

Integrating heterogeneous data to predict lamb feed efficiency

Quentin Le Graverand, supervised by Christel Marie-Etancelin & Flavie Tortereau

October 2020 – September 2023





Introduction

Feed efficiency

Omics

Models





- ✓ **Environmental stakes:** 84 % of livestock GHG emissions
≈ feed supply chain & enteric methane
(Gerber et al., 2013)



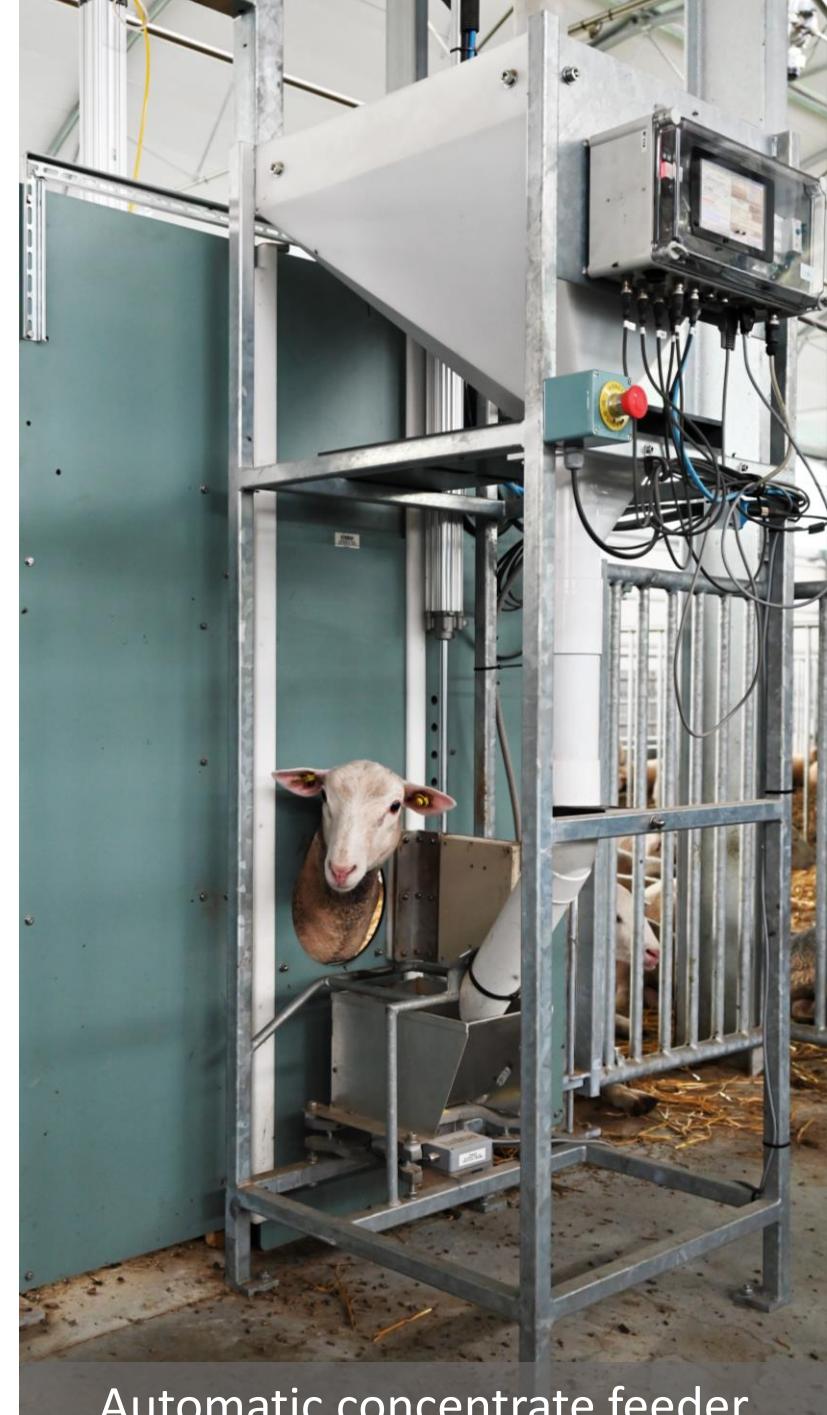
- ✓ **Societal stakes:** feed/food competition



- ✓ **Economic stakes:** feeding is the 1st to 2nd biggest cost

→ Selecting for feed efficiency would increase the sustainability

→ Selecting for feed efficiency requires feed intake records



Automatic concentrate feeder

- Feed efficiency criterion -

Many criteria exist but let's focus on : residual feed intake (**RFI**) (Koch et al., 1963)

RFI was computed by regressing (thesis formula):

$$\underbrace{\text{Average feed intake}}_{\text{Observed}} = \underbrace{\mu + \beta_1 \text{ Weight gain} + \beta_2 \text{ Metabolic weight} + \beta_3 \text{ Muscle} + \beta_4 \text{ Fat}}_{\text{Expected based on production and maintenance}} + \epsilon$$

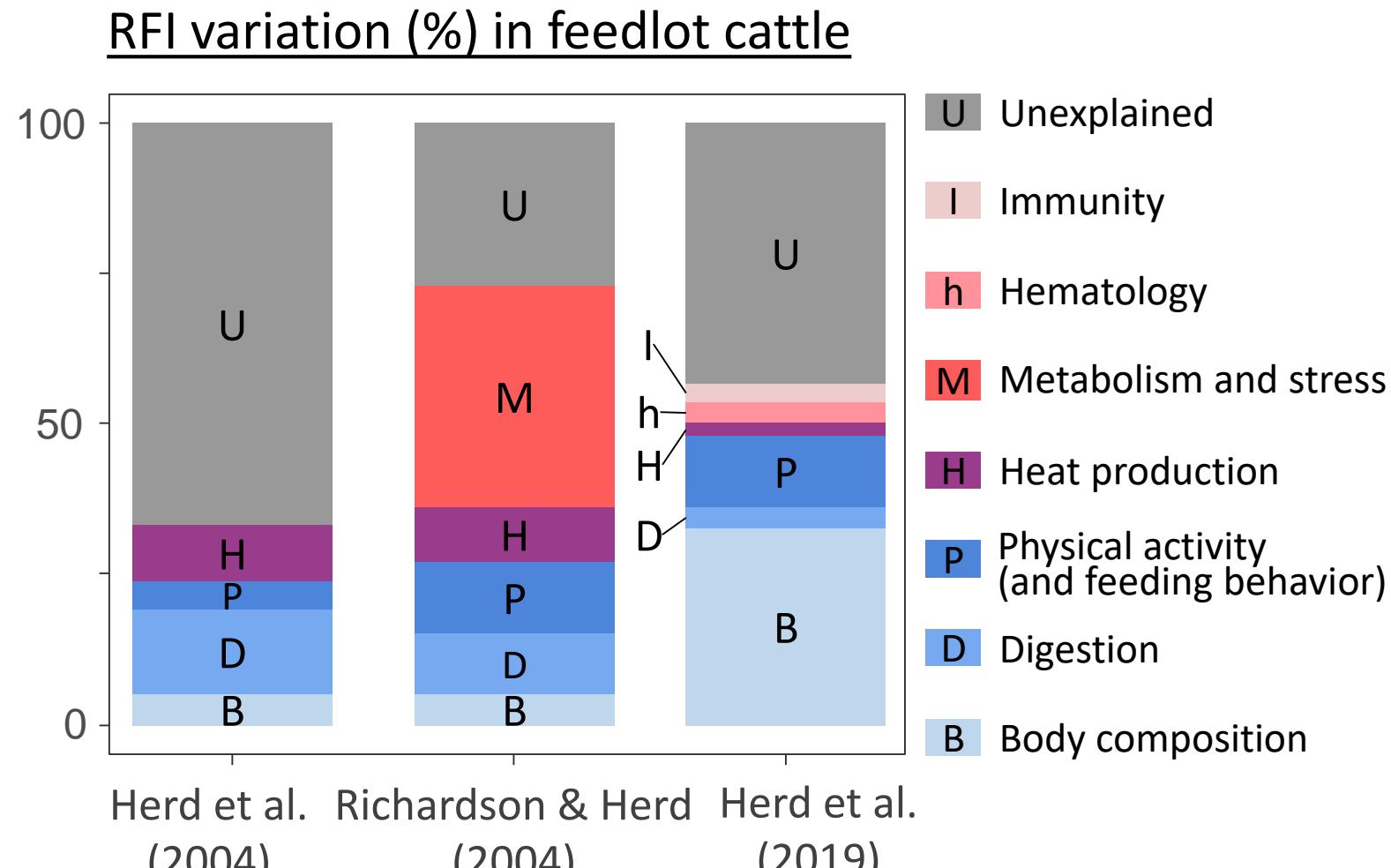


 RFI <0 : efficient

Observed intake

 RFI >0 : inefficient

- Main determinisms in cattle (3 studies):
body composition > digestion >
metabolism > activity
- Some determinisms are still unknown:
27-58% of variations were unexplained in cattle



➡ Past studies focused on the main biological functions

➡ More and more studies dissect traits at the molecular level, with omics data

- Potential of omics as proxies -

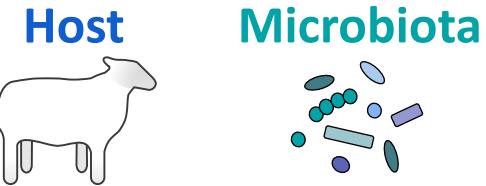


Genomics

What is the host's potential?

$$-0.01 < r_{\text{predictions/RFI}} < 0.67$$

(Pryce et al., 2012; Lu et al., 2016;
Silva et al., 2016; Brunes et al., 2021)

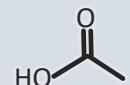


Metabarcoding

Who is there to degrade fiber?

$$0.55 < r_{\text{predictions/RFI}} < 0.71$$

(Ellison et al., 2019; Tapiro et al., 2023)

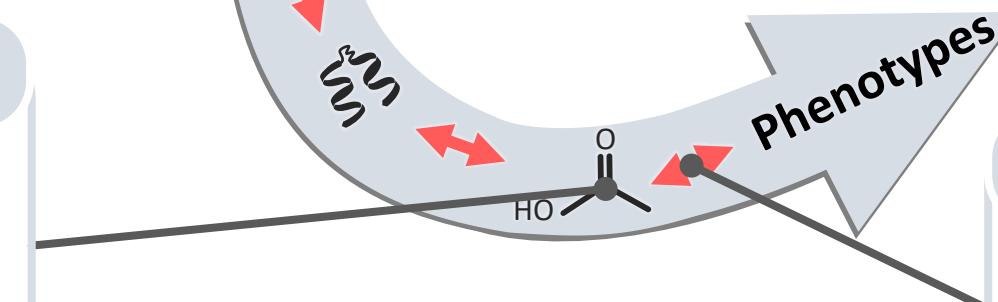


Metabolomics/Lipidomics

What are the host and microbiota producing?

$$0.80 < \text{AUROC}_{\text{RFI extremes}} < 0.87$$

(Goldansaz et al., 2020; Touitou et al., 2022)



Infrared spectra

What is the visible?

$$0.28 < r_{\text{predictions/RFI}} < 0.68$$

(Shetty et al., 2017)

Prediction accuracies vary a lot between studies

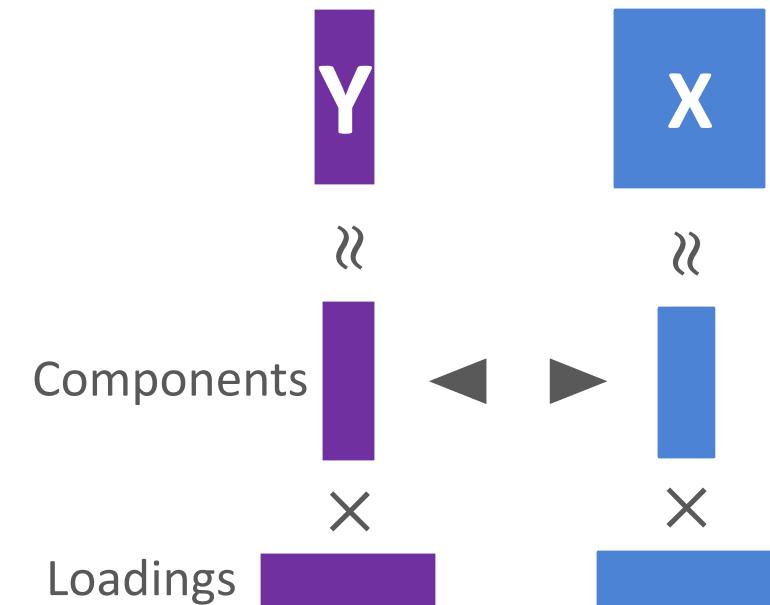
→ Difficult to determine which are the most promising proxies

Many models were used with omics: mixed models, Bayesian models, Random Forest, ...

Partial least squares regressions (PLSR)

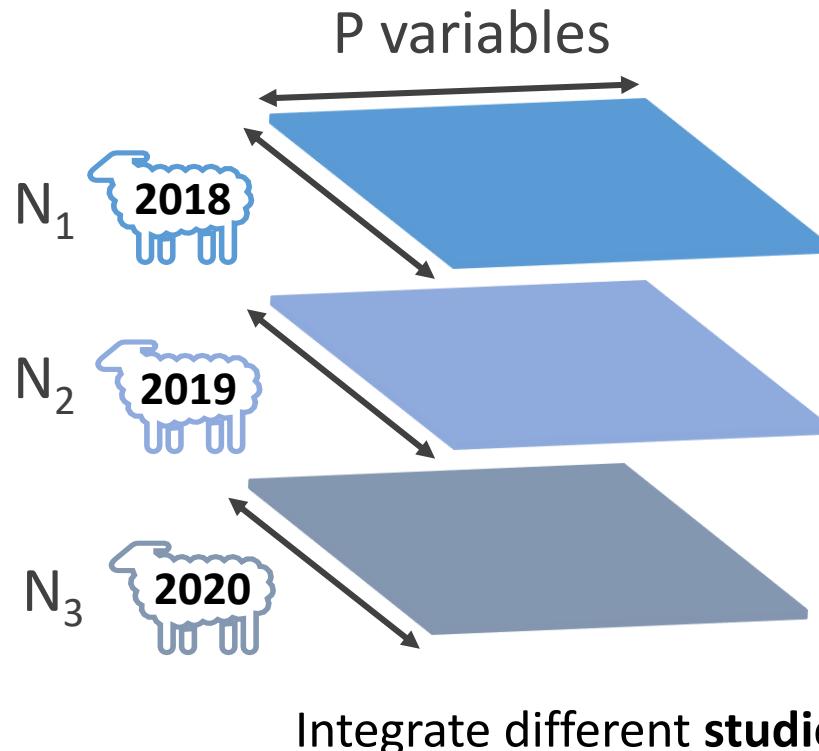
- ✓ Common in chemometrics
- ✓ Easy interpretation of loadings
- ✓ Integration methods were proposed in mixOmics

(Rohart et al., 2017)

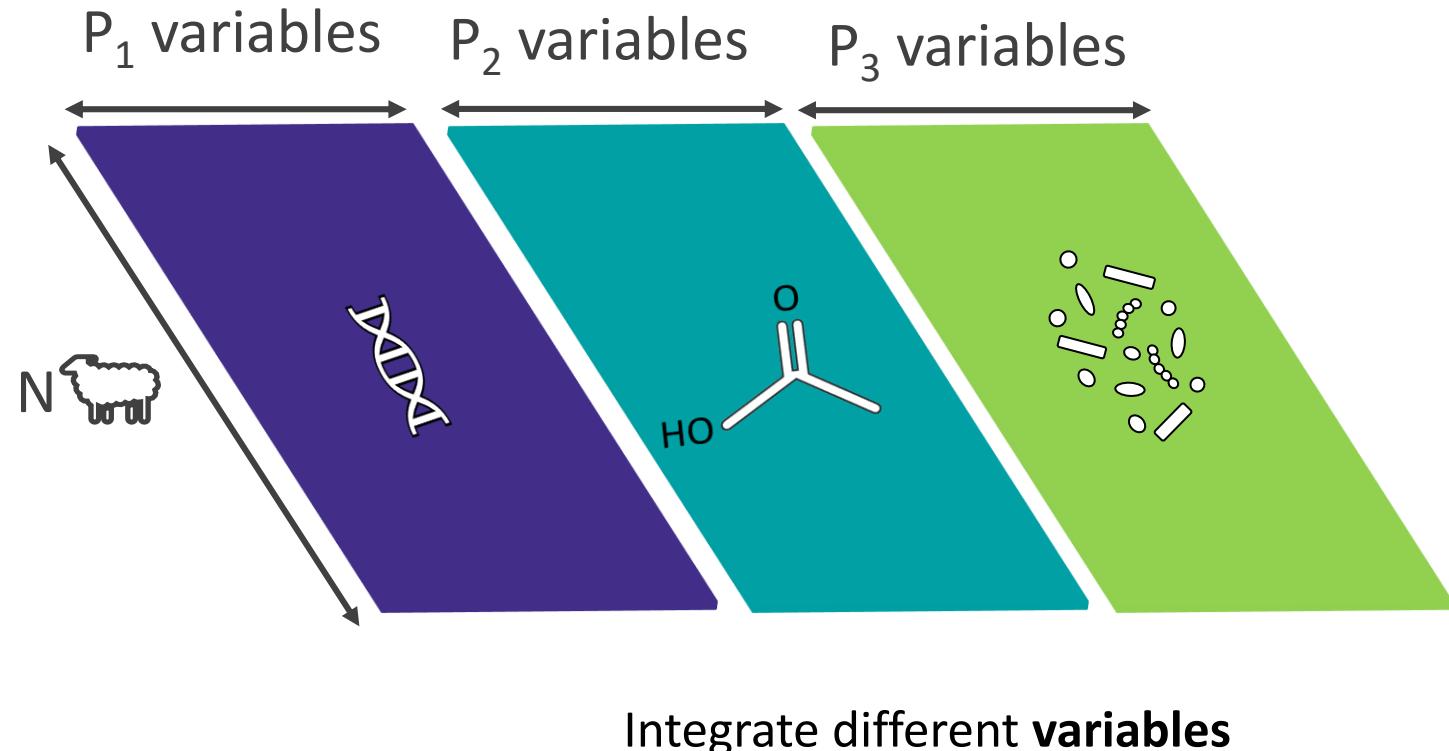


© adapted from Lê Cao & Welham

➡ **Integration strategies** (omics, years) will be tested to increase the prediction accuracy

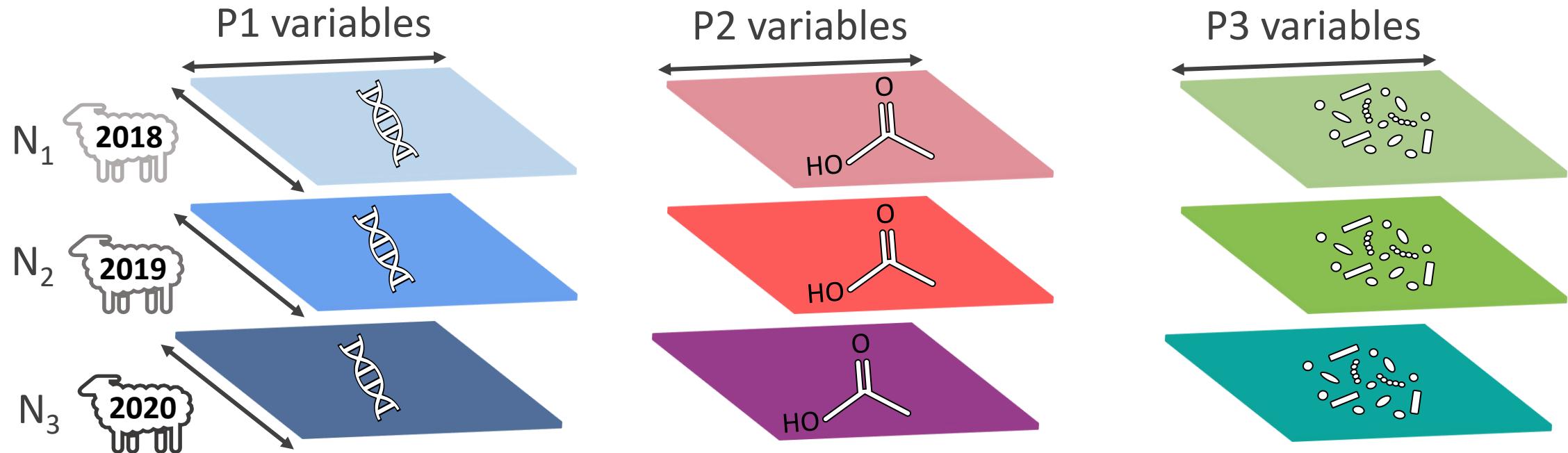
P-integration (MINT.sPLSR)

→ To look for signals generalizable
from one study to another

N-integration (Block.sPLSR)

→ To look for signals from
different biological layers

NP-integration (MINT.block.sPLSR)



Integrate different **studies** and **variables**

→ To look for signals which are generalizable and part of different biological layers

Gaps of knowledge

What determinisms underly feed efficiency ?

Which proxies could predict accurately feed efficiency in meat sheep ?

How integration strategy can increase the prediction accuracy in meat sheep ?

Goal: checking the two following hypotheses

H1: Feed efficiency can be predicted from omics data

H2: Integrating several omics can improve feed efficiency predictions

> Material & methods

- Overall experimental design -



- RFI divergent lines -

Divergent selection is used to:

Exacerbate genetic differences

Anticipate selection consequences



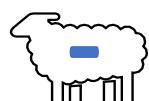
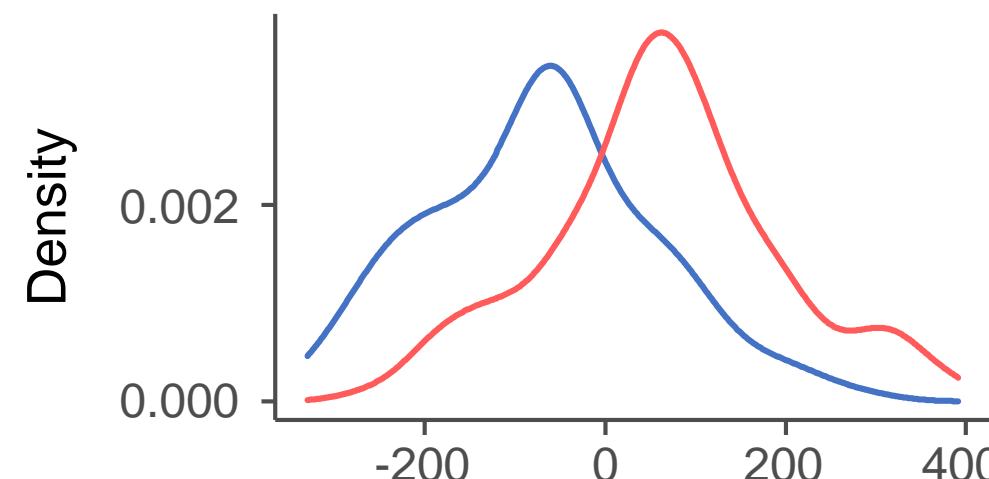
RFI- : most efficient line



RFI+ : least efficient line

Sheep divergent lines on RFI

255 ♂ Romane lambs (2nd and 3rd generations raised between 2018 and 2020)



Most efficient:

$$\mu_{RFI-} = -68 \text{ g/d}$$

RFI phenotype

$$\Delta \text{ between lines} = 132 \text{ g/d}$$

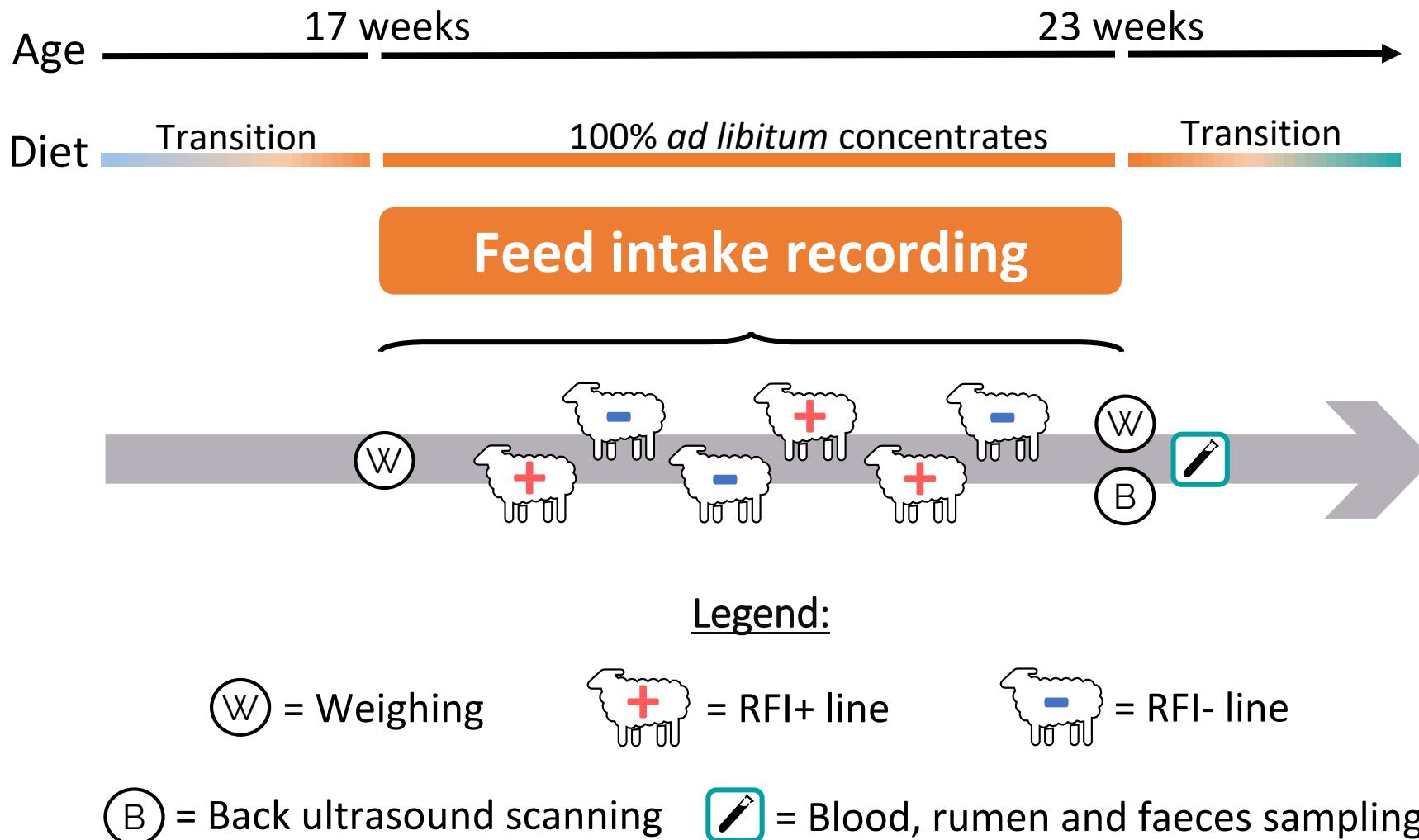


Least efficient:

$$\mu_{RFI+} = 64 \text{ g/d}$$

- Phenotyping and sampling -

255 male lambs with complete phenotypes and omics records



Fixed effects (year, pen, suckling) and covariates (age, weight) are already routinely recorded

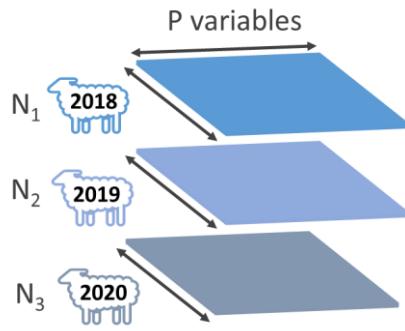


Blood, ruminal, and faecal omics under study

Omics	Technique (+Inference)	Variables	Number	Nature
Genomics		DNA microarray	Genotypes	30 000 Categorical
Metabolomics		NMR spectroscopy	Bucket areas	900 Continuous/Compositional
Lipidomics		GC(-MS)	Fatty acid concentrations	80 Continuous/Compositional
Metabarcoding		16S V4-V5	Prokaryote abundances	600 Discrete/Compositional
Infrared spectra		Near infrared spectroscopy	Absorbances	1 050 Continuous

Total ≈ 32 600 variables after filtering

14 / 47



P-integration

- Predicting feed intake from rumen omics -



Why predict feed intake ?

- Essential to estimate feed efficiency
- Measurable
- Correlated with feed efficiency ($r_{\text{phenotypic}} = 0.62 \pm 0.02$) (Tortereau et al., 2020)

Why predict from rumen omics ?

- The rumen is essential for digestion in ruminants

Two questions

Q1: Does P-integration (//years) improve the prediction accuracy from rumen data ?

→ Two PLS models

Q2: Is it worth it to collect rumen variables ?

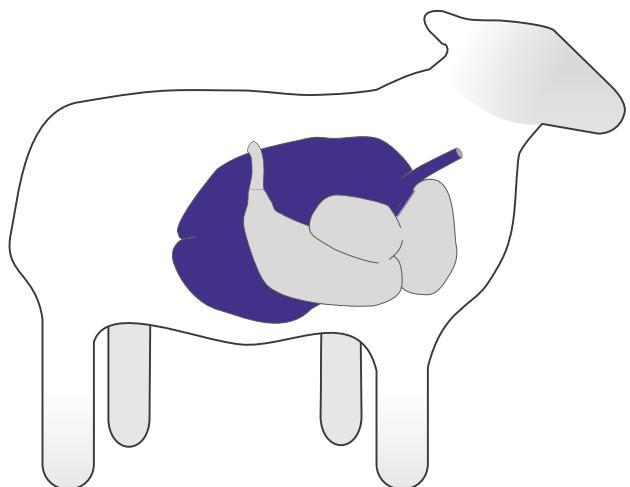
→ Compare prediction accuracies with a gold standard



Gold standard

Farm records

Fixed effects (year, pen, suckling) and covariates (age, weight)



Rumen fluid samples

Microbiota

Prokaryote abundances

Metabolomics

NMR spectra buckets

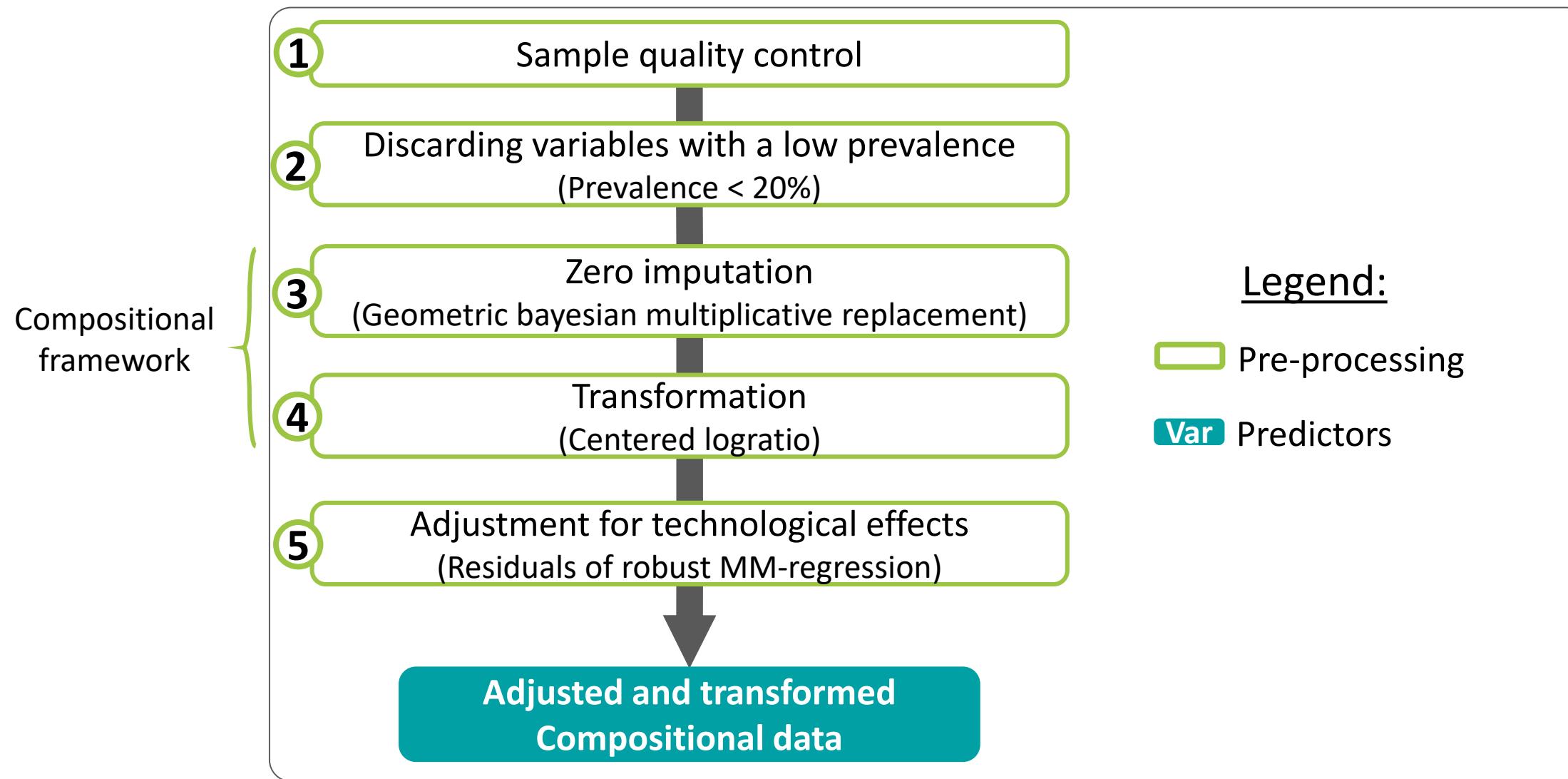
Lipidomics

Volatile fatty acids (VFAs)

Long-chain fatty acids (LCFAs)

5 blocks modelled separately

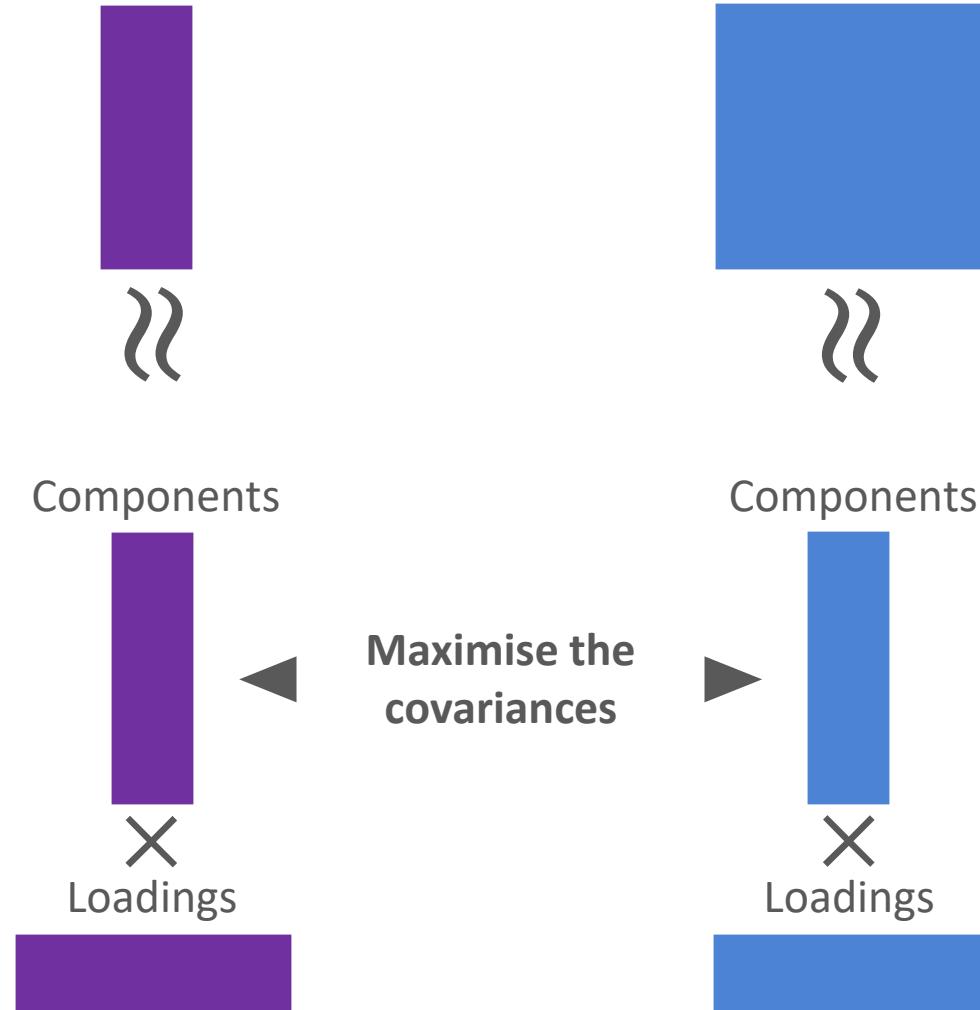
No block integration

Pre-processing of compositional data

Two different sparse partial least squares regressions (sPLSRs)

Y = feed intake

X = transformed abundances of OTUS

**Classic sPLSR**

- Do not account for the year of phenotyping
- Maximizes the covariance between X and Y components
- LASSO selection

Multivariate integrative sPLSR (MINT-sPLSR)

- Account for the 3 years of phenotyping
- Maximizes the sum of covariance between X and Y components, per year
- LASSO selection

(Rohart et al., 2017)

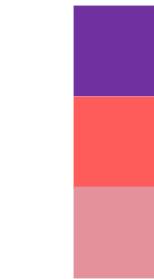
Two different sparse partial least squares regressions (sPLSRs)

Y = feed intake

X = transformed abundances of OTUS



Global components



Maximise the sum
of covariances
per year



Global loadings



Global components



Global loadings



Classic sPLSR

- Do not account for the year of phenotyping
- Maximizes the covariance between X and Y components
- LASSO selection

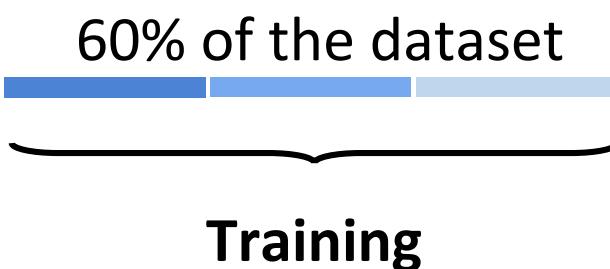
Multivariate integrative sPLSR (MINT-sPLSR)

- Account for the 3 years of phenotyping
- Maximizes the sum of covariance between X and Y components per year
- LASSO selection

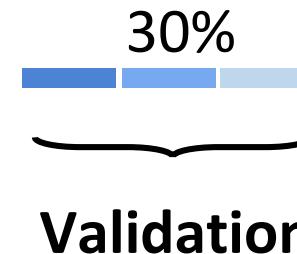
(Rohart et al., 2017)

Cross-validation: repeated random subsampling

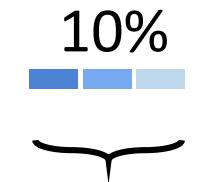
- Contemporaneous animals in training and testing sets
- Stratification year x pen x line
- Repeated 100 times



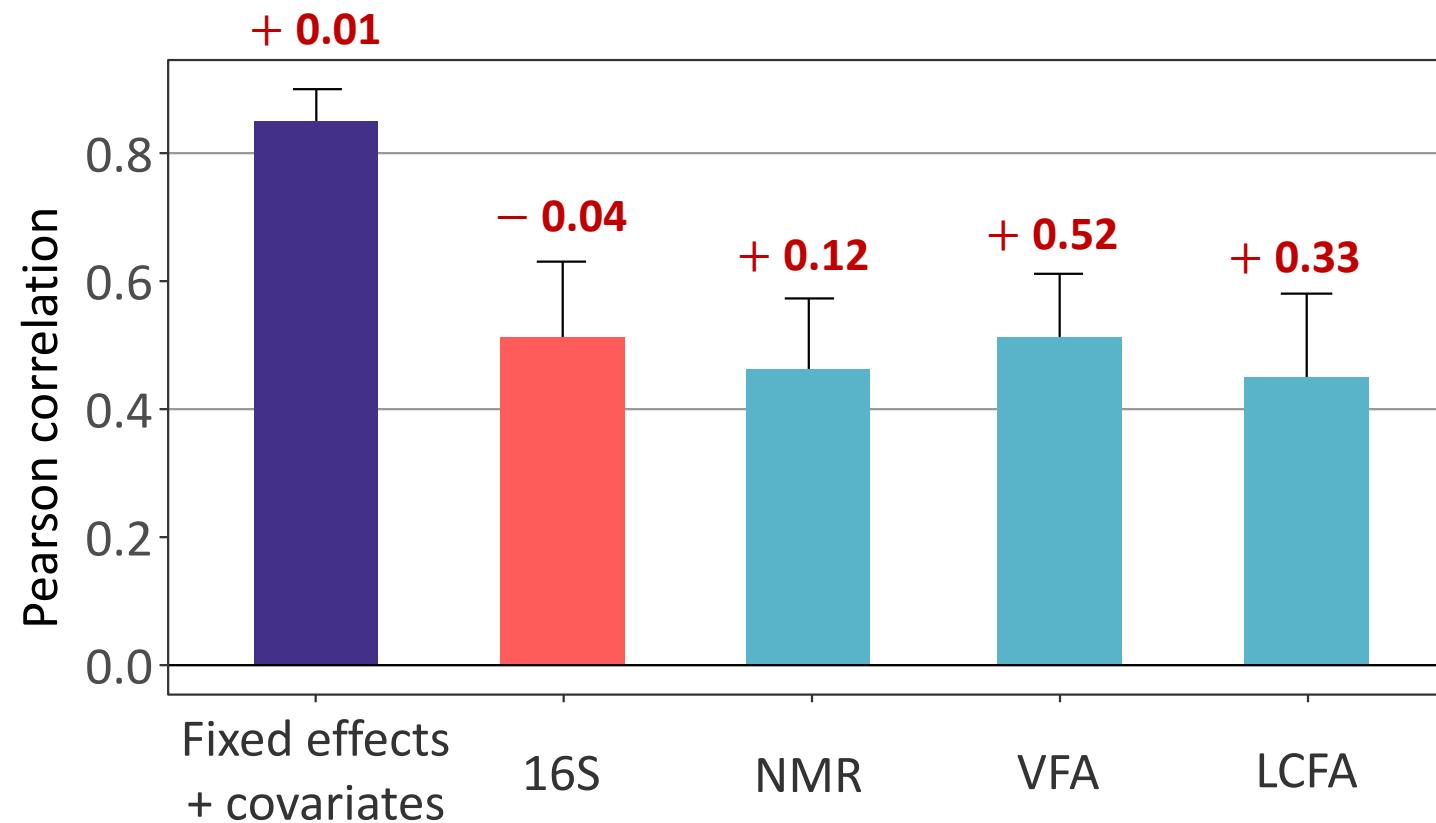
Fit sPLSR or MINT-sPLSR
models



Select
hyperparameters



Compute the prediction accuracy
(=Pearson correlation)

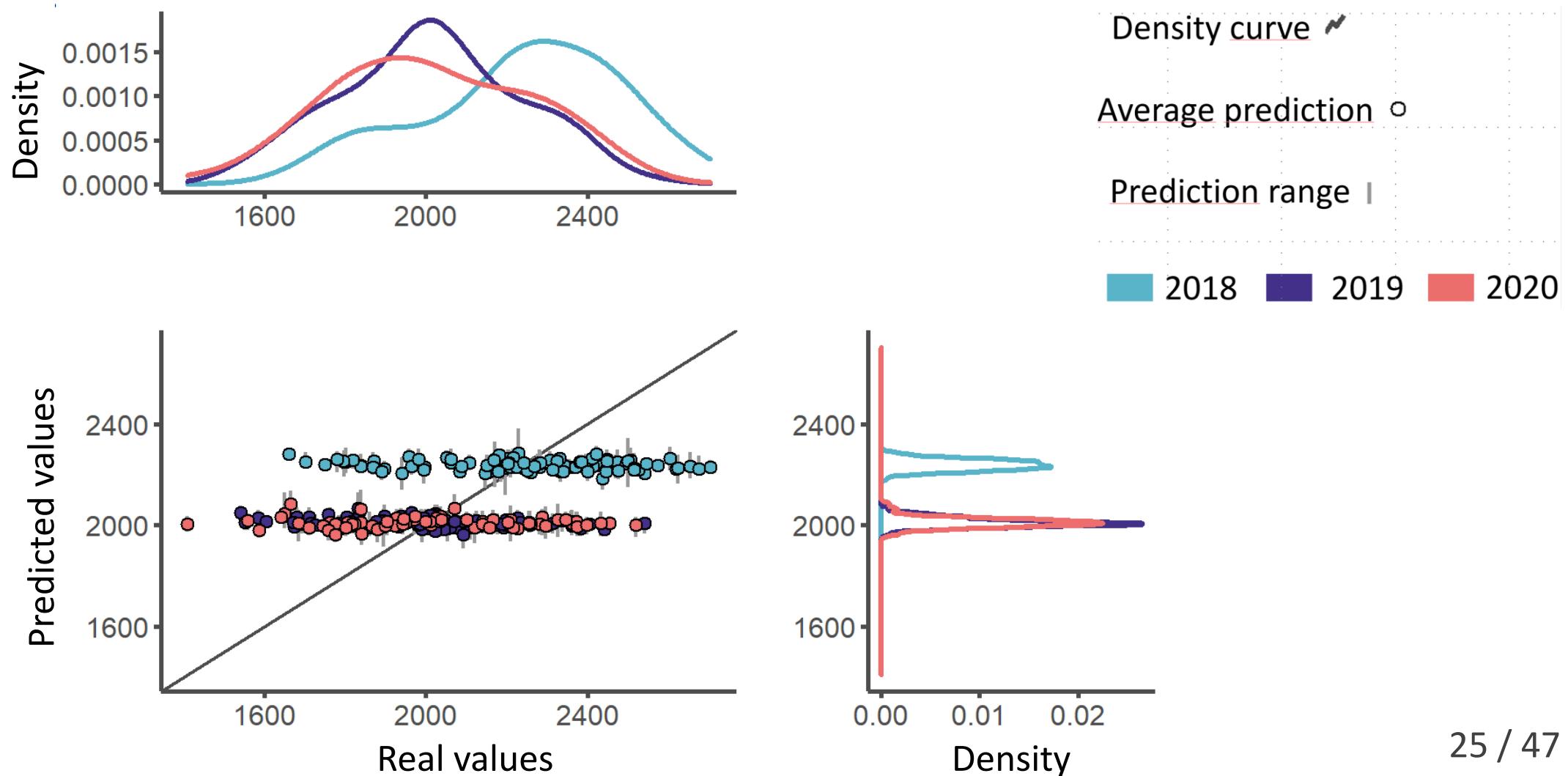
MINT-sPLSR prediction accuracies

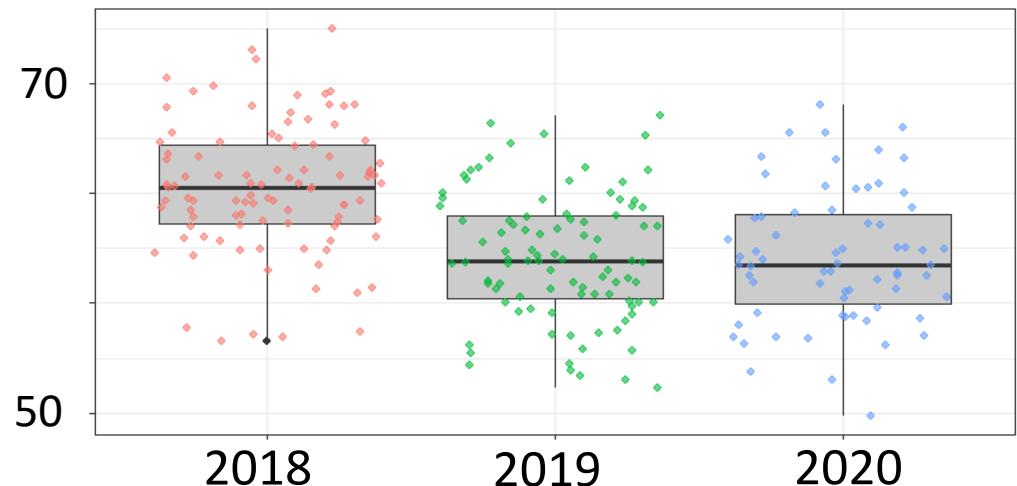
±: MINT-sPLSR accuracy compared to the sPLSR accuracy

Discussion hypotheses

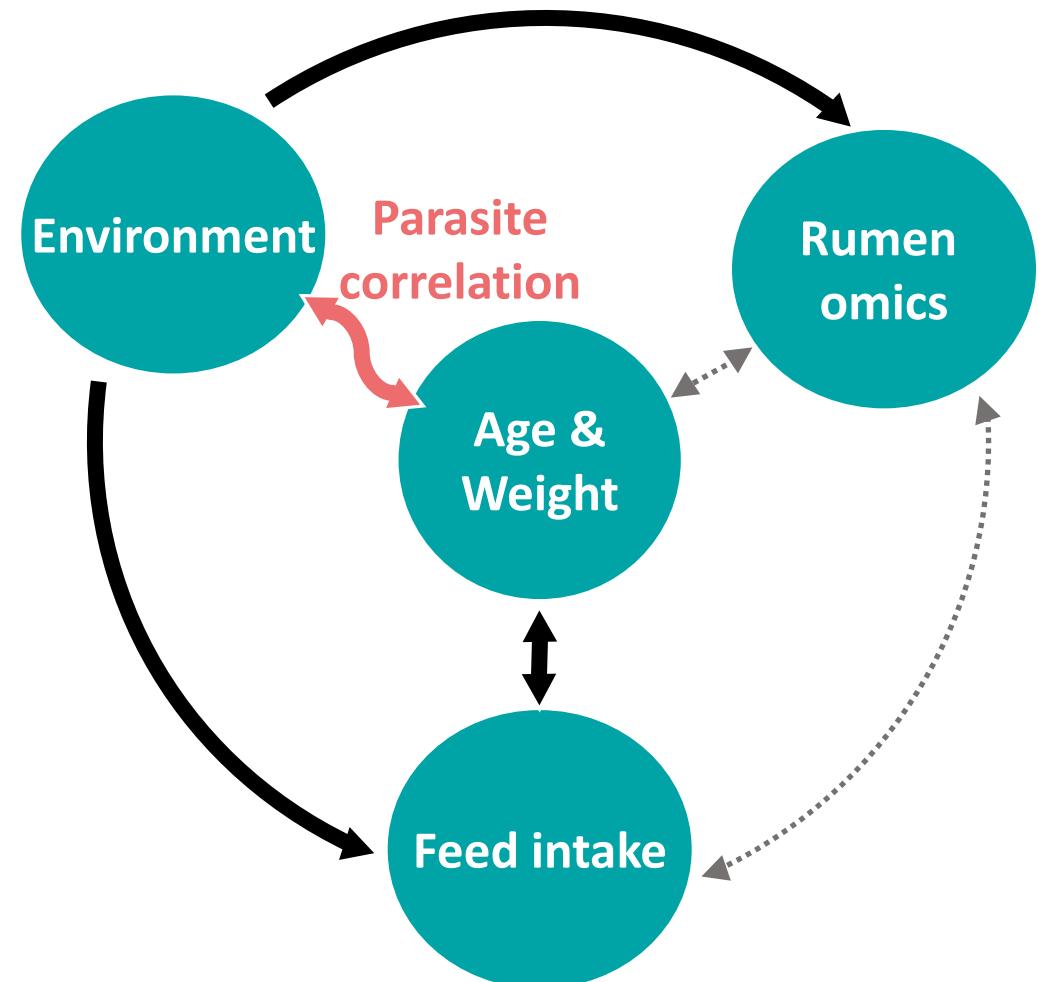
- No accuracy gain when:
study effect < proxies effects
- No or negligible gain when:
there are no generalisable proxies
- Accuracy gain when:
there are generalisable proxies ??

→ Let's look into the proxies having the biggest accuracy increase: VFAs

Distribution of MINT-sPLSR predictions of feed intake from VFAs

Annual distribution of body weights (kg)

- 1-Animals were older in 2018
- 2- Animals were allocated by weight into groups !!

Spurious correlations

Feed efficiency cannot be estimated if intake is predicted from proxies of body weight differences

Reminder that: correlation \Leftrightarrow causation

Heavy cats



Dr Claire UA Millington (Twitter)

■ **Q1: Does P-integration (//years) improve the prediction accuracy from rumen data ?**

No

Fixed effects and covariates: the year of study did not affect much proxies

Microbiota: proxies do not generalize well from one study to another

Yes

Metabolomics and lipomics: some proxies are generalizable and sensitive to the year of study

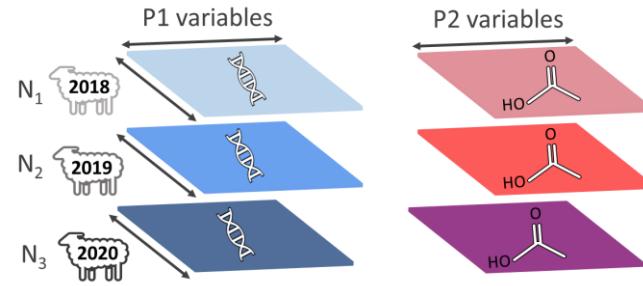
■ **Q2: Is it worth it to collect rumen variables ?**

No

Body weight is easier to record, was a better proxy

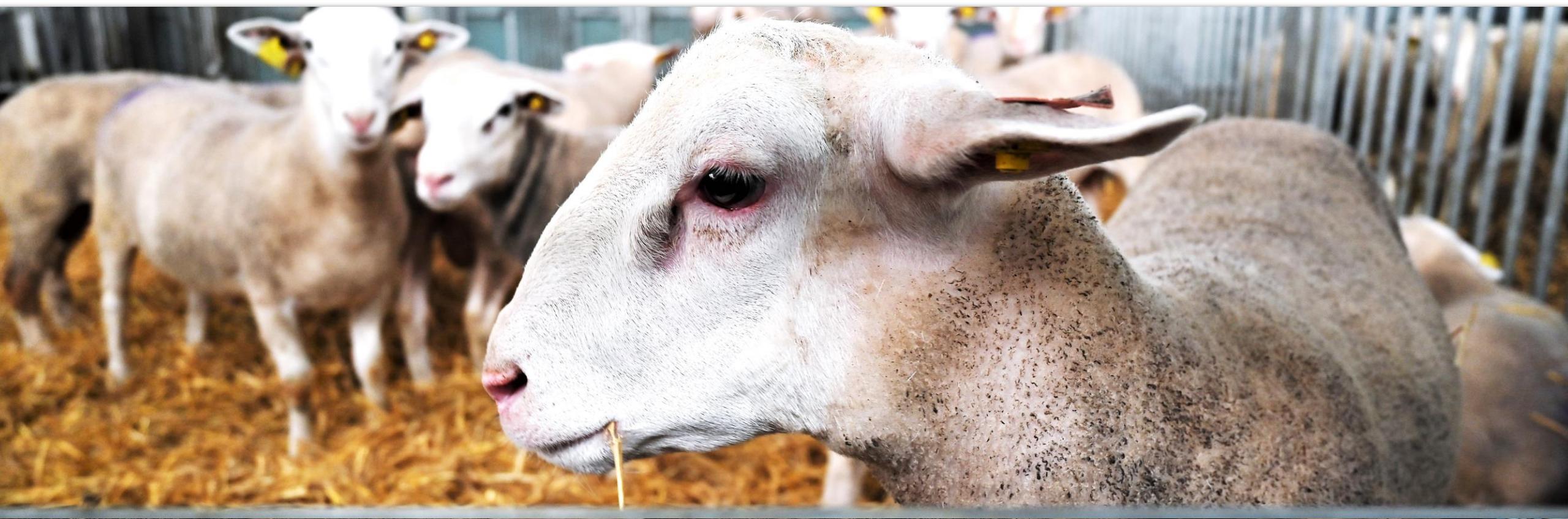
Rumen variables are too noisy ? Sampling + environmental + technological effects ?

➡ Is it the same for feed efficiency ?



NP-integration

- Predicting RFI by integrating heterogeneous data -



Context

Previously: efficiency cannot be estimated if the intake is predicted from microbiota, fixed effects and covariates

Other omics may provide proxies for feed efficiency

Few studies compared omics in one population, with individuals raised different years

Two questions

Q1: Which are the best proxies of feed efficiency ? → **Genomics, metabolomics, lipidomics, phenomics, inferred data**

Q2: Does NP-integration (//blocks, years) improve the prediction accuracy of feed efficiency ? → **Single vs multi-block**



Farm records

Fixed effects (year, pen, suckling) and covariates (age, weight)



Pedigree (relatedness matrix)

Blood samples**Genomics**

SNPs

Plasma metabolomics

NMR spectra buckets

Inferred metabolite concentrations

ASICS

**Faeces samples****Infrared spectra**

Near infrared spectra absorbances

Spectra first derivative

**Rumen fluid samples****Microbiota**

Prokaryote abundances

Inferred prokaryote functions

FROGSFunc

**Metabolomics**

NMR spectra buckets

Inferred metabolite concentrations

ASICS

**Lipidomics**

Volatile fatty acids (VFAs)

Long-chain fatty acids (LCFAs)

Total: 13 blocks of predictors

Early integration (concatenation) may struggle with:

- The **dimensionality curse**: 255 sheepx 35 000 variables
- The **data heterogeneity**:
 - discrete, continuous, categorical
 - compositionality
 - variable number per block

(Picard et al., 2021)



Late integration may help:

- 1- Pre-process all blocks separately
- 2-Fit one submodel per block of predictors
- 3- Fit a metamodel on the predictions of submodels

Dimensionality curse





mixOmics partition: training and testing sets

mixOmics proposed the MINT.block.sPLSR

(Rohart et al., 2017)

- Block weights depend on the training errors

16S weight > SNPs weight

⚠ Overfitting

Thesis (in collaboration with KA Lê Cao): training, validation and testing sets

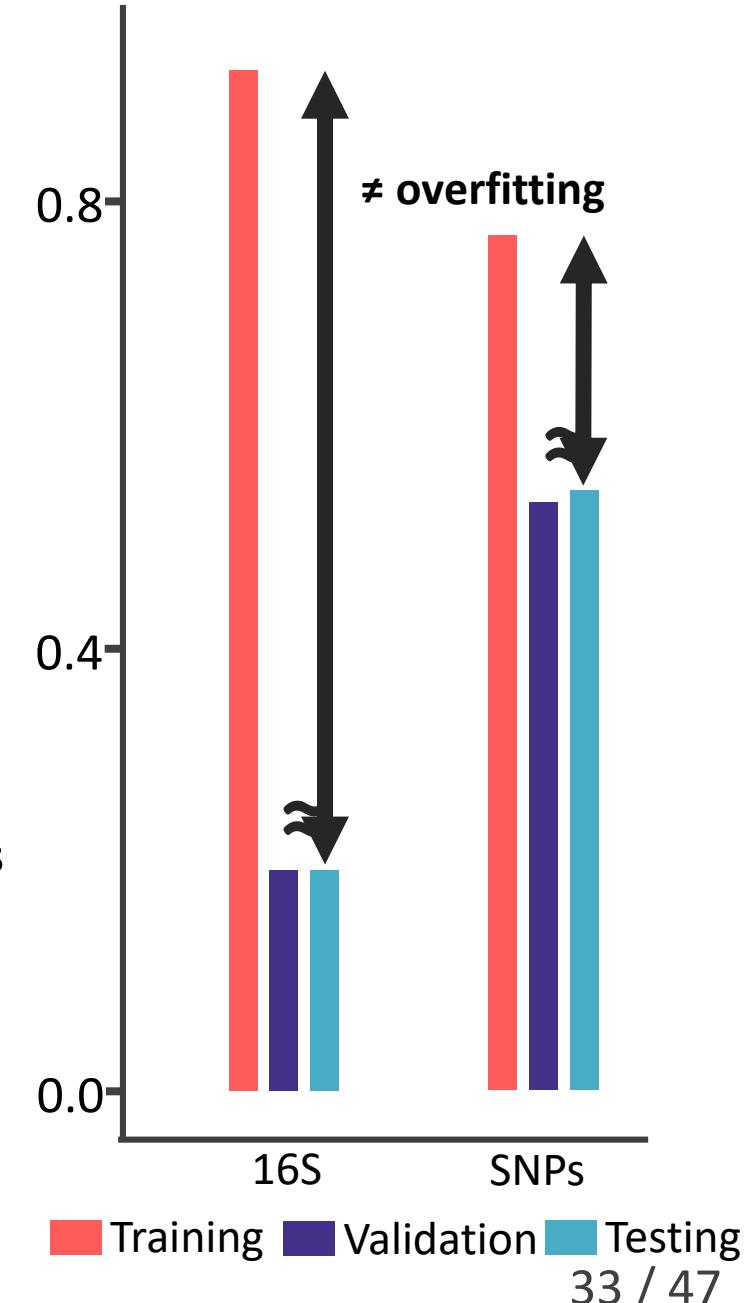
Competitive machine learning showcased the blending strategy

(Töscher and Jahrer, 2009)

- Block weights depend on the validation errors

SNPs weight > 16S weight

Average Pearson correlation



Predicting RFI

- Integration strategy -

Stratification year x pen x line

Training 60%

Validation 30%

Testing 10%

13 separate
blocks

Try different MINT-sPLSR
hyperparameters

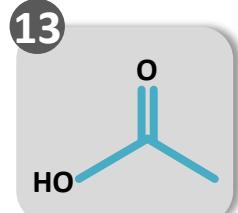
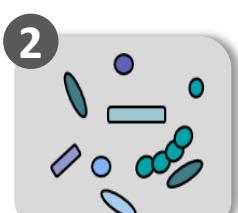
P-integration



Select the best
hyperparameters

Blend the
13 predictions

Evaluate



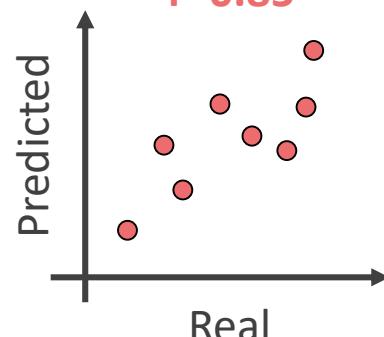
Predictions
(block 1)

Predictions
(block 2)

Predictions
(block 13)

NP-integration

Meta-
model
**MINT-sPLSR
(Nested CV)**



$\times 100$ times

Prediction accuracy of RFI from different sets of predictors

Source	Variables	Average (SD)
Farm records 	Fixed effects + covariates	0.35 ^{de} (0.13)
	Pedigree	0.54 ^g (0.13)
Blood 	Genotypes	0.54 ^g (0.13)
	Buckets	0.37 ^{ef} (0.14)
	Metabolites	0.34 ^d (0.15)
Rumen 	Prokaryote abundances	0.20 ^{ab} (0.15)
	Prokaryote functions	0.18 ^a (0.17)
	Buckets	0.23 ^b (0.17)
	Metabolites	0.27 ^c (0.17)
	VFAs	0.24 ^b (0.13)
Faeces 	LCFAs	0.23 ^b (0.15)
	Spectral absorbances	0.39 ^f (0.14)
	Spectral first derivative	0.39 ^f (0.14)

^{a,b}: pairwise permutation tests (paired Welch's t-tests)

Prediction accuracy of RFI from different sets of predictors

Source	Variables	Average (SD)
Farm records 	Fixed effects + covariates	0.35 ^{de} (0.13)
	Pedigree	0.54 ^g (0.13)
	Genotypes	0.54 ^g (0.13)
Blood 	Buckets	0.37 ^{ef} (0.14)
	Metabolites	0.34 ^d (0.15)
	Prokaryote abundances	0.20 ^{ab} (0.15)
Rumen 	Prokaryote functions	0.18 ^a (0.17)
	Buckets	0.23 ^b (0.17)
	Metabolites	0.27 ^c (0.17)
	VFAs	0.24 ^b (0.13)
	LCFAs	0.23 ^b (0.15)
Faeces 	Spectral absorbances	0.39 ^f (0.14)
	Spectral first derivative	0.39 ^f (0.14)

- Best predictors: genomics and pedigree (lambs are closely related in divergent lines !)

^{a,b}: pairwise permutation tests (paired Welch's t-tests)

Prediction accuracy of RFI from different sets of predictors

Source	Variables	Average (SD)
Farm records 	Fixed effects + covariates	0.35 ^{de} (0.13)
	Pedigree	0.54 ^g (0.13)
Blood 	Genotypes	0.54 ^g (0.13)
	Buckets	0.37 ^{ef} (0.14)
	Metabolites	0.34 ^d (0.15)
Rumen 	Prokaryote abundances	0.20 ^{ab} (0.15)
	Prokaryote functions	0.18 ^a (0.17)
	Buckets	0.23 ^b (0.17)
	Metabolites	0.27 ^c (0.17)
	VFAs	0.24 ^b (0.13)
Faeces 	LCFAs	0.23 ^b (0.15)
	Spectral absorbances	0.39 ^f (0.14)
	Spectral first derivative	0.39 ^f (0.14)

- Best predictors: genomics and pedigree (lambs are closely related in divergent lines !)
- Intermediate predictors : faecal phenomics, plasma metabolomics, fixed effects and covariates

^{a,b}: pairwise permutation tests (paired Welch's t-tests)

Prediction accuracy of RFI from different sets of predictors

Source	Variables	Average (SD)
Farm records 	Fixed effects + covariates	0.35 ^{de} (0.13)
	Pedigree	0.54 ^g (0.13)
Blood 	Genotypes	0.54 ^g (0.13)
	Buckets	0.37 ^{ef} (0.14)
	Metabolites	0.34 ^d (0.15)
Rumen 	Prokaryote abundances	0.20 ^{ab} (0.15)
	Prokaryote functions	0.18 ^a (0.17)
	Buckets	0.23 ^b (0.17)
	Metabolites	0.27 ^c (0.17)
	VFAs	0.24 ^b (0.13)
	LCFAs	0.23 ^b (0.15)
Faeces 	Spectral absorbances	0.39 ^f (0.14)
	Spectral first derivative	0.39 ^f (0.14)

- Best predictors: genomics and pedigree (lambs are closely related in divergent lines !)
- Intermediate predictors : faecal phenomics, plasma metabolomics, fixed effects and covariates
- Worst predictors: rumen variables

^{a,b}: pairwise permutation tests (paired Welch's t-tests)

Prediction accuracy of RFI from different sets of predictors

Source	Variables	Average (SD)
Farm records 	Fixed effects + covariates	0.35 ^{de} (0.13)
	Pedigree	0.54 ^g (0.13)
	Genotypes	0.54 ^g (0.13)
Blood 	Buckets	0.37 ^{ef} (0.14)
	Metabolites	0.34 ^d (0.15)
	Prokaryote abundances	0.20 ^{ab} (0.15)
	Prokaryote functions	0.18 ^a (0.17)
Rumen 	Buckets	0.23 ^b (0.17)
	Metabolites	0.27 ^c (0.17)
	VFAs	0.24 ^b (0.13)
	LCFAs	0.23 ^b (0.15)
Faeces 	Spectral absorbances	0.39 ^f (0.14)
	Spectral first derivative	0.39 ^f (0.14)

- Best predictors: genomics and pedigree (lambs are closely related in divergent lines !)
 - Intermediate predictors : faecal phenomics, plasma metabolomics, fixed effects and covariates
- Worst predictors: rumen variables
- Inference: conflicting results

^{a,b}: pairwise permutation tests (paired Welch's t-tests)

Prediction accuracy of RFI from different sets of predictors

Source	Variables	Average (SD)
Farm records 	Fixed effects + covariates	0.35 ^{de} (0.13)
	Pedigree	0.54 ^g (0.13)
	Genotypes	0.54 ^g (0.13)
Blood 	Buckets	0.37 ^{ef} (0.14)
	Metabolites	0.34 ^d (0.15)
	Prokaryote abundances	0.20 ^{ab} (0.15)
Rumen 	Prokaryote functions	0.18 ^a (0.17)
	Buckets	0.23 ^b (0.17)
	Metabolites	0.27 ^c (0.17)
	VFAs	0.24 ^b (0.13)
	LCFAs	0.23 ^b (0.15)
Faeces 	Spectral absorbances	0.39 ^f (0.14)
	Spectral first derivative	0.39 ^f (0.14)
All 	Predictions	0.59^h (0.12)

^{a,b}: pairwise permutation tests (paired Welch's t-tests)

- Best predictors: genomics and pedigree (lambs are closely related in divergent lines !)
 - Intermediate predictors : faecal phenomics, plasma metabolomics, fixed effects and covariates
- Worst predictors: rumen variables
 - Inference: conflicting results



Data integration significantly improved the prediction accuracy

- **Block selection**

On average 6-7 blocks are selected (out of 13)



- **Block selection**

On average 6.64 blocks were selected (out of 13)

- **Value importance in the projection (VIP):**

1 value per block and training set

The most important blocks have the highest VIP

Pedigree > Genotypes > Faecal phenomics

> Fixed effects and covariates > Plasma buckets

- **Contribution and accuracy rankings were similar**

Distribution of VIP per block

Fixed effects+covariates

Pedigree relatedness

Genotypes

Plasma buckets

Plasma concentrations

Prokaryote abundances

Prokaryote functions

Rumen buckets

Rumen concentrations

VFA

LCFA

NIRS absorbances

NIRS first derivative

Value importance in the projection

■ Sampling

Our results do not advocate for collecting rumen data: too complex and noisy ?

Faeces represent the end-product of all digestion and assimilation processes (Monteiro et al., 2022)

Blood and faeces can be sampled in larger cohorts than rumen fluids

■ Inference

Prediction accuracies did not differ consistently between original blocks and inferred blocks (e.g. buckets or metabolites)

The meta-model selection did not consistently favor the original block or inferred block

Inference may be used to ease the interpretation of the biology underlying proxies

■ Blending integration

Do not rely on inter-omic associations

■ Q1: Which are the best proxies of feed efficiency ?

Genetics > Faecal phenomics > Fixed effects and covariates > Plasma buckets

⚠ Genomic and genetic accuracies are likely inflated in divergent lines !

■ Q2: Does NP-integration improve the prediction accuracy of feed efficiency ?

Yes

Integrating omics improved the prediction accuracy of RFI by +0.05 (from 0.54 to 0.59)

Similar results were observed in  :

Accuracy increased by +0.07 maximum when metagenomics/metabolomics are added to genomics
(Hess et al., 2022; Ross et al., 2020)

➡ “More is better” but is it enough in practice?
(Huang et al., 2017)



Thesis perspectives

How omics could help in animal or plant breeding?

What gaps of knowledge remain?



Use of omics in animal and plant breeding

To increase the genomic prediction accuracy with bivariate models

(Hayes et al., 2017)

To deal with missing phenotypes by training mixed models handling omics

(Christensen et al., 2021)

To select traits by working on the hologenome

(Larzul et al., 2023; Gonzalez-Recio et al., 2023)

Gaps of knowledge

When should we sample ? In pigs, timing influenced the accuracy of growth predictions from the faecal microbiota

(Maltecca et al., 2019)

How often should we sample ? Longitudinal analysis was suggested to account for omics dynamics and feed efficiency

(Maltecca et al., 2019; Martin et al., 2021)

Which feed efficiency determinisms can be unraveled with omics integration?

Acknowledgements



Flavie Tortereau

Christel Marie-Etancelin

Annabelle Meynadier

Jean-Louis Weisbecker



P3R Experimental Unit



Kim-Anh Lê Cao





Question
time !

quentin.le-graverand@inrae.fr

