selon la lignée RFI et l'année de phénotypage (a) et distribution des valeurs de RFI dans les différents jeux de données (b)

2.3 Prédire la lignée RFI et la RFI à partir d'un type de données

Des modèles de classification sont utilisés pour prédire la lignée RFI, et des modèles de régression sont utilisés pour prédire la RFI. Dans certains cas une réduction de dimensions est appliquée aux données en amont du modèle prédictif. Toutes ces méthodes se divisent en deux catégories : linéaires et non-linéaires. Dans les sections suivantes, un jeu de données est représenté par une matrice X, de taille n individus (en lignes) et p variables (SNP ou abondances résiduelles d'OTU, en colonnes). Le vecteur $x_i = (x_{i1}, x_{i2}, ..., x_{ip})$ représente les observations de l'individu i pour les p variables et y_i est la valeur de Y de l'individu i.

2.3.1 Méthodes linéaires

Régression linéaire

Une régression linéaire considère que la relation entre les p variables de X et la réponse Y est linéaire. Un modèle de régression linéaire est de la forme suivante :

$$E[Y_i|x_i] = f(x_i) = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j$$

Où les coefficients du modèle $\beta = (\beta_0, \ldots, \beta_p)$ sont inconnus.
L'objectif de la régression est d'estimer au mieux les β_j afin de minimiser une fonction de perte correspondant à la somme des résidus au carré (residual sum of squares, RSS) entre la sortie Y (continue) et la sortie prédite $\hat{Y}$:

$$RSS(\beta) = \sum_{i=1}^{n}(y_i - f(x_i))^2 = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ji}\beta_j\right)^2$$

Cette fonction de perte permet de mesurer l'erreur de prédiction du modèle. Le problème de minimisation de la fonction de perte est résolu par un algorithme d'optimisation, ou dans certains cas via une formule exacte pour la valeur optimale de β.
Avec les données étudiées, $p >> n$, en particulier pour les données de génotype (environ 30 957 SNP). Cela a pour conséquence de créer des modèles complexes sujets au sur-apprentissage. Un moyen de lutter contre ce phénomène est d'appliquer une pénalité sur la fonction de perte du modèle. Nous utiliserons deux pénalité dans ces travaux, la pénalité L1 aussi appelée Lasso, et la pénalité L2 aussi appelée Ridge. Ces pénalités poussent la régression à trouver un équilibre entre faire une bonne prédiction et limiter le sur-apprentissage. Elles appliquent toutes deux une contrainte sur la taille des coefficients de la régression.
Avec la pénalité L1, le problème de minimisation de la fonction de perte devient :

$$\min_{\beta} \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ji}\beta_j\right)^2 + \boldsymbol{\lambda \sum_{j=1}^{p} |\beta_j|}$$

avec $\lambda \geq 0$, le paramètre de régularisation. Plus λ est grand, plus le modèle met à zéro les coefficients des variables qui contribuent le moins à la somme des résidus. Les variables les plus informatives sont ainsi sélectionnées et la complexité du modèle diminue.
Avec la pénalité L2, le problème de minimisation de la fonction de perte devient :

$$\min_{\beta} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \boldsymbol{\lambda \sum_{j=1}^{p} \beta_j^2}$$

avec $\lambda \geq 0$ le paramètre de régularisation. De manière similaire à la pénalité L1, plus λ est grand, plus les coefficients des variables peu informatives sont faibles et la complexité du modèle diminue [16]. Le paramètre λ a été fixé à 1 pour toutes les analyses du stage.

Un modèle de régression linéaire avec pénalité L2 est utilisé pour prédire la RFI et est implémenté grâce à la fonction `Ridge` Utiliser une fonte monospace pour les noms de fonctions et de logiciels, par exemple avec \texttt du package python Scikit-Learn [1]. Pour prédire la lignée RFI, un modèle de régression linéaire avec pénalité L2 est également utilisé. Prédire la lignée RFI est un problème de classification, néanmoins la fonction `RidgeClassifier` de Scikit-Learn permet de résoudre le problème comme un problème de régression, en transformant les valeurs de lignée RFI (RFI- ou RFI+) en -1 et 1 au préalable. Une pénalité L2 est aussi utilisée sur cette classification.

Analyse en composantes principale

L'analyse en composantes principales, ou Principal Component Analysis (PCA), est une méthode linéaire non supervisée de réduction de dimensions. Son objectif est de créer un ensemble de variables latentes, les composantes principales, qui conservent le maximum de variance observée dans les variables initiales. Les variables latentes sont des combinaisons linéaires orthogonales des variables initiales et sont déterminées de manière à être indépendantes les unes des autres. Après PCA, les données sont représentées par z variables latentes avec $z << p$, on dit que les données sont projetées [17]. Dans les travaux de ce stage, plusieurs PCA sont réalisées à l'aide de la fonction PCA de Scikit-Learn. Ces nouvelles représentations des données sont ensuite utilisées pour prédire la lignée RFI ou la RFI.

Partial Least Squares

La technique Partial Least Squares (PLS), aussi appelée Projection to Latent Structures, est une méthode linéaire de régression. Elle construit comme la PCA, un ensemble de combinaisons linéaires des variables initiales. Cependant cette méthode est supervisée, car elle utilise la variable de sortie y pour guider la projection. La PLS cherche à créer des variables latentes conservant la variance des variables initiales et ayant forte covariance avec la sortie y. Ensuite les variables latentes sont utilisées pour prédire y avec une régression linéaire [16]. Le package R `mixOmics` [2]

1. `https://scikit-learn.org/stable/api/index.html`
2. `http://mixomics.org/`

propose un ensemble de fonctions permettant d'utiliser la méthode PLS pour des problèmes de régression et également de classification [9]. La méthode sPLSDA (sparse Partial Least Squares Discriminant Analysis) est utilisée pour prédire la lignée RFI et la méthode sPLSR (sparse Partial Least Squares Regression) est utilisée pour prédire la RFI. Le terme "sparse" désigne la sélection de variables effectuée à l'aide d'une pénalité L1 appliquée sur les coefficients des combinaisons linéaires formant les variables latentes.

2.3.2 Méthodes non linéaires

Régression logistique

Un régression logistique applique une fonction logistique sur la sortie d'une régression linéaire afin d'obtenir des valeurs de sorties comprises entre 0 et 1. La fonction logistique standard tend vers 0 en $-\infty$ et vers 1 en $+\infty$. Elle est définie de la façon suivante :

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

Un modèle de régression logistique est donc de la forme :

$$H(x_i) = \frac{1}{1 + e^{-f(x_i)}} \quad \text{avec} \quad f(x_i) = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j$$

Lorsque la variable qualitative à prédire est binaire, les deux labels possibles prennent la valeur 0 et 1 (ici RFI- devient 0 et RFI+ devient 1). La régression logistique modélise les probabilités d'appartenir à la classe 1 (RFI+) des individus. Ainsi les probabilités que l'individu x_i soit RFI+ ou RFI- correspondent respectivement à :

$$P(Y_i = 1|x_i) = H(x_i)$$

$$P(Y_i = 0|x_i) = 1 - H(x_i)$$

L'objectif de la régression logistique est d'estimer au mieux β afin de minimiser la fonction de perte suivante, appelée entropie croisée binaire (BCE) :

$$BCE(\beta) = \sum_{i=1}^{n} -y_i \, log(H(x_i)) - (1 - y_i) \, log(1 - H(x_i))$$

Si un individu appartient à la classe 0, le terme $-y_i log(H(x_i)$ est égale à 0. La fonction est alors égale à $log(1 - H(x_i))$. Ainsi, si le modèle prédit 0 ($H(x_i) = 0$), la fonction est égale à 0 ($log(1) = 0$), et le modèle ne fait pas d'erreurs. Si le modèle prédit une valeur plus grande que 0 ($H(x_i) > 0$), alors $log(1 - H(x_i)) \neq 1$ et la fonction augmente [18].
Un modèle de régression logistique avec pénalité L2 est utilisé pour prédire la lignée RFI et est implémenté grâce à la fonction `LogisticRegression` de Scikit-Learn.

Réseau de neurones

Les réseaux de neurones artificiels sont des modèles de régression non linéaires ayant une structure inspirée des réseaux de neurones biologiques. Ils peuvent être utilisés pour prédire des variables qualitatives ou quantitatives. Dans les travaux du stage, un réseau de neurones est utilisé pour prédire la RFI (quantitative). Un réseau de neurones constitué d'un seul neurone est une fonction de la forme :

$$\hat{y}_i = \Phi\left(\sum_{j=1}^{p} x_{ij}\omega_j + \omega_0\right)$$

Où $\hat{y}_i$ est la valeur de sortie prédite par le neurone pour l'individu i, x_{ij} est le vecteur d'observations de l'individu i pour les p variables, Φ est une fonction d'activation (généralement non linéaire), $(\omega_j)_{j=1...n}$ et ω_0 sont les poids et le biais du neurone. Il existe plusieurs choix pour la fonction d'activation. Les réseaux de neurones utilisés pour prédire la RFI appliquent la fonction d'activation ReLU (Rectified Linear Unit) qui est définie par $f(x) = \max(0, x)$.

Un réseau de neurone peut posséder un nombre de couches variable et chacune de ces couches peut également contenir un nombre variable de neurones. Dans notre cas de prédiction de la RFI, tous les neurones d'une couches sont connectés aux neurones de la couche précédente et de la couche suivante. Les neurones de la première couche d'entrée prennent en entrée les valeurs du vecteur x_i. Les neurones des couches suivantes prennent en entrée les sorties des neurones des couches qui les précèdent. La figure 2 montre en détail le passage de la première couche à la suivante.

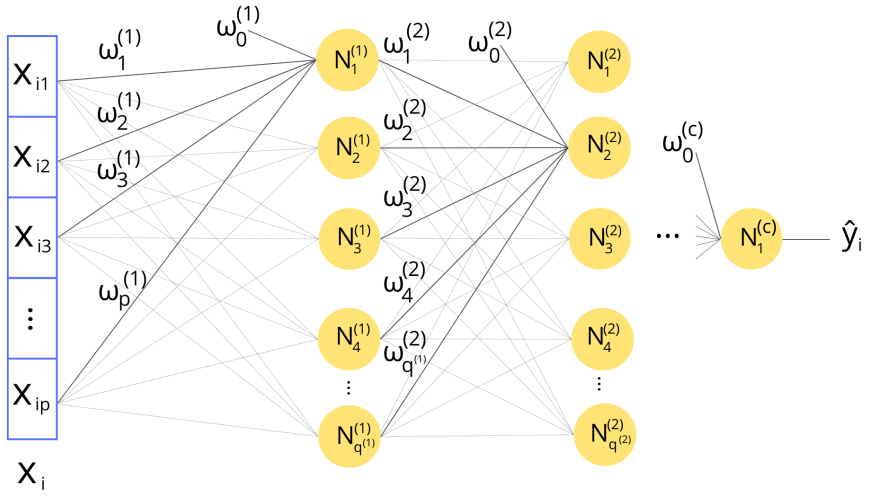

Figure 2. Schéma d'un réseau de neurone utilisé pour prédire une variable $\hat{y}_i$ à partir d'un vecteur x_i.

La sortie d'un neurone $N^{(c)}$ de la couche c est égale à l'activation de la somme des sorties des neurones de la couche $c-1$ multipliés par les poids de la couche c, plus le biais de la couche c.

L'apprentissage du réseau de neurone consiste à déterminer les poids et les biais des neurones qui minimisent une fonction de perte entre Y et $\hat{Y}$. La fonction de perte utilisée pour prédire la

RFI avec des réseaux des neurones est l'erreur quadratique moyenne (mean squared error, MSE) entre le vecteur y_i et la prédiction $\hat{y_i}$:

$$MSE(\omega) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2$$

L'apprentissage du modèle est la répétition de deux phases successives, la phase de propagation vers l'avant et la phase de rétropropagation. La propagation vers l'avant est le calcul de $\hat{y_i}$ grâce aux calculs, couches après couches, de toutes les sorties de neurones. À partir de $\hat{y_i}$, la fonction de perte est déterminée. Ensuite grâce à un algorithme d'optimisation, les poids et les biais des neurones sont ajustés, de la dernière couche jusqu'à la première, de manière à minimiser la fonction de perte. Une fois les paramètres ajustés, il y a de nouveau une propagation vers l'avant et donc une nouvelle valeur de perte, et ainsi de suite. Le nombre de répétition de ces deux étapes peut être variable, une époque désigne une répétition [19]. L'apprentissage est terminé lorsque que toutes les époques sont réalisées. Pour prédire la RFI, une classe `NeuralNetwork` est implémentée à l'aide du package python PyTorch [3].

Auto-Encodeur Variationnel

Un auto-encodeur est un réseau de neurones artificiels utilisé pour faire de la réduction de dimensions. Il est constitué de deux parties, l'encodeur $e(x_i)$, qui projette les données dans un espace latent de plus petites dimensions, et le décodeur $d(e(x_i))$, qui tente de reconstruire les données initiales à partir de la projection. L'entraînement d'un auto-encoder consiste à minimiser la MSE entre X et la reconstruction $\hat{X}$:

$$MSE(\omega) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x_i})^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{p}\sum_{j=1}^{p}(x_{ij} - d(e(x_{ij})))^2\right)$$

Ainsi, un auto-encoder permet de trouver une projection des données qui conserve au maximum les informations initiales. A partir de l'espace latent créé par l'auto-encoder, un processus génératif peut-être réalisé. N'importe quel point de l'espace latent peut être reconstruit en un vecteur $\hat{x}$ avec le décodeur, même s'il ne faisait pas partie des données fournies en entrée. Ce processus permet de générer de nouvelles données. Néanmoins l'espace latent créé par l'auto-encoder peut être irrégulier, car l'auto-encoder tente d'obtenir la meilleure reconstruction, sans contrainte sur la cohérence de l'espace latent. C'est-à-dire que des points de l'espace latent proches peuvent donner des données reconstruites très éloignées, ou certains points peuvent être reconstruits en données qui n'ont pas de sens par rapport aux données initiales. Un auto-encodeur variationnel (VAE) est un auto-encodeur ayant une méthode de projection particulière qui permet de résoudre le problème d'irrégularité de l'espace latent. L'encoder du VAE encode le vecteur x_i en une distribution de paramètres μ_i et σ_i. A partir de cette distribution latente, un vecteur z_i est échantillonné. Le vecteur z_i contient les variables latentes de la projection. Le décodeur utilise z_i pour reconstruire x_i. La figure 3 montre le cas d'un VAE ayant deux variables latentes.

3. `https://pytorch.org/`

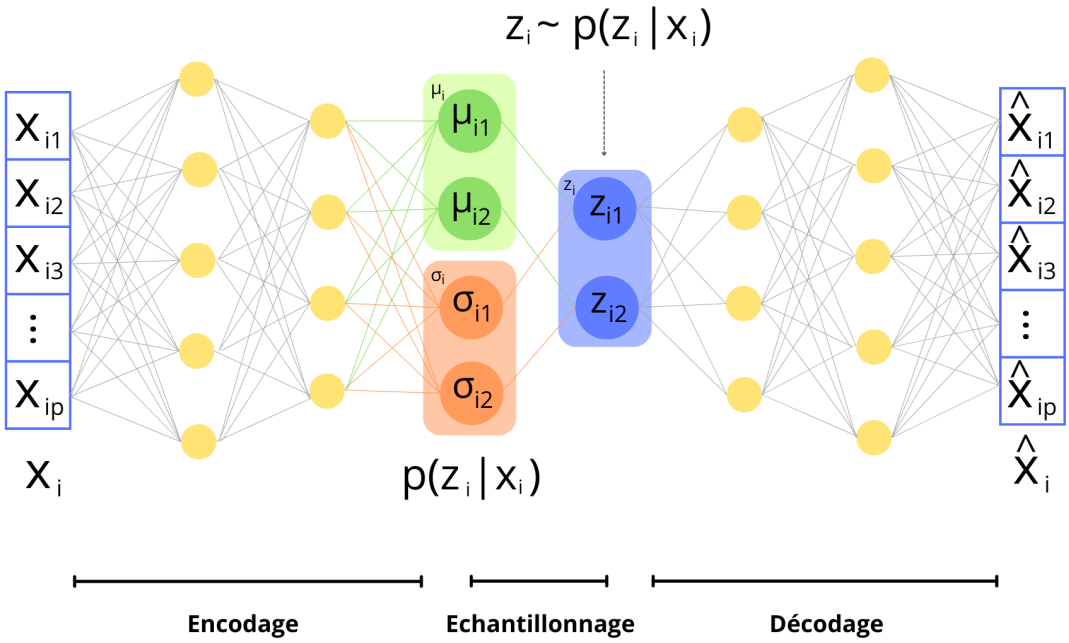

Figure 3. Schéma d'un auto-encoder variationnel ayant un espace latent composé de deux variables latentes.

Les paramètres μ_i et σ_i de la distribution encodée sont forcés à être similaires aux paramètres d'une distribution suivant une loi normale N(0, I), où I est une matrice d'identité. Cette contrainte pousse le modèle à encoder les données selon des distributions similaires et ainsi créer une continuité de l'espace latent. Pour appliquer cette contrainte, la fonction de perte du VAE est celle d'un auto-encoder classique à laquelle est ajoutée un terme de régularisation (en rouge ci-dessous) appelé divergence de Kullback-Leibler :

$$\min_{\omega} \frac{1}{n} \sum_{i=1}^{n} \left(\boldsymbol{D_{KL}[N(\mu_i, \sigma_i), N(0, I)]} + \frac{1}{p} \sum_{j=1}^{p} (x_{ij} - \hat{x}_i)^2 \right)$$

La divergence de Kullback-Leibler mesure la différence entre deux distributions. Le VAE cherche un compromis entre obtenir une bonne reconstruction des données et construire un espace latent ayant une distribution proche d'une distribution normale [20].
Pour projeter les données de génotype et de micobiote, une classe `VariationalAutoencoder` est implémentée à l'aide de PyTorch. Les variables latentes sont ensuite utilisées pour prédire la lignée RFI ou la RFI.

2.4 Prédire la lignée RFI et la RFI avec l'intégration de données

2.4.1 DIABLO

La méthode DIABLO (Data Integration Analysis for Biomarker discovery using Latent variable approaches for Omics studies) Citer la ref, est proposée par `mixOmics` et permet de faire de

la N-integration. Cette méthode linéaire réalise une classification en prenant en entrée plusieurs blocs de données. DIABLO cherche à créer des variables latentes pour chaque bloc de données, qui maximisent la covariance entre les variables des blocs, y compris avec la variable de sortie Y qui est considérée comme un des blocs. En d'autres termes, DIABLO tente de capturer l'information partagée entre les différents types de données et avec la variable à prédire. Ensuite, les variables latentes de chaque bloc sont utilisées pour prédire Y et les prédictions des différents blocs sont agrégées pour créer la prédiction finale [9]. La méthode DIABLO est utilisée pour prédire la lignée RFI.

2.4.2 Concaténation

La concaténation consiste à projeter les données de génotype et de microbiote afin de créer un nouveau jeu de données constitué de leurs variables latentes (figure 4). Les projections peuvent être réalisées avec une PCA ou avec un VAE. Les jeux de données obtenus avec la concaténation des projections de PCA sont utilisées pour prédire la lignée RFI ou la RFI à l'aide d'une régression linéaire. Ceux obtenus avec la concaténation des projections de VAE sont utilisés pour prédire la lignée RFI ou la RFI à l'aide d'une régression logistique et d'un réseau de neurones.

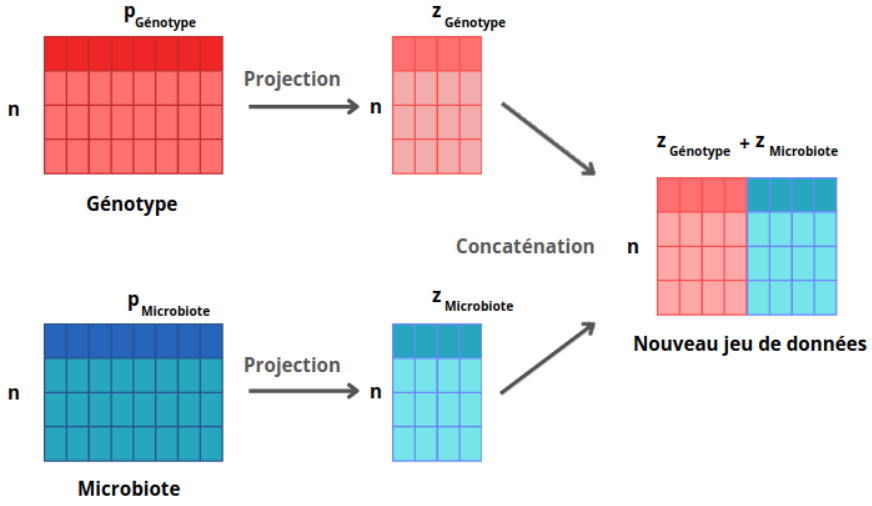

Figure 4. Schéma de la concaténation des variables latentes obtenues par projection des données de génotype et de microbiote.

2.5 Entraînement et évaluation des modèles

2.5.1 Critères d'évaluation

AUC

L'AUC (Area Under The Curve) est une métrique permettant d'évaluer les performances d'un modèle de classification binaire. Il indique la capacité du modèle à classer correctement

des échantillons dans les deux classes possibles. Lorsqu'un modèle classe des échantillons selon deux catégories, 0 ou 1, il y a plusieurs possibilités :

- Les échantillons 0 sont classés comme appartenant à la classe 0 (vrais négatifs (TN)).
- Les échantillons 0 sont classés comme appartenant à la classe 1 (faux positifs (FP)).
- Les échantillons 1 sont classés comme appartenant à la classe 1 (vrais positifs (TP)).
- Les échantillons 1 sont classés comme appartenant à la classe 0 (faux négatifs (FN)).

Ce serait plus lisible avec un tableau je pense

Les valeurs TN, FP, TP et FN permettent de tracer la courbe ROC (Receiver operating characteristic). Cette courbe représente le taux de vrai positif (TPR) en fonction de taux de faux positif (FPR) :

$$TPR = \frac{TP}{TP + FN} \qquad FPR = \frac{FP}{FP + TN}$$

L'AUC est l'aire sous la courbe ROC. Si l'AUC est égale à 1, TPR est égale à 1 et FPR est égale à 0, le modèle classe parfaitement les échantillons. Si l'AUC est égale à 0, TPR est égale à 0 et FPR est égale à 1, le modèle classe les échantillons dans leur classe opposée. Lorsque l'AUC est à 0.5, le modèle n'a aucune capacité à discriminer les échantillons selon leur classe. En dessous d'un AUC de 0.5, le modèle à de moins bonnes performances qu'un modèle qui classe aléatoirement les échantillons. Les AUC des modèles de classification seront calculées afin de comparer les performances [16].

MSE

Pour évaluer les performances des modèles de régression, l'erreur quadratique moyenne (MSE) est calculée :

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Plus la MSE est proche de 0, plus le modèle est performant. Plus la MSE est élevée, plus les erreurs de prédictions du modèles sont grandes.

TODO : R^2

2.5.2 Validation croisée

Avant d'être évalués sur le jeu de test, les modèles sont entraînés et validés sur le jeu d'entraînement avec de la validation croisée K-fold répétée 10 fois. Le jeu d'entraînement est divisé en K parties, appelées fold, de tailles à peu près égales et stratifiées selon la lignée et l'année de phénotypage. Le modèle est entraîné sur K-1 folds et est évalué sur le fold restant suivant les métriques précédemment décrites. Tous les folds servent de fold test une fois. Le processus est ensuite répété, 10 fois au total, avec des nouvelles répatitions des individus dans les folds. Les performances obtenues sur chaque fold sont ensuite combinées. Cette étape permet d'obtenir la variabilité des performances sur le jeu d'entraînement du modèle et de tester sa capacité à généraliser ses performances sur des données avec lesquelles il ne s'est pas entraîné [16]. Une fois validés, les modèles sont évalués à l'aide du jeu de test précédemment sanctuarisé.

2.6 Récapitulatif

La table 1 résume les analyses réalisées sur le génotype et le microbiote afin de prédire la lignée RFI.

	Méthode linéaire	**Méthode non linéaire**
Sans projection	Régression linéaire classifiante[1]	Régression logistique[1]
Avec projection	PCA puis régression linéaire classifiante[1]	VAE puis régression logistique[1]
Avec projection guidée par la lignée RFI	sPLSDA[2]	VAELR[3] puis régression logistique[1]

Table 1. Résumé des analyses réalisées pour prédire la lignée RFI (qualitative).

[1] Pénalité L2 (Ridge) sur les coefficients du modèle
[2] Pénalité L1 (Lasso) sur les coefficients du modèle
[3] Le modèle VAELR a été développé pendant le stage et sera présenté dans la partie résultats 3.1.

La table 2 résument les analyses réalisées sur le génotype et le microbiote afin de prédire la RFI.

	Méthode linéaire	**Méthode non linéaire**
Sans projection	Régression linéaire[1]	Réseau de neurones
Avec projection	PCA puis régression linéaire[1]	VAE puis réseau de neurones
Avec projection guidée par la RFI	sPLSR[2]	VAER[3] puis réseau de neurones

Table 2. Résumé des analyses réalisées pour prédire la RFI (quantitative).

[1] Pénalité L2 (Ridge) sur les coefficients du modèle
[2] Pénalité L1 (Lasso) sur les coefficients du modèle
[3] Le modèle VAER a été développé pendant le stage et sera présenté dans la partie résultats 3.1.

Les méthodes d'intégration de données utilisées pour prédire la lignée RFI et la RFI à partir des données de génotypes et de microbiote sont DIABLO et la concaténation des projections de

PCA et de VAE.

Les hyperparamètres des modèles, comme le nombre de couches cachées des auto-encoders variationnel ou encore le poids λ de la pénalité L2, sont identiques entre ceux utilisés pour le génotype et ceux utilisés pour le microbiote. Les différents hyperparamètres ont été fixés, soit en laissant les valeurs par défaut des fonctions utilisées, soit en regardant dans la littérature ce qui se fait habituellementce serait bien ici de mettre un lien vers un dépôt github ou un notebook qui donne les détails des modèles qui ont été ajustés. Le notebook permet de montrer le code sans fournir les données. Ces hyperparamètres peuvent être optimisés pour améliorer encore plus les performances des modèles, néanmoins les différentes valeurs possibles n'ont pas été explorées dans les travaux du stage.

3 Résultats

Les analyses réalisées ont pour objectifs de :

- Comparer les performances de modèles linéaires et non linéaires concernant la prédiction de la lignée RFI (qualitative), avec :
  - Les données de génotype,
  - Les données de microbiote,
  - L'intégration des deux types de données.
- Voir si les données permettent d'obtenir des performances de prédiction similaires pour la prédiction de la RFI (quantitative), en comparant encore une fois plusieurs modèles linéaires et non linéaires.

Parmi les méthodes non-linéaires, j'ai implémenté une nouvelle façon de faire des projections prenant en compte la sortie Y, nommée VAELR dans le cas d'une sortie qualitative, et VAER dans le cas d'une sortie quantitative. Je commence d'abord par présenter ces 2 méthodes dans la section 3.1, puis je présente les résultats de prédiction de la lignée RFI section 3.2 et les résultats de prédiction de la RFI section 3.3.

3.1 Projection par auto-encoder variationnel guidée par Y

Comme expliqué dans la section 2.3.1 partie Partial Least Squares, les modèles de PLSDA et PLSR sont des méthodes permettant de prédire respectivement, une variable qualitative et une variable quantitative, à partir de la projection des données. Cette projection est guidée par Y. Ces deux modèles sont des modèles linéaires et nous avons réfléchis à implémenter un VAE qui, sur le même principe que la PLS, oriente sa projection en fonction de la sortie Y. Cela nous permet d'avoir une sorte d'équivalent non-linéaire de la PLS. Pour la classification, la régression logistique peut être implémentée en pytorch de la même manière qu'un réseau de neurone. Le modèle de régression logistique est alors composé d'une couche linéaire et d'une couche d'activation utilisant une fonction sigmoïde. J'ai donc implémenté une nouvelle classe qui assemble un VAE classique et une régression logistique, appelée VAELR (Variational Auto-Encoder Logistic Regression). Comme pour le VAE classique, l'entraînement du modèle vise à trouver les poids et les

biais des neurones qui permettent de minimiser une fonction de perte. Pour le VAELR la fonction de perte est :

$$\frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{-y_i log(\hat{y}_i) - (1-y_i) log(1-\hat{y}_i)} + D_{KL}[N(\mu_i,\sigma_i), N(0,I)] + \frac{1}{p}\sum_{j=1}^{p}(x_{ij}-\hat{x}_{ij})^2\right)$$

Où y_i est la valeur de Y pour l'individu i, $\hat{y}_i$ est la prédiction de la régression logistique, x_{ij} est la valeur de la variable j pour l'individu i, $\hat{x}_{ij}$ est la reconstruction de x_{ij} à partir de la projection du VAELR. Le terme écrit en rouge représentent l'erreur de prédiction de la régression logistique. Les paramètres à optimiser sont les poids et biais de la régression logistique et du VAE qui composent le VAELR.
Le cas de la régression est similaire. J'ai implémenté une classe qui assemble un VAE classique et un réseau de neurones, appelée VAER (Variational Auto-Encoder Regression). Pour le VAER la fonction de perte est :

$$\frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{(y_i-\hat{y}_i)^2} + D_{KL}[N(\mu_i,\sigma_i), N(0,I)] + \frac{1}{p}\sum_{j=1}^{p}(x_{ij}-\hat{x}_{ij})^2\right).$$

Où y_i est la valeur de Y pour l'individu i, $\hat{y}_i$ est la prédiction du réseau de neurones, x_{ij} est la valeur de la variable j pour l'individu i, $\hat{x}_{ij}$ est la reconstruction de x_{ij} à partir de la projection du VAER. Le terme écrit en rouge représentent l'erreur de prédiction du réseau de neurones. Les paramètres à optimiser sont les poids et biais du réseau de neurones et du VAE qui composent le VAER. Ainsi les modèles VAELR et VAER cherchent à construire un espace de représentation condensés qui permet à la fois de bien reconstruire les données et de minimiser l'erreur de prédiction. Comme pour les autres projections non linéaires, les variables latentes sont récupérées afin d'être soumises aux modèles de régression logistique et de réseau de neurones.

3.2 Prédiction de la lignée RFI

J'ai commencé par étudier les performances de prédiction de la lignée RFI de plusieurs modèles. Je me suis d'abord demandée si le génotype puis le microbiote permettent de prédire la lignée, et ensuite si l'intégration des deux permet d'obtenir de meilleures performances. Les dimensions de projection des modèles de réduction de dimensions ont été déterminées en fixant un seuil de variance à conserver lors d'une analyse en composantes principales. Ce seuil est de 60%. Il est nécessaire de conserver 44 composantes pour les données de microbiote et 90 composantes pour les données de génotype afin d'obtenir au moins 60 % de variance expliquée (annexe). Comme évoqué précédemment, nous voulions des modèles de même taille et ayant les mêmes paramètres entre les modèles utilisés pour le génotype et ceux utilisés pour le microbiote. Nous avons donc choisis de conserver 90 variables latentes dans les projections des données de génotype et de microbiote. Cependant, pour les projections guidées par la lignée RFI, le nombre de variables latentes a été fixé à deux. Cela s'explique par le fait qu'au delà de deux composantes, une des étapes nécessaires aux modèles de sPLSDA prend plus de temps d'exécution sans voir

les performances s'améliorer (test avec 2 composantes : AUC = 0.99, temps = 3 min et test avec 10 composantes : AUC = 0.99, temps = 25 min). Ainsi pour pouvoir comparer leurs performances avec celles des modèles de VAELR, les dimensions de projection de ces derniers ont également été fixées à deux variables latentes. Cela permet de comparer aisément les méthodes linéaires et non-linéaires mais rend la comparaison entre les méthodes de projection avec et sans guide plus difficile.

Pour l'entraînement et la validation des modèles de sPLSDA, la validation croisée est stratifiée seulement selon la lignée RFI. La raison est que la fonction perf qui évalue les performances du modèle sPLSDA stratifie automatiquement les folds de la validation croisée selon la variable à prédire. D'après la documentation, cette fonction accepte une liste de vecteurs contenant les indices définissant chaque folds [4]. Néanmoins, il semble que la fonction n'autorise pas cet usage. J'ai pensé à une correction possible en regardant le code source mais sa mise en oeuvre aurait demandé du temps.

Les performances de prédiction à partir des données de génotype et de microbiote sont présentées dans le table 3 et seront abordées dans les section 3.2.1 et 3.2.2.

Méthodes d'analyse	**Génotype**		**Microbiote**	
	AUC	**Temps (s)**	**AUC**	**Temps (s)**
Régression linéaire classifiante	0.99	0.27	0.51	0.20
Régression logistique	0.99	1.40	0.57	0.14
PCA puis régression linéaire classifiante	**1.00**	**4.05**	**0.59**	**2.49**
VAE puis régression logistique	0.99	60.4	0.49	6.31
sPLSDA	0.99	1.89	0.48	0.16
VAELR puis régression logistique	1.00	72.9	0.58	6.76

Table 3. Résultats des prédictions de la lignée RFI à partir du génotype ou du microbiote.

Le génotype semble être un bon prédicteur de la lignée RFI au contraire du microbiote. La méthode qui propose la meilleure performance de prédiction de la lignée RFI à partir du génotype ou du microbiote est la régression linéaire classifiante précédée d'une projection par PCA. Pour tenter de comprendre les différences et les similarités de performances entre les méthodes, les projections des variables utilisées pour la prédiction sont étudiées dans les sections suivantes. Ensuite, les performances avec l'intégration des deux types de données sont présentées. Même si le génotype semble être suffisant pour prédire la lignée, nous voulions trouver des méthodes d'intégration permettant d'obtenir des performances aussi élevées afin de les utiliser par la suite pour la prédiction de la RFI.

3.2.1 Génotypes

Le génotype est utilisé pour prédire la lignée RFI. La figure 5 montre les espaces de projection des différentes analyses.

4. `https://www.rdocumentation.org/packages/mixOmics/versions/6.3.2/topics/perf`

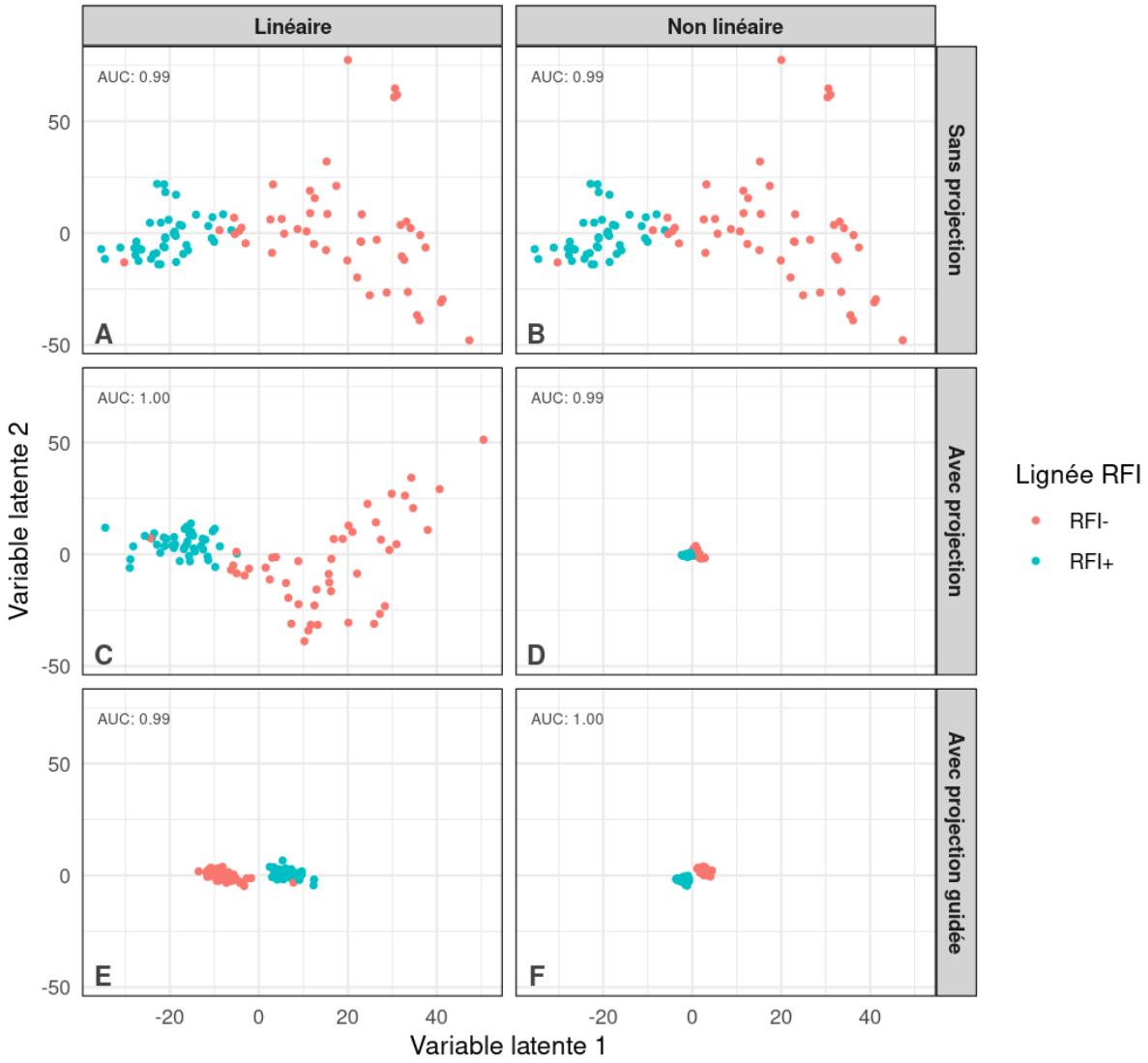

Figure 5. Projections et résultats de la classification effectuées sur le jeu de test de génotypage SNP des agneaux pour six méthodes différentes (voir table 1).
A : PCA réalisée sur le jeu de tests de génotype et résultats de la régression linéaire classifiante.
B : PCA réalisée sur le jeu de tests de génotype et résultats de la régression logistique.
C : PCA réalisée sur l'espace latent créé par projection PCA du jeu de tests de génotype.
D : PCA réalisée sur l'espace latent créé par projection VAE du jeu de test de génotype.
E : Espace latent de la sPLSDA réalisée sur le jeu des tests de génotype.
F : Espace latent du VAELR réalisé sur le jeu de tests de génotype.
C montre les résultats de la régression linéaire classifiante à partir de l'espace latent créé par projection PCA. E montre les résultats de la classification sPLSDA. D et F montrent les résultats de la régression logistique à partir des espaces latents créés par projection VAE et VAELR.

Sur chaque graphique, les individus forment deux clusters verticalement séparables, un cluster d'individus RFI- et un cluster d'individus RFI+. Le cluster d'individus RFI- est situé à droite sur tous les graphiques, exceptés sur le graphique E où il est situé à gauche, et inversement pour

le cluster d'individus RFI+. Une plus grande variance des variables latentes est observée sur les graphiques A, B (sans projection) et C (projection par PCA) comparée à celle observée sur les graphiques D (projection par VAE), E (sPLSDA) et F (VAELR). Les graphiques D et F montrent que les projections par VAE construisent des espaces latents beaucoup plus denses que ceux obtenus avec les autres méthodes de projection. Un individu RFI- semble être toujours projeté dans la même zone que les individus RFI+. D'après les projections montrées par ces graphiques et les résultats d'AUC (élevés) des modèles de classification correspondant, le génotype semble être un bon prédicteur de la lignée RFI. Les performances de prédiction sont similaires entre les méthodes linéaires ou non linéaires, et avec projection ou sans projection. Les temps d'exécution des programmes sont relativement petit et similaires entre toutes les méthodes (table 3). La méthode qui propose la meilleure performance de prédiction de la lignée RFI à partir des données génomiques, avec le temps d'exécution le plus petit, est la régression linéaire classifiante précédée d'une projection par PCA.

3.2.2 Microbiote

Le microbiote est utilisé pour prédire la lignée RFI. La figure 6 montre les espaces de projection des différentes analyses. Chaque graphique montre un nuage de points constitué d'individus RFI- et d'individus RFI+. Les individus des deux lignées ne sont pas visuellement séparables. Une plus grande variance des variables latentes est observées sur les graphiques A, B (sans projection) et C (projection par PCA) comparée à celle observée sur les graphiques D, (projection par VAE), E (sPLSDA) et F (VAELR). Le graphique D (projection par VAE) présente un espace latent beaucoup plus dense que ceux présents sur les autres graphiques.

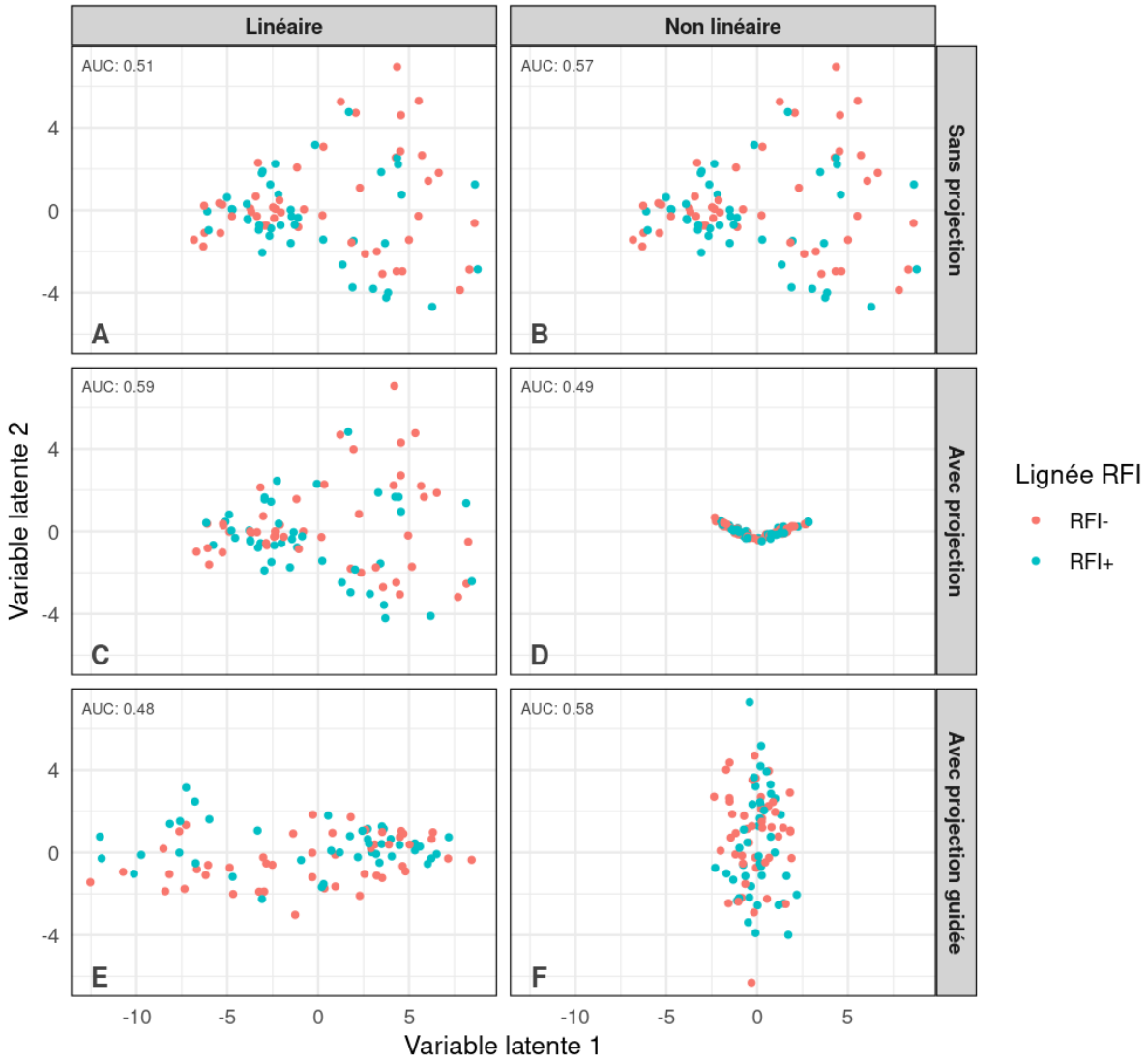

Figure 6. Projections et résultats de la classification effectués sur le jeu de test de microbiote ruminal des agneaux pour six méthodes différentes (voir table 2).
A : PCA réalisée sur le jeu de tests de microbiote et résultats de la régression linéaire classifiante.
B : PCA réalisée sur le jeu de tests de microbiote et résultats de la régression logistique.
C : PCA réalisée sur l'espace latent créé par projection PCA du jeu de tests de microbiote.
D : PCA réalisée sur l'espace latent créé par projection VAE du jeu test de microbiote.
E : Espace latent de la sPLSDA réalisé sur le jeu des tests de microbiote.
F : Espace latent du VAELR réalisé sur le jeu de tests de microbiote.
C montre les résultats de la régression linéaire classifiante à partir de l'espace latent créé par projection PCA. E montre les résultats de la classification sPLSDA. D et F montrent les résultats de la régression logistique à partir des espaces latents créés par projection VAE et VAELR.

D'après les projections montrées par ces graphiques et les résultats d'AUC (faibles) des modèles de classification correspondant, le microbiote ne semble pas être un bon prédicteur de la lignée RFI. L'exécution des programmes est très rapide pour toutes les méthodes (table 3). La

méthode qui propose la meilleure performance de prédiction de la lignée RFI à partir des données de microbiote, avec le temps d'exécution le plus petit, est la régression linéaire classifiante précédée d'une projection par PCA.

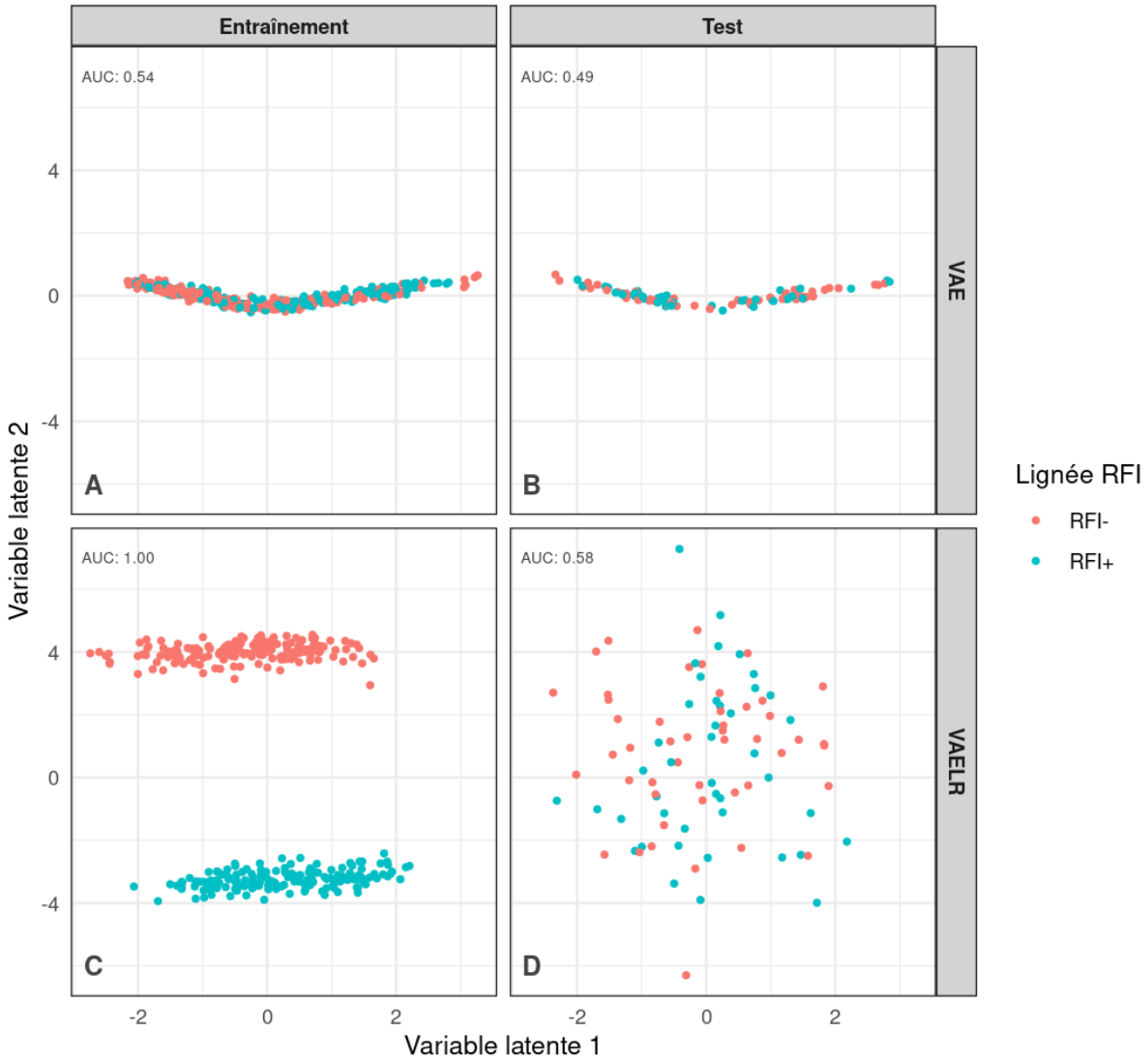

Figure 7. Plots of projections and classification results performed on lamb ruminal microbiota train and test set for VAE and VAELR à mettre en français.
A : PCA réalisée sur l'espace latent créé par projection VAE du jeu d'entraînement de microbiote.
B : PCA réalisée sur l'espace latent créé par projection VAE du jeu de test de microbiote.
C : Espace latent du VAELR réalisé sur le jeu d'entraînement de microbiote.
D : Espace latent du VAELR réalisé sur le jeu de tests de microbiote.
A, B, C and D montre les résultats de la régression logistique réalisée sur les espaces latents

La figure 7 montre les projections des espaces latents obtenus avec l'auto-encoder variationnel et l'auto-encoder variationnel combiné à une régression logistique pour le jeu d'entraînement et le

jeu de test des données de microbiote. On remarque que les espaces latents du VAE (graphique A et B) sont plus denses que ceux du VAELR (graphique C et D). Les espaces latents (entraînement et test) du VAE sont similaires sur leur répartition des individus RFI- et RFI+, alors que celles du VAELR différent. Pour le VAELR, les individus des données d'entraînement sont très bien discriminés selon la lignée RFI, ils sont linéairement séparables et l'AUC est égale à 1.00. Cette performance n'est pas retrouvée sur le jeu de test, l'AUC est égale à 0.58. Le modèle de VAELR fait probablement du sur-apprentissage, mais semble mieux réussir la classification que le VAE sur le jeu de test dont l'AUC est de 0.49. Cela montre que la structure "naturelle" des données, identifiée par le VAE, n'est que très peu associée avec la RFI et que celle identifiée par le VAELR, bien que plus propice à la prédiction, se généralise quand même mal. Attention, tu reprends encore la légende des figures dans le texte (cf commentaire de Quentin.

3.2.3 Intégration du génotype et du microbiote

Les données de génotype et de microbiote sont utilisées ensemble pour prédire la lignée RFI.s Les résultats de classification sont présentés dans la table 4.

Méthodes d'analyse	AUC	Temps (s)
DIABLO	0.99	2.62
PCA puis concaténation puis régression linéaire classifiante	**1.00**	**6.80**
VAE puis concaténation puis régression logistique	0.99	67.8
VAELR puis concaténation puis régression logistique	0.99	87.5

Table 4. Résultats des prédictions de la lignée RFI à partir de l'intégration du génotype et du microbiote.

Encore une fois, les performances de prédiction sont similaires entre les méthodes linéaires ou non linéaires. Les temps d'exécution sont néanmoins plus long avec les méthodes non linéaires. La méthode qui propose la meilleure performance de prédiction de la lignée RFI à partir de l'intégration des données de génotype et de microbiote, avec le temps d'exécution le plus petit, est la régression linéaire classifiante précédée d'une concaténation des projections par PCA. L'intégration des deux types de données permet de bien prédire la lignée. Au vu des résultats présentés précédemment, c'est le génotype qui importe le plus pour obtenir ces performances.

3.3 Prédiction de la RFI

3.3.1 Génotype et microbiote

J'ai ensuite étudié les performances de prédiction de la RFI avec des modèles de régression. Les dimensions de projection des modèles de réduction de dimensions sont les mêmes que celles des modèles utilisés pour la partie classification (90 variables latentes pour la PCA et le VAE, 2 variables latentes pour la sPLSR et le VAER). Les performances de prédiction des modèles à partir des données de génotype et de microbiote sont présentées dans le table 5.

Méthodes d'analyse	Génotype		Microbiote	
	MSE	Temps (s)	MSE	Temps (s)
Régression linéaire	0.22	0.23	0.30	0.18
Réseau de neurones	0.18	19.3	**0.20**	5.60
PCA puis régression linéaire	**0.17**	3.53	0.23	4.02
VAE puis réseau de neurones	0.19	87.8	0.23	10.6
sPLSDR	0.19	1.77	0.23	0.07
VAER puis réseau de neurones	0.19	88.0	0.24	11.0

Table 5. Résultats des prédictions de la RFI à partir du génotype ou du microbiote.

La prédiction à partir du génotype permet d'obtenir une MSE plus petite que celle obtenue avec la prédiction à partir du microbiote. L'écart de performance entre les deux types de données ne semble pas être aussi élevé que celui constaté lors de la classification. Les performances de prédiction sont similaires entre les méthodes linéaires ou non linéaires, et avec projection ou sans projection. La méthode qui propose la meilleure performance de prédiction de la RFI à partir des données de génotype est la régression linéaire précédée d'une projection par PCA. Pour le microbiote, la méthode qui propose la meilleure performance de prédiction de la RFI est le réseau de neurones.

3.3.2 Intégration du génotype et du microbiote

Les données de génotype et de microbiote sont utilisées ensemble pour prédire la RFI. Les résultats de régression sont présentés dans la table 6.

Méthodes d'analyse	MSE	Temps (s)
PCA puis concaténation puis régression linéaire	0.23	8.57
VAE puis concaténation puis réseau de neurones	**0.18**	108
VAER puis concaténation puis réseau de neurones	0.26	130

Table 6. Résultats des prédictions de la RFI à partir de l'intégration du génotype et du microbiote.

Encore une fois, les performances de prédiction sont similaires entre les méthodes linéaires ou non linéaires, mais les méthodes non linéaires ont des temps d'exécution plus long que celui de la méthode linéaire. La méthode qui propose la meilleure performance de prédiction de la lignée RFI à partir de l'intégration des données de génotype et de microbiote, avec le temps d'exécution le plus petit, est le réseau de neurones précédé d'une concaténation des projections par VAE. Les prédictions obtenues avec l'intégration des deux types de données ne sont pas meilleures à celles obtenues avec le génotype seul. Néanmoins il y a une amélioration entre les prédictions à partir des projections de VAE séparément et avec l'intégration des projections (0.19 et 0.23 pour génotype et microbiote seuls, 0.18 pour l'intégration des deux). Il faut remettre ces valeurs en contexte en les comparant aux RFI à prédire : grosse ou petite erreur ? Tu peux par exemple caluler l'erreur relative : MSE / valeur moyenne de Y.

La MSE permet de comparer les performances des modèles mais étant une moyenne des résidus, elle est très sensible aux valeurs extrêmes et difficilement interprétable. Ainsi nous avons étudié le R^2 des modèles linéaires de la RFI prédite en fonction de la RFI. La figure 8 montre

les modèles linéaires pour les valeurs de RFI prédites obtenue avec un réseau de neurones sur les projections de VAE. Il y a une amélioration du R^2 dans le cas de l'intégration du génotype et du microbiote. De nombreuses valeurs de RFI prédites avec l'intégration des données ont des valeurs similaires, aux alentours de 0. Il est difficile de savoir si l'amélioration de la MSE et du R^2 pour l'intégration n'est pas due à une saturation des valeurs de RFI prédites et non un apport des données de microbiote dans la prédiction. Une hypothèse possible expliquant ces valeurs serait un entraînement trop court du réseau de neurone et que l'algorithme d'optimisation n'a pas encore convergé.

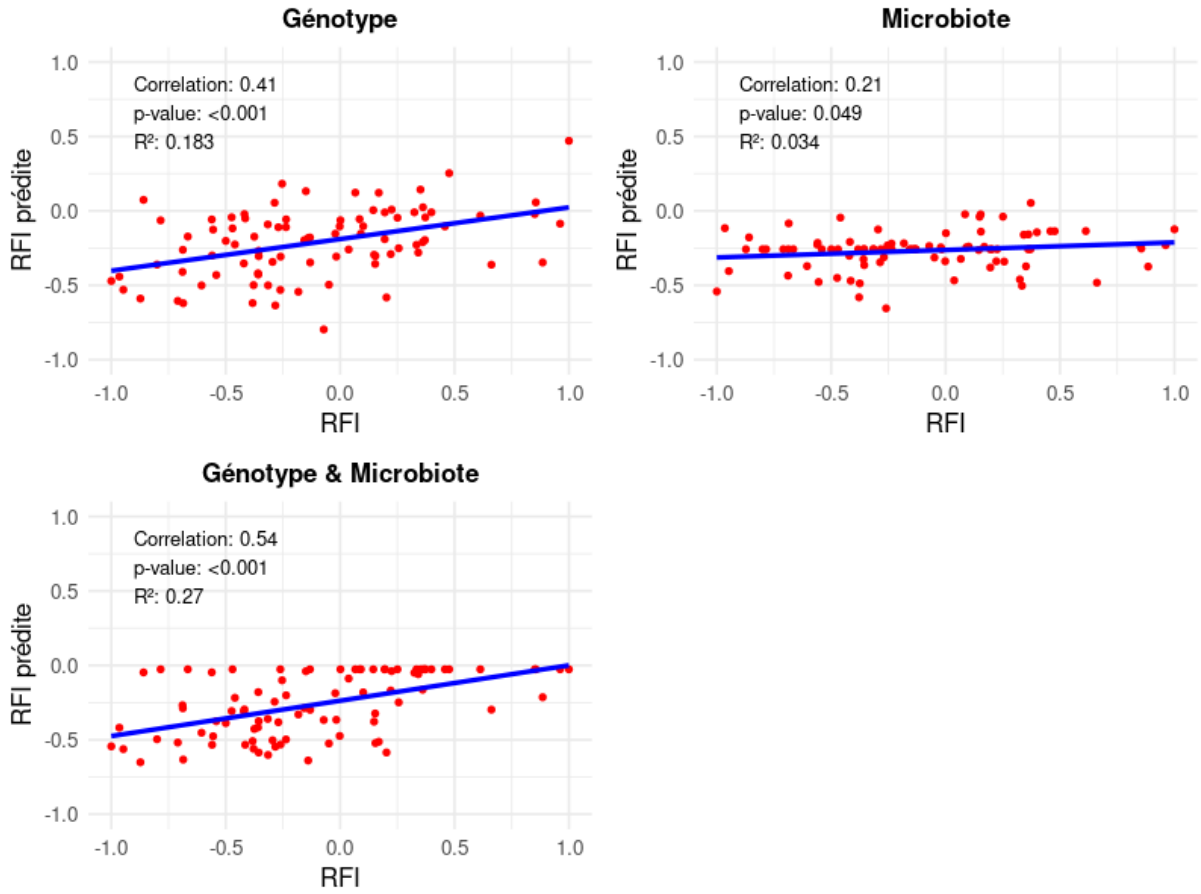

Figure 8. Régression linéaire de la RFI prédite à partir des projections de VAE en fonction de la RFI. Les valeurs de corrélations et de p-value sont les résultats d'un test de corrélation de Spearman réalisé entre la RFI prédite et la RFI. La RFI prédite est obtenue par prédiction d'un réseau de neurones à partir du génotype, du microbiote, et de l'intégration des deux. Tu as mis la droite de régresion sur la figure, c'est utile mais pour les graphes observé versus prédit, il faut surtout mettre la droite $y = x$ pour voir à quel point les prédictions sont loin de la réalité et si on on fait des erreurs systématiques dans la prédiction. Il y a par exemple ici un biais à prédire de faibles valeurs de RFI comparativement à la réalité.

4 Discussion

Les résultats montrent que les individus des deux lignées se distinguent très bien grâce à leur génotype. Ce résultat était attendu car les deux populations sont issues d'une sélection divergente, ce qui a pour effet d'accentuer les différences génétiques entre les lignées divergentes [5]. Les modèles de classifications linéaires ou non linéaires, avec ou sans projection, sont performants pour prédire la lignée de la population étudiée à partir du génotype (AUC moyen = 0.99, table 3). Les résultats montrent également que les individus des deux lignées ne sont pas discriminables grâce à leur microbiote du rumen (AUC moyen = 0.54, table 3). La sélection divergente pour l'efficience alimentaire n'est pas reflétée par les données de microbiote du rumen. Les performances de prédictions de lignée avec le génotype ou le microbiote sont en adéquation avec les résultats de la thèse de Quentin Le Graverand. Dans les travaux de la thèse, des modèles de sPLSDA ont été utilisés pour prédire la lignée RFI. Prédire à partir du génotype a permis d'obtenir un BER (Balanced error rate) moyen égale à 0.00. Prédire à partir du microbiote donne un BER moyen de 0.46. Le BER est un autre critère d'évaluation des performances de classification. Il est déterminé avec le calcul suivant : $BER = 0.5 \times (FP/(TN + FP) + FN/(FN + TP))$ [5].

Utiliser des modèles non linéaires n'a pas permis d'augmenter les performances des modèles. Néanmoins, le modèle VAELR a montré un bon potentiel de prédiction de la lignée à partir du microbiote sur le jeu d'entraînement (AUC=1.00). Pour envisager d'utiliser ce modèle, il est nécessaire de diminuer le sur-apprentissage. Une solution pourrait être d'étudier les valeurs des hyperparamètres du modèle. De manière générale, les performances des prédictions à partir du microbiote peuvent potentiellement être améliorées avec une recherche des hyperparamètres optimaux pour chaque modèle.

Les modèles d'intégration de données linéaires ou non linéaires distinguent très bien les individus RFI- des individus RFI+. Une performance aussi élevée que celle obtenue avec le génotype seul était souhaitée avant d'envisager d'utiliser les mêmes modèles pour prédire la RFI.

Pour le problème de classification de la lignée RFI, la régression linéaire classifiante des projections de PCA, ou de la concaténation des projections de PCA, donne les meilleurs résultats avec un temps d'exécution plus petit que les autres modèles, ou un temps d'exécution relativement court.

Les résultats de prédiction de la RFI montrent que le génotype prédit plus efficacement que le microbiote. Les méthodes linéaires ou non linéaires, et avec projection ou sans projection ont des performances proches pour la prédiction de la RFI. La méthode qui propose la meilleure performance de prédiction de la RFI à partir des données de génotype est la régression linéaire précédée d'une projection par PCA. Pour le microbiote, c'est le réseau de neurones. L'intégration des données permet d'améliorer les prédictions dans le cas de la concaténation des projections de VAE par rapport aux projections seules. Néanmoins la concaténation comme méthode d'intégration n'est peut-être pas adaptée pour prédire la RFI. Les deux types de données n'ont pas la même capacité prédictive et la concaténation réalisée ne permet pas de jouer sur la contribution des deux modalités. D'autres méthodes d'intégration existent et pourraient montrer de meilleures capacités prédictives, comme l'alignement des espaces latents des VAE. Il serait également pertinent de réaliser des analyses perturbatrices sur les prédictions afin de déterminer les SNP et les OTU expliquant le mieux les variations de RFI. Perturber les entrées des VAELR et VAER,

dont les projections sont directement liées au phénotype, pourrait permettre d'identifier les facteurs de génotype ou de microbiote qui influencent le phénotype. Enfin, il y a une saturation des prédictions de la RFI à partir de la concaténation des projections de VAE dont l'origine n'est pas certaine. Résoudre le problème de saturation permettrait de s'assurer de l'amélioration des prédictions obtenues avec l'intégration des données.

5 Conclusion

Améliorer l'efficience alimentaire des moutons permettrait de rendre la filière ovine plus durable. Sélectionner les animaux pour leur efficience alimentaire nécessitent néanmoins de connaître leur phénotypes dont les moyens de mesures coûtent cher. Une solution à ce problème serait de prédire l'efficience alimentaire des moutons à l'aide de modèle de machine learning. Lors de ce stage plusieurs modèles de machine learning linéaire et non linéaire ont été utilisés pour prédire la lignée RFI des animaux et leur RFI à partir des données de génotypes et des données de microbiote. Le problème de prédiction de la lignée RFI est très bien résolue par le génotype mais la lignée RFI ne semble pas impacter les données de microbiote du rumen. Le modèle le plus performant pour la lignée RFI est la régression linéaire classifiante appliquée aux projections ou à la concaténation des projections obtenues par PCA des données de génotype et de microbiote. Le génotype est également plus performant que le microbiote pour prédire la RFI, même si l'intégration des deux types de données avec les projections de VAE semble montrer un potentiel d'amélioration des prédictions. Des analyses supplémentaires sont cependant nécessaires pour conclure. D'autres pistes de méthode d'intégration de données sont à envisager afin de prendre en compte les différences de capacité à prédire la RFI du génotype et du microbiote. Les modèles VAELR et VAER implémentés lors du stage pourront être utilisés pour tenter d'identifier les SNPs et OTUs influençant le plus la lignée RFI et la RFI.

Bibliographie

[1] Gavin Edwards. *Machine Learning | An Introduction*. en. https://towardsdatascience.com/machine-learning-an-introduction-23b84d51e6d0. Jan. 2020.

[2] Chunming Xu et Scott A. Jackson. "Machine learning and complex biological data". In : *Genome Biology* 20.1 (avr. 2019), p. 76. issn : 1474-760X. doi : 10.1186/s13059-019-1689-0. url : https://doi.org/10.1186/s13059-019-1689-0.

[3] Konstantina Kourou et al. "Machine learning applications in cancer prognosis and prediction". In : *Computational and Structural Biotechnology Journal* 13 (jan. 2015), p. 8-17. issn : 2001-0370. doi : 10.1016/j.csbj.2014.11.005. url : https://www.sciencedirect.com/science/article/pii/S2001037014000464.

[4] Indhupriya Subramanian et al. "Multi-omics Data Integration, Interpretation, and Its Application". en. In : *Bioinformatics and Biology Insights* 14 (jan. 2020), p. 117793221989905. issn : 1177-9322, 1177-9322. doi : 10.1177/1177932219899051. url : http://journals.sagepub.com/doi/10.1177/1177932219899051.

[5] Quentin Le Graverand. "Integrating heterogeneous data to predict lamb feed efficiency". en. Thèse de doct. Institut National Polytechnique de Toulouse - INPT, sept. 2023. url : https://theses.hal.science/tel-04309422.

[6] Mary K. Wojczynski et Hemant K. Tiwari. "Definition of Phenotype". In : *Genetic Dissection of Complex Traits*. T. 60. Advances in Genetics. Academic Press, 2008, p. 75-105. doi : https://doi.org/10.1016/S0065-2660(07)00404-X. url : https://www.sciencedirect.com/science/article/pii/S006526600700404X.

[7] Vladimir Gligorijević et Nataša Pržulj. "Methods for biological data integration : perspectives and challenges". en. In : *Journal of The Royal Society Interface* 12.112 (nov. 2015), p. 20150571. issn : 1742-5689, 1742-5662. doi : 10.1098/rsif.2015.0571. url : https://royalsocietypublishing.org/doi/10.1098/rsif.2015.0571.

[8] Marylyn D. Ritchie et al. "Methods of integrating data to uncover genotype–phenotype interactions". en. In : *Nature Reviews Genetics* 16.2 (fév. 2015), p. 85-97. issn : 1471-0056, 1471-0064. doi : 10.1038/nrg3868. url : https://www.nature.com/articles/nrg3868.

[9] Florian Rohart et al. "mixOmics : An R package for 'omics feature selection and multiple data integration". en. In : *PLOS Computational Biology* 13.11 (nov. 2017). Sous la dir. de Dina Schneidman, e1005752. issn : 1553-7358. doi : 10.1371/journal.pcbi.1005752. url : https://dx.plos.org/10.1371/journal.pcbi.1005752.

[10] Mitra Mazinani et Brian Rude. "Population, World Production and Quality of Sheep and Goat Products". In : *American Journal of Animal and Veterinary Sciences* 15.4 (avr. 2020), p. 291-299. issn : 1557-4555. doi : 10.3844/ajavsp.2020.291.299. url : http://thescipub.com/abstract/10.3844/ajavsp.2020.291.299 (visité le 19/06/2024).

[11] WG Hill. "Selective breeding". In : (2017).

[12] Gonzalo Cantalapiedra-Hijar et al. “Efficience Alimentaire : comment mieux la comprendre et en faire un élément de durabilité de l’élevage”. In : *INRAE Productions Animales* (mars 2021), p. 235-248. issn : 2273-7766, 2273-774X. doi : 10.20870/productions-animales.2020.33.4.4594. url : https://productions-animales.org/article/view/4594 (visité le 19/06/2024).

[13] Q. Le Graverand et al. “Predicting feed efficiency traits in growing lambs from their ruminal microbiota”. en. In : *animal* 17.6 (juin 2023), p. 100824. issn : 17517311. doi : 10.1016/j.animal.2023.100824. url : https://linkinghub.elsevier.com/retrieve/pii/S1751731123001209.

[14] D. P. Berry et J. J. Crowley. “CELL BIOLOGY SYMPOSIUM : Genetics of feed efficiency in dairy and beef cattle1”. en. In : *Journal of Animal Science* 91.4 (avr. 2013), p. 1594-1613. issn : 0021-8812, 1525-3163. doi : 10.2527/jas.2012-5862. url : https://academic.oup.com/jas/article/91/4/1594/4716972 (visité le 05/06/2024).

[15] Sobin Kim et Ashish Misra. “SNP genotyping : technologies and biomedical applications”. In : *Annu. Rev. Biomed. Eng.* 9 (2007), p. 289-320.

[16] Trevor Hastie et al. *The elements of statistical learning : data mining, inference, and prediction*. T. 2. Springer, 2009.

[17] Markus Ringnér. “What is principal component analysis ?” In : *Nature biotechnology* 26.3 (2008), p. 303-304.

[18] David G Kleinbaum et al. *Logistic regression*. Springer, 2002.

[19] Muktha Sai Ajay. *Introduction to Artificial Neural Networks*. en. Mai 2020. url : https://towardsdatascience.com/introduction-to-artificial-neural-networks-ac338f4154e5 (visité le 18/06/2024).

[20] Joseph Rocca. *Understanding Variational Autoencoders (VAEs)*. en. Mars 2021. url : https://towardsdatascience.com/understanding-variational-autoencoders-vaes-f70510919f73 (visité le 25/03/2024).

Annexes

TODO : Mettre sélection de dimension en annexe

TODO : Mettre performance sur jeu d'entraînement, avec BER et accuracy (expliquer le calcul