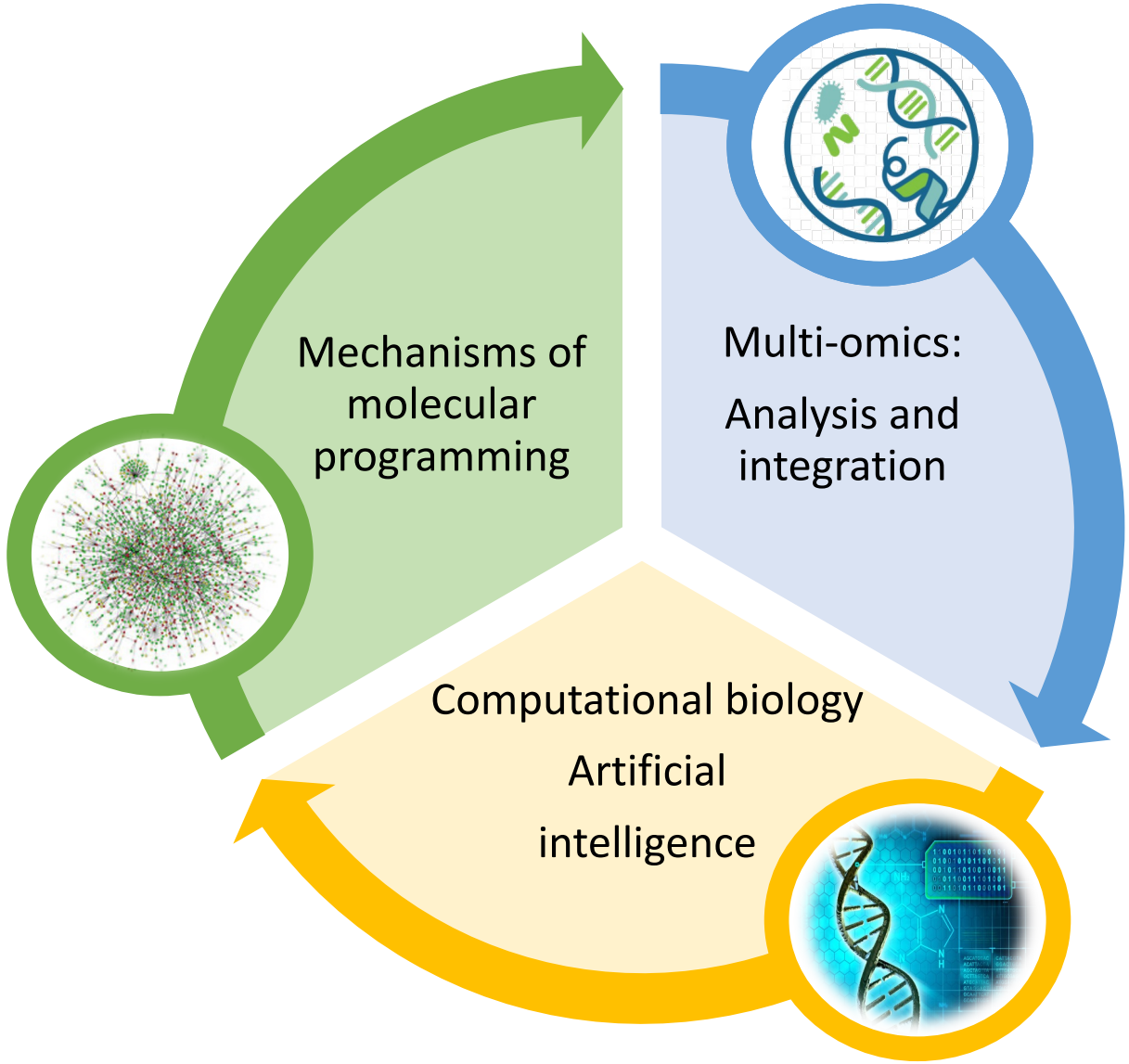
HIVE: a new model for Horizontal multi-omics Integration with Variational autoEncoders

Silvia Bottini, PhD

Junior Professor Chair INRAE

Team "M2P2" – Institut Sophia Agrobiotech – Sophia Antipolis

Scientific interests

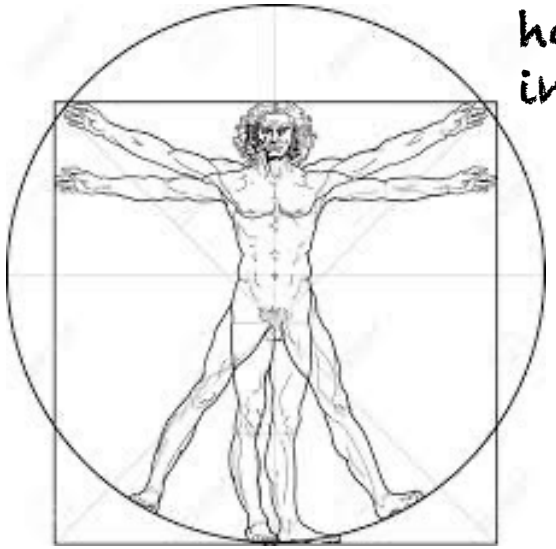


From human to plant health

vaccine development

cancer

host-pathogens interactions



fundamental molecular biology mechanisms

neurological diseases

rare diseases

Silvia Bottini, junior professor chair - ISA UMR1355 - INRAE - CNRS - UniCA - silvia.bottini@inrae.fr

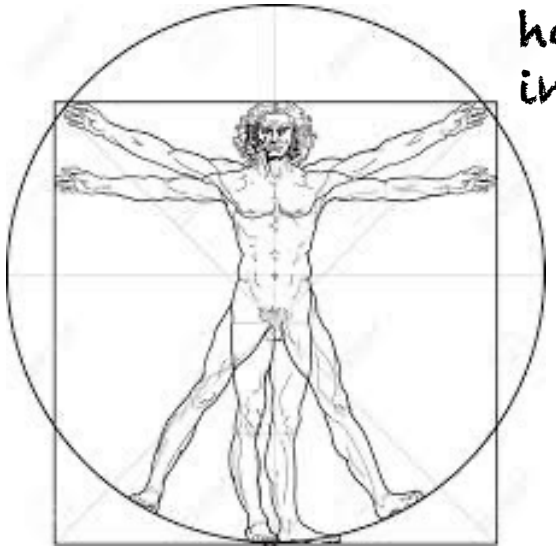
From human to plant health

vaccine development

cancer

host-pathogens interactions

fundamental molecular biology mechanisms



neurological diseases

rare diseases

Plant immunity response

Plants play an important role in human existence

98% of **oxygen** is produced by plants

82% of the **calories** in **human food** supply are provided by terrestrial plants

66% of crop production accounted by **9** species of **plant variety**

Climate change

Environmental degradation

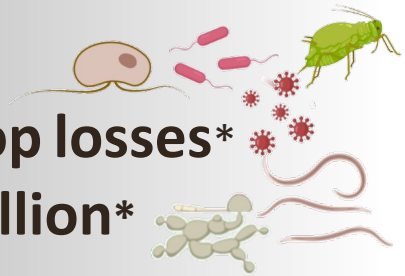
Population growth

Poor agripractice

Biotic stressors

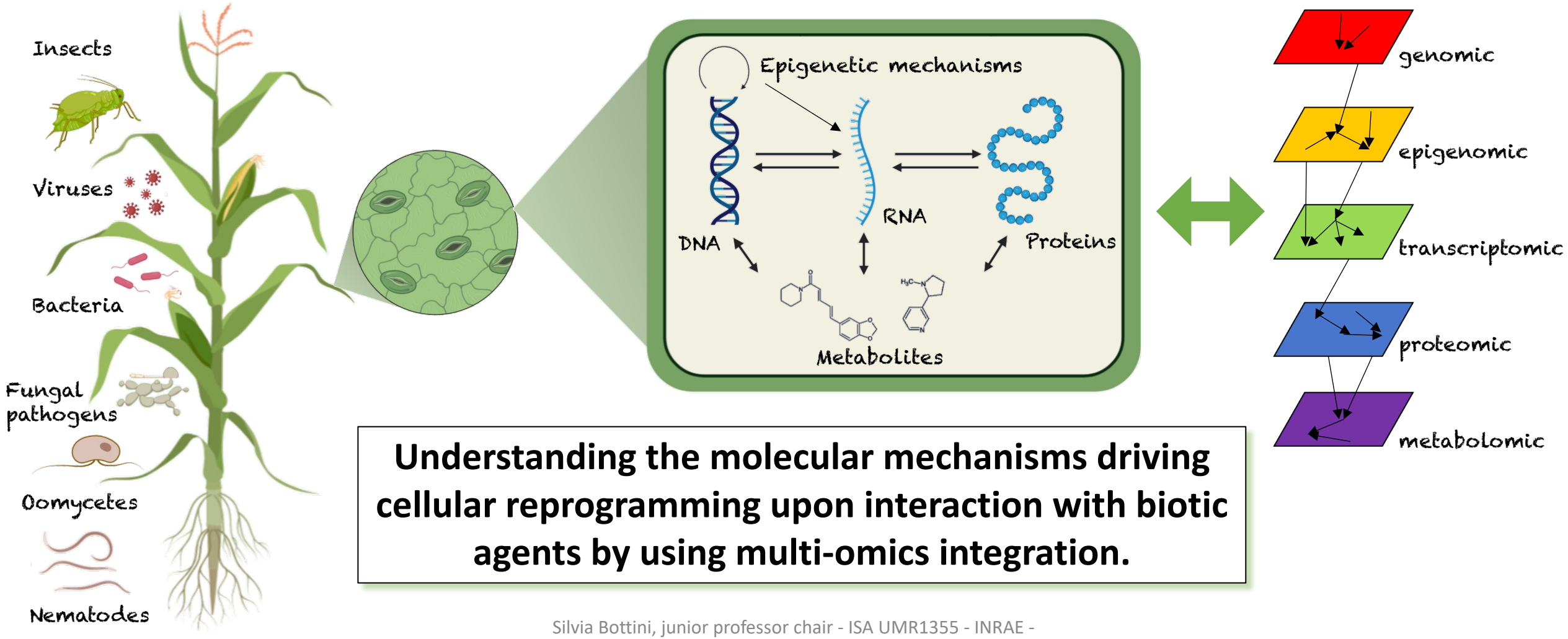
Between **20% to 40%** of **crop losses***

Economic loss around **$220 billion***



*according to the Food and Agriculture Organization of the United Nations (2023)

How to improve plant health?

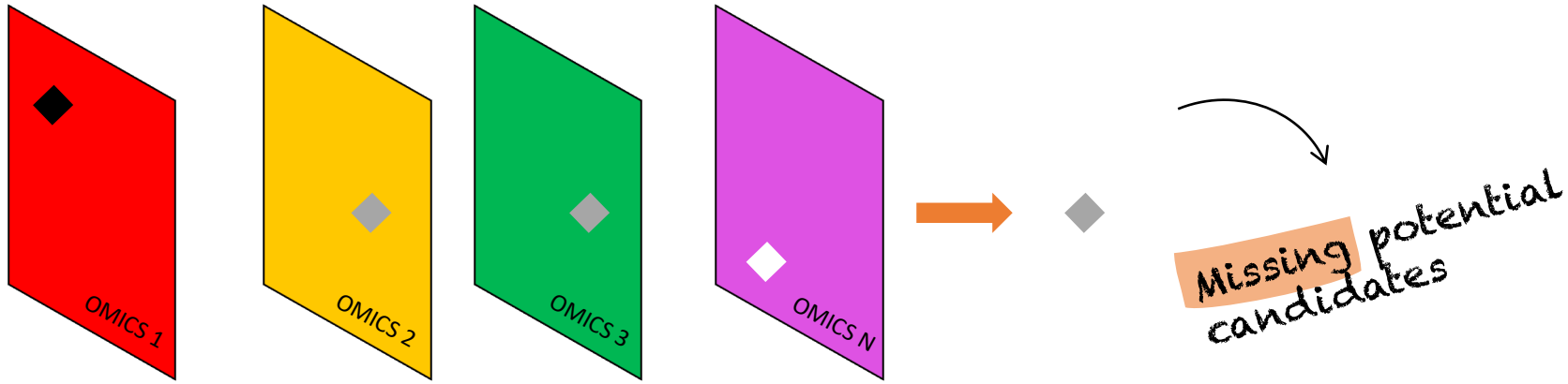


Understanding the molecular mechanisms driving cellular reprogramming upon interaction with biotic agents by using multi-omics integration.

Silvia Bottini, junior professor chair - ISA UMR1355 - INRAE - CNRS - UniCA - silvia.bottini@inrae.fr

Why multi-omics integration?

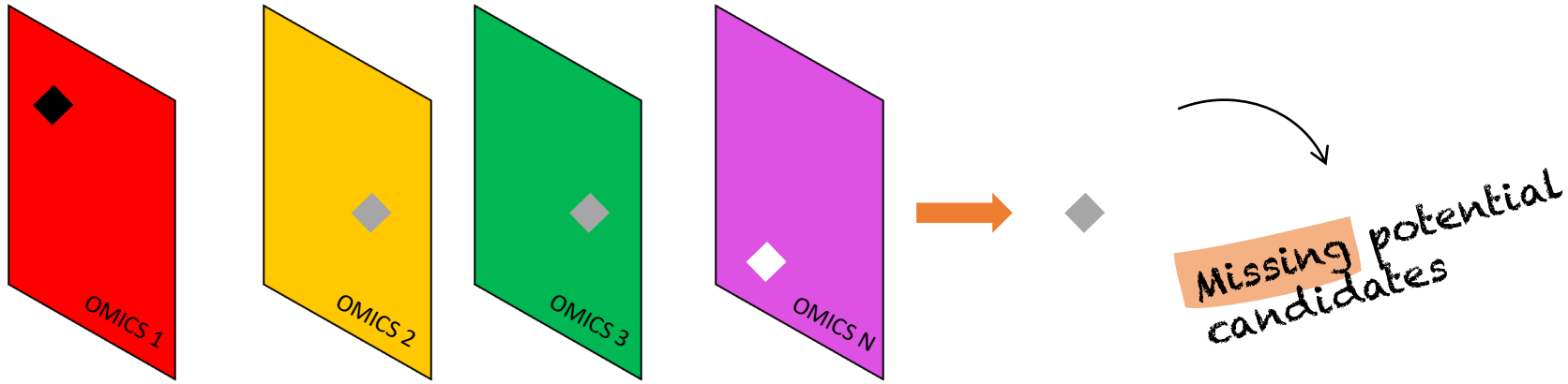
Multi-omics approach **without** integration



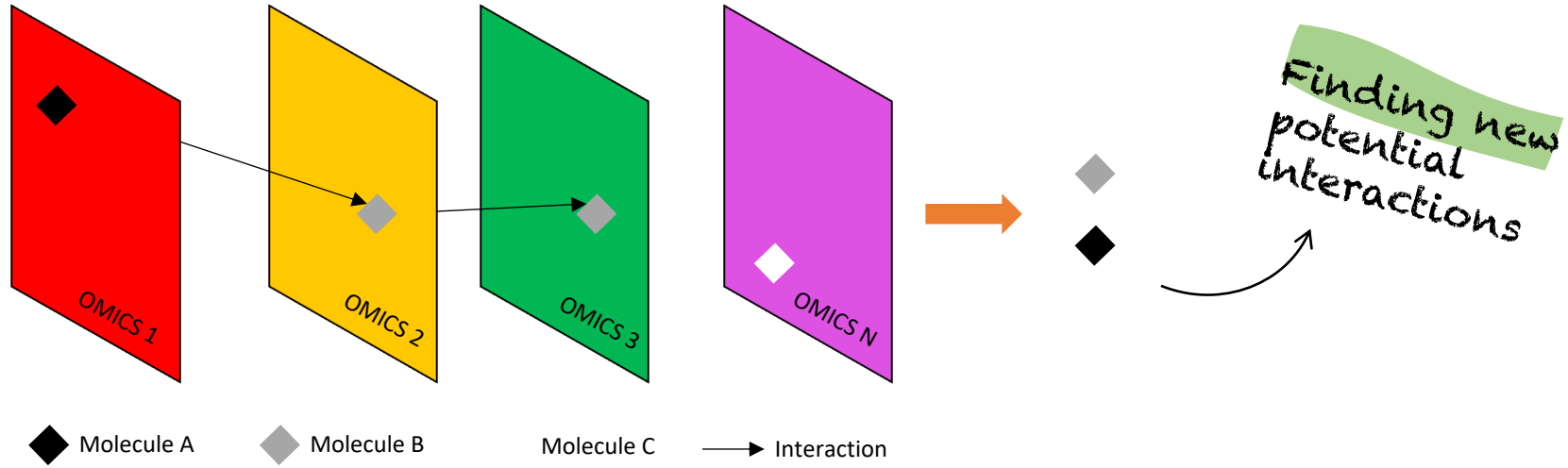
Molecule A Molecule B Molecule C Interaction

Why multi-omics integration?

Multi-omics approach **without** integration

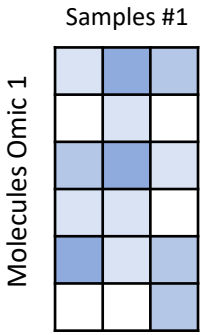


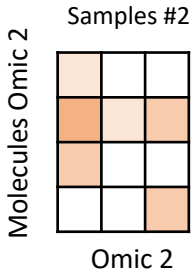
Multi-omics approach **with** integration



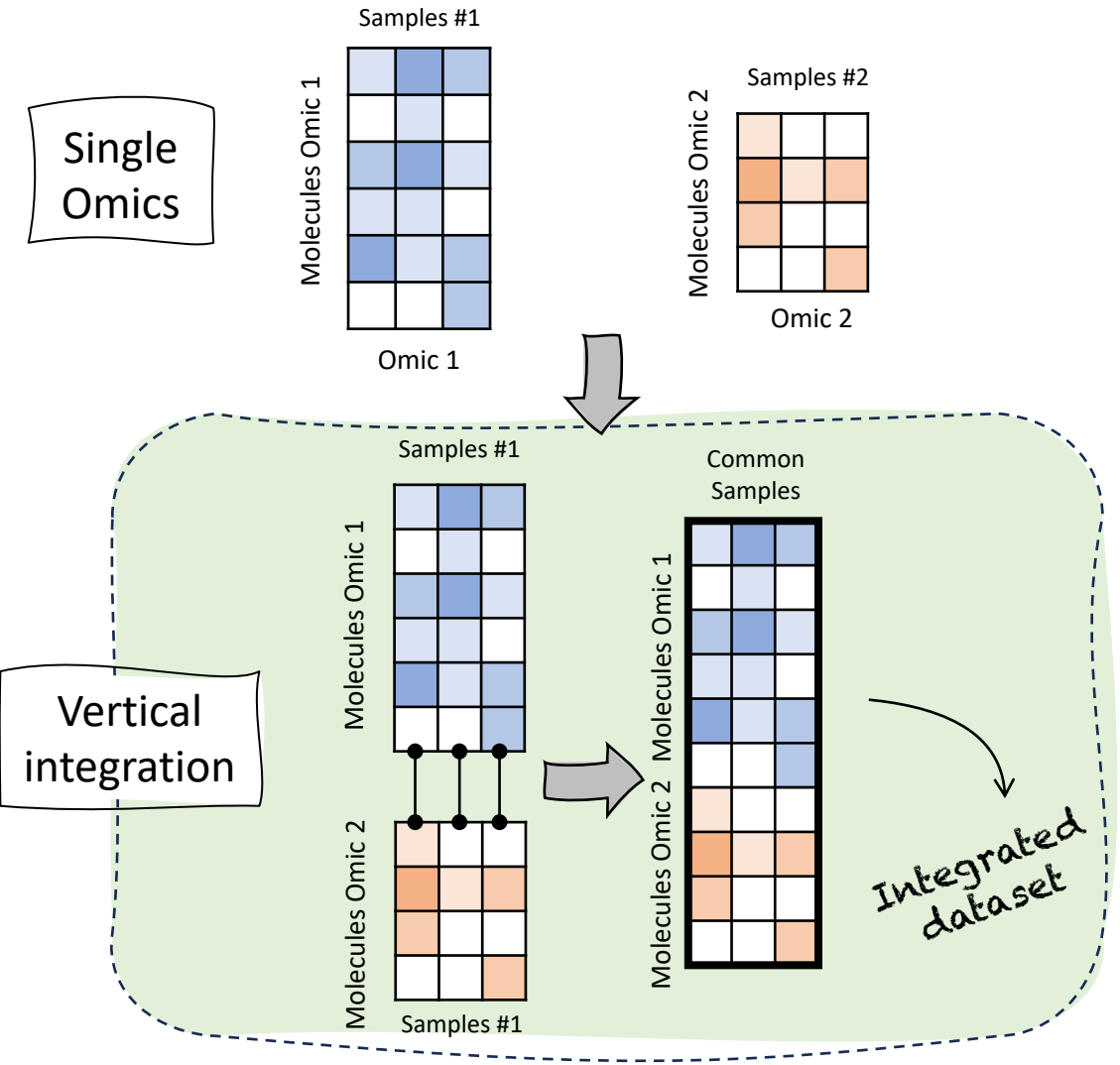
How to do multi-omics integration?

Single Omics

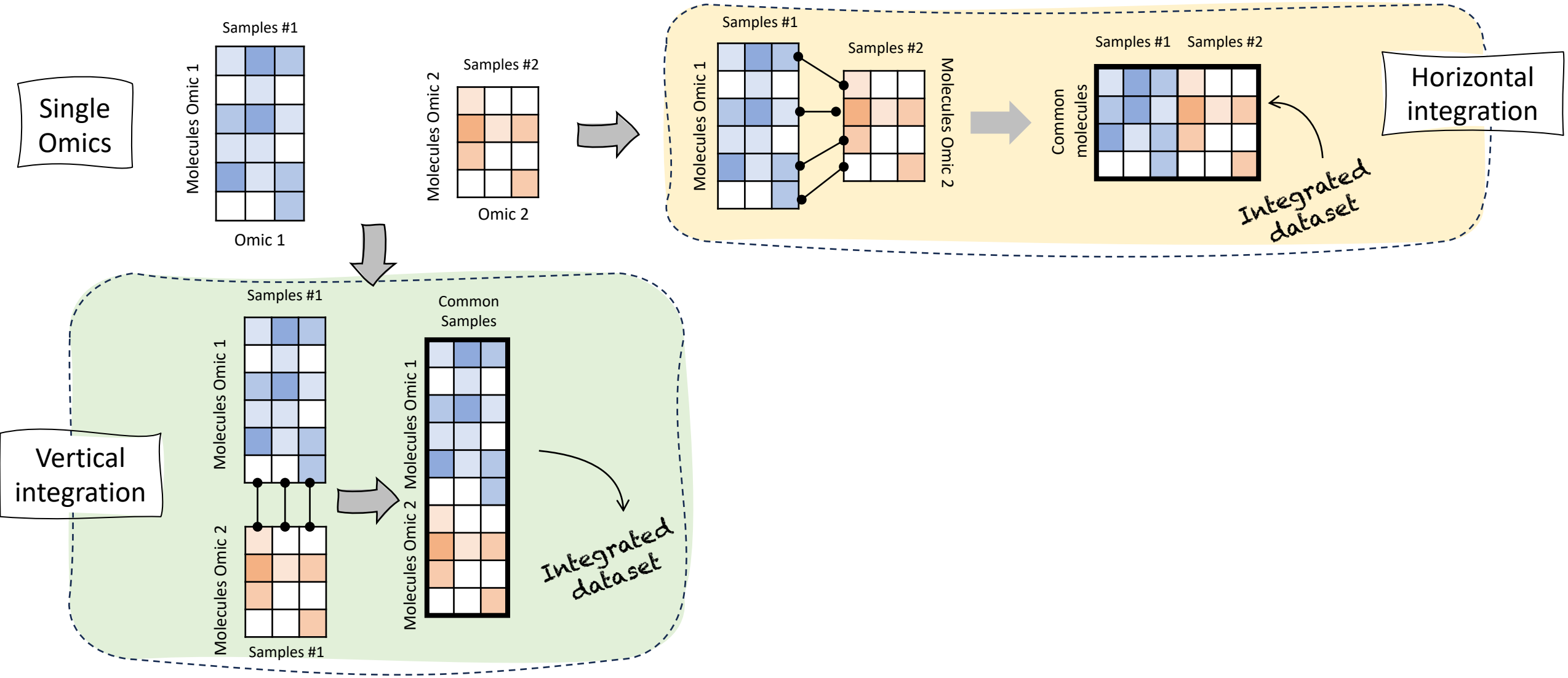




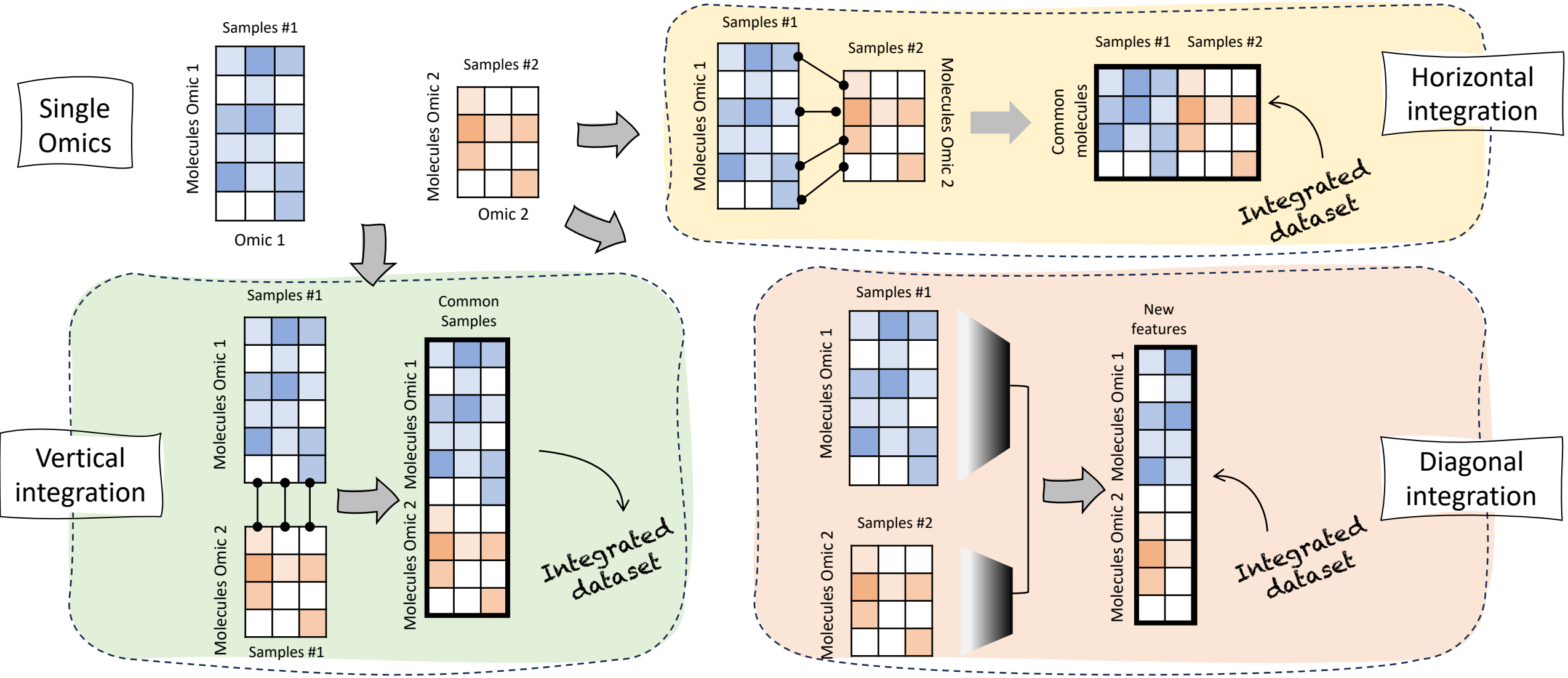
How to do multi-omics integration?



How to do multi-omics integration?



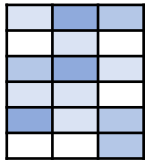
How to do multi-omics integration?



Challenges in multi-omics integration analysis

Silvia Bottini, junior professor chair - ISA UMR1355 - INRAE - CNRS - UniCA - silvia.bottini@inrae.fr

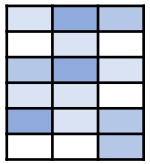
Challenges in multi-omics integration analysis



Heterogeneity, sparsity and uneven datasets.

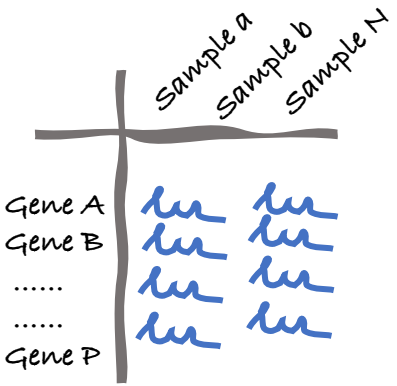
Silvia Bottini, junior professor chair - ISA UMR1355 - INRAE - CNRS - UniCA - silvia.bottini@inrae.fr

Challenges in multi-omics integration analysis



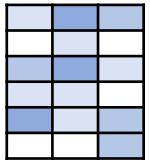
Heterogeneity, sparsity and uneven datasets.

More features than data (p >> n).

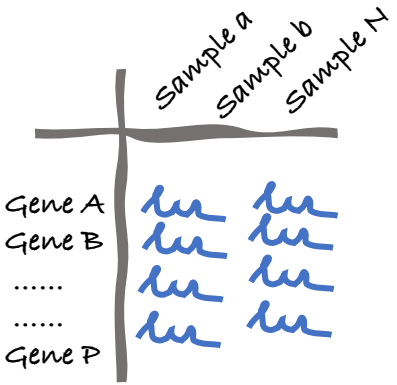


Silvia Bottini, junior professor chair - ISA UMR1355 - INRAE - CNRS - UniCA - silvia.bottini@inrae.fr

Challenges in multi-omics integration analysis



Heterogeneity, sparsity and uneven datasets.



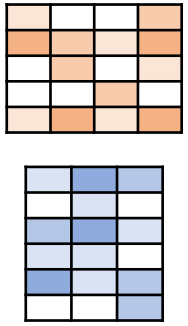
More features than data (p >> n).

NaN

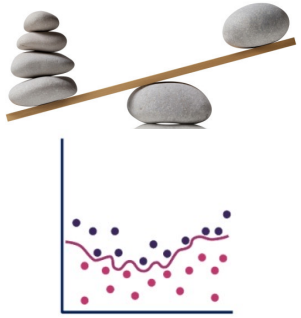
Missing data.

Silvia Bottini, junior professor chair - ISA UMR1355 - INRAE - CNRS - UniCA - silvia.bottini@inrae.fr

Challenges in multi-omics integration analysis

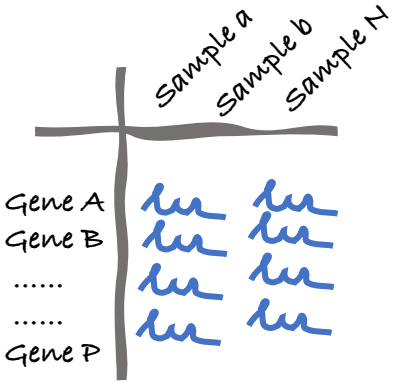


Heterogeneity, sparsity and uneven datasets.



Class imbalance and overfitting.

More features than data (p >> n).

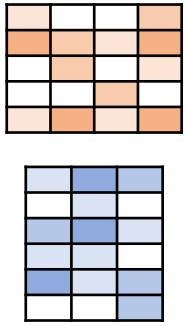


NaN

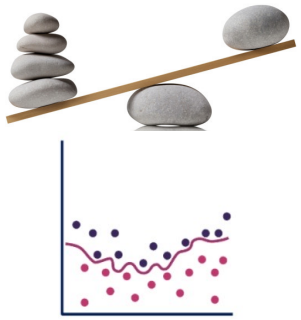
Missing data.

Silvia Bottini, junior professor chair - ISA UMR1355 - INRAE - CNRS - UniCA - silvia.bottini@inrae.fr

Challenges in multi-omics integration analysis

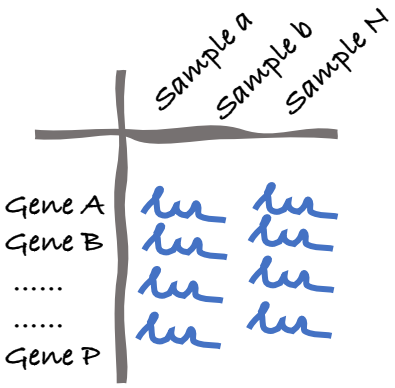


Heterogeneity, sparsity and uneven datasets.



Class imbalance and overfitting.

More features than data (p >> n).



How to choose a good model to perform the integration?

NaN

Missing data.

Silvia Bottini, junior professor chair - ISA UMR1355 - INRAE - CNRS - UniCA - silvia.bottini@inrae.fr

Multi-OMICS integration

Data-driven

Concatenation first

- Easy
- Straightforward

- Need proper normalization
- Do not take into account the different distribution of each omic
- Computationally intensive

Feature selection first

- Reduce computational costs

- Inter-omics relationships are lost
- Weak signals could be lost

Conversion first

- Allow easy representation of data

- May prevents the identification of indirect mechanisms
- Transformations can be challenging

Hypothesis-driven

Interactome first

- Allow inter-omics connections

- Strongly depends on existing knowledge

Pathway as network

- Incorporate pathway topology

- Hard to extend
- Computationally intense

Labory J, Bottini S. The multi-omics revolution in the era of deep learning: allies or enemies? 2023, invited book chapter, in press

Multi-OMICS integration

Data-driven

Concatenation first

- Easy
- Straightforward

- Need proper normalization
- Do not take into account the different distribution of each omic
- Computationally intensive

Hypothesis-driven

Pathway as network

- Incorporate pathway topology

- Hard to extend
- Computationally intense

Model choice depends on the available data and the biological question.

Peanut – *Arachis spp.*

- Family: fabaceae
- Native: South America
- Important legume in Africa and Asia
- High protein content

Wild species

- More than 80 species
- Originated exclusively from South America
- Important to study as a source of resistance characteristics to biotic and abiotic stresses.

Giulia Calia
PhD student

Ana Zotta Mota
Post-doc

Patricia Messenberg Guimaraes

Biological model: two wild peanuts species

Arachis duranensis

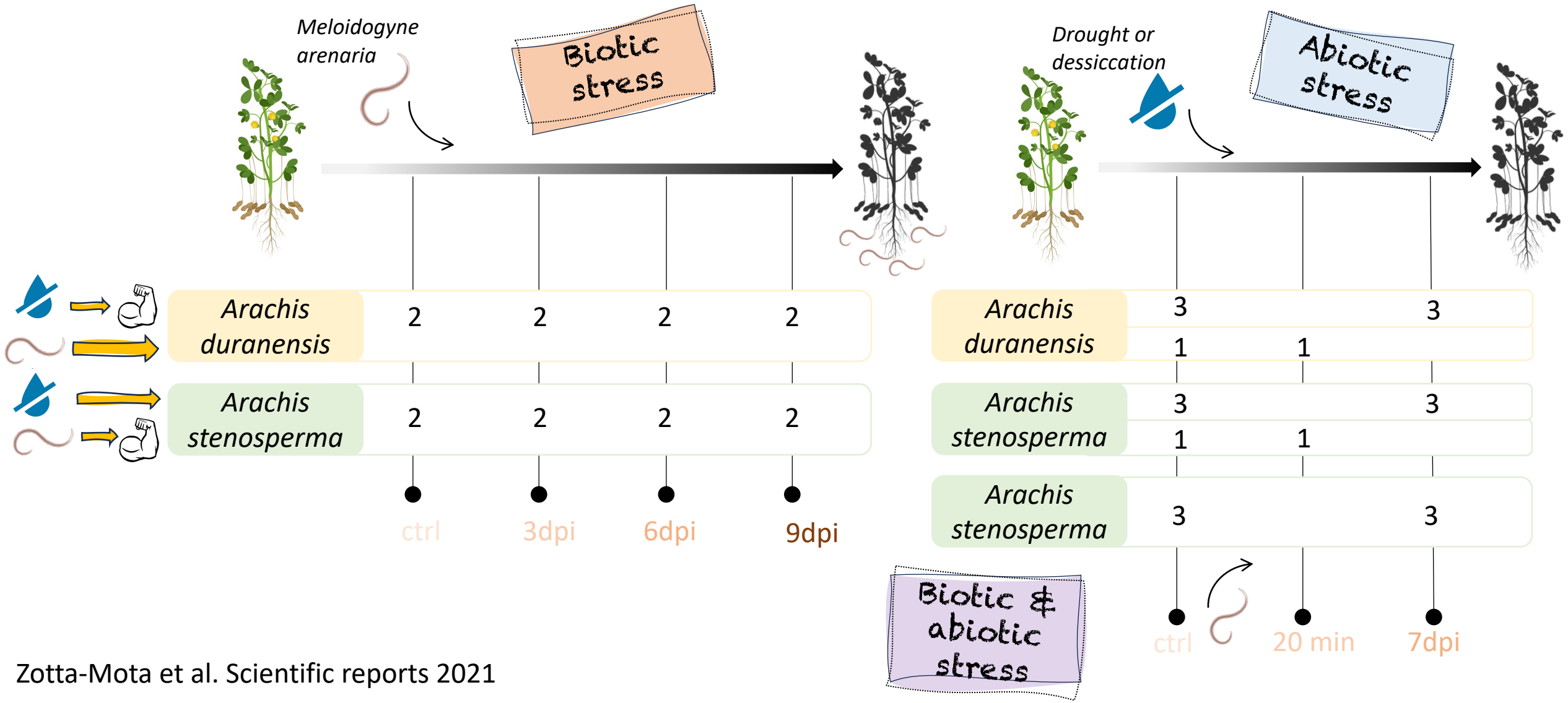
Arachis stenosperma

Abiotic stress:
drought

Biotic stress:
root knot nematodes
Meloydogine arenaria

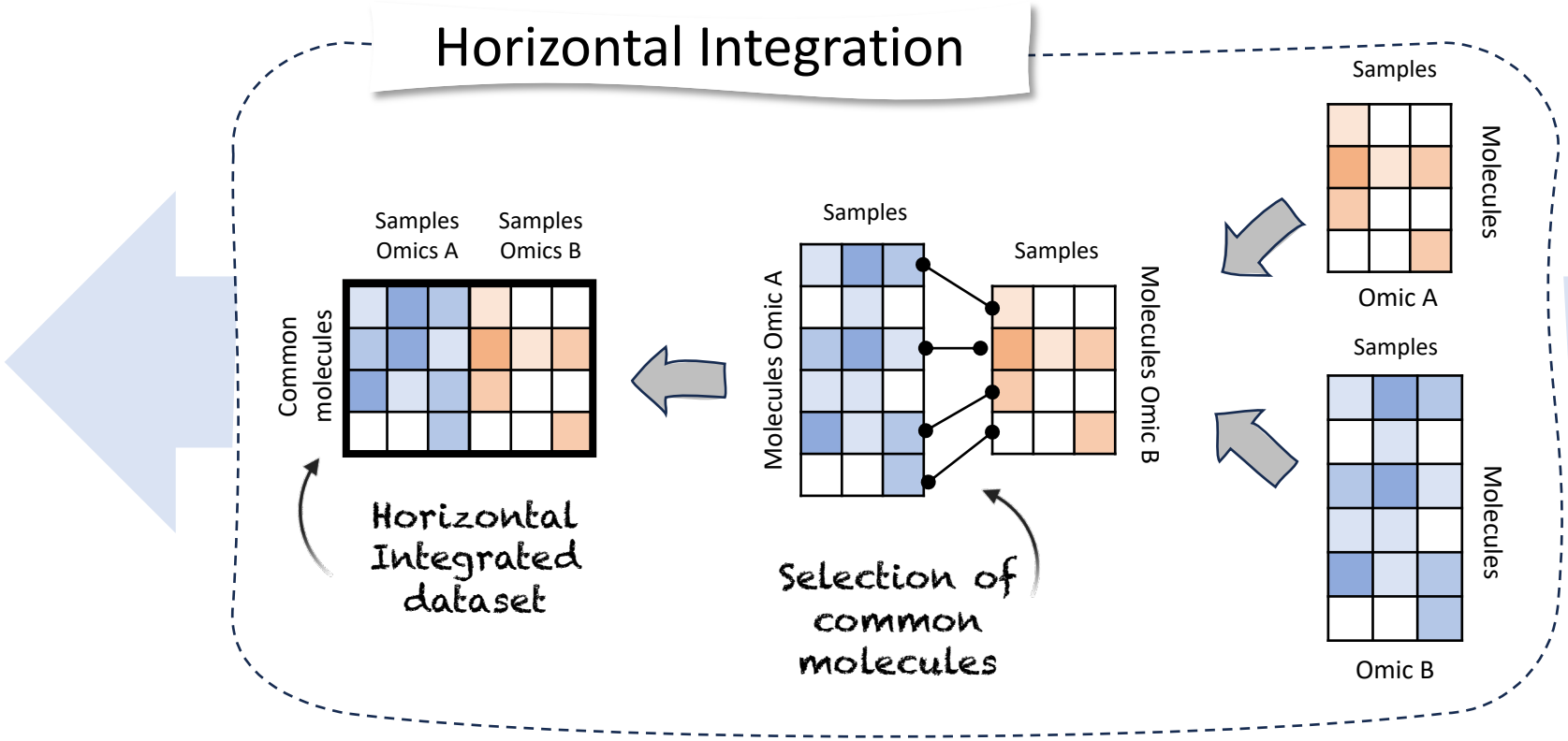
Silvia Bottini, junior professor chair - ISA UMR1355 - INRAE - CNRS - UniCA - silvia.bottini@inrae.fr

The experimental design

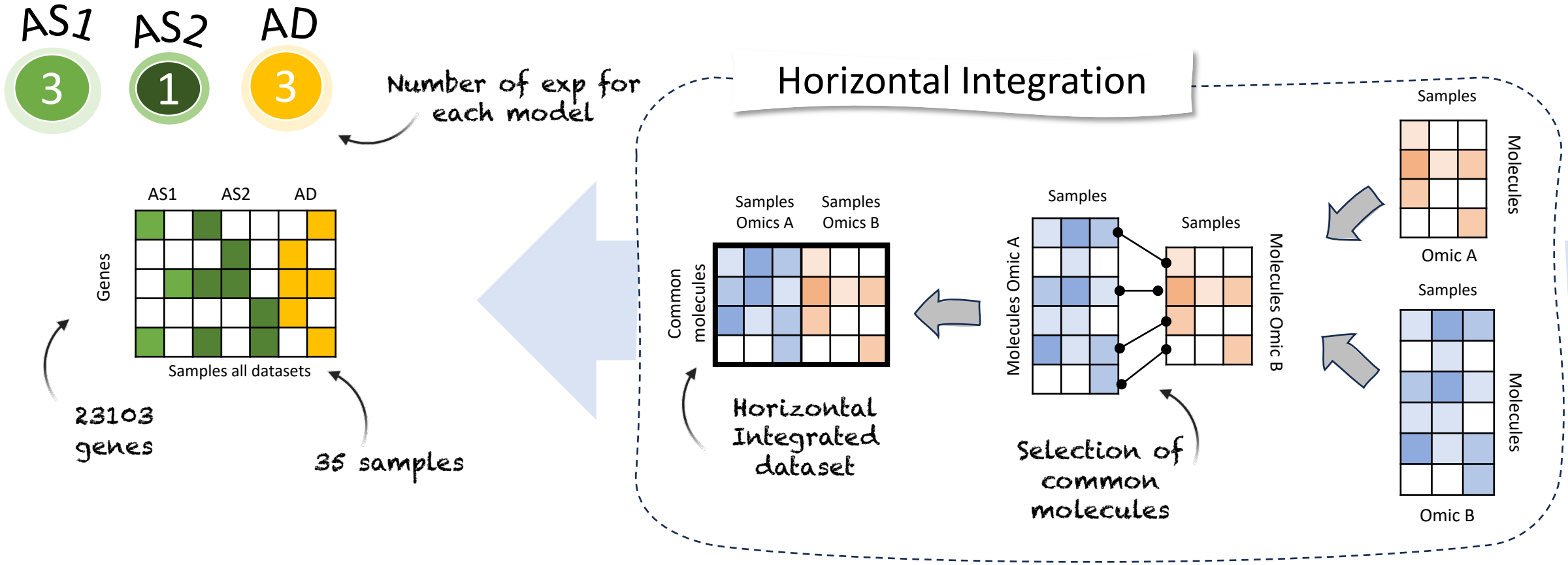


Zotta-Mota et al. Scientific reports 2021

The pilot integrated dataset



The pilot integrated dataset



AS1: A. astenosperma #1

AS2: A. astenosperma #2

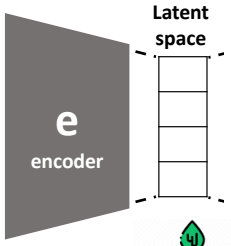
AD: A. duranensis

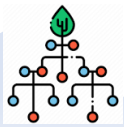
HIVE: a general framework to analyse integrated multi-omics data

Secret code
=
Multi-omics data

HIVE

Variational Autoencoder





Random forest regression (RFR)

explainer

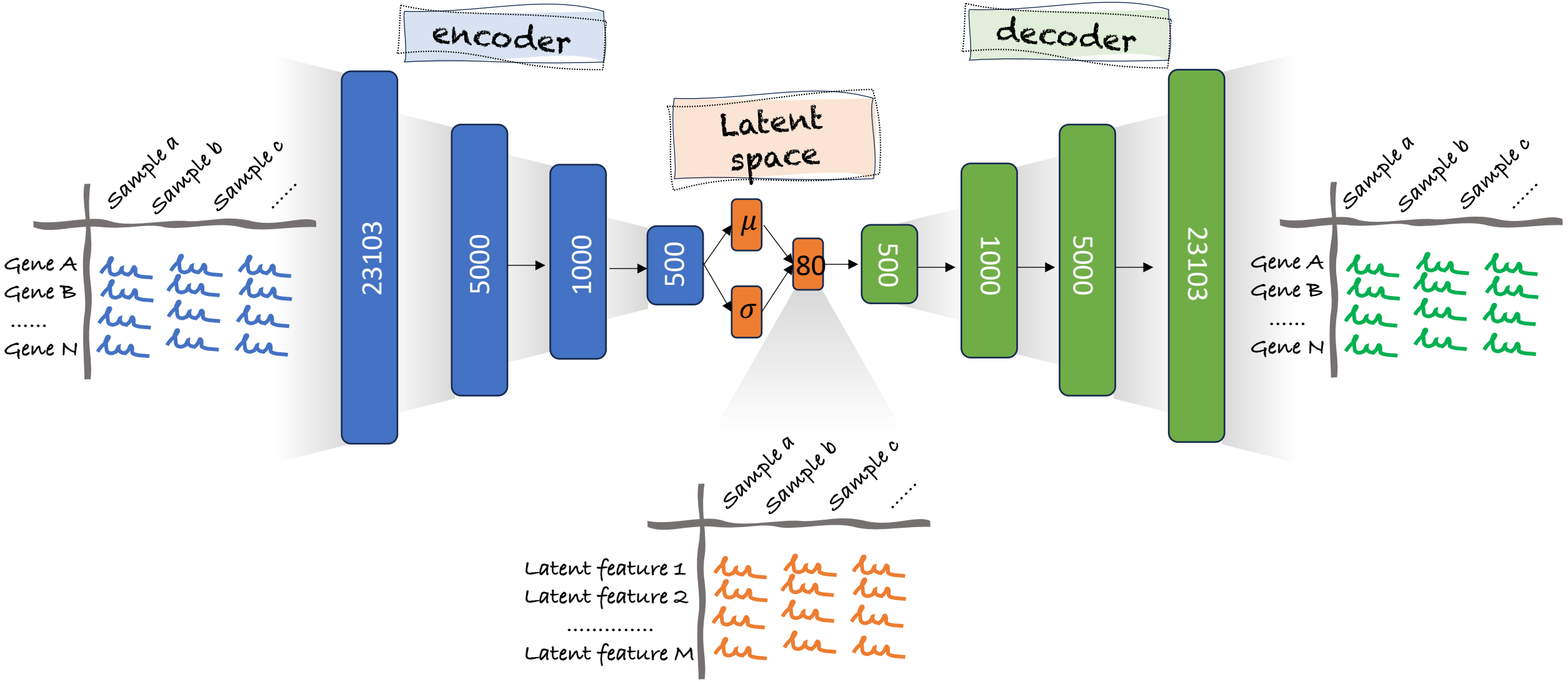
SHAP

Decrypted code
=
Latent features

Molecules contribution to each latent feature
=
The key to convert the secret code

Silvia Bottini, junior professor chair - ISA UMR1355 - INRAE - CNRS - UniCA - silvia.bottini@inrae.fr

Variational autoencoder



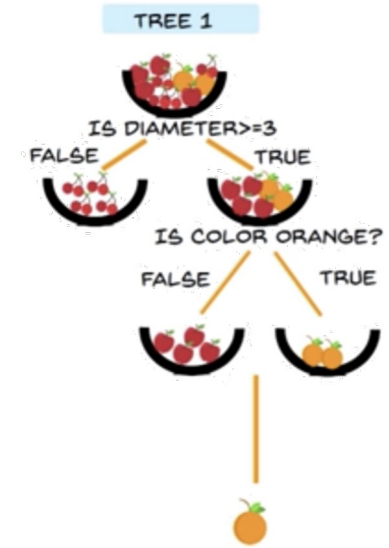
Random forest

The random forest algorithm makes use of multiple decision trees.
It can solve both regression and classification problems.

Silvia Bottini, junior professor chair - ISA UMR1355 - INRAE - CNRS - UniCA - silvia.bottini@inrae.fr

Random forest

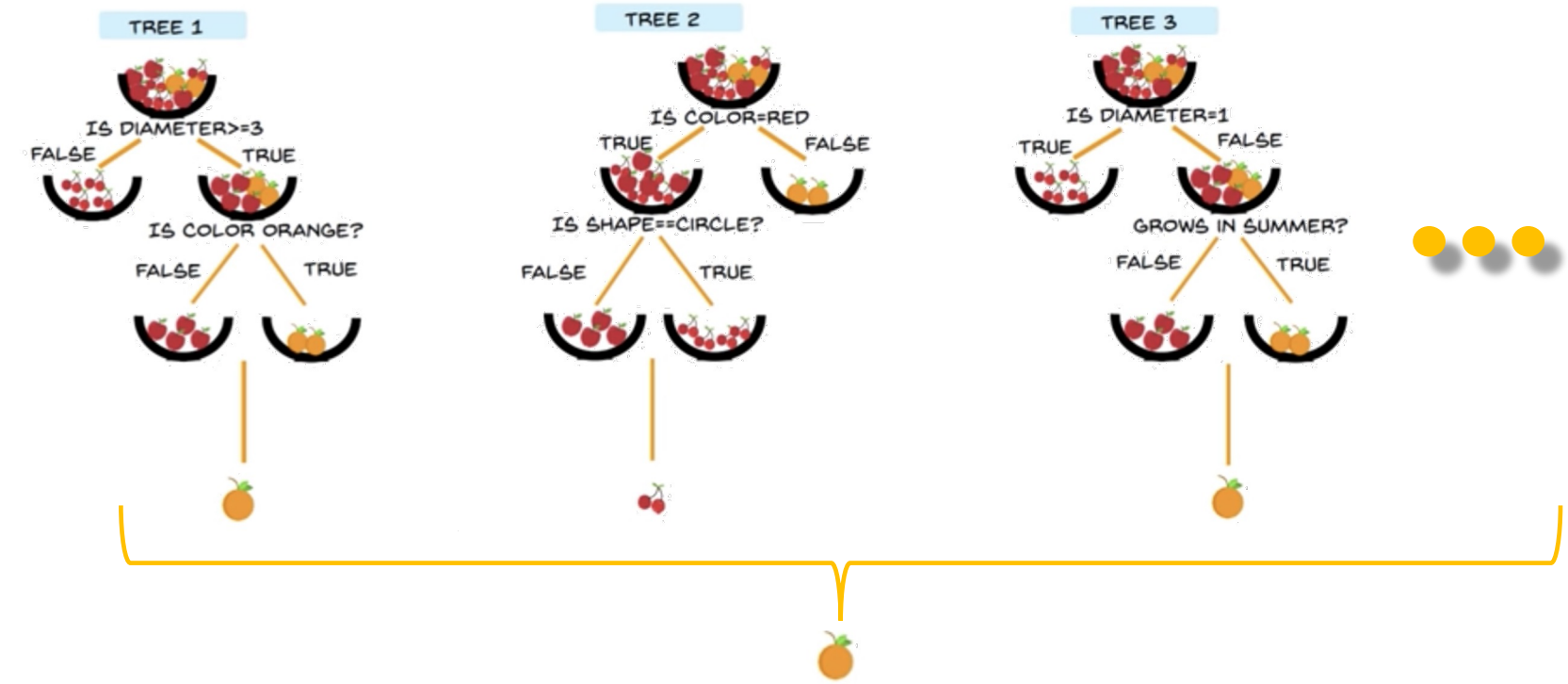
The random forest algorithm makes use of multiple decision trees.
It can solve both regression and classification problems.



Silvia Bottini, junior professor chair - ISA UMR1355 - INRAE - CNRS - UniCA - silvia.bottini@inrae.fr

Random forest

The random forest algorithm makes use of multiple decision trees.
It can solve both regression and classification problems.



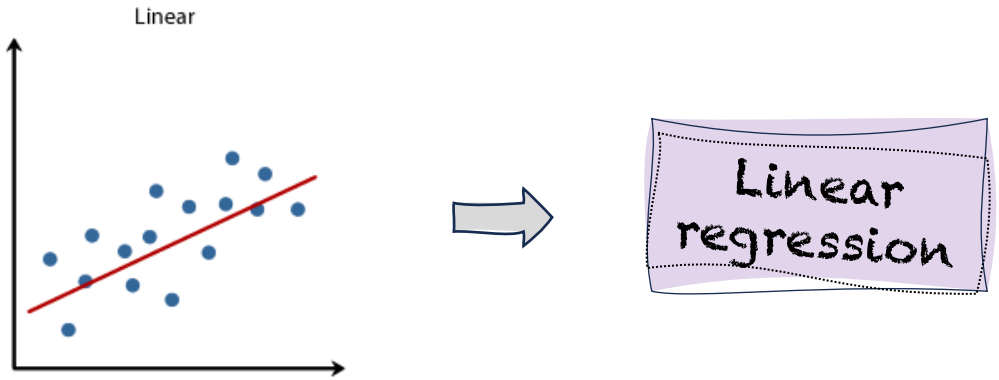
Robustness

Computationally efficient

Immune to the curse of dimensionality

Silvia Bottini, junior professor chair - ISA UMR1355 - INRAE - CNRS - UniCA - silvia.bottini@inrae.fr

Random forest regression



Linear regression

coefficients

noise

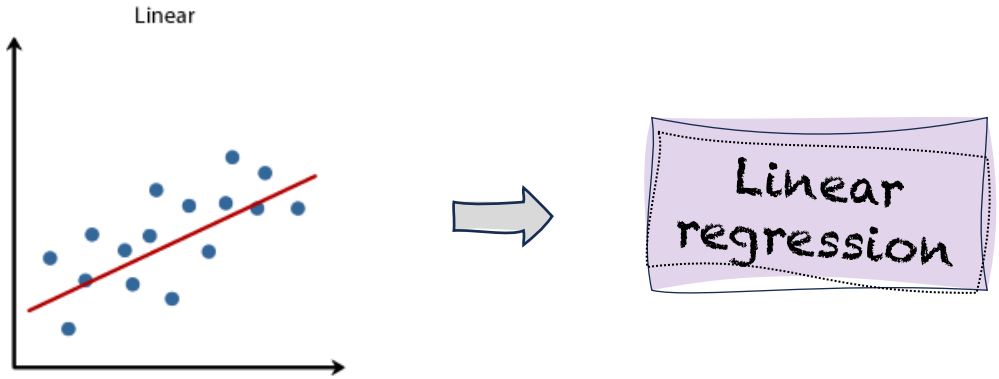
Response, dependent variable

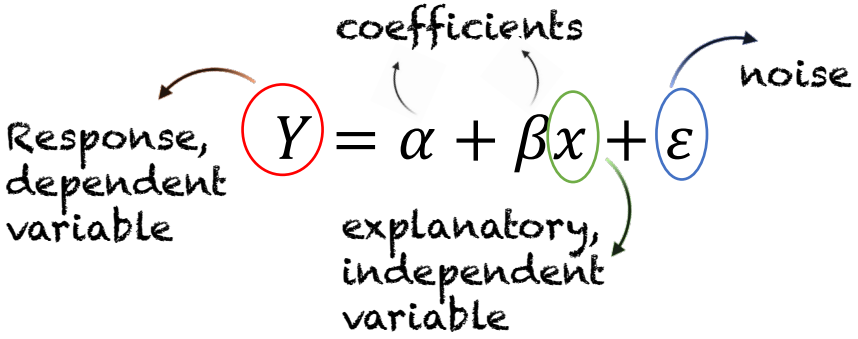
$Y = \alpha + \beta x + \varepsilon$

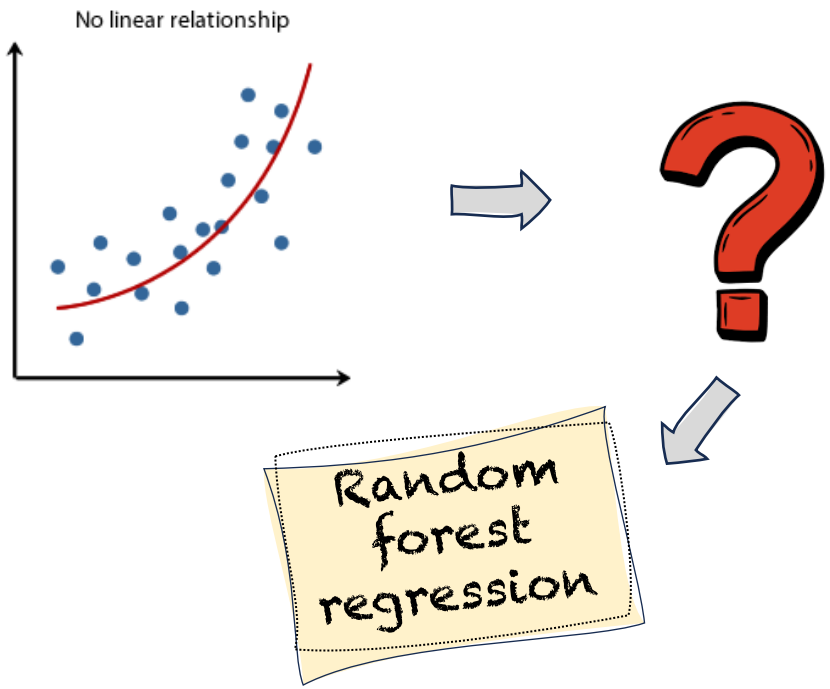
explanatory, independent variable

- Large sample size and low variance in the data set.
- Assumptions about the relations between predictors and target variables, such as constant variance, normality of errors, etc.

Random forest regression







- Large sample size and low variance in the data set.
- Assumptions about the relations between predictors and target variables, such as constant variance, normality of errors, etc.

- Does not make any assumption about the underlying data distribution.
- Handle nonlinear relationships better.
- Less prone to overfitting.

The Shapley value

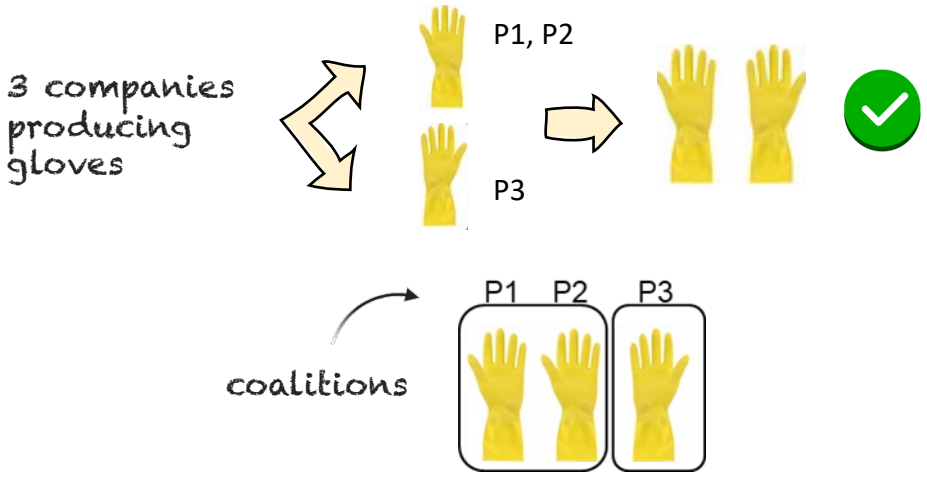
Def. In a *cooperative* game, “players” have the possibility to forge coalitions to achieve a common goal. One difficulty in the theory of cooperative games is the distribution of benefits among the players.

Lloyd Shapley, introduced the concept in 1953 and received the Nobel Prize in Economics in 2012.

Silvia Bottini, junior professor chair - ISA UMR1355 - INRAE - CNRS - UniCA - silvia.bottini@inrae.fr

The Shapley value

Def. In a *cooperative* game, "players" have the possibility to forge coalitions to achieve a common goal. One difficulty in the theory of cooperative games is the distribution of benefits among the players.



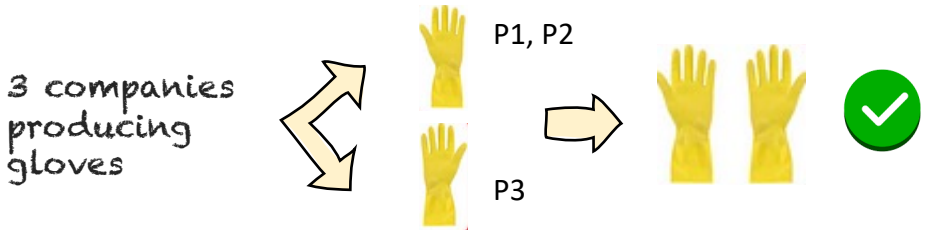
Lloyd Shapley, introduced the concept in 1953 and received the Nobel Prize in Economics in 2012.

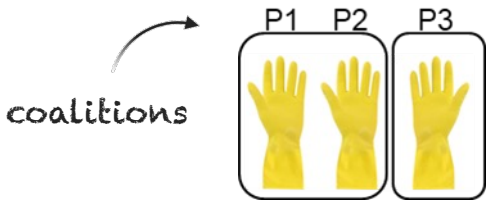
Silvia Bottini, junior professor chair - ISA UMR1355 - INRAE - CNRS - UniCA - silvia.bottini@inrae.fr

The Shapley value

Def. In a *cooperative* game, "players" have the possibility to forge coalitions to achieve a common goal. One difficulty in the theory of cooperative games is the distribution of benefits among the players.

Lloyd Shapley, introduced the concept in 1953 and received the Nobel Prize in Economics in 2012.





More formally:

$$v(P1) = v(P2) = v(P3) = 0$$

$$v(P1, P2) = 0$$

$$v(P1, P3) = v(P2, P3) = v(P1, P2, P3) = 1$$

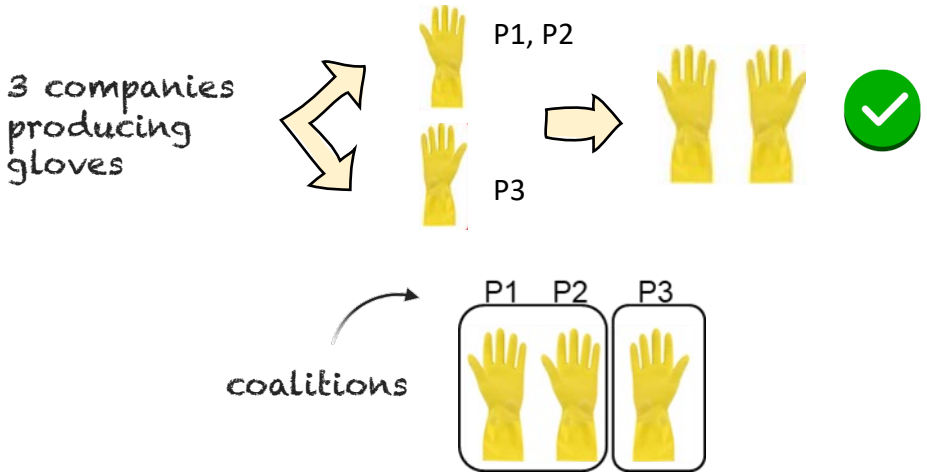
Maximum gain

Silvia Bottini, junior professor chair - ISA UMR1355 - INRAE - CNRS - UniCA - silvia.bottini@inrae.fr

The Shapley value

Def. In a *cooperative* game, "players" have the possibility to forge coalitions to achieve a common goal. One difficulty in the theory of cooperative games is the distribution of benefits among the players.

Lloyd Shapley, introduced the concept in 1953 and received the Nobel Prize in Economics in 2012.



How can the revenue be redistributed fairly?

	Left-handed glove		Right-handed glove
σ	**P1**	**P2**	**P3**
$P1, P2, P3$	0	0	1
$P1, P3, P2$	0	0	1
$P2, P1, P3$	0	0	1
$P2, P3, P1$	0	0	1
$P3, P1, P2$	1	0	0
$P3, P2, P1$	0	1	0
$\emptyset_i(N, v)$	**1/6**	**1/6**	**4/6**

More formally:

$$v(P1) = v(P2) = v(P3) = 0$$

$$v(P1, P2) = 0$$

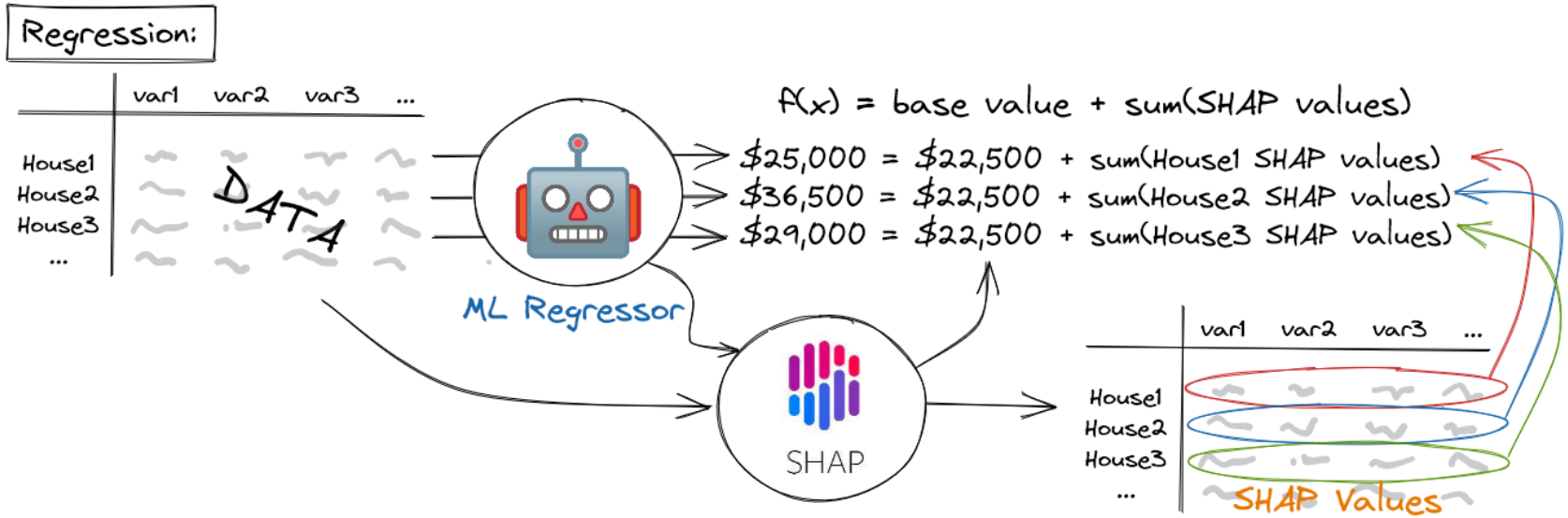
$$v(P1, P3) = v(P2, P3) = v(P1, P2, P3) = 1$$

Maximum gain

Silvia Bottini, junior professor chair - ISA UMR1355 - INRAE - CNRS - UniCA - silvia.bottini@inrae.fr

SHAP explainer

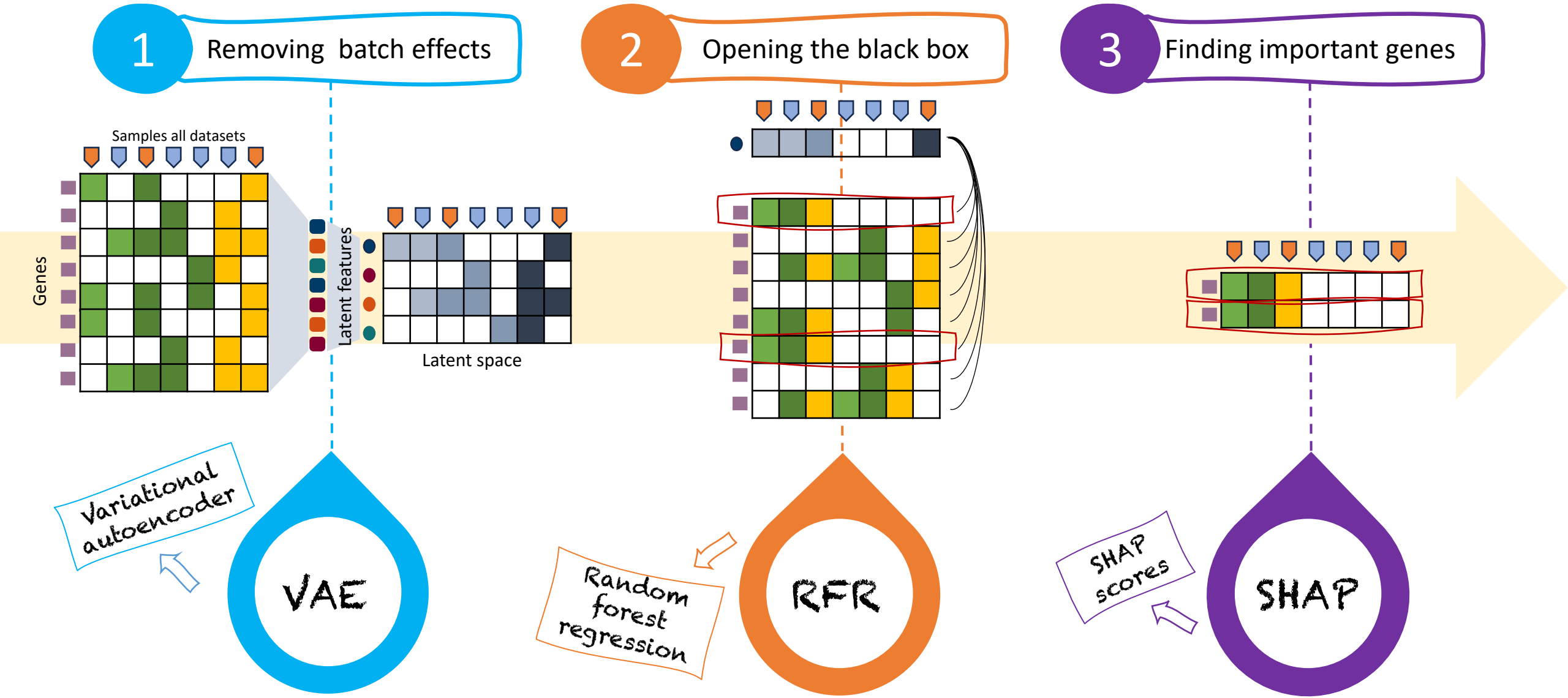
SHAP is an explainable artificial intelligence technique based on Shapley value and used in machine learning to determine how input variables contribute to output predictions.



- SHAP can be applied to any machine learning model as a *post hoc* interpretation technique
- it is particularly efficient to compute SHAP for tree-based models, such as random forests and gradient boosted trees.

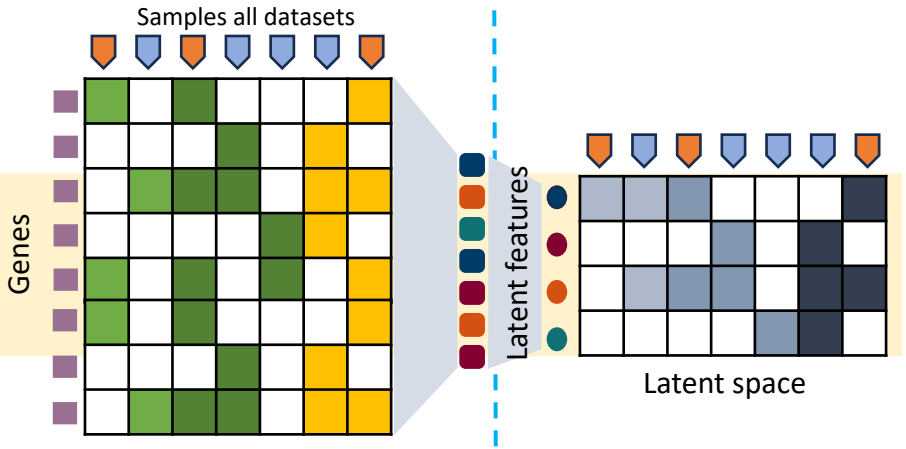
Silvia Bottini, junior professor chair - ISA UMR1355 - INRAE - CNRS - UniCA - silvia.bottini@inrae.fr

HIVE pipeline



HIVE pipeline

1 Removing batch effects



Variational autoencoder

VAE

2 Opening the black box

Random forest regression

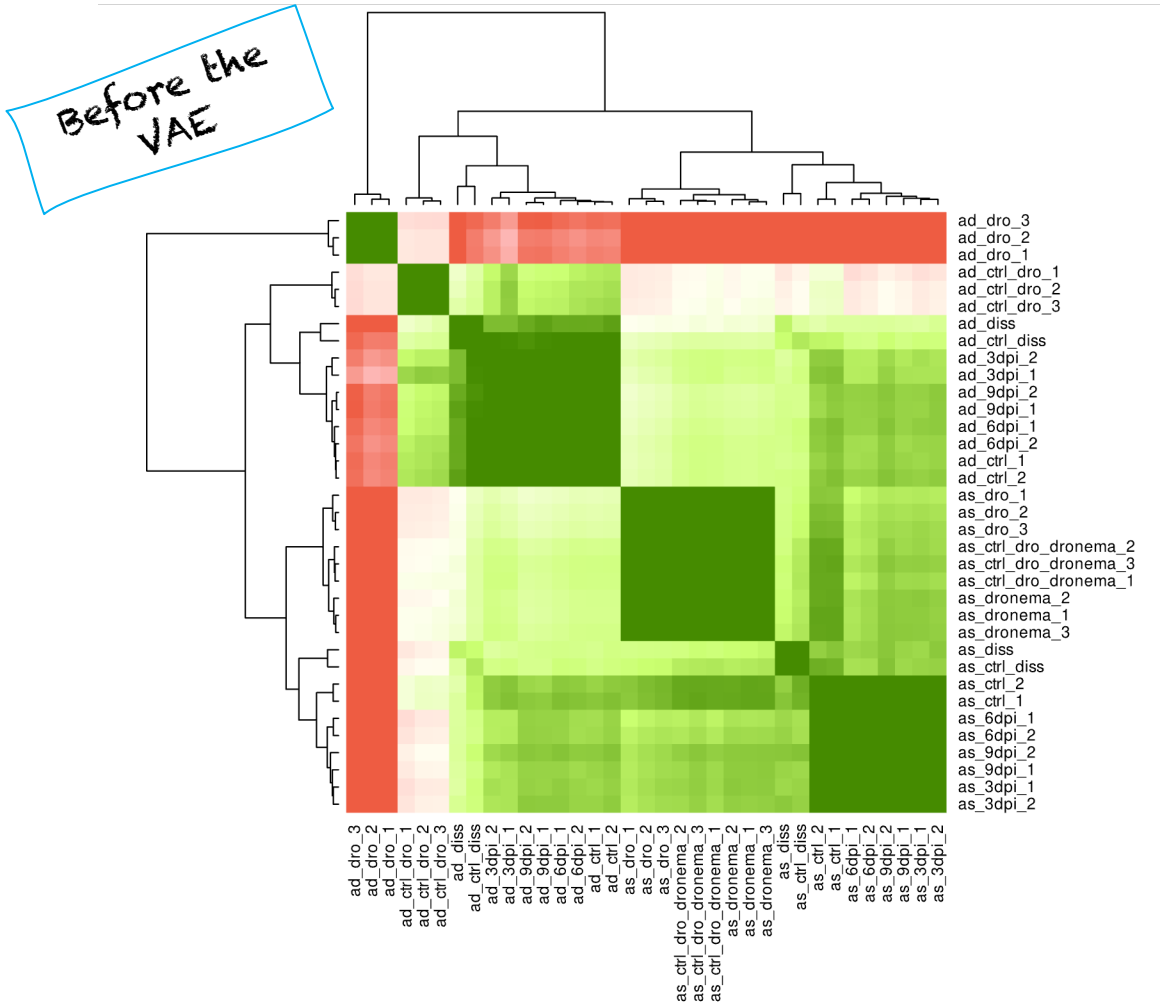
RFR

3 Finding important genes

SHAP scores

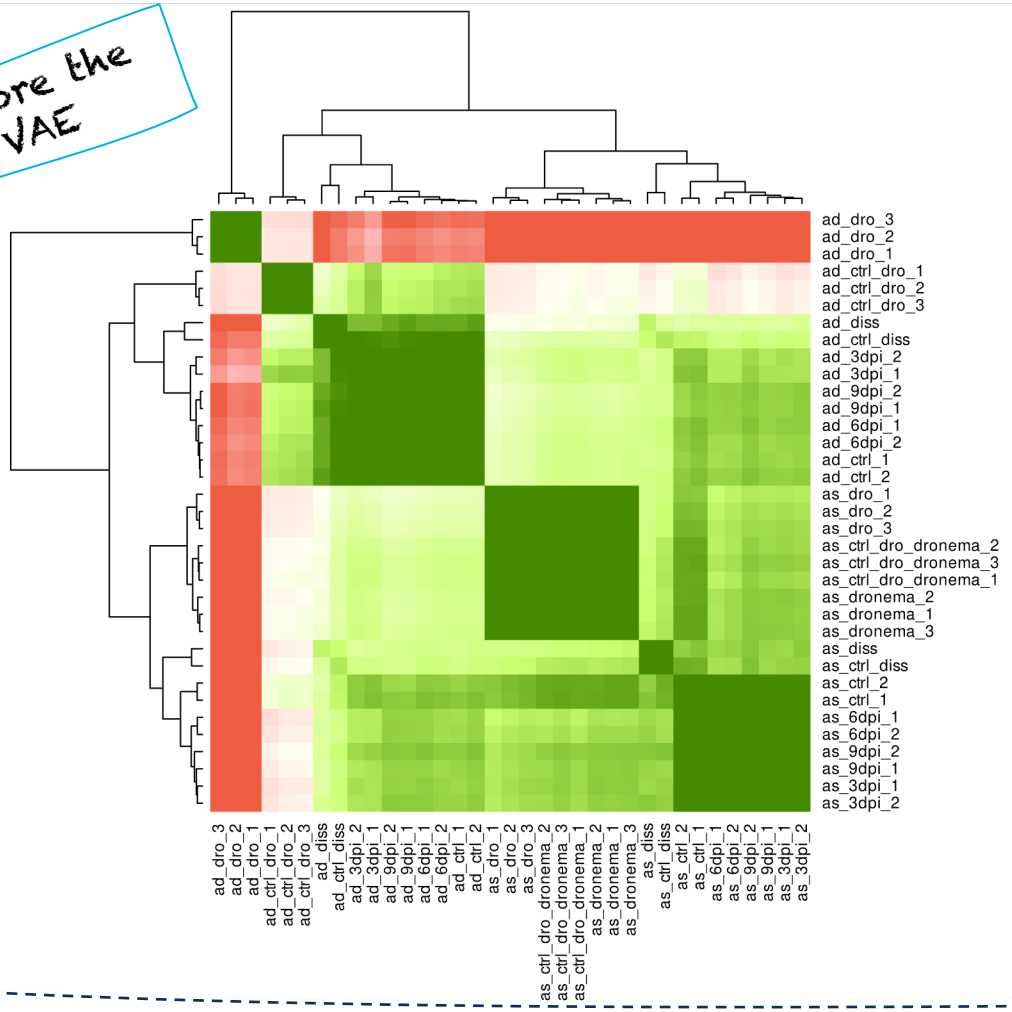
SHAP

Why do we need the VAE?

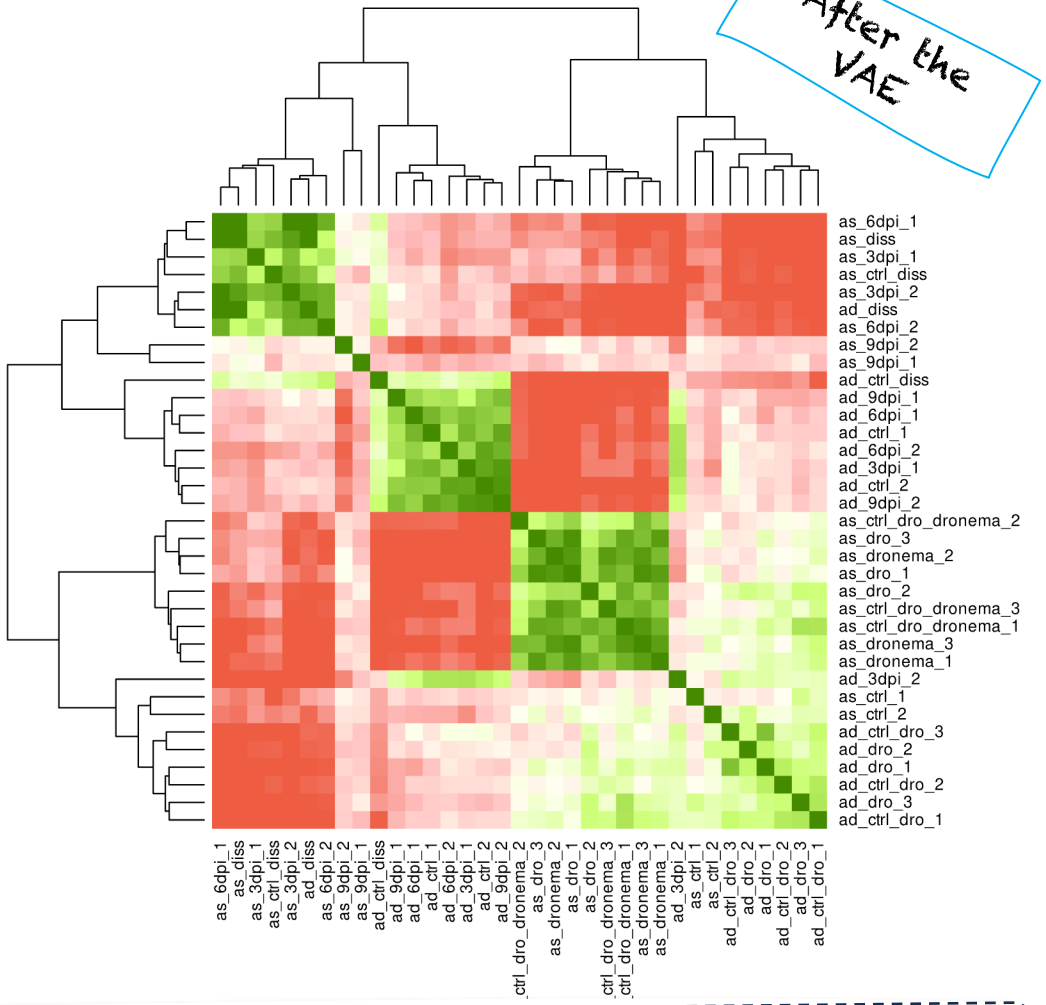


Why do we need the VAE?

Before the VAE



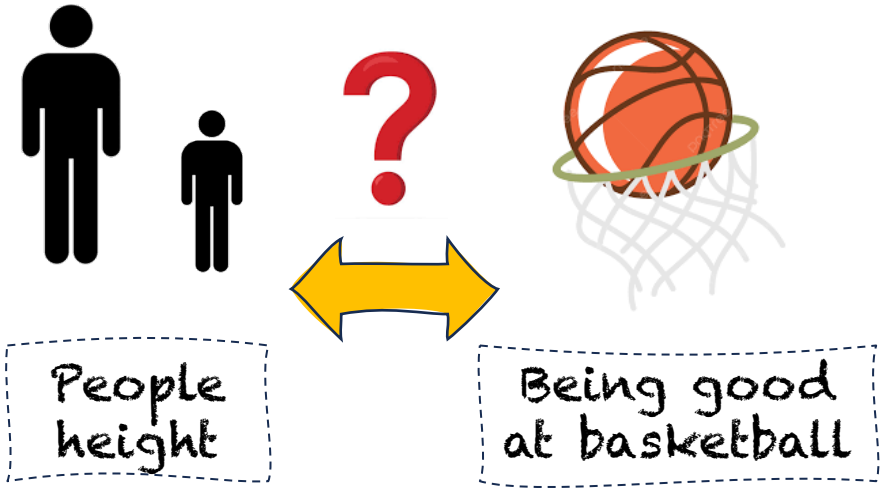
After the VAE

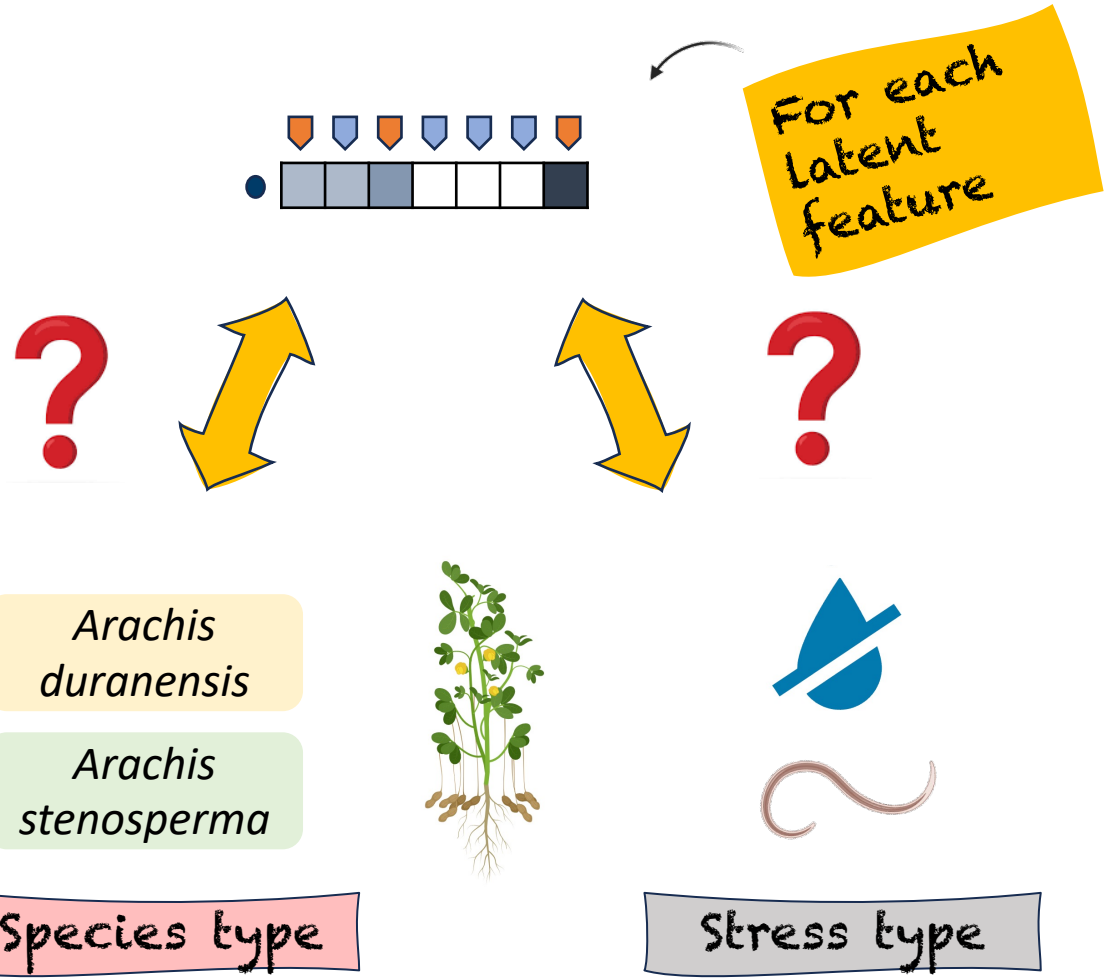


The VAE reduces the batch effects allowing unpaired experiments integrated analysis.

Which is the meaning of the latent features?

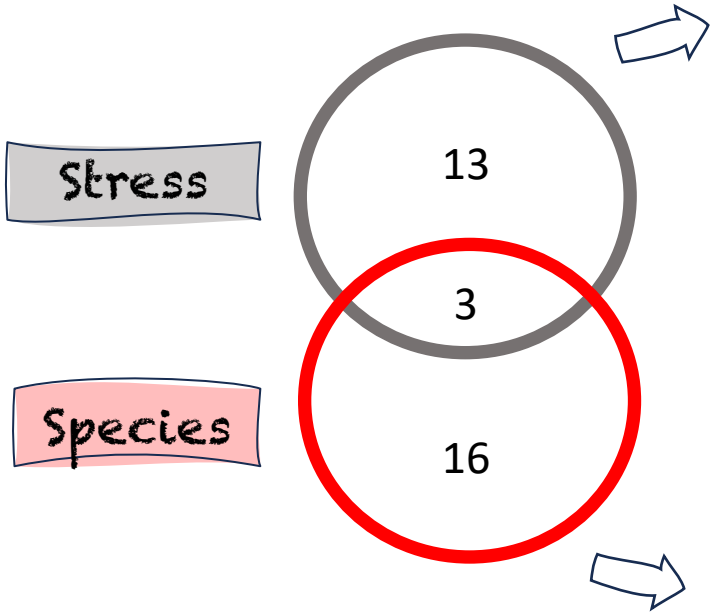
Point bi-serial correlation is a special case of the Pearson Correlation and is used when you want to measure the relationship between a continuous variable and a dichotomous (or binary) variable.

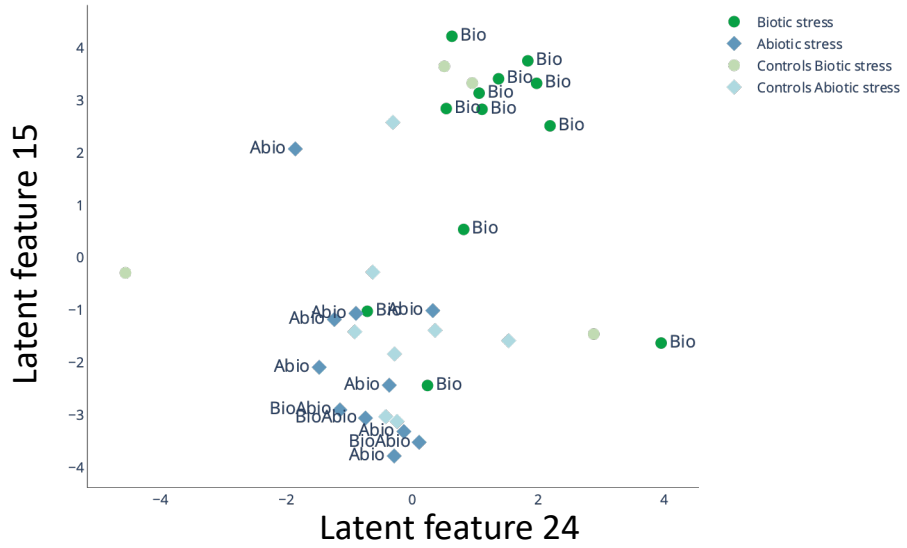


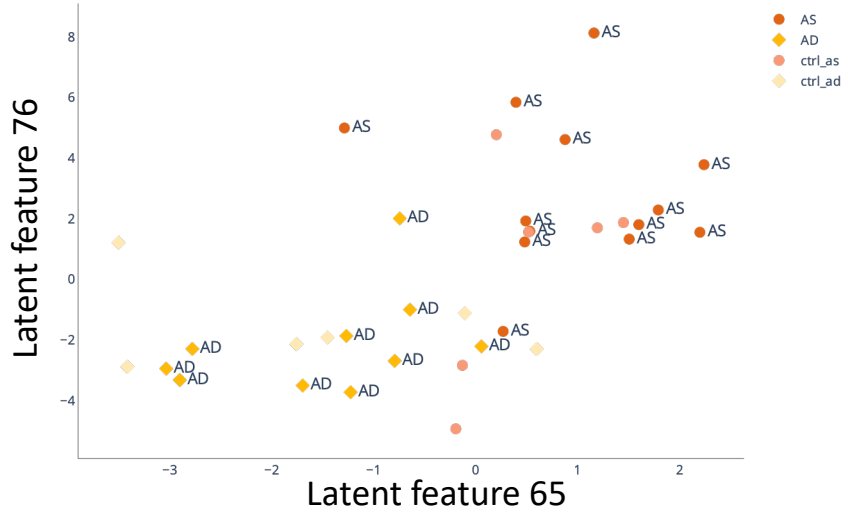


Silvia Bottini, junior professor chair - ISA UMR1355 - INRAE - CNRS - UniCA - silvia.bottini@inrae.fr

Latent features are associated with phenotypic characteristics

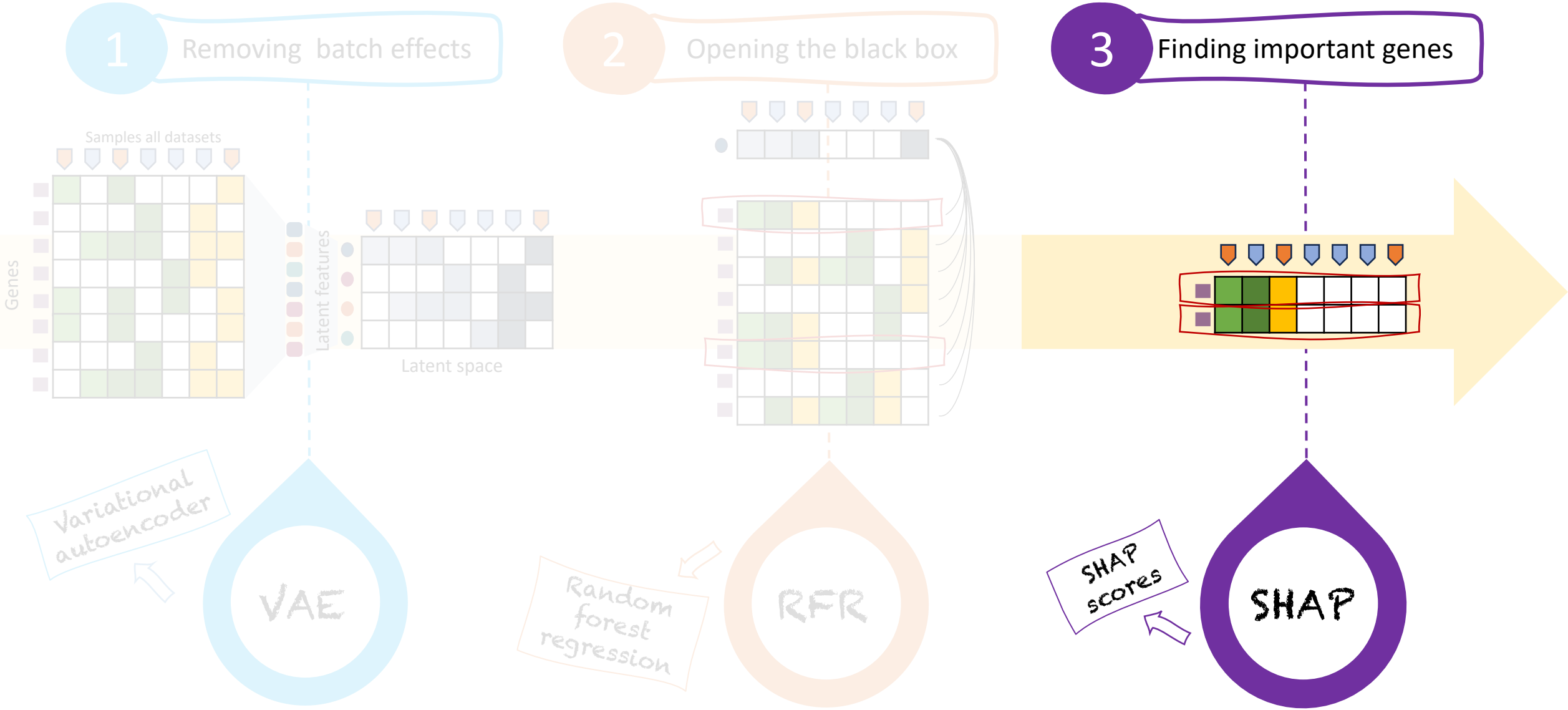




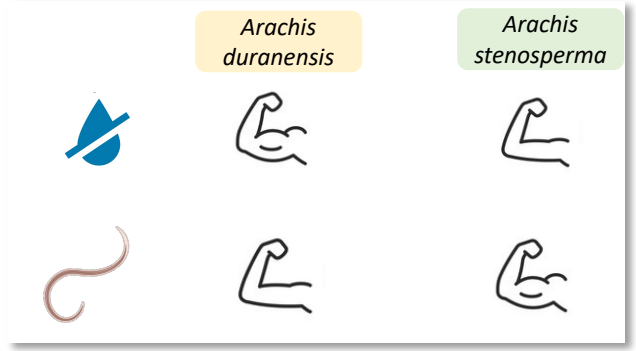


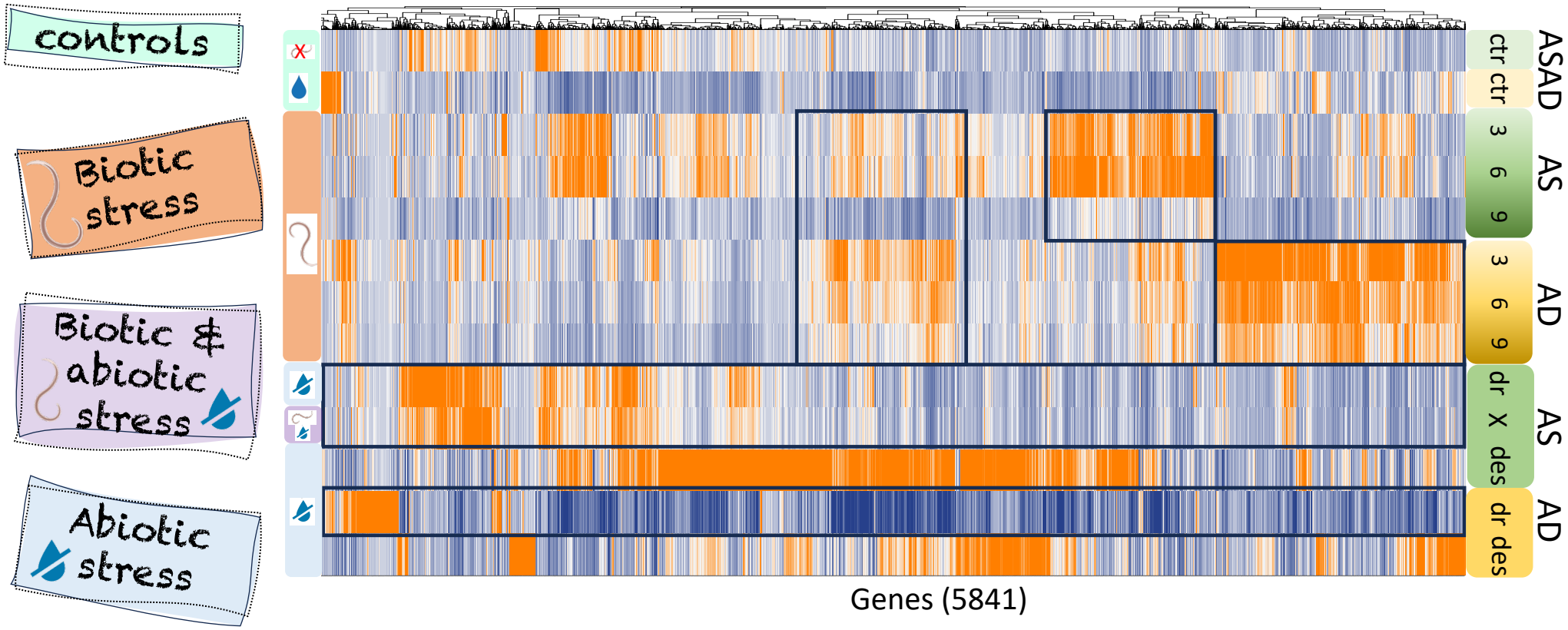
Top 2 associated latent features

HIVE pipeline

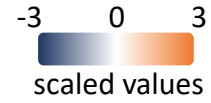


Global expression profiles of genes selected by HIVE

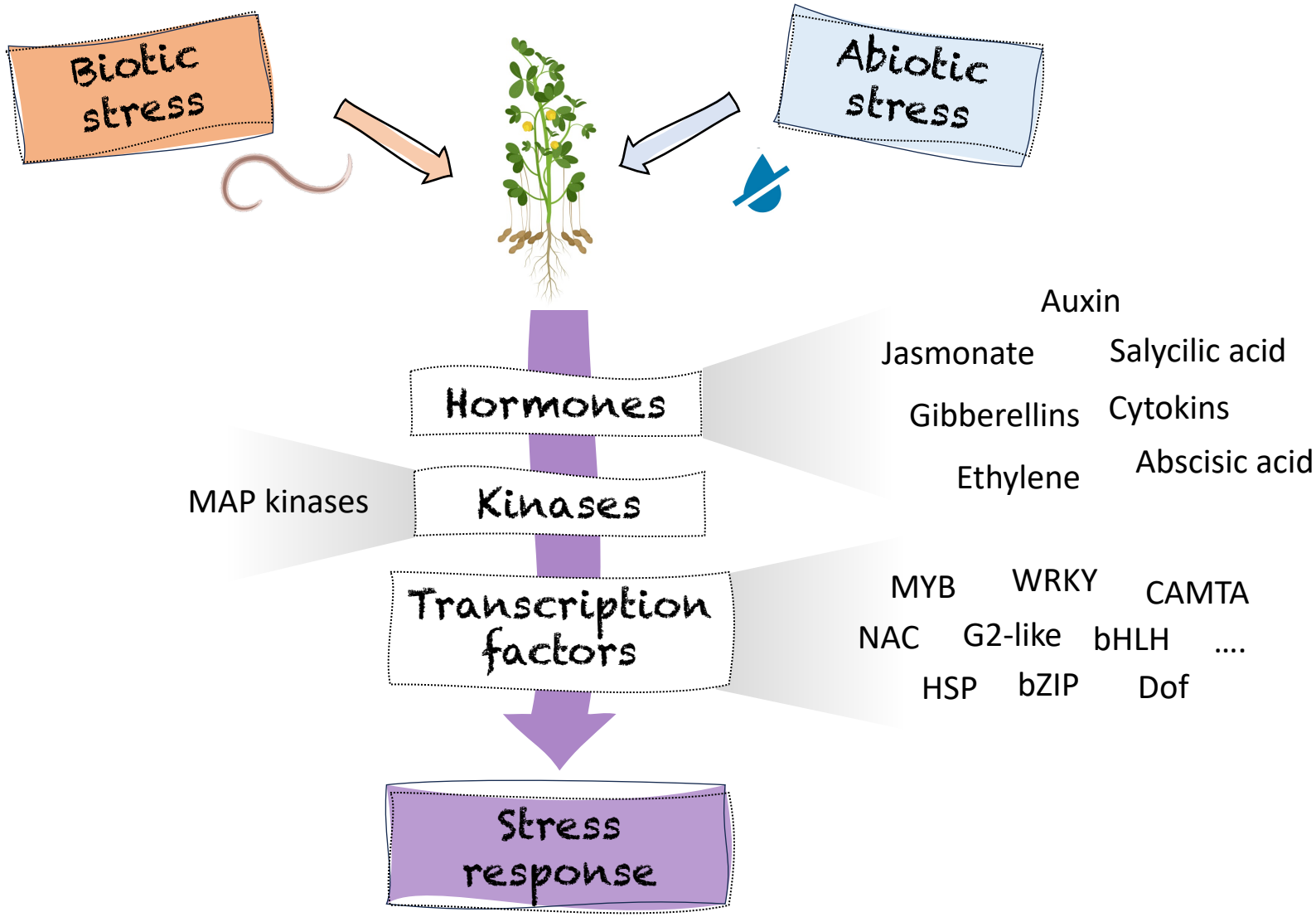




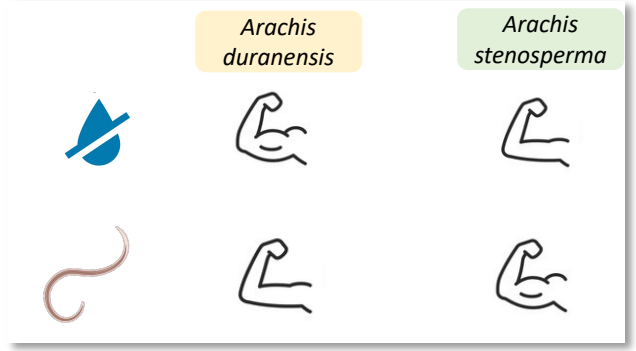
Silvia Bottini, junior professor chair - ISA UMR1355 - INRAE - CNRS - UniCA - silvia.bottini@inrae.fr

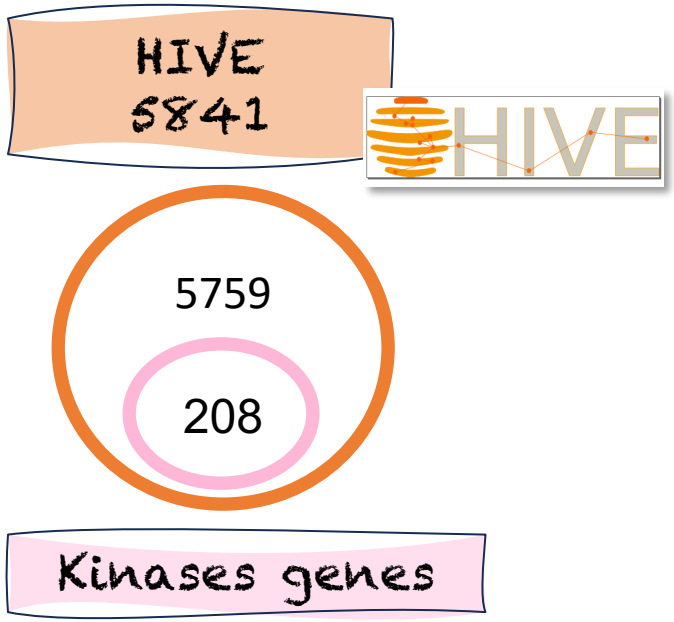


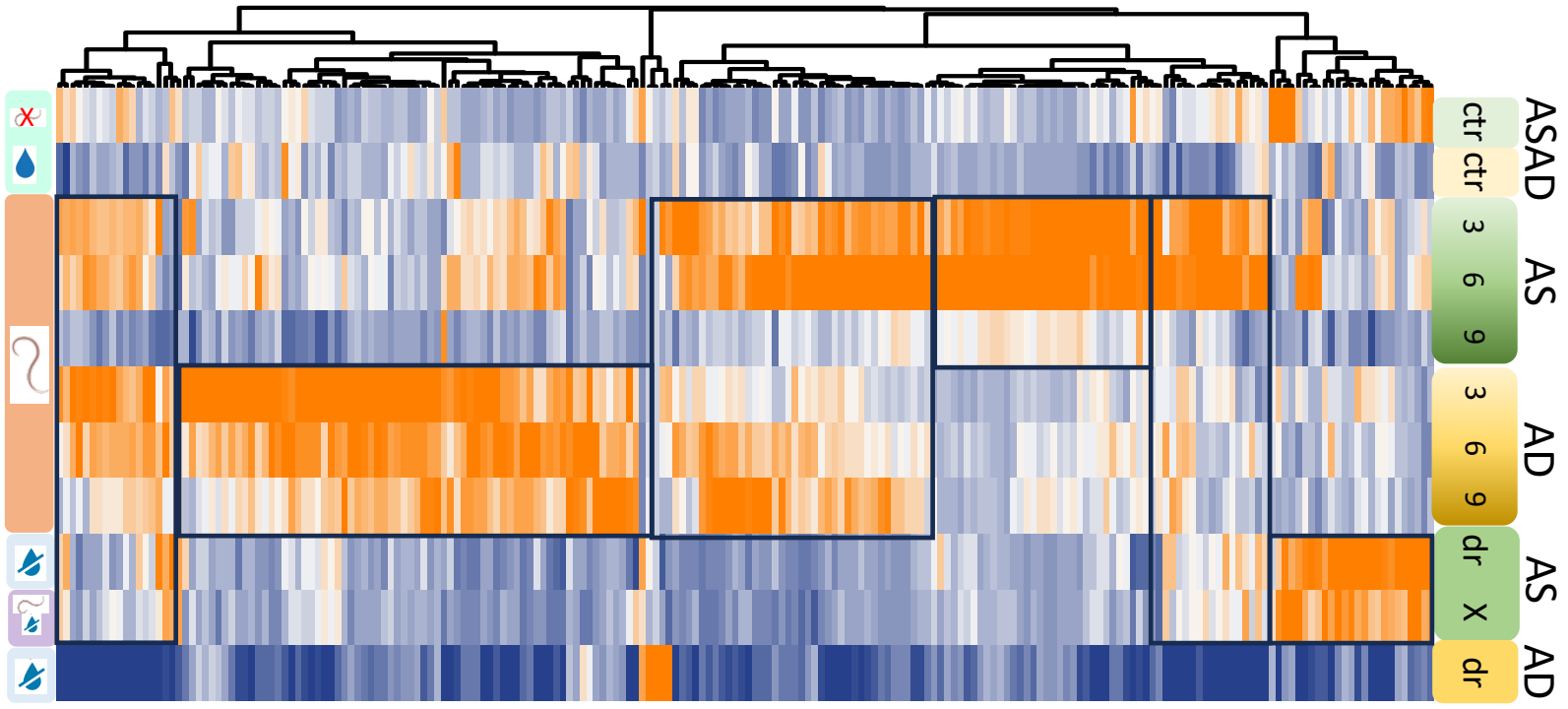
Plant molecular response to stress



HIVE identifies genes coding for kinases



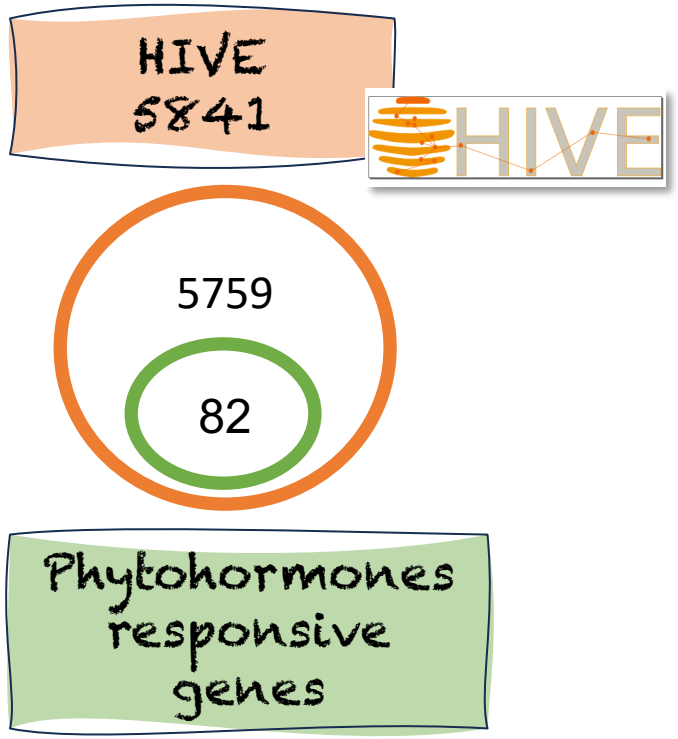


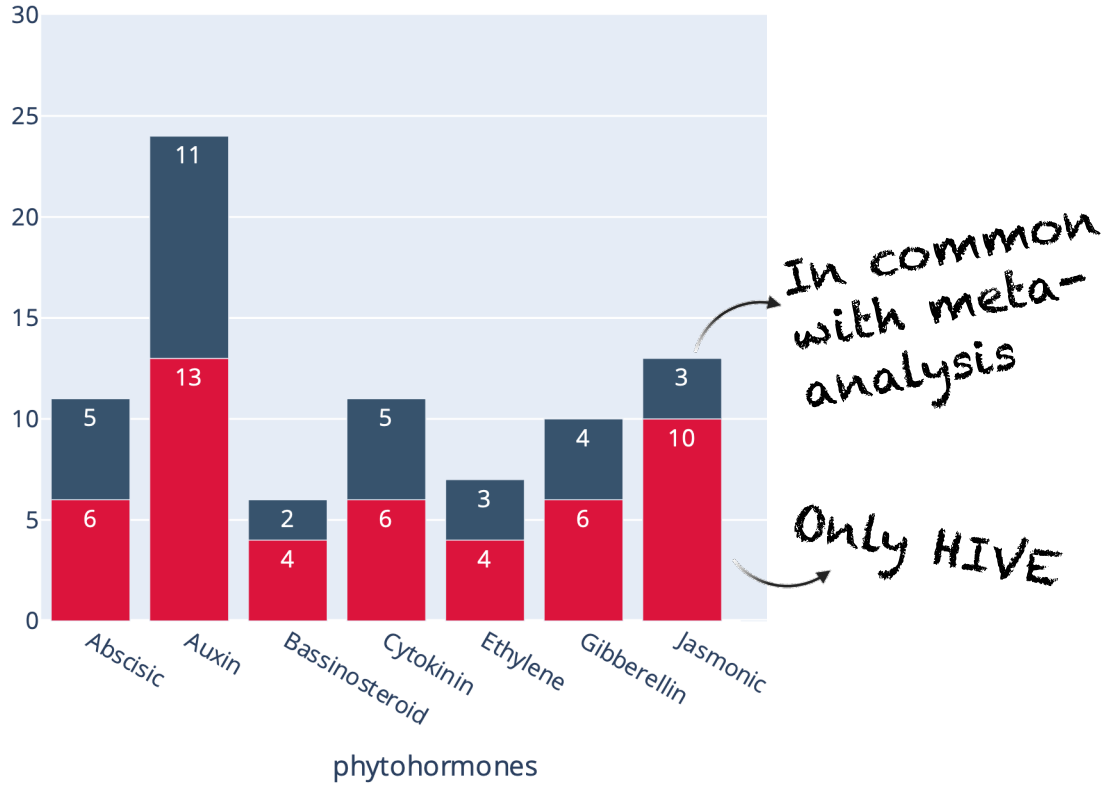


Stress sensing and signal transduction by protein kinases play crucial roles in plant responses to different stress conditions

Silvia Bottini, junior professor chair - ISA UMR1355 - INRAE - CNRS - UniCA - silvia.bottini@inrae.fr

HIVE identifies genes involved in the main hormone signaling pathways in response to stress

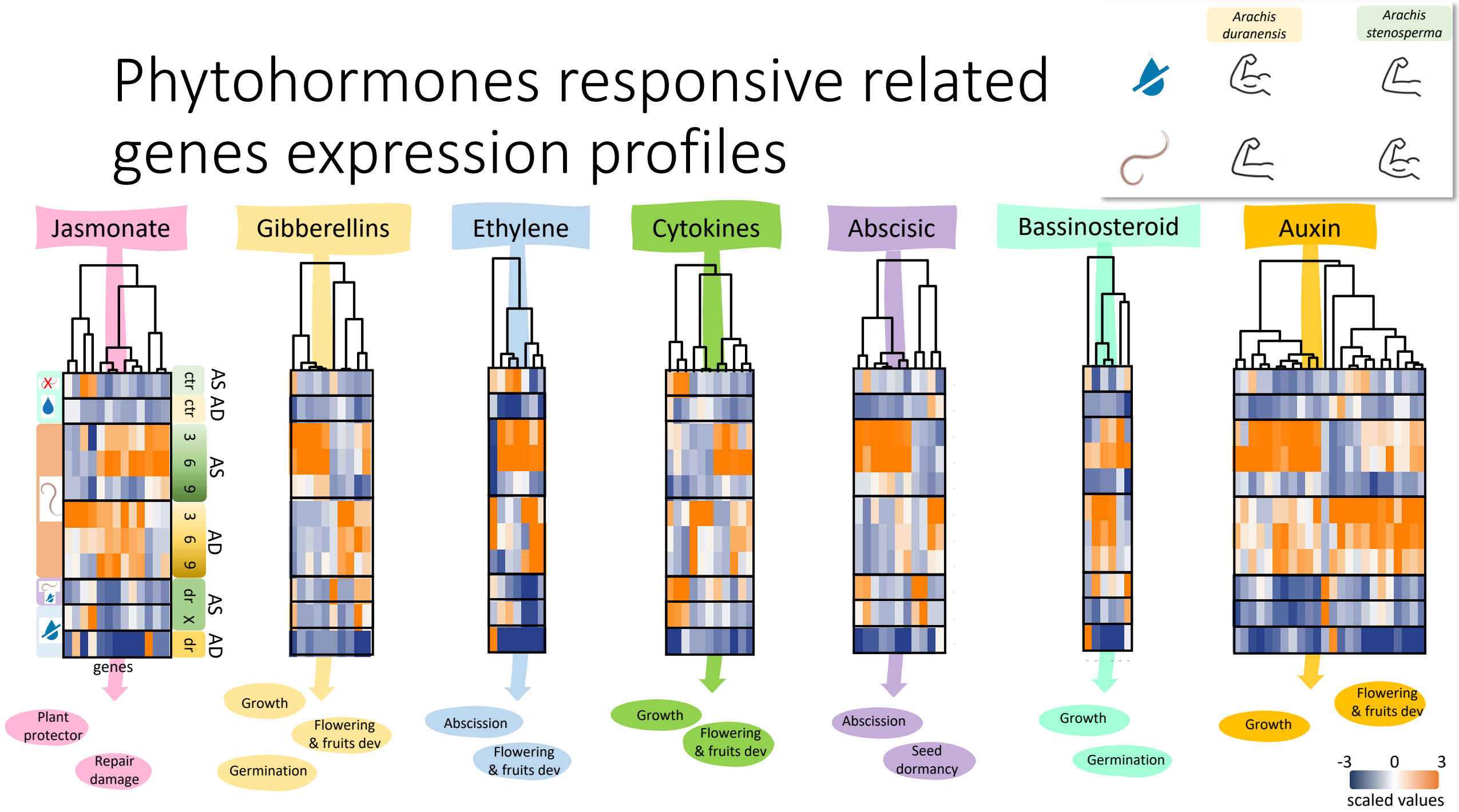




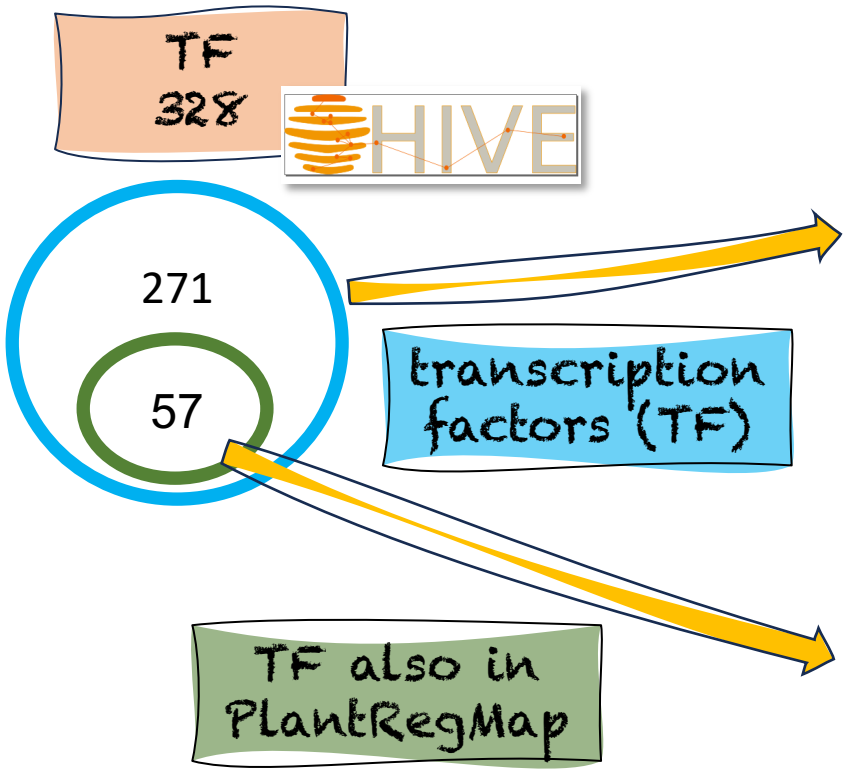
60% were not found in the previous analysis

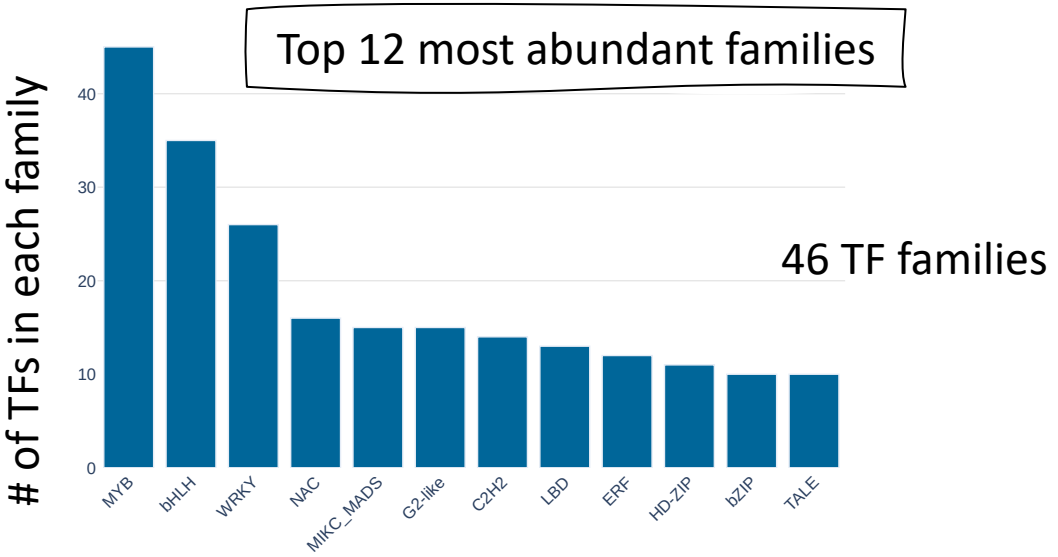
Silvia Bottini, junior professor chair - ISA UMR1355 - INRAE - CNRS - UniCA - silvia.bottini@inrae.fr

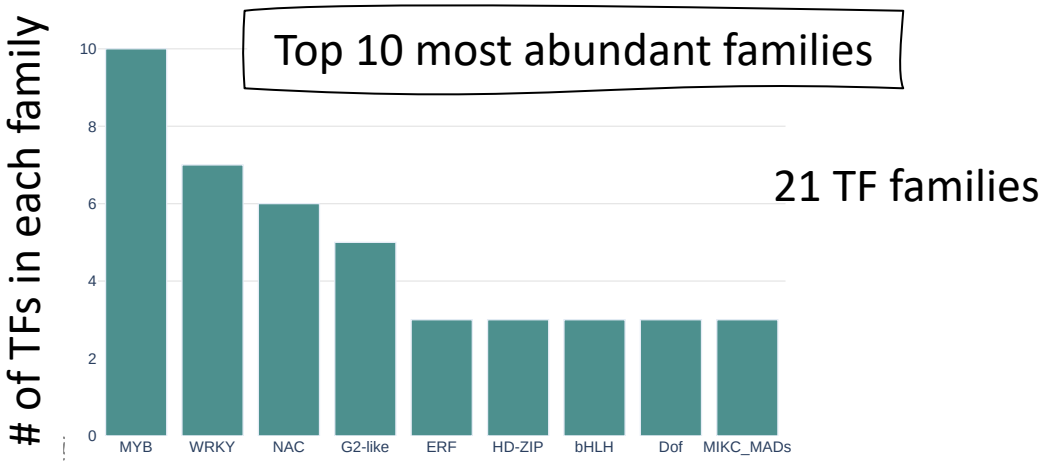
Phytohormones responsive related genes expression profiles



The integrated analysis allowed to retrieve transcription factors induced upon biotic and/or abiotic stress

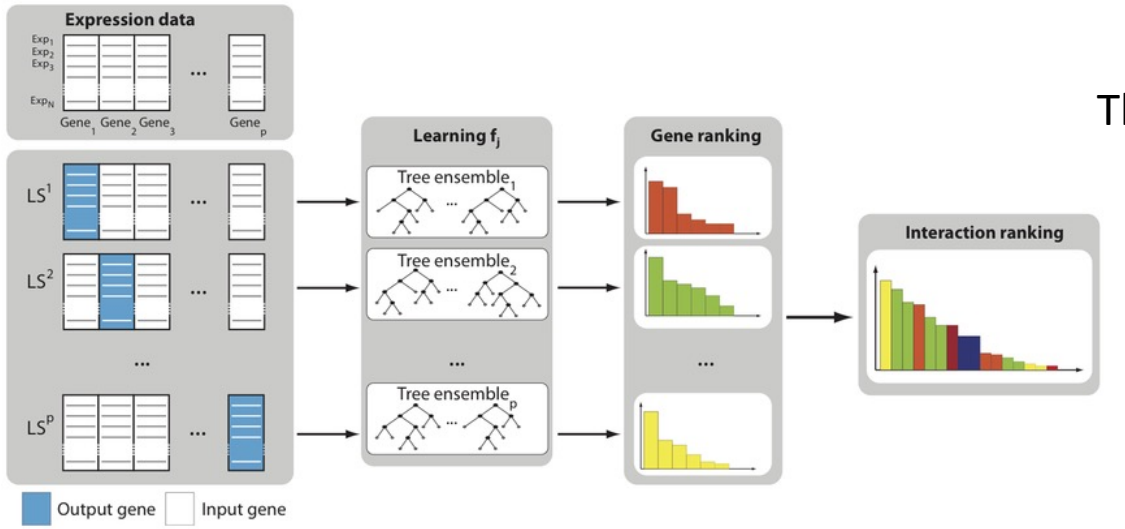






Finding THE targets

GENIE3 Huynh-Thu et al. Plos One 2010



Threshold 0.09

2916 pairs

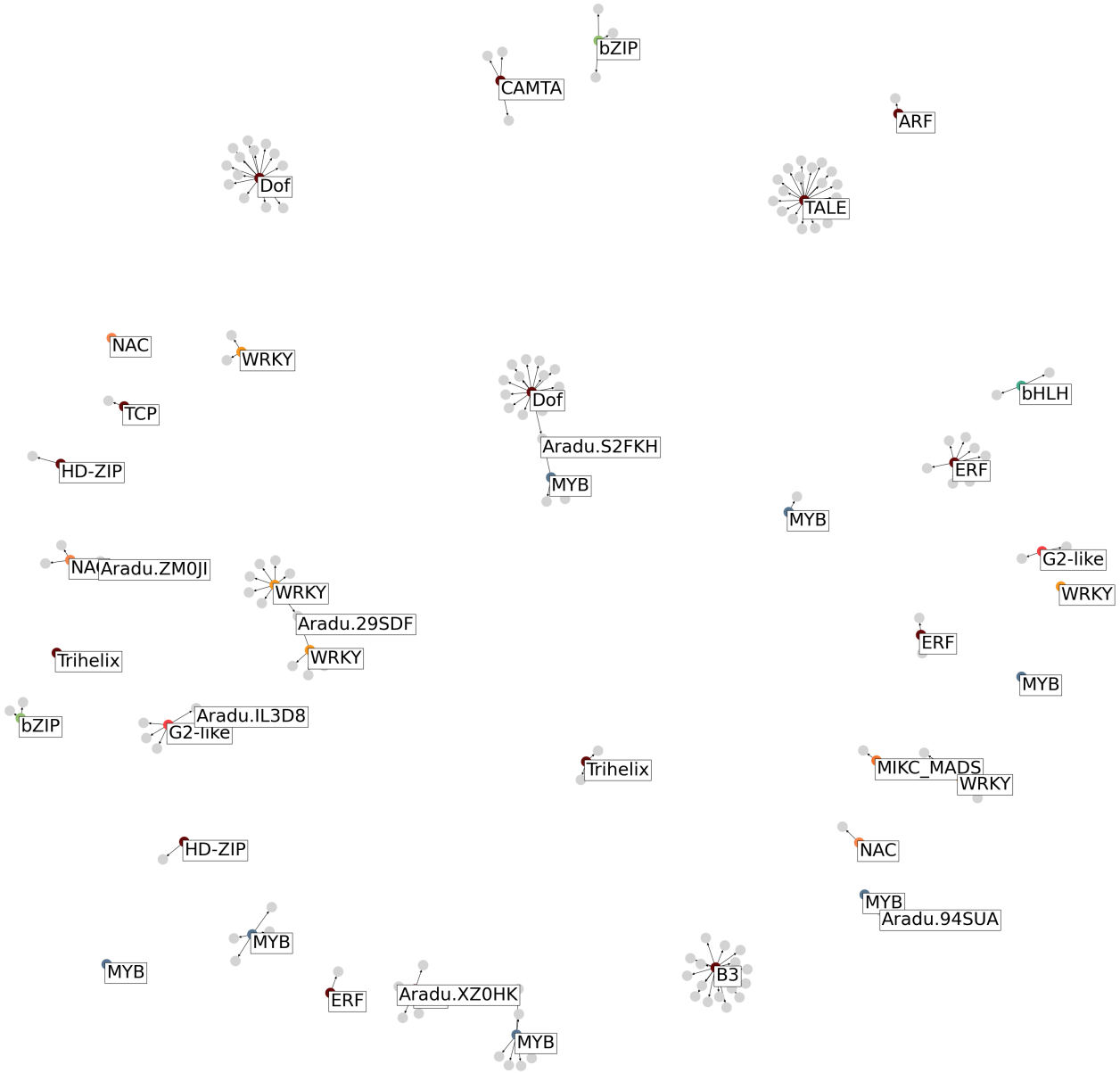
Binding motif filter From plantRegMap

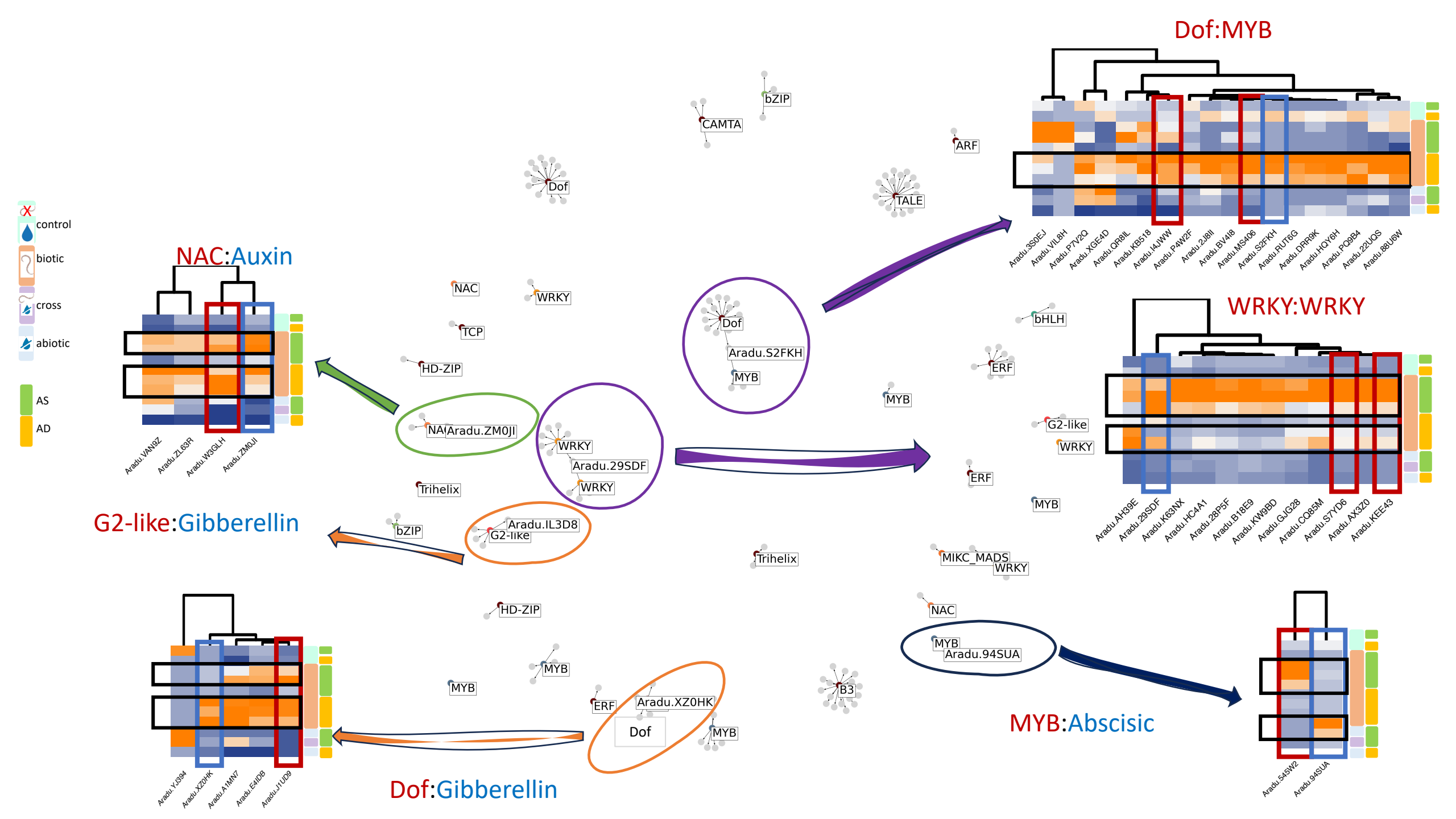
146 pairs

57 TF

5513 putative targets

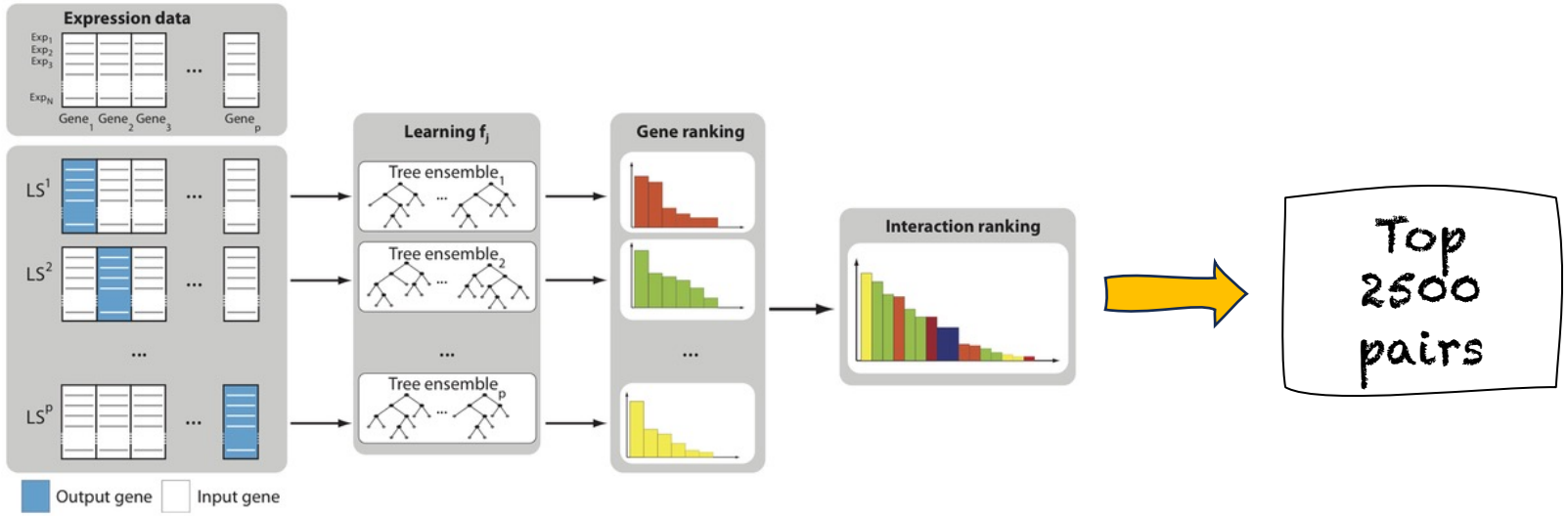
Silvia Bottini, junior professor chair - ISA UMR1355 - INRAE - CNRS - UniCA - silvia.bottini@inrae.fr


bZIP
CAMTA
ARF
Dof
TALE
NAC
WRKY
Dof
bHLH
TCP
Aradu.S2FKH
ERF
HD-ZIP
MYB
MYB
NAC
Aradu.ZM0JI
G2-like
WRKY
WRKY
Aradu.29SDF
ERF
Trihelix
WRKY
MYB
bZIP
Aradu.IL3D8
G2-like
Trihelix
MIKC_MADS
WRKY
HD-ZIP
NAC
MYB
Aradu.94SUA
MYB
MYB
B3
ERF
Aradu.XZ0HK
MYB


control
biotic
cross
abiotic
AS
AD
NAC:Auxin
Aradu.VAN9Z
Aradu.ZL63R
Aradu.W3GLH
Aradu.ZM0JI
G2-like:Gibberellin
Aradu.YJ594
Aradu.XZ0HK
Aradu.A1MN7
Aradu.E4IDB
Aradu.J1UD9
Dof:Gibberellin
Dof:MYB
Aradu.3S0EJ
Aradu.VIL8H
Aradu.P7V2Q
Aradu.XGE4D
Aradu.QR8IL
Aradu.KB518
Aradu.I4JWW
Aradu.P4W2F
Aradu.2J8II
Aradu.BV4I8
Aradu.MS406
Aradu.S2FKH
Aradu.RUT6G
Aradu.DRR9K
Aradu.HQY6H
Aradu.PQ9B4
Aradu.22UQS
Aradu.88U6W
WRKY:WRKY
Aradu.AH39E
Aradu.29SDF
Aradu.K63NX
Aradu.HC4A1
Aradu.28P5F
Aradu.B18E9
Aradu.KW9BD
Aradu.GJG28
Aradu.CQ85M
Aradu.S7YD6
Aradu.AX3Z0
Aradu.KEE43
MYB:Abscisic
Aradu.545W2
Aradu.94SUA
NAC
TCP
HD-ZIP
WRKY
bZIP
Trihelix
MYB
ERF
bHLH
G2-like
MIKC_MADS
B3
ARF
TALE
CAMTA
Aradu.ZM0JI
Aradu.29SDF
Aradu.IL3D8
Aradu.S2FKH
Aradu.XZ0HK
Aradu.94SUA

Finding THE targets

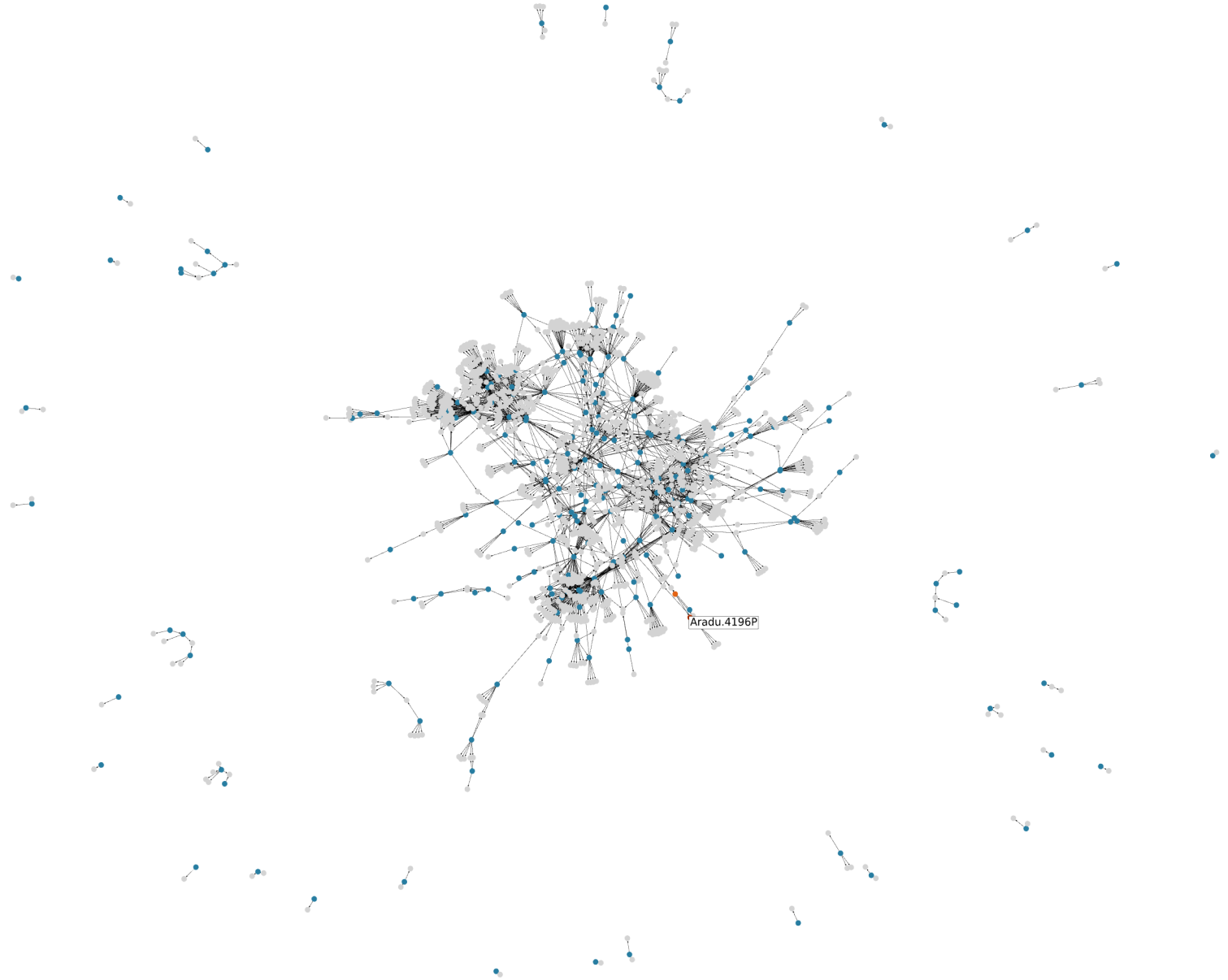
GENIE3 Huynh-Thu et al. Plos One 2010



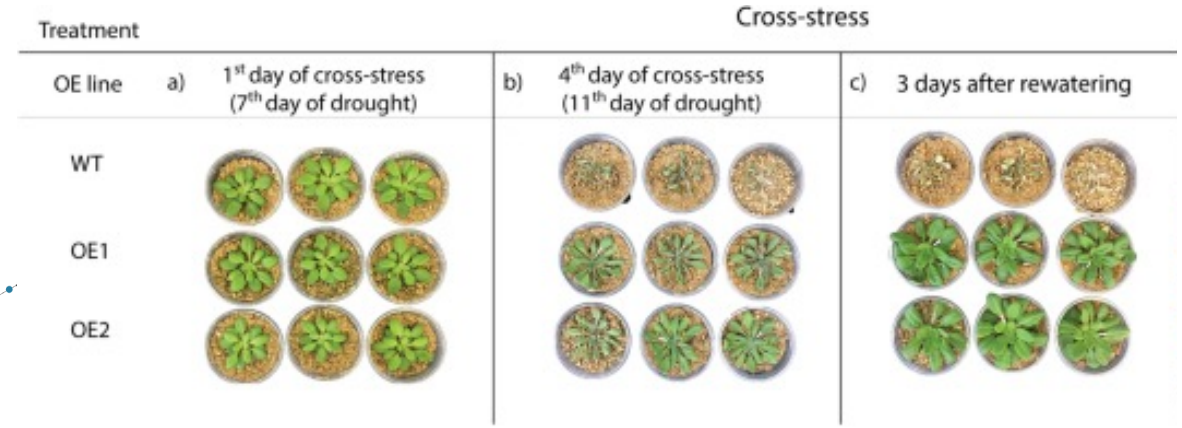
Top 2500 pairs

271 TF

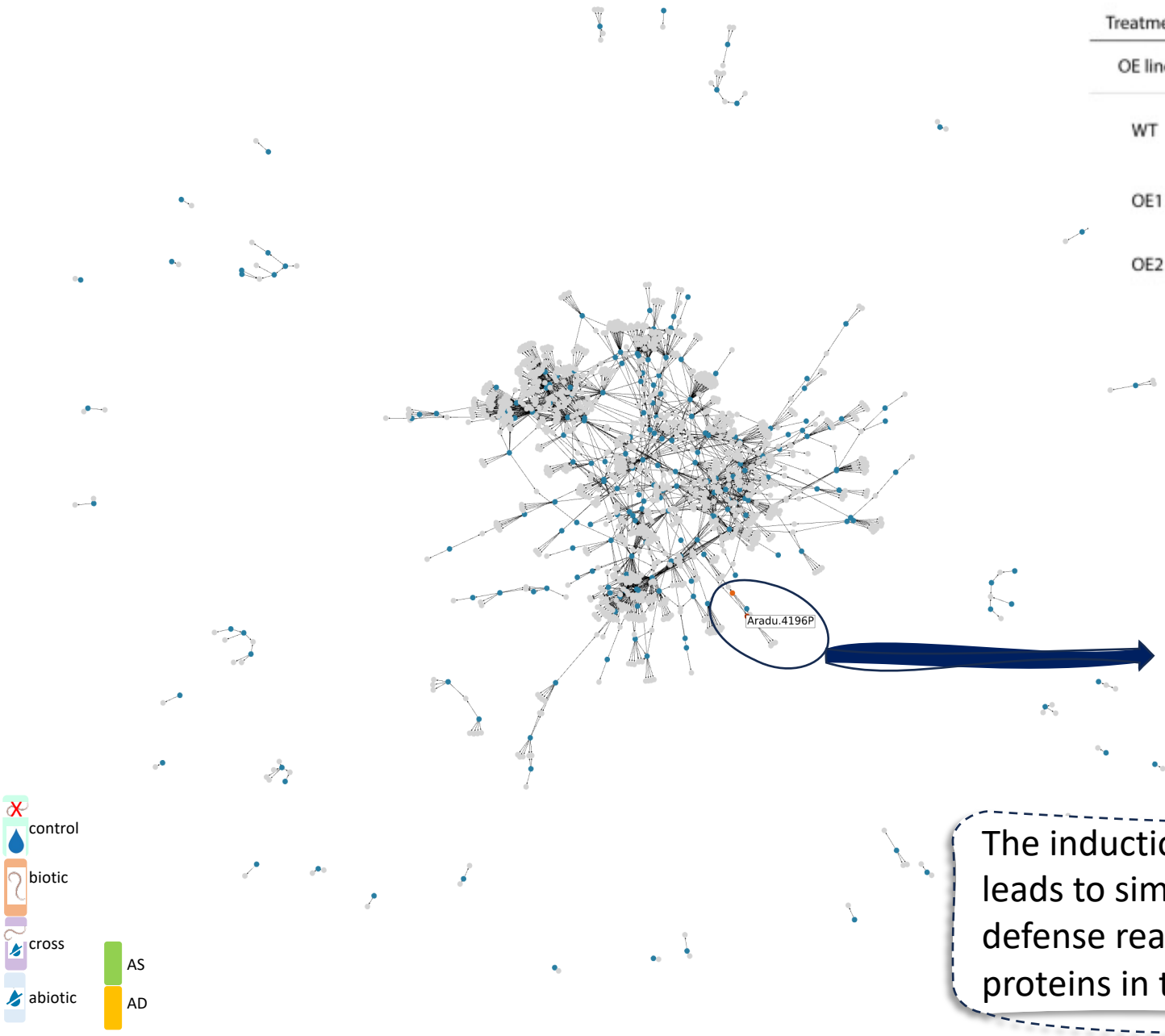
5513 putative targets


Aradu.4196P

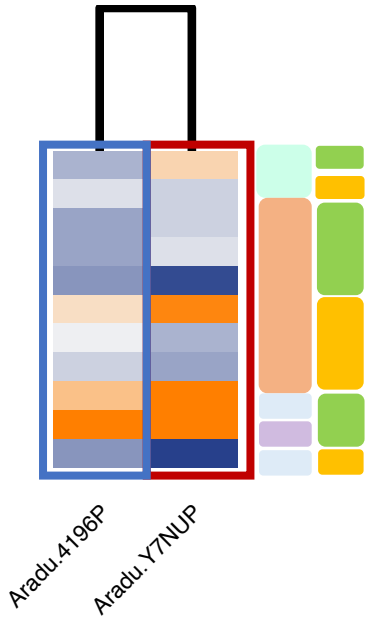
Zotta-Mota et al. Sci. Rep. 2021



Overexpression *in planta* showed high resistance to both stresses, separately or combined.

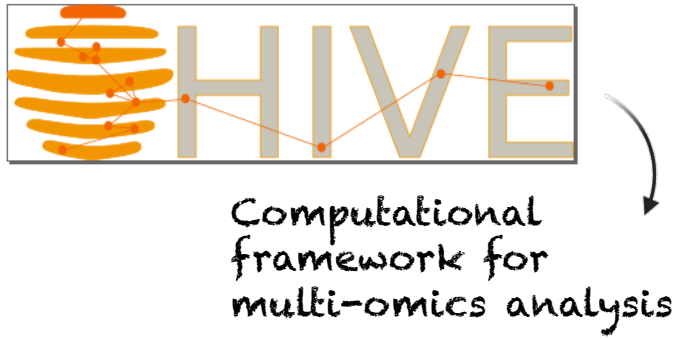


G2-like:Endochitinase



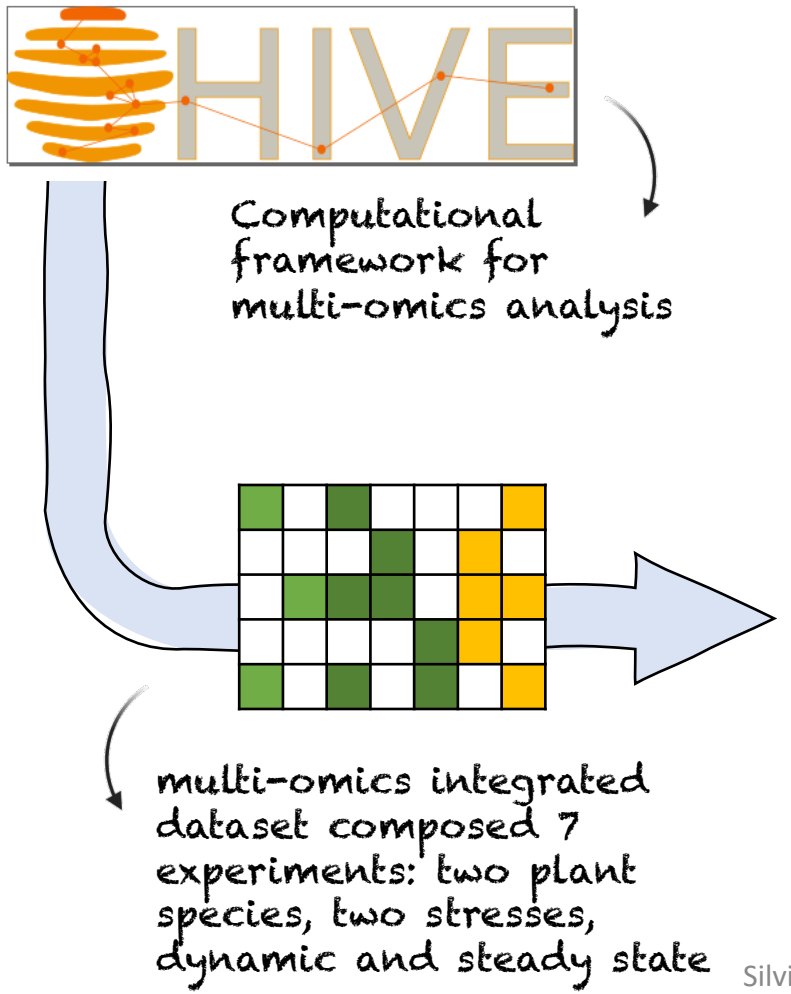
The induction of chitinases by specific agents leads to simultaneous enhancement of other defense reactions and pathogenesis related proteins in the same plant tissues.

Conclusions

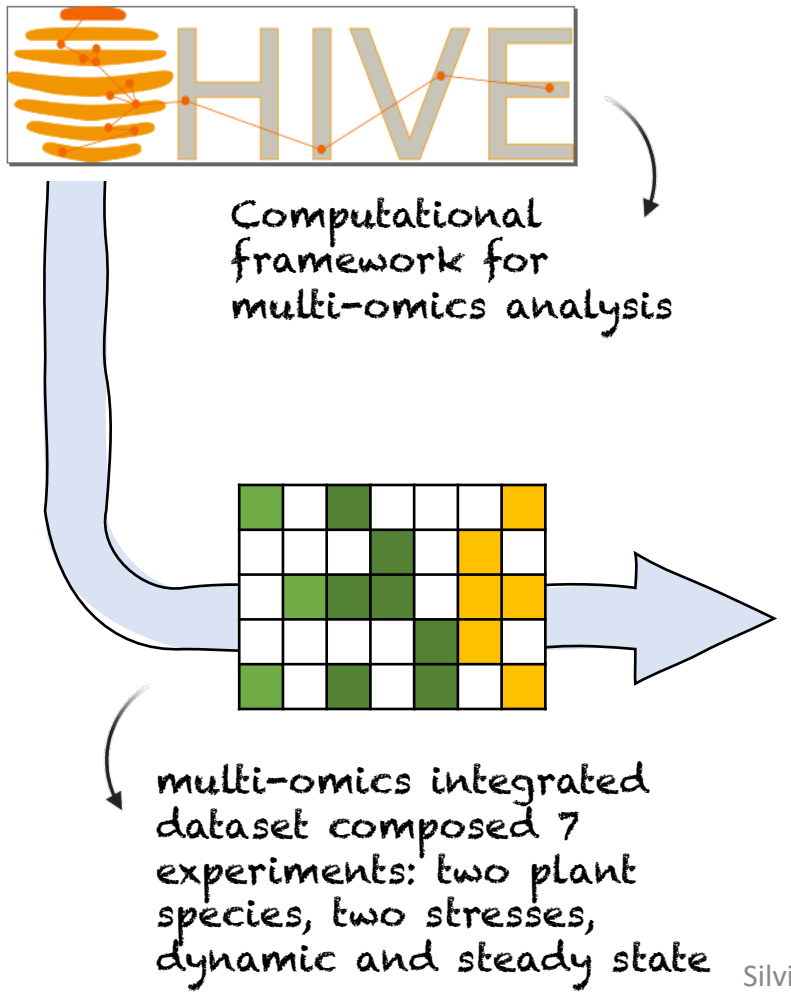


Silvia Bottini, junior professor chair - ISA UMR1355 - INRAE - CNRS - UniCA - silvia.bottini@inrae.fr

Conclusions



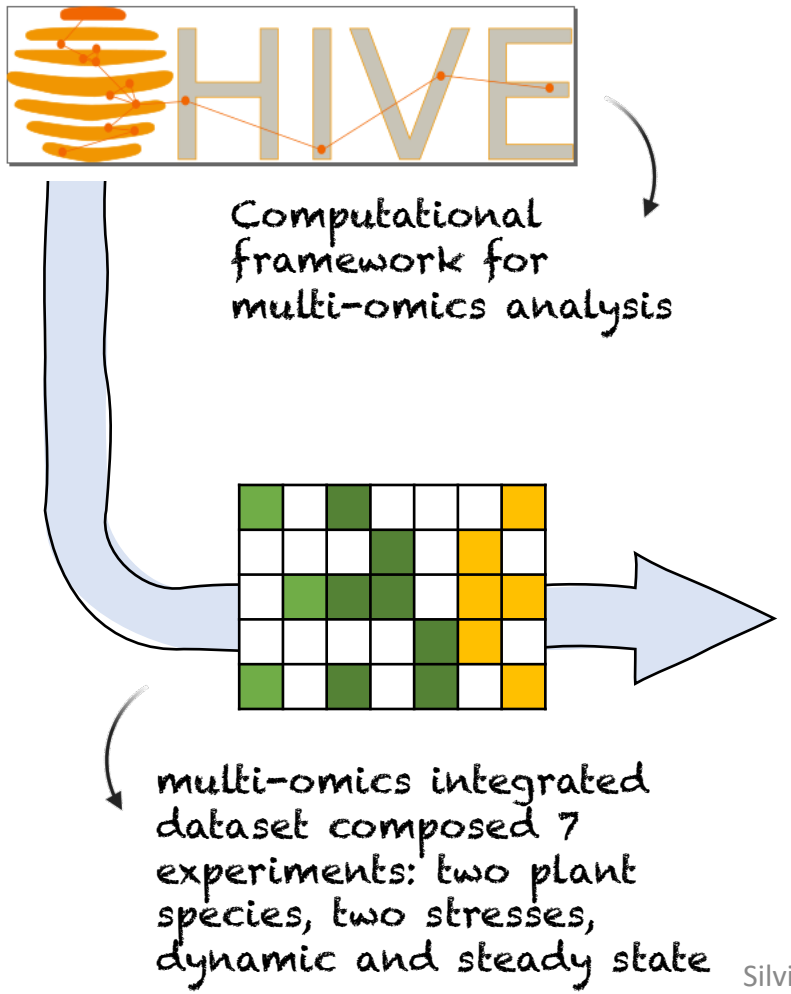
Conclusions



Identification of **5841 genes** deregulated in at least one condition, including **several experimentally validated genes**.

Silvia Bottini, junior professor chair - ISA UMR1355 - INRAE - CNRS - UniCA - silvia.bottini@inrae.fr

Conclusions



Identification of **5841 genes** deregulated in at least one condition, including **several experimentally validated genes**.

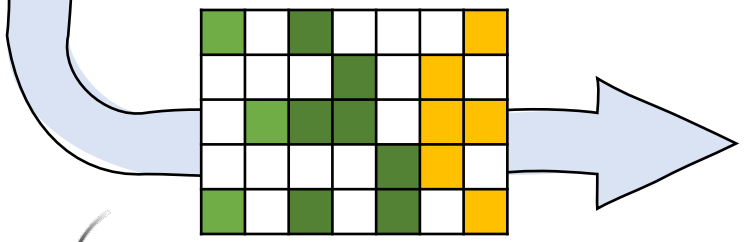
Identification of several **genes** connected with **plant-stress response** (phytohormones, kinases, transcription factors, R-genes).

Silvia Bottini, junior professor chair - ISA UMR1355 - INRAE - CNRS - UniCA - silvia.bottini@inrae.fr

Conclusions

HIVE

Computational framework for multi-omics analysis



multi-omics integrated dataset composed 7 experiments: two plant species, two stresses, dynamic and steady state

Identification of **5841 genes** deregulated in at least one condition, including **several experimentally validated genes**.

Identification of several **genes** connected with **plant-stress response** (phytohormones, kinases, transcription factors, R-genes).

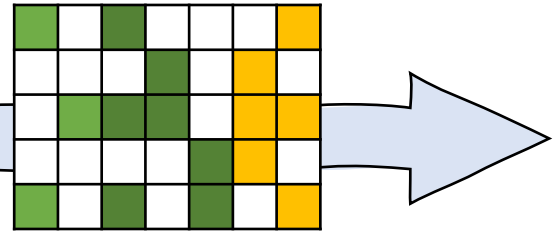
Identification of **potential candidates** to better understand the responses to stresses.

Silvia Bottini, junior professor chair - ISA UMR1355 - INRAE - CNRS - UniCA - silvia.bottini@inrae.fr

Conclusions

HIVE

Computational framework for multi-omics analysis



multi-omics integrated dataset composed 7 experiments: two plant species, two stresses, dynamic and steady state

Identification of **5841 genes** deregulated in at least one condition, including **several experimentally validated genes**.

Identification of several **genes** connected with **plant-stress response** (phytohormones, kinases, transcription factors, R-genes).

Identification of **potential candidates** to better understand the responses to stresses.

HIVE can be applied on any biological model!

Silvia Bottini, junior professor chair - ISA UMR1355 - INRAE - CNRS - UniCA - silvia.bottini@inrae.fr

Acknowledgments

Giulia Calia
Justine Labory

All team M2P2

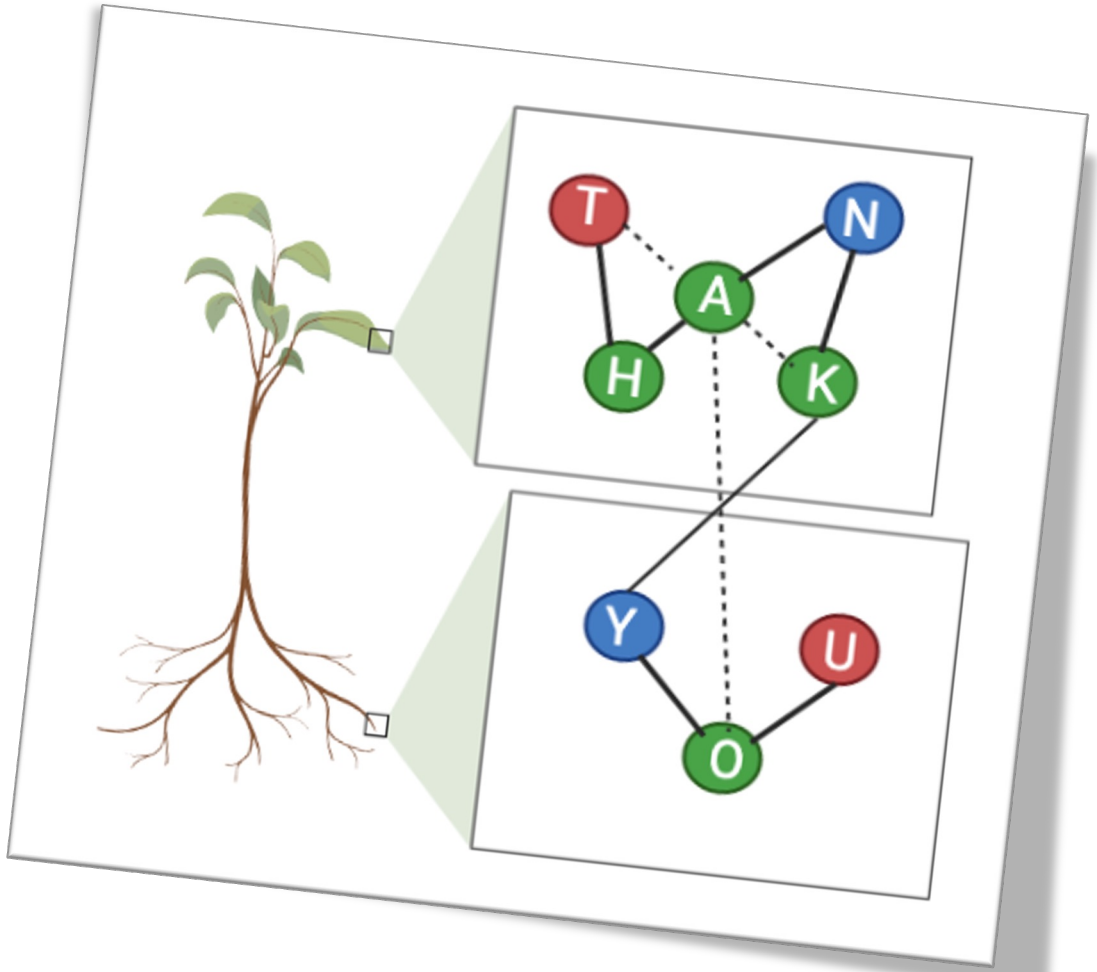
Ana Zotta Mota
Etienne Danchin

All team GAME

Patricia Messenberg Guimaraes
& all the team

Le comité de pilotage du métaprogramme DIGIT-BIO

Questions



Email: silvia.bottini@inrae.fr

Benchmarking tools for genes selection from integrated datasets

PLS-DA Partial Least Squares Discriminant Analysis

Supervised method

Iterative method that constructs *H* successive artificial (latent) components, where each component is a linear combination of the variables.

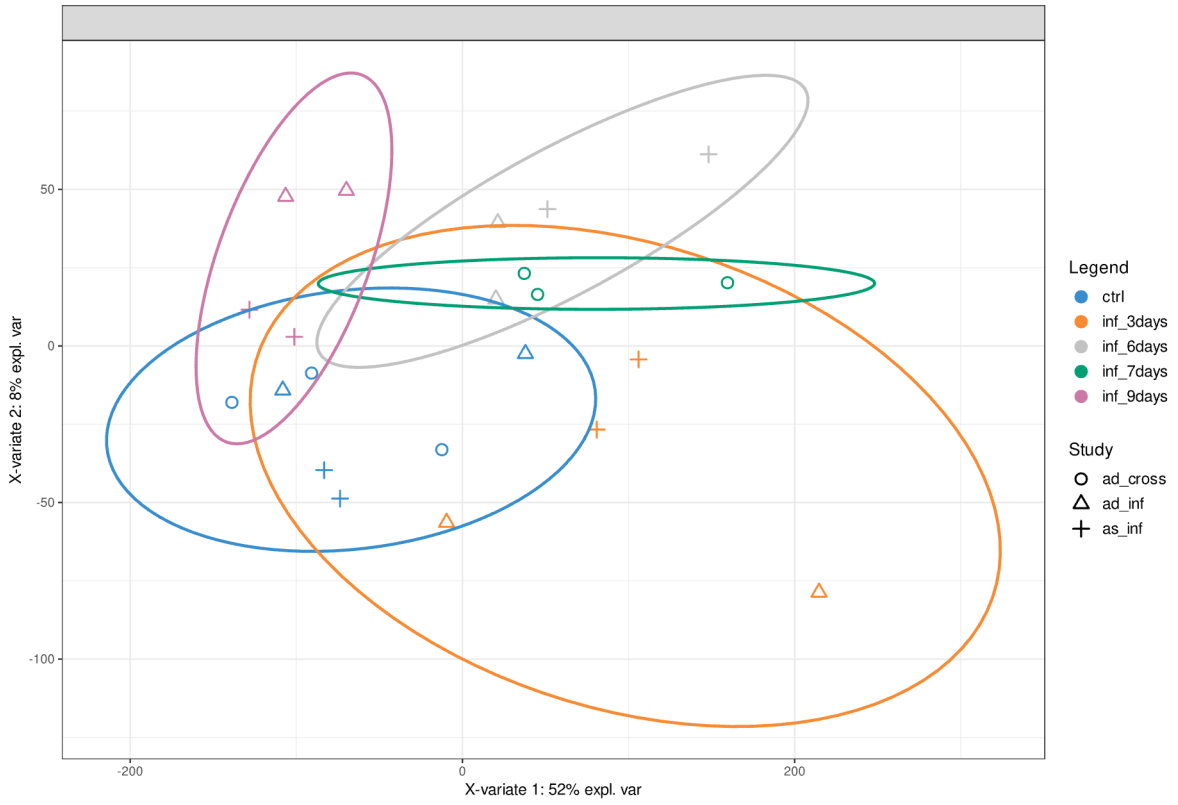
Silvia Bottini, junior professor chair - ISA UMR1355 - INRAE - CNRS - UniCA - silvia.bottini@inrae.fr

Benchmarking tools for genes selection from integrated datasets

PLS-DA Partial Least Squares Discriminant Analysis

Supervised method

Iterative method that constructs *H* successive artificial (latent) components, where each component is a linear combination of the variables.



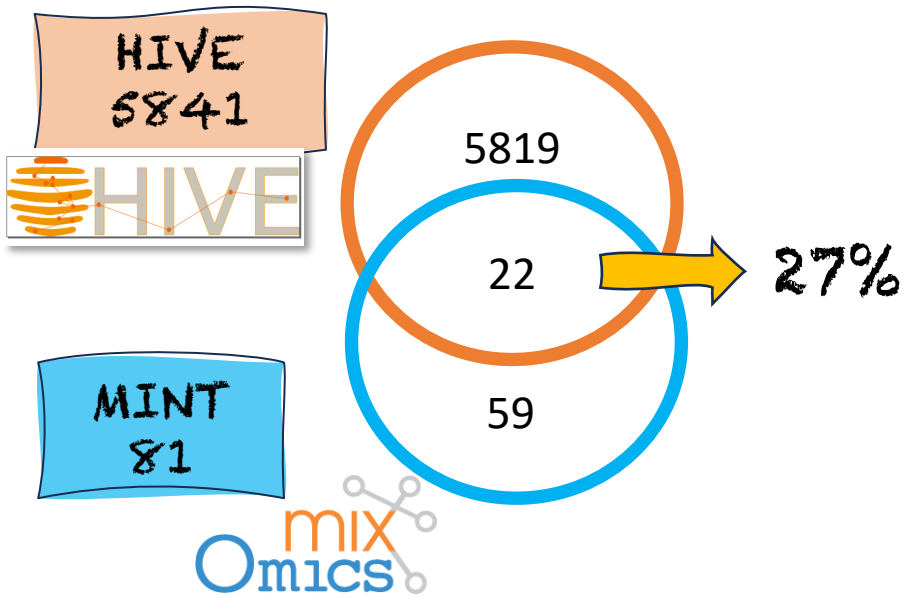
Component 1 = 22 genes

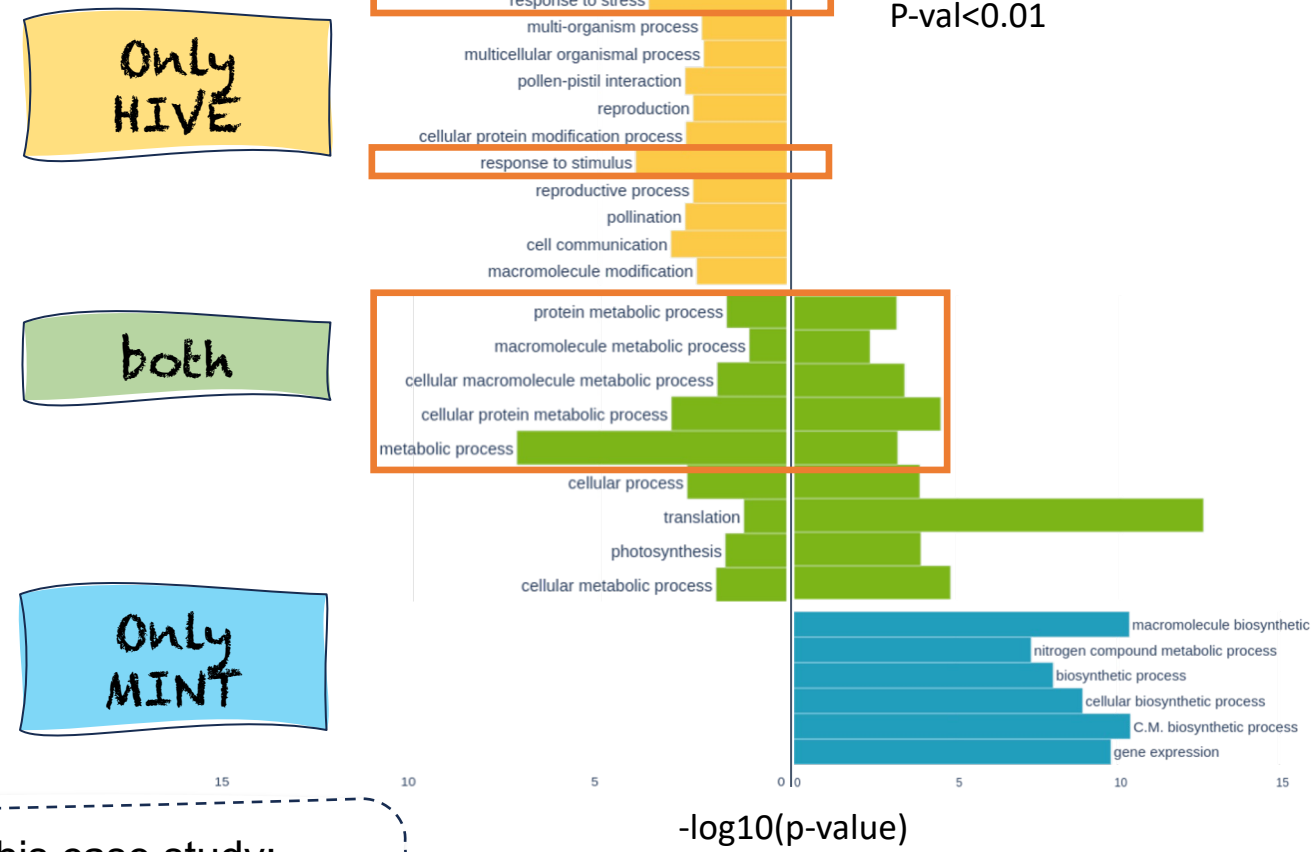
Component 2 = 59 genes

Total = 81 genes

Silvia Bottini, junior professor chair - ISA UMR1355 - INRAE - CNRS - UniCA - silvia.bottini@inrae.fr

HIVE vs MINT

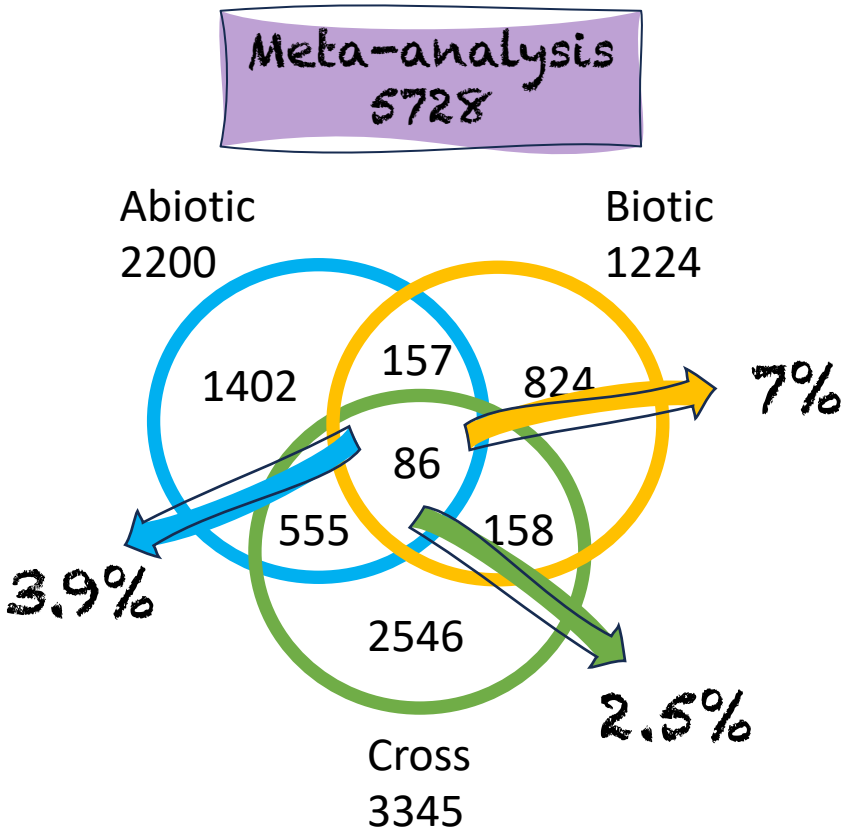




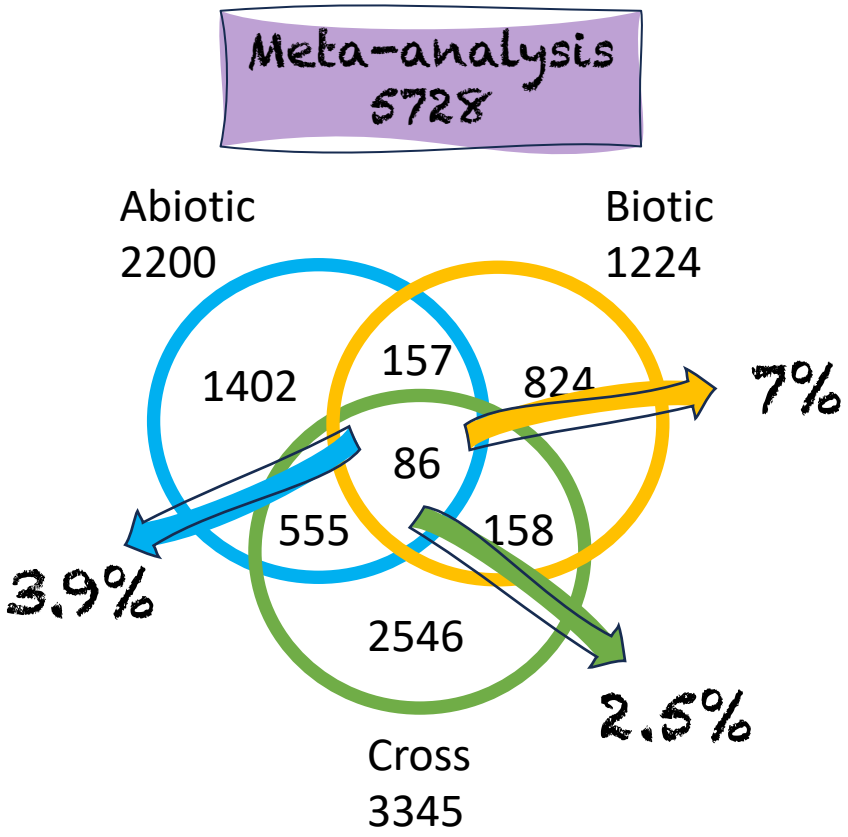
GO Term,
Biological processes
P-val<0.01

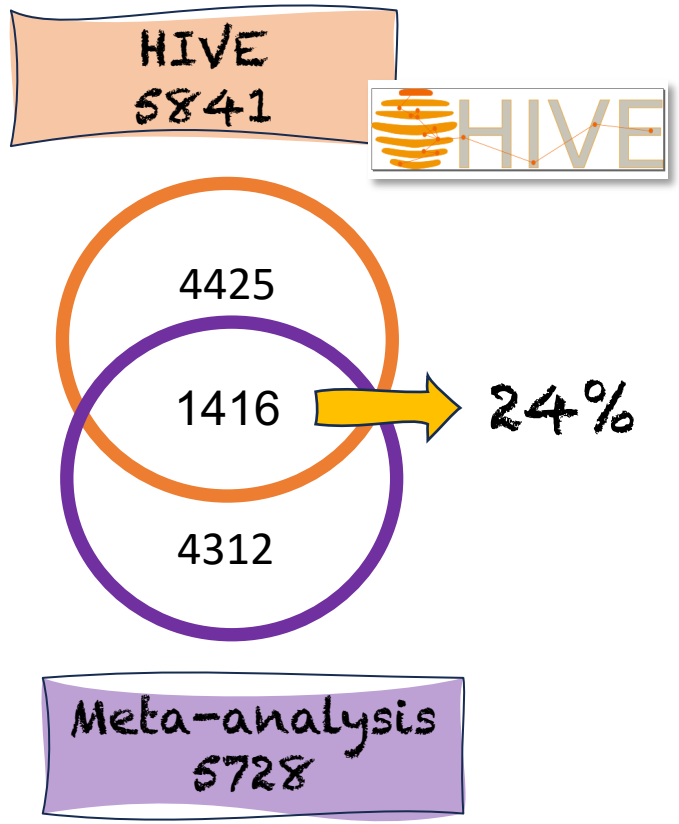
The possible limitations in the performances of MINT on this case study:
- the complex experimental design making difficult to choose the classes to perform the classification
- the limited number of replicates of some classes

HIVE vs meta-analysis



HIVE vs meta-analysis





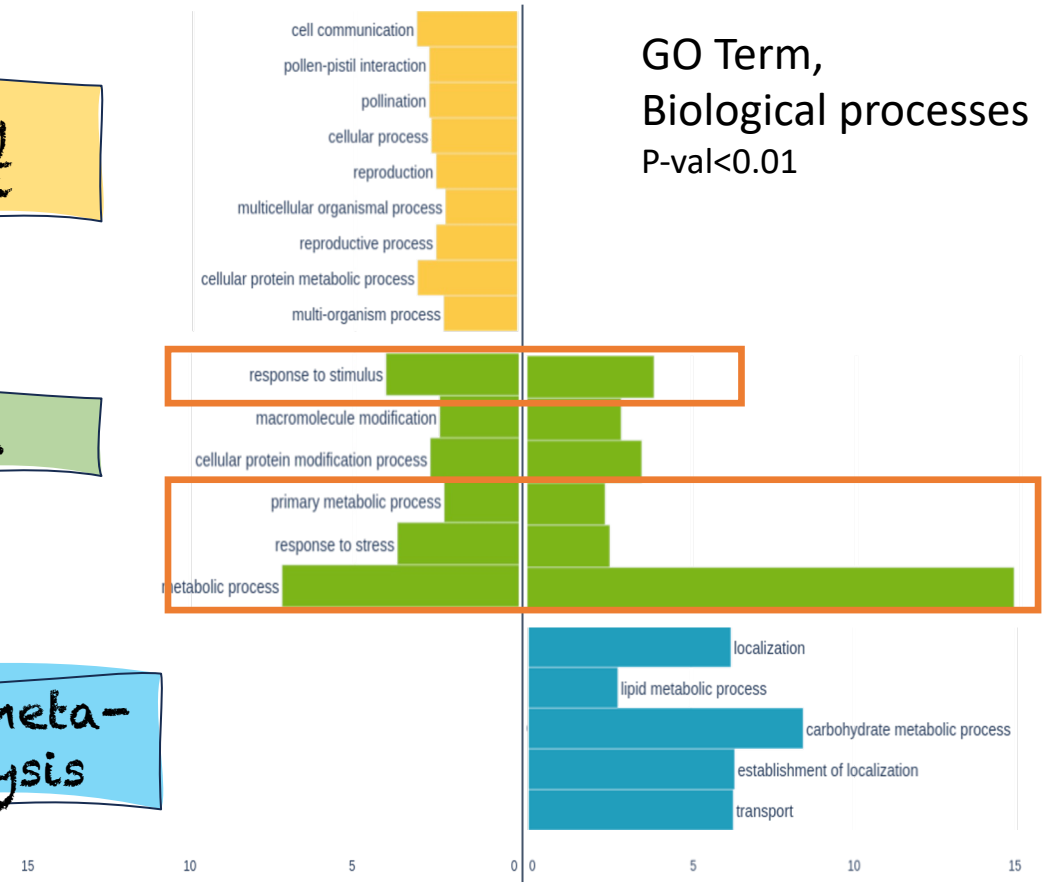
HIVE and meta-analysis have a higher overlap than the three separated meta-analysis

HIVE vs meta-analysis

Only HIVE

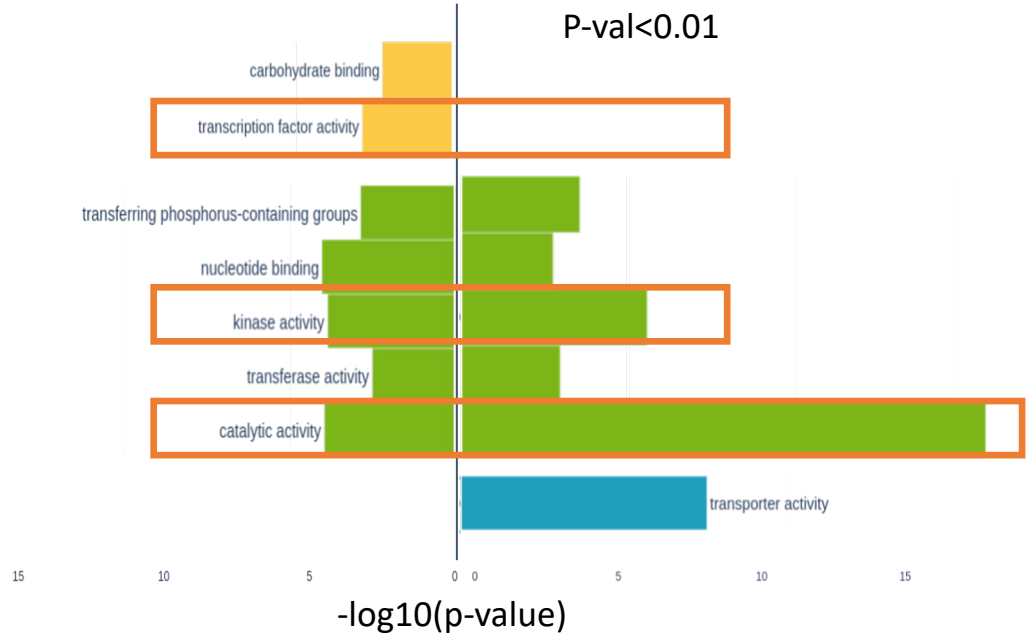
both

Only meta-analysis



GO Term,
Biological processes
P-val<0.01

-log10(p-value)



GO Term,
Molecular function
P-val<0.01

-log10(p-value)

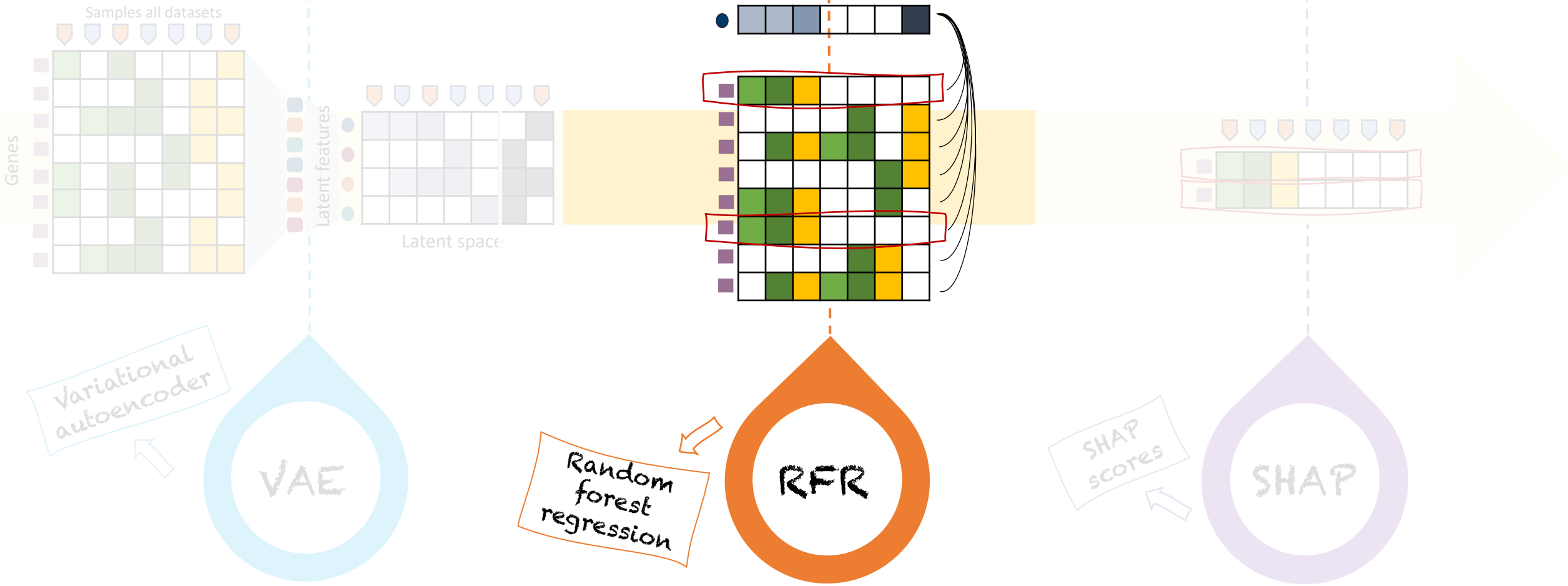
HIVE selected the genes in common with the meta-analysis that bring most of the biological information and found additional genes with novel biological information

HIVE pipeline

1 Removing batch effects

2 Opening the black box

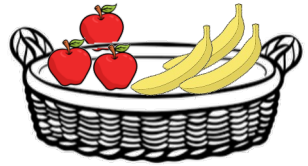
3 Finding important genes



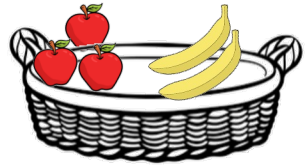
Feature importance

Random forest uses Mean Decrease in Impurity (MDI) to calculate Feature importance.

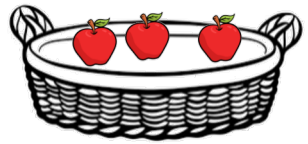
It calculates for each feature the mean decrease in impurity it introduced across all the decision trees while constructing them.



Gini=0.5
High impurity



Gini=0.47
Lower impurity



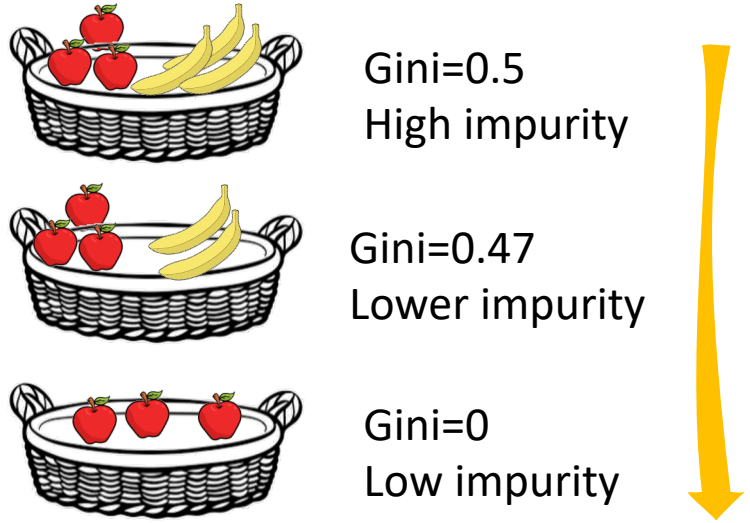
Gini=0
Low impurity

Silvia Bottini, junior professor chair - ISA UMR1355 - INRAE - CNRS - UniCA - silvia.bottini@inrae.fr

Feature importance

Random forest uses Mean Decrease in Impurity (MDI) to calculate Feature importance.

It calculates for each feature the mean decrease in impurity it introduced across all the decision trees while constructing them.



For every feature, it tries every split and calculates the Gini impurity, compares it to the current impurity, and picks the best one that reduces the impurity.

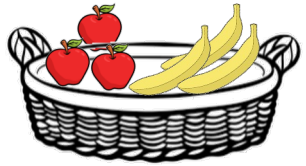
Silvia Bottini, junior professor chair - ISA UMR1355 - INRAE - CNRS - UniCA - silvia.bottini@inrae.fr

Feature importance

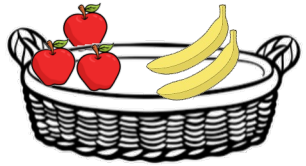
Random forest uses Mean Decrease in Impurity (MDI) to calculate Feature importance.

It calculates for each feature the mean decrease in impurity it introduced across all the decision trees while constructing them.

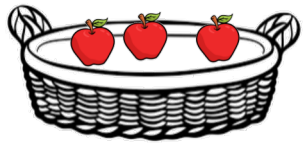
Gini index



Gini=0.5
High impurity



Gini=0.47
Lower impurity



Gini=0
Low impurity

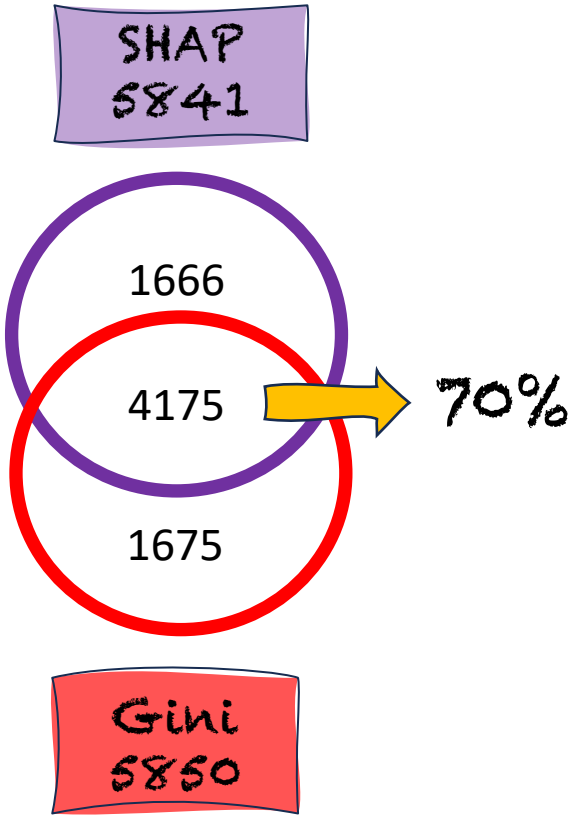
For every feature, it tries every split and calculates the Gini impurity, compares it to the current impurity, and picks the best one that reduces the impurity.

Two features, one has more splitting points than the other, but they both decrease the impurity by the same amount, they will get the same importance!

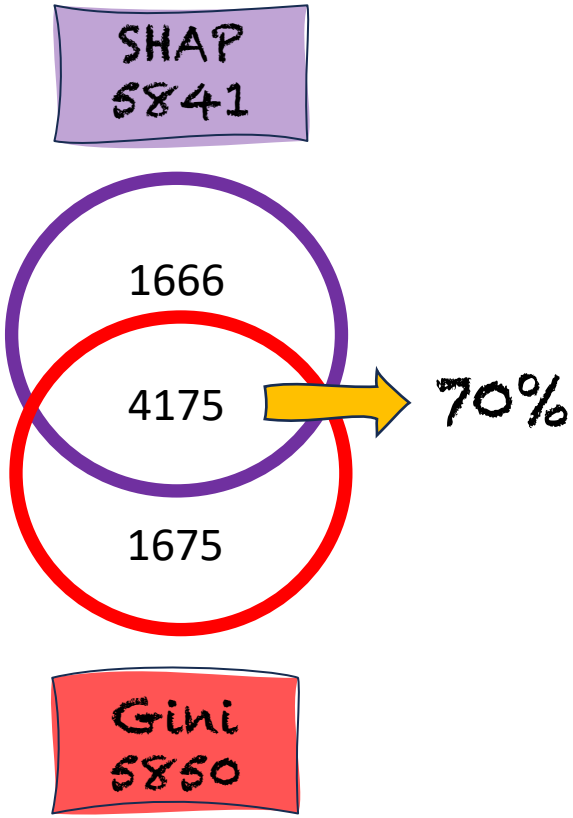
biased towards high cardinal features

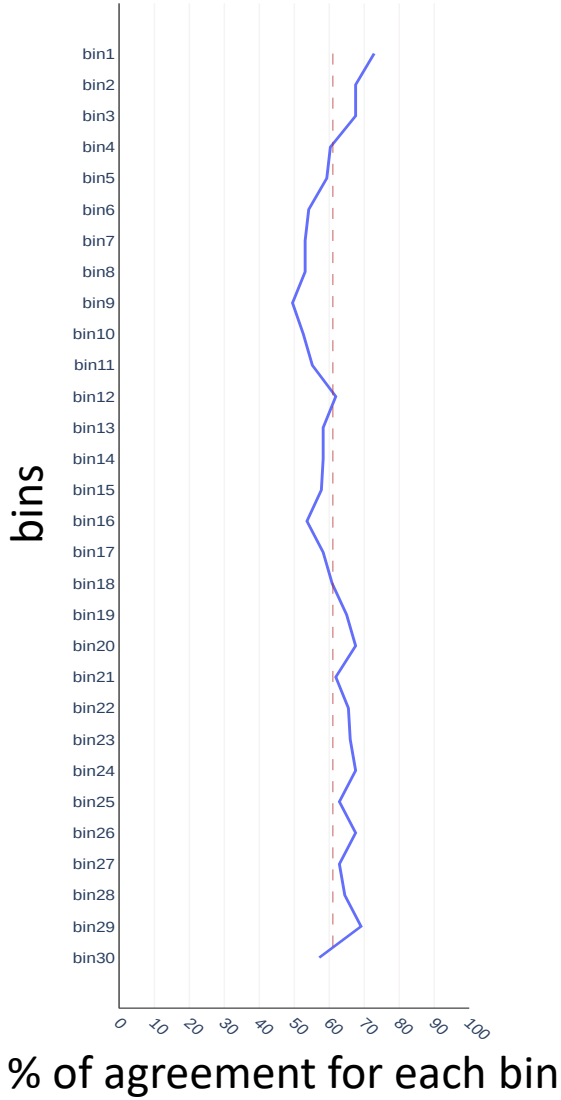
Silvia Bottini, junior professor chair - ISA UMR1355 - INRAE - CNRS - UniCA - silvia.bottini@inrae.fr

Comparing importance scores

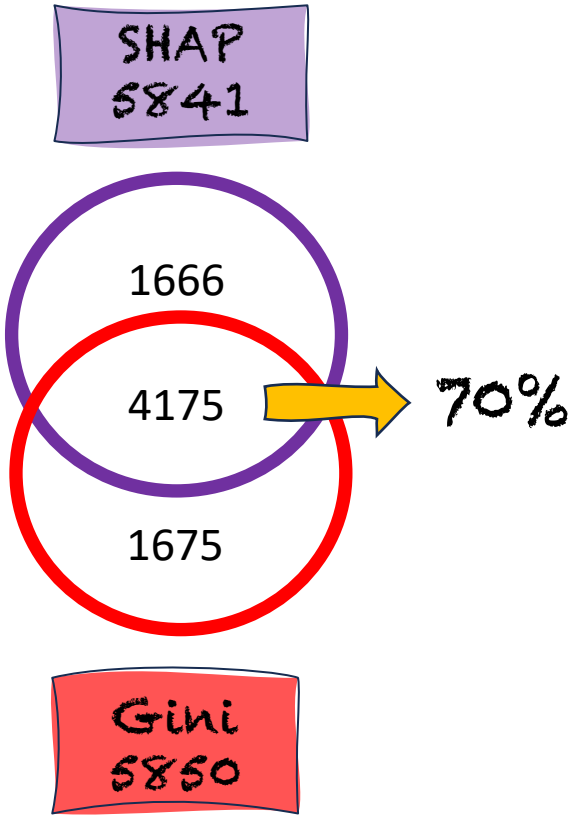


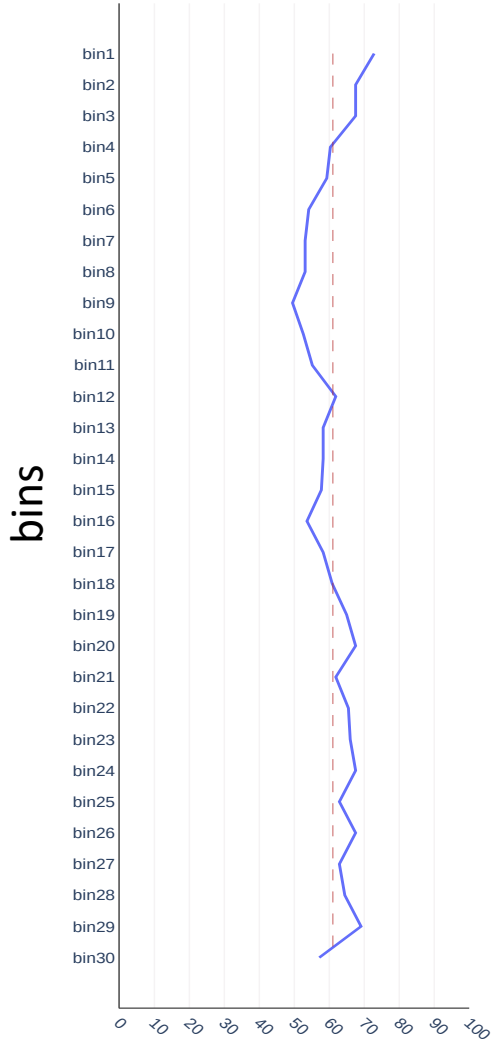
Comparing importance scores





Comparing importance scores



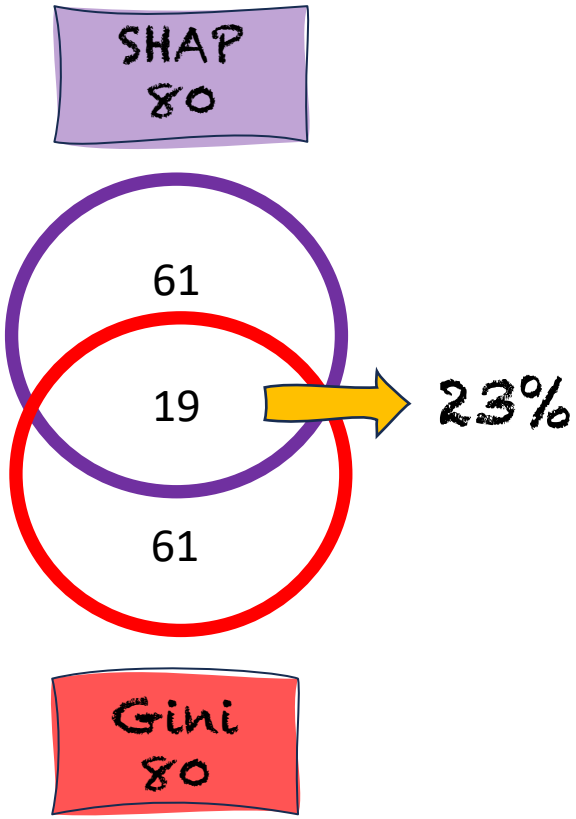


% of agreement for each bin

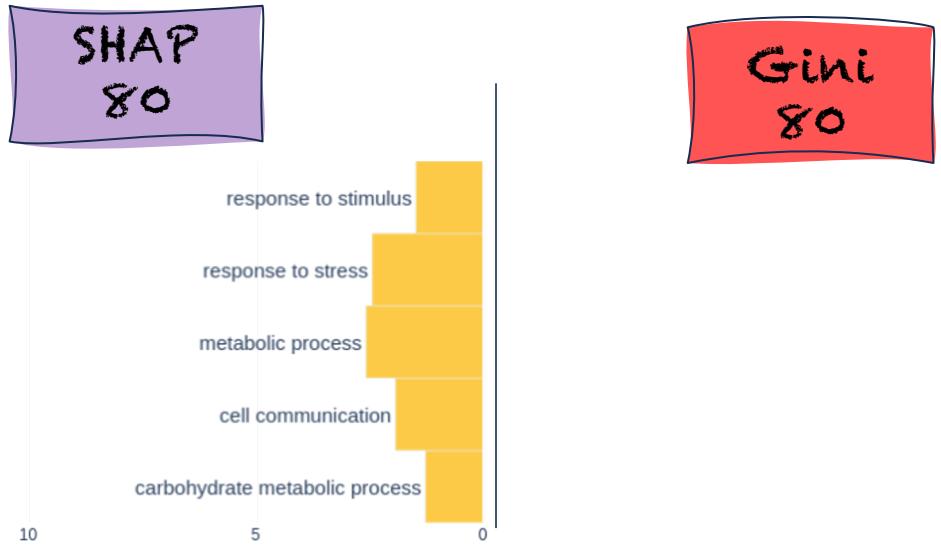
GO Term, Biological process (p-value<0.01)

-log10(p-value)

SHAP score performs better in selecting genes with biological meaning



GO Term, Biological process (p-value<0.1)

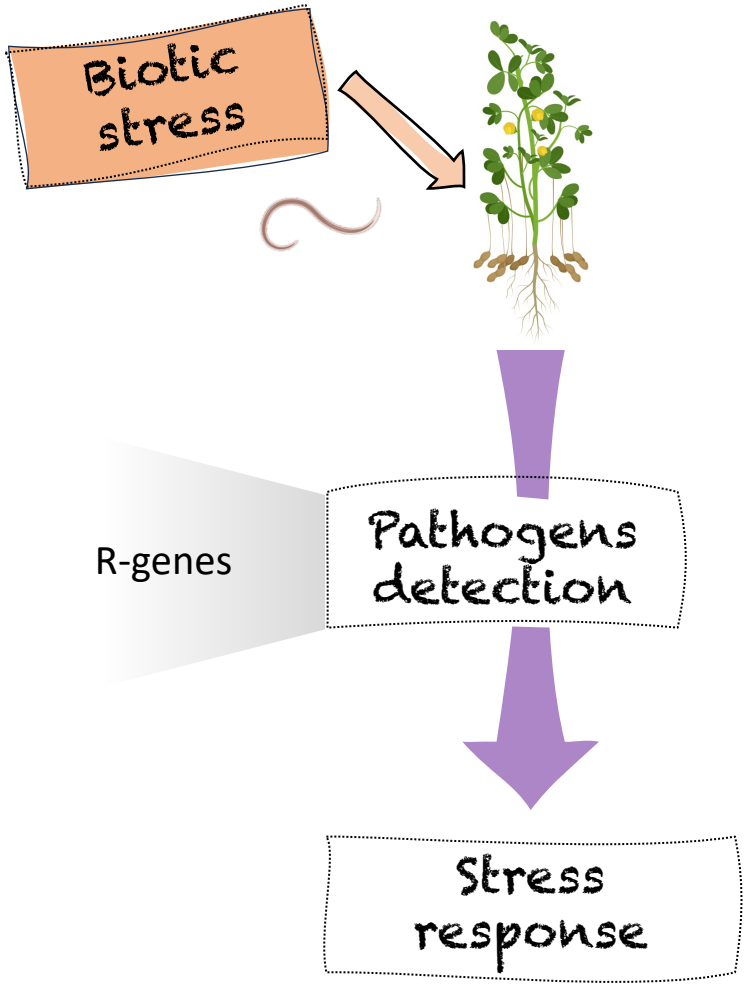


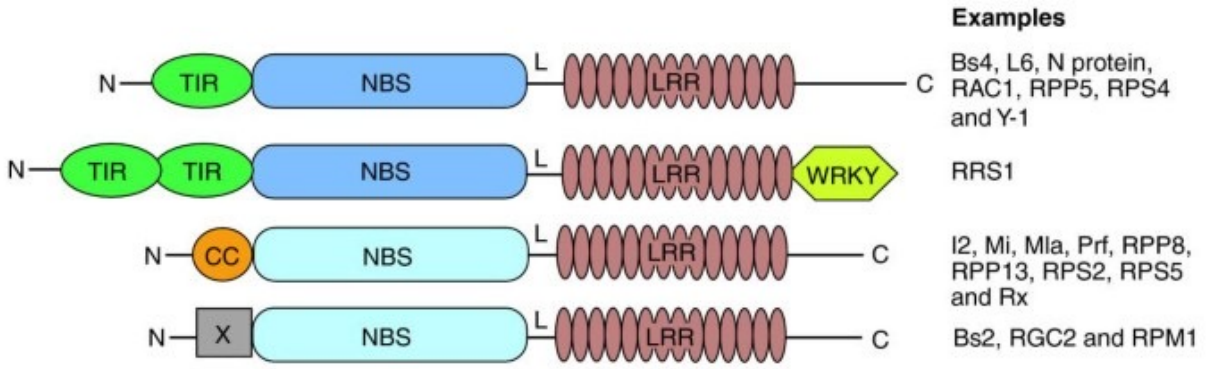
... soon:
LIME & SAGE

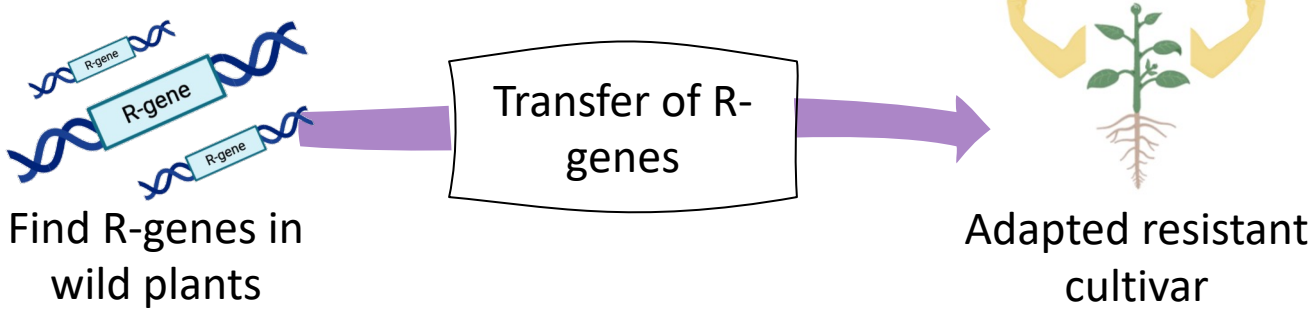
Silvia Bottini, junior professor chair - ISA UMR1355 - INRAE - CNRS - UniCA - silvia.bottini@inrae.fr

Plant NBS-LRR proteins: adaptable guards

Most of disease resistance genes (R-genes) in plants encode nucleotide-binding site leucine-rich repeat (NBS-LRR) proteins.
This large family is encoded by hundreds of diverse genes per genome and can be subdivided into the functionally distinct TIR-domain-containing (TNL) and CC-domain-containing (CNL) subfamilies.







NBS-LRR genes

Zotta-Mota et al. BMC Plant Bio. 2018

NBS-LRR
66

HIVE

44

22

33

33%

Meta-analysis
NBS-LRR
55